经济学、管理学类研究生教学用书

高级计量经济学及Stata应用（第二版）

陈 强 编著



高等教育出版社

统计学、计量经济学主要课程教材

统计学（第四版）（送教师课件）	袁卫 等
统计学习题与案例	袁卫 等
统计学原理（送教师课件）	范秀荣 等
统计学案例分析	苏继伟
统计学（第二版）	邱东
统计学（送教师课件）	马敏娜
统计学（送教师课件）	费宇 等
统计学（送教师课件）	卞毓宁
统计学基础	吴启富
统计学实验——SPSS和R软件应用与实例（送教师课件）	费宇
应用统计学实验教程	吴先华 等
应用时间序列分析（送教师课件）	史代敏
国民经济统计学（第二版）	邱东
应用多元统计分析（送教师课件）	傅德印
应用计量经济学：时间序列分析（第2版）	Walterenders 著 杜江 等译
计量经济学（第三版）（送教师课件）	李子奈 等
计量经济学学习指南与练习	潘文卿 等
计量经济学（送教师课件）	王少平
计量经济学导论（第四版，英文改编）	Jeffrey M. Wooldridge 著，王少平改编
高级计量经济学	洪永淼
高级计量经济学及Stata应用（第二版）（送教师课件）	陈强

ISBN 978-7-04-032983-4

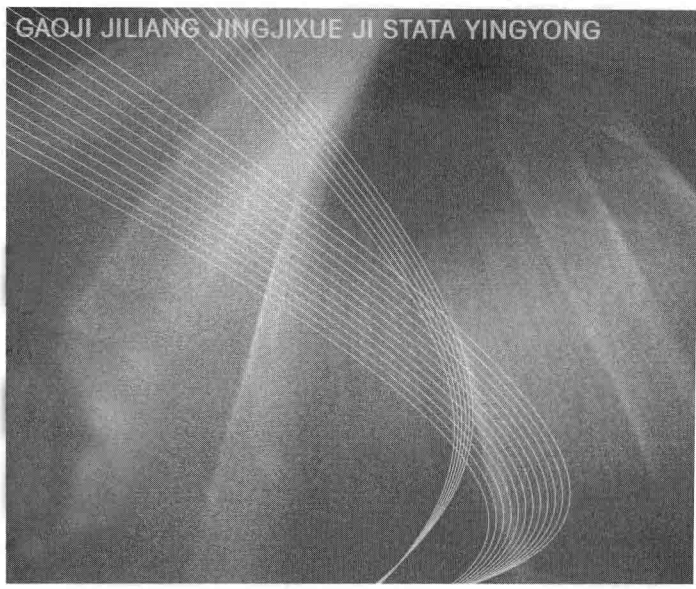


定价 59.00 元

经济学、管理学类

高级计量经济学及Stata应用（第二版）

陈强 编著



高等教育出版社·北京

内容简介

本书较多地借鉴了现代计量经济学的最新发展，内容全面，除了介绍传统的横截面数据外，对面板数据（含长面板、动态面板、非线性面板）、时间序列（含VAR、单位根、协整）、自然实验、重复截面数据、GMM、自助法、蒙特卡罗法、分位数回归、门限回归、非参数估计、处理效应、空间计量、久期分析、贝叶斯估计等均做了较深入的分析。本书力图以生动的语言、较多的插图与经济意义来直观地解释计量方法，而又不失数学的严谨性。同时，结合目前欧美最为流行的Stata计量软件，及时地介绍相应的Stata命令与实例，为读者提供“一站式”服务。

本书适合普通高等学校经济学、管理学类或社科类硕士生、博士生与研究人员使用。为便于读者学习高级计量经济学，本书在内容安排上，假设读者已经学过微积分、线性代数与概率统计，但不要求学过本科阶段的计量经济学（学过更好）。

图书在版编目（CIP）数据

高级计量经济学及Stata应用 / 陈强编著. -- 2版
.-- 北京 : 高等教育出版社，2014.4
ISBN 978-7-04-032983-4

Ⅰ. ①高… Ⅱ. ①陈… Ⅲ. ①计量经济学－高等学校－教材②经济计量分析－应用软件－高等学校－教材
Ⅳ. ①F224.0

中国版本图书馆CIP数据核字(2014)第023987号

策划编辑 施春花　责任编辑 施春花　封面设计 赵 阳　版式设计 马敬茹
插图绘制 尹文军　责任校对 李大鹏　责任印制 张泽业

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京市四季青双青印刷厂
开　　本 787 mm×1092 mm　1/16
印　　张 42.5　插页 1
字　　数 1110千字
购书热线 010-58581118
咨询电话 400-810-0598

网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.landraco.com
　　http://www.landraco.com.cn

版　　次 2010年10月第1版
　　2014年4月第2版
印　　次 2014年4月第1次印刷
定　　价 59.00元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 32983-00

作者简介



陈强，男，1971年出生，山东大学经济学院教授，泰岳经济研究中心副主任（主持工作）。分别于1992年、1995年获北京大学经济学学士、硕士学位，后留校任教。2007年获美国Northern Illinois University数学硕士与经济学博士学位。主要研究领域为发展经济学、计量经济学、经济史与制度经济学。已独立发表论文于*Economica*, *Journal of Comparative Economics*,《经济学（季刊）》、《世界经济》等国内外期刊。曾获中国数量经济学年会、中国制度经济学年会优秀论文奖、山东省高等学校优秀科研成果论文一等奖。现为美国经济学会、英国皇家经济学会会员，*Applied Economics*，《经济学（季刊）》与《产业经济评论》的匿名审稿人。2010年入选教育部新世纪优秀人才支持计划。

第一版
读者反馈

■ 在图书馆比较了很多本书，最后决定买这本，讲解很清楚很好。

——京东读者

■ 这是一本难得的优秀教材，对研究生而言是莫大的幸事，对我们学习经典的计量经济学理论和软件实现都起着不可估量的作用，感谢陈强老师。

——京东读者

■ 这本书虽然面世时间不长，但是字里行间透着作者的严谨与对内容讲解的透彻。即使仅学习它的理论部分都让人受益匪浅。Stata部分讲得也很实用。

——卓越读者

■ 慢慢品读，慢慢演习。先说说优点：

1. 语言顺畅，版面充实。如果按照一般教材比较宽的边距，这本书会很厚；
2. 讲解非常之贴心——时不时地告诉你一些小技巧，语言非常柔和，一改一般教材严苛死板的说辞语气；
3. 慢慢发现。

——卓越读者

■ 本书接轨现代计量经济学，语言生动，例题翔实，尤其适合Stata初学者和提高者。

——当当读者

■ 国内对于深度专题探讨Stata的书很少，所以看到这本就果断买了，不过感觉涉及内容非常多，但是每章还是比较简洁，回去好好研究。

——当当读者

■ 陈老师的书不错，强烈推荐，我出国留学就带了一本书，就是您这本书。

——人大经济论坛读者

■ 这本书真的难得，陈老师是真用心写的，复杂理论通俗化，Stata操作很详细，现在第二遍阅读，有好多收获。强烈推荐，强烈推荐。

——人大经济论坛读者

第二版前言

《高级计量经济学及 Stata 应用》自 2010 年 10 月出版以来，受到广大读者热烈欢迎，在此表示特别感谢。随着时间的推移，尽管网上依旧好评如潮，但第一版的不足越发清晰。当代计量经济学博大精深且发展迅猛，厚爱本书的读者也提出了不少合理化建议，要求增加这样或那样的内容。为此，从 2012 年暑期即着手第二版写作，冬去春来，到第二版初稿完成时，竟又接近济南的盛夏了。

第二版是第一版的重大升级(a significant upgrade)。第二版新增了四章全新内容，即非线性面板(Nonlinear Panels)、处理效应(Treatment Effects)、空间计量经济学(Spatial Econometrics)与久期分析(Duration Analysis)。已有章节也得到不少充实与完善，不胜枚举，部分新增内容参见二级子目录。比如，原来的“离散被解释变量”那一章，由于增加了太多内容，不得不分为三章。另外，Stata 软件也更新到 Stata 12，功能更为强大。

曾国藩云：带勇以能打仗为要。类似地，论文以创新为要，而教材则以易懂为要。为此，第二版秉承第一版的写作风格，在深入浅出、通俗易懂方面痛下功夫。时而感叹何必自苦，顶着海归的科研压力，却将宝贵时间投入此无底洞式的教材写作；但一想到我之费时费力可使得数以千计的学子省时省力，心下也便释然了。

自从 2012 年“千人计划学者”波士顿学院肖志杰教授在山东大学执教以来，给我教益良多；山东大学经济学院王永副教授、韩青博士对第二版新增内容进行了仔细的校对，在此一并感谢(当然文责自负)。第二版的修订得到山东大学研究生精品课程项目的资助，以及高等教育出版社编辑一如既往的支持，在此表示衷心感谢。第二版的错漏与不足之处，依然恳请读者指正。本书用到的所有数据集均可在我的个人网页(www. econ. sdu. edu. cn/tree/Faculty. php)下载，或通过 qiang2chen2@126. com 联系索取。正是您的一贯支持，给予我不断前行的动力。

陈　强

2013 年 10 月

第一版前言

本书是在山东大学经济学院硕士生、博士生《高级计量经济学》教案的基础上编著而成，适合高等学校经济管理类或社科类研究生与研究人员使用。

本书的主要特色如下：

（1）接轨现代计量经济学。本书较多地借鉴了 Angrist and Pischke(2008)，Baum(2006)，Cameron and Trivedi(2005，2010)，Greene(2012)，Hamilton(1994)，Hayashi(2000)，Hsiao(2003)，Kennedy(2003)，Poirier(1995)，Verbeek(2004)，Wooldridge(2010)，其中尤以 Hayashi(2000)对本书的影响最深。

（2）内容全面。除了介绍传统的横截面数据外，本书对面板数据（含长面板、动态面板）、时间序列（含 VAR、单位根、协整）、自然实验、重复截面数据、GMM、蒙特卡罗法、自助法、分位数回归、门限回归、非参数估计、贝叶斯估计等方法均进行了较深入的介绍。

（3）计量理论与软件操作相结合。学习计量的学生，既需要了解计量原理，也需要知道如何在电脑上实现。为此，本书提供了“一站式”服务，在讲解每个估计方法后，随即介绍相应的 Stata 电脑操作及实例（Stata 为目前欧美最为流行的计量软件）。

（4）本书力图以生动的语言、较多的插图与经济意义来直观地解释计量方法，而不仅仅是从数学推导到数学推导；另一方面，又不失数学的严谨性（部分证明放在附录）。

（5）先修课不包括本科水平的计量经济学。在中国的国情下，不少经济类研究生并未学过本科阶段的计量经济学。因此，本书在内容安排上，假设读者已经学过微积分、线性代数与概率统计，但不要求学过本科阶段的计量经济学（当然，如果学过更好）。

学习计量经济学不是一件容易的事（我也经历过，以后还要经历），但回报却很丰厚（其为实证研究不可或缺之工具），可以说是“高投入、高产出”。对于许多初学者而言，或许计量经济学难就难在使用了较多的数学[①]。但数学只是一种语言，而任何数学符号原则上都可以“翻译”为汉语。事实上，看似复杂的数学公式后面，常常有着非常直观的道理。因此，只要渐渐地掌握数学这门语言，学会看数学符号背后的含义，学习计量也就不难了。

套用一个参禅的故事，学习计量大致可以分为三个境界。第一境界是“见山是山，见水是水”，第二境界是“见山不是山，见水不是水”，第三境界是“见山又是山，见水又是水”。在第一阶段，以为计量就是作最小二乘回归而已，自然不在话下。在第二阶段，开始体会到计量的精妙之处，心中时时产生疑问。在第三阶段，通过考前复习及实践应用，对计量的理论与方法逐渐融会贯通，进而内化为熟练掌握的工具。其中，尤以第二阶段最为漫长。“取法乎上，仅得其中”。貌似难懂之处，其实正是取得进步的地方。这是一个“痛并快乐着”的过程，时常伴有顿悟之喜悦。我曾为学生们写了一首打油诗，收录在此，以博一笑。

---

① 对于从理科转学经济学的同学来说，可能面临另一问题，即如何更快地建立经济学的直觉，加深对经济意义的理解。

计量啊计量

辛苦读研学计量，推来导去费思量。
只因成绩盼优良，折腾数据叫爹娘。
爱恨交加为那般，实证研究是桥梁。
此情可待成佳酿，奈何当下心已凉。

在本书出版之际，特别要感激以下曾教授过我统计学或计量经济学的授业恩师们（以时间先后为序）：范培华、胡健颖、靳云汇、陈良焜（北京大学）；Dale Poirier（University of California, Irvine）；Susan Porter-Hudak, Nader Ebrahimi, Mohsen Pourahmadi（Northern Illinois University）。没有他们的谆谆教诲，本书是绝不可能完成的。

山东大学经济学院的领导与同事们对本书的写作给予了大力支持与鼓励，另外，2008 级与 2009 级硕士与博士生在听课过程中提出了很多好建议，在此一并感谢。山东大学经济学院王永老师，博士生韩青，硕士生李欢、李晶、林兴兰、戚传萍、吴振华、张甜等参与了校对，陈丽云同学协助制作了部分插图，在此表示衷心感谢（当然，文责自负）。最后，要特别感谢高等教育出版社的于明编辑、边晓娜编辑及其同仁们，为本书的撰写提出了许多宝贵意见，并付出了辛勤的劳动。

正如苏格拉底所说，学习是学生自我发现的过程，而教师不过是助产婆，但愿这本教科书能起到这个作用。

当然，由于本人知识有限，对于本书中的错误与不足之处，恳请各位老师与同学及时指出，以便在本书的网站上公布勘误表，并在未来的版本中更新。邮箱为 qiang2chen2@126.com。

陈　强

2010 年 2 月于济南

郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》，其行为人将承担相应的民事责任和行政责任；构成犯罪的，将被依法追究刑事责任。为了维护市场秩序，保护读者的合法权益，避免读者误用盗版书造成不良后果，我社将配合行政执法部门和司法机关对违法犯罪的单位和个人进行严厉打击。社会各界人士如发现上述侵权行为，希望及时举报，本社将奖励举报有功人员。

反盗版举报电话　(010) 58581897　58582371　58581879
反盗版举报传真　(010) 82086060
反盗版举报邮箱　dd@hep.com.cn
通信地址　北京市西城区德外大街4号　高等教育出版社法务部
邮政编码　100120

目　录

第 1 章　绪论 …… 1

1.1　什么是计量经济学 …… 1

1.2　经济数据的特点与类型 …… 2

第 2 章　概率统计回顾 …… 3

2.1　概率与条件概率 …… 3

2.2　分布与条件分布 …… 4

2.3　随机变量的数字特征 …… 5

2.4　迭代期望定律 …… 8

2.5　随机变量无关的三个层次概念 …… 9

2.6　常用连续型统计分布 …… 9

2.7　统计推断的思想 …… 11

习题 …… 12

附录 …… 12

第 3 章　小样本 OLS …… 13

3.1　古典线性回归模型的假定 …… 13

3.2　OLS 的代数推导 …… 14

3.3　OLS 的几何解释 …… 17

3.4　拟合优度 …… 17

3.5　OLS 的小样本性质 …… 18

3.6　对单个系数的 t 检验 …… 20

3.7　对线性假设的 F 检验 …… 23

3.8　F 统计量的似然比原理表达式 …… 25

3.9　分块回归与偏回归(选读) …… 26

3.10　预测 …… 27

习题 …… 28

附录 …… 29

第 4 章　Stata 简介 …… 30

4.1　为什么使用 Stata …… 30

4.2　Stata 的窗口 …… 30

4.3　Stata 操作实例 …… 31

4.4　Stata 命令库的更新 …… 46

4.5　进一步学习 Stata 的资源 …… 47

习题 …… 48

第 5 章　大样本 OLS …… 49

5.1　为何需要大样本理论 …… 49

5.2　随机收敛 …… 49

5.3　大数定律与中心极限定理 …… 51

5.4　统计量的大样本性质 …… 52

5.5　渐近分布的推导 …… 53

5.6　随机过程的性质 …… 53

5.7　大样本 OLS 的假定 …… 57

5.8　OLS 的大样本性质 …… 58

5.9　线性假设的大样本检验 …… 60

5.10　大样本 OLS 的 Stata 命令及实例 …… 61

习题 …… 63

附录 …… 63

第 6 章　最大似然估计法 …… 66

6.1　最大似然估计法的定义 …… 66

6.2　线性回归模型的最大似然估计 …… 68

6.3　最大似然估计的数值解 …… 69

6.4　信息矩阵与无偏估计的最小方差 …… 70

6.5　最大似然法的大样本性质 …… 71

6.6　最大似然估计量的渐近协方差矩阵 …… 74

6.7　三类渐近等价的统计检验 …… 75

6.8　准最大似然估计法 …… 78

6.9　对正态分布假设的检验 …… 80

6.10　最大似然估计法的 Stata 命令及实例 …… 80

习题 …… 84

附录 …… 84

第 7 章　异方差与 GLS …… 87

7.1　异方差的后果 …… 87

7.2　异方差的例子 …… 87

7.3　异方差的检验 …… 88

7.4　异方差的处理 …… 90

7.5　处理异方差的 Stata 命令及实例 …… 93

7.6　Stata 命令的批处理 …… 96

习题 …… 98

附录 …… 98

第 8 章　自相关 …… 100

8.1　自相关的后果 …… 100

8.2 自相关的例子 …… 101
8.3 自相关的检验 …… 101
8.4 自相关的处理 …… 103
8.5 处理自相关的 Stata 命令及实例 …… 108
习题 …… 115
第 9 章 模型设定与数据问题 …… 116
9.1 遗漏变量 …… 116
9.2 无关变量 …… 117
9.3 建模策略:“由小到大”还是“由大到小” …… 118
9.4 解释变量个数的选择 …… 118
9.5 对函数形式的检验 …… 120
9.6 多重共线性 …… 123
9.7 极端数据 …… 124
9.8 虚拟变量 …… 126
9.9 经济结构变动的检验 …… 127
9.10 缺失数据与线性插值 …… 132
9.11 变量单位的选择 …… 133
习题 …… 133
附录 …… 133
第 10 章 工具变量,2SLS 与 GMM …… 135
10.1 解释变量与扰动项相关的例子 …… 135
10.2 工具变量法作为一种矩估计 …… 138
10.3 二阶段最小二乘法 …… 140
10.4 有关工具变量的检验 …… 141
10.5 GMM 的假定 …… 146
10.6 GMM 的推导 …… 147
10.7 GMM 的大样本性质 …… 148
10.8 如何获得工具变量 …… 151
10.9 MLE 也是 GMM …… 152
10.10 工具变量法的 Stata 命令及实例 …… 153
习题 …… 167
附录 …… 167
第 11 章 二值选择模型 …… 169
11.1 离散被解释变量的例子 …… 169
11.2 二值选择模型 …… 169
11.3 二值选择模型的微观基础 …… 177
11.4 二值选择模型中的异方差问题 …… 178
11.5 稀有事件偏差(选读) …… 179
11.6 含内生变量的 Probit 模型(选读) …… 183
11.7 双变量 Probit 模型(选读) …… 187
11.8 部分可观测的双变量 Probit 模型(选读) …… 189
习题 …… 190
第 12 章 多值选择模型 …… 192
12.1 多项 Logit 与多项 Probit …… 192
12.2 条件 Logit 模型 …… 193
12.3 混合 Logit 模型 …… 193
12.4 嵌套 Logit …… 205
习题 …… 208
第 13 章 排序与计数模型 …… 209
13.1 排序模型 …… 209
13.2 泊松回归 …… 211
13.3 负二项回归 …… 213
13.4 零膨胀泊松回归与负二项回归 …… 215
13.5 计数模型的 Stata 实例 …… 215
习题 …… 222
第 14 章 受限被解释变量 …… 223
14.1 断尾回归 …… 223
14.2 零断尾泊松回归与负二项回归 …… 226
14.3 随机前沿模型(选读) …… 228
14.4 偶然断尾与样本选择 …… 235
14.5 归并回归 …… 238
14.6 归并数据的两部分模型 …… 243
14.7 含内生解释变量的 Tobit 模型(选读) …… 246
习题 …… 248
附录 …… 248
第 15 章 短面板 …… 250
15.1 面板数据的特点 …… 250
15.2 面板数据的估计策略 …… 251
15.3 混合回归 …… 252
15.4 个体固定效应模型 …… 252
15.5 时间固定效应 …… 253
15.6 一阶差分法 …… 254
15.7 随机效应模型 …… 254
15.8 组间估计量 …… 255
15.9 拟合优度的度量 …… 255
15.10 非平衡面板 …… 256
15.11 究竟该用固定效应还是随机效应模型 …… 257
15.12 个体时间趋势 …… 257
15.13 短面板的 Stata 命令及实例 …… 258

习题 …… 271
第 16 章 长面板与动态面板…… 272
16.1 长面板的估计策略 …… 272
16.2 面板校正标准误 …… 272
16.3 仅解决组内自相关的 FGLS …… 274
16.4 全面 FGLS …… 278
16.5 组间异方差的检验 …… 279
16.6 组内自相关的检验 …… 280
16.7 组间同期相关的检验 …… 281
16.8 变系数模型 …… 283
16.9 面板工具变量法 …… 287
16.10 豪斯曼 - 泰勒估计量(选读) …… 288
16.11 动态面板 …… 289
16.12 动态面板的 Stata 命令及实例 …… 291
16.13 偏差校正 LSDV 法 …… 300
16.14 重复截面数据与组群分析 …… 301
习题 …… 302
第 17 章 非线性面板…… 303
17.1 面板二值选择模型 …… 303
17.2 面板二值选择模型的随机效应估计 …… 304
17.3 面板二值选择模型的固定效应估计 …… 305
17.4 面板二值选择模型的 Stata 实例 …… 307
17.5 面板泊松回归 …… 313
17.6 面板负二项回归 …… 314
17.7 面板计数模型的 Stata 实例 …… 315
17.8 面板 Tobit …… 325
17.9 面板随机前沿模型 …… 327
习题 …… 332
第 18 章 随机实验与自然实验…… 334
18.1 实验数据 …… 334
18.2 理想的随机实验 …… 335
18.3 引入更多的解释变量 …… 335
18.4 随机实验执行过程中可能出现的问题 …… 336
18.5 自然实验 …… 337
18.6 双重差分法 …… 339
18.7 三重差分法 …… 343
18.8 观测数据的处理效应 …… 344
习题 …… 345
第 19 章 蒙特卡罗法与自助法…… 346
19.1 蒙特卡罗法的思想与用途 …… 346
19.2 蒙特卡罗法实例:模拟中心极限定理 …… 347
19.3 蒙特卡罗法实例:服从卡方分布的扰动项 …… 348
19.4 蒙特卡罗积分 …… 349
19.5 最大模拟似然法与模拟矩估计 …… 350
19.6 自助法的思想与用途 …… 351
19.7 自助法的分类 …… 352
19.8 使用自助法估计标准误 …… 352
19.9 使用自助法进行区间估计 …… 353
19.10 使用自助法进行假设检验 …… 353
19.11 自助法的一致性(选读) …… 354
19.12 异方差情况下的自助法 …… 354
19.13 面板数据与时间序列的自助法 …… 355
19.14 自助法的 Stata 命令 …… 355
19.15 使用自助法进行稳健的豪斯曼检验 …… 356
习题 …… 358
附录 …… 358
第 20 章 平稳时间序列…… 361
20.1 时间序列的数字特征 …… 361
20.2 自回归模型 …… 362
20.3 移动平均模型 …… 364
20.4 ARMA …… 364
20.5 自回归分布滞后模型 …… 365
20.6 ARMA 模型的 Stata 命令及实例 …… 366
20.7 误差修正模型 …… 371
20.8 MA(∞)与滞后算子 …… 372
20.9 向量自回归过程 …… 375
20.10 VAR 的脉冲响应函数 …… 377
20.11 预测误差的方差分解 …… 380
20.12 格兰杰因果检验 …… 381
20.13 面板格兰杰因果检验 …… 381
20.14 VAR 的 Stata 命令及实例 …… 381
20.15 季节调整 …… 399
习题 …… 407
第 21 章 单位根与协整…… 409
21.1 非平稳序列 …… 409
21.2 ARMA 的平稳性 …… 410
21.3 VAR 的平稳性 …… 411
21.4 单位根所带来的问题 …… 411

21.5 单位根检验与平稳性检验 …… 414
21.6 单位根检验的 Stata 实例 …… 418
21.7 面板单位根检验 …… 422
21.8 协整的思想与初步检验 …… 432
21.9 Beveridge-Nelson 分解公式 …… 433
21.10 协整的定义与最大似然估计 …… 434
21.11 协整分析的 Stata 实例 …… 437
习题 …… 445
附录 …… 445
第 22 章 自回归条件异方差模型 …… 447
22.1 条件异方差模型的例子 …… 447
22.2 ARCH 模型的性质 …… 448
22.3 ARCH 模型的 MLE 估计 …… 449
22.4 GARCH 模型 …… 450
22.5 何时使用 ARCH 或 GARCH 模型 …… 451
22.6 ARCH 与 GARCH 模型的扩展 …… 451
22.7 ARCH 与 GARCH 的 Stata 命令及实例 …… 453
22.8 多维 GARCH 模型(选读) …… 460
习题 …… 467
第 23 章 似不相关回归 …… 468
23.1 单一方程估计与系统估计 …… 468
23.2 似不相关回归的假定 …… 468
23.3 SUR 的 FGLS 估计 …… 470
23.4 SUR 的假设检验 …… 471
23.5 似不相关回归的 Stata 命令及实例 …… 471
23.6 变系数面板数据的 SUR 估计 …… 475
习题 …… 478
附录 …… 479
第 24 章 联立方程模型 …… 482
24.1 联立方程模型的结构式与简化式 …… 482
24.2 联立方程模型的识别 …… 483
24.3 单一方程估计法 …… 486
24.4 三阶段最小二乘法 …… 487
24.5 三阶段最小二乘法的 Stata 实例 …… 489
24.6 结构 VAR …… 493
24.7 SVAR 的 Stata 实例 …… 496
习题 …… 502
第 25 章 非线性回归与门限回归 …… 503
25.1 非线性最小二乘法 …… 503
25.2 非线性回归的 Stata 命令及实例 …… 504
25.3 门限回归 …… 505
25.4 面板数据的门限回归 …… 507
25.5 门限回归的计算机操作 …… 508
习题 …… 508
第 26 章 分位数回归 …… 509
26.1 为什么需要分位数回归 …… 509
26.2 总体分位数 …… 509
26.3 样本分位数 …… 510
26.4 分位数回归的估计方法 …… 512
26.5 分位数回归的 Stata 命令及实例 …… 513
习题 …… 517
第 27 章 非参数与半参数估计 …… 518
27.1 为什么需要非参数与半参数估计 …… 518
27.2 对密度函数的非参数估计 …… 518
27.3 核密度估计的性质 …… 520
27.4 最优带宽 …… 521
27.5 多元密度函数的核估计 …… 523
27.6 非参数核回归 …… 523
27.7 多元核回归 …… 525
27.8 k 近邻回归 …… 525
27.9 局部线性回归 …… 526
27.10 非参数估计的 Stata 命令及实例 …… 526
27.11 半参数估计 …… 530
习题 …… 533
附录 …… 533
第 28 章 处理效应 …… 537
28.1 处理效应与选择难题 …… 537
28.2 通过随机分组解决选择难题 …… 539
28.3 依可测变量选择 …… 539
28.4 匹配估计量的思想 …… 540
28.5 倾向得分匹配 …… 542
28.6 倾向得分匹配的 Stata 实例 …… 545
28.7 偏差校正匹配估计量 …… 555
28.8 双重差分倾向得分匹配 …… 557
28.9 断点回归的思想 …… 559
28.10 精确断点回归 …… 561
28.11 模糊断点回归 …… 563
28.12 断点回归的 Stata 实例 …… 565

28.13　处理效应模型 …… 570
习题 …… 574
第 29 章　空间计量经济学 …… 575
29.1　地理学第一定律 …… 575
29.2　空间权重矩阵 …… 575
29.3　空间自相关 …… 578
29.4　空间自回归模型 …… 583
29.5　空间杜宾模型 …… 586
29.6　空间误差模型 …… 586
29.7　一般的空间计量模型 …… 589
29.8　含内生解释变量的 SARAR 模型 …… 592
29.9　空间面板模型 …… 593
29.10　空间计量方法的局限性 …… 598
第 30 章　久期分析 …… 599
30.1　久期数据的处理方法 …… 599
30.2　风险函数 …… 599
30.3　久期数据的归并问题 …… 601
30.4　描述性分析 …… 602
30.5　久期模型的最大似然估计 …… 603
30.6　比例风险模型 …… 604
30.7　加速失效时间模型 …… 606
30.8　Cox 模型 …… 607
30.9　比例风险模型的设定检验 …… 610
30.10　分层 Cox 模型 …… 611
30.11　随时间而变的解释变量 …… 612
30.12　不可观测的异质性 …… 613
30.13　其他久期分析模型 …… 614
30.14　久期分析的 Stata 命令及实例 …… 615
习题 …… 630
第 31 章　贝叶斯估计简介 …… 631
31.1　贝叶斯估计的思想 …… 631
31.2　贝叶斯定理 …… 631
31.3　贝叶斯估计的一个例子 …… 632
31.4　基于后验分布的统计推断 …… 634
31.5　先验分布的选择 …… 635
31.6　多元回归的贝叶斯分析 …… 637
31.7　马尔可夫链蒙特卡罗法 …… 639
习题 …… 640
第 32 章　如何做规范的实证研究 …… 641
32.1　计量理论与现实数据 …… 641
32.2　实证研究的主要步骤 …… 642
32.3　实证论文的结构 …… 644
32.4　计量实践的十诫 …… 645
32.5　结束语 …… 646
习题 …… 646
附录：常用数据来源 …… 647
参考书目 …… 649
数学符号 …… 664
英文缩写 …… 666

第1章 绪 论

1.1 什么是计量经济学

顾名思义,“计量经济学”(Econometrics,也译为“经济计量学”)就是运用概率统计的方法对经济变量之间的(因果)关系进行定量分析的科学。之所以把“因果”两个字加括号,是因为计量经济学常常不足以确定经济变量之间的因果关系(由于实验数据的缺乏)[①],另一方面,大多数实证分析的目的恰恰正是要确定变量之间的因果关系(即是否 X 导致 Y),而非仅仅是相关关系[②]。因此,在学习与应用计量经济学的过程中,很有必要时时以“因果关系”作为思考的框架与指引。

比如,你看到街上人们带伞,于是预测今天要下雨。这是一种相关关系。然而,“人们带伞”并不是造成“下雨”的原因。因此,计量分析必须建立在经济理论的基础之上。然而,即使有理论基础,因果关系依然不好分辨。首先,可能存在“逆向因果关系”(reverse causality)。比如,FDI(外商直接投资)能促进经济增长,但也可能是 FDI 被吸引到增长潜力高的国家或地区。其次,也可能是被遗漏的第三个变量(Z)对这两个变量(X,Y)同时产生了作用,参见图 1.1。

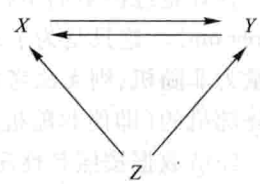


图 1.1 可能的因果关系

作为一个例子,考虑决定教育投资回报率(returns to schooling)的因素:

$$\ln W_i = \alpha + \beta S_i + \varepsilon_i \tag{1.1}$$

其中,$\ln W$(工资收入的自然对数)为“被解释变量”(dependent variable),S(教育年限)为“解释变量”(explanatory variable 或 regressor)、“自变量”(independent variable)或“协变量”(covariate),ε 为“随机扰动项”(stochastic disturbance)或“误差项”(error term),而下标 i 表示第 i 个观测值(即个体 i)。

如果用数据估计这个简单的一元回归,其结果一般会显示,对数工资收入与受教育年限显著正相关,而且教育投资回报率 β 还挺高。然而,一个人的工资收入也与能力有关,但能力不能直接观测,而能力高的人通常选择接受更多教育。因此,在这个简单的回归中,教育的高回报率其实包含了对能力的回报。

另外,影响工资收入的因素还可能包括工作经验、毕业学校、人种、性别、外貌等。因此,需要引入更多的“控制变量”(control variables),也就是多元回归的方法,才能较准确地估计我们“感兴趣的参数”(parameters of interest),即本例中的教育投资回报率 β。然而,现实中总有某些相关

---

① 计量经济学家 Guy Orcutt 曾说过,“做计量经济学就像试图通过播收音机来研究电的规律”(Doing econometrics is like trying to learn the laws of electricity by playing the radio),可见其难度。

② 如果使用计量经济学做预测,则只需要相关关系,不必顾及因果关系。

的变量无法观测，即存在“遗漏变量”（omitted variables），而这些遗漏变量统统被纳入随机扰动项 ε_i 中了。

随机扰动项 ε_i 中还可能包含哪些因素呢？如果真实模型（true model）为

$$\ln W_i = \alpha + \beta S_i + \gamma S_i^2 + \varepsilon_i \tag{1.2}$$

那么 γS_i^2 也被纳入扰动项中了（可以视为广义的遗漏变量）。如果变量测量得不准确，则测量误差也被放入扰动项中了。总之，扰动项就像是一个“垃圾桶”，所有你不想要、无法把握的东西都往里面扔。另一方面，我们又希望扰动项有很好的性质。在很多情况下，这是自相矛盾的。西方有个谚语“The devil is in the details”，意即“魔鬼就在细节中”[①]。套用到计量经济学上来，或许可以说“The devil is in the error term”，意即魔鬼就在扰动项中。计量经济学的很多玄妙之处就在于扰动项。如果真正理解了扰动项，也就加深了对计量经济学的理解。

1.2 经济数据的特点与类型

由于在经济学中通常无法像自然科学那样做“控制实验”（controlled experiment），故经济数据一般不是“实验数据”（experimental data）[②]，而是自然发生的“观测数据”（observational data）。由于个人行为的随机性，所有经济变量原则上都是随机变量[③]。

在计量经济学的本科课程中，为了简单起见，有时假设解释变量是非随机的、固定的（fixed regressors）。这只是为了教学法上的方便，却给更深入的理论探讨带来了不便。比如，如果解释变量为非随机，则无法考虑其与扰动项的相关性。因此，在这本研究生水平的教材中，所有变量都是随机的（即便非随机的常数，也可以视为退化的随机变量）。

经济数据按照其性质，可大致分成以下三种类型。

（1）横截面数据（cross-sectional data，简称截面数据）：指的是多个经济个体的变量在同一时点上的取值。比如，2012 年中国各省的 GDP。

（2）时间序列数据（time series data）：指的是某个经济个体的变量在不同时点上的取值。比如，在 1978—2012 年山东省每年的 GDP。

（3）面板数据（panel data）：指的是多个经济个体的变量在不同时点上的取值。比如，在 1978—2012 年中国各省每年的 GDP。

本书介绍的计量经济理论将包括以上三种数据类型，并使用国际上最为流行的 Stata 计量软件。在此之前，我们首先要对概率统计进行简短的回顾，并引入一些新概念（比如，均值独立、迭代期望定律）。

---

① 比如，对于一份冠冕堂皇的合同，可能让你以后吃尽苦头的正是那些合同附录中的小字部分。

② 第 18 章将讨论随机实验与自然实验数据。

③ 你能举出哪些经济数据（变量）不是随机变量吗？

第 2 章　概率统计回顾

2.1　概率与条件概率

1. 概率

假如街上有个老太太问你："什么是概率？"那么你会怎么回答呢？若回答"事情发生的可能性"，老太太可能反问你："说'可能性'不就行了，为什么又造了一个新词'概率'？"也许她会问你一个更具体的问题："天气预报说明天 70% 概率下雨。这是啥意思？"也许你想说，这表明"明天 70% 的时间会下雨"，但更好的答案则是："如果有 100 天的天气预报都报了 70% 的概率明天降雨，则大约有 70 天会下雨。"

总之，可以将"概率"理解为在大量重复实验下，事件发生的频率趋向的某个稳定值。记事件"下雨"为 A，其发生的"概率"(probability)为 $\mathrm{P}(A)$[①]。

2. 条件概率

例　已知明天会出太阳，下雨的概率有多大？

记事件"出太阳"为 B，则在出太阳的前提条件下降雨的"条件概率"(conditional probability)为

$$\mathrm{P}(A\mid B)\equiv\frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)} \tag{2.1}$$

其中，"∩"表示事件的交集(intersection)，故 $\mathrm{P}(A\cap B)$ 为"太阳雨"的概率，参见图 2.1。条件概率是计量经济学的重要概念之一。

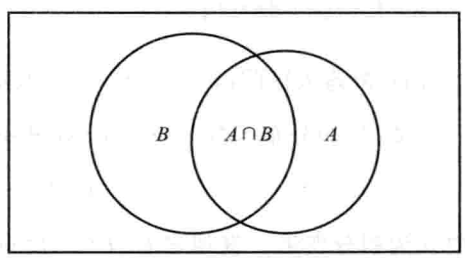


图 2.1　条件概率示意图

例　股市崩盘的可能性为无条件概率；而在已知经济陷入严重衰退的情况下，股市崩盘的可能性则为条件概率。

3. 独立事件

如果条件概率等于无条件概率，即 $\mathrm{P}(A\mid B)=\mathrm{P}(A)$，即 B 是否发生不影响 A 的发生，则称 A，B 为相互独立的随机事件。此时，$\mathrm{P}(A\mid B)\equiv\frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}=\mathrm{P}(A)$，故

$$\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B) \tag{2.2}$$

也可以将此式作为独立事件的定义。

---

① 这里不讨论概率的严格公理化定义。

4. 全概率公式

如果事件组 $\{B_1, B_2, \cdots, B_n\}$ $(n \geq 2)$ 两两互不相容，$P(B_i) > 0$ $(\forall i = 1, \cdots, n)$，且 $B_1 \cup B_2 \cup \cdots \cup B_n$ 为必然事件（即在 $B_1, B_2, \cdots, B_n$ 中必然有某个 B_i 发生，“$\cup$”表示事件的并集，union），则对任何事件 A 都有（无论 A 与 $\{B_1, B_2, \cdots, B_n\}$ 是否有任何关系），

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(B_i)\mathrm{P}(A \mid B_i) \tag{2.3}$$

全概率公式把世界分成了 n 个可能的情形，再把每种情况下的条件概率“加权平均”而汇总成无条件概率（权重为每种情形发生的概率）。该公式有助于理解后面的迭代期望定律。

2.2 分布与条件分布

1. 离散型概率分布

假设随机变量 X 的可能取值为 $\{x_1, x_2, \cdots, x_k, \cdots\}$，其对应的概率为 $\{p_1, p_2, \cdots, p_k, \cdots\}$，即 $p_k \equiv \mathrm{P}(X = x_k)$，则称 X 为离散型随机变量，其分布律可以表示为

$$\begin{array}{cccccc} X & x_1 & x_2 & \cdots & x_k & \cdots \\ p & p_1 & p_2 & \cdots & p_k & \cdots \end{array} \tag{2.4}$$

其中，$p_k \geq 0$，$\sum_k p_k = 1$。常见的离散分布有“两点分布”（Bernoulli）、“二项分布”（Binomial）、“泊松分布”（Poisson）与“负二项分布”（Negative Binomial）等（参见标准的概率统计本科教材）。

2. 连续型概率分布

连续型随机变量可以取任意实数，其“概率密度函数”（probability density function，简记 pdf）$f(x)$ 满足，

（i）$f(x) \geq 0, \forall x$；

（ii）$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$；

（iii）X 落入区间 $[a, b]$ 的概率为 $P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$。

定义“累积分布函数”（cumulative distribution function，简记 cdf）为

$$F(x) \equiv \mathrm{P}(-\infty < X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \tag{2.5}$$

其中，t 为积分变量。直观来看，$F(x)$ 度量的是，从 $-\infty$ 至 x 为止，概念密度函数 $f(t)$ 曲线下的面积。

3. 多维随机向量的概率分布

为了研究经济变量之间的关系，常需要同时考虑两个或多个随机变量，即“随机向量”（random vector）。二维连续型随机向量 (X, Y) 的“联合密度函数”（joint pdf）$f(x, y)$ 满足：

（i）$f(x, y) \geq 0, \forall x, y$；

（ii）$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}x\mathrm{d}y = 1$；

（iii）(X, Y) 落入平面某区域 D 的概率为 $\mathrm{P}\{(X, Y) \in D\} = \iint_D f(x, y)\,\mathrm{d}x\mathrm{d}y$。

直观来看，二维随机向量的联合密度函数就像是一顶倒扣的草帽，而落入平面某区域 D 的

概率就是这顶草帽下在区域 D 之上的体积。更一般地，n 维连续型随机向量 $(X_1, X_2, \cdots, X_n)$ 可以由联合密度函数 $f(x_1, x_2, \cdots, x_n)$ 来描述。

从二维联合密度函数 $f(x,y)$，可以计算 X 的（一维）边缘密度函数（marginal pdf）：

$$f_x(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y \tag{2.6}$$

这个公式的直观含义与“全概率公式”相似，即给定 x，把所有 y 取值的可能性都“加”起来（积分的本质就是加总）。类似地，可以计算 Y 的（一维）边缘密度函数：

$$f_y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x \tag{2.7}$$

定义二维随机向量 (X,Y) 的累积分布函数为

$$F(x,y) \equiv \mathrm{P}(-\infty < X \leqslant x; -\infty < Y \leqslant y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t,s)\,\mathrm{d}t\mathrm{d}s \tag{2.8}$$

4. 条件分布

“条件分布”（conditional distribution）的概念对于计量经济学至关重要。考虑在 $X=x$ 条件下 Y 的条件分布，记为 $Y|X=x$。然而，如果 X 为连续型随机变量，事件 $\{X=x\}$ 发生的概率为 0，该如何计算 $Y|X=x$ 的“条件概率密度”（conditional pdf）呢？为此，考虑 $X \in [x-\varepsilon, x+\varepsilon]$，然后让 $\varepsilon \to 0^+$，可以证明条件密度函数为（参见附录）

$$f(y|x) = \frac{f(x,y)}{f_x(x)} \tag{2.9}$$

直观上，这个公式与条件概率的公式十分类似。

2.3 随机变量的数字特征

虽然随机变量的密度函数或累积分布函数能够完整地描述随机变量，但我们常希望用少数几个常数来刻画其主要特征（称为“数字特征”），比如该随机变量平均来说在什么位置（即平均值，或集中趋势），波动幅度（即离散趋势）有多大，与其他变量是否存在“协动”（即相关性）。为此，引入期望、方差、协方差、相关系数、原点矩、中心矩、偏度、峰度、协方差矩阵、条件期望、条件方差等概念。

定义 对于分布律为 $p_k \equiv \mathrm{P}(X=x_k)$ 的离散型随机变量 X，其“**期望**”（expectation）[①]为

$$\mathrm{E}(X) \equiv \mu \equiv \sum_{k=1}^{\infty} x_k p_k \tag{2.10}$$

由上式可知，期望的直观含义就是加权平均，权重即为概率。

定义 对于概率密度函数为 $f(x)$ 的连续型随机变量 X，其“**期望**”为

$$\mathrm{E}(X) \equiv \mu \equiv \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \tag{2.11}$$

容易证明，对随机变量求期望的这种运算（即“期望算子”，expectation operator）满足“线性性”（linearity），即 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$，$\mathrm{E}(kX) = k\mathrm{E}(X)$，其中 k 为任意常数。

定义 随机变量 X 的“**方差**”（variance）为

$$\mathrm{Var}(X) \equiv \sigma^2 \equiv \mathrm{E}[X - \mathrm{E}(X)]^2 \tag{2.12}$$

---

① 也称为“数学期望”（mathematical expectation）或“均值”（mean）。

方差越大则随机变量取值的波动幅度越大。称方差的算术平方根为“**标准差**”(standard deviation),通常记为 σ。在计算方差时,常使用以下简便公式。

命题 $\mathrm{Var}(X)=\mathrm{E}[X^2]-[\mathrm{E}(X)]^2$

证明: $\mathrm{Var}(X)\equiv\mathrm{E}[X-\mathrm{E}(X)]^2=\mathrm{E}\{X^2-2\mathrm{E}(X)X+[\mathrm{E}(X)]^2\}$

$$=\mathrm{E}[X^2]-2[\mathrm{E}(X)]^2+[\mathrm{E}(X)]^2=\mathrm{E}[X^2]-[\mathrm{E}(X)]^2$$

我们常需要考虑两个变量之间的相关性,即一个随机变量的取值会对另一随机变量的取值有多大影响。

定义 随机变量 X 与 Y 的“**协方差**”(covariance)为

$$\mathrm{Cov}(X,Y)\equiv\sigma_{XY}\equiv\mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] \tag{2.13}$$

如果当随机变量 X 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时,随机变量 Y 的取值也倾向于大于(小于)其期望值 $\mathrm{E}(Y)$,则 $\mathrm{Cov}(X,Y)>0$,二者存在正相关;反之,如果当随机变量 X 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时,随机变量 Y 的取值反而倾向于小于(大于)其期望值 $\mathrm{E}(Y)$,则 $\mathrm{Cov}(X,Y)<0$,二者存在负相关。如果 $\mathrm{Cov}(X,Y)=0$,则说明二者“线性不相关”,但不一定“相互独立”,因为二者还可能存在非线性的相关关系。在计算协方差时,常使用以下简便公式:

$$\begin{aligned}\mathrm{Cov}(X,Y)&\equiv\mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))]\\&=\mathrm{E}[XY-X\mathrm{E}(Y)-\mathrm{E}(X)Y+\mathrm{E}(X)\mathrm{E}(Y)]\\&=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)\end{aligned} \tag{2.14}$$

协方差的运算也满足线性性,可以证明(参见习题)

$$\mathrm{Cov}(X,Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z) \tag{2.15}$$

协方差的一个缺点是,它受 X 与 Y 的计量单位的影响。为了将其标准化,引入相关系数的定义。

定义 随机变量 X 与 Y 的“**相关系数**”(correlation)为

$$\rho\equiv\mathrm{Corr}(X,Y)\equiv\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{2.16}$$

可以证明,$-1\leqslant\rho\leqslant1$。需要注意的是,如果以上各定义式中的积分不收敛,则随机变量的数字特征可能不存在。比如,自由度为 1 的 t 分布变量,其期望与方差都不存在。更一般地,对于随机变量 X,可以定义一系列的数字特征,即各阶“矩”(moment)的概念。

定义 一阶原点矩为 $\mathrm{E}(X)$(即期望),二阶原点矩为 $\mathrm{E}(X^2)$,三阶原点矩为 $\mathrm{E}(X^3)$,四阶原点矩为 $\mathrm{E}(X^4)$,等等。

定义 二阶中心矩为 $\mathrm{E}[X-\mathrm{E}(X)]^2$(即方差),三阶中心矩为 $\mathrm{E}[X-\mathrm{E}(X)]^3$,四阶中心矩为 $\mathrm{E}[X-\mathrm{E}(X)]^4$,等等。

其中,一阶原点矩(期望)表示随机变量的平均值,二阶中心矩(方差)表示随机变量的波动程度,三阶中心矩表示随机变量密度函数的不对称性(偏度),而四阶中心矩表示随机变量密度函数的最高处(山峰)有多“尖”及尾部有多“厚”(峰度)。然而,三、四阶中心矩还取决于变量的单位。为此,首先将变量“标准化”(即减去其期望 μ,再除以其标准差 σ),并引入以下定义。

定义 随机变量 X 的“**偏度**”(skewness)为 $\mathrm{E}[(X-\mu)/\sigma]^3$。

显然,如果随机变量为对称分布(比如,正态分布),则其偏度为 0(奇函数在关于原点对称的区间上积分为 0)。

定义 随机变量 X 的“**峰度**”(kurtosis)为 $\mathrm{E}[(X-\mu)/\sigma]^4$。

对于正态分布,其峰度为 3。如果随机变量 X 的峰度大于 3(比如 t 分布),则其密度函数的

最高处(山峰)比正态分布更“尖”,而两侧尾部则更“厚”(参见2.6节的图2.3)。

定义 随机变量 X 的“**超额峰度**”(excess kurtosis)为 $\mathrm{E}[(X-\mu)/\sigma]^4-3$。

第6章将使用正态分布的偏度与峰度性质来检验某个分布是否为正态分布。

更一般地,对于任意函数 $g(\cdot)$,称随机变量函数的期望 $\mathrm{E}[g(X)]=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x$ 为“**矩**”(moment)。

定义 “**条件期望**”(conditional expectation)就是条件分布 $Y|x$ 的期望,即

$$\mathrm{E}(Y|X=x)\equiv\mathrm{E}(Y|x)=\int_{-\infty}^{+\infty}yf(y|x)\mathrm{d}y \tag{2.17}$$

在上式中,由于 y 已被积分积掉,故 $\mathrm{E}(Y|x)$ 只是 x 的函数,参见图2.2。

定义 “**条件方差**”(conditional variance)就是条件分布 $Y|x$ 的方差,即

$$\begin{aligned}\mathrm{Var}(Y|X=x)&\equiv\mathrm{Var}(Y|x)\\&=\int_{-\infty}^{+\infty}[y-\mathrm{E}(Y|x)]^2f(y|x)\mathrm{d}y\end{aligned} \tag{2.18}$$

同样地,在上式中,y 已被积分积掉,故 $\mathrm{Var}(Y|x)$ 只是 x 的函数,参见图2.2。

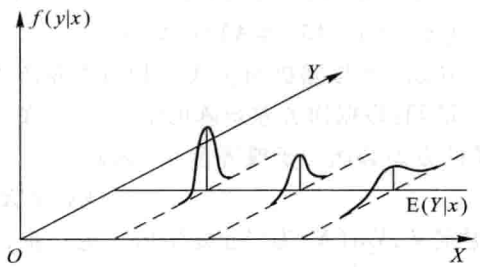


图2.2 条件期望与条件方差示意图

为了引入随机向量的数字特征,首先需要回顾关于矩阵半正定与正定的概念。

定义 对于 $n\times n$ 对称矩阵 $\boldsymbol{A}$(symmetric),如果对于任意 n 维非零列向量 $\boldsymbol{c}$,都有二次型 $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}\geqslant 0$,则称 $\boldsymbol{A}$ 为**半正定矩阵**。

定义 对于 $n\times n$ 对称矩阵 $\boldsymbol{A}$,如果对于任意 n 维非零列向量 $\boldsymbol{c}$,都有二次型 $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}>0$,则称 $\boldsymbol{A}$ 为**正定矩阵**。

根据线性代数知识,正定矩阵的行列式一定不等于0,故其逆矩阵一定存在。从几何意义上看,对于正定矩阵,可以通过坐标变换变为一个主对角线上元素全部为正数的对角矩阵(特征值全部为正)。特别地,在一维的情形下,正定矩阵就相当于正数。

类似地,可以定义半负定与负定矩阵。

命题 对于任意矩阵 $\boldsymbol{D}$,$\boldsymbol{D}'\boldsymbol{D}$ 为半正定矩阵。

证明:首先,由于 $\boldsymbol{D}'\boldsymbol{D}=(\boldsymbol{D}'\boldsymbol{D})'$,故 $\boldsymbol{D}'\boldsymbol{D}$ 为对称矩阵。不失一般性,假设 $\boldsymbol{D}'\boldsymbol{D}$ 为 n 阶矩阵。其次,对于任意 n 维非零列向量 $\boldsymbol{c}$,二次型

$$\boldsymbol{c}'(\boldsymbol{D}'\boldsymbol{D})\boldsymbol{c}=(\boldsymbol{c}'\boldsymbol{D}')(\boldsymbol{D}\boldsymbol{c})=\underbrace{(\boldsymbol{D}\boldsymbol{c})'\boldsymbol{D}\boldsymbol{c}}_{\text{平方和}}\geqslant 0 \tag{2.19}$$

因此,$\boldsymbol{D}'\boldsymbol{D}$ 为半正定矩阵。

定义 设 $\boldsymbol{X}=(X_1X_2\cdots X_n)'$ 为 n 维随机向量,则其“**协方差矩阵**”(covariance matrix)为 $n\times n$ 的对称半正定矩阵:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{X})&\equiv\mathrm{Var}(\boldsymbol{X})\equiv\mathrm{E}[(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))']\\&=\mathrm{E}\left[\begin{pmatrix}X_1-\mathrm{E}(X_1)\\\vdots\\X_n-\mathrm{E}(X_n)\end{pmatrix}(X_1-\mathrm{E}(X_1)\quad\cdots\quad X_n-\mathrm{E}(X_n))\right]\end{aligned}$$

$$= \mathrm{E}\begin{pmatrix} [X_1 - \mathrm{E}(X_1)]^2 & \cdots & [X_1 - \mathrm{E}(X_1)][X_n - \mathrm{E}(X_n)] \\ \vdots & & \vdots \\ [X_1 - \mathrm{E}(X_1)][X_n - \mathrm{E}(X_n)] & \cdots & [X_n - \mathrm{E}(X_n)]^2 \end{pmatrix} \tag{2.20}$$

$$= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$$

其中，主对角线元素 $\sigma_{ii} \equiv \mathrm{Var}(X_i)$，而非主对角线元素 $\sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j)$。

假设 $\boldsymbol{A}$ 为 $m \times n$ 常数矩阵（不含随机变量），可以证明以下重要性质（参见习题）：

(i) $\mathrm{E}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{X})$；　　（期望算子的线性性）

(ii) $\mathrm{Var}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{XX}') - \mathrm{E}(\boldsymbol{X})[\mathrm{E}(\boldsymbol{X})]'$；　　（一维随机变量公式的推广）

(iii) $\mathrm{Var}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}'$。

命题 n 维随机向量 $\boldsymbol{X}$ 的协方差矩阵 $\mathrm{Var}(\boldsymbol{X})$ 为半正定矩阵。

证明：根据协方差矩阵的定义，$\mathrm{Var}(\boldsymbol{X})$ 为 $n \times n$ 对称矩阵。对于 n 维非零列向量 $\boldsymbol{c}$，随机变量 $\boldsymbol{c}'\boldsymbol{X}$ 的方差必然大于或等于 0。因此，

$$\mathrm{Var}(\boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{c}'\mathrm{Var}(\boldsymbol{X})\boldsymbol{c} \geqslant 0 \tag{2.21}$$

根据定义，$\mathrm{Var}(\boldsymbol{X})$ 为半正定矩阵。进一步，如果希望任意线性组合 $\boldsymbol{c}'\boldsymbol{X}$ 的方差为正数（不考虑方差为 0 的退化情形），则一般假设 $\mathrm{Var}(\boldsymbol{X})$ 为正定矩阵。

对于两个随机向量，也可以类似地考虑两个随机向量之间的协方差。

定义 设 $\boldsymbol{X} = (X_1 X_2 \cdots X_n)'$ 为 n 维随机向量，$\boldsymbol{Y} = (Y_1 Y_2 \cdots Y_m)'$ 为 m 维随机向量，则这两个随机向量之间的协方差矩阵为

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) \equiv \mathrm{E}[(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))(\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y}))'] = \mathrm{E}(\boldsymbol{XY}') - \mathrm{E}(\boldsymbol{X})\mathrm{E}(\boldsymbol{Y}') \tag{2.22}$$

2.4 迭代期望定律

定理（迭代期望定律） 对于条件期望的运算，有以下重要的“迭代期望定律”（Law of iterated expectation），

$$\mathrm{E}(Y) = \mathrm{E}_X[\mathrm{E}(Y|x)] \tag{2.23}$$

上式表明，无条件期望 $\mathrm{E}(Y)$ 等于，对于给定 $X = x$ 情况下 Y 的条件期望 $\mathrm{E}(Y|x)$ 再对 X 求期望。下面以连续型变量为例来证明。

证明：$\mathrm{LHS} = \mathrm{E}(Y) = \int_{-\infty}^{+\infty} y f_y(y)\,\mathrm{d}y$

$$\mathrm{RHS} = \mathrm{E}_X\left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_x(x)}\mathrm{d}y\right] = \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_x(x)}\mathrm{d}y\right] f_x(x)\,\mathrm{d}x$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x,y)\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{+\infty} y\left[\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\right]\mathrm{d}y = \mathrm{LHS}$$

其中，LHS 指“方程左边”（Left Hand Side），而 RHS 指“方程右边”（Right Hand Side）。从上面的证明可以看出，迭代期望定律很像全概率公式。直观来看，无条件期望等于条件期望之加权平均，而权重为条件“$X = x$”的概率密度（取值可能性）。在离散随机变量的情形下，可以看得更为清楚：

$$\mathrm{E}(Y)=\sum_{x_i}\mathrm{P}(X=x_i)\mathrm{E}(Y|x_i) \tag{2.24}$$

推而广之,对于任意函数 $g(\cdot)$,可以得到

$$\mathrm{E}[g(Y)]=\mathrm{E}_X\mathrm{E}[g(Y)|x] \tag{2.25}$$

有时期望算子 E_X 的下标被省去,此时需注意对什么变量求期望。

2.5 随机变量无关的三个层次概念

定义 对于连续型随机变量 X 与 Y,如果其联合密度等于边缘密度的乘积,即 $f(x,y)=f_x(x)f_y(y)$,则称 X 与 Y 相互独立。

直观来看,如果 X 与 Y 相互独立,则 X 的取值不对 Y 的取值产生任何影响,反之亦然。这是有关随机变量"无关"的最强概念。线性不相关的概念则更弱,仅要求协方差为 0,即 $\mathrm{Cov}(X,Y)=0$。显然,"相互独立"意味着"线性不相关",但反之不然。事实上,在二者之间还有一个中间层次的无关概念,即"均值独立"(mean-independent),在计量经济学中很有用。

定义 假设条件期望 $\mathrm{E}(Y|x)$ 存在。如果 $\mathrm{E}(Y|x)$ 不依赖于 X,则称"Y 均值独立于 X"(Y is mean-independent of X)。

注意:均值独立不是一种对称的关系,即"Y 均值独立于 X"并不意味着"X 均值独立于 Y"。

命题 "Y 均值独立于 X"当且仅当 $\mathrm{E}(Y|x)=\mathrm{E}(Y)$(即条件期望等于无条件期望)。

证明:(1) 假设"Y 均值独立于 X",则 $\mathrm{E}(Y|x)$ 不依赖于 X,故 $\mathrm{E}_X[\mathrm{E}(Y|x)]=\mathrm{E}(Y|x)$。根据迭代期望定律,$\mathrm{E}(Y)=\mathrm{E}_X[\mathrm{E}(Y|x)]=\mathrm{E}(Y|x)$。

(2) 假设 $\mathrm{E}(Y|x)=\mathrm{E}(Y)$,则显然 $\mathrm{E}(Y|x)$ 不依赖于 X。

命题 如果 X 与 Y 相互独立,则 Y 均值独立于 X,且 X 均值独立于 Y。

证明:X 与 Y 相互独立意味着,X 与 Y 没有任何关系,故条件期望 $\mathrm{E}(Y|x)$ 也不依赖于 X。证明参见习题。

定理(均值独立意味着不相关) 如果 Y 均值独立于 X 或 X 均值独立于 Y,则 $\mathrm{Cov}(X,Y)=0$。

证明:

$$\begin{aligned}
\mathrm{Cov}(X,Y)&=\mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] && \text{(协方差的定义)}\\
&=\mathrm{E}_X\mathrm{E}_Y[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))|x] && \text{(迭代期望定律)}\\
&=\mathrm{E}_X[(X-\mathrm{E}(X))\mathrm{E}_Y(Y-\mathrm{E}(Y)|x)] && \text{(将 } X-\mathrm{E}(X)\text{ 视为常数提出)}\\
&=\mathrm{E}_X[(X-\mathrm{E}(X))(\mathrm{E}(Y|x)-\mathrm{E}(Y))] && \text{(期望算子的线性性)}\\
&=\mathrm{E}_X[(X-\mathrm{E}(X))\cdot 0]=0 && \text{(均值独立的定义)}
\end{aligned}$$

总之,"相互独立"⇒"均值独立"⇒"线性不相关"。

2.6 常用连续型统计分布

在计量经济学中常用的连续型统计分布包括正态分布、χ^2 分布、t 分布与 F 分布等。

1. 正态分布

如果随机变量 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \tag{2.26}$$

则称 X 服从正态分布，记为 $X \sim N(\mu, \sigma^2)$，其中 μ 为期望，而 σ^2 为方差。将 X 进行标准化，定义 $Z \equiv \frac{X-\mu}{\sigma}$，则 Z 服从标准正态分布，记为 $Z \sim N(0,1)$。标准正态分布的概率密度以过原点的垂线为对称轴，呈钟形(bell-shaped)(参见图 2.3)，通常记为 $\phi(x)$；其累积分布函数则记为 $\Phi(x)$。据说，当高斯(Gauss)发现标准正态分布的“核”(kernel)$\exp\{-x^2/2\}$时欣喜若狂，因为全世界只有一个标准正态分布，正态分布因此也叫“高斯分布”(Gaussian distribution)。

如果 n 维随机向量 $\boldsymbol{X} = (X_1 X_2 \cdots X_n)'$ 的联合密度函数为

$$f(x_1, \cdots, x_n) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \right\} \tag{2.27}$$

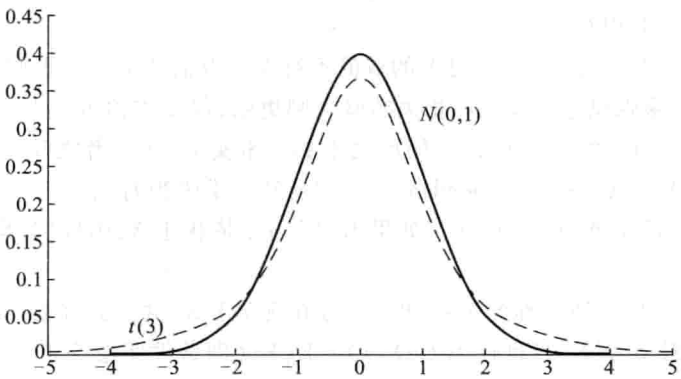


图 2.3 $N(0,1)$ 与 $t(3)$ 的概率密度

则称 $\boldsymbol{X}$ 服从期望为 $\boldsymbol{\mu}$、协方差矩阵为 $\boldsymbol{\Sigma}$ 的 n 维正态分布，记为 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。正态分布具有良好的性质。比如，多维正态的每个分量都是正态，其分量的任意线性组合仍然是正态。反之，每个分量均为一维正态并不足以保证其联合分布也是多维正态的。

命题 对于 n 个随机变量 $X_1, X_2, \cdots, X_n$，如果任意线性组合 $k_1 X_1 + \cdots + k_n X_n$(其中 $k_1, \cdots, k_n$ 不全为 0)都服从一维正态分布，则 $(X_1 X_2 \cdots X_n)$ 服从 n 维正态分布。

利用这个性质，可以证明如下命题。

命题 如果 $(X_1 X_2 \cdots X_n)$ 服从 n 维正态分布，设 $Y_1, \cdots, Y_m$ 分别是 $(X_1 X_2 \cdots X_n)$ 的线性函数，则 $(Y_1 \cdots Y_m)$ 也服从多维正态分布。

证明：考虑 $Y_1, \cdots, Y_m$ 的任意线性组合 $k_1 Y_1 + \cdots + k_m Y_m$，其中 $k_1, \cdots, k_m$ 不全为 0。由于 $Y_1, \cdots, Y_m$ 分别是 $(X_1 X_2 \cdots X_n)$ 的线性函数，故 $k_1 Y_1 + \cdots + k_m Y_m$ 也是 $(X_1 X_2 \cdots X_n)$ 的线性函数。由于多维正态的线性组合仍为正态，故 $k_1 Y_1 + \cdots + k_m Y_m$ 服从一维正态分布。根据 $k_1, \cdots, k_m$ 的任意性可知，$(Y_1 \cdots Y_m)$ 服从多维正态分布。

另外，如果 $(X_1 X_2 \cdots X_n)$ 服从 n 维正态分布，则“$X_1, X_2, \cdots, X_n$ 相互独立”与“$X_1, X_2, \cdots, X_n$ 两两不相关”等价。换言之，对于正态分布，“不相关”就意味着“相互独立”。利用这个性质，有时可以很容易地证明正态变量的独立性。

2. χ^2 分布(卡方分布，Chi-square)

如果 $Z \sim N(0,1)$，则 $Z^2 \sim \chi^2(1)$，即自由度为 1 的 χ^2 分布。如果 $\{Z_1, \cdots, Z_k\}$ 为独立同分布的标准正态分布，则其平方和满足自由度为 k 的 χ^2 分布

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k) \tag{2.28}$$

其中，参数 k 为“自由度”(degree of freedom)，表示它由 k 个相互独立(自由)的随机变量构成。

由于 χ^2 分布来自标准正态的平方和，故其取值只能为正数，参见图 2.4。

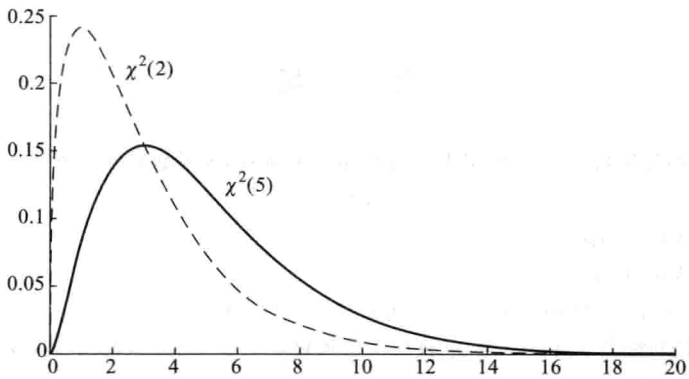


图 2.4 χ^2 分布的概率密度

3. t 分布

假设 $Z \sim N(0,1)$，$Y \sim \chi^2(k)$，而且 Z 与 Y 相互独立，则 $\dfrac{Z}{\sqrt{Y/k}} \sim t(k)$，其中 k 为自由度。t 分布也以过原点的垂线为对称，但与标准正态分布相比，中间的“山峰”更低（但更尖），且两侧有“厚尾”（fat tails），参见图 2.3。当自由度 $k \to \infty$ 时，t 分布收敛于标准正态分布。

4. F 分布

假设 $Y_1 \sim \chi^2(k_1)$，$Y_2 \sim \chi^2(k_2)$，而且 Y_1, Y_2 相互独立，则 $\dfrac{Y_1/k_1}{Y_2/k_2} \sim F(k_1,k_2)$。$F$ 分布的取值也只能为正数，其概率密度的形状与 χ^2 分布相似。如果 $X \sim t(k)$，则可以证明 $X^2 \sim F(1,k)$（参见本章习题）。

2.7 统计推断的思想

计量经济学所使用的主要方法是数理统计中的“统计推断”（statistical inference），其基本思想如下。将我们感兴趣的研究对象的全体称为“总体”（population），其中每个研究对象称为“个体”（individual）。由于总体包含的个体可能很多，进行普查的成本可能很高，故常需要从总体中抽取部分个体，称为“样本”（sample），而样本中包含个体的数目称为“样本容量”（sample size），参见图 2.5。

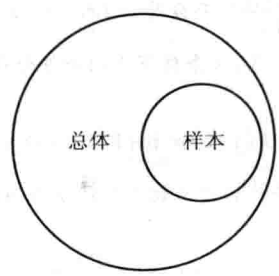


图 2.5 总体与样本

通常希望样本为“随机样本”（random sample），即总体中的每个个体都有相同的概率被抽中，而且被抽中的概率相互独立，即满足“独立同分布”（independently identically distributed，简记 iid）的假定。

由于样本来自总体，必然带有总体的信息。而统计推断就是根据样本数据对总体性质进行推断的科学。统计推断的主要形式有参数估计（含点估计与区间估计）、假设检验及预测等。

注意:对于随机样本,通常被抽中的个体散布于总体的各处,而不像图 2.5 中那样"样本数据"都集中在一起。

习 题

2.1 假设 $\boldsymbol{X}$ 为 n 维随机列向量,其期望为 $\mathrm{E}(\boldsymbol{X})=\boldsymbol{\mu}$。$\boldsymbol{A}$ 为 $m\times n$ 常数矩阵。证明:

(1) $\mathrm{E}(\boldsymbol{AX})=\boldsymbol{A\mu}$;

(2) $\mathrm{Var}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{XX}')-\boldsymbol{\mu\mu}'$;

(3) $\mathrm{Var}(\boldsymbol{AX})=\boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}'$。

2.2 假设 X,Y,Z 分别为一维随机变量,证明:$\mathrm{Cov}(X,Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z)$。

2.3 假设连续型随机变量 X,Y 相互独立,证明 Y 均值独立于 X。(提示:使用随机变量相互独立的定义。)

2.4 假设 $X\sim t(k)$,证明 $X^2\sim F(1,k)$。

2.5 (选做题)假设 $\boldsymbol{X}$ 为 $n\times K$ 常数矩阵,$n>K$,且矩阵 $\boldsymbol{X}$ 满列秩,证明 $\boldsymbol{X'X}$ 为正定矩阵。(提示:基于 $\boldsymbol{X'X}$ 半正定,使用反证法进一步证明其为正定。)

附 录

A2.1 证明条件密度函数为 $f(y\mid x)=\dfrac{f(x,y)}{f_x(x)}$。

证明:记二维连续型随机向量 (X,Y) 的联合密度函数为 $f(x,y)$,累积分布函数为 $F(x,y)$,随机变量 X 的边缘密度函数为 $f_x(x)$,边缘分布函数为 $F_x(x)$。

为了求解在 $X=x$ 条件下 Y 的条件密度函数,首先考虑在 $X\in[x-\varepsilon,x+\varepsilon]$ 的条件下,Y 的条件分布函数,参见图 2.6。

$$
\begin{aligned}
&\mathrm{P}\{Y\leqslant y\mid X\in[x-\varepsilon,x+\varepsilon]\}\\
&=\frac{\mathrm{P}\{Y\leqslant y,X\in[x-\varepsilon,x+\varepsilon]\}}{\mathrm{P}\{X\in[x-\varepsilon,x+\varepsilon]\}} && \text{(根据条件概率的定义)}\\
&=\frac{F(x+\varepsilon,y)-F(x-\varepsilon,y)}{F_x(x+\varepsilon)-F_x(x-\varepsilon)} && \text{(根据分布函数的性质)}\\
&=\frac{[F(x+\varepsilon,y)-F(x-\varepsilon,y)]/2\varepsilon}{[F_x(x+\varepsilon)-F_x(x-\varepsilon)]/2\varepsilon} && \text{(分子分母同除以 } 2\varepsilon\text{)}
\end{aligned}
$$

若使 $\varepsilon\to0^+$,则分子为 $\lim\limits_{\varepsilon\to0^+}[F(x+\varepsilon,y)-F(x-\varepsilon,y)]/2\varepsilon=\dfrac{\partial F(x,y)}{\partial x}$,而分母为 $\lim\limits_{\varepsilon\to0^+}[F_x(x+\varepsilon)-F_x(x-\varepsilon)]/2\varepsilon=f_x(x)$。

故在 $X=x$ 条件下 Y 的条件分布函数为

$$
F(y\mid x)\equiv\lim_{\varepsilon\to0^+}\mathrm{P}\{Y\leqslant y\mid X\in[x-\varepsilon,x+\varepsilon]\}=\frac{\dfrac{\partial F(x,y)}{\partial x}}{f_x(x)}
$$

而条件密度函数为条件分布函数的导数,故

$$
f(y\mid x)=\frac{\partial F(y\mid x)}{\partial y}=\frac{\dfrac{\partial^2F(x,y)}{\partial x\partial y}}{f_x(x)}=\frac{f(x,y)}{f_x(x)}
$$

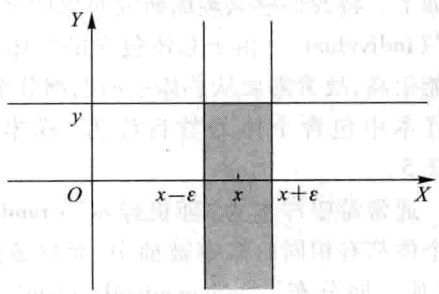


图 2.6 条件密度函数的计算

第 3 章　小样本 OLS

3.1　古典线性回归模型的假定

“最小二乘法”(Ordinary Least Square,简记 OLS)是单一方程线性回归模型最常见、最基本的估计方法。“古典线性回归模型”(Classical Linear Regression Model,简记 CLRM)的假定如下。

假定 3.1　线性假定(linearity)。总体(population)模型为

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{3.1}$$

其中,n 为样本容量,解释变量 x_{ik} 的第一个下标表示第 i 个“观测值”(observation),而第二个下标则表示第 k 个解释变量($k = 1, \cdots, K$),共有 K 个解释变量。如果有常数项,则通常令第一个解释变量为单位向量,即 $x_{i1} \equiv 1, \forall i$。$\beta_1, \beta_2, \cdots, \beta_K$ 均为待估参数,被称为“回归系数”(regression coefficients)。线性假设的含义是每个解释变量 x_{ik} 对被解释变量 y_i 的边际效应均为常数,比如 $\dfrac{\partial \mathrm{E}(y_i)}{\partial x_{i1}} = \beta_1$。如果认为某解释变量的边际效应是可变的,则可以加入平方项(x_{ik}^2)、三次方项(x_{ik}^3),或交互项($x_{ik}x_{im}$)。比如,教育投资的回报率可能随着教育程度的提高而递减①。如果回归方程为 $y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma x_{ik} x_{im} + \varepsilon_i$,则 x_{ik} 对 y_i 的平均边际效应为 $\dfrac{\mathrm{E}(y_i)}{x_{ik}} = \beta_k + \gamma x_m$,它随着 x_{im} 取值的变化而不同。此时,只要把高次项也作为解释变量来看待,则依然满足线性假定。

总体模型也称为“数据生成过程”(Data Generating Process,简记 DGP)。为了更简洁地表达,下面引入矩阵符号。把方程(3.1)的所有解释变量和参数都写成向量,记第 i 个观测数据为 $\boldsymbol{x}_i \equiv (x_{i1}\ x_{i2} \cdots\ x_{iK})'$,$\boldsymbol{\beta} \equiv (\beta_1\ \beta_2 \cdots\ \beta_K)'$,则方程(3.1)为

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{3.2}$$

把所有观测($i = 1, \cdots, n$)所对应的方程叠放(stack)在一起可得

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3.3}$$

定义 $\boldsymbol{y} \equiv (y_1\ y_2\ \cdots\ y_n)'$,数据矩阵 $\boldsymbol{X} \equiv (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n)'$,$\boldsymbol{\varepsilon} \equiv (\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n)'$,则

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4}$$

当看到矩阵方程(3.2),(3.3),(3.4)时,应学会观想其背后对应的代数方程。

假定 3.2　严格外生性(strict exogeneity)

$$\mathrm{E}(\varepsilon_i \mid \boldsymbol{X}) = \mathrm{E}(\varepsilon_i \mid \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = 0 \quad (i = 1, \cdots, n) \tag{3.5}$$

---

① 或许读研的回报率比读本科更低?希望不是如此。

即在给定数据矩阵 $\boldsymbol{X}$ 的情况下，扰动项 ε_i 的条件期望为 0。这意味着，ε_i 必须均值独立于(mean-independent)所有解释变量的观测数据，而不仅仅是同一解释变量 $\boldsymbol{x}_i$ 中的观测数据。根据均值独立的性质，ε_i 与所有解释变量都不相关，即 $\text{Cov}(\varepsilon_i, x_{jk}) = 0, \forall j,k$。这是一个很强的假定，但在第 5 章的大样本 OLS 理论中可以减弱。

事实上，均值独立仅要求 $\text{E}(\varepsilon_i|\boldsymbol{X}) = c$，其中 c 为某常数，但不一定为 0。但当回归方程中有常数项时，要求 $\text{E}(\varepsilon_i|\boldsymbol{X}) = 0$ 并不会带来过多限制，因为如果 $\text{E}(\varepsilon_i|\boldsymbol{X}) = c \neq 0$，总可以把扰动项的非零期望 c 归入常数项中，即只要定义新的扰动项为 $(\varepsilon_i - c)$ 就可以满足严格外生性。

命题 $\text{E}(\varepsilon_i) = 0$，即扰动项的无条件期望为 0。

证明：根据迭代期望定律，$\text{E}(\varepsilon_i) = \text{E}_{\boldsymbol{X}}\underbrace{E(\varepsilon_i|\boldsymbol{X})}_{=0} = \text{E}_{\boldsymbol{X}}(0) = 0$。

定义 如果随机变量 X, Y 满足 $\text{E}(XY) = 0$，则称 X, Y“**正交**”(orthogonal)①。

命题 解释变量与扰动项正交。

证明：$0 = \text{Cov}(x_{jk}, \varepsilon_i) = \text{E}(x_{jk}\varepsilon_i) - \underbrace{\text{E}(x_{jk})\text{E}(\varepsilon_i)}_{=0} = \text{E}(x_{jk}\varepsilon_i)$。

假定 3.3 不存在“严格多重共线性”(strict multicolinearity)，即数据矩阵 $\boldsymbol{X}$ 满列秩，$\text{rank}(\boldsymbol{X}) = K$，其中“rank”表示矩阵的秩。

如果不满足此条件，则 $\boldsymbol{\beta}$“不可识别”(unidentified)，因为 $\boldsymbol{X}$ 中某个或多个变量为多余。在后面将看到，根据 OLS 估计，$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$。如果 $\boldsymbol{X}$ 满列秩，则对称矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 为正定矩阵(参见第 2 章习题)，故 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在；反之，则 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在。在实际数据中，一般不容易出现严格多重共线性的问题，除非你设了过多的“虚拟变量”(参见第 9 章)，同时在回归方程中又包括常数项。即使如此，Stata 软件也会自动识别，并删去多余的解释变量。

假定 3.4 球型扰动项(spherical disturbance)，即扰动项满足“同方差”、“无自相关”的性质，

$$\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix} \tag{3.6}$$

其中，$\boldsymbol{I}_n$ 为 n 阶单位矩阵。之所以称为“球型扰动项”，因为扰动项的协方差矩阵与单位矩阵 $\boldsymbol{I}_n$ 成正比。一方面，在球型扰动项的假设下，协方差矩阵 $\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})$ 的主对角线元素都等于 σ^2，即满足“条件同方差”(conditional homoskedasticity)；如果不完全相等，则存在“条件异方差”(conditional heteroskedasticity)。另一方面，在球型扰动项的假设下，协方差矩阵 $\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})$ 的非主对角线元素都等于 0，即不同个体的扰动项之间没有“自相关”(autocorrelation)②；反之，则存在自相关。

3.2 OLS 的代数推导

可以从两个角度来看待方程(3.1)中的被解释变量 y_i 与解释变量 $\{x_{i1}, x_{i2}, \cdots, x_{iK}\}$：在抽样之前(事前)，可以看作是随机变量；而在抽样之后(事后)，则又代表该随机变量的实现值，可以

① 这与线性代数中“向量正交”的概念不同，后者指的是常数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的内积为 0，即 $\boldsymbol{x}'\boldsymbol{y} = 0$。

② 因为是扰动项自己与自己的相关性，故称为“自相关”。

用其进行代数计算。

为了估计未知参数向量 $\boldsymbol{\beta}$，对于 $\boldsymbol{\beta}$ 的一个任意假想值（hypothetical value）$\tilde{\boldsymbol{\beta}}$（读为 beta tilde），记第 i 个数据的拟合误差（即残差，residual）为 $e_i = y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}$。类比方程（3.4）可知，残差向量 $\boldsymbol{e} \equiv (e_1 \quad e_2 \quad \cdots \quad e_n)' = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$。最小二乘法寻找能使残差平方和（Sum of Squared Residuals，简记 SSR）$\sum_{i=1}^{n} e_i^2$ 最小的 $\tilde{\boldsymbol{\beta}}$。从几何上来看，如果是一元回归，就是要寻找最佳拟合的回归直线，使观测值 y_i 到该回归直线的距离平方和最小；如果是二元回归，就是要寻找最佳拟合的回归平面；如果是更多元的回归，则是寻找最佳拟合的回归超平面（superplane）。可以将此最小化问题写为

$$
\begin{aligned}
\min_{\tilde{\boldsymbol{\beta}}} \mathrm{SSR}(\tilde{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} && \text{（将平方和写成向量内积的形式）}\\
&= (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) && \text{（残差向量的表达式）}\\
&= (\boldsymbol{y}' - \tilde{\boldsymbol{\beta}}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) && \text{（矩阵转置的运算性质①）}\\
&= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} && \text{（乘积展开）}\\
&= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}} && \text{（合并同类项）}
\end{aligned}
$$

其中，$(\boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{\beta}})' = \tilde{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$（对称矩阵），且二者均为 1×1 常数（由目标函数 $\sum_{i=1}^{n} e_i^2$ 为常数可知），故二者相等，可以合并为 $2\boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}$。目标函数 $\mathrm{SSR}(\tilde{\boldsymbol{\beta}})$ 实际上是 $\tilde{\boldsymbol{\beta}}$ 的二次函数（二次型），参见图 3.1。

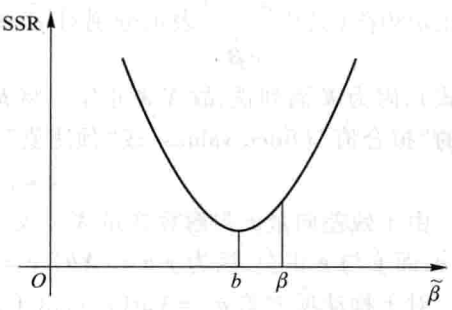


图 3.1 参数的假想值 $\tilde{\boldsymbol{\beta}}$、真实值 $\boldsymbol{\beta}$ 与 OLS 估计值 $\boldsymbol{b}$

为了对向量 $\tilde{\boldsymbol{\beta}}$ 求导，首先介绍以下两个向量微分的规则。假设列向量 $\boldsymbol{a} = (a_1 \quad a_2 \quad \cdots \quad a_K)'$，则 $\boldsymbol{a}'\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{K} a_i\tilde{\beta}_i$，故

$$
\frac{\partial(\boldsymbol{a}'\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} \equiv \left(\frac{\partial(\boldsymbol{a}'\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_1} \quad \frac{\partial(\boldsymbol{a}'\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_2} \cdots \frac{\partial(\boldsymbol{a}'\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_K}\right)' = (a_1 a_2 \cdots a_K)' = \boldsymbol{a} \tag{3.7}
$$

这个规则类似于对一次函数求导。从上式可知，对向量 $\tilde{\boldsymbol{\beta}}$ 求导，就是分别对 $\tilde{\boldsymbol{\beta}}$ 的每个分量求偏导，然后再把这些偏导排成列向量的形式（故向量微分并不神秘！）。假设 $\boldsymbol{A}$ 为 K 阶对称矩阵，同理可证（但较烦琐）：

$$
\frac{\partial(\tilde{\boldsymbol{\beta}}'\boldsymbol{A}\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} \equiv \left(\frac{\partial(\tilde{\boldsymbol{\beta}}'\boldsymbol{A}\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_1} \quad \frac{\partial(\tilde{\boldsymbol{\beta}}'\boldsymbol{A}\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_2} \cdots \frac{\partial(\tilde{\boldsymbol{\beta}}'\boldsymbol{A}\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_K}\right)' = 2\boldsymbol{A}\tilde{\boldsymbol{\beta}} \tag{3.8}
$$

这个规则类似于对二次函数求导。使用这两个向量微分规则，可得最小化的一阶条件：

---

① 即 $(\boldsymbol{A}+\boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$，以及 $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}'\boldsymbol{A}'$。

$$\frac{\partial(\mathrm{SSR})}{\partial\tilde{\boldsymbol{\beta}}}=-2\boldsymbol{X}'\boldsymbol{y}+2\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}=0 \tag{3.9}$$

其中，由于 $\boldsymbol{y}'\boldsymbol{y}$ 不包含 $\tilde{\boldsymbol{\beta}}$（相当于常数），故在微分时消去。通过移项可知，最小二乘估计量 $\boldsymbol{b}$ 满足

$(\boldsymbol{X}'\boldsymbol{X})_{K\times K}\boldsymbol{b}_{K\times 1}=\boldsymbol{X}'_{K\times n}\boldsymbol{y}_{n\times 1}$ （称为“正规方程组”，含 K 个方程，K 个未知数）

$\boldsymbol{X}'\boldsymbol{y}-(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b}=\boldsymbol{0}$ （移项）

$\boldsymbol{X}'\underbrace{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})}_{=\boldsymbol{e}}=\boldsymbol{0}$ （向左提取共同的矩阵因子 $\boldsymbol{X}'$）

因此，$\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0}$，其中残差向量 $\boldsymbol{e}\equiv\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\equiv\boldsymbol{y}-\hat{\boldsymbol{y}}$。残差向量 $\boldsymbol{e}$ 与解释变量 $\boldsymbol{X}$ 正交，这是 OLS 的一大特征。最后求解可得 OLS 估计量：

$$\boldsymbol{b}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.10}$$

二阶条件要求黑赛矩阵（Hessian）$\dfrac{\partial^2(\mathrm{SSR})}{\partial\tilde{\boldsymbol{\beta}}\,\partial\tilde{\boldsymbol{\beta}}'}\equiv\dfrac{\partial\left(\dfrac{\partial\mathrm{SSR}}{\partial\tilde{\boldsymbol{\beta}}}\right)}{\partial\tilde{\boldsymbol{\beta}}'}\equiv\begin{pmatrix}\dfrac{\partial^2\mathrm{SSR}}{\partial^2\tilde{\beta}_1} & \cdots & \dfrac{\partial^2\mathrm{SSR}}{\partial\tilde{\beta}_1\partial\tilde{\beta}_K}\\ \vdots & & \vdots\\ \dfrac{\partial^2\mathrm{SSR}}{\partial\tilde{\beta}_K\partial\tilde{\beta}_1} & \cdots & \dfrac{\partial^2\mathrm{SSR}}{\partial^2\tilde{\beta}_K}\end{pmatrix}=2\boldsymbol{X}'\boldsymbol{X}$

为正定矩阵（其中 $\dfrac{\partial(\cdot)}{\partial\tilde{\boldsymbol{\beta}}'}$ 表示分别对 $\tilde{\boldsymbol{\beta}}$ 的每一个分量求偏导，然后把这些偏导数排成行向量的形式），因为 $\boldsymbol{X}$ 满列秩，故 $\boldsymbol{X}'\boldsymbol{X}$ 正定。将 $\boldsymbol{b}$ 替代方程“$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$”中的 $\boldsymbol{\beta}$ 并令 $\boldsymbol{\varepsilon}=\boldsymbol{0}$，可得被解释变量的“拟合值”（fitted values）或“预测值”（predicted values）

$$\hat{\boldsymbol{y}}\equiv(\hat{y}_1\ \ \hat{y}_2\ \ \cdots\ \ \hat{y}_n)'\equiv\boldsymbol{X}\boldsymbol{b} \tag{3.11}$$

由于残差向量 $\boldsymbol{e}$ 与解释变量 $\boldsymbol{X}$ 正交，故可以把被解释变量 $\boldsymbol{y}$ 分解为两个正交的部分，即 $\boldsymbol{y}=\hat{\boldsymbol{y}}+\boldsymbol{e}$，而 $\hat{\boldsymbol{y}}$ 与 $\boldsymbol{e}$ 正交，因为 $\hat{\boldsymbol{y}}'\boldsymbol{e}=(\boldsymbol{X}\boldsymbol{b})'\boldsymbol{e}=\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{b}'\cdot\boldsymbol{0}=\boldsymbol{0}$。

对于扰动项方差 $\sigma^2=\mathrm{Var}(\varepsilon_i)$，由于总体扰动项 $\boldsymbol{\varepsilon}$ 不可观测，而样本残差 $\boldsymbol{e}$ 可以近似地看成是 $\boldsymbol{\varepsilon}$ 的实现值[①]，故使用以下统计量作为对方差 σ^2 的估计：

$$s^2\equiv\frac{1}{n-K}\sum_{i=1}^{n}e_i^2 \tag{3.12}$$

其中，$(n-K)$ 为自由度。为什么除以 $(n-K)$ 而不除以 n？因为随机变量 $\{e_1,e_2,\cdots,e_n\}$ 必须满足 K 个正规方程 $\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0}$，故必有其中 $(n-K)$ 个 e_i 是相互独立的。经过这样校正后，才是“无偏估计”（unbiased estimator），即满足 $\mathrm{E}(s^2)=\sigma^2$。当然，如果样本容量 n 很大（$n\to\infty$），则 $\dfrac{n-K}{n}\to 1$，是否进行“小样本校正”（small sample adjustment）并没有多少差别。另外，称 $s=\sqrt{s^2}$ 为“回归方程的标准误差”（standard error of the regression），简称“回归方程的标准误”。更一般地，通常称某统计量的标准差为该统计量的“标准误”（standard error）。

① 我们永远无法知道 $\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\}$ 的真正实现值，因为扰动项不可观测（unobservable）。

3.3 OLS的几何解释

利用OLS的正交性，可以给予OLS估计量直观的几何解释，参见图3.2。其中，$\hat{\boldsymbol{y}}$ 是 $\boldsymbol{y}$ 向超平面 $\boldsymbol{X}$ 的投影(projection)，因为 $\boldsymbol{e}$ 与 $\boldsymbol{X}$ 正交。

由于 $\hat{\boldsymbol{y}} \equiv \boldsymbol{Xb} = \boldsymbol{X}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}_{=\boldsymbol{b}} \equiv \boldsymbol{Py}$，故 $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 被称为“投影矩阵”(projection matrix)，因为用 $\boldsymbol{P}$ 左乘任何向量就可得到该向量在超平面 $\boldsymbol{X}$ 上的投影。另一方面，$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Py} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} \equiv \boldsymbol{My}$，其中 $\boldsymbol{M} \equiv \boldsymbol{I}_n - \boldsymbol{P}$ 被称为“消灭矩阵”(annihilator matrix)，因为用消灭矩阵 $\boldsymbol{M}$ 左乘任何向量，则得到该向量对超平面 $\boldsymbol{X}$ 投影后的残差向量。

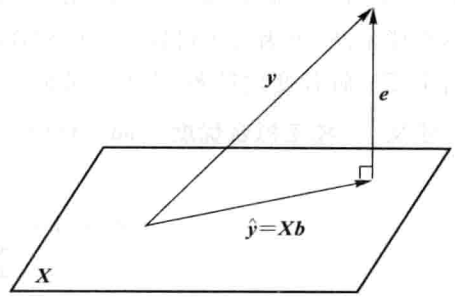


图3.2 最小二乘法的正交性

对于矩阵 $\boldsymbol{P}$ 与 $\boldsymbol{M}$，可以证明以下性质(参见习题)：

(i) $\boldsymbol{PX} = \boldsymbol{X}$；（自己的投影还是自己）

(ii) $\boldsymbol{Pe} = \boldsymbol{0}$；（垂直于 $\boldsymbol{X}$ 的向量 $\boldsymbol{e}$ 投影于 $\boldsymbol{X}$ 则退化为一个点）

(iii) $\boldsymbol{MX} = \boldsymbol{0}$；（自己对自己投影，其残差为 $\boldsymbol{0}$）

(iv) $\boldsymbol{P}$ 与 $\boldsymbol{M}$ 都是对称阵；

(v) $\boldsymbol{P}^2 = \boldsymbol{P}$；（再次投影的效果等于一次投影）

(vi) $\boldsymbol{M}^2 = \boldsymbol{M}$。（再次消灭的效果等于一次消灭）

利用消灭矩阵的性质，可以把残差写成总体扰动项 $\boldsymbol{\varepsilon}$ 的函数：

$$\boldsymbol{e} = \boldsymbol{My} = \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \underbrace{\boldsymbol{MX\beta}}_{=\boldsymbol{0}} + \boldsymbol{M\varepsilon} = \boldsymbol{M\varepsilon} \tag{3.13}$$

进一步，可以把残差平方和也写为 $\boldsymbol{\varepsilon}$ 的函数：

$$\mathrm{SSR} = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{M\varepsilon})'\boldsymbol{M\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}^2\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M\varepsilon} \tag{3.14}$$

3.4 拟合优度

如果有常数项，则可将被解释变量的离差平方和 $\sum_{i=1}^{n}(y_i - \bar{y})^2$ 分解为

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}e_i^2 \tag{3.15}$$

其中，$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n}y_i$ 为样本均值。上式表明，导致被解释变量 y_i 偏离其样本均值 $\bar{y}$ 的因素可以分为两部分，即可以由模型解释的部分 $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，以及无法由模型解释的残差部分 $\sum_{i=1}^{n}e_i^2$。这个“平方和分解公式”之所以成立正是由于OLS的正交性质(参见习题)。利用此公式，可以定义“拟合优度”，用以衡量线性回归模型对样本数据的拟合程度。

定义 “**拟合优度**”(goodness of fit) R^2 为

$$0 \leqslant R^2 \equiv \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \leqslant 1 \tag{3.16}$$

拟合优度 R^2 也称为“可决系数”(coefficient of determination)。可以证明(参见习题),在有常数项的情况下,拟合优度就等于被解释变量 y_i 与拟合值 $\hat{y}_i$ 之间相关系数的平方,即 $R^2 = [\mathrm{Corr}(y_i, \hat{y}_i)]^2$。显然,$R^2$ 越高,则说明拟合程度越好。如果向回归方程中增加解释变量,则 R^2 必然只增不减,因为至少可以让新增解释变量的系数为 0 从而保持 R^2 不变。为此,可以通过调整自由度对解释变量过多(模型不够简洁)进行惩罚。

定义 “**校正拟合优度**”(adjusted R^2)$\overline{R}^2$ 为

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-K)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 / (n-1)} \tag{3.17}$$

$\overline{R}^2$ 的一个缺点是它可能为负数。无论 R^2 还是 $\overline{R}^2$,只是反映了拟合程度的好坏(即观测值距离回归超平面的远近),除此以外没有太多意义。我们并不知道它们的统计分布。评估一个回归方程是否显著,更多地应该看后面介绍的 F 检验(尽管 R^2 与 F 统计量也有联系)。

如果回归模型中没有常数项,则上述“离差形式”的平方和分解公式不成立(参见习题),但仍可以将被解释变量的平方和 $\sum_{i=1}^{n} y_i^2$ 分解:

$$\sum_{i=1}^{n} y_i^2 = \boldsymbol{y}'\boldsymbol{y} = (\hat{\boldsymbol{y}} + \boldsymbol{e})'(\hat{\boldsymbol{y}} + \boldsymbol{e}) = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + 2\underbrace{\hat{\boldsymbol{y}}'\boldsymbol{e}}_{=0} + \boldsymbol{e}'\boldsymbol{e} = \sum_{i=1}^{n}\hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 \tag{3.18}$$

其中,由于 $\hat{\boldsymbol{y}}$ 与 $\boldsymbol{e}$ 正交,故 $\hat{\boldsymbol{y}}'\boldsymbol{e} = 0$。因此,仍可计算“非中心 R^2”(Uncentered R^2):

$$R_{uc}^2 \equiv \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}} \tag{3.19}$$

事实上,如果线性回归没有常数项,Stata 软件汇报的 R^2 正是 R_{uc}^2。另外,可以证明,$R_{uc}^2 = \dfrac{\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}}$(参见习题),这个结果在第 6 章推导拉格朗日乘子(LM)检验时会用到。

3.5 OLS 的小样本性质

(1) 线性性:OLS 估计量 $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 为 $\boldsymbol{y}$ 的线性组合。

(2) 无偏性:$\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta}$,即 $\boldsymbol{b}$ 不会系统地高估或低估 $\boldsymbol{\beta}$。

证明:“抽样误差”(sampling error)为

$$\boldsymbol{b} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \equiv \boldsymbol{A}\boldsymbol{\varepsilon}$$

其中,记 $\boldsymbol{A} \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$。所以

$$\mathrm{E}(\boldsymbol{b} - \boldsymbol{\beta}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{A}\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{A}\underbrace{\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})}_{=\boldsymbol{0}} = \boldsymbol{0} \text{ (严格外生性)}$$

将上式移项,即可得 $\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta}$。在此证明中,严格外生性是不可缺少的关键假定。

推论 无条件期望 $\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta}$。

证明:$\mathrm{E}(\boldsymbol{b}) = \mathrm{E}_X\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \mathrm{E}_X(\boldsymbol{\beta}) = \boldsymbol{\beta}$ (常数的期望仍为常数本身)。

(3) 估计量 $\boldsymbol{b}$ 的方差为 $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$。

证明:$\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})=\mathrm{Var}(\boldsymbol{b}-\boldsymbol{\beta}|\boldsymbol{X})$ (因为 $\boldsymbol{\beta}$ 是常数)

$$=\mathrm{Var}(\boldsymbol{A}\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{A}\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})\boldsymbol{A}'=\boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A}'$$

$$=\sigma^2\boldsymbol{A}\boldsymbol{A}'=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

在这个证明过程中,球型扰动项的假定是关键。如果扰动项存在"条件异方差",则估计量的方差表达式与上式不同,应使用"异方差稳健标准误"(heteroskedasticity-robust standard error),参见第 5 章。

(4) "高斯 - 马尔可夫定理"(Gauss-Markov Theorem):最小二乘法是最佳线性无偏估计(Best Linear Unbiased Estimator,简记 BLUE),即在所有线性无偏估计中,最小二乘法的方差最小。

证明:如上所述,OLS 估计量 $\boldsymbol{b}$ 为线性无偏估计。假设 $\hat{\boldsymbol{\beta}}$ 为任一线性无偏估计,需要证明 $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})\geqslant\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$,即 $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})-\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ 为半正定矩阵。$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})-\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ 为半正定矩阵的一个含义是,$\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ 的主对角线元素(即方差)一定小于或等于 $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ 的主对角线上对应的元素(参见习题)。为了比较 $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ 与 $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ 的大小,首先要找到 $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$ 的表达式,因为已知 $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$。由于 $\hat{\boldsymbol{\beta}}$ 为线性估计,故存在常数矩阵 $\boldsymbol{C}_{K\times n}$,使得 $\hat{\boldsymbol{\beta}}=\boldsymbol{C}\boldsymbol{y}$。我们知道,$\boldsymbol{b}=\boldsymbol{A}\boldsymbol{y}$,其中 $\boldsymbol{A}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$。定义 $\boldsymbol{D}\equiv\boldsymbol{C}-\boldsymbol{A}$,则

$$\hat{\boldsymbol{\beta}}=\boldsymbol{C}\boldsymbol{y}=(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{y}=\boldsymbol{D}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})+\boldsymbol{b}=\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{b} \tag{3.20}$$

利用 $\hat{\boldsymbol{\beta}}$ 的无偏性可得,

$$\boldsymbol{\beta}=\mathrm{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})=\mathrm{E}(\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{b}|\boldsymbol{X})=\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{D}\underbrace{\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})}_{=\boldsymbol{0}}+\underbrace{\mathrm{E}(\boldsymbol{b}|\boldsymbol{X})}_{=\boldsymbol{\beta}}=\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\beta} \tag{3.21}$$

因此,对于任意 $\boldsymbol{\beta}$,都有 $\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{0}$。故 $\boldsymbol{D}\boldsymbol{X}=\boldsymbol{0}$,这意味着矩阵 $\boldsymbol{D}$ 必须满足这个性质,才能保证 $\hat{\boldsymbol{\beta}}=\boldsymbol{C}\boldsymbol{y}$ 为线性无偏估计。利用 $\boldsymbol{D}\boldsymbol{X}=\boldsymbol{0}$,可以简化 $\hat{\boldsymbol{\beta}}$ 的表达式为

$$\hat{\boldsymbol{\beta}}=\underbrace{\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}}_{=\boldsymbol{0}}+\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{b}=\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{b}$$

如果使用 $\hat{\boldsymbol{\beta}}$ 作为 $\boldsymbol{\beta}$ 估计量,则抽样误差为

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{b}-\boldsymbol{\beta}=\boldsymbol{D}\boldsymbol{\varepsilon}+\boldsymbol{A}\boldsymbol{\varepsilon}=(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{\varepsilon} \tag{3.22}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})=\mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|\boldsymbol{X})=\mathrm{Var}[(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{\varepsilon}|\boldsymbol{X}]=(\boldsymbol{D}+\boldsymbol{A})\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})(\boldsymbol{D}+\boldsymbol{A})'$$

$$=\sigma^2(\boldsymbol{D}+\boldsymbol{A})(\boldsymbol{D}'+\boldsymbol{A}')=\sigma^2(\boldsymbol{D}\boldsymbol{D}'+\underbrace{\boldsymbol{A}\boldsymbol{D}'}_{=\boldsymbol{0}}+\underbrace{\boldsymbol{D}\boldsymbol{A}'}_{=\boldsymbol{0}}+\boldsymbol{A}\boldsymbol{A}')$$

$$=\sigma^2[\boldsymbol{D}\boldsymbol{D}'+(\boldsymbol{X}'\boldsymbol{X})^{-1}] \quad (\text{因为 } \boldsymbol{D}\boldsymbol{A}'=\underbrace{\boldsymbol{D}\boldsymbol{X}}_{=\boldsymbol{0}}(\boldsymbol{X}'\boldsymbol{X})^{-1}=\boldsymbol{0})$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})-\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})=\sigma^2[\boldsymbol{D}\boldsymbol{D}'+(\boldsymbol{X}'\boldsymbol{X})^{-1}]-\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}=\sigma^2\boldsymbol{D}\boldsymbol{D}'$$

由于 $\boldsymbol{D}\boldsymbol{D}'$ 为半正定矩阵,故高斯 - 马尔可夫定理成立。

应该注意的是,如果没有球型扰动项的假定,则最小二乘法不是 BLUE,还存在其他更优的线性无偏估计,参见第 7 章的广义最小二乘法(Generalized Least Square,简记 GLS)。

(5) 方差的无偏估计:$\mathrm{E}(s^2|\boldsymbol{X})=\sigma^2$。

证明:因为 $\mathrm{E}(s^2|\boldsymbol{X})=\mathrm{E}\left(\dfrac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\middle|\boldsymbol{X}\right)=\mathrm{E}\left(\dfrac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{n-K}\middle|\boldsymbol{X}\right)=\dfrac{1}{n-K}\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X})$,故只要证明

$E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) = (n-K)\sigma^2$ 即可。为方便证明，引入矩阵“迹”的运算。

定义 任意方阵 $\boldsymbol{A}$ 的“迹”(trace)就是其主对角线元素之和，记为 $\text{trace}(\boldsymbol{A})$。

可以看出，迹运算具有线性性，即 $\text{trace}(\boldsymbol{A}+\boldsymbol{B}) = \text{trace}(\boldsymbol{A}) + \text{trace}(\boldsymbol{B})$，而 $\text{trace}(k\boldsymbol{A}) = k\text{trace}(\boldsymbol{A})$，其中 k 为常数。还可以证明，只要 $\boldsymbol{AB}$ 与 $\boldsymbol{BA}$ 都存在，则 $\text{trace}(\boldsymbol{AB}) = \text{trace}(\boldsymbol{BA})$(对于迹运算，矩阵乘法可以交换次序)。另外，如果 $\boldsymbol{A}$ 为 1×1 矩阵(常数)，则显然 $\text{trace}(\boldsymbol{A}) = \boldsymbol{A}$。这些良好的性质为下面的证明提供了方便。

$$
\begin{aligned}
E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) &= E[\text{trace}(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X})] && (\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}\text{ 为 }1\times1\text{ 矩阵})\\
&= E[\text{trace}(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})] && (\text{将 }\boldsymbol{\varepsilon}'\text{ 与 }\boldsymbol{M}\boldsymbol{\varepsilon}\text{ 交换次序})\\
&= \text{trace}[E(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})] && (\text{期望算子与迹算子可交换次序})\\
&= \text{trace}[\boldsymbol{M}\sigma^2\boldsymbol{I}_n] && (\text{球型扰动项})\\
&= \sigma^2\text{trace}(\boldsymbol{M}) && (\text{迹运算的线性性})\\
\text{trace}(\boldsymbol{M}) &= \text{trace}[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] && (\text{消灭矩阵 }\boldsymbol{M}\text{ 的定义})\\
&= \text{trace}(\boldsymbol{I}_n) - \text{trace}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] && (\text{迹运算的线性性})\\
&= n - \text{trace}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}] && (\boldsymbol{X}\text{ 与 }(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{ 互换})\\
&= n - \text{trace}(\boldsymbol{I}_K) && (\boldsymbol{X}'\boldsymbol{X}\text{ 为 }K\times K\text{ 矩阵})\\
&= n - K && (K\text{ 阶单位阵的迹为 }K)
\end{aligned}
$$

因此，对协方差阵 $\text{Var}(\boldsymbol{b}|\boldsymbol{X})$ 的无偏估计为 $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，在 Stata 中记为“VCE”(Variance-Covariance Matrix Estimated)。Stata 所汇报的协方差矩阵就是根据 $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 来计算的。

3.6 对单个系数的 t 检验

计量经济学的一大用途是用来检验经济理论及其推论。比如，“资本市场有效理论”认为，所有已知的信息都已经体现在股价中了，所以股价是不可预测的。那么，股价与过去的宏观经济指标之间就不应该有相关性。这是一个可以检验的假设。为了进行“假设检验”(hypothesis testing)，我们必须对回归方程扰动项的具体概率分布进行假设。

假定 3.5 在给定 $\boldsymbol{X}$ 的情况下，$\boldsymbol{\varepsilon}|\boldsymbol{X}$ 的条件分布为正态，即 $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$。

如果把扰动项看成是许多遗漏变量与测量误差之和，则中心极限定理(Central Limit Theorem，参见第 5 章)表明，扰动项可能近似地服从正态分布。这是假定 3.5 成立的理论基础。从第 2 章我们知道，正态分布具有以下特点：密度函数完全由均值及协方差矩阵决定；两个随机变量不相关就意味着相互独立；正态分布变量的线性函数仍然是正态分布。

本节首先考虑最简单的假设检验，即对单个系数进行检验。需要检验的“原假设”(null hypothesis，也称为“零假设”)为 $H_0: \beta_k = \overline{\beta}_k$，其中 $\overline{\beta}_k$ 为给定的常数。通常 $\overline{\beta}_k = 0$，此时即检验是否变量 x_{ik} 的系数显著地不等于 0。假设检验的实质是一种概率意义上的反证法，即首先假设原假设成立，然后看在原假设成立的前提下，是否导致不太可能发生的“小概率事件”在一次抽样中出现。如果小概率事件竟然在一次抽样实验中被观测到，则说明原假设不可信，应该拒绝原假设，接受“替代假设”(alternative hypothesis，也称为“备择假设”)$H_1: \beta_k \neq \overline{\beta}_k$。

直观来说，如果未知参数 β_k 的估计值 b_k 离 $\overline{\beta}_k$ 较远，则应倾向于拒绝原假设。这类检验称为“沃尔德检验”(Wald test)。在衡量距离远近时，由于绝对距离依赖于变量的单位，故需要以标准差为基准来考虑相对距离。

由于 $\boldsymbol{\varepsilon}|\boldsymbol{X}\sim N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$，而 $\boldsymbol{b}-\boldsymbol{\beta}=\boldsymbol{A}\boldsymbol{\varepsilon}$ 为 $\boldsymbol{\varepsilon}$ 的线性函数，故 $(\boldsymbol{b}-\boldsymbol{\beta})|\boldsymbol{X}$ 服从正态分布。更进一步，$\mathrm{E}(\boldsymbol{b}-\boldsymbol{\beta}|\boldsymbol{X})=\mathbf{0}$，$\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，故 $(\boldsymbol{b}-\boldsymbol{\beta})|\boldsymbol{X}\sim N(\mathbf{0},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$。因此，在原假设“$H_0:\beta_k=\overline{\beta}_k$”成立的情况下，其第 k 个分量 $(b_k-\overline{\beta}_k)|\boldsymbol{X}\sim N(0,\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk})$，其中 $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}$ 为矩阵 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的第 (k,k) 个元素，而 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}$ 为 b_k 的方差。如果 σ^2 已知，则统计量 $z_k\equiv\dfrac{b_k-\overline{\beta}_k}{\sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}\sim N(0,1)$。但通常 σ^2 未知，此时，σ^2 称为“厌恶参数”(nuisance parameter)，因为虽然我们对 σ^2 不感兴趣，但 σ^2 却出现在统计量的表达式中。一个合格的“检验统计量”(test statistic)必须满足两个条件：首先，它必须能够根据样本数据计算出来；其次，它的概率分布是已知的。如果以估计值 s^2 来替代 σ^2，则可得到满足这两个条件的 t 统计量。

定理(t 统计量的分布) 在假定 3.1—3.5 均满足，且原假设“$H_0:\beta_k=\overline{\beta}_k$”也成立的情况下，$t$ 统计量 $t_k\equiv\dfrac{b_k-\overline{\beta}_k}{\mathrm{SE}(b_k)}\equiv\dfrac{b_k-\overline{\beta}_k}{\sqrt{s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}\sim t(n-K)$。其中，$\mathrm{SE}(b_k)$ 是 b_k 的“估计标准误差”(estimated standard error)，简称“标准误”。

证明：我们知道，如果 $Z\sim N(0,1)$，$Y\sim\chi^2(k)$，而且 Z 与 Y 相互独立，则 $\dfrac{Z}{\sqrt{Y/k}}\sim t(k)$，其中 k 为自由度。因此，将统计量 t_k 往这个方向变形。

$$t_k\equiv\frac{b_k-\overline{\beta}_k}{\sqrt{s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}=\frac{b_k-\overline{\beta}_k}{\sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}\cdot\sqrt{\frac{\sigma^2}{s^2}}=\frac{z_k}{\sqrt{s^2/\sigma^2}}=\frac{z_k}{\sqrt{\dfrac{\boldsymbol{e}'\boldsymbol{e}}{(n-K)\sigma^2}}}\equiv\frac{z_k}{\sqrt{q/(n-K)}}$$

其中，$z_k\equiv\dfrac{b_k-\overline{\beta}_k}{\sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}$，$q\equiv\dfrac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2}$。

已知 $z_k\sim N(0,1)$，下面将证明：

(1) $q|\boldsymbol{X}\sim\chi^2(n-K)$；

(2) $z_k|\boldsymbol{X}$ 与 $q|\boldsymbol{X}$ 相互独立，

则根据 t 分布的定义，$\dfrac{z_k}{\sqrt{q/(n-K)}}\sim t(n-K)$。

(1) $q\equiv\dfrac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2}=\dfrac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{\sigma^2}=\dfrac{\boldsymbol{\varepsilon}'}{\sigma}\boldsymbol{M}\dfrac{\boldsymbol{\varepsilon}}{\sigma}$（二次型）。

由于 $\boldsymbol{\varepsilon}|\boldsymbol{X}\sim N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$，故 $\left.\dfrac{\boldsymbol{\varepsilon}}{\sigma}\right|\boldsymbol{X}\sim N(\mathbf{0},\boldsymbol{I}_n)$。已知消灭矩阵 $\boldsymbol{M}$ 为“幂等矩阵”(idempotent matrix，即 $\boldsymbol{M}^2=\boldsymbol{M}$)。根据线性代数知识，对于幂等矩阵 $\boldsymbol{M}$，$\mathrm{rank}(\boldsymbol{M})=\mathrm{trace}(\boldsymbol{M})=n-K$。根据数理统计知识可知[①]，$q|\boldsymbol{X}\sim\chi^2(n-K)$。由于 $\boldsymbol{M}$ 不是满秩矩阵，故 $q|\boldsymbol{X}$ 的自由度降为 $(n-K)$。

(2) $z_k\equiv\dfrac{b_k-\overline{\beta}_k}{\sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}}$ 是 $\boldsymbol{b}$ 的函数，而 q 是 $\boldsymbol{e}$ 的函数。因此，为了证明 z_k 与 q 相互独立，只要证明 $\boldsymbol{b}$ 与 $\boldsymbol{e}$ 相互独立即可。由于 $\boldsymbol{b}=\boldsymbol{\beta}+\boldsymbol{A}\boldsymbol{\varepsilon}$ 与 $\boldsymbol{e}=\boldsymbol{M}\boldsymbol{\varepsilon}$ 都是正态扰动项 $\boldsymbol{\varepsilon}$ 的线性函数，故 $(\boldsymbol{b},\boldsymbol{e})$ 的联合分布也是正态分布(参见第 2 章命题)。因此，要证明 $\boldsymbol{b}$ 与 $\boldsymbol{e}$ 相互独立，只要证明 $\mathrm{Cov}(\boldsymbol{b},$

---

① 如果 $\boldsymbol{x}\sim N(\mathbf{0},\boldsymbol{I}_n)$(这意味着 $\boldsymbol{x}$ 的各分量相互独立)，且 $\boldsymbol{A}$ 为幂等矩阵，则二次型 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 服从自由度为 $\mathrm{rank}(\boldsymbol{A})$ 的 χ^2 分布。如果 $\boldsymbol{A}=\boldsymbol{I}_n$，则 $\boldsymbol{x}'\boldsymbol{x}\sim\chi^2(n)$，这是大家所熟知的特殊情形。更一般的情形则是这种特殊情形的推广。

$\boldsymbol{e}) = \mathbf{0}$ 即可，而这又由OLS的正交性所保证。

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{b},\boldsymbol{e}\,|\,\boldsymbol{X}) &= \mathrm{Cov}(\boldsymbol{\beta}+\boldsymbol{A\varepsilon},\boldsymbol{M\varepsilon}\,|\,\boldsymbol{X}) && \text{（代入 } \boldsymbol{b} \text{ 与 } \boldsymbol{e} \text{ 的表达式）}\\
&= \mathrm{Cov}(\boldsymbol{A\varepsilon},\boldsymbol{M\varepsilon}\,|\,\boldsymbol{X}) && \text{（去掉常数 } \boldsymbol{\beta} \text{ 不改变结果）}\\
&= \mathrm{E}(\boldsymbol{A\varepsilon\varepsilon'M}) - \underbrace{\mathrm{E}(\boldsymbol{A\varepsilon})}_{=\mathbf{0}}\underbrace{\mathrm{E}(\boldsymbol{M\varepsilon})'}_{=\mathbf{0}'} && \text{（根据协方差矩阵公式）}\\
&= \boldsymbol{A}\mathrm{E}(\boldsymbol{\varepsilon\varepsilon'})\boldsymbol{M} = \sigma^2\boldsymbol{AM} = \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'M} = \sigma^2(\boldsymbol{X'X})^{-1}\underbrace{(\boldsymbol{MX})'}_{=\mathbf{0}} = \mathbf{0}
\end{aligned}$$

1. t 检验的步骤

第一步：计算 t_k。如果 $|t_k|$ 很大，则表明原假设 H_0 较不可信。因为如果原假设 H_0 为真，则 $|t_k|$ 很大的概率将很小（为小概率事件），不应该在一次抽样中被观测到。

第二步：计算“显著性水平”（significance level）为 α 的“临界值”（critical value）$t_{\alpha/2}(n-K)$，其中 $t_{\alpha/2}(n-K)$ 的定义为

$$\mathrm{P}\{t(n-K) > t_{\alpha/2}(n-K)\} = \mathrm{P}\{t(n-K) < -t_{\alpha/2}(n-K)\} = \alpha/2 \tag{3.23}$$

其中，$t(n-K)$ 指的是自由度为 $(n-K)$ 的 t 分布变量。上式表明，$t(n-K)$ 大于 $t_{\alpha/2}(n-K)$，或小于 $-t_{\alpha/2}(n-K)$ 的概率都是 $\alpha/2$。在实践中，通常取 $\alpha = 5\%$，则 $\alpha/2 = 2.5\%$。有时也使用 $\alpha = 1\%$ 或 $\alpha = 10\%$，参见图3.3。

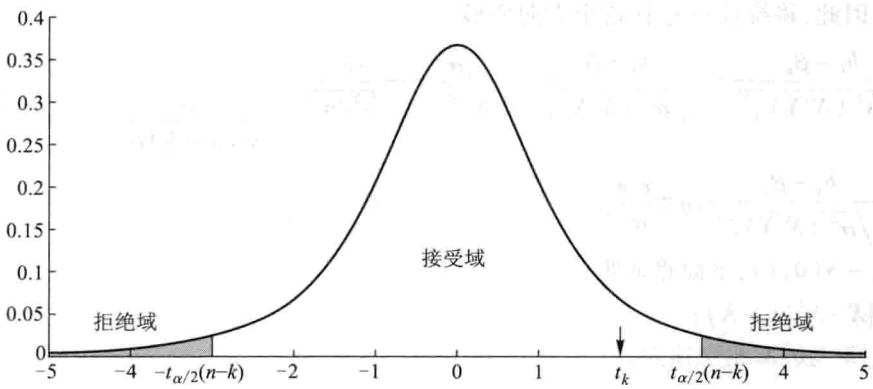


图3.3 t 检验

第三步：如果 $|t_k| > t_{\alpha/2}(n-K)$，即 t_k 落入两边的“拒绝域”（rejection region，图中阴影部分），则拒绝 H_0；反之，t_k 落入中间的“接受域”（acceptance region，图中空白部分）则接受 H_0。由于拒绝域分布在 t 分布的两边，也称此检验为“双边检验”（two-tailed）。

2. 计算 p 值

除了根据临界值进行假设检验外，还可以通过 p 值进行检验。

定义 给定检验统计量的样本观测值，称原假设可被拒绝的最小显著性水平为此假设检验问题的 p 值（probability value，即 p-value）。

在上面的 t 检验中，p 值为 $\mathrm{P}(t > |t_k|) \times 2$，其中 t_k 为检验统计量的样本观测值。显然，p 值越小则越倾向于拒绝原假设。比如，p 值 $= 0.03$，则可以在5%的显著性水平上拒绝原假设。更进一步，“p 值 $= 0.03$”提供了比“在5%的显著性水平上拒绝原假设”更多的信息，因为还可以“在3%的显著性水平上拒绝原假设”。总之，使用 p 值进行假设检验一般比临界值更有信息量。当Stata直接给出 p 值时，就不需要知道临界值了。

由于现代的假设检验一般在计算机中进行，故本书没有附统计分布的临界值表（在一般的

概率统计教材都可以找到)。

3. 计算置信区间

参数估计可以分为两种:点估计与区间估计。前面已经介绍了 OLS 的点估计。这里介绍对 β_k 的区间估计。假设"置信度"(confidence level)为$(1-\alpha)$(比如 $\alpha=5\%$,则 $1-\alpha=95\%$),区间估计的目的就是要找到一个"置信区间"(confidence interval),使得该区间覆盖真实参数 β_k 的概率为$(1-\alpha)$。由于$\dfrac{b_k-\beta_k}{\mathrm{SE}(b_k)}\sim t(n-K)$,故

$$\mathrm{P}\left\{-t_{\alpha/2}<\frac{b_k-\beta_k}{\mathrm{SE}(b_k)}<t_{\alpha/2}\right\}=1-\alpha \quad (\text{根据 } t_{\alpha/2}\text{的定义})$$

$$\mathrm{P}\{b_k-t_{\alpha/2}\mathrm{SE}(b_k)<\beta_k<b_k+t_{\alpha/2}\mathrm{SE}(b_k)\}=1-\alpha \quad (\text{不等式变形})$$

即置信区间为$[b_k-t_{\alpha/2}\mathrm{SE}(b_k),b_k+t_{\alpha/2}\mathrm{SE}(b_k)]$,以点估计 b_k 为中心,区间半径为 $t_{\alpha/2}\mathrm{SE}(b_k)$。需要注意的是,这个置信区间是个随机区间,随着样本的不同而不同。置信区间的直观含义是,如果置信度为95%,而同样的抽样进行了100次,得到100个置信区间,则其中大约有95个置信区间能覆盖到真实参数 β_k。

4. 第Ⅰ类错误与第Ⅱ类错误

根据样本信息对总体进行推断,有可能犯错误。特别地,在进行假设检验时,可能犯两类性质不同的错误。

定义 **"第Ⅰ类错误"**(Type Ⅰ error)指的是,虽然原假设为真,但却根据观测数据做出了拒绝原假设的错误判断,即"弃真"。第Ⅰ类错误的发生概率为 $\mathrm{P}(\text{reject } H_0\mid H_0)$。

定义 **"第Ⅱ类错误"**(Type Ⅱ error)指的是,虽然原假设为假(替代假设为真),但却根据观测数据做出了接受原假设的错误判断,即"存伪"。第Ⅱ类错误的发生概率为 $\mathrm{P}(\text{accept } H_0\mid H_1)$。

自然地,我们希望这两类错误的发生概率越小越好。但一般来说,除非增加样本容量,二者存在此消彼长的关系,即减少第Ⅰ类错误的发生概率,必然导致第Ⅱ类错误的发生概率增加,反之亦然。传统上,在进行假设检验时,一般先指定我们可以接受的发生第Ⅰ类错误的最大概率,即"显著性水平",比如5%;而不指定第Ⅱ类错误的发生概率(通常更难计算)。

定义 称"1减去第Ⅱ类错误的发生概率"为统计检验的"功效"或"势"(power),即"$1-\mathrm{P}(\text{accept } H_0\mid H_1)$"。换言之,功效为在原假设为错误的情况下,拒绝原假设的概率。

由于在进行假设检验时,通常知道第Ⅰ类错误的发生概率,而不知道第Ⅱ类错误的发生概率。因此,如果拒绝原假设,可以比较理直气壮,因为知道犯错误的概率就是显著性水平(比如5%);另一方面,如果接受原假设,则比较没有把握,因为我们通常并不知道第Ⅱ类错误的发生概率(可能很高)。

3.7 对线性假设的 F 检验

我们常常想知道整个回归方程是否显著,即需要检验原假设"$H_0:\beta_2=\cdots=\beta_K=0$"($\beta_1$ 为常数项)。更一般地,想知道回归系数的 m 个线性假设是否同时成立:

$$H_0:\underbrace{\boldsymbol{R}}_{m\times K}\underbrace{\boldsymbol{\beta}}_{K\times 1}=\underbrace{\boldsymbol{r}}_{m\times 1}$$

其中,$\boldsymbol{r}$ 为 m 维列向量,$\boldsymbol{R}$ 为 $m\times K$ 矩阵,$\mathrm{rank}(\boldsymbol{R})=m$,即 $\boldsymbol{R}$ 满行秩,没有多余的方程或自相矛盾

的方程。

例 对于模型 $y=\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\varepsilon$，检验“$H_0:\beta_2=\beta_3$ 且 $\beta_4=0$”。则

$$\boldsymbol{R}=\begin{pmatrix}0&1&-1&0\\0&0&0&1\end{pmatrix},\boldsymbol{r}=\begin{pmatrix}0\\0\end{pmatrix},\text{因为}\begin{pmatrix}0&1&-1&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\end{pmatrix}=\begin{pmatrix}\beta_2-\beta_3\\\beta_4\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}。$$

直观来看，由于 $\boldsymbol{b}$ 是 $\boldsymbol{\beta}$ 的估计量，故如果 H_0 成立，则 $(\boldsymbol{Rb}-\boldsymbol{r})$ 应该比较接近于 $\boldsymbol{0}$ 向量。因此，可以使用以下的沃尔德检验。

定理（F 统计量的分布） 在假定 3.1—3.5 均满足，且原假设“$H_0:\boldsymbol{R\beta}=\boldsymbol{r}$”也成立的情况下，则 F 统计量

$$F\equiv\frac{(\boldsymbol{Rb}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})/m}{s^2}\sim F(m,n-K)\tag{3.24}$$

证明：由于 $s^2=\boldsymbol{e}'\boldsymbol{e}/(n-K)$，可将 F 写成

$$F\equiv\frac{(\boldsymbol{Rb}-\boldsymbol{r})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})/m}{(\boldsymbol{e}'\boldsymbol{e}/\sigma^2)/(n-K)}\equiv\frac{w/m}{q/(n-K)}$$

其中，$w\equiv(\boldsymbol{Rb}-\boldsymbol{r})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})$。下面将证明：

（1）$w|\boldsymbol{X}\sim\chi^2(m)$；

（2）$q|\boldsymbol{X}\sim\chi^2(n-K)$；（已在 t 统计量定理中证明）

（3）$w|\boldsymbol{X}$ 与 $q|\boldsymbol{X}$ 相互独立，

则根据 F 分布的定义，$\frac{w/m}{q/(n-K)}\sim F(m,n-K)$。

（1）定义 $\boldsymbol{v}\equiv\boldsymbol{Rb}-\boldsymbol{r}$。在 H_0 成立的情况下，$\boldsymbol{v}\equiv\boldsymbol{Rb}-\boldsymbol{r}=\boldsymbol{Rb}-\boldsymbol{R\beta}=\boldsymbol{R}(\boldsymbol{b}-\boldsymbol{\beta})$。由于 $\boldsymbol{b}$ 为正态分布，故 $\boldsymbol{v}|\boldsymbol{X}$ 为 m 维正态分布，且 $\mathrm{E}(\boldsymbol{v}|\boldsymbol{X})=\boldsymbol{0}$，其方差为

$$\mathrm{Var}(\boldsymbol{v}|\boldsymbol{X})=\mathrm{Var}[\boldsymbol{R}(\boldsymbol{b}-\boldsymbol{\beta})|\boldsymbol{X}]=\boldsymbol{R}\mathrm{Var}(\boldsymbol{b})\boldsymbol{R}'=\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$$

根据数理统计知识知道[①]，

$$w\equiv(\boldsymbol{Rb}-\boldsymbol{r})'[\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})=\boldsymbol{v}'[\mathrm{Var}(\boldsymbol{v}|\boldsymbol{X})]^{-1}\boldsymbol{v}\sim\chi^2(m)$$

注：由于 $\boldsymbol{R}$ 满行秩，故 $[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}$ 存在。

（2）w 是 $\boldsymbol{b}$ 的函数，而 q 是 $\boldsymbol{e}$ 的函数，由于 $\boldsymbol{b}$ 与 $\boldsymbol{e}$ 相互独立，故 $w|\boldsymbol{X}$ 与 $q|\boldsymbol{X}$ 相互独立。

F 检验的步骤

第一步：计算 F 统计量。如果 F 统计量很大，则表明原假设 H_0 较不可信。因为如果原假设 H_0 为真，则“F 统计量很大”的概率将很小（为小概率事件），不应该在一次抽样中被观测到。

第二步：计算显著性水平为 α 的临界值 $F_\alpha(m,n-K)$，$F_\alpha(m,n-K)$ 满足

$$\mathrm{P}\{F(m,n-K)>F_\alpha(m,n-K)\}=\alpha\tag{3.25}$$

其中，$F(m,n-K)$ 指的是自由度为 $(m,n-K)$ 的 F 分布变量。上式表明，$F(m,n-K)$ 大于临界值 $F_\alpha(m,n-K)$ 的概率恰好为 α。

第三步：如果 $F>F_\alpha(m,n-K)$，即 F 统计量落入右边拒绝域（即阴影部分），则拒绝 H_0；反之，F 统计量落入接受域（即空白部分），则接受 H_0，参见图 3.4。由于拒绝域只在右侧，故这是一个“单边右侧检验”（one-tailed）。

---

① 如果 m 维随机变量 $\boldsymbol{x}$ 服从正态分布 $\boldsymbol{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma}$ 为非退化矩阵（满秩），则二次型 $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sim\chi^2(m)$。

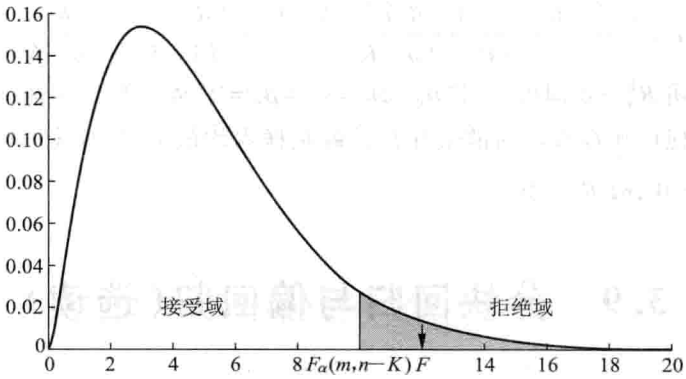


图 3.4 F 检验

3.8 F统计量的似然比原理表达式

如果使用约束条件下的最小二乘法，即“约束最小二乘法”(Restricted OLS，简记 RLS；或 Constrained OLS)，可以得到 F 统计量的另一方便表达式。为此，考虑以下约束极值问题：

$$\min_{\tilde{\boldsymbol{\beta}}} \text{SSR}(\tilde{\boldsymbol{\beta}})$$
$$s.t.\ \boldsymbol{R}\tilde{\boldsymbol{\beta}} = \boldsymbol{r} \tag{3.26}$$

其基本思想是，如果“$H_0: \boldsymbol{R\beta} = \boldsymbol{r}$”正确，则加上此约束应该不会使残差平方和 SSR($\tilde{\boldsymbol{\beta}}$)的最小值增大很多。通过引入拉格朗日函数，可以求解这个约束极值问题，并证明(参见附录)：

$$F = \frac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*} - \boldsymbol{e}'\boldsymbol{e})/m}{\boldsymbol{e}'\boldsymbol{e}/(n-K)} \tag{3.27}$$

其中，$\boldsymbol{e}$ 为无约束的残差向量，$\boldsymbol{e}^{*}$ 为有约束的残差向量，而 m 为约束条件的个数。这个 F 统计量表达式有时更容易计算。这种通过比较“条件极值”与“无条件极值”而进行的检验，统称为“似然比检验”(Likelihood ratio test)，参见第 6 章。利用 F 统计量的似然比原理表达式，可以建立 F 统计量检验与拟合优度 R^2 的关系。

命题 对于线性回归方程“$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$”，检验原假设“$H_0: \beta_2 = \cdots = \beta_K = 0$”(即该方程的显著性)的 F 统计量等于$\dfrac{R^2/(K-1)}{(1-R^2)/(n-K)}$。

证明：由于共有$(K-1)$个约束，故根据似然比原理的 F 统计量为

$$F = \frac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*} - \boldsymbol{e}'\boldsymbol{e})/(K-1)}{\boldsymbol{e}'\boldsymbol{e}/(n-K)} = \frac{\dfrac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*} - \boldsymbol{e}'\boldsymbol{e})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\Big/(K-1)}{\dfrac{\boldsymbol{e}'\boldsymbol{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\Big/(n-K)} \tag{3.28}$$

其中，$\boldsymbol{e}$ 为无约束 OLS 的残差向量，而 $\boldsymbol{e}^{*}$ 为在 H_0 约束下 OLS 的残差向量。记约束回归的拟合优度为 R_*^2，由于$\dfrac{\boldsymbol{e}'\boldsymbol{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - R^2$，$\dfrac{\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - R_*^2$，故

$$F=\frac{[(1-R_*^2)-(1-R^2)]/(K-1)}{(1-R^2)/(n-K)}=\frac{(R^2-R_*^2)/(K-1)}{(1-R^2)/(n-K)} \tag{3.29}$$

现在只需要证明 $R_*^2=0$ 即可。当“$H_0:\beta_2=\cdots=\beta_K=0$”成立时，$y_i=\beta_1+\varepsilon_i$，而 $b_1^*=\bar{y}$（若只对常数项 β_1 进行回归，则对常数项的最佳估计就是样本均值 $\bar{y}$）①，故 $\hat{y}_i^*=b_1^*=\bar{y}$。由此可知，$\sum_{i=1}^{n}(\hat{y}_i^*-\bar{y})^2=0$，故 $R_*^2=0$。

3.9 分块回归与偏回归（选读）

在多元回归中，只要增加一个变量就会对所有的回归系数产生影响。这种影响究竟如何发生？仅从 OLS 估计量的表达式 $\boldsymbol{b}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 很难看出不同变量的影响（比如，看不出其中的某个元素 b_k 是如何决定的）。为此，将数据矩阵分为两个部分，即 $\boldsymbol{X}=(\boldsymbol{X}_1\ \boldsymbol{X}_2)$，分别对应于两组解释变量。则多元回归模型可以写为，

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon} \tag{3.30}$$

其中，回归系数也相应地分组，即 $\boldsymbol{\beta}\equiv\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}$。对上式进行回归，可得 $\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}$ 的 OLS 估计量 $\boldsymbol{b}=\begin{pmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{pmatrix}$。为了知道 $\boldsymbol{X}_2$ 的边际影响，考虑以下回归。首先，把 $\boldsymbol{y}$ 对 $\boldsymbol{X}_1$ 进行回归，将所得的残差记为 $\boldsymbol{e}_1$，即 $\boldsymbol{y}$ 中不能由 $\boldsymbol{X}_1$ 解释的部分；其次，把 $\boldsymbol{X}_2$ 中的每个变量分别对 $\boldsymbol{X}_1$ 进行回归，将所得的残差列为一个残差矩阵 $\boldsymbol{e}_2$（其中的每列均为残差向量），即 $\boldsymbol{X}_2$ 中不能由 $\boldsymbol{X}_1$ 解释的部分；最后，将 $\boldsymbol{e}_1$ 对 $\boldsymbol{e}_2$ 进行回归，即考察“$\boldsymbol{X}_2$ 中与 $\boldsymbol{X}_1$ 无关的部分”对“$\boldsymbol{y}$ 中与 $\boldsymbol{X}_1$ 无关的部分”的解释力。在最后这一步回归中，$\boldsymbol{e}_2$ 的回归系数将是什么？正好是 $\boldsymbol{b}_2$！这个结论被称为“弗里希 - 沃 - 洛弗尔定理”（Frisch-Waugh-Lovell Theorem），而上述回归被称为“分块回归”（partitioned regression）或“偏回归”（partial regression）。证明如下。

上述回归的正规方程组 $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b}=\boldsymbol{X}'\boldsymbol{y}$ 可以写为，

$$\begin{pmatrix}\boldsymbol{X}_1'\\ \boldsymbol{X}_2'\end{pmatrix}(\boldsymbol{X}_1\ \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}_1'\\ \boldsymbol{X}_2'\end{pmatrix}\boldsymbol{y} \tag{3.31}$$

$$\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2\\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{y}\\ \boldsymbol{X}_2'\boldsymbol{y}\end{pmatrix} \tag{3.32}$$

$$\begin{cases}\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{b}_1+\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2=\boldsymbol{X}_1'\boldsymbol{y}\\ \boldsymbol{X}_2'\boldsymbol{X}_1\boldsymbol{b}_1+\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2=\boldsymbol{X}_2'\boldsymbol{y}\end{cases} \tag{3.33}$$

由此方程组的第一个方程可得，

$$\boldsymbol{b}_1=(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}-(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2 \tag{3.34}$$

从上式可知，如果 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 正交，即 $\boldsymbol{X}_1'\boldsymbol{X}_2=0$，则 $\boldsymbol{b}_1=(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$，相当于将 $\boldsymbol{y}$ 单独对 $\boldsymbol{X}_1$ 进行回归。把 $\boldsymbol{b}_1$ 的表达式代入该方程组的第二个方程，求解 $\boldsymbol{b}_2$ 可得，

---

① 当唯一的解释变量为常数项时，数据矩阵为单位向量，即 $\boldsymbol{X}=\mathbf{1}_{n\times1}\equiv(1\ \ 1\ \cdots\ 1)'$，故最小二乘估计为 $b_1^*=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{y}=n^{-1}\sum_{i=1}^{n}y_i=\bar{y}$。

$$\begin{aligned}\boldsymbol{b}_2 &= [\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1')\boldsymbol{y}_2] \\ &\equiv [\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}_2] \\ &= [\boldsymbol{X}_2'\boldsymbol{M}_1'\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{M}_1'\boldsymbol{M}_1\boldsymbol{y}_2] \\ &= (\boldsymbol{e}_2'\boldsymbol{e}_2)^{-1}\boldsymbol{e}_2'\boldsymbol{e}_1\end{aligned} \tag{3.35}$$

其中，$\boldsymbol{M}_1 \equiv \boldsymbol{I}-\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$为以 $\boldsymbol{X}_1$ 为解释变量的消灭矩阵（$\boldsymbol{M}_1$ 为对称幂等矩阵，故 $\boldsymbol{M}_1'\boldsymbol{M}_1 = \boldsymbol{M}_1$）。因此，$\boldsymbol{M}_1\boldsymbol{X}_2$ 即为把 $\boldsymbol{X}_2$ 对 $\boldsymbol{X}_1$ 进行回归所得的残差矩阵 $\boldsymbol{e}_2$，而 $\boldsymbol{M}_1\boldsymbol{y}_2$ 为把 $\boldsymbol{y}$ 对 $\boldsymbol{X}_1$ 进行回归所得的残差 $\boldsymbol{e}_1$。由上式可知，$\boldsymbol{b}_2$ 正是把 $\boldsymbol{e}_1$ 对 $\boldsymbol{e}_2$ 进行回归所得的回归系数。上述过程也被称为“过滤掉 $\boldsymbol{X}_1$ 的影响”（partialing out or netting out the effects of $\boldsymbol{X}_1$）。

显然，如果 $\boldsymbol{X}_2$ 只包括一个变量 z，则可以用分块回归来计算此变量 z 的单个回归系数。因此，多元回归中的回归系数也被称为“偏回归系数”（partial regression coefficients）。此结果揭示了变量 z 的回归系数的含义，即表示“滤去$\boldsymbol{X}_1$ 影响的 z”对“滤去$\boldsymbol{X}_1$ 影响的 $\boldsymbol{y}$”的作用。此时，z 对 $\boldsymbol{X}_1$ 回归所得的 $\boldsymbol{e}_2$ 为残差向量。将两个残差向量（$\boldsymbol{e}_2$，$\boldsymbol{e}_1$）画成散点图，则可以直观地看出滤去 $\boldsymbol{X}_1$ 后 z 对于 $\boldsymbol{y}$ 的偏影响。该散点图被称为“新增变量图”（added-variable plot），反映新增变量对于被解释变量的偏影响。对于每个解释变量，都可以画出其相应的新增变量图。因此，它把多维空间中的多元回归关系通过一系列的二维散点图来直观地展现。在此散点图上，还可以画出 $\boldsymbol{e}_1$ 对 $\boldsymbol{e}_2$ 的回归直线，其斜率就是变量 z 的偏回归系数。如果此回归直线接近于水平线，则说明 z 对于 $\boldsymbol{y}$ 的偏影响很小（回归系数的绝对值很小）。另一方面，如果散点图中的多数散点偏离此回归直线较远，则说明 z 对于 $\boldsymbol{y}$ 的线性作用关系不强（回归系数在统计上较不显著）。具体应用参见第 4 章“Stata 简介”。

3.10 预　测

有时，建立计量模型的目的并不仅仅是参数估计与假设检验。在某些情形下（特别对于时间序列数据而言），还常常进行预测（prediction or forecasting），即给定解释向量$\boldsymbol{x}_0$ 的（未来）取值，预测被解释变量 y_0 的取值。假设以上计量模型对所有观测值都成立（包括外推到未来的观测值），则

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0 \tag{3.36}$$

显然，我们可以用 $\hat{y}_0 \equiv \boldsymbol{x}_0'\boldsymbol{b}$ 来对 y_0 作点预测，其中 $\boldsymbol{b}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计量。“预测误差”（prediction error）$(\hat{y}_0 - y_0)$可以写为

$$\hat{y}_0 - y_0 = \boldsymbol{x}_0'\boldsymbol{b} - \boldsymbol{x}_0'\boldsymbol{\beta} - \varepsilon_0 = \boldsymbol{x}_0'(\boldsymbol{b}-\boldsymbol{\beta}) - \varepsilon_0 \tag{3.37}$$

由于 $\boldsymbol{b}$ 是 $\boldsymbol{\beta}$ 的无偏估计，故 $\mathrm{E}(\hat{y}_0 - y_0) = \boldsymbol{x}_0'\mathrm{E}(\boldsymbol{b}-\boldsymbol{\beta}) - \mathrm{E}(\varepsilon_0) = 0$，因此 $\hat{y}_0$ 是 y_0 的“无偏预测”（unbiased predictor），即用 $\hat{y}_0$ 来预测 y_0 不会系统地高估或低估[①]。此预测量的方差为

$$\mathrm{Var}(\hat{y}_0) = \mathrm{Var}(\boldsymbol{x}_0'\boldsymbol{b}) = \boldsymbol{x}'_0\mathrm{Var}(\boldsymbol{b})\boldsymbol{x}_0 = \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 \tag{3.38}$$

这个方差反映的是，由于抽样误差$(\boldsymbol{b}-\boldsymbol{\beta})$所带来的预测量 $\hat{y}_0$ 的波动。特别地，如果我们精确地知道$\boldsymbol{\beta}$，则 $\mathrm{Var}(\hat{y}_0) = \mathrm{Var}(\boldsymbol{x}_0'\boldsymbol{\beta}) = 0$。然而，通常我们更关心的是预测误差$(\hat{y}_0 - y_0)$的方差

① 注意 $\hat{y}_0$ 与 y_0 都是随机变量。

$$\begin{aligned}\mathrm{Var}(\hat{y}_0 - y_0) &= \mathrm{Var}[\boldsymbol{x}_0'(\boldsymbol{b}-\boldsymbol{\beta}) - \varepsilon_0] \\ &= \mathrm{Var}(\varepsilon_0) + \mathrm{Var}[\boldsymbol{x}_0'(\boldsymbol{b}-\boldsymbol{\beta})] \\ &= \sigma^2 + \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\end{aligned} \tag{3.39}$$

其中，假设 ε_0 与 $\boldsymbol{b}$ 不相关（因为估计 $\boldsymbol{b}$ 没有用到 ε_0 的信息），故 $\mathrm{Cov}[\boldsymbol{x}_0'(\boldsymbol{b}-\boldsymbol{\beta}), \varepsilon_0] = 0$。从这个表达式可以清楚地看出，预测误差的方差有两个来源，即抽样误差 $\sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$（由于不能精确地知道参数 $\boldsymbol{\beta}$ 所导致），以及 y_0 本身的不确定性（ε_0 的方差 σ^2，即使精确地知道 $\boldsymbol{\beta}$ 也无济于事）。

如果假设扰动项服从正态分布，则 $\hat{y}_0 - y_0 \sim N(0, \sigma^2 + \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)$。由于 σ^2 未知，可以用 s^2 来估计，得到下面的 t 统计量：

$$\frac{\hat{y}_0 - y_0}{s\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t(n-K) \tag{3.40}$$

因此，y_0 的置信度为 $(1-\alpha)$ 的置信区间为

$$\left(\hat{y}_0 - t_{\alpha/2}\, s\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}, \hat{y}_0 + t_{\alpha/2}\, s\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right) \tag{3.41}$$

即这个区间以 $(1-\alpha)$ 的概率可以覆盖 y_0。

习　题

3.1 定义投影矩阵为 $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$，消灭矩阵为 $\boldsymbol{M} \equiv \boldsymbol{I}_n - \boldsymbol{P}$。证明：$\boldsymbol{PX} = \boldsymbol{X}, \boldsymbol{Pe} = \boldsymbol{0}, \boldsymbol{MX} = \boldsymbol{0}, \boldsymbol{P}' = \boldsymbol{P}, \boldsymbol{M}' = \boldsymbol{M}, \boldsymbol{P}^2 = \boldsymbol{P}, \boldsymbol{M}^2 = \boldsymbol{M}$。

3.2 假定一元回归方程为 $y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i (i = 1, \cdots, n)$。写出其正规方程组，并解出最小二乘估计值 b_1, b_2。

3.3 使用最小二乘法估计含截距项的回归方程 $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i (i = 1, \cdots, n)$。

证明：

（1）残差和 $\sum_{i=1}^n e_i = 0$，$\sum_{i=1}^n x_{ik} e_i = 0$ $(k = 2, \cdots, K)$；（提示：使用正规方程组。）

（2）被解释变量的均值等于其预测值的均值，即 $\bar{y} = \bar{\hat{y}}$；

（3）平方和分解公式，即被解释变量的变动（离差平方和）等于预测值的变动加上残差平方和，$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^n e_i^2$。[提示：$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$，并使用上面的结果。]

解释为什么在没有常数项的情况下，平方和分解公式不再成立？

3.4 证明在有常数项的情况下，$R^2 = [\mathrm{Corr}(y_i, \hat{y}_i)]^2 \equiv \dfrac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] \cdot \left[\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2\right]}$

3.5 证明 $R_{uc}^2 = \dfrac{\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}}$。

3.6 假设 n 阶对称矩阵 $\boldsymbol{A}$ 半正定。证明：$\boldsymbol{A}$ 的任何主对角线元素均为非负。这个结论有助于理解高斯－马尔可夫定理，即 $\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) - \mathrm{Var}(\boldsymbol{b} \mid \boldsymbol{X})$ 为半正定矩阵。

3.7 在假定3.1—3.4之下，是否存在 $\boldsymbol{\beta}$ 的一个线性估计量（不一定是无偏估计），其方差小于OLS估计量 $\boldsymbol{b}$ 的方差？如果存在，此方差最小能取什么值？

3.8 证明：$\mathrm{Var}(s^2 \mid \boldsymbol{X}) = \dfrac{2\sigma^4}{n-K}$。[提示：如果随机变量服从 $\chi^2(m)$ 分布，则其期望为 m，而方差为 $2m$。]

3.9 （单边 t 检验）对于原假设 $H_0: \beta_k = \bar{\beta}_k$，替代假设 $H_1: \beta_k > \bar{\beta}_k$，写出其检验步骤。

附　录

A3.1　约束最小二乘法(RLS)

考虑有约束的最小二乘问题:

$$\min_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})$$

$$s.t.\ \boldsymbol{R}\tilde{\boldsymbol{\beta}}=\boldsymbol{r}$$

其中,$\boldsymbol{r}$ 为 $\boldsymbol{m}$ 维列向量,$\boldsymbol{R}$ 为 $m\times K$ 矩阵,$\mathrm{rank}(\boldsymbol{R})=m$(矩阵 $\boldsymbol{R}$ 满行秩)。引入拉格朗日函数:

$$L(\tilde{\boldsymbol{\beta}},\lambda)\equiv(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})-\lambda'(\boldsymbol{r}-\boldsymbol{R}\tilde{\boldsymbol{\beta}}) \tag{3.42}$$

其中,λ 为 m 维拉格朗日乘子列向量。一阶条件为

$$\frac{\partial L(\tilde{\boldsymbol{\beta}},\lambda)}{\partial\tilde{\boldsymbol{\beta}}}=-2\boldsymbol{X}'\boldsymbol{y}+2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}^*+\boldsymbol{R}'\lambda^*=\boldsymbol{0}_{K\times 1} \tag{3.43}$$

$$\frac{\partial L(\tilde{\boldsymbol{\beta}},\lambda)}{\partial\lambda}=-(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}^*)=\boldsymbol{0}_{m\times 1} \tag{3.44}$$

其中,$\boldsymbol{b}^*$ 与 λ^* 分别表示 $\tilde{\boldsymbol{\beta}}$ 与 λ 的最优值。为了建立"约束 OLS 估计量"$\boldsymbol{b}^*$ 与"无约束 OLS 估计量"$\boldsymbol{b}$ 之间的关系,在方程(3.43)两边同时左乘 $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 可得,

$$-2\underbrace{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}_{=\boldsymbol{b}}+2\underbrace{\boldsymbol{R}\boldsymbol{b}^*}_{=\boldsymbol{r}}+[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']\lambda^*=\boldsymbol{0} \tag{3.45}$$

由于矩阵 $\boldsymbol{R}$ 满行秩,故 $[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}$ 存在。根据方程(3.38),$\boldsymbol{R}\boldsymbol{b}^*=\boldsymbol{r}$。因此,从上式求解 λ^* 可得

$$\lambda^*=-2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}) \tag{3.46}$$

将 λ^* 的表达式代入方程(3.43),并在方程两边同时左乘 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 可得,

$$\boldsymbol{b}^*=\boldsymbol{b}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}) \tag{3.47}$$

这就是"约束 OLS 估计量"(RLS)。从这个式子可以看出,约束 OLS 与无约束 OLS 之差$(\boldsymbol{b}^*-\boldsymbol{b})$,是$(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b})$的线性函数,而$(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b})$衡量的是无约束 OLS 估计量 $\boldsymbol{b}$ 偏离约束条件 $\boldsymbol{R}\tilde{\boldsymbol{\beta}}=\boldsymbol{r}$ 的程度。如果 $\boldsymbol{b}$ 恰好满足这些约束,则$\boldsymbol{b}^*=\boldsymbol{b}$。

记 $\boldsymbol{e}$ 为无约束的残差向量,$\boldsymbol{e}^*$ 为有约束的残差向量,则

$$\boldsymbol{e}^*=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}^*=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}-\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b})=\boldsymbol{e}-\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b}) \tag{3.48}$$

$$\begin{aligned}\boldsymbol{e}^{*\prime}\boldsymbol{e}^*&=[\boldsymbol{e}-\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b})]'[\boldsymbol{e}-\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b})]\\&=\boldsymbol{e}'\boldsymbol{e}-(\boldsymbol{b}^*-\boldsymbol{b})'\underbrace{\boldsymbol{X}'\boldsymbol{e}}_{=\boldsymbol{0}}-\underbrace{\boldsymbol{e}'\boldsymbol{X}}_{=\boldsymbol{0}}(\boldsymbol{b}^*-\boldsymbol{b})+(\boldsymbol{b}^*-\boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b})\\&=\boldsymbol{e}'\boldsymbol{e}+(\boldsymbol{b}^*-\boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}^*-\boldsymbol{b})\end{aligned} \tag{3.49}$$

其中,根据无约束 OLS 的性质,$\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0}$。根据方程(3.47),

$$\boldsymbol{b}^*-\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}) \tag{3.50}$$

将上式代入方程(3.49)可得

$$\boldsymbol{e}^{*\prime}\boldsymbol{e}^*-\boldsymbol{e}'\boldsymbol{e}=(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{r}) \tag{3.51}$$

根据 F 检验定理,F 统计量为 $F=\dfrac{(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{r})/m}{\boldsymbol{e}'\boldsymbol{e}/(n-K)}$。根据方程(3.51),$F$ 统计量可以写为 $F=\dfrac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^*-\boldsymbol{e}'\boldsymbol{e})/m}{\boldsymbol{e}'\boldsymbol{e}/(n-K)}$,这正是似然比原理的 F 统计量表达式。

第 4 章　Stata 简介

4.1　为什么使用 Stata

Stata 是目前在欧美最为流行的计量软件，具有操作简单、功能强大的特点。由于使用 Stata 的用户很多，对于最新的计量方法，常常可以下载由用户写的 Stata 命令程序（user-written Stata commands）[①]，十分方便。而官方的 Stata 版本也经常更新，以适应计量经济学迅猛发展的需要。尽管 Stata 13 已于 2013 年 6 月发布，但由于在中国普遍使用的仍是 Stata 12 或更低版本，故本书主要介绍 Stata 12。

4.2　Stata 的窗口

安装好 Stata 后，点击计算机桌面上的 Stata 图标，即可打开 Stata。此时可以看到，在最上方有一排菜单，即“File Edit Data Graphics Statistics User Window Help”。在菜单之下，则为一系列图标，起着快捷键的作用。在图标之下，有五个窗口，分别为（如图 4.1）

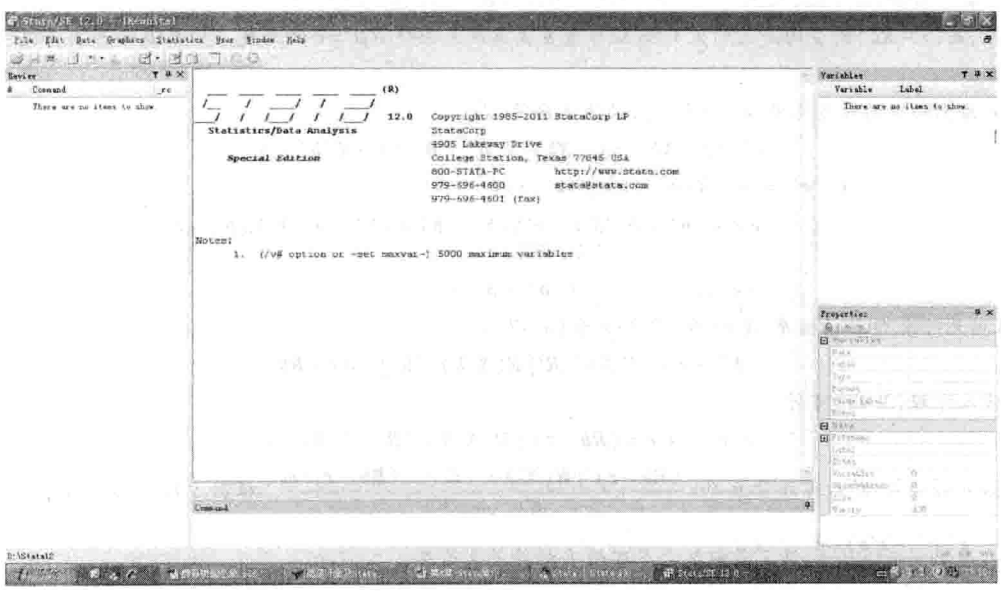


图 4.1　Stata 12 的主要窗口

① 称为 Stata“外部命令”或“非官方命令”。

左上“Review”(历史窗口):此窗口记录着自启动 Stata 以来执行过的命令。

中上“Results”(结果窗口):此窗口显示执行 Stata 命令后的输出结果。

中下“Command”(命令窗口):在此窗口输入想要执行的 Stata 命令。

右上“Variables”(变量窗口):此窗口记录着目前 Stata 内存中的所有变量。

右下“Properties”(性质窗口):此窗口显示当前数据文件与变量的性质。

为了使屏幕分割更美观实用,可以用鼠标将以上窗口拉到任意大小与位置。然后点击菜单“Edit”→“Preferences”→“General Preferences”→“Windowing”→“Lock splitter”,就可以锁定当前画面,而在以后重启 Stata 时自动显示这个画面设置。

4.3 Stata 操作实例

学习 Stata 的最方便方法大概是通过实例来学习。因此,这里选取 Nerlove(1963)对电力行业规模报酬的经典研究来介绍 Stata 的实际操作①。从作者的个人网页(www. econ. sdu. edu. cn/tree/Faculty. php)可以下载 Nerlove(1963)论文原文,以及数据集“nerlove. xls”(Excel 文件)。该数据集包括了 1955 年美国 145 家电力企业的横截面数据。

1. 将数据导入 Stata

打开 Stata 软件后,点击 Data Editor (Edit)图标②(也可以点击菜单“Window”→“Data Editor”),即可打开一个类似 Excel 的空白表格。然后,用 Excel 打开文件“nerlove. xls”,复制文件中的所有数据,并粘贴到 Data Editor 中。此时,Stata 会问你“第一行为数据还是变量名”(Is the first row data or variable names?),点击相应的选择即可(对于此数据集,因为 Excel 表的第一行为变量名,故应选“Treat first row as variable names”)。

导入数据的另一方法是(特别在数据量很大的情况下),点击菜单“File”→“Import”,然后导入各种格式的数据。但这种方法有时不如直接从 Excel 表中粘贴数据方便直观。

关闭 Data Editor (Edit)后,即会看到右上方的“Variables”窗口出现了 5 个变量,分别为 tc(total cost,总成本),q(total output,总产量),pl(price of labor,小时工资率),pf(price of fuel,燃料价格),与 pk(user cost of capital,资本的租赁价格③)。

此时,可以点击 Save 图标④(也可以点击菜单“File”→“Save”),将数据存为 Stata 格式的文件(扩展名为 dta),比如 nerlove. dta。这样,以后就可以用 Stata 直接打开这个数据集了(不需要再从 Excel 表中粘贴过来)。打开的方式有两种。可以点击 Open 图标⑤(也可以点击菜单“File”→“Open”),然后寻找要打开的 dta 文件的位置。另一种方法是在命令窗口输入以下命令(假设文件在 E 盘的根目录)并回车(按 Enter 键):

```
. use E:\nerlove.dta,clear
```

其中,逗号“,”之后的“clear”为“选择项”(options),表示可以替代内存中的已有数据。如果

---

① 此例来自 Hayashi(2000)。

② 看上去像一个微型表格(上面有只笔),排在所有图标的倒数第五个。

③ 以“电力企业发行的长期公司债利率”乘以“该企业所在地区的电力建筑成本指数”来计算,详见 Nerlove(1963), Appendix。

④ 看上去像一张微型软盘,排在所有图标的第二个。

⑤ 排在所有图标的第一个。

要关闭一个数据集(如果对数据集进行了改动,别忘了先存盘Save),以便使用另外一个数据集,可以在命令窗口输入

```
. clear
```

这样,内存中所有的当前数据都被清空,然后可以再打开另外一个数据集。

2. 日期数据的导入(可暂时跳过此部分)

如果数据中含有格式为"1949-10-01"或"1949/10/01"的时间变量,在导入Stata后,可能被视为"字符串"(string),而非"数字"(numeric),无法直接对其进行运算。

对于日度数据(daily data),可以使用命令"gen newvar = date (varname,"YMD")"将其转换为"整数日期变量"(integer date variable),其中,命令"generate"表示生成新变量(可缩写为gen或g),函数"date"表示转换为日期变量,而"YMD"告诉Stata,原始数据的格式为"年-月-日"。如果原始数据的格式为"月-日-年",则应该为"MDY",以此类推。然后,可以用命令"format newvar %td"让该时间变量仍然以日期格式在Stata中显示。在Stata内部,所有日期变量的存储格式均为"elapsed dates",即计算从1960年1月1日以来过了多少天。

类似地,对于月度数据(monthly data),可以使用命令"gen newvar = monthly (varname,"YM")"进行转换;其中,"YM"告诉Stata,原始数据的格式为"年-月"。然后用命令"format newvar %tm"让该变量仍以日期格式在Stata中显示。此时,Stata内部的日期变量存储格式为"elapsed months",即计算从1960年1月以来过了多少月。

对于季度数据(quarterly data),可以使用命令"gen newvar = quarterly (varname,"YQ")"进行转换;其中,"YQ"告诉Stata,原始数据的格式为"年-季"。然后用命令"format newvar %tq"让该变量仍以日期格式在Stata中显示。此时,Stata内部的日期变量存储格式为"elapsed quarters",即计算从1960年第1季度以来过了多少季度。

如果在原始数据中,年、月、日分别以数字(numeric)变量"Y,M,D"来表示,则可用以下命令将其合成为单一的日期变量,"gen newvar = mdy(M,D,Y)"。

在Stata中,还可以定义年度数据、半年度数据、周数据、时钟数据(可精确到千分之一秒,适用于高频数据,比如每时每刻变化的股票价格),相应的命令为"yearly,halfyearly,weekly,clock"。更多相关说明,参见"help date"。

3. 变量的标签

在变量窗口,每个变量的"名字"(Name)旁边显示了其"标签"(label)。但目前的标签过于简略,缺乏变量的解释信息。点击倒数第3个图标,即可打开变量管理器(Variables Manager)(或点击菜单"Data"→"Variables Manager"),然后很方便地对同时编辑所有变量的变量名、标签以及变量的存储格式。你可以试着把tc,q,pl,pf与pk的标签分别改为"total cost","total output","price of labor","price of fuel"与"user cost of capital"。

Stata中字母的大小写是严格区分的(case sensitive),因此Stata建议对于变量名一律使用小写字母。

4. 审视数据

一个数据集可能很大,而我们常希望看到数据的概貌。想看数据集中的变量名单、标签等,可以在命令窗口输入

```
. describe
```

其中,"describe"中的下划线表示,可将该命令简写为"d"而得到同样的效果。如果想给整个数据集加上一个标签,以说明此数据集来自"Nerlove 1963 paper",可输入命令:

```
. label data "Nerlove 1963 paper"
```

再次使用命令“describe”,就会看到数据集的标签“Nerlove 1963 paper”。运行结果如下:

```
Contains data
  obs:           145                          Nerlove 1963 paper
 vars:             5
 size:         2,320
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
tc              float  %8.0g                  total cost
q               int    %8.0g                  total output
pl              float  %8.0g                  price of labor
pf              float  %8.0g                  price of fuel
pk              int    %8.0g                  user cost of capital
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

如果想看变量 tc 与 q 的具体数据,可使用命令:

```
. list tc q
```

由于样本容量为 145,要花一些时间才能显示完毕。如果想中途停止该命令的执行,则可以点击 Break 图标[①],或直接在键盘上同时按“Ctrl + Break”。运行结果如下:

```
     +-------------+
     |     tc    q |
     |-------------|
  1. |   .082    2 |
  2. |   .661    3 |
  3. |    .99    4 |
  4. |   .315    4 |
  5. |   .197    5 |
     |-------------|
  6. |   .098    9 |
  7. |   .949   11 |
  8. |   .675   13 |
  9. |   .525   13 |
 10. |   .501   22 |
     |-------------|
 11. | 1.194   25 |
 12. |   .67   25 |
 13. |   .349   35 |
 14. |   .423   39 |
 15. |   .501   43 |
     |-------------|
 16. |    .55   63 |
     +-------------+
--Break--
r(1);
```

如果改变主意,仍然希望显示变量 tc 与 q 的全部数据,那么只要把光标放在命令窗口,并按键盘上的“Page Up”键即可调用上一个命令(反之,使用“Page Down”键可调用下一个命令)。另一种简便的方法是,在左上角的历史窗口点击任何曾用过的命令:如果用鼠标单击旧命令,则会把旧命令重新调入命令窗口,按回车后即执行,或将旧命令进行编辑后再执行;如果用鼠标双击旧命令,则将马上自动执行。这两种方法可以节省写命令的时间。

有时我们想对数据集的一部分执行命令,比如只想看变量 tc 与 q 的前 5 个数据,则可输入命令

① 所有图标中的最后一个,图形为圆形中间带一个叉。

```
. list tc q in 1/5
```

	tc	q
1.	.082	2
2.	.661	3
3.	.99	4
4.	.315	4
5.	.197	5

同理,如果要罗列从第32~36个观测值,则可输入命令

```
. list tc q in 32/36
```

	tc	q
32.	3.154	214
33.	2.599	220
34.	3.298	234
35.	2.441	235
36.	2.031	253

也可以通过逻辑关系来定义数据集的子集。如果要列出所有满足条件"$q \geq 10\ 000$"的变量tc与q的数据[①],则可以使用以下命令

```
. list tc q if q >=10000
```

	tc	q
142.	67.12	11477
143.	73.05	11796
144.	139.422	14359
145.	119.939	16719

其中,">="表示"大于等于"。其他表示关系的逻辑符号为"=="(等于)[②],">"(大于),"<"(小于),"<="(小于等于),"~="(不等于)。查看具体数据的一个直接方法是,点击Data Editor (Edit)图标,或者点击该图标右边的Data Editor (Browse)图标[③]。

如果想删除满足"$q \geq 10\ 000$"条件的观测值,则可使用命令

```
.drop if q >=10000
```

反之,如果只想保留满足"$q \geq 10\ 000$"条件的观测值,而删去所有其他观测值,则可使用命令

```
.keep if q >=10000
```

5. 考察变量的统计特征

如果想看变量q的统计特征,可输入命令

```
. summarize q
```

Variable	Obs	Mean	Std. Dev.	Min	Max
q	145	2133.083	2931.942	2	16719

这将显示变量q的样本容量、平均值、标准差、最小值与最大值。如果要计算满足条件"$q \geq$

① Nerlove(1963)的数据按照总产量q的升序排列,故这里显示的是最后四个观测值。

② 正如在一般计算机程序中的约定,如果只有一个等号"=",则表示赋值,而非"等于"。

③ 看上去像一个微型表格(上面有放大镜),在所有图标中排在倒数第四位。

10 000”的子样本的统计指标,则可使用命令

```
. su q if q > =10000
```

Variable	Obs	Mean	Std. Dev.	Min	Max
q	4	13587.75	2453.921	11477	16719

如果想看更多的统计指标,则可使用命令

```
. su q,detail
```

total output

Percentiles		Smallest		
1%	3	2		
5%	13	3		
10%	43	4	Obs	145
25%	279	4	Sum of Wgt.	145
50%	1109		Mean	2133.083
		Largest	Std. Dev.	2931.942
75%	2507	11477		
90%	5819	11796	Variance	8596285
95%	8642	14359	Skewness	2.398202
99%	14359	16719	Kurtosis	9.474916

新增的统计指标有百分位数(percentiles),方差(variance),偏度(skewness)与峰度(kurtosis)[①]。如果不指明变量,则将显示数据集中所有变量的统计指标。

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
tc	145	12.9761	19.79458	.082	139.422
q	145	2133.083	2931.942	2	16719
pl	145	1.976552	.2300404	1.5	2.3
pf	145	26.17655	7.876071	10.3	42.8
pk	145	174.4966	18.20948	138	233

如果要显示变量 pl 的经验累积分布函数(empirical cumulative distribution function),可使用命令

```
. tabulate pl
```

price of labor	Freq.	Percent	Cum.
1.5	7	4.83	4.83
1.6	4	2.76	7.59
1.7	15	10.34	17.93
1.8	26	17.93	35.86
1.9	12	8.28	44.14
2	12	8.28	52.41
2.1	32	22.07	74.48
2.2	17	11.72	86.21
2.3	20	13.79	100.00
Total	145	100.00	

① 有关“偏度”与“峰度”的定义,参见第 2 章。

如果要显示内存中3个价格变量之间的相关系数,可输入命令,

```
. pwcorr pl pf pk,sig star(.05)
```

其中,"pwcorr"表示"pairwise correlation"(两两相关),选择项"sig"表示显示相关系数的显著性水平(即 p 值,列在相关系数的下方),选择项"star(.05)"表示给所有显著性水平小于或等于5%的相关系数打上星号。如果命令 pwcorr 之后没有指定变量,则显示数据集中所有变量的相关系数。

	pl	pf	pk
pl	1.0000		
pf	0.3310*	1.0000	
	0.0000		
pk	-0.1845*	0.1254	1.0000
	0.0263	0.1328	

结果显示,pf 与 pl 的相关系数为 0.331,且在5%水平上显著不为零(p 值为 0.000 0);pk 与 pl 的相关系数为 -0.184 5,且在5%水平上显著(p 值为 0.026 3);pk 与 pf 的相关系数为 0.125 4,但在5%水平上不显著(p 值为 0.132 8)。

6. 画图

Stata 具有很强的画图功能。如果想看变量 q 的直方图(假定组宽为 1 000),可输入以下命令,其运行结果如图 4.2:

```
. histogram q,width(1000) frequency
```

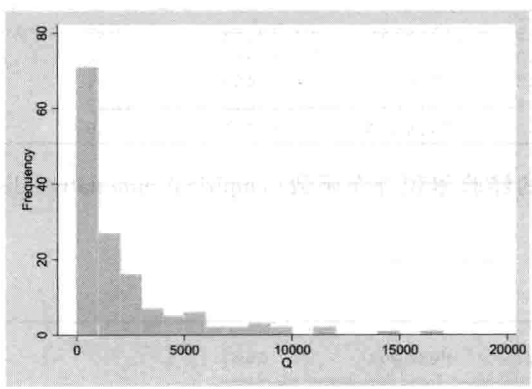


图 4.2 直方图

其中,逗号",",之后的"width(1000)"与"frequency"都是"选择项"(options),分别表示将组宽设为 1 000,将纵坐标定为频数(落入每组的个体数)。由于直方图不连续,如果想看连续的经验分布图[①],可使用以下命令,结果如图 4.3:

```
. kdensity q
```

如果要画 tc 与 q 之间的散点图,则可输入以下命令,结果如图 4.4:

```
. scatter tc q
```

① 即核密度图,参见第27章。

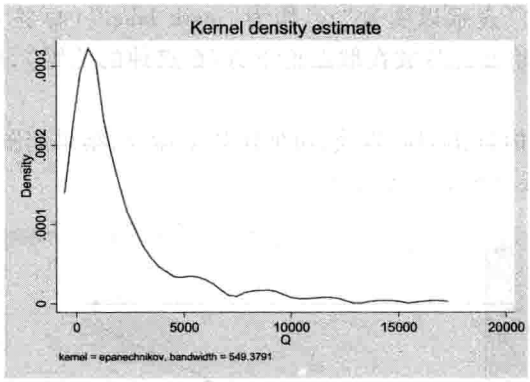


图 4.3 核密度图

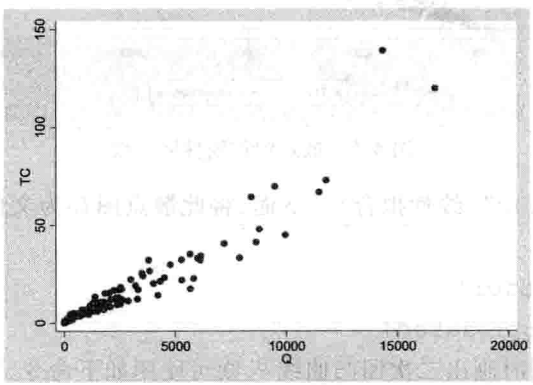


图 4.4 散点图

然而,在上面的散点图中,我们无法知道其中的每个点分别对应于哪个观测值。为此,首先定义一个新变量“n”来表示第 n 个观测值。

```
. gen n = _n
```

其中,“_n”即表示第 n 个观测值。然后运行以下命令,结果见图 4.5。

```
. scatter tc q,mlabel(n) mlabpos(6)
```

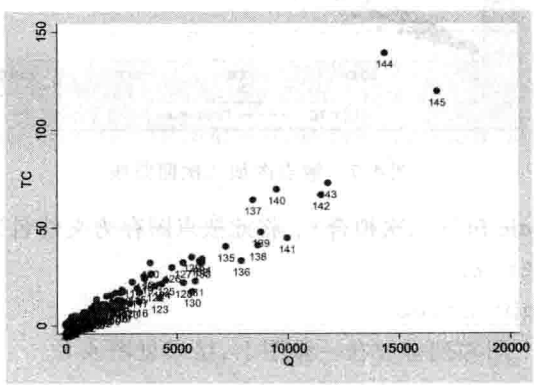


图 4.5 带标签的散点图

其中，选择项“mlabel(n)”表示以变量“n”作为“mark label”（标签）；选择项“mlabpos(6)”（mark label position）表示将此标签放在散点正下方（6点钟的位置），默认位置为散点的右边（3点钟的位置）。

如果想在散点图上同时画出回归直线，可使用如下命令，结果如图4.6：

```
. twoway (scatter tc q)(lfit tc q)
```

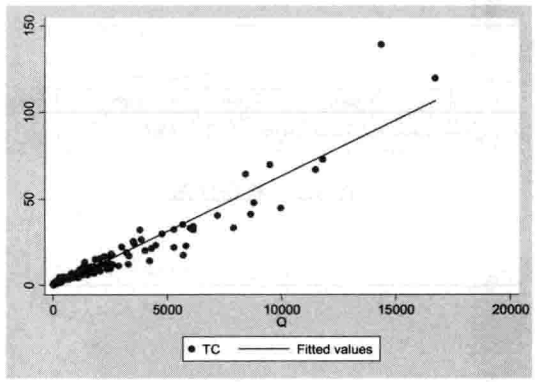


图4.6 散点图加线性回归线

其中，“lfit”表示“linear fit”（线性拟合）。下面，将此散点图存为文件名为“scatter1”的图像文件，以便以后调用。

```
. graph save scatter1
(file scatter1.gph saved)
```

如果想在散点图上同时画出二次回归曲线[①]，则可使用如下命令，结果如图4.7：

```
. twoway (scatter tc q)(qfit tc q)
```

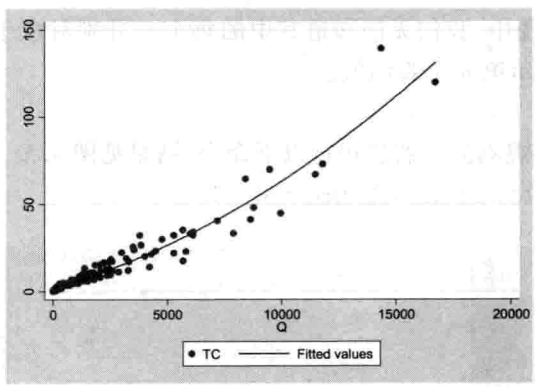


图4.7 散点图加二次回归线

其中，“qfit”表示“quadratic fit”（二次拟合）。将此散点图存为文件名为“scatter2”的图像文件。

```
. graph save scatter2
(file scatter2.gph saved)
```

下面，我们将上述两个图并列排放在一张图上，结果见图4.8。

① 即 $y=\alpha+\beta x+\gamma x^2+\varepsilon$。

shuangqing

```
. graph combine scatter1.gph scatter2.gph
```

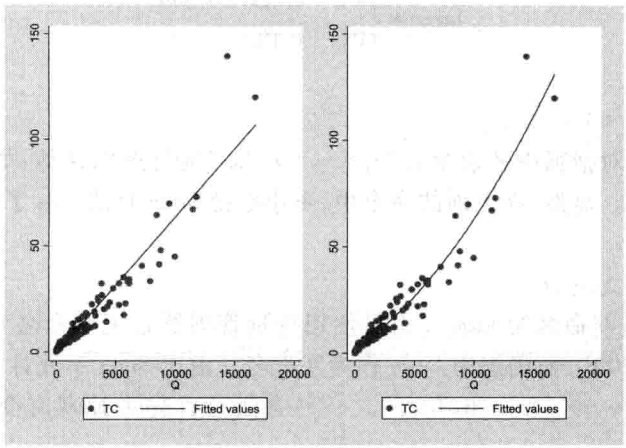


图 4.8 散点图加二次回归线

更多的作图方法及选择项，参见菜单“Graphics”。对于任何命令，只要输入“help command”（比如，help histogram），即可看到对该命令的详细说明。

7. 生成新变量

Nerlove（1963）假设第 i 个企业的生产函数为 Cobb-Douglas 函数：

$$Q_i = A_i L_i^{\alpha_1} K_i^{\alpha_2} F_i^{\alpha_3} \tag{4.1}$$

其中，A, L, K, F 分别为生产率、劳动力、资本与燃料。记 $r \equiv \alpha_1 + \alpha_2 + \alpha_3$ 为规模效应（degree of returns to scale）。如果 $r = 1$，则规模报酬不变；如果 $r > 1$，则规模报酬递增；如果 $r < 1$，则规模报酬递减。Nerlove（1963）的主要目的就是要确定 20 世纪 50 年代美国电力行业的规模经济。假设企业追求成本最小化，可以证明成本函数也为 Cobb-Douglas 函数[①]：

$$TC_i = \delta_i Q_i^{1/r} (P_{L,i})^{\alpha_1/r} (P_{K,i})^{\alpha_2/r} (P_{F,i})^{\alpha_3/r} \tag{4.2}$$

其中，δ_i 是 $A_i, \alpha_1, \alpha_2, \alpha_3$ 的函数。取自然对数后得到如下模型，

$$\ln TC_i = \beta_1 + \frac{1}{r}\ln Q_i + \frac{\alpha_1}{r}\ln P_{L,i} + \frac{\alpha_2}{r}\ln P_{K,i} + \frac{\alpha_3}{r}\ln P_{F,i} + \varepsilon_i \tag{4.3}$$

为了估计这个方程，需要在 Stata 中对原变量取自然对数，可使用命令generate。

```
. g lntc = log(tc)
. g lnq = log(q)
. g lnpl = log(pl)
. g lnpf = log(pf)
. g lnpk = log(pk)
```

如果需要 q 的非线性平方项，则可以使用命令

```
. g q2 = q^2
```

如果要生成 lnpl 与 lnpk 的互动项（interaction term），则可以使用命令

```
. g lnplpk = lnpl * lnpk
```

假设希望定义“$q \geq 10\ 000$”为大企业，并使用“虚拟变量”（dummy variable，也称为“哑变

① 参见标准的微观经济学教材。

量”)large来表示,即

$$\text{large} \equiv \begin{cases} 1, & \text{若 } q \geqslant 10\,000 \\ 0, & \text{其他} \end{cases} \tag{4.4}$$

则可使用命令

```
. g larg = (q > =10000)
```

其中,括弧“( )”表示对括弧中的表达式“q > =10000”进行逻辑评估:如果为真,则取值为1;如果为假,则取值为0。显然,在上面的命令中,不小心把large打成larg了。为此,可以将变量重新命名:

```
. rename larg large
```

这样,变量larg就被重新命名为large(也可使用变量管理器来重新命名)。假设我们想改变大企业的定义为“$q \geqslant 6\,000$”,而仍想用large作为变量名。由于Stata不允许变量名重复,故无法直接使用命令“g large = (q > =6000)”。一种方法是首先去掉现有变量large,然后再定义一次:

```
. drop large
. g large = (q > =6000)
```

更简洁的方法则只需使用一个命令

```
. replace large = (q > =6000)
```

该命令将原来的变量($q \geqslant 10\,000$)直接替换为新变量($q \geqslant 6\,000$)。

在执行Stata命令时,有时需要调用许多变量,而某些变量名可能很长。此时,如果在命令窗口一一输入变量名,可能较费事。解决方法之一是,可以直接在右上角的变量窗口单击需要的变量,则该变量名就会显现在命令窗口。解决方法之二是,如果有以下变量lnq1,lnq2,…,lnq30,而只想使用其中的前15个变量,则可以用lnq1—lnq15来简略地表示这15个变量。解决方法之三是,用“*”号来节省变量名的书写。假设想将内存中所有以“ln”开头的变量都去掉,则可输入命令

```
. drop ln*
```

这将去掉内存中的lntc,lnq,lnpl,lnpf,lnpk变量。如果你后悔删除这些变量,Stata并没有类似文字编辑器Word的“undo”命令,无法撤销此命令。唯一的弥补方法是,重新使用命令generate,再去生成这些变量(当然,可以从历史窗口直接点击旧命令)。

8. Stata的计算器功能

Stata也可以作为计算器来使用。只要输入命令“display expression”即可。比如,如果计算ln2,可输入以下命令,

```
. display log(2)
. 69314718
```

如果要计算标准正态变量小于1.96的概率,则可使用命令,

```
. di normal(1.96)
. 9750021
```

其中,“normal”表示标准正态分布的累积分布函数(cdf)。有关常见概率分布的累积分布函数、密度函数等,参见“help density function”。

9. 线性回归分析

使用OLS对方程(4.3)进行估计,可输入命令

```
. regress lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs = 145
Model	269.524728	4	67.3811819	F(4, 140) = 437.90
Residual	21.5420958	140	.153872113	Prob > F = 0.0000
				R-squared = 0.9260
Total	291.066823	144	2.02129738	Adj R-squared = 0.9239
				Root MSE = .39227

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

上表中的“_cons”表示常数项，“R-squared”显示 $R^2=0.926\ 0$，“Adj R-squared”显示 $\bar{R}^2=0.923\ 9$。检验整个方程显著性的 F 统计量之 p 值（Prob > F）为 0.000 0，显示这个回归方程是高度显著的。然而，lnpl 与 lnpk 这两个变量均不显著，其 p 值（P > |t|）分别为 0.131 与 0.528。特别地，变量 lnpk 的系数（Coef.）符号为负，与经济理论的预测相反。Nerlove（1963）认为，这是由于“资本使用成本”的数据不太可靠。表上方的回归结果还显示，残差平方和 $\sum_{i=1}^{n} e_i^2 = 21.542$，而方程的标准误差（Root MSE）为 $s=0.392$。

如果要显示估计系数的协方差矩阵，可输入命令

```
. vce
```

Covariance matrix of coefficients of regress model

e(V)	lnq	lnpl	lnpk	lnpf	_cons
lnq	.00030393				
lnpl	-.00035938	.08988127			
lnpk	.00034967	.02497537	.11548412		
lnpf	.00030089	-.01124831	-.00669535	.01006447	
_cons	-.00451909	-.15095534	-.59317676	.00784373	3.1662023

其中，“vce”表示“variance covariance matrix estimated”。

在进行回归时，如果不要常数项[①]，可以加上选择项“noconstant”。

```
. reg lntc lnq lnpl lnpk lnpf,noc
```

如果只对“大企业”这个子样本进行回归，则可以输入命令

```
. reg lntc lnq lnpl lnpk lnpf if q > =6000
```

或者使用虚拟变量 large：

```
. reg lntc lnq lnpl lnpk lnpf if large
```

即只对“large = 1”的子样本进行回归。如果想对“小企业”（即除了“大企业”以外的所有企业）进行回归，可以使用命令

```
. reg lntc lnq lnpl lnpk lnpf if large ==0
```

或者输入命令

① 对于这个回归而言，常数项本身在5%的水平上显著，而且经济理论也要求有常数项。因此，进行无常数项的回归仅仅起示范的作用。

```
. reg lntc lnq lnpl lnpk lnpf if ~large
```

其中,“~”表示逻辑的“否”(not)运算。

如果要计算被解释变量的拟合值($\hat{y}$),并将其记为 lntchat,可输入命令

```
. predict lntchat
```

如果要计算“残差”(residual),并将其记为 e1,可输入命令

```
. predict e1,residual
```

其中,选择项“residual”表示预测残差。如果没有任何选择项,则“默认值”(default)为计算拟合值 $\hat{y}$。由于 lnq 的系数为 $1/r$,即规模报酬的倒数。故可以估计规模报酬为

```
. display 1/_b[lnq]
1.387129
```

其中,“_b[lnq]”表示“lnq”的 OLS 系数估计值。

由于 $\hat{r}=1.387\ 129>1$,故认为可能存在规模报酬递增。为此,检验规模报酬不变的原假设“$H_0: r=1$”,输入命令

```
. test lnq=1
```

这个命令检验的原假设为,变量 lnq 的系数等于 1。F 检验的结果为

```
( 1)  lnq = 1

      F(  1,    140) =  256.27
           Prob > F =    0.0000
```

即以很小的 p 值拒绝原假设,故可以认为存在规模报酬递增。

方程(4.3)还显示,变量 lnpl,lnpk 与 lnpf 的系数之和应该等于 1(无论“$H_0: r=1$”成立与否)。为此,可以检验以下联合假设

```
. test (lnq=1)(lnpl+lnpk+lnpf=1)
```

```
( 1)  lnq = 1
( 2)  lnpl + lnpk + lnpf = 1

      F(  2,    140) =  128.15
           Prob > F =    0.0000
```

由于 p 值 =0.000 0,故强烈拒绝此联合假设。另外,由于 lnpl 与 lnpk 均不显著,我们希望对其显著性进行联合检验:

```
. test lnpl lnpk
```

```
( 1)  lnpl = 0
( 2)  lnpk = 0

      F(  2,    140) =    1.69
           Prob > F =    0.1874
```

结果显示,由于 p 值很大(0.187 4),故可以接受二者的系数皆为 0 的联合假设。

Stata 也可以检验“非线性假设”(参见第 5 章附录)。比如,检验变量 lnpl 的系数是 lnq 的系数的平方:

```
. testnl _b[lnpl]=_b[lnq]^2
```

```
(1)  _b[lnpl] = _b[lnq]^2

           F(1, 140) =       0.04
            Prob > F =       0.8334
```

由于 p 值很大(0.833 4),故我们无法拒绝这个原假设。

在进行回归后,如果要画变量 lnq 的新增变量图(参见第 3 章),可以使用以下命令,结果见图 4.9。

```
. avplot lnq
```

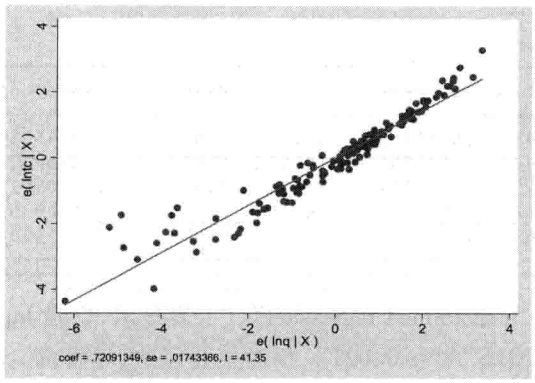


图 4.9 lnq 的新增变量图

从图 4.9 可以看出,lnq 对 lntc 的偏回归系数为正(即图中回归直线的斜率);而且散点比较集中地分布在回归直线附近,表明此关系很显著。如果要同时画出所有变量的新增变量图,则可用以下命令,结果见图 4.10。

```
. avplots
```

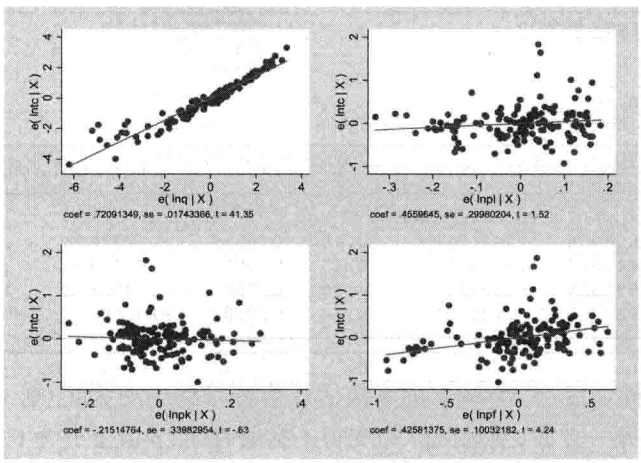


图 4.10 所有变量的新增散点图

从图 4.10 可以看出,解释变量 lnpl,lnpk,lnpf 与被解释变量 lntc 的偏关系都没有 lnq 与 lntc 的偏关系那么紧密(统计显著性更弱些),而且 lnpk 的偏回归系数为负数(图中回归线接近水平线,且斜率为负)。

10. 约束回归

Stata 也提供了进行“约束 OLS 回归”(RLS)的命令。由于经济理论要求变量 lnpl,lnpk 与 lnpl 的系数之和为 1,故考虑在此约束下重新估计方程(4.3)。首先定义“约束条件 1”如下:

```
. constraint def 1 lnpl + lnpk + lnpf = 1
```

然后进行有约束的OLS估计：

```
. cnsreg lntc lnq lnpl lnpk lnpf,c(1)
```

```
Constrained linear regression                   Number of obs   =        145
                                                Root MSE        =     0.3915

 ( 1)  lnpl + lnpk + lnpf = 1

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |   .7213365   .0173912    41.48   0.000     .6869553    .7557176
        lnpl |   .6064693    .207239     2.93   0.004      .196772    1.016167
        lnpk |  -.0208375   .1933394    -0.11   0.914    -.4030563    .3613813
        lnpf |   .4143682   .0987832     4.19   0.000     .2190805    .6096559
       _cons |  -4.636069   .8949922    -5.18   0.000    -6.405408   -2.866731
------------------------------------------------------------------------------
```

其中，“cnsreg”表示“constrained regression”。上表显示，变量lnpk的系数估计值从无约束OLS的“-0.22”变为约束OLS的“-0.021”，相对更合理些，尽管仍为负数，也依然不显著（p值为0.914）。

如果希望加上约束条件“$H_0: r = 1$”（由于前面已经拒绝了这个假设，故此处仅为演示目的），可以定义“约束条件2”如下：

```
. cons def 2 lnq = 1
```

然后在同时满足约束条件1，2的情况下进行回归：

```
. cnsreg lntc lnq lnpl lnpk lnpf,c(1 -2)
```

```
Constrained linear regression                   Number of obs   =        145
                                                Root MSE        =     0.6553

 ( 1)  lnpl + lnpk + lnpf = 1
 ( 2)  lnq = 1

------------------------------------------------------------------------------
        lntc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnq |          1  (constrained)
        lnpl |   .1558956   .3436328     0.45   0.651    -.5234015    .8351927
        lnpk |   .1443526   .3231175     0.45   0.656    -.4943898    .7830949
        lnpf |   .6997518   .1626168     4.30   0.000     .3782892    1.021214
       _cons |  -7.926918    1.45791    -5.44   0.000    -10.80893   -5.044905
------------------------------------------------------------------------------
```

加上两个约束条件之后，变量lnpk的系数估计值符号终于变为正（0.144 352 6），但依然不显著（p值为0.656）；另一方面，变量lnpl却变得不显著性了（p值上升为0.651）。

11. Stata的日志

如果希望在每次使用Stata时，储存其运行结果，可点击菜单“File”→“Log”→“Begin”，然后输入日志（log）的文件名，并存储在指定的位置。从此以后，你在Stata中的所有操作及其输出结果，都将被记录在此日志中，直至选择退出。

如果要暂时关闭日志（不再记录输出结果），可输入命令“log off”。如果要恢复使用日志，可输入命令“log on”。如果要彻底退出日志，只要输入命令“log close”即可。如果要查看日志文件中的内容，只要点击存储位置上的日志文件图标即可。

12. Stata 命令运行结果的存储与调用

所有的 Stata 命令可以分为两种，即 e - 类命令（e-class commands）与 r - 类命令（r-class commands）。e - 类命令为“估计命令”（estimation commands），比如“regress”；而所有其他命令为 r - 类命令，比如，“summarize”。r - 类命令的运行结果都存储在“r()”，可以通过输入“return list”来显示，比如

```
. summarize q
```

Variable	Obs	Mean	Std. Dev.	Min	Max
q	145	2133.083	2931.942	2	16719

```
. return list
```

```
scalars:
                  r(N) =  145
              r(sum_w) =  145
               r(mean) =  2133.08275862069
                r(Var) =  8596284.659770114
                 r(sd) =  2931.942131040467
                r(min) =  2
                r(max) =  16719
                r(sum) =  309297
```

上表列出了在运行命令“summarize q”之后，Stata 所存储的结果，其中包括未显示的“r(Var)”（方差）、“r(sum)”（求和）等。而且，我们可以调用这些结果来作进一步的计算。比如，为了计算“变异系数”（coefficient of variation，即标准差除以平均值），可使用以下命令：

```
. display r(sd)/r(mean)
1.3745093
```

写得更漂亮些，可以用命令：

```
. display "The coefficient of variation is " r(sd)/r(mean)
The coefficient of variation is 1.3745093
```

另一方面，e - 类命令的运行结果都存储在“e()”，可以通过输入“ereturn list”来显示，比如，

```
. reg tc q
```

Source	SS	df	MS	Number of obs = 145
Model	51190.3707	1	51190.3707	F(1, 143) = 1399.00
Residual	5232.46776	143	36.5906836	Prob > F = 0.0000
				R-squared = 0.9073
Total	56422.8385	144	391.825267	Adj R-squared = 0.9066
				Root MSE = 6.049

tc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
q	.0064307	.0001719	37.40	0.000	.0060908	.0067705
_cons	-.741095	.6219699	-1.19	0.235	-1.970538	.4883481

```
. ereturn list
```

```
scalars:
                  e(N) =  145
               e(df_m) =  1
               e(df_r) =  143
                  e(F) =  1399.000119377508
                 e(r2) =  .9072633016022542
               e(rmse) =  6.049023360142635
                e(mss) =  51190.37074066489
                e(rss) =  5232.467756451834
               e(r2_a) =  .9066147932218505
                 e(ll) =  -465.7241671687837
               e(ll_0) =  -638.1285147616334
               e(rank) =  2

macros:
            e(cmdline) : "regress tc q"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
            e(depvar) : "tc"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "regress_estat"

matrices:
                  e(b) :  1 x 2
                  e(V) :  2 x 2

functions:
             e(sample)
```

上表列出了运行命令 reg 后 Stata 存储的结果，包括标量(scalars)、宏(macros)[①]、矩阵(matrices，即系数矩阵 e(b)与协方差矩阵 e(V))以及函数(functions)[②]。

4.4 Stata 命令库的更新

由于 Stata 版本的不同(即使同为 Stata 12)，如果你发现本书中极少数命令无法运行，可在命令窗口输入，

```
. update all
```

这将更新你的 Stata 命令库(包括 Stata 的“ado”程序文件与其他可执行文件)。

Stata 用户还写了大量的外部命令或非官方命令(user-written software)，可以直接下载到 Stata 中使用。最流行的 Stata 非官方命令下载平台为“统计软件成分”(Statistical Software Components，简记 SSC)，是一个由 Boston College 维护的非官方统计程序集散地，网址为 http://ideas.repec.org/s/boc/bocode.html。在 Stata 中可输入如下相关命令，

```
. ssc new   (罗列 SSC 提供的最新非官方 Stata 命令及简介)
. ssc hot   (罗列 SSC 提供的最流行非官方 Stata 命令)
```

---

① “宏”是 Stata 编程中使用的一种缩写方式，它以一个简洁的字符串来代指一个通常更为复杂的字符串。

② 有时，我们可能只用数据集中的一个子样本进行估计。这里的函数“e(sample)”是个虚拟变量，即如果一个观测值在样本中，则取值为 1；反之，则取值为 0。

. ssc install newcommand （安装 SSC 提供的非官方命令“newcommand”）

. help ssc （有关 SSC 提供的帮助信息）

如果使用“ssc install newcommand”来下载非官方程序，则所有下载与安装过程将自动完成（包括新命令的帮助文件）。如果你要使用某种估计方法，但不知道它是否存在，可使用以下命令来搜索。

. search keyword （搜索帮助文件、常见问题（FAQs）①、例子②、*Stata Journal*（SJ）③，*Stata Technical Bulletin*（STB）④等）

. findit keyword （搜索以上内容，以及 Stata 的网络资源）

命令 findit 的搜索范围比命令 search 更广些。事实上，findit 等价于“search, all”。而且，命令 search 的搜索结果（通常比较少）直接在 Stata 的结果窗口显示，而 findit 的搜索结果（通常比较多）将打开另一页面显示。

发现有用的非官方命令之后，如果该命令不来自 SSC，则一般需要自行安装。此时，需要将所有相关文件下载到指定的 Stata 文件夹中（通常是 ado\plus\）。如果不清楚应把文件复制到哪个文件夹，可在 Stata 中输入以下命令，以显示 Stata 的系统路径（system directories）：

. sysdir

此时，你会看到类似于以下的结果（取决于 Stata 的安装位置），

```
   STATA:  D:\Stata12\
 UPDATES:  D:\Stata12\ado\updates\
    BASE:  D:\Stata12\ado\base\
    SITE:  D:\Stata12\ado\site\
    PLUS:  c:\ado\plus\
PERSONAL:  c:\ado\personal\
OLDPLACE:  c:\ado\
```

然后，将下载的新命令文件复制到 PLUS 所指示的文件夹即可（此处为“c:\ado\plus\”）。

4.5 进一步学习 Stata 的资源

更多有关 Stata 的知识，将在本书的以后章节中逐步介绍。

有关 Stata 的英文参考书包括 Baum（2006），Cameron and Trivedi（2010），以及 Stata 出版社（Stata Press）出版的系列书籍⑤。加州大学洛杉矶分校（UCLA）网站（http://www.ats.ucla.edu/stat/stata/）提供了大量有关 Stata 的资源及实例（搜索“Stata UCLA”即可找到此网站）。

中文参考书包括陈传波《Stata 十八讲》⑥，胡咏梅（2010），兰草（2012），劳伦斯·汉密尔顿（2008），李春涛、张璇（2009），王群勇（2007，2008），王天夫、李博柏（2008），杨菊华（2012），张鹏伟、李嫣怡（2011）等。

Stata 本身的“帮助”（Help）菜单包含了详细的信息。在使用 Stata 命令时（比如，reg），宜养

① 网址为：http://www.stata.com/support/faqs/。

② 网址为：http://www.stata.com/links/examples-and-datasets/。

③ 网址为：http://www.stata.com/bookstore/stata-journal/。

④ 网址为：http://www.stata.com/products/stb/。

⑤ 网址为：http://www.stata.com/bookstore/books-on-stata/

⑥ 可自行搜索下载免费电子书。

成习惯,经常看其相应的帮助信息(输入命令"help reg"即可)。更进一步的学习,则可查看Stata手册(Stata manuals)。在Stata 11中,每个命令的帮助页面(比如"help reg")底部均有相应的Stata手册链接。

习　题

中国的GDP(以购买力平价计)何时能超过美国?从Penn World Table(权威的跨国宏观数据集)下载中美两国1978—2010年有关人口与人均GDP的数据,导入Stata中,将两国log(GDP)的时间趋势画在一张图上,并进行简单外推预测(假设未来的增长率与1978—2010年间相同)。下载地址为(或搜索"Penn World Table"):http://pwt.econ.upenn.edu/php_site/pwt_index.php。下载时选csv格式,根据网站说明存储数据。

第 5 章　大样本 OLS

5.1　为何需要大样本理论

“大样本理论”(large sample theory),也称为“渐近理论”(asymptotic theory),研究的是当样本容量 n 趋于无穷大时统计量的性质。大样本理论近年来大受欢迎,主要原因如下。

(1) 小样本理论的假设过强。比如,小样本理论的严格外生性假设要求解释变量与所有的扰动项均正交。在时间序列模型中,这意味着解释变量与扰动项的过去、现在与未来值全部正交!在以被解释变量的滞后值作为解释变量的自回归模型中,必然违背这个假定。而在大样本理论中,只要求解释变量与同期的扰动项不相关即可。

例　$y_t = \beta y_{t-1} + \varepsilon_t$。

假设 $E(y_{t-1}\varepsilon_t) = 0$。然而,由于 ε_t 是 y_t 的一部分,故二者必然相关,即

$$E(y_t\varepsilon_t) = E[(\beta y_{t-1} + \varepsilon_t)\varepsilon_t] = \beta E(y_{t-1}\varepsilon_t) + E(\varepsilon_t^2) = E(\varepsilon_t^2) > 0$$

另外,小样本理论的统计推断必须假定扰动项为正态分布。大样本理论则无需此限制性假定。

(2) 在小样本下,我们必须研究统计量的精确分布(exact distribution),但常常难以推导。在大样本下,只要研究其渐近分布就可以了,而渐近分布较容易推导。

(3) 使用大样本理论的代价是要求样本容量较大,一般要求至少 $n \geqslant 30$,最好 n 在 100 以上。由于现代的数据集越来越大,经常成百上千,当 $n \to \infty$ 时才严格成立的渐近理论就是对实际数据的很好近似。

5.2　随 机 收 敛

1. 确定性序列的收敛

定义　确定性序列 $\{a_n\}_{n=1}^{\infty} = \{a_1, a_2, a_3, \cdots\}$ “**收敛**”(converges)于常数 a,记为 $\lim\limits_{n\to\infty} a_n = a$ 或 $a_n \to a$,如果 $\forall \varepsilon > 0$,存在 $N > 0$,只要 $n > N$,就有 $|a_n - a| < \varepsilon$,即 $\{a_{N+1}, a_{N+2}, \cdots\}$ 均落入区间 $(a-\varepsilon, a+\varepsilon)$ 内,参见图 5.1。

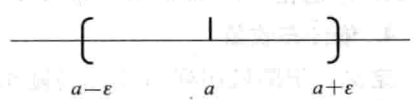


图 5.1　确定性序列的收敛

2. 随机序列的收敛

定义　随机序列 $\{x_n\}_{n=1}^{\infty} = \{x_1, x_2, x_3, \cdots\}$ “**依概率收敛**”(converges in probability)于常数 a,记为 $\operatorname{plim}\limits_{n\to\infty} x_n = a$,或 $x_n \xrightarrow{p} a$,如果 $\forall \varepsilon > 0$,当 $n \to \infty$ 时,都有 $\lim\limits_{n\to\infty} P(|x_n - a| > \varepsilon) = 0$。这意味着,任意给定 $\varepsilon > 0$,当

n 越来越大时，随机变量 x_n 落在区间 $(a-\varepsilon, a+\varepsilon)$ 之外的概率收敛于0，参见图5.2。

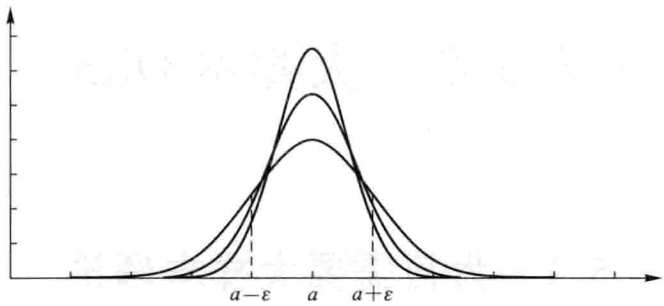


图5.2 随机序列的收敛

对于随机向量与随机矩阵，也可以定义依概率收敛，只要定义其每个元素都依概率收敛即可。

定义 随机序列 $\{x_n\}_{n=1}^{\infty}$ "**依概率收敛**"于随机变量 x，记为 $x_n \xrightarrow{p} x$，如果随机序列 $\{x_n - x\}_{n=1}^{\infty}$ 依概率收敛于0。

命题（连续函数与依概率收敛可交换运算次序，preservation of convergence for continuous transformation） 假设 $g(\cdot)$ 为连续函数，则 $\operatorname{plim}_{n\to\infty} g(x_n) = g(\operatorname{plim}_{n\to\infty} x_n)$。

证明：使用连续函数与依概率收敛的定义（略）。

从直观上可以理解为，当 x_n 的分布越来越集中于某 x 附近时，$g(x_n)$ 的分布自然也就越来越集中于 $g(x)$ 附近。这个命题意味着概率收敛算子 $\operatorname{plim}_{n\to\infty}$ 与连续函数 $g(\cdot)$ 可交换运算次序。在这一点上，概率极限 $\operatorname{plim}_{n\to\infty}$ 与普通极限 $\lim_{n\to\infty}$ 的性质相同。而期望算子E却没有这么好的性质。例如，一般来说，$\mathrm{E}(x^2) \neq [\mathrm{E}(x)]^2$。这也是小样本的精确分布比大样本的渐近分布更难推导的原因之一。

例 如果 $\operatorname{plim} s^2 = \sigma^2$，则 $\operatorname{plim} s = \operatorname{plim}(s^2)^{1/2} = (\operatorname{plim} s^2)^{1/2} = (\sigma^2)^{1/2} = \sigma$（因为开根号是连续函数。解读：如果样本方差是方差的一致估计，则样本标准差也是标准差的一致估计）。

3. 依均方收敛

定义 随机序列 $\{x_n\}_{n=1}^{\infty}$ "**依均方收敛**"（converges in mean square）于常数 a，如果 $\lim_{n\to\infty} \mathrm{E}(x_n) = a$，$\lim_{n\to\infty} \mathrm{Var}(x_n) = 0$。

命题 依均方收敛是依概率收敛的充分条件。

证明：使用切比雪夫不等式（参见附录）。

直观上比较显然，即当 x_n 的均值越来越趋于 a，而方差越来越小并趋于0时，就有 $\operatorname{plim}_{n\to\infty} x_n = a$，即在极限处 x_n 退化（degenerate）为常数 a。这个命题是依均方收敛概念的主要用途。

4. 依分布收敛

定义 记随机序列 $\{x_n\}_{n=1}^{\infty}$ 与随机变量 x 的累积分布函数（cdf）分别为 $F_n(\cdot)$ 与 $F(\cdot)$。如果对于任意实数 c，都有 $\lim_{n\to\infty} F_n(c) = F(c)$，则称随机序列 $\{x_n\}_{n=1}^{\infty}$ "**依分布收敛**"（converges in distribution）于随机变量 x，记为 $x_n \xrightarrow{d} x$。比如，当 t 分布的自由度越来越大时，其累积分布函数收敛于标准正态的累积分布函数，参见图5.3。

如果 x 为正态分布，而 $x_n \xrightarrow{d} x$，则称 $\{x_n\}_{n=1}^{\infty}$ 为“渐近正态”(asymptotically normal)。

直观上，依分布收敛意味着，两个随机变量的概率密度函数长得越来越像。容易看出，“依概率收敛”比“依分布收敛”更强(前者是后者的充分条件)，即“$x_n \xrightarrow{p} x$”⇔“$x_n - x \xrightarrow{p} 0$”⇒“$x_n \xrightarrow{d} x$”；但反之不然，因为当 x_n 与 x 的分布函数很接近时，x_n 与 x 的实际取值仍然可以很不相同(比如，x_n 与 x 可以是相互独立的随机变量)。

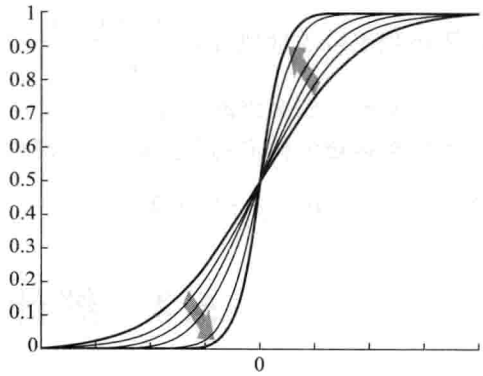


图 5.3 依分布收敛

命题 假设 $g(\cdot)$ 为连续函数，且 $x_n \xrightarrow{d} x$，则 $g(x_n) \xrightarrow{d} g(x)$。

这个命题再次显示了大样本理论的方便之处。从直观上可以理解为，当 x_n 的分布越来越像 x 的分布时，$g(x_n)$ 的分布自然也就越来越像 $g(x)$ 的分布。

例 假设 $x_n \xrightarrow{d} z$，其中 $z \sim N(0,1)$；则 $x_n^2 \xrightarrow{d} z^2$，其中 $z^2 \sim \chi(1)$，即 $x_n^2 \xrightarrow{d} \chi(1)$(因为平方是连续函数。解读：渐近标准正态的平方服从渐近 $\chi(1)$ 分布)。

5.3 大数定律与中心极限定理

1. 弱大数定律(Weak Law of Large Numbers)

假定 $\{x_n\}_{n=1}^{\infty}$ 为独立同分布的随机序列，且 $\mathrm{E}(x_1)=\mu$，$\mathrm{Var}(x_1)=\sigma^2$ 存在，则样本均值 $\bar{x}_n \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu$。

证明：因为 $\mathrm{E}(\bar{x}_n)=\mu$，而 $\mathrm{Var}(\bar{x}_n)=\mathrm{Var}\left(\frac{x_1+\cdots+x_n}{n}\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}\to 0$。故 $\bar{x}_n$ 依均方收敛于 μ。因此，$\bar{x}_n \xrightarrow{p} \mu$。该定理表明，样本无限大时，样本均值趋于总体均值，故名“大数定律”。

2. 中心极限定理(Central Limit Theorem，简记 CLT)

定理 假定 $\{x_n\}_{n=1}^{\infty}$ 为独立同分布的随机序列，且 $\mathrm{E}(x_1)=\mu$，$\mathrm{Var}(x_1)=\sigma^2$ 存在，则 $\sqrt{n}(\bar{x}_n-\mu) \xrightarrow{d} N(0,\sigma^2)$。

根据弱大数定律，$(\bar{x}_n-\mu) \xrightarrow{p} 0$(退化的随机变量)，而 $\sqrt{n}\to\infty$，故用 $\sqrt{n}(\bar{x}_n-\mu)$(即“$\infty\cdot 0$”型)得到一个非退化的分布。这表明，不仅 $(\bar{x}_n-\mu)$ 依概率收敛到 0，而且我们还知道其收敛到 0 的速度与 $\frac{1}{\sqrt{n}}$ 收敛到 0 的速度类似(故二者乘积为非退化分布)，这被称为“$\sqrt{n}$ 收敛”(root-n convergence)。从直观上看，可以视为 $\bar{x}_n \xrightarrow{d} N\left(\mu,\frac{\sigma^2}{n}\right)$，即样本均值近似地服从正态分布。但这是不严格的写法，因为 $\frac{\sigma^2}{n}\to 0$，不是正数(非退化随机变量的方差必须为正)。在一维情况下，中

心极限定理也可等价地写为$\frac{\bar{x}_n-\mu}{\sqrt{\sigma^2/n}}\xrightarrow{d}N(0,1)$（这是本科概率统计教材常用的形式）。但这一形式不易推广到多维的情形。

推广到多维的情形：假定$\{\boldsymbol{x}_n\}_{n=1}^{\infty}$为独立同分布的随机向量序列，且$\mathrm{E}(\boldsymbol{x}_1)=\boldsymbol{\mu}$，$\mathrm{Var}(\boldsymbol{x}_1)=\boldsymbol{\Sigma}$存在，则$\sqrt{n}(\bar{\boldsymbol{x}}_n-\boldsymbol{\mu})\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma})$。

5.4 统计量的大样本性质

1. 均方误差

假设$\hat{\beta}$是一维参数β的估计量。我们希望抽样误差（sampling error）$(\hat{\beta}-\beta)$尽量地小，即$\hat{\beta}$离真实参数β越近越好。因此，可以考虑以误差平方（squared error）$(\hat{\beta}-\beta)^2$作为度量。但$\hat{\beta}$是随机变量，故引入“均方误差”的概念。

定义 以估计量$\hat{\beta}$来估计参数β，则其“**均方误差**”（Mean Squared Error，简记 MSE）为$\mathrm{MSE}(\hat{\beta})\equiv\mathrm{E}[(\hat{\beta}-\beta)^2]$。

在最理想的情况下，一个最优的估计量应该在所有的估计量中具有最小的均方误差。

另外，我们不希望$\hat{\beta}$系统地高估或低估β，即没有系统误差（systematic error）。

定义 以估计量$\hat{\beta}$来估计参数β，则其“**偏差**”为$\mathrm{Bias}(\hat{\beta})\equiv\mathrm{E}(\hat{\beta})-\beta$。

定义 如果偏差$\mathrm{Bias}(\hat{\beta})=0$，则称$\hat{\beta}$为“**无偏估计量**”（unbiased estimator）。

命题 均方误差可以分解为方差与偏差平方之和，即

$$\mathrm{MSE}(\hat{\beta})=\mathrm{Var}(\hat{\beta})+[\mathrm{Bias}(\hat{\beta})]^2 \tag{5.1}$$

证明：
$$\begin{aligned}\mathrm{MSE}(\hat{\beta})&\equiv\mathrm{E}[(\hat{\beta}-\beta)^2]=\mathrm{E}\{[\hat{\beta}-\mathrm{E}(\hat{\beta})+\mathrm{E}(\hat{\beta})-\beta]^2\}\quad(\text{加、减 }E(\hat{\beta}))\\&=\mathrm{E}[\hat{\beta}-\mathrm{E}(\hat{\beta})]^2+2\mathrm{E}\{[\hat{\beta}-\mathrm{E}(\hat{\beta})][\mathrm{E}(\hat{\beta})-\beta]\}+\mathrm{E}[\mathrm{E}(\hat{\beta})-\beta]^2\\&=\mathrm{Var}(\hat{\beta})+2\mathrm{E}\{[\hat{\beta}-\mathrm{E}(\hat{\beta})][\mathrm{E}(\hat{\beta})-\beta]\}+[\mathrm{Bias}(\hat{\beta})]^2\end{aligned}$$

而上式中的交叉项为

$$\mathrm{E}\{[\hat{\beta}-\mathrm{E}(\hat{\beta})][\mathrm{E}(\hat{\beta})-\beta]\}=[\mathrm{E}(\hat{\beta})-\beta]\mathrm{E}[\hat{\beta}-\mathrm{E}(\hat{\beta})]=[\mathrm{E}(\hat{\beta})-\beta]\cdot 0=0$$

在多维情况下，也有类似的结论，

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}})\equiv\mathrm{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})']=\mathrm{Var}(\hat{\boldsymbol{\beta}})+[\mathrm{Bias}(\hat{\boldsymbol{\beta}})][\mathrm{Bias}(\hat{\boldsymbol{\beta}})]' \tag{5.2}$$

因此，使均方误差最小化，可以视为是在“估计量方差”与“偏差”之间进行权衡（trade-off）。比如，一个无偏估计量（偏差为0），如果方差很大，则可能不如一个虽然有偏差但方差却很小的估计量。

2. 一致估计量

定义 如果$\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n=\boldsymbol{\beta}$，则估计量$\hat{\boldsymbol{\beta}}_n$是参数$\boldsymbol{\beta}$的“**一致估计量**”（consistent estimator）。

一致性（consistency）（也译为“相合性”）意味着，当样本容量足够大时，$\hat{\boldsymbol{\beta}}_n$依概率收敛到真实参数$\boldsymbol{\beta}$。这是对估计量最基本，也是最重要的要求。在大样本理论中，无偏性不再重要。如果估计方法不一致，则意味着你的研究没有太大意义；因为无论样本容量多大，估计量也不会收敛到真实的参数值。

3. 渐近正态分布与渐近方差

定义 如果$\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma})$，其中$\boldsymbol{\Sigma}$为半正定矩阵，则称$\hat{\boldsymbol{\beta}}_n$为“**渐近正态分布**”

(asymptotically normally distributed),而称 $\boldsymbol{\Sigma}$ 为"**渐近方差**"(asymptotic variance),记为 $\text{Avar}(\hat{\boldsymbol{\beta}}_n)$。直观上,可以近似地认为 $\hat{\boldsymbol{\beta}}_n \xrightarrow{d} N(\boldsymbol{\beta}, \boldsymbol{\Sigma}/n)$。由于 $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ 收敛到一个非退化的分布,故 $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ 收敛到 0 的速度与 $\frac{1}{\sqrt{n}}$ 收敛到 0 的速度大致相同,因此也称为"$\sqrt{n}$收敛"。

4. 渐近有效

假设 $\hat{\boldsymbol{\beta}}_n$ 与 $\tilde{\boldsymbol{\beta}}_n$ 都是 $\boldsymbol{\beta}$ 的渐近正态估计量,其渐近方差分别为 $\boldsymbol{\Sigma}$ 与 $\boldsymbol{V}$。如果 $(\boldsymbol{V} - \boldsymbol{\Sigma})$ 为半正定矩阵,则称 $\hat{\boldsymbol{\beta}}_n$ 比 $\tilde{\boldsymbol{\beta}}_n$"**更为渐近有效**"(asymptotically more efficient)。

5.5 渐近分布的推导

下面介绍推导渐近分布的常用技巧,主要涉及依概率收敛与依分布收敛的交叉运算,统称"斯拉斯基定理"(Slutsky Theorem)。下面的 x_n, x, y_n 可以是随机变量或随机向量。

(1) $x_n \xrightarrow{d} x, y_n \xrightarrow{p} a \Rightarrow x_n + y_n \xrightarrow{d} x + a$。

在极限处,y_n 退化为常数 a,不再是正常的随机变量。故 $x_n + y_n$ 在极限处只是将 x_n 的渐近分布 x 位移到 $x + a$。不必考虑 x_n 与 y_n 是否相互独立。

特例:如果 $a = 0$,则 $x_n + y_n \xrightarrow{d} x$,即 y_n 的作用可以忽略。

(2) $x_n \xrightarrow{d} x, y_n \xrightarrow{p} 0 \Rightarrow x_n y_n \xrightarrow{p} 0$。

在极限处,y_n 退化为 0,而 x_n 有一个正常的渐近分布 x,故 $x_n y_n$ 退化为 0。

(3) 随机向量 $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}$,随机矩阵 $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$, $\boldsymbol{A}_n \boldsymbol{x}_n$ 可以相乘 $\Rightarrow \boldsymbol{A}_n \boldsymbol{x}_n \xrightarrow{d} \boldsymbol{A}\boldsymbol{x}$。

特别地,如果 $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$,则 $\boldsymbol{A}_n \boldsymbol{x}_n \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$。

注:在极限处,随机矩阵 $\boldsymbol{A}_n$ 退化为常数矩阵 $\boldsymbol{A}$。正态分布的线性组合仍然服从正态分布,而且 $\text{Var}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\text{Var}(\boldsymbol{x})\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$。

(4) 随机向量 $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}$,随机矩阵 $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$, $\boldsymbol{A}_n \boldsymbol{x}_n$ 可以相乘, $\boldsymbol{A}^{-1}$ 存在

$\Rightarrow$ 二次型 $\boldsymbol{x}_n' \boldsymbol{A}_n^{-1} \boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}' \boldsymbol{A}^{-1} \boldsymbol{x}$。

注:在极限处,随机矩阵 $\boldsymbol{A}_n$ 退化为常数矩阵 $\boldsymbol{A}$。

总之,在渐近理论中,如果一个随机变量(向量)在极限处退化为一个常数(向量),则常常可以将其视为一个常数(向量)来处理(比如,加法或乘法)。

5.6 随机过程的性质

随机序列 $\{x_n\}_{n=1}^{\infty}$ 有个更好听的名字,叫做"随机过程"(stochastic process)。如果下标为时间,则记为 $\{x_t\}_{t=1}^{\infty}$,也称为"时间序列"(time series)。

1. 严格平稳过程

考察中国 1978—2007 年的通货膨胀率这一随机过程,即 $\{\pi_{1978}, \pi_{1979}, \cdots, \pi_{2007}\}$,参见图 5.4。

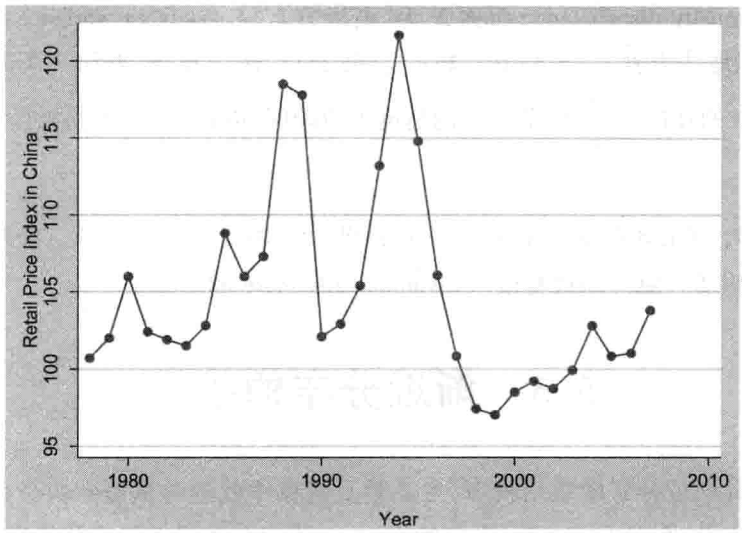


图 5.4 中国零售物价环比指数,1978—2007

数据来源:国家统计局网站

假如每年的通货膨胀率作为一个随机变量都有自己不同的分布,那么我们如何估计$\mathrm{E}(\pi_{1978})$与$\mathrm{Var}(\pi_{1978})$呢?每年的通货膨胀率的样本容量仅为1,而且历史不能重演!但如果这30年的通货膨胀率的概率分布都不变,则可以将$\bar{\pi} \equiv \frac{1}{30}\sum_{t=1978}^{2007}\pi_t$作为$\mathrm{E}(\pi_t)$的估计量。

从理论上讲,可以把随机过程$\{x_1, x_2, \cdots, x_t, \cdots\}$看成是一个无穷维的随机向量。但无穷维随机向量不易把握,故考虑其有限维随机向量的分布,而下面介绍的"严格平稳过程"即要求其有限维分布不随时间的推移而改变。比如,x_i的分布与x_j的分布相同$(\forall i, j)$;(x_1, x_4)的分布与(x_2, x_5)的分布相同;(x_1, x_2, x_3)的分布与(x_5, x_6, x_7)的分布相同。

定义 随机过程$\{x_t\}_{t=1}^{\infty}$是"**严格平稳过程**"(strictly stationary process),简称平稳过程,如果对任意m个时期的时间集合$\{t_1, t_2, \cdots, t_m\}$,随机向量$\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\}$的联合分布等于随机向量$\{x_{t_1+k}, x_{t_2+k}, \cdots, x_{t_m+k}\}$的联合分布,其中$k$为任意整数。也就是说,将$\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\}$中每个变量的时间下标全部前移或后移$k$期,不会改变其分布。换言之,$\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\}$的联合分布仅依赖于$\{t_1, t_2, \cdots, t_m\}$各个时期之间的相对距离,而不依赖于其绝对位置。

例 如果随机过程$\{x_t\}_{t=1}^{\infty}$为iid,则$\{x_t\}_{t=1}^{\infty}$是平稳过程,且不存在序列相关。

例 如果随机过程$\{x_t\}_{t=1}^{\infty} = \{x_1, x_1, x_1, \cdots\}$(即$x_t \equiv x_1$),则$\{x_t\}_{t=1}^{\infty}$是平稳过程,且存在最强的序列相关。

例 考虑以下一阶自回归过程(first order autoregression,简记AR(1)),

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \mathrm{Cov}(y_{t-1}, \varepsilon_t) = 0 \tag{5.3}$$

其中,$\{\varepsilon_t\}$为独立同分布。

命题 如果$\rho = 1$,则$\{y_t\}$不是平稳过程。如果$|\rho| < 1$,则$\{y_t\}$是严格平稳过程。

证明:如果$\rho = 1$,则$y_t = y_{t-1} + \varepsilon_t$。因此,$y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t$。故当$t \to \infty$时,$\mathrm{Var}(y_t) = t\sigma_\varepsilon^2 \to \infty$,其中$\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_t)$,即方差越来越大,以至无穷。因此,$\{y_t\}$不是平稳过程。此时,$\{y_t\}$称为"随机游走"(random walk),存在"单位根"(unit root),参见第21章。

如果$|\rho| < 1$,对该方程两边同时取方差,可得$\mathrm{Var}(y_t) = \rho^2 \mathrm{Var}(y_{t-1}) + \sigma_\varepsilon^2$。这是一阶线性差

分方程。容易看出，由于 $\rho^2<1$，故 $\mathrm{Var}(y_t)$ 将收敛于 $\frac{\sigma_\varepsilon^2}{1-\rho^2}$①，参见图 5.5。进一步可以证明，$\{y_t\}_{t=0}^{\infty}$ 是严格平稳过程（参见 Stock and Watson，2011，p. 578）。

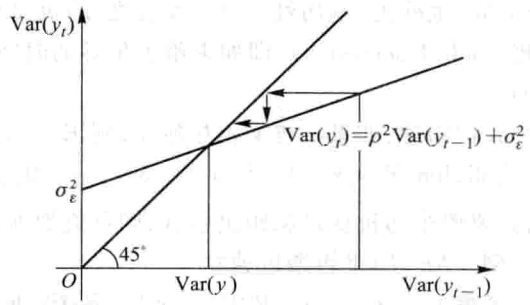


图 5.5 平稳一阶自回归过程的方差收敛

有时我们仅仅关心随机过程的期望、方差及协方差是否稳定，故引入以下“弱平稳过程”的概念。

定义 随机过程 $\{x_t\}_{t=1}^{\infty}$ 是“**弱平稳过程**”（weakly stationary process）或“**协方差平稳过程**”（covariance stationary process），如果 $\mathrm{E}(x_t)$ 不依赖于 t，而且 $\mathrm{Cov}(x_t,x_{t+k})$ 仅依赖于 k（即 x_t 与 x_{t+k} 在时间上的相对距离）而不依赖于其绝对位置 t。

显然，弱平稳过程的期望与方差均为常数（只要在 $\mathrm{Cov}(x_t,x_{t+k})$ 中令 $k=0$，即可知方差为常数）。

定义 一个协方差平稳过程 $\{x_t\}_{t=1}^{\infty}$ 被称为“**白噪声过程**”（white noise process），如果对于 $\forall t$，都有 $\mathrm{E}(x_t)=0$，而且 $\mathrm{Cov}(x_t,x_{t+k})=0$，$\forall k\neq 0$。

注：白噪声过程不一定独立同分布，也不一定是严格平稳过程。“白噪声”是性质比较好的噪声②，即该噪声的期望值为 0，而不同期之间的噪声互不相关。

显然，严格平稳过程③是弱平稳过程的充分条件；但反之则不然，因为弱平稳过程只要求二阶矩平稳（即期望、方差、协方差等不随时间而变），而概率分布可能还依赖于更高阶的矩。

对于随机向量过程 $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$，可以类似地给出其为平稳过程或弱平稳过程的定义（只要将上述定义中的 x 置换为 $\boldsymbol{x}$ 即可）。显然，如果 $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ 为（弱）平稳过程，则其每个分量都是（弱）平稳过程；反之，则不然。

例 假设 $\{\varepsilon_t\}_{t=1}^{\infty}$ 为 iid 过程。定义 $\boldsymbol{x}_t\equiv\begin{pmatrix}\varepsilon_t\\ \varepsilon_1\end{pmatrix}$。显然，$\{\boldsymbol{x}_t\}_{t=1}^{\infty}=\left\{\begin{pmatrix}\varepsilon_1\\ \varepsilon_1\end{pmatrix},\begin{pmatrix}\varepsilon_2\\ \varepsilon_1\end{pmatrix},\begin{pmatrix}\varepsilon_3\\ \varepsilon_1\end{pmatrix},\cdots\right\}$ 的第一个分量 $\{\varepsilon_t\}_{t=1}^{\infty}$ 与第二个分量 $\{\varepsilon_1,\varepsilon_1,\varepsilon_1,\cdots\}$ 都是平稳过程（参见上文的例子）。然而，$\boldsymbol{x}_1\equiv\begin{pmatrix}\varepsilon_1\\ \varepsilon_1\end{pmatrix}$ 的联合分布却与 $\boldsymbol{x}_2\equiv\begin{pmatrix}\varepsilon_2\\ \varepsilon_1\end{pmatrix}$ 的联合分布不同，故 $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ 不是平稳过程。进一步，由于 $\mathrm{Var}(\boldsymbol{x}_1)=\begin{pmatrix}\sigma_\varepsilon^2 & \sigma_\varepsilon^2\\ \sigma_\varepsilon^2 & \sigma_\varepsilon^2\end{pmatrix}\neq\begin{pmatrix}\sigma_\varepsilon^2 & 0\\ 0 & \sigma_\varepsilon^2\end{pmatrix}=\mathrm{Var}(\boldsymbol{x}_2)$，故 $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ 也不是弱平稳过程。

2. 渐近独立性

然而，仅为“严格平稳过程”（相当于“同分布”的假定）还不足以应用大数定律或中心极限定理，因为它们都要求独立同分布，即序列中各变量还要相互独立。显然，“相互独立”的假定对于大多数经济变量而言过强了。比如，今年的通胀率显然与去年的通胀率相关，不会相互独立。然而，今年的通胀率与 100 年前的通胀率或许可以近似地视为相互独立，称为“**渐近独立**”

① 只要在方程 $\mathrm{Var}(y_t)=\rho^2\mathrm{Var}(y_{t-1})+\sigma_\varepsilon^2$ 中，令 $\mathrm{Var}(y_t)=\mathrm{Var}(y_{t-1})$ 即可求出方差的均衡值 $\mathrm{Var}(y)$。

② 噪声本来是一种听觉，但偏偏又可以有视觉效果（白色），这是统计学中的一个奇妙术语。

③ 假设严格平稳过程的期望、方差、协方差存在。

(ergodic,也称为"遍历性"①)。换言之,随机过程没有长记忆(long memory),或没有长期的路径依赖(path dependence),即如果给予足够的时间,则系统的演化将忘记自己是从什么初始条件起步的。

我们知道,如果 x 与 y 相互独立,则 $E(xy)=E(x)E(y)$。利用这个性质,可以定义渐近独立,比如,$\lim_{n\to\infty}[E(x_t x_{t+n})-E(x_t)E(x_{t+n})]=0$,其严格定义参见附录。直观来说,渐近独立意味着,只要两个随机变量相距足够远,则可近似地认为它们相互独立。

例 AR(1)是否渐近独立?

考虑 $y_t=\rho y_{t-1}+\varepsilon_t$,其中 $|\rho|<1$。是否时间间隔越大,则两个随机变量之间的协方差越来越小,并收敛于0呢?当时间间隔为1时,

$$\mathrm{Cov}(y_t,y_{t-1})=\mathrm{Cov}(\rho y_{t-1}+\varepsilon_t,y_{t-1})=\rho\sigma_y^2 \tag{5.4}$$

当时间间隔为2时,$y_t=\rho y_{t-1}+\varepsilon_t=\rho(\rho y_{t-2}+\varepsilon_{t-1})+\varepsilon_t=\rho^2 y_{t-2}+\rho\varepsilon_{t-1}+\varepsilon_t$,

故

$$\mathrm{Cov}(y_t,y_{t-2})=\mathrm{Cov}(\rho^2 y_{t-2}+\rho\varepsilon_{t-1}+\varepsilon_t,y_{t-2})=\rho^2\sigma_y^2 \tag{5.5}$$

以此类推,当时间间隔为 j 时,$\mathrm{Cov}(y_t,y_{t-j})=\rho^j\sigma_y^2$。由于 $|\rho|<1$,故当 $j\to\infty$ 时,$\mathrm{Cov}(y_t,y_{t-j})\to 0$。因此,AR(1)为渐近独立。

渐近独立定理(Ergodic Theorem) 假设 $\{x_i\}_{i=1}^{\infty}$ 为渐近独立的严格平稳过程,且 $E(x_i)=\mu$,则 $\bar{x}_n\equiv\frac{1}{n}\sum_{i=1}^{n}x_i\xrightarrow{p}\mu$,即样本均值 $\bar{x}_n$ 是总体均值 $E(x_i)$ 的一致估计。

这是对大数定律的一个重要推广。大数定律要求每个 x_i 相互独立,而渐近独立定理允许 $\{x_i\}_{i=1}^{\infty}$ 存在"序列相关"(serial correlation)②,只要这种相关关系在极限处消失即可。大数定律要求每个 x_i 的分布相同,而渐近独立定理要求 $\{x_i\}_{i=1}^{\infty}$ 为严格平稳过程(故也是同分布的)。在相当程度上,渐近独立定理对经济数据的意义要比大数定律更大。

命题 如果 $\{x_i\}_{i=1}^{\infty}$ 为渐近独立的严格平稳过程,则对于任何连续函数 $f(\cdot)$,$\{f(x_i)\}_{i=1}^{\infty}$ 也是渐近独立的严格平稳过程。

渐近独立定理意味着,渐近独立平稳过程 $\{x_i\}_{i=1}^{\infty}$ 的任何"总体矩"(population moment) $E[f(x_i)]$,都可以由其对应的"样本矩"(sample moment) $\frac{1}{n}\sum_{i=1}^{n}f(x_i)$ 一致地估计。比如,$E(\boldsymbol{x}_i\boldsymbol{x}_i')_{K\times K}$ 可以由 $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$ 一致地估计,其中 $(\boldsymbol{x}_i\boldsymbol{x}_i')_{K\times K}$ 为随机矩阵。

除了平稳性与渐近独立性,为了使用中心极限定理,还需要另一条件,即鞅差分序列。

定义 称随机过程 $\{x_i\}_{i=1}^{\infty}$ 为"**鞅**"(martingale)③,如果它满足 $E(x_i|x_{i-1},\cdots,x_1)=x_{i-1}$,$\forall i\geqslant 2$。

例 随机游走过程 $x_t=x_{t-1}+\varepsilon_t$。显然,$E(x_t|x_{t-1},\cdots,x_1)=x_{t-1}$。

例 资本市场有效理论认为,所有有关未来价格的已知信息均已反映在当期价格上,故 $E(p_{t+1}|p_t,\cdots,p_1)=p_t$。因此,试图预测未来价格的走势是徒劳的。然而,如果市场参与者的信息不对称,这个结论不一定成立。

---

① "遍历性"这一术语最早来自于物理学。但对计量经济学而言,翻译为"渐近独立性"更贴切些。

② "序列相关"(serial correlation)与"自相关"(autocorrelation)是同义词。

③ "Martingale"的原意为"马颔缰"。也许因为与赛马有关,"martingale"在英文中也指一种赌博的方法,即不断地把赌注翻倍。后来,"martingale"被数学家用来指一种随机过程。

定义 称随机过程$\{x_i\}_{i=1}^{\infty}$为"**鞅差分序列**"(Martingale Difference Sequence,简记 MDS),如果它满足$\mathrm{E}(x_i|x_{i-1},\cdots,x_1)=0$,$\forall i\geqslant 2$。

显然,这意味着x_i均值独立于它的所有过去值。因此,$\mathrm{Cov}(x_i,x_{i-j})=0$,$\forall j\neq 0$。而且,根据迭代期望定律可知,鞅差分序列的无条件期望$\mathrm{E}(x_i)=\mathrm{E}_{x_{i-1},\cdots,x_1}\mathrm{E}(x_i|x_{i-1},\cdots,x_1)=0$。

命题 对鞅序列进行一阶差分,就得到鞅差分序列。

证明:假设$\{x_i\}_{i=1}^{\infty}=\{x_1,x_2,x_3,\cdots\}$为鞅过程。定义其差分为$g_1\equiv x_1$,$g_i\equiv x_i-x_{i-1}$,$\forall i\geqslant 2$。对$\forall i\geqslant 2$,条件期望

$$
\begin{aligned}
&\mathrm{E}(g_i|g_{i-1},\cdots,g_1)\\
&=\mathrm{E}(g_i|x_{i-1},\cdots,x_1) && (\{g_{i-1},\cdots,g_1\}\text{与}\{x_{i-1},\cdots,x_1\}\text{包含同样的信息})\\
&=\mathrm{E}(x_i-x_{i-1}|x_{i-1},\cdots,x_1) && (\text{定义 } g_i\equiv x_i-x_{i-1})\\
&=\mathrm{E}(x_i|x_{i-1},\cdots,x_1)-x_{i-1} && (\text{期望算子的线性性})\\
&=x_{i-1}-x_{i-1}=0 && (\text{鞅过程的定义})
\end{aligned}
$$

因此,$\{g_i\}_{i=1}^{\infty}$是鞅差分序列。

鞅差分序列的中心极限定理(Central Limit Theorem for Ergodic Stationary MDS) 假设$\{\boldsymbol{g}_i\}_{i=1}^{\infty}$为渐近独立的平稳鞅差分随机向量过程,且其协方差矩阵为$\mathrm{Cov}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\boldsymbol{\Sigma}$[①],记$\bar{\boldsymbol{g}}\equiv\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i$,则$\sqrt{n}\bar{\boldsymbol{g}}\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{\Sigma})$。

普通的中心极限定理仅适用于独立同分布的情形,而这个定理适用于更一般的渐近独立的平稳鞅差分序列。

5.7 大样本 OLS 的假定

假定 5.1 线性假定

$$y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\quad(i=1,\cdots,n) \tag{5.6}$$

假定 5.2 渐近独立的平稳过程(ergodic stationarity)

$(K+1)$维随机过程$\{y_i,\boldsymbol{x}_i\}$为渐近独立的平稳过程。作为一个最简单的例子,如果样本为随机样本,则$\{y_i,\boldsymbol{x}_i\}$独立同分布,故是渐近独立的平稳过程。

假定 5.3 前定解释变量(predetermined regressors)

所有解释变量均为"前定"(predetermined),即它们与同期的扰动项[②]正交,即$\mathrm{E}(x_{ik}\varepsilon_i)=0$,$\forall i,k$。由于回归方程通常有常数项,故总可以假设$\mathrm{E}(\varepsilon_i)=0$,故$\mathrm{Cov}(x_{ik},\varepsilon_i)=\mathrm{E}(x_{ik}\varepsilon_i)-\mathrm{E}(x_{ik})\mathrm{E}(\varepsilon_i)=0-0=0$。这意味着$\boldsymbol{x}_i$与$\varepsilon_i$不相关,仿佛在$\varepsilon_i$产生之前,$\boldsymbol{x}_i$便已经确定,故名"前定解释变量"。定义如下列向量:

$$\boldsymbol{g}_i\equiv\boldsymbol{x}_i\varepsilon_i=\begin{pmatrix}x_{i1}\\ \vdots\\ x_{iK}\end{pmatrix}\varepsilon_i \tag{5.7}$$

---

① 因为$\mathrm{E}(\boldsymbol{g}_i)=\boldsymbol{0}$,故$\mathrm{Cov}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{g}_i\boldsymbol{g}_i')$。

② 也就是在同一方程中的扰动项。

显然，$\mathrm{E}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{x}_i\varepsilon_i)=\mathbf{0}$。这个假定比第3章“小样本OLS”中的严格外生性假定更弱，因为后者要求扰动项与过去、现在及未来的解释变量都不相关（对于时间序列数据而言）。

假定5.4 秩条件（rank condition）

$K\times K$矩阵$\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$为非退化矩阵，即其逆矩阵$[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$存在。这个条件保证了在大样本下，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$存在。

假定5.5 $\boldsymbol{g}_i$为鞅差分序列，且其协方差矩阵$\boldsymbol{S}\equiv\mathrm{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\mathrm{E}(\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i')$为非退化矩阵。

由于鞅差分序列的无条件期望为0，故假定5.5比假定5.3更强。

在以上假定中，我们不需要假设“严格外生性”与“正态随机扰动项”，这使得模型具有更大的适用性与稳健性。

5.8 OLS的大样本性质

为了便于使用渐近理论，下面把OLS估计量$\boldsymbol{b}$写成$\{y_i,\boldsymbol{x}_i\}$的函数。

由于$\boldsymbol{X}\equiv\begin{pmatrix}\boldsymbol{x}_1'\\ \boldsymbol{x}_2'\\ \vdots\\ \boldsymbol{x}_n'\end{pmatrix}$，故$\boldsymbol{X}'\boldsymbol{X}=(\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\boldsymbol{x}_n)\begin{pmatrix}\boldsymbol{x}_1'\\ \boldsymbol{x}_2'\\ \vdots\\ \boldsymbol{x}_n'\end{pmatrix}=\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$。

其中，每个$\boldsymbol{x}_i\boldsymbol{x}_i'$都是一个$K\times K$的矩阵。

定义 $\boldsymbol{S}_{XX}\equiv\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$（随机矩阵$\boldsymbol{x}_i\boldsymbol{x}_i'$的样本均值）。

另一方面，$\boldsymbol{X}'\boldsymbol{y}=(\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\boldsymbol{x}_n)\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\sum_{i=1}^{n}\boldsymbol{x}_iy_i$。

其中，每个$\boldsymbol{x}_iy_i$都是一个$K\times 1$的向量。

定义 $\boldsymbol{S}_{Xy}\equiv\frac{1}{n}\boldsymbol{X}'\boldsymbol{y}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_iy_i$（随机向量$\boldsymbol{x}_iy_i$的样本均值）。因此，

$$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}'\boldsymbol{y}}{n}=\left(\frac{\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'}{n}\right)^{-1}\left(\frac{\sum_{i=1}^{n}\boldsymbol{x}_iy_i}{n}\right)=\boldsymbol{S}_{XX}^{-1}\boldsymbol{S}_{Xy}\tag{5.8}$$

对于大样本理论而言，关注的重点是当$n\to\infty$时，$\boldsymbol{S}_{XX}^{-1}$与$\boldsymbol{S}_{Xy}$的概率收敛性质。

定理（OLS估计量的大样本性质）

（1）（$\boldsymbol{b}$为一致估计量）在假定5.1～5.4之下，$\operatorname{plim}_{n\to\infty}\boldsymbol{b}=\boldsymbol{\beta}$。

（2）（$\boldsymbol{b}$为渐近正态）如果把假定5.3（即$\mathrm{E}(\boldsymbol{g}_i)=\mathbf{0}$）强化为假定5.5（即$\{\boldsymbol{g}_i\}$为MDS），则$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\mathrm{Avar}(\boldsymbol{b}))$。其中，

$$\mathrm{Avar}(\boldsymbol{b})=[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\boldsymbol{S}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1},\text{而}\ \boldsymbol{S}\equiv\mathrm{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\mathrm{E}(\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i')$$

（3）（$\mathrm{Avar}(\boldsymbol{b})$的一致估计量）假设$\hat{\boldsymbol{S}}$为$\boldsymbol{S}$的一致估计量，则$\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\boldsymbol{S}_{XX}^{-1}$是$\mathrm{Avar}(\boldsymbol{b})$的一致估计量。

证明:(1) 抽样误差可以写为

$$\boldsymbol{b}-\boldsymbol{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}=\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n}=\left(\frac{\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'}{n}\right)^{-1}\left(\frac{\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i}{n}\right)=\boldsymbol{S}_{XX}^{-1}\bar{\boldsymbol{g}} \tag{5.9}$$

其中,$\bar{\boldsymbol{g}}\equiv\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i$,$\boldsymbol{g}_i\equiv\boldsymbol{x}_i\varepsilon_i$。假定 5.2 意味着$\{\boldsymbol{x}_i\boldsymbol{x}_i'\}$也是渐近独立的平稳序列,故根据渐近独立定理,$\boldsymbol{S}_{XX}\equiv\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\xrightarrow{p}\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$。假定 5.4 意味着$[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$存在,故$\boldsymbol{S}_{XX}^{-1}\xrightarrow{p}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$。由于$\{\boldsymbol{g}_i\equiv\boldsymbol{x}_i\varepsilon_i=\boldsymbol{x}_i(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\}$,故$\{\boldsymbol{g}_i\}$也是渐近独立的平稳序列。假定 5.3 意味着,$\bar{\boldsymbol{g}}\equiv\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\xrightarrow{p}\mathrm{E}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{x}_i\varepsilon_i)=\boldsymbol{0}$。因此,$\boldsymbol{S}_{XX}^{-1}\bar{\boldsymbol{g}}\xrightarrow{p}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\cdot\boldsymbol{0}=\boldsymbol{0}$,故$\operatorname*{plim}_{n\to\infty}(\boldsymbol{b}-\boldsymbol{\beta})=\boldsymbol{0}$。

显然,扰动项与同期解释变量的不相关(假定 5.3)是保证 OLS 为一致估计量的最重要条件。为什么扰动项与解释变量相关就会导致不一致的估计[①]呢?下面我们以一元回归的示意图(参见图 5.6)来做直观的解释。

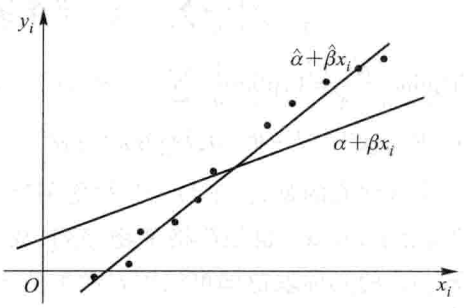


图 5.6 扰动项与解释变量相关导致不一致估计

假设 $y_i=\alpha+\beta x_i+\varepsilon_i$,而且 $\mathrm{Cov}(x_i,\varepsilon_i)>0$。真实的回归线($\alpha+\beta x_i$)与样本回归线($\hat{\alpha}+\hat{\beta}x_i$)参见图 5.6。由于 x_i 与 ε_i 正相关,故当 x_i 较小时,ε_i 也倾向于较小;而当 x_i 较大时,ε_i 也倾向于较大。因此,样本回归线很可能比真实回归线更陡峭,即最小二乘法 $\hat{\beta}$ 将倾向于高估 β。反之,如果 $\mathrm{Cov}(x_i,\varepsilon_i)<0$,则 $\hat{\beta}$ 将倾向于低估 β。增大样本容量($n\to\infty$)能使这种偏差(bias)消失吗?不能!即便使用人口普查的海量数据,偏差也依然存在!

什么情况下可能出现 $\mathrm{Cov}(x_i,\varepsilon_i)\neq0$ 呢?在存在遗漏变量、内生解释变量或解释变量测量误差的情况下常会出现(参见第 10 章)。

(2) 由于抽样误差 $\boldsymbol{b}-\boldsymbol{\beta}=\boldsymbol{S}_{XX}^{-1}\bar{\boldsymbol{g}}$,故$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})=\boldsymbol{S}_{XX}^{-1}(\sqrt{n}\bar{\boldsymbol{g}})$。根据假定 5.5 及鞅差分序列的中心极限定理,$\sqrt{n}\bar{\boldsymbol{g}}\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{S})$,其中 $\boldsymbol{S}\equiv\mathrm{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\mathrm{E}(\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i')$。由于$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})=\boldsymbol{S}_{XX}^{-1}(\sqrt{n}\bar{\boldsymbol{g}})$是$\sqrt{n}\bar{\boldsymbol{g}}$的线性组合,故$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})\xrightarrow{d}N(\boldsymbol{0},\mathrm{Avar}(\boldsymbol{b}))$。由于$\boldsymbol{S}_{XX}\equiv\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\xrightarrow{p}\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$,故 $\mathrm{Avar}(\boldsymbol{b})=[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\boldsymbol{S}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$(此处使用了公式 $\mathrm{Var}(\mathbf{AY})=\mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^{\mathrm{t}}$),其中 $[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$为对称矩阵。注意:此处不需要假设扰动项服从正态分布。

(3) 如果存在 $\hat{\boldsymbol{S}}\xrightarrow{p}\boldsymbol{S}$,已知$\boldsymbol{S}_{XX}^{-1}\xrightarrow{p}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$,故估计量$\widehat{\mathrm{Avar}}(\boldsymbol{b})\equiv\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\boldsymbol{S}_{XX}^{-1}\xrightarrow{p}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\boldsymbol{S}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}$,是 $\mathrm{Avar}(\boldsymbol{b})$的一致估计量。由于$\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\boldsymbol{S}_{XX}^{-1}$的形式为两个$\boldsymbol{S}_{XX}^{-1}$("两片面包")夹着一个 $\hat{\boldsymbol{S}}$("中间的菜"),故被称为"夹心估计量"或"三明治估计量"(sandwich estimator)。

---

① 有时称之为"随机解释变量问题"。但此称呼并不准确,因为事实上,所有解释变量都是随机的。关键的问题是,随机的解释变量可能与同期的扰动项相关,导致不一致估计。

为了得到 $\boldsymbol{S}\equiv \mathrm{E}(\varepsilon_i^2 \boldsymbol{x}_i\boldsymbol{x}_i')$ 的一致估计量，需要对解释变量的四阶矩进行假设。

假定 5.6（解释变量的四阶矩存在） $\mathrm{E}[(x_{ik}x_{ij})^2]$ 存在且为有限（$\forall i,j,k$）。

这只是一个技术性的假定，不必太在意。在假定 5.6 下，可以证明①，$\hat{\boldsymbol{S}}\equiv\frac{1}{n}\sum_{i=1}^{n} e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'$ 是 $\boldsymbol{S}$ 的一致估计量，其中 $\{e_i\}_{i=1}^n$ 为最小二乘法的残差。进一步，可以证明，s^2 是 σ^2 的一致估计。

命题 s^2 是无条件方差 $\mathrm{E}(\varepsilon_i^2)=\sigma^2$ 的一致估计量。

证明：

$$s^2 \equiv \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}=\frac{\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}}{n-K}=\frac{\boldsymbol{\varepsilon}'[\boldsymbol{I}_n-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{\varepsilon}}{n-K} \qquad \text{（参见第 3 章）}$$

$$=\frac{1}{n-K}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}] \qquad \text{（乘积展开）}$$

$$=\frac{n}{n-K}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n}-\frac{\boldsymbol{\varepsilon}'\boldsymbol{X}}{n}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n}\right] \qquad \text{（同时乘除 } n\text{）}$$

$$=\frac{n}{n-K}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2-\bar{\boldsymbol{g}}'\boldsymbol{S}_{XX}^{-1}\bar{\boldsymbol{g}}\right] \qquad \text{（}\bar{\boldsymbol{g}}\text{ 与 }\boldsymbol{S}_{XX}\text{ 的定义）}$$

由于 $\operatorname{plim}_{n\to\infty}\frac{n}{n-K}=1$，$\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2=\mathrm{E}(\varepsilon_i^2)=\sigma^2$（因为 $\{\varepsilon_i\}$ 为渐近独立的平稳序列），$\operatorname{plim}_{n\to\infty}\bar{\boldsymbol{g}}'\boldsymbol{S}_{XX}^{-1}\bar{\boldsymbol{g}}=\boldsymbol{0}'\cdot[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\cdot\boldsymbol{0}=0$，故 $\operatorname{plim}_{n\to\infty}s^2=\sigma^2$。

应该注意的是，由于 $\{\varepsilon_i\}$ 为严格平稳序列，故无条件方差 $\mathrm{E}(\varepsilon_i^2)$ 是一个常数，不依赖于 i。进一步，由于 $\{\varepsilon_i,\boldsymbol{x}_i\}$ 也是严格平稳序列，故条件方差 $\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)$ 的函数形式也不依赖于 i②；但是，$\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)$ 的具体取值却可以因 $\boldsymbol{x}_i$ 的取值不同而不同，因为到目前为止我们并没有假设“条件同方差”。

5.9 线性假设的大样本检验

1. 检验单个系数： $H_0:\beta_k=\bar{\beta}_k$。

在原假设 H_0 成立的情况下，$\sqrt{n}(b_k-\bar{\beta}_k)\xrightarrow{d}N(0,\mathrm{Avar}(b_k))$，其中 b_k 为 OLS 估计量 $\boldsymbol{b}$ 的第 k 个元素，而 $\mathrm{Avar}(b_k)$ 为矩阵 $\mathrm{Avar}(\boldsymbol{b})$ 的第 (k,k) 个元素。另一方面，$\widehat{\mathrm{Avar}}(b_k)$ 是对 $\mathrm{Avar}(b_k)$ 的一致估计，即 $\widehat{\mathrm{Avar}}(b_k)\xrightarrow{p}\mathrm{Avar}(b_k)$。定义 t 统计量为

$$t_k\equiv\frac{\sqrt{n}(b_k-\bar{\beta}_k)}{\sqrt{\widehat{\mathrm{Avar}}(b_k)}}=\frac{b_k-\bar{\beta}_k}{\sqrt{\frac{1}{n}\widehat{\mathrm{Avar}}(b_k)}}\equiv\frac{b_k-\bar{\beta}_k}{\mathrm{SE}^*(b_k)}\xrightarrow{d}N(0,1) \tag{5.10}$$

其中，$\mathrm{SE}^*(b_k)\equiv\sqrt{\frac{1}{n}\widehat{\mathrm{Avar}}(b_k)}=\sqrt{\frac{1}{n}(\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\boldsymbol{S}_{XX}^{-1})_{kk}}$ 被称为“异方差稳健的标准误”（heteroskedasticity-consistent standard errors），简称“稳健标准误”（robust standard errors，也称为 White's standard errors，Huber-White standard errors，或 Eicker-Huber-White standard errors），最早由

① 参见 Hayashi(2000, p. 124)。

② 例如，$\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)=\boldsymbol{x}_i'\boldsymbol{x}_i=\sum_{k=1}^{K}x_{ik}^2$，$\forall i=1,\cdots,n$。

Eicker(1967),Huber(1967)与 White(1980)提出,之后得到广泛应用。之所以这样称呼,是因为在前面的推导过程中并未用到"条件同方差"的假定,故在"条件异方差"的情况下也适用。统计量 t_k 被称为"稳健 t 比值",服从标准正态分布,而不是 t 分布。显然,$|t_k|$越大,则越倾向于拒绝 H_0。比如,对于显著性水平 5%,如果 $|t_k|$大于临界值 1.96,则拒绝 H_0。

命题 在条件同方差的假定下,稳健标准误还原为普通(非稳健)标准误。

证明:假设 $\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)=\sigma^2>0$(条件同方差),则根据迭代期望定律,

$$\boldsymbol{S}\equiv\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i'\varepsilon_i^2)=\mathrm{E}_{\boldsymbol{x}_i}\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i'\varepsilon_i^2|\boldsymbol{x}_i)=\mathrm{E}_{\boldsymbol{x}_i}[\boldsymbol{x}_i\boldsymbol{x}_i'\mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)]=\sigma^2\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$$

由于 $s^2\xrightarrow{p}\sigma^2$,$\boldsymbol{S}_{XX}\xrightarrow{p}\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$,故 $s^2\boldsymbol{S}_{XX}$是 $\boldsymbol{S}$ 的一致估计量,因此

$$\widehat{\mathrm{Avar}}(\boldsymbol{b})=\boldsymbol{S}_{XX}^{-1}(s^2\boldsymbol{S}_{XX})\boldsymbol{S}_{XX}^{-1}=s^2\boldsymbol{S}_{XX}^{-1}=s^2\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}=ns^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\tag{5.11}$$

$$\mathrm{SE}^*(b_k)=\sqrt{\frac{1}{n}\widehat{\mathrm{Avar}}(b_k)}=\sqrt{\frac{1}{n}\cdot ns^2(\boldsymbol{X}'\boldsymbol{X})_{kk}^{-1}}=\sqrt{s^2(\boldsymbol{X}'\boldsymbol{X})_{kk}^{-1}}\tag{5.12}$$

这个公式正是第 3 章"小样本 OLS"中普通(非稳健)标准误的公式。

2. 检验线性假设:$H_0:\underbrace{\boldsymbol{R}}_{m\times K}\underbrace{\boldsymbol{\beta}}_{K\times 1}=\underbrace{\boldsymbol{r}}_{m\times 1}$,**其中 $\boldsymbol{R}$ 满行秩。**

根据沃尔德检验原理,考察$(\boldsymbol{Rb}-\boldsymbol{r})$的大小,譬如其二次型 $(\boldsymbol{Rb}-\boldsymbol{r})'(\boldsymbol{Rb}-\boldsymbol{r})$。在 H_0 成立的情况下,统计量 $W\equiv n(\boldsymbol{Rb}-\boldsymbol{r})'[\boldsymbol{R}\widehat{\mathrm{Avar}}(\boldsymbol{b})\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})\xrightarrow{d}\chi^2(m)$。将这个表达式中的 n 拆成$\sqrt{n}\cdot\sqrt{n}$,则可以把 W 更直观地写为

$$W\equiv[\sqrt{n}(\boldsymbol{Rb}-\boldsymbol{r})]'[\boldsymbol{R}\widehat{\mathrm{Avar}}(\boldsymbol{b})\boldsymbol{R}']^{-1}[\sqrt{n}(\boldsymbol{Rb}-\boldsymbol{r})]\xrightarrow{d}\chi^2(m)\tag{5.13}$$

证明:记$\boldsymbol{c}_n\equiv\sqrt{n}(\boldsymbol{Rb}-\boldsymbol{r})$,$\boldsymbol{Q}_n\equiv\boldsymbol{R}\widehat{\mathrm{Avar}}(\boldsymbol{b})\boldsymbol{R}'$,则 $W=\boldsymbol{c}_n'\boldsymbol{Q}_n^{-1}\boldsymbol{c}_n$。在 H_0 成立的情况下,$\boldsymbol{c}_n\equiv\sqrt{n}(\boldsymbol{Rb}-\boldsymbol{r})=\sqrt{n}(\boldsymbol{Rb}-\boldsymbol{R\beta})=\sqrt{n}\boldsymbol{R}(\boldsymbol{b}-\boldsymbol{\beta})=\boldsymbol{R}[\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})]$。

因为$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})\xrightarrow{d}N(\boldsymbol{0},\mathrm{Avar}(\boldsymbol{b}))$,而$\boldsymbol{c}_n$是$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})$的线性组合,故$\boldsymbol{c}_n\xrightarrow{d}\boldsymbol{c}$,其中 $\boldsymbol{c}\sim N(\boldsymbol{0},\boldsymbol{R}\mathrm{Avar}(\boldsymbol{b})\boldsymbol{R}')$。

定义 $\boldsymbol{Q}\equiv\mathrm{Var}(\boldsymbol{c})=\boldsymbol{R}\mathrm{Avar}(\boldsymbol{b})\boldsymbol{R}'$,由于$\widehat{\mathrm{Avar}}(\boldsymbol{b})\xrightarrow{p}\mathrm{Avar}(\boldsymbol{b})$,故$\boldsymbol{Q}_n\xrightarrow{p}\boldsymbol{Q}$。

因此,$W=\boldsymbol{c}_n'\boldsymbol{Q}_n^{-1}\boldsymbol{c}_n\xrightarrow{d}\boldsymbol{c}'\boldsymbol{Q}^{-1}\boldsymbol{c}=\boldsymbol{c}'[\mathrm{Var}(\boldsymbol{c})]^{-1}\boldsymbol{c}\sim\chi^2(m)$①。

注:由于 $\boldsymbol{R}$ 满行秩,且 $\mathrm{Avar}(\boldsymbol{b})$为正定矩阵,故$\boldsymbol{Q}^{-1}$存在。

有关"非线性假设"的检验,参见本章附录。

5.10 大样本 OLS 的 Stata 命令及实例

在 Stata 中,可以很方便地得到 OLS 估计的稳健标准误,其命令为

```
reg y x1 x2 x3,robust
```

其中,选择项"robust"表示稳健标准误。仍以 Nerlove(1963)数据为例。为了方便对比,把使用普通标准误的结果复制如下。

---

① 根据数理统计知识,如果 m 维随机变量 $\boldsymbol{x}$ 服从正态分布 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$,其中 $\boldsymbol{\Sigma}$ 为非退化矩阵(满秩),则二次型$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sim\chi^2(m)$。

```
. use nerlove.dta,clear
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs = 145
				F(4, 140) = 437.90
Model	269.524728	4	67.3811819	Prob > F = 0.0000
Residual	21.5420958	140	.153872113	R-squared = 0.9260
				Adj R-squared = 0.9239
Total	291.066823	144	2.02129738	Root MSE = .39227

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

检验变量 lnq 的系数是否为1。

```
. test lnq =1
```

```
 ( 1)  lnq = 1

       F(  1,   140) =  256.27
            Prob > F =    0.0000
```

其次,使用稳健标准误重新进行回归。

```
. reg lntc lnq lnpl lnpk lnpf,r
```

Linear regression				Number of obs = 145
				F(4, 140) = 177.19
				Prob > F = 0.0000
				R-squared = 0.9260
				Root MSE = .39227

lntc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0325376	22.16	0.000	.656585	.785242
lnpl	.4559645	.260326	1.75	0.082	-.0587139	.9706429
lnpk	-.2151476	.3233711	-0.67	0.507	-.8544698	.4241745
lnpf	.4258137	.0740741	5.75	0.000	.2793653	.5722622
_cons	-3.566513	1.718304	-2.08	0.040	-6.963693	-.1693331

对比以上两个结果可知,使用选择项"robust"所得到的OLS回归系数完全相同,只是标准误(Std. Err.)变为稳健标准误(Robust Std. Err.),与普通标准误不同[①]。对于变量 lnq 的系数,其稳健标准误(0.033)几乎是普通标准误(0.017)的两倍。另一方面,其他变量系数的稳健标准误与普通标准误比较接近。如果认为存在异方差,则应使用稳健标准误。在异方差的情况下,如果使用普通标准误,将大大低估变量 lnq 系数的真实标准误,从而导致不正确的统计推断。如何检验异方差,将在第7章进行。

① 不同的软件包所汇报的标准误可能略有不同,可能的原因在于是否作"自由度调整"(degree of freedom adjustment),即是用 n 还是 $(n-K)$ 作为"标准误估计量"的分母。

在 Stata 中使用稳健标准误，即可进行大样本检验。对单个变量系数显著性的检验，可以使用上表中的稳健 t 统计量（服从正态分布）来进行。更直观地，可以直接看表中所列的 p 值（即“P > |t|”）。对于更一般的线性假设，仍可使用命令 test 来检验。比如，检验变量 lnq 的系数是否为 1：

```
. test lnq =1
```

```
( 1)  lnq = 1

      F(  1,   140) =   73.57
           Prob > F =    0.0000
```

由于 p 值为 0.0000，故即使使用稳健标准误，也仍然强烈拒绝“变量 lnq 的系数为 1”的原假设。需要注意的是，在使用稳健标准误的情况下，Stata 仍然汇报 F 统计量（服从 F 分布），即依然使用小样本理论中的 F 统计量公式，但将协方差矩阵换成“稳健的协方差矩阵”。事实上，F 分布与 χ^2 分布在大样本下是等价的，参见本章附录。

对于非线性假设的检验，Stata 命令为 testnl。比如，检验变量 lnpl 的系数是 lnq 的系数的平方：

```
. testnl _b[lnpl] =_b[lnq]^2
```

```
(1)  _b[lnpl] = _b[lnq]^2

           F(1, 140) =        0.05
            Prob > F =        0.8164
```

由于 p 值为 0.82，故无法拒绝原假设。

习　题

5.1　确定性序列 $\{x_n\}_{n=1}^{\infty}$ 可以被看成是退化的随机序列。证明：如果 $\lim\limits_{n\to\infty} x_n = a$，则 $\operatorname{plim}\limits_{n\to\infty} x_n = a$。

5.2　假设 $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0,\sigma^2)$，证明 $\hat{\beta}_n \xrightarrow{p} \beta$。

5.3　假设 $\hat{\beta}$ 是 β 的一致估计量。定义 $\theta \equiv \log(\beta)$，证明 $\hat{\theta} \equiv \log(\hat{\beta})$ 是 θ 的一致估计量。

5.4　鞅差分过程一定是白噪声吗？为什么？

5.5　假设 $\{x_i\}$ 是确定性序列，其取值随 i 而变化。$\{\varepsilon_i\}$ 是独立同分布的随机序列，期望为 0，方差存在。回答以下问题，并解释原因。

（1）序列 $\{x_i\varepsilon_i\}$ 是独立同分布的吗？

（2）序列 $\{x_i\varepsilon_i\}$ 存在自相关吗？

（3）序列 $\{x_i\varepsilon_i\}$ 是鞅差分序列吗？

（4）序列 $\{x_i\varepsilon_i\}$ 是严格平稳序列吗？

5.6　当 $n\to\infty$ 时，是否 $\mathrm{SE}^*(b_k) \xrightarrow{p} 0$？为什么？

附　录

A5.1　依均方收敛是依概率收敛的充分条件

证明：设随机序列 $\{x_n\}_{n=1}^{\infty}$ 依均方收敛于常数 a，即 $\lim\limits_{n\to\infty}\mathrm{E}(x_n) = a$，$\lim\limits_{n\to\infty}\mathrm{Var}(x_n) = 0$。根据切比雪夫不等式，对

于任意 $\varepsilon>0$,都有

$$P(|x_n-E(x_n)|\geqslant\varepsilon)\leqslant\frac{Var(x_n)}{\varepsilon^2}$$

当 $n\to\infty$ 时,对此不等式两边同时取极限可得,

$$\lim_{n\to\infty}P(|x_n-a|\geqslant\varepsilon)=\lim_{n\to\infty}P(|x_n-E(x_n)|\geqslant\varepsilon)\leqslant\lim_{n\to\infty}\frac{Var(x_n)}{\varepsilon^2}=0$$

根据依概率收敛的定义可知,$\{x_n\}_{n=1}^{\infty}$ 依概率收敛于 a。

A5.2 渐近独立过程的定义

如果对于任意两个有界函数 $f:\mathbf{R}^{k+1}\to\mathbf{R}$ 与 $g:\mathbf{R}^{l+1}\to\mathbf{R}$,都有

$$\lim_{n\to\infty}\{|E[f(x_i,\cdots,x_{i+k})g(x_{i+n},\cdots,x_{i+l+n})]|-|E[f(x_i,\cdots,x_{i+k})]|\cdot|E[g(x_{i+n},\cdots,x_{i+l+n})]|\}=0$$

则称随机过程 $\{x_i\}_{i=1}^{\infty}$ 是一个"渐近独立过程"。

A5.3 F 分布与 χ^2 分布在大样本下是等价的

命题 假设 $F\sim F(m,n-K)$ 分布,则当 $n\to\infty$ 时,$mF\xrightarrow{d}\chi^2(m)$。

证明:因为 $F\sim F(m,n-K)$,故可设 $F=\dfrac{\chi^2(m)/m}{\chi^2(n-K)/(n-K)}$。

根据 χ^2 分布的性质,$E[\chi^2(n-K)]=n-K$,而 $Var[\chi^2(n-K)]=2(n-K)$,故 F 统计量分母的期望值为 $E[\chi^2(n-K)/(n-K)]=1$,而 F 统计量分母的方差为 $Var[\chi^2(n-K)/(n-K)]=\dfrac{2(n-K)}{(n-K)^2}=\dfrac{2}{n-K}\to0$(当 $n\to\infty$ 时)。因此,分母依均方收敛于1,进而分母依概率收敛于1,即 $\chi^2(n-K)/(n-K)\xrightarrow{p}1$。

所以,$F\xrightarrow{d}\chi^2(m)/m$,$mF\xrightarrow{d}\chi^2(m)$。

A5.4 非线性假设的大样本检验

记要检验的非线性原假设为"$H_0:\boldsymbol{a}(\boldsymbol{\beta})=\mathbf{0}$",其中 $\boldsymbol{a}(\cdot)$ 为向量函数(vector-valued function)且其一阶导数连续。假设 $\boldsymbol{a}(\cdot)$ 的维度是 m,即共有 m 个非线性约束。记 $\boldsymbol{A}(\boldsymbol{\beta})\equiv\dfrac{\partial\boldsymbol{a}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$ 为雅可比矩阵(Jacobian),即由 $\boldsymbol{a}(\beta)$ 的一阶偏导数组成的 $m\times K$ 矩阵,其中 K 是 $\boldsymbol{\beta}$ 的维度。假设矩阵 $\boldsymbol{A}(\boldsymbol{\beta})$ 满行秩(在检验线性假设时也有类似要求)。为了检验这个非线性假设,首先要介绍一个命题。

命题(Delta法,The Delta Method) 假设 $\{\boldsymbol{x}_n\}$ 是一个 K 维随机向量序列,满足 $\boldsymbol{x}_n\xrightarrow{p}\boldsymbol{\beta}$,并且 $\sqrt{n}(\boldsymbol{x}_n-\boldsymbol{\beta})\xrightarrow{d}z$,其中 z 为某随机变量;假设向量函数 $\boldsymbol{a}(\cdot):\mathbf{R}^K\to\mathbf{R}^m$ 有连续的一阶导数,且记其在 $\boldsymbol{\beta}$ 处的雅可比矩阵为 $\underbrace{\boldsymbol{A}(\boldsymbol{\beta})}_{m\times K}\equiv\dfrac{\partial\boldsymbol{a}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$,则

$$\sqrt{n}[\boldsymbol{a}(\boldsymbol{x}_n)-\boldsymbol{a}(\boldsymbol{\beta})]\xrightarrow{d}\boldsymbol{A}(\boldsymbol{\beta})z$$

证明:由于 $\boldsymbol{x}_n\xrightarrow{p}\boldsymbol{\beta}$,故 $\boldsymbol{a}(\boldsymbol{x}_n)\xrightarrow{p}\boldsymbol{a}(\boldsymbol{\beta})$,因此,$\boldsymbol{a}(\boldsymbol{x}_n)-\boldsymbol{a}(\boldsymbol{\beta})\xrightarrow{p}\mathbf{0}$。根据微分中值定理,存在一个位于 $\boldsymbol{x}_n$ 与 $\boldsymbol{\beta}$ 之间的向量 $\boldsymbol{y}_n$(参见图5.7),使得

$$\boldsymbol{a}(\boldsymbol{x}_n)-\boldsymbol{a}(\boldsymbol{\beta})=\frac{\partial\boldsymbol{a}(\boldsymbol{y}_n)}{\partial\boldsymbol{y}_n'}(\boldsymbol{x}_n-\boldsymbol{\beta})=\underbrace{\boldsymbol{A}(\boldsymbol{y}_n)}_{m\times K}\underbrace{(\boldsymbol{x}_n-\boldsymbol{\beta})}_{K\times1}$$

上式两边同时乘以 $\sqrt{n}$ 可得,

$$\sqrt{n}[\boldsymbol{a}(\boldsymbol{x}_n)-\boldsymbol{a}(\boldsymbol{\beta})]=\boldsymbol{A}(\boldsymbol{y}_n)\sqrt{n}(\boldsymbol{x}_n-\boldsymbol{\beta})$$

由于 $\boldsymbol{y}_n$ 介于 $\boldsymbol{x}_n$ 与 $\boldsymbol{\beta}$ 之间,而 $\boldsymbol{x}_n\xrightarrow{p}\boldsymbol{\beta}$,故 $\boldsymbol{y}_n\xrightarrow{p}\boldsymbol{\beta}$。由于 $\boldsymbol{A}(\cdot)$ 为连续函数,故 $\boldsymbol{A}(\boldsymbol{y}_n)\xrightarrow{p}\boldsymbol{A}(\boldsymbol{\beta})$。而根据题设,$\sqrt{n}(\boldsymbol{x}_n-\boldsymbol{\beta})\xrightarrow{d}z$。因此,

$$\sqrt{n}[\boldsymbol{a}(\boldsymbol{x}_n)-\boldsymbol{a}(\boldsymbol{\beta})]\xrightarrow{d}\boldsymbol{A}(\boldsymbol{\beta})z$$

介绍了 Delta 法之后，就可以证明如下命题。

命题　在上述原假设“$H_0: \boldsymbol{a}(\boldsymbol{\beta}) = \mathbf{0}$”成立的前提下，统计量

$$W \equiv n \cdot \boldsymbol{a}(\boldsymbol{b})' [\boldsymbol{A}(\boldsymbol{b}) \widehat{\mathrm{Avar}}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{b})']^{-1} \boldsymbol{a}(\boldsymbol{b}) \xrightarrow{d} \chi^2(m) \tag{5.14}$$

证明：由于 $\sqrt{n}[\boldsymbol{b} - \boldsymbol{\beta}] \xrightarrow{d} N(\mathbf{0}, \mathrm{Avar}(\boldsymbol{b}))$，根据 Delta 法，

$$\sqrt{n}[\boldsymbol{a}(\boldsymbol{b}) - \boldsymbol{a}(\boldsymbol{\beta})] \xrightarrow{d} \boldsymbol{c}$$

其中 $\boldsymbol{c} \sim N(\mathbf{0}, \boldsymbol{A}(\boldsymbol{\beta}) \mathrm{Avar}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{\beta})')$。

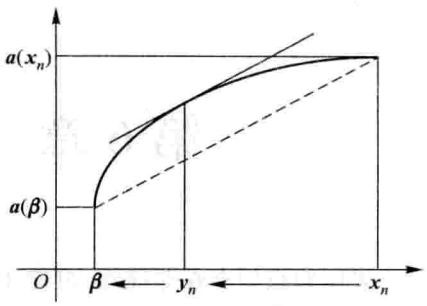


图 5.7　微分中值定理示意图

根据原假设 $H_0: \boldsymbol{a}(\boldsymbol{\beta}) = \mathbf{0}$，故 $\sqrt{n}\boldsymbol{a}(\boldsymbol{b}) \xrightarrow{d} \boldsymbol{c}$，其中

$$\boldsymbol{c} \sim N(\mathbf{0}, \boldsymbol{A}(\boldsymbol{\beta}) \mathrm{Avar}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{\beta})')$$

由于 $\boldsymbol{A}(\boldsymbol{b}) \xrightarrow{p} \boldsymbol{A}(\boldsymbol{\beta})$，$\widehat{\mathrm{Avar}}(\boldsymbol{b}) \xrightarrow{p} \mathrm{Avar}(\boldsymbol{b})$，故

$$\boldsymbol{A}(\boldsymbol{b}) \widehat{\mathrm{Avar}}(\boldsymbol{b}) A(\boldsymbol{b})' \xrightarrow{p} \boldsymbol{A}(\boldsymbol{\beta}) \mathrm{Avar}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{\beta}) = \mathbf{Var}(\boldsymbol{c})$$

由于 $\boldsymbol{A}(\boldsymbol{\beta})$ 满行秩，而 $\mathrm{Avar}(\boldsymbol{b})$ 为正定矩阵，故矩阵 $\mathbf{Var}(\boldsymbol{c})$ 可逆。因此，

$$\begin{aligned} W &\equiv n \cdot \boldsymbol{a}(\boldsymbol{b})' [\boldsymbol{A}(\boldsymbol{b}) \widehat{\mathrm{Avar}}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{b})']^{-1} \boldsymbol{a}(\boldsymbol{b}) \\ &= \sqrt{n}\boldsymbol{a}(\boldsymbol{b})' [\boldsymbol{A}(\boldsymbol{b}) \widehat{\mathrm{Avar}}(\boldsymbol{b}) \boldsymbol{A}(\boldsymbol{b})']^{-1} \sqrt{n}\boldsymbol{a}(\boldsymbol{b}) \xrightarrow{d} \boldsymbol{c}' \mathbf{Var}(\boldsymbol{c})^{-1} \boldsymbol{c} \sim \chi^2(m) \end{aligned}$$

第 6 章　最大似然估计法

如果回归模型存在非线性，可能不方便使用 OLS，这时常常采用最大似然估计法（MLE）或非线性最小二乘法（NLS，参见第 25 章）。

6.1　最大似然估计法的定义

假设随机向量 $\boldsymbol{y}$ 的概率密度函数为 $f(\boldsymbol{y};\boldsymbol{\theta})$，其中 $\boldsymbol{\theta}$ 为 K 维未知参数向量，$\boldsymbol{\theta}\in\boldsymbol{\Theta}$，其中 $\boldsymbol{\Theta}$ 为参数空间，即参数 $\boldsymbol{\theta}$ 所有可能取值所构成的集合。通过抽取随机样本 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 来估计 $\boldsymbol{\theta}$。假设 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 为独立同分布（iid），则样本数据的联合密度函数为 $f(\boldsymbol{y}_1;\boldsymbol{\theta})f(\boldsymbol{y}_2;\boldsymbol{\theta})\cdots f(\boldsymbol{y}_n;\boldsymbol{\theta})$。

在抽样之前，$\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 被视为随机向量。抽样之后，$\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 就有了特定的样本值。因此，可以将样本的联合密度函数视为在 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 给定的情况下，未知参数 $\boldsymbol{\theta}$ 的函数。定义“**似然函数**”（likelihood function）为

$$L(\boldsymbol{\theta};\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)=\prod_{i=1}^{n}f(\boldsymbol{y}_i;\boldsymbol{\theta}) \tag{6.1}$$

由此可知，似然函数与联合密度函数完全相等，只是 $\boldsymbol{\theta}$ 与 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 的角色互换，即把 $\boldsymbol{\theta}$ 作为自变量，而视 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\}$ 为已给定的。为了运算方便，常把似然函数取对数，将乘积的形式转化为求和的形式

$$\ln L(\boldsymbol{\theta};\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)=\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}) \tag{6.2}$$

“最大似然估计法”（Maximum Likelihood Estimation，简记 MLE 或 ML）来源于一个简单而深刻的思想：给定样本取值后，该样本最有可能来自参数 $\boldsymbol{\theta}$ 为何值的总体。换言之，寻找 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$，使得观测到样本数据的可能性最大，即最大化对数似然函数（loglikelihood function）[①]：

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\ln L(\boldsymbol{\theta};\boldsymbol{y}) \tag{6.3}$$

在数学上，常把最大似然估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ 写为

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\equiv\operatorname{argmax}\ln L(\boldsymbol{\theta};\boldsymbol{y}) \tag{6.4}$$

其中，“argmax”（即 argument of the maximum）表示能使 $\ln L(\boldsymbol{\theta};\boldsymbol{y})$ 最大化的 $\boldsymbol{\theta}$ 取值。假设存在唯一内点解，则该无约束极值问题的一阶条件为

$$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y})\equiv\frac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}}\equiv\begin{pmatrix}\dfrac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\theta_1}\\ \vdots\\ \dfrac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\theta_K}\end{pmatrix}=\boldsymbol{0} \tag{6.5}$$

① 由于对数函数为单调函数，故“对数似然函数最大化”等价于“似然函数最大化”。

此一阶条件要求,对数似然函数的梯度向量(gradient,即一阶偏导数、斜率)$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y})$为$\mathbf{0}$。这实际上是一个由$K$个未知参数$(\theta_1,\theta_2,\cdots,\theta_K)$、$K$个方程构成的方程组。该梯度向量也称为"得分函数"(score function)或"得分向量"(score vector)。得分函数$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y})$本身是$\boldsymbol{y}$的函数,故也是随机向量。在本章下面的讨论中,记真实参数为$\boldsymbol{\theta}_0$,而$\boldsymbol{\theta}$为该参数的任何可能取值。可以证明,得分函数的期望值也为$\mathbf{0}$。

命题(得分函数的期望为0) 如果似然函数正确(correctly specified),则$\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]=\mathbf{0}$,其中$\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})$表示得分函数$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y})$在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处的取值,证明参见本章附录。

例 假设随机样本$y_i\sim N(\theta_0,1)$,$i=1,\cdots,n$。则样本数据的对数似然函数为

$$L(\theta)=-\frac{n}{2}\ln 2\pi-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2 \tag{6.6}$$

其得分函数为

$$s(\theta)=\frac{\partial L(\theta)}{\partial\theta}=\sum_{i=1}^{n}(y_i-\theta) \tag{6.7}$$

故得分函数的期望值为

$$\mathrm{E}[s(\theta)]=E\left[\frac{\partial L(\theta)}{\partial\theta}\right]=\sum_{i=1}^{n}[\mathrm{E}(y_i)-\theta]=\sum_{i=1}^{n}[\theta_0-\theta] \tag{6.8}$$

在$\theta=\theta_0$处

$$E[s(\theta)]\big|_{\theta=\theta_0}=\sum_{i=1}^{n}[\theta_0-\theta]\Big|_{\theta=\theta_0}=0 \tag{6.9}$$

显然,方程(6.9)的结果与上述命题是一致的。

进一步,可以将得分函数分解为

$$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y})\equiv\frac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}}=\frac{\partial\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\sum_{i=1}^{n}\frac{\partial\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\equiv\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\theta};\boldsymbol{y}_i) \tag{6.10}$$

其中,$\boldsymbol{s}_i(\boldsymbol{\theta};\boldsymbol{y}_i)\equiv\frac{\partial\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$为第$i$个观测值对得分函数的贡献。回到上面的例子,则$s(\theta;\boldsymbol{y})=\sum_{i=1}^{n}(y_i-\theta)$,而$\boldsymbol{s}_i(\theta,y_i)=(y_i-\theta)$。二阶条件要求,对数似然函数的黑赛矩阵(Hessian matrix)$\frac{\partial^2\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\equiv\frac{\partial\left(\frac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}'}$①为负定矩阵,即对数似然函数必须为严格凹函数(strictly concave function)。类似地,黑赛矩阵可以分解为

$$\boldsymbol{H}(\boldsymbol{\theta};\boldsymbol{y})\equiv\frac{\partial^2\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=\sum_{i=1}^{n}\frac{\partial^2\ln L(\boldsymbol{\theta};\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\equiv\sum_{i=1}^{n}\boldsymbol{H}_i(\boldsymbol{\theta};\boldsymbol{y}_i) \tag{6.11}$$

其中,$\boldsymbol{H}_i(\boldsymbol{\theta};\boldsymbol{y}_i)$为第$i$个观测值对黑赛矩阵的贡献。在上文的一维例子中,$\boldsymbol{H}(\theta;\boldsymbol{y})=\frac{\partial s(\theta)}{\partial\theta}=\frac{\partial\sum_{i=1}^{n}(y_i-\theta)}{\partial\theta}=-n$,而$\boldsymbol{H}_i(\theta;y_i)=\frac{\partial s_i(\theta)}{\partial\theta}=\frac{\partial(y_i-\theta)}{\partial\theta}=-1$。

---

① 其中,$\partial\boldsymbol{\theta}$表示对$\boldsymbol{\theta}$的各个元素分别求偏导,然后排成一个列向量;而$\partial\boldsymbol{\theta}'$表示对$\boldsymbol{\theta}$的各个元素分别求偏导,然后排成一个行向量。一个多元函数的黑赛矩阵包含了其所有二阶混合偏导数。

6.2 线性回归模型的最大似然估计

有关最大似然估计的计算，先来看一个数理统计中的简单例子。

例 假设 $X \sim N(\mu, \sigma^2)$，其中 σ^2 已知，得到一个样本容量为 1 的样本 $x_1 = 2$，求对 μ 的最大似然估计。似然函数为 $L(\mu) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\dfrac{-(2-\mu)^2}{2\sigma^2}\right\}$。显然，似然函数在 $\hat{\mu} = 2$ 处取最大值，参见图 6.1。

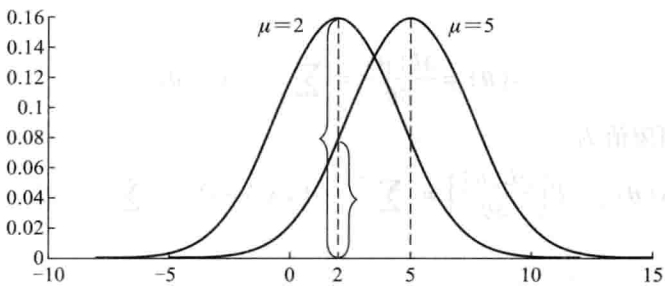


图 6.1 选择参数使观测到样本的可能性最大

例（非正式） 某人操一口浓重的四川口音，则判断他最有可能来自四川。

下面以线性回归模型为例，来说明最大似然法估计的实际操作。假设线性回归模型为（参见第 3 章）

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.12}$$

为了使用 MLE，首先要对扰动项的条件概率分布进行假设，即假设 $\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$。因此，被解释变量的条件分布为 $\boldsymbol{y} \mid \boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$，其条件概率密度函数为（参见第 2 章，多维正态分布的联合密度公式）

$$f(\boldsymbol{y} \mid \boldsymbol{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\} \tag{6.13}$$

用假想值 $\tilde{\boldsymbol{\beta}}$，$\tilde{\sigma}^2$ 代替真实值 $\boldsymbol{\beta}$，σ^2，并取对数可得

$$\ln L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \tag{6.14}$$

其中，右边第一项与 $\tilde{\boldsymbol{\beta}}$，$\tilde{\sigma}^2$ 无关，可以忽略。这个最大化问题可以分两步进行。第一步，在给定 $\tilde{\sigma}^2$ 的情况下，选择最优 $\tilde{\boldsymbol{\beta}}$[①]。第二步，代入第一步中得到的最优 $\tilde{\boldsymbol{\beta}}$，选择最优 $\tilde{\sigma}^2$。

在第一步，选择 $\tilde{\boldsymbol{\beta}}$ 使得 $\ln L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ 最大。由于 $\tilde{\boldsymbol{\beta}}$ 只出现在第三项中，故这等价于使 $(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$ 最小。而这正是最小二乘法的目标函数 $\boldsymbol{e}'\boldsymbol{e}$。因此

---

① 一般来说，最优的 $\tilde{\boldsymbol{\beta}}$ 依赖于 $\tilde{\sigma}^2$，即 $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\tilde{\sigma}^2)$。但对于这个问题，最优的 $\tilde{\boldsymbol{\beta}}$ 与 $\tilde{\sigma}^2$ 无关。这使得第二步的最优化计算大大简化。

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}}=\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{6.15}$$

在第二步，对数似然函数变为 $-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\tilde{\sigma}^2-\frac{1}{2\tilde{\sigma}^2}\boldsymbol{e}'\boldsymbol{e}$，被称为“集中对数似然函数”（concentrated log likelihood function），因为其 $\tilde{\boldsymbol{\beta}}$ 的取值已在第一步中固定，称为“concentrated with respect to $\tilde{\boldsymbol{\beta}}$”。对 $\tilde{\sigma}^2$ 求导可得

$$-\frac{n}{2}\frac{1}{\tilde{\sigma}^2}+\frac{1}{2\tilde{\sigma}^4}\boldsymbol{e}'\boldsymbol{e}=0 \tag{6.16}$$

求解 σ^2 的 MLE 估计量为

$$\hat{\sigma}^2_{\mathrm{ML}}=\frac{\boldsymbol{e}'\boldsymbol{e}}{n}\neq\hat{\sigma}^2_{\mathrm{OLS}}=\frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\equiv s^2 \tag{6.17}$$

由此可知，MLE 对回归系数 $\boldsymbol{\beta}$ 的估计与 OLS 是完全一样的，对扰动项方差 σ^2 的估计则略有不同，但此差别在大样本下消失。由于 OLS 估计量 s^2 是对 σ^2 的无偏估计，故 MLE 估计量 $\hat{\sigma}^2_{\mathrm{ML}}$ 是有偏的（小样本性质）。MLE 的主要优点是大样本性质良好，比如一致性、最小渐近方差。

6.3 最大似然估计的数值解

如果模型存在非线性，最大似然估计通常没有解析解，而只能寻找“数值解”（numerical solution）。方法之一为“网格搜索”（grid search）。如果待估参数 θ 是一维的，且大致知道其取值范围，比如 $\theta\in(0,1)$，则可将此区间划分为 10 个等分（0.1，…，0.9），分别计算目标函数的取值，以得到 $\hat{\theta}_{\max}$；然后再把（$\hat{\theta}_{\max}-0.1$，$\hat{\theta}_{\max}+0.1$）划分为 10 个等分，以此类推，直至达到理想的精度为止。然而，如果待估参数 $\boldsymbol{\theta}$ 为多维，或者我们对 $\boldsymbol{\theta}$ 的取值范围所知不多，则网格搜索法变得不现实。因此，在实践中，一般使用“迭代法”（iteration）进行数值求解。常用的迭代法为“高斯－牛顿法”（Gauss-Newton method），其基本思想如下。

假设欲求非线性方程 $f(x)=0$ 的解，而 $f(x)$ 的导数 $f'(x)$ 处处存在。记该方程的解为 x^*，满足 $f(x^*)=0$，参见图 6.2。首先预测一个初始值 x_0，在点 $(x_0,f(x_0))$ 处作一条曲线 $f(x)$ 的切线，记此切线与横轴的交点为 x_1。然后在点 $(x_1,f(x_1))$ 处再作一条曲线 $f(x)$ 的切线，记此切线与横轴的交点为 x_2。以此类推，不断迭代（iteration），可以得到序列 $\{x_0,x_1,x_2,x_3,\cdots\}$，其递推公式为①

$$x_{i+1}=x_i-\frac{f(x_i)}{f'(x_i)} \tag{6.18}$$

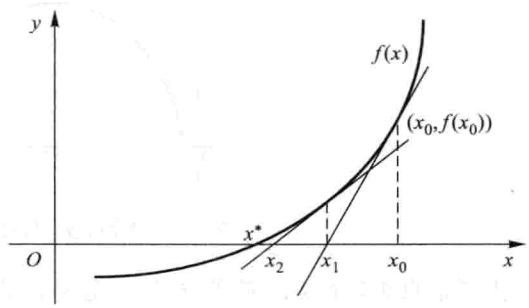


图 6.2 高斯－牛顿法

在一般情况下，该序列将收敛至 x^*（给定一个精确度，收敛到这个精确度范围内即停止）。高斯－牛顿法之所以常用，原因之一是它的收

① 连接 $(x_i,f(x_i))$ 与 $(x_{i+1},0)$ 的切线斜率为 $f'(x_i)$，故 $\frac{f(x_i)-0}{x_i-x_{i+1}}=f'(x_i)$，等式变形即得到迭代公式。

敛速度很快，是二次的。比如，如果本次迭代的误差为0.1，则下次迭代的误差约为0.1^2，而下下次迭代的误差约为0.1^4，等等。因此，常常只需要迭代几次就够了。当然，如果初始值x_0选择不当，也可能出现迭代不收敛的情形。另外，使用高斯－牛顿法得到的可能只是“局部最大值”(local maximum)，而非“整体最大值”(global maximum)。

高斯－牛顿法也适用于多元函数的情形$f(\boldsymbol{x})=0$，只要在上述迭代过程中，将切线替换为(超)切平面即可。高斯－牛顿法相当于对原函数$f(x)$作了一阶近似。一个改进方法是，在每次迭代过程中，对原函数$f(x)$作二阶近似(二阶泰勒展开)，这被称为“牛顿－拉弗森法”(Newton-Raphson method)。

6.4 信息矩阵与无偏估计的最小方差

为了研究MLE的大样本性质，定义“**信息矩阵**”(information matrix)为对数似然函数的黑赛矩阵之期望值(对$\boldsymbol{y}$求期望)的负数，即

$$\boldsymbol{I}(\boldsymbol{\theta}) \equiv -\mathrm{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] \tag{6.19}$$

在一维情形下，$-\dfrac{\partial^2 \ln L}{\partial \theta^2}$即为对数似然函数的二阶导数之负数。由于对数似然函数为凹函数，故其二阶导数为负数，因此加上负号以得到正数。更一般地，$-\dfrac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$表示的是对数似然函数在$\boldsymbol{\theta}$空间中的曲率(curvature)，取期望值之后的$\boldsymbol{I}(\boldsymbol{\theta})$即表示平均曲率(对$\boldsymbol{y}$进行平均)。如果曲率大，对数似然函数很陡峭，则较容易根据样本分辨真实$\boldsymbol{\theta}$的位置；反之，如果曲率小，对数似然函数很平坦，则不易根据样本判断真实$\boldsymbol{\theta}$的位置，参见图6.3。在极端情况下，如果似然函数完全平坦，则似然函数不存在唯一的最大值，即MLE没有唯一解；此时，无法根据样本数据来判断$\boldsymbol{\theta}$的位置。由于$\boldsymbol{I}(\boldsymbol{\theta})$包含了$\boldsymbol{\theta}$是否容易估计的信息，故称为“信息矩阵”。

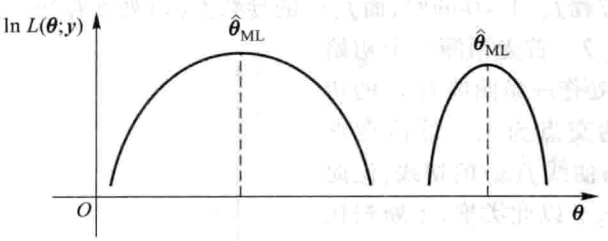


图6.3 平坦(左)与陡峭(右)的对数似然函数

由于信息矩阵涉及二阶偏导数，不容易计算，故常将其表达为一阶偏导数的乘积形式。

命题(信息矩阵等式) 在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处，以下“信息矩阵等式”(information matrix equality)成立，

$$\boldsymbol{I}(\boldsymbol{\theta}_0) \equiv -\mathrm{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = \mathrm{E}\left[\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}}\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}'}\right] = \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})'] \tag{6.20}$$

证明参见本章附录。进一步，可以证明以下命题。

命题(得分函数的方差为信息矩阵) 在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处，信息矩阵$\boldsymbol{I}(\boldsymbol{\theta}_0)$就是得分函数的协方差

矩阵 $\mathrm{Var}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]$。

证明：

$$\begin{aligned}\mathrm{Var}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})] &= \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})'] - \underbrace{\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]}_{=0}\underbrace{\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]'}_{=0} \\ &= \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})'] \\ &= \boldsymbol{I}(\boldsymbol{\theta}_0)\end{aligned} \tag{6.21}$$

其中，$\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]=\boldsymbol{0}$，而上式的最后一步用到了信息矩阵等式。

在统计学中有一个著名的结论：假设 $\hat{\boldsymbol{\theta}}$ 是对真实参数 $\boldsymbol{\theta}_0$ 的任意无偏估计，则在一定的正则条件（regularity conditions）下，$\hat{\boldsymbol{\theta}}$ 的方差不会小于 $[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$，即 $\mathrm{Var}(\hat{\boldsymbol{\theta}}) \geqslant [\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$。其中，$[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$称为“克莱默－劳下限”（Cramer-Rao Lower Bound），证明参见附录。因此，无偏估计所能达到的最小方差与信息矩阵（即对数似然函数的平均曲率）有关。曲率 $\boldsymbol{I}(\boldsymbol{\theta}_0)$ 越大，则 $[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$越小，无偏估计可能达到的最小方差越小。在古典线性回归模型中，根据信息矩阵的定义可以证明（参见附录）

$$[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1} = \begin{pmatrix} \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & 2\sigma^4/n \end{pmatrix} \tag{6.22}$$

其中，$\boldsymbol{\theta}_0=(\boldsymbol{\beta}\ \sigma^2)'$。由于 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，故 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}=\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$均达到了无偏估计的最小方差。

命题 在高斯－马尔可夫定理中，如果加上扰动项为正态分布的假定，则 OLS 是“最佳无偏估计”（Best Unbiased Estimator，简记 BUE），而不仅仅是 BLUE。

当然，MLE 并不一定是无偏估计（比如，对于古典线性回归扰动项方差 σ^2 的 MLE 估计就是有偏的，参见本章第 2 节）。但上述克莱默－劳下限的结论也可以推广到渐近分布的情形，即在一定的正则条件下，对于真实参数 $\boldsymbol{\theta}_0$ 的渐近正态一致估计（Consistent and Asymptotically Normally distributed estimators，简记 CAN）所能达到的最小方差为 $[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$，即克莱默－劳下限[①]。

6.5 最大似然法的大样本性质

MLE 之所以被广泛应用，是因为 MLE 估计量拥有良好的大样本性质。

定理（MLE 的大样本性质） 在一定的正则条件下[②]，MLE 估计量拥有以下良好的大样本性质。

（1）一致性，即 $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\theta}}_{\mathrm{ML}}=\boldsymbol{\theta}_0$。

（2）渐近有效性，即渐近协方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})=n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$，在大样本下达到了克莱默－劳下限。

---

① 在极少数个别情况下，仍可能出现渐近方差比 $[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$更小的渐近正态一致估计量，但这些例外情形在现实中都不重要。

② 这些正则条件包括，参数空间 Θ 为紧集（即有界闭集），真实参数 θ 在参数空间 Θ 的内部，样本为独立同分布的，样本的取值范围（support）不依赖于参数 θ，对数似然函数的三阶导数存在且有界，信息矩阵 $I(\theta)$ 在真实参数 θ 附近为正定矩阵等。这些正则条件一般都能满足，故不必太在意。

(3) 渐近正态，即$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}-\boldsymbol{\theta}_0)\xrightarrow{d}N(\boldsymbol{0},n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1})$，可以近似地认为$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\xrightarrow{d}N(\boldsymbol{\theta}_0,[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1})$。这表明，$(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}-\boldsymbol{\theta}_0)$趋于0的速度与$\frac{1}{\sqrt{n}}$趋于0的速度类似，即“$\sqrt{n}$收敛”(root-n convergence)。

此定理的严格证明需用到较多数学，下面为证明梗概。

证明(选读)：为了使用大数定律，将对数似然函数$\ln L(\boldsymbol{\theta};\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)=\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})$除以$n$，并定义

$$Q_n(\boldsymbol{\theta})\equiv\frac{1}{n}\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}) \tag{6.23}$$

显然，可以使用$Q_n(\boldsymbol{\theta})$作为MLE最大化的目标函数($Q_n(\boldsymbol{\theta})$既是参数$\boldsymbol{\theta}$的函数，又是样本容量为n的样本数据的函数，故记为$Q_n(\boldsymbol{\theta})$)。对于任何$\boldsymbol{\theta}$，$\{\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}),i=1,\cdots,n\}$为iid，故根据大数定律，其样本均值将收敛到总体均值，即

$$Q_n(\boldsymbol{\theta})\equiv\frac{1}{n}\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})\xrightarrow{p}\mathrm{E}[\ln f(\boldsymbol{y};\boldsymbol{\theta})]\equiv Q(\boldsymbol{\theta}) \tag{6.24}$$

其中，$Q(\boldsymbol{\theta})\equiv\mathrm{E}[\ln f(\boldsymbol{y};\boldsymbol{\theta})]$只依赖于$\boldsymbol{\theta}$，因为期望算子$\mathrm{E}(\cdot)$已将样本数据$\boldsymbol{y}$积分掉。下面，我们将证明$Q(\boldsymbol{\theta})$在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处存在唯一的最大值，即对数似然函数的期望值在真实参数$\boldsymbol{\theta}_0$处达到最大值。为此，考虑随机变量$\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}$。由于对数函数$\ln(\cdot)$为严格凹函数，故根据概率统计中的詹森不等式(Jensen's inequality)可知，对于任意$\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$，

$$\mathrm{E}\left[\ln\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}\right]<\ln\left\{\mathrm{E}\left[\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}\right]\right\} \tag{6.25}$$

其中，如果$\boldsymbol{\theta}=\boldsymbol{\theta}_0$，则随机变量$\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}$退化为常数1，上式变为等式。不等式(6.25)右边的期望可以写为，

$$\mathrm{E}\left[\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}\right]=\int\left[\frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta}_0)}\right]f(\boldsymbol{y};\boldsymbol{\theta}_0)\,\mathrm{d}y=\int f(\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}y=1 \tag{6.26}$$

其中，由于$f(\boldsymbol{y};\boldsymbol{\theta})$为概率密度函数，故其积分为1。因此，不等式(6.25)的右边等于0，经整理可得

$$\mathrm{E}[\ln f(\boldsymbol{y};\boldsymbol{\theta})]-\mathrm{E}[\ln f(\boldsymbol{y};\boldsymbol{\theta}_0)]<0 \tag{6.27}$$

由于$Q(\boldsymbol{\theta})\equiv\mathrm{E}[\ln f(\boldsymbol{y};\boldsymbol{\theta})]$，故上式表明，对于任意$\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$，都有

$$Q(\boldsymbol{\theta})<Q(\boldsymbol{\theta}_0) \tag{6.28}$$

因此，$Q(\boldsymbol{\theta})$在$\boldsymbol{\theta}=\boldsymbol{\theta}_0$处达到唯一的最大值。综合以上结果，我们知道：

(i) $Q_n(\boldsymbol{\theta})$在$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$处达到唯一的最大值(MLE的定义)；

(ii) $Q(\boldsymbol{\theta})$在$\boldsymbol{\theta}_0$处达到唯一的最大值(方程(6.28))；

(iii) $Q_n(\boldsymbol{\theta})\xrightarrow{p}Q(\boldsymbol{\theta})$(方程(6.24))；

直观来看，由于函数$Q_n(\boldsymbol{\theta})$越来越接近于$Q(\boldsymbol{\theta})$，故$Q_n(\boldsymbol{\theta})$取最大值处$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$也越来越接近$Q(\boldsymbol{\theta})$取最大值处$\boldsymbol{\theta}_0$，参见图6.4。

MLE一致性的另一证明方法是通过“极值估计量”(extremum estimator)来证明(更为抽象，但基本思想与上述证明相同)，参见Amemiya (1985)，Newey and McFadden (1994)或Hayashi

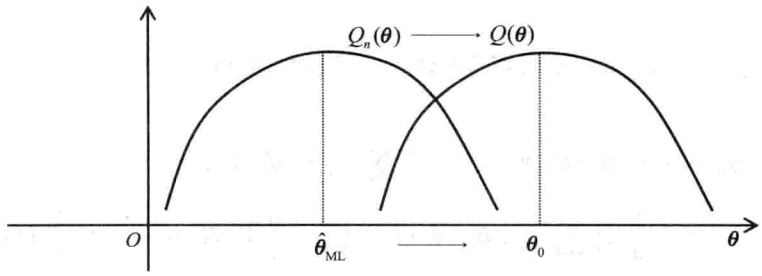


图 6.4 MLE 估计量的一致性示意图

(2000, Chapter 7)。最小二乘法、最大似然估计以及以后要介绍的 NLS 与 GMM 都是极值估计量的特例,即都是通过求解一个极值问题来获得估计量。

只要证明了(3)渐近正态性,即$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)\xrightarrow{d}N(\mathbf{0},n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1})$,也就证明了(2)渐近有效性,即 $\text{Avar}(\hat{\boldsymbol{\theta}}_{ML})=n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$。根据 MLE 的一阶条件

$$\mathbf{s}(\hat{\boldsymbol{\theta}}_{ML};\boldsymbol{y})=\frac{\partial \ln L(\hat{\boldsymbol{\theta}}_{ML};\boldsymbol{y})}{\partial \boldsymbol{\theta}}=\mathbf{0} \tag{6.29}$$

即 $\hat{\boldsymbol{\theta}}_{ML}$ 满足 MLE 的一阶条件。利用微分中值定理(Mean Value Theorem),将 $\boldsymbol{s}(\hat{\boldsymbol{\theta}}_{ML};\boldsymbol{y})$ 在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处进行泰勒展开可得

$$\boldsymbol{s}(\hat{\boldsymbol{\theta}}_{ML};\boldsymbol{y})=\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})+\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{y})(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)=\mathbf{0} \tag{6.30}$$

其中,$\boldsymbol{\theta}^*$ 介于 $\boldsymbol{\theta}_0$ 与 $\hat{\boldsymbol{\theta}}_{ML}$ 之间,即存在 $\lambda\in(0,1)$,使得 $\boldsymbol{\theta}^*=\lambda\boldsymbol{\theta}_0+(1-\lambda)\hat{\boldsymbol{\theta}}_{ML}$;而 $\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{y})\equiv\frac{\partial^2 \ln L(\boldsymbol{\theta}^*;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ 为在 $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ 处的黑赛矩阵。将方程(6.30)移项,然后两边同乘以$\sqrt{n}$可得

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)=[-\boldsymbol{H}(\boldsymbol{\theta}^*;\boldsymbol{y})]^{-1}\sqrt{n}\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y}) \tag{6.31}$$

根据 MLE 的一致性,$\underset{n\to\infty}{\text{plim}}\hat{\boldsymbol{\theta}}_{ML}=\boldsymbol{\theta}_0$;而$\boldsymbol{\theta}^*$ 夹在$\boldsymbol{\theta}_0$ 与$\hat{\boldsymbol{\theta}}_{ML}$之间,故$\underset{n\to\infty}{\text{plim}}\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$。由此可知

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)\xrightarrow{d}[-\boldsymbol{H}(\boldsymbol{\theta}_0;\boldsymbol{y})]^{-1}\sqrt{n}\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y}) \tag{6.32}$$

其中,$\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})=\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)$,即可将得分函数分解为各个观测值的贡献,参见方程(6.10)。类似地,黑赛矩阵也可分解为 $\boldsymbol{H}(\boldsymbol{\theta}_0;\boldsymbol{y})=\sum_{i=1}^{n}\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)$,参见方程(6.11)。因此,方程(6.32)可以写为

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)\xrightarrow{d}\left[-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\right]^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\right) \tag{6.33}$$

由于样本为 iid,故$\{\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\}_{i=1}^{n}$与$\{\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\}_{i=1}^{n}$都是 iid。因此,一方面,根据大数定律,$-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\xrightarrow{p}-\text{E}[\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]\equiv\boldsymbol{A}_0$。另一方面,由于 $\text{E}[\boldsymbol{s}_i(\boldsymbol{\theta};\boldsymbol{y}_i)]=\mathbf{0}$(得分函数的期望为$\mathbf{0}$),故根据中心极限定理,$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\right]\xrightarrow{d}N(\mathbf{0},\boldsymbol{B}_0)$,其中$\boldsymbol{B}_0\equiv\text{Var}(\boldsymbol{s}_i)$。由此可知

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta}_0)\xrightarrow{d}N(\mathbf{0},\boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}) \tag{6.34}$$

其中,渐近方差协方差矩阵为 $\boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}$,表现为夹心估计量(sandwich estimator)。下面分别求

$\boldsymbol{A}_0$ 与 $\boldsymbol{B}_0$。

由于 $\{\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\}_{i=1}^n$ 是 iid，故 $\mathrm{E}[\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathrm{E}[\boldsymbol{H}_j(\boldsymbol{\theta}_0;\boldsymbol{y}_j)]$，$\forall i,j$（iid 随机变量的期望值相同）。因此

$$\begin{aligned}\boldsymbol{A}_0 &\equiv -\mathrm{E}[\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=-\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]\\ &=-\frac{1}{n}\mathrm{E}\Big\{\sum_{i=1}^{n}[\boldsymbol{H}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]\Big\}=-\frac{1}{n}\mathrm{E}[\boldsymbol{H}(\boldsymbol{\theta}_0;\boldsymbol{y})]=\frac{1}{n}\boldsymbol{I}(\boldsymbol{\theta}_0)\end{aligned} \tag{6.35}$$

另一方面，$\{\boldsymbol{s}_i(\boldsymbol{\theta};\boldsymbol{y}_i)\}_{i=1}^n$ 也是 iid，故 $\mathrm{Var}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathrm{Var}[\boldsymbol{s}_j(\boldsymbol{\theta}_0;\boldsymbol{y}_j)]$，$\forall i,j$（iid 随机变量的方差相同）。由于得分函数的方差为信息矩阵，故

$$\begin{aligned}\boldsymbol{I}(\boldsymbol{\theta}_0)&=\mathrm{Var}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]=\mathrm{Var}\Big[\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\Big]\\ &=\sum_{i=1}^{n}\mathrm{Var}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=n\mathrm{Var}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]\end{aligned} \tag{6.36}$$

因此

$$\boldsymbol{B}_0=\mathrm{Var}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\frac{1}{n}\boldsymbol{I}(\boldsymbol{\theta}_0)=\boldsymbol{A}_0 \tag{6.37}$$

将 $\boldsymbol{A}_0$ 与 $\boldsymbol{B}_0$ 的表达式(6.37)代入方程(6.34)可知，

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}-\boldsymbol{\theta}_0)\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{A}_0^{-1})=N(\boldsymbol{0},n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}) \tag{6.38}$$

由于 $\boldsymbol{A}_0=\boldsymbol{B}_0$，故渐近方差的表达式简化为 $n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$。

除了良好的大样本性质，最大似然估计量还具有“不变性”(invariance)的优点。

定理(不变性) 如果将参数 $\boldsymbol{\theta}$“参数变换”(reparameterize)为 $\boldsymbol{\alpha}\equiv\boldsymbol{g}(\boldsymbol{\theta})$，则对 $\boldsymbol{\alpha}$ 的最大似然估计就是 $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}=\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$，其中 $\boldsymbol{g}(\cdot)$ 可以是多维函数，也不要求 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\theta}$ 有一一对应的函数关系，证明参见附录。利用最大似然估计的不变性，有时可以大大简化计算，比如，对 $(\mu^2+\sigma^2)$ 的最大似然估计就是 $(\hat{\mu}_{\mathrm{ML}}^2+\hat{\sigma}_{\mathrm{ML}}^2)$。然而，MLE 的不变性也从另一方面说明了 MLE 估计量较有可能不是无偏估计。比如，假设 $\mathrm{E}(\hat{\theta}_{\mathrm{MLE}})=\theta$（无偏估计），但对于一般的函数 $g(\cdot)$ 来说，$\mathrm{E}[g(\hat{\theta}_{\mathrm{MLE}})]\neq g[\mathrm{E}(\theta)]=g(\theta)$，故在参数变换后，$g(\hat{\theta}_{\mathrm{MLE}})$ 将不再是 $g(\theta)$ 的无偏估计。

虽然 MLE 的大样本性质很优越，但其缺点是必须事先假设随机变量的概率分布（通常假设为正态分布），而研究者有时对此并无把握。因此，当 MLE 所依赖的分布假设为正确时，MLE 是有效率的；但它不够稳健。相反，“广义矩估计”(Generalized Method of Moments，简记 GMM)通过总体矩条件来进行估计，并不依赖于任何概率分布的假设，故一般来说更为稳健（参见第 10 章）。

6.6 最大似然估计量的渐近协方差矩阵

在大样本下，最大似然估计量的渐近协方差矩阵为

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})=n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}=n\left\{-\mathrm{E}\left[\frac{\partial^2\ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right\}^{-1} \tag{6.39}$$

显然，这个表达式依赖于未知参数 $\boldsymbol{\theta}_0$。对于 MLE 的渐近协方差矩阵，文献中有以下三种估计方法。

(1) 期望值法。如果知道黑赛矩阵期望值的具体函数形式,则直接以 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ 替代 $\boldsymbol{\theta}_0$ 可得

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = n\left\{-\mathrm{E}\left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\mathrm{ML}};\boldsymbol{y})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right]\right\}^{-1} \tag{6.40}$$

然而,由于黑赛矩阵通常包含复杂的非线性函数,其期望值可能没有解析解,故此法很少用。

(2) 观测信息矩阵法。以 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ 替代 $\boldsymbol{\theta}_0$ 后,干脆将期望算子忽略掉,即 $\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = n\left[-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\mathrm{ML}};\boldsymbol{y})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right]^{-1}$。这种方法在 Stata 中被称为"观测信息矩阵"(Observed Information Matrix,简记 OIM)法,即直接使用观测到的信息矩阵。其缺点是,二阶偏导数可能不容易计算。OIM 法通常是 Stata 的默认方法。

(3) 梯度向量外积或 BHHH 法。利用信息矩阵等式,用 $\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'$ 来估计 $\boldsymbol{I}(\boldsymbol{\theta}_0)$,即 $\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = n\left(\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'\right)^{-1}$,其中 $\hat{\boldsymbol{s}}_i \equiv \frac{\partial \ln f(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_{\mathrm{ML}})}{\partial\boldsymbol{\theta}}$ 为第 i 个观测值对得分函数的贡献之估计值,证明参见附录。该方法称为"梯度向量外积"(Outer Product of Gradients,简记 OPG)或 BHHH 法[①]。它的优点是,只需要计算一阶偏导数,十分简便。BHHH 法的另一优点是,该协方差估计量总是非负定的(nonnegative definite),而 OIM 法的协方差估计量则没有此保证。

以上三种估计渐近协方差矩阵的方法在大样本下是渐近等价的(asymptotically equivalent)。然而,在有限样本中,这三种方法的计算结果可能差别较大,甚至可能导致统计推断作出不同的结论,参见 Greene (2012, p. 522)的一个例子。

另外,以上三种计算渐近方差的方法都建立在似然函数正确的前提上。如果似然函数不正确(比如,真实分布并非正态分布,却误设为正态分布),则这三种方法都失效。此时,可以考虑使用稳健标准误,详见本章第 8 节。

6.7 三类渐近等价的统计检验

在计量经济学中,经常使用以下三类在大样本下渐近等价的统计检验。对于线性回归模型,检验原假设 H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$[②],其中 $\boldsymbol{\beta}_{K\times 1}$ 为未知参数,$\boldsymbol{\beta}_0$ 已知,共有 K 个约束。

(1) 沃尔德检验(Wald Test):通过研究 $\boldsymbol{\beta}$ 的无约束估计量 $\hat{\boldsymbol{\beta}}_U$ 与 $\boldsymbol{\beta}_0$ 的距离来进行检验。其基本思想是,如果 H_0 正确,则 $(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0)$ 的绝对值不应该很大。沃尔德统计量为

$$W \equiv (\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0)'[\mathrm{Var}(\hat{\boldsymbol{\beta}}_U)]^{-1}(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(K) \tag{6.41}$$

其中,K 为约束条件的个数(即解释变量的个数),其证明类似于第 5 章对于线性假设"H_0: $\boldsymbol{R\beta} = \boldsymbol{r}$"的大样本检验之证明。第 5 章所介绍的 t 检验、F 检验都是 Wald 检验。

(2) 似然比检验(Likelihood Ratio Test,简记 LR):通常来说,无约束的似然函数最大值

① "BHHH 法"的命名来自 Berndt, Hall, Hall and Hausman (1974)。

② 对于一般的线性假设"H_0: $\boldsymbol{R\beta} = \boldsymbol{r}$"或非线性假设"$H_0$: $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{0}$",也有类似的结果。在此,为了集中阐述这三类检验的思想,仅考虑最简单但也最常见的情形"H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$"。

$\ln L(\hat{\boldsymbol{\beta}}_U)$ 比有约束的似然函数最大值 $\ln L(\hat{\boldsymbol{\beta}}_R)$ 更大，因为在无约束条件下的参数空间 $\boldsymbol{\Theta}$ 比有约束条件下（即 H_0 成立时）参数的取值范围更大，参见图 6.5。

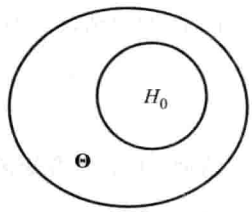


图 6.5 无约束与有约束的参数空间

LR 检验的基本思想是，如果 H_0 正确，则 $\ln L(\hat{\boldsymbol{\beta}}_U) - \ln L(\hat{\boldsymbol{\beta}}_R)$ 不应该很大。在这个简单例子中，有约束的估计量 $\hat{\boldsymbol{\beta}}_R = \boldsymbol{\beta}_0$。LR 统计量为

$$\mathrm{LR} \equiv -2\ln\left[\frac{L(\hat{\boldsymbol{\beta}}_R)}{L(\hat{\boldsymbol{\beta}}_U)}\right] = 2[\ln L(\hat{\boldsymbol{\beta}}_U) - \ln L(\hat{\boldsymbol{\beta}}_R)] \xrightarrow{d} \chi^2(K) \tag{6.42}$$

证明的基本方法是，将对数似然函数作二阶泰勒展开（根据 MLE 一阶条件，此泰勒展开的一阶项为 0），参见 Amemiya (1985, p. 142)。第 3 章介绍的 F 统计量的另一表达式 $F = \dfrac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^{*} - \boldsymbol{e}'\boldsymbol{e})/(K-1)}{\boldsymbol{e}'\boldsymbol{e}/(n-K)}$，就可以看成是依据似然比原理而设计的①。

（3）拉格朗日乘子检验（Lagrange Multiplier Test，简记 LM）：考虑有约束条件的对数似然函数最大化问题：

$$\max_{\tilde{\boldsymbol{\beta}}} \ln L(\tilde{\boldsymbol{\beta}}) \quad s.t.\ \boldsymbol{\beta} = \boldsymbol{\beta}_0 \tag{6.43}$$

引入拉格朗日乘子函数，

$$\max_{\tilde{\boldsymbol{\beta}},\lambda} \ln L(\tilde{\boldsymbol{\beta}}) - \lambda'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{6.44}$$

其中，$\boldsymbol{\lambda}$ 为拉格朗日乘子向量。如果 $\hat{\boldsymbol{\lambda}} \approx \mathbf{0}$，则说明此约束条件不“紧”（tight）或不是“硬约束”（binding constraint），加上这个约束条件并不会使似然函数的最大值下降很多，即原假设 H_0 很可能成立。根据一阶条件（对 $\tilde{\boldsymbol{\beta}}$ 求导）可知，$\hat{\boldsymbol{\lambda}} = \dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}}$，即最优的拉格朗日乘子向量等于对数似然函数在 $\hat{\boldsymbol{\beta}}_R$ 处的梯度向量。LM 统计量为

$$\mathrm{LM} \equiv \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}}\right)' [\boldsymbol{I}(\hat{\boldsymbol{\beta}}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}}\right) \xrightarrow{d} \chi^2(K) \tag{6.45}$$

其中，$\boldsymbol{I}(\hat{\boldsymbol{\beta}}_R)$ 为信息矩阵在 $\hat{\boldsymbol{\beta}}_R$ 处的取值。由于 $\dfrac{\partial \ln L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 被称为“得分函数”（score function），故这个检验也称为“得分检验”（score test）；而 $\boldsymbol{I}(\hat{\boldsymbol{\beta}}_R)$ 正是得分函数的协方差矩阵。另一直观理解是，由于在无约束估计量 $\hat{\boldsymbol{\beta}}_U$ 处，$\dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}_U)}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$，如果原假设 H_0 成立，则在约束估计量 $\hat{\boldsymbol{\beta}}_R$ 处，这个梯度向量也应该接近于 $\mathbf{0}$，即 $\dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \approx \mathbf{0}$，而 LM 统计量反映的就是此接近程度。

① 可以证明，似然函数是残差平方和的单调减函数。

可以把这三类统计检验的思想表现在同一张图上，参见图 6.6。

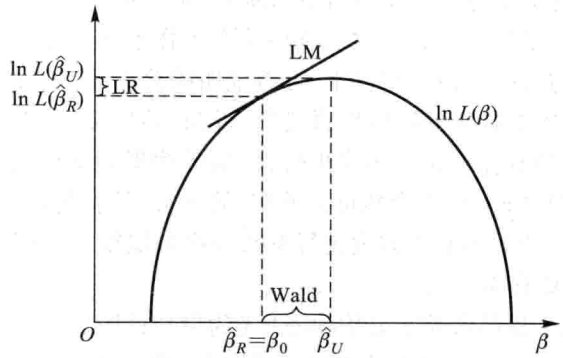


图 6.6 三类渐近等价的统计检验

LM 统计量的计算（选读）

由于有时不方便直接计算 LM 统计量，在实践上常进行以下辅助回归（auxiliary regression），并计算其 nR_{uc}^2，其中 R_{uc}^2 为“非中心 R^2”（参见第 3 章）。显然，$\dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \boldsymbol{s}_i$ 可以用 $\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i$ 来估计，其中 $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i(\hat{\boldsymbol{\beta}}_R)$ 为第 i 个观测值对得分函数的贡献在 $\hat{\boldsymbol{\beta}}_R$ 处的取值。另一方面，根据 BHHH 法，$\boldsymbol{I}(\hat{\boldsymbol{\beta}}_R)$ 可以用 $\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i'$ 来估计。因此，LM 统计量可以写为

$$\mathrm{LM} = \left(\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i'\right)\left(\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i'\right)^{-1}\left(\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i\right) \tag{6.46}$$

定义 $n \times K$ 矩阵 $\boldsymbol{S} \equiv \begin{pmatrix} \hat{\boldsymbol{s}}_1' \\ \hat{\boldsymbol{s}}_2' \\ \vdots \\ \hat{\boldsymbol{s}}_n' \end{pmatrix} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{1K} \\ \hat{s}_{21} & \hat{s}_{22} & \cdots & \hat{s}_{2K} \\ \vdots & \vdots & & \vdots \\ \hat{s}_{n1} & \hat{s}_{n2} & \cdots & \hat{s}_{nK} \end{pmatrix}$，常数向量 $\mathbf{1} \equiv \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$，则

$$\sum_{i=1}^{n} \hat{\boldsymbol{s}}_i' = (1 \quad 1 \quad \cdots \quad 1)\begin{pmatrix} \hat{\boldsymbol{s}}_1' \\ \hat{\boldsymbol{s}}_2' \\ \vdots \\ \hat{\boldsymbol{s}}_n' \end{pmatrix} = \mathbf{1}'\boldsymbol{S},\ \sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i' = (\hat{\boldsymbol{s}}_1 \quad \hat{\boldsymbol{s}}_2 \quad \cdots \quad \hat{\boldsymbol{s}}_n)\begin{pmatrix} \hat{\boldsymbol{s}}_1' \\ \hat{\boldsymbol{s}}_2' \\ \vdots \\ \hat{\boldsymbol{s}}_n' \end{pmatrix} = \boldsymbol{S}'\boldsymbol{S}$$，故

$$\mathrm{LM} = \mathbf{1}'\boldsymbol{S}\,(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1} \tag{6.47}$$

为此，可以进行如下辅助回归[①]：

$$1 \xrightarrow{\mathrm{OLS}} \hat{\boldsymbol{s}}_i'\boldsymbol{\gamma} + v_i \tag{6.48}$$

其中，被解释变量为常数向量 $\mathbf{1}$（相当于线性回归中的 $\boldsymbol{y}$），而解释变量为得分函数的 K 个分量，没有常数项。这个辅助回归的数据矩阵（data matrix）正是 $\boldsymbol{S}$（相当于线性回归中的 $\boldsymbol{X}$）。根据第 3 章，有

$$R_{uc}^2 = \frac{\boldsymbol{y}'\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{\mathbf{1}'\boldsymbol{S}\,(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\mathbf{1}}{\mathbf{1}'\mathbf{1}} = \frac{\mathrm{LM}}{n} \tag{6.49}$$

① 也称“人工回归”，即只为计算统计量而进行的回归，辅助回归本身通常没有什么经济意义。

故 $LM = nR_{uc}^2$。

总之，Wald 检验仅利用无约束估计的信息，LM 检验仅利用有约束估计的信息，而 LR 检验同时利用无约束与有约束估计的信息。这三类检验在大样本下是渐近等价的，但在小样本下性质则不同。在正态分布与线性假设的情况下，可以证明检验统计量 $Wald \geqslant LR \geqslant LM$，即给定显著性水平 α，Wald 检验比 LR 检验更可能拒绝原假设，而 LR 检验又比 LM 检验更可能拒绝原假设。

另外，如果不对模型的具体概率分布作出假设，则无法得到似然函数。此时，一般就无法使用 LR 检验与 LM 检验；但 Wald 检验依然可以使用，故 Wald 检验的使用范围最广。Wald 检验的缺点是，它不具有不变性，即如果对原假设进行参数变换可能得到不同的 Wald 统计量取值①；而 LR 检验及某些 LM 检验具有不变性。

在实际应用中，究竟采取哪种检验常取决于“无约束估计”与“有约束估计”。如果无约束估计更方便，则常使用 Wald 检验；如果有约束估计更方便，则常使用 LM 检验（参见第 7、8 章对异方差、自相关的检验）。

6.8 准最大似然估计法

如果随机变量不服从正态分布，但却使用了以正态分布为前提的最大似然估计法，该估计量还可能是一致估计量吗？仍然有可能！例如，对于线性模型，MLE 估计量等价于 OLS 估计量，而 OLS 估计量的一致性并不依赖于正态分布的假定。

定义 使用不正确的似然函数而得到的最大似然估计，称为“**准最大似然估计**”（Quasi MLE，简记 QMLE）或“伪最大似然估计”（Pseudo MLE）②。

之所以在某些情况下可以“歪打正着”地得到一致的准最大似然估计，是因为 MLE 也可以被视为 GMM，而后者一般不需要对随机变量的具体分布作假定（参见第 10 章）。这也说明，虽然 MLE 要求随机变量服从正态分布，但这个假定可能并不那么强。

如果 QMLE 估计量满足以下两个条件，则依然是一致估计量。

（i）模型设定的概率密度函数属于“线性指数分布族”（linear exponential family），即概率密度函数可以写为 $f(y;\boldsymbol{\theta}) = \dfrac{p(y)\mathrm{e}^{r(\boldsymbol{\theta})}}{q(\boldsymbol{\theta})}$ 的形式。线性指数分布族包括正态分布，二项分布（第 11 章 Probit 与 Logit 回归），泊松分布（第 13 章泊松回归），负二项分布（第 13 章负二项回归），Γ 分布（第 30 章久期分析，含指数分布），以及逆高斯分布（inverse Gaussian）等。

（ii）条件期望 $\mathrm{E}(y|\boldsymbol{x})$ 的函数形式设定正确。

然而，在更一般的情况下，QMLE 并非一致估计，譬如第 14 章的 Tobit 回归。即使 QMLE 碰巧为一致估计，但 $\hat{\boldsymbol{\theta}}_{\mathrm{QML}}$ 的渐近方差也通常不再是 $n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$。

假设正确的对数似然函数为 $\ln L(\boldsymbol{\theta};\boldsymbol{y})$，而被误设为 $\ln L^*(\boldsymbol{\theta};\boldsymbol{y})$，称为“准对数似然函数”（pseudo log likelihood function）。最大化 $\ln L^*(\boldsymbol{\theta};\boldsymbol{y})$ 的结果即 QMLE 估计量

$$\hat{\boldsymbol{\theta}}_{\mathrm{QML}} \equiv \arg\max \ln L^*(\boldsymbol{\theta};\boldsymbol{y}) \tag{6.50}$$

---

① 比如，选择检验“$H_0: \theta_1 = \theta_2$”还是检验“$H_0: (\theta_1/\theta_2) - 1 = 0$”可能导致不同的结果，尽管这两个原假设是完全等价的。为了检查其稳健性（as a robustness check），可以对几种形式不同但等价的原假设均进行 Wald 检验，并对比其结果。

② 文献中对于 QMLE 的另一种定义是，只有当它是一致估计量时，才称为 QMLE 估计量。

遵循类似于 MLE 一致性的证明步骤，可以证明 $\hat{\boldsymbol{\theta}}_{\text{QML}} \xrightarrow{p} \boldsymbol{\theta}^*$，其中 $\boldsymbol{\theta}^*$ 称为“准真实值”（pseudo-true value），但通常 $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$。对于 $\hat{\boldsymbol{\theta}}_{\text{QML}}$ 的大样本分布，使用类似于 MLE 的推导可证明

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A}_0^{*-1}\boldsymbol{B}_0^{*}\boldsymbol{A}_0^{*-1}) \tag{6.51}$$

其中，$\boldsymbol{A}_0^*$ 与 $\boldsymbol{B}_0^*$ 的表达式类似于 $\boldsymbol{A}_0$ 与 $\boldsymbol{B}_0$，但在 $\boldsymbol{\theta}^*$ 处取值，且使用正确的概率密度函数来求概率极限。另外，由于 $\ln L^*(\boldsymbol{\theta}; \boldsymbol{y})$ 并非真正的对数似然函数，信息矩阵等式也不再成立，故一般 $\boldsymbol{A}_0^* \neq \boldsymbol{B}_0^*$，因此公式(6.51)中的夹心估计量 $\boldsymbol{A}_0^{*-1}\boldsymbol{B}_0^{*}\boldsymbol{A}_0^{*-1}$ 无法进一步简化。基于 $\boldsymbol{A}_0^{*-1}\boldsymbol{B}_0^{*}\boldsymbol{A}_0^{*-1}$ 的标准误差被称为“胡贝尔 - 怀特稳健标准误”（Huber-White robust standard errors），最早由 Huber (1967) 与 White (1982) 提出。

胡贝尔 - 怀特稳健标准误也简称为“稳健标准误”，因为它与第 5 章介绍的异方差稳健标准误是一致的。假设用 MLE 来估计古典线性回归模型，但真实模型其实存在异方差，而我们在同方差的错误设定下来求 MLE 估计量。此时，得到的就是 QMLE 估计量幸运的是，此 $\hat{\boldsymbol{\beta}}_{\text{QML}}$ 依然是真实参数 $\boldsymbol{\beta}$ 的一致估计，而胡贝尔 - 怀特稳健标准误就是异方差稳健的标准误。。

在使用 MLE 估计非线性模型时，如果对模型的正确设定没有把握，而且 QMLE 估计量依然是一致估计量，则应考虑使用（胡贝尔 - 怀特）稳健标准误。在 Stata 中，仍以选择项“r”或“vce(robust)”来实现。反之，如果对于模型设定很有信心，则直接使用 OIM 或 OPG 法来估计渐近方差会更有效率，没有必要使用稳健标准误。在实践中，可同时估计这两种标准误，如果二者相差不大，则也验证了模型设定的正确性；反之，如果二者相差很大，则应考察模型设定的准确性。

需要注意的是，当 QMLE 估计量不一致时，即使采用（胡贝尔 - 怀特）稳健标准误也无济于事。此时，$\hat{\boldsymbol{\theta}}_{\text{QML}} \xrightarrow{p} \boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$，故首先应担心估计量的一致性。稳健标准误只是帮助我们更精确地估计一个错误的“准真实参数”$\boldsymbol{\theta}^*$，而且通常不知道 $\boldsymbol{\theta}^*$ 的经济意义。换言之，在这种情况下，（胡贝尔 - 怀特）稳健标准误只是一致地估计了一个不一致估计量的方差（a consistent estimator of the variance of an inconsistent estimator）。

另外，无论 OIM、OPG 法，还是（胡贝尔 - 怀特）稳健标准误都假设样本数据为 iid。如果样本数据可分为若干组，而同一组内的观测值存在自相关，则应使用“聚类稳健标准误”（cluster-robust standard errors），在 Stata 中由选择项“vce(cluster *clustvar*)”来实现，其中“*clustvar*”为聚类变量（详见第 8 章 8.4 节）。

总之，对于线性回归模型，通常建议总是使用稳健标准误。而对于非线性模型，可以分为以下四种情况来考虑。

（i）如果对模型设定较有信心或模型拟合得较好，则可以不用稳健标准误。

（ii）如果对模型设定缺乏信心，而且 QMLE 为一致估计，则应使用稳健标准误。

（iii）如果对模型设定缺乏信心，但 QMLE 也不一致，则应首先担心 QMLE 估计量的一致性，仅仅使用稳健标准误进行校正是无济于事的。

（iv）对于聚类样本，应使用聚类稳健的标准误。

6.9 对正态分布假设的检验

对于线性回归模型，如果扰动项不服从正态分布，则无法使用小样本 OLS 进行统计推断；然而，OLS 估计量依然是一致估计的，而且服从渐近正态分布，故可以用大样本 OLS 进行统计推断。在这种情形下，检验扰动项是否服从正态分布似乎意义不大（但如果接受扰动项为正态，则可以使用小样本 OLS 进行统计推断）。

然而，对非线性模型使用 MLE 时，由于正态分布假定是推导 MLE 的前提，故检验扰动项是否服从正态分布就可能比较重要。尽管在某些情况下，“准最大似然估计”也是一致的，但毕竟不如真正的 MLE 有效率。

为了考察扰动项是否为正态，最直观的方法是画图。可以把残差画成直方图（histogram），并与正态分布的密度函数比较。但直方图是不连续的。为了得到对密度函数的光滑估计，可以使用“核密度估计法”（kernel density estimation），参见第 27 章，并与正态密度相比较。

另外一种画图方法是，将正态分布的分位数（quantiles）与残差的分位数画成散点图（scatter plot）。如果残差来自正态分布，则该图上的散点应该集中在 45°线附近。称这种图为“分位数 - 分位数图”（Quantile-Quantile plot，简记 QQ plot）。

通过画图可以初步判断扰动项是否为正态，但最终的结论仍要通过严格的统计检验。常用的检验方法利用了正态分布的偏度与峰度性质。回顾第 2 章，随机变量 X 的偏度为 $\mathrm{E}[(X-\mu)/\sigma]^3$，峰度为 $\mathrm{E}[(X-\mu)/\sigma]^4$，而超额峰度（excess kurtosis）为 $\mathrm{E}[(X-\mu)/\sigma]^4-3$。因此，对于残差 $\{e_1, \cdots, e_n\}$，其偏度与超额峰度的样本估计值分别为 $\frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n} e_i^3$ 与 $\left(\frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n} e_i^4\right)-3$（注意 $\bar{e}=0$）。在扰动项服从正态分布的原假设下，这两个统计量服从正态分布。较常用的“雅克 - 贝拉检验”（Jarque and Bera，1987，简记 JB）使用了它们的平方之加权平均作为检验统计量：

$$\mathrm{JB}\equiv\frac{n}{6}\left[\left(\frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n} e_i^3\right)^2+\frac{1}{4}\left(\frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n} e_i^4-3\right)^2\right]\xrightarrow{d}\chi^2(2) \tag{6.52}$$

由于 JB 统计量本质上是两个正态分布的平方和，故其自由度为 2。JB 检验虽然常用，但其统计量的收敛速度较慢，对样本容量的要求较高。为此，在 Stata 的官方程序中提供了 D'Agostino et al（1990）的改进方法，基于偏度与峰度设计了更复杂的检验统计量（参见“help sktest”。另外，Stata 也提供检验正态性的两种非参数方法（参见“help swilk”）。

如果发现某变量不服从正态分布，有时可以通过取对数，使之变得更接近于正态分布。

6.10 最大似然估计法的 Stata 命令及实例

1. 最大似然估计

Stata 提供了一个“ml”的命令，可自行定义似然函数来执行最大似然估计，详见“help ml”。但通常并不需要这样做，因为对于多数需要使用最大似然估计的情形，Stata 一般都已给出

了专门的命令，比如，第 11 ~ 13 章的离散被解释变量模型。

2. LR 检验

LR 检验可通过 Stata 命令 `lrtest` 来实现。在第 15 章，将使用命令 `lrtest` 对面板数据中的异方差进行检验。

3. 正态分布检验

以 Stata 系统自身的数据集 auto. dta 为例：

```
. sysuse auto  (调用系统数据集 auto. dta)
(1978 Automobile Data)
. hist mpg,normal  (画变量 mpg 的直方图，并与正态密度比较，结果如图 6.7 所示)
(bin =8,start =12,width =3.625)
```

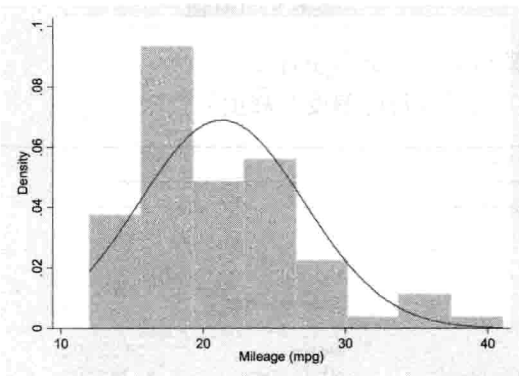


图 6.7 直方图与正态密度

直方图显示，变量 mpg 的分布与正态分布有一定差距。

```
. kdensity mpg,normal lpattern(" - ")  (画变量 mpg 的核密度图，并与正态密度比较，选择项“lpattern(" - ")”表示用虚线来画核密度曲线，结果如图 6.8 所示)
```

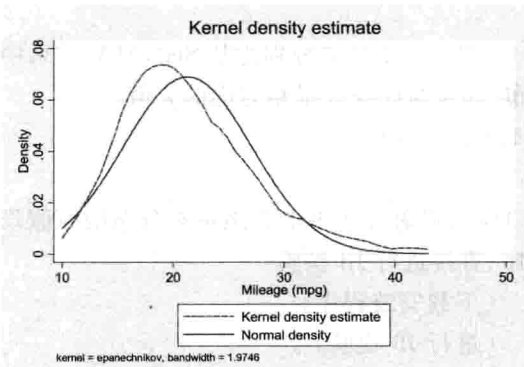


图 6.8 核密度与正态密度

```
. qnorm mpg  (画变量 mpg 的 QQ 图，结果如图 6.9 所示)
```

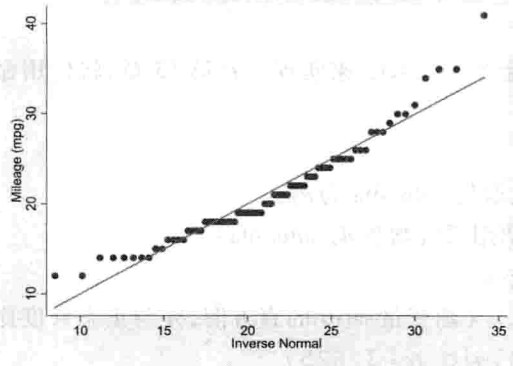


图 6.9 QQ 图

对于 JB 检验,可以比较容易地手工计算其统计量:

```
. su mpg,detail  (显示变量的偏度与峰度)
```

```
                    Mileage (mpg)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           12              12
 5%           14              12
10%           14              14      Obs                  74
25%           18              14      Sum of Wgt.          74

50%           20                      Mean            21.2973
                      Largest         Std. Dev.       5.785503
75%           25              34
90%           29              35      Variance        33.47205
95%           34              35      Skewness        .9487176
99%           41              41      Kurtosis        3.975005
```

根据以上显示的偏度与峰度,可以计算 JB 统计量为

```
. di (r(N)/6) * ((r(skewness)^2) + [(1/4) * (r(kurtosis) - 3)^2])
14.031924
```

其中,r(N)、r(skewness)、r(kurtosis)分别为从 Stata 计算结果中提取的样本容量、偏度、峰度的取值。根据 $\chi^2(2)$ 分布,计算与此统计量相对应的 p 值。

```
.di chi2tail(2,14.031924)
.00089744
```

该 p 值小于 1%,故可以在 1% 的显著性水平上拒绝正态分布的原假设。

也可以下载非官方程序,直接进行 JB 检验:

```
.ssc install jb6  (下载安装程序)
.jb6 mpg          (进行 JB 检验)
```

```
Jarque-Bera normality test:  14.03 Chi(2)  9.0e-04
Jarque-Bera test for Ho: normality: (mpg)
```

该程序的计算结果与手工计算一致。

```
.sktest mpg  (进行 D'Agostino 检验)
```

```
                    Skewness/Kurtosis tests for Normality
                                                              joint
Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)   Prob>chi2
---------+-----------------------------------------------------------
     mpg |    74       0.0015         0.0804        10.95       0.0042
```

D'Agostino 检验的 p 值为 0.0042，故在 1% 的显著水平上强烈拒绝正态分布的原假设。

```
. swilk mpg   (进行非参数 Shapiro-Wilk 检验)
```

```
                    Shapiro-Wilk W test for normal data
Variable |   Obs        W           V          z       Prob>z
---------+-----------------------------------------------------
     mpg |    74    0.94821       3.335      2.627     0.00430
```

Shapiro-Wilk 检验的 p 值与 D'Agostino 检验非常接近，也强烈拒绝原假设。

```
. sfrancia mpg   (进行非参数 Shapiro-Francia 检验)
```

```
                    Shapiro-Francia W' test for normal data
Variable |   Obs        W'          V'         z       Prob>z
---------+-----------------------------------------------------
     mpg |    74    0.94872       3.650      2.510     0.00604
```

Shapiro-Francia 检验同样在 1% 的显著性水平上拒绝正态假设。下面考虑对变量 mpg 取自然对数，使之更接近于正态分布：

```
. gen lnmpg = log(mpg)                          (取对数)
. kdensity lnmpg,normal lpattern(dash)   (画核密度图，如图 6.10)
```

其中，选择项“lpattern(dash)”的效果等同于“lpattern("-")”。

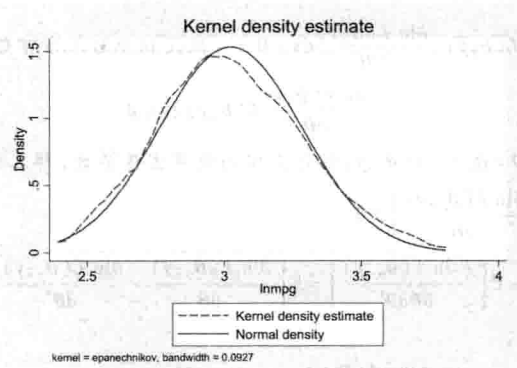


图 6.10 取对数后的核密度与正态密度

核密度图显示，取对数后的变量已较接近于正态分布。

```
. jb6 lnmpg   (进行 JB 检验)
```

```
Jarque-Bera normality test:   .8632 Chi(2)   .6495
Jarque-Bera test for Ho: normality: (lnmpg)
```

```
. sktest lnmpg   (进行 D'Agostino 检验)
```

Skewness/Kurtosis tests for Normality				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
lnmpg	74	0.3586	0.9446	0.87	0.6474

JB 检验与 D'Agostino 检验均接受取对数后的变量为正态分布。

习 题

6.1 假设离散型随机变量 y 服从泊松分布，其概率分布律为 $P(y=k)=\frac{\lambda^k e^{-\lambda}}{k!}$，$k=0,1,2,\cdots$。对于随机样本 $\{y_1,y_2,\cdots,y_n\}$，求 λ 的 MLE 估计量。

6.2 假设随机变量 y 服从在 $[0,\theta]$ 区间的均匀分布。对于随机样本 $\{y_1,y_2,\cdots,y_n\}$，求 θ 的 MLE 估计量。

6.3 假设样本 $\{y_1,y_2,\cdots,y_n\}$ 为独立同分布的，且 $y_i\sim N(\mu,1)$。分别用 Wald 检验、LR 检验、LM 检验来检验原假设 $H_0:\mu=\mu_0$。证明这三个检验统计量相等。

附 录

A6.1 如果似然函数正确，则得分函数在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处的期望值为 0，即 $E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right]=\mathbf{0}$。

证明：因为似然函数 $L(\boldsymbol{\theta};\boldsymbol{y})$ 是概率密度函数，故似然函数的积分为 1，即

$$\int L(\boldsymbol{\theta};\boldsymbol{y})\,d\boldsymbol{y}=1$$

$$\int \exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\,d\boldsymbol{y}=1 \quad (\text{引入对数似然函数})$$

方程两边对 $\boldsymbol{\theta}$ 求导可得

$$\int \exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}d\boldsymbol{y}=\mathbf{0} \quad (\text{假设积分与求导可交换次序})$$

$$\int \frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}L(\boldsymbol{\theta};\boldsymbol{y})\,d\boldsymbol{y}=\mathbf{0}$$

由于似然函数正确，故在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处，$L(\boldsymbol{\theta}_0;\boldsymbol{y})$ 就是真实的概率密度函数（样本数据 $\boldsymbol{y}$ 来自于此总体）。因此，在上式中令 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 即可得，$E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right]=\mathbf{0}$。

A6.2 证明在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处，$-E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]=E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}}\cdot\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}'}\right]$。

证明：根据附录 A6.1 可知，

$$\int \frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\,d\boldsymbol{y}=\mathbf{0}$$

该方程两边对 $\boldsymbol{\theta}'$ 求导可得（假设积分与求导可交换次序）

$$\int\left\{\frac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]+\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}'}\right\}d\boldsymbol{y}=\mathbf{0}$$

移项可得

$$-\int\left\{\frac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\right\}d\boldsymbol{y}=\int\left\{\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\exp[\ln L(\boldsymbol{\theta};\boldsymbol{y})]\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}'}\right\}d\boldsymbol{y}$$

直接使用似然函数（而不用对数似然函数），可得

$$-\int\left\{\frac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}L(\boldsymbol{\theta};\boldsymbol{y})\right\}d\boldsymbol{y}=\int\left\{\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\cdot\frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}'}L(\boldsymbol{\theta};\boldsymbol{y})\right\}d\boldsymbol{y}$$

由于似然函数正确，故在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ 处，似然函数就是联合密度函数。因此，在上式中令 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$，可得

$$-\mathrm{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]=\mathrm{E}\left[\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}'}\right]$$

A6.3　证明“克莱默－劳下限”(Cramer-Rao Lower Bound)。

证明：假设 $\hat{\boldsymbol{\theta}}$ 是对真实参数 $\boldsymbol{\theta}_0$ 的任意无偏估计，则在一定的正则条件下，$\mathrm{Var}(\hat{\boldsymbol{\theta}})\geqslant[I(\boldsymbol{\theta}_0)]^{-1}$。为简单起见，只证明一维情形，多维情形可类似证明。

由于 $\hat{\theta}(\boldsymbol{y})$ 是 θ_0 的无偏估计(其中 $\boldsymbol{y}$ 为样本数据)，故

$$\theta_0=\mathrm{E}[\hat{\theta}(\boldsymbol{y})]=\int\hat{\theta}(\boldsymbol{y})L(\theta_0;\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\quad(\text{随机变量函数 }\hat{\theta}(\boldsymbol{y})\text{的期望定义})$$

将上式两边同时对 θ_0 求导可得，

$$\begin{aligned}
1&=\int\hat{\theta}(\boldsymbol{y})\frac{\partial L(\theta_0;\boldsymbol{y})}{\partial\theta}\mathrm{d}\boldsymbol{y}=\int\hat{\theta}(\boldsymbol{y})\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\cdot L(\theta_0;\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\
&=\mathrm{E}\left[\hat{\theta}(\boldsymbol{y})\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right]\qquad(\text{期望的定义})\\
&=\mathrm{Cov}\left[\hat{\theta}(\boldsymbol{y}),\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right]\qquad\left(\text{根据附录 A6.1},\mathrm{E}\left[\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right]=0\right)\\
&\leqslant\mathrm{Var}[\hat{\theta}(\boldsymbol{y})]\cdot\mathrm{Var}\left[\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right]\qquad(\text{相关系数小于或等于 1})\\
&=\mathrm{Var}[\hat{\theta}(\boldsymbol{y})]\cdot\mathrm{E}\left[\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right]^2\qquad(\text{根据信息矩阵等式})
\end{aligned}$$

因此，$\mathrm{Var}[\hat{\theta}(\boldsymbol{y})]\geqslant\left[\mathrm{E}\left(\frac{\partial \ln L(\theta_0;\boldsymbol{y})}{\partial\theta}\right)^2\right]^{-1}=I(\theta_0)^{-1}$。

A6.4　证明古典线性回归模型的信息矩阵。

证明：对数似然函数为 $\ln L(\boldsymbol{\beta},\sigma^2)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$。为了求导方便，记 $\gamma\equiv\sigma^2$，则 $\ln L(\boldsymbol{\beta},\gamma)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\gamma-\frac{1}{2\gamma}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$。因此，一阶导数为

$$\frac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}}=\frac{1}{\gamma}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X\beta})$$

$$\frac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\gamma}=-\frac{n}{2\gamma}+\frac{1}{2\gamma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$$

二阶导数为

$$\frac{\partial^2 \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=-\frac{1}{\gamma}\boldsymbol{X}'\boldsymbol{X}$$

$$\frac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}\partial\gamma}=-\frac{1}{\gamma^2}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X\beta})=-\frac{1}{\gamma^2}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$\frac{\partial^2 \ln L(\boldsymbol{\beta},\gamma)}{\partial\gamma^2}=\frac{n}{2\gamma^2}-\frac{1}{\gamma^3}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})=\frac{n}{2\gamma^2}-\frac{1}{\gamma^3}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

根据严格外生性 $\mathrm{E}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=0$ 可知，$\mathrm{E}(\boldsymbol{X}'\boldsymbol{\varepsilon})=0$，故 $\mathrm{E}\left(\frac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}\partial\gamma}\right)=-\frac{1}{\gamma^2}\mathrm{E}(\boldsymbol{X}'\boldsymbol{\varepsilon})=0$。另外，$\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})=\mathrm{E}\left(\sum_{i=1}^{n}\varepsilon_i^2\right)=n\sigma^2=n\gamma$，故 $\mathrm{E}\left(\frac{\partial^2 \ln L(\boldsymbol{\beta},\gamma)}{\partial\gamma^2}\right)=\frac{n}{2\gamma^2}-\frac{1}{\gamma^3}\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})=\frac{n}{2\gamma^2}-\frac{n}{\gamma^2}=-\frac{n}{2\gamma^2}$。将 $\gamma\equiv\sigma^2$ 代回表达式可得，

$$I(\boldsymbol{\theta})\equiv-\mathrm{E}\left(\frac{\partial^2 \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)=-\mathrm{E}\begin{pmatrix}\dfrac{\partial^2 \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \dfrac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\boldsymbol{\beta}\partial\gamma}\\ \dfrac{\partial \ln L(\boldsymbol{\beta},\gamma)}{\partial\gamma\partial\boldsymbol{\beta}} & \dfrac{\partial^2 \ln L(\boldsymbol{\beta},\gamma)}{\partial\gamma^2}\end{pmatrix}=\begin{pmatrix}\dfrac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{0}\\ \boldsymbol{0} & \dfrac{n}{2\sigma^4}\end{pmatrix}$$

因此，信息矩阵的逆矩阵为

$$I(\boldsymbol{\theta})^{-1}=\begin{pmatrix}\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & \mathbf{0}\\ \mathbf{0} & \frac{n}{2\sigma^4}\end{pmatrix}^{-1}=\begin{pmatrix}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{2\sigma^4}{n}\end{pmatrix}$$

A6.5 证明最大似然估计量的不变性。

证明：将 $\boldsymbol{\alpha}$ 视为对原参数 $\boldsymbol{\theta}$ 的参数变换(reparameterization)，则可以将似然函数写为 $\boldsymbol{\alpha}$ 函数。记 $L(\boldsymbol{\theta};\boldsymbol{y})$ 为以 $\boldsymbol{\theta}$ 为参数的似然函数，其最大似然估计为 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$。定义 $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}\equiv\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$，需要证明 $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ 为对 $\boldsymbol{\alpha}$ 的最大似然估计。

定义以 $\boldsymbol{\alpha}$ 为参数的似然函数为

$$\begin{aligned}
L_g(\boldsymbol{\alpha};\boldsymbol{y}) &\equiv \sup_{\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{\alpha}\}} L(\boldsymbol{\theta};\boldsymbol{y}) && \text{(可能多个 } \boldsymbol{\theta} \text{ 满足“}\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{\alpha}\text{”，择其最大者)}\\
&\leqslant \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\boldsymbol{y}) && \text{(去掉约束条件“}\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{\alpha}\text{”，不会减小最大值)}\\
&= L(\hat{\boldsymbol{\theta}}_{\mathrm{ML}};\boldsymbol{y}) && \text{(根据最大似然估计量的定义)}\\
&= \sup_{\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})=\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}\}} L(\boldsymbol{\theta};\boldsymbol{y}) && \text{(由于 } \hat{\boldsymbol{\theta}}_{\mathrm{ML}}\in\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\}\text{，故可加约束条件)}\\
&= L_g(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}};\boldsymbol{y}) && \text{(根据以 } \boldsymbol{\alpha} \text{ 为参数的似然函数的定义)}
\end{aligned}$$

因此，$L_g(\boldsymbol{\alpha};\boldsymbol{y})\leqslant L_g(\hat{\boldsymbol{\alpha}}_{\mathrm{ML}};\boldsymbol{y})$，$\forall\boldsymbol{\alpha}$，即 $L_g(\boldsymbol{\alpha};\boldsymbol{y})$ 在 $\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ 处达到最大值，故 $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ 是对 $\boldsymbol{\alpha}$ 的最大似然估计。在上面的推导中，sup 为“上确界”(least upper bound)算子。如果上确界可以达到，则上确界等于最大值。

A6.6 关于 BHHH 法：为什么用 $\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'$ 来估计信息矩阵 $\boldsymbol{I}(\boldsymbol{\theta})$。

因为 $\dfrac{\partial\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}}=\dfrac{\partial\left[\sum_{i=1}^{n}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})\right]}{\partial\boldsymbol{\theta}}=\sum_{i=1}^{n}\dfrac{\partial\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\equiv\sum_{i=1}^{n}\boldsymbol{s}_i$，

所以 $\dfrac{\partial^2\ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=\sum_{i=1}^{n}\dfrac{\partial^2\ln f(\boldsymbol{y}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$。因此，

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}_0) &\equiv -\mathrm{E}\left[\frac{\partial^2\ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] && \text{(信息矩阵的定义)}\\
&= -\mathrm{E}\left[\sum_{i=1}^{n}\frac{\partial^2\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] && \text{(样本信息矩阵等于每个观测值信息矩阵之和)}\\
&= \sum_{i=1}^{n}\left[-\mathrm{E}\left(\frac{\partial^2\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\right] && \text{(求和与期望算子交换次序)}\\
&= \sum_{i=1}^{n}\left[\mathrm{E}\left(\frac{\partial\ln L(\boldsymbol{\theta}_0;\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}}\cdot\frac{\partial\ln L(\boldsymbol{\theta}_0;\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}'}\right)\right] && \text{(信息矩阵等式)}\\
&= \sum_{i=1}^{n}\mathrm{E}(\boldsymbol{s}_i\boldsymbol{s}_i') && (\boldsymbol{s}_i \text{ 的定义})\\
&= n\mathrm{E}(\boldsymbol{s}_i\boldsymbol{s}_i') && \text{(iid 的假设)}
\end{aligned}$$

由于 $\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'=n\left(\dfrac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'\right)$，而 $\dfrac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'\xrightarrow{p}\mathrm{E}(\boldsymbol{s}_i\boldsymbol{s}_i')$，故 $\sum_{i=1}^{n}\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i'$ 是 $n\mathrm{E}(\boldsymbol{s}_i\boldsymbol{s}_i')$ 的一致估计。

第 7 章　异方差与 GLS

古典线性回归模型(CLRM)假设的是一种理想状态。但现实的数据千奇百怪,常常不符合古典模型的某些假定。从这一章开始,我们将逐步减弱古典模型的各项假定。

7.1　异方差的后果

古典模型假设球型扰动项。"异方差"(heteroskedasticity)是违背球型扰动项假设的一种情形,即扰动项方差 $\mathrm{Var}(\varepsilon_i|\boldsymbol{X})$ 依赖于 i,而不是常数 σ^2。在存在异方差的情况下:

(1) OLS 估计量依然是无偏、一致且渐近正态的。这是因为,在证明以上性质时,并未用到"同方差"的假定。

(2) OLS 估计量方差 $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ 的表达式不再是 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$,因为 $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})\neq\sigma^2\boldsymbol{I}$。因此,通常的 t 检验、F 检验也失效了。

(3) 高斯 - 马尔可夫定理不再成立,即 OLS 不再是 BLUE。在存在异方差的情况下,本章将要介绍的 GLS 才是 BLUE。为了直观地理解为何 OLS 不再是 BLUE,假设 $\mathrm{Var}(\varepsilon_i|\boldsymbol{X})$ 是某解释变量 x_i 的增函数,即 x_i 越大则 $\mathrm{Var}(\varepsilon_i|\boldsymbol{X})$ 越大,参见图 7.1。

显然,OLS 回归线在 x_i 较小时可以较精确地估计,而在 x_i 较大时则难以精确估计。方差较大的数据包含的信息量较小,但 OLS 却对所有的数据等量齐观地进行处理。因此,从整体而言,异方差的存在使得 OLS 的效率降低。GLS 及其特例"加权最小二乘法"(Weighted Least Square,简记 WLS)正是通过对不同数据所包含信息量的不同进行相应的处理以提高估计效率(比如,给予信息量大的数据更大的权重)。

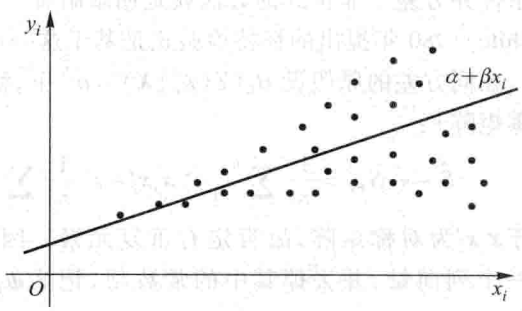


图 7.1　异方差的一种情形

7.2　异方差的例子

一般来说,截面数据较容易产生异方差现象。举例如下。

(1) 消费函数:

$$C_i=\alpha+\beta Y_i+\varepsilon_i$$

其中,C 为消费,Y 为收入。一般来说,富人的消费计划较有弹性,而穷人的消费多为必需品,很少变动。另一方面,富人的消费支出可能更难测量,故包含较多测量误差。因此,$\mathrm{Var}(\varepsilon_i|Y_i)$ 可能随

Y_i 的上升而变大。

(2) 企业的投资、销售收入与利润:大型企业的商业活动可能动辄以亿元计,而小型企业则以万元计,因此,扰动项的规模也不相同;如果将大、中、小型企业放在一起回归,就可能存在异方差。

(3) 组间异方差:如果样本包含两组(类)数据,则可能存在组内同方差,但组间异方差的情形,比如,第一组为自我雇佣者(企业主、个体户)的收入,第二组为打工族的收入,自我雇佣者的收入波动可能比打工族更大。

(4) 组平均数:如果数据本身就是组平均数,则大组平均数的方差要比小组平均数的方差小。比如,考虑全国各省的人均GDP,每个省一个数据。显然,人口较多的省份其方差较小,方差与人口数成反比。

(5) 时间序列数据中也可能出现条件异方差,比如第22章的ARCH模型。

7.3 异方差的检验

1. 看残差图(residual plot)

具体来说,可以看"残差 e_i 与拟合值 $\hat{y}_i$ 的散点图"(residual-versus-fitted plot),也可以看"残差 e_i 与某个解释变量 x_{ik} 的散点图"(residual-versus-predictor plot)。这是最直观的方法,但不严格。有关异方差的统计检验不少,下面介绍最流行的两种方法。

2. 怀特检验(White test)

既然在条件同方差下,稳健标准误还原为普通标准误,那么这二者之间的差别就可以用来度量条件异方差。非正式的方法就是用眼睛看一下稳健标准误与普通标准误是否相差不多。怀特(White)1980年提出的怀特检验正是基于这一思想。

在同方差的原假设 $H_0: \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{X}) = \sigma^2$ 下,稳健协方差矩阵与普通协方差矩阵之差收敛到一个零矩阵[①]:

$$\hat{\boldsymbol{S}} - s^2 \boldsymbol{S}_{XX} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' - s^2 \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \frac{1}{n}\sum_{i=1}^{n} (e_i^2 - s^2)\boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{p} \boldsymbol{0}_{K\times K} \tag{7.1}$$

由于 $\boldsymbol{x}_i\boldsymbol{x}_i'$ 为对称矩阵,故肯定有重复元素。因此,将随机矩阵 $\boldsymbol{x}_i\boldsymbol{x}_i'$ 中不重复的元素取出,排列成一个列向量,并去掉其中的常数项,记为 $\boldsymbol{\psi}_i$。$\boldsymbol{\psi}_i$ 包含了所有解释变量及其平方项与交叉项,记其维度为 m。比如,假设 $\boldsymbol{x}_i = (1 \quad x_{i2} \quad x_{i3})'$,则 $\boldsymbol{x}_i\boldsymbol{x}_i' = \begin{pmatrix} 1 & x_{i2} & x_{i3} \\ x_{i2} & x_{i2}^2 & x_{i2}x_{i3} \\ x_{i3} & x_{i2}x_{i3} & x_{i3}^2 \end{pmatrix}$,故 $\boldsymbol{\psi}_i = (x_{i2} \quad x_{i3} \quad x_{i2}^2 \quad x_{i3}^2 \quad x_{i2} \quad x_{i3})'$,$m=5$。为了得到一维的检验统计量。定义

$$\boldsymbol{c}_n \equiv \frac{1}{n}\sum_{i=1}^{n} (e_i^2 - s^2)\boldsymbol{\psi}_i \xrightarrow{p} \boldsymbol{0}_{m\times 1} \tag{7.2}$$

在一定的条件下可以使用中心极限定理,则 $\sqrt{n}\,\boldsymbol{c}_n \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{B})$,其中,$\boldsymbol{B}$ 为 $\boldsymbol{c}_n$ 的渐近方差。假设 $\hat{\boldsymbol{B}}$ 为 $\boldsymbol{B}$ 的一致估计量,则

① 在同方差的条件下,二者都是一致的,即收敛到同一个总体协方差矩阵,故二者之差收敛到零矩阵。

$$n\,\boldsymbol{c}_n'\hat{\boldsymbol{B}}^{-1}\boldsymbol{c}_n \xrightarrow{d} \chi^2(m) \tag{7.3}$$

其中，m 为 $\boldsymbol{c}_n$（也就是 $\boldsymbol{\psi}_i$）的维度。在实际操作上，一般进行如下的辅助回归：

$$e_i^2 \xrightarrow{\text{OLS}} \text{常数} + \boldsymbol{\psi}_i'\boldsymbol{\gamma} \tag{7.4}$$

并检验 $\boldsymbol{\psi}_i$ 中所有变量的系数 $\boldsymbol{\gamma}$ 均为 0。显然，如果上述回归的拟合优度 R^2 很低，即意味着 e_i^2 无法由解释变量及其平方项与交叉项来解释，故倾向于接受同方差的原假设。可以证明，$nR^2 \xrightarrow{d} \chi^2(m)$。如果 nR^2 很大（超过临界值），则拒绝原假设 H_0。为什么检验统计量是 nR^2 呢？事实上，在大样本中，nR^2 与检验整个回归方程显著性的 F 统计量是渐近等价的。

根据第 3 章的命题，对于回归方程“$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$”，检验原假设“$H_0: \beta_2 = \cdots = \beta_K = 0$”，则 F 统计量 $= \dfrac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$。进一步，在大样本情况下，$F$ 分布与 χ^2 分布是等价的（参见第 5 章附录），即 $(K-1)F = \dfrac{(n-K)R^2}{(1-R^2)} \xrightarrow{d} \chi^2(K-1)$。在原假设成立的情况下，当 $n \to \infty$ 时，$n-K \to n$[①]，$(1-R^2) \to 1$[②]，故 $(K-1)F \to nR^2$，因此 F 检验与 nR^2 检验在大样本下是等价的。

怀特检验的优点是，它可以检验任何形式的异方差；其缺点则是，如果 H_0 被拒绝，怀特检验并不提供有关异方差具体形式的信息。

3. BP 检验（Breusch and Pagan，1979）

假设回归模型为 $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$，检验以下原假设：

$$H_0: \mathrm{E}(\varepsilon_i^2 \mid x_2, \cdots, x_K) = \sigma^2 \tag{7.5}$$

如果 H_0 不成立，则条件方差 $\mathrm{E}(\varepsilon_i^2 \mid x_2, \cdots, x_K)$ 是 $(x_2, \cdots, x_K)$ 的函数，称为“条件方差函数”（conditional variance function）。BP 检验假设此条件方差函数为线性函数[③]：

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + u_i \tag{7.6}$$

如果认为异方差只与部分解释变量有关，则可以仅使用部分解释变量。也可以添加其他变量，如拟合值 $\hat{y}$，或不在回归方程中的变量 z。根据方程(7.6)，原假设简化为

$$H_0: \delta_2 = \cdots = \delta_K = 0 \tag{7.7}$$

由于扰动项 ε_i 不可观测，故使用残差平方 e_i^2 对解释变量进行辅助回归：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \tag{7.8}$$

仍然使用 nR^2 统计量：

$$nR^2 \xrightarrow{d} \chi^2(K-1) \tag{7.9}$$

其中，R^2 为辅助回归的 R^2。BP 检验与怀特检验的区别在于，后者还包括平方项与交叉项。因此，BP 检验可以看成是怀特检验的特例。BP 检验的优点在于其建设性，即可以帮助确认异方差的具体形式。

---

① 严格来说，应为 $(n-K)/n \to 1$。

② 在原假设“$H_0: \beta_2 = \cdots = \beta_K = 0$”成立的情况下，总体 $R^2 = 0$（约束回归，参见第 3 章），而样本 R^2（无约束回归）收敛于总体 R^2，故 $R^2 \to 0$，因此 $1 - R^2 \to 1$。

③ 更一般地，这个检验对于模型设定“$\varepsilon_i^2 = h(\delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK}) + u_i$”也成立，其中 $h(\cdot)$ 是任意的一个可导函数。

7.4 异方差的处理

1. 使用"OLS + 稳健标准误"

如果发现存在异方差,一种处理方法是,仍然进行OLS回归,但使用稳健标准误。这是最简单,也是目前通用的方法。只要样本容量较大,即使在异方差的情况下,若使用稳健标准误,则所有参数估计、假设检验均可照常进行。换言之,只要使用了稳健标准误,就可以与异方差"和平共处"了。

然而,还可能存在比OLS更有效的方法,比如下面介绍的GLS。

2. 广义最小二乘法(GLS)

假设 $\mathrm{Var}(\varepsilon|\boldsymbol{X})=\sigma^2\boldsymbol{V}(\boldsymbol{X})\neq\sigma^2\boldsymbol{I}_n$,其中 $\boldsymbol{V}(\boldsymbol{X})$ 为对称正定矩阵且已知,可能依赖于 $\boldsymbol{X}$。GLS的基本思想是,通过变量转换,使得转换后的模型满足球型扰动项的假定。为了作这个变换,首先介绍一个命题。

命题 对于对称正定矩阵 $\boldsymbol{V}_{n\times n}$,存在非退化矩阵 $\boldsymbol{C}_{n\times n}$,使得 $\boldsymbol{V}^{-1}=\boldsymbol{C}'\boldsymbol{C}$。

证明:参见附录。该命题的直观含义是,在一维情况下,"$\boldsymbol{V}$ 正定"即要求 $\boldsymbol{V}$ 为正数,故 $\frac{1}{\boldsymbol{V}}$ 也是正数,可以分解为 $\frac{1}{\sqrt{\boldsymbol{V}}}\cdot\frac{1}{\sqrt{\boldsymbol{V}}}$;但如果 $\boldsymbol{V}$ 为负数,则无法进行此分解。推广到多维的情形,就是以上命题。

需要注意的是,上述命题中的矩阵 $\boldsymbol{C}$ 不唯一,但这并不影响GLS的最终结果。

将原回归模型 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 两边同时左乘矩阵 $\boldsymbol{C}$ 可得

$$\boldsymbol{C}\boldsymbol{y}=\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{C}\boldsymbol{\varepsilon} \tag{7.10}$$

定义以下变量转换:

$$\tilde{\boldsymbol{y}}\equiv\boldsymbol{C}\boldsymbol{y},\quad \tilde{\boldsymbol{X}}\equiv\boldsymbol{C}\boldsymbol{X},\quad \tilde{\boldsymbol{\varepsilon}}\equiv\boldsymbol{C}\boldsymbol{\varepsilon} \tag{7.11}$$

则可以将模型写为

$$\tilde{\boldsymbol{y}}=\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}} \tag{7.12}$$

容易验证,变换后的回归模型仍然满足严格外生性,因为

$$\mathrm{E}(\tilde{\boldsymbol{\varepsilon}}\mid\tilde{\boldsymbol{X}})=\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{C}\boldsymbol{X})=\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{C}\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{0} \tag{7.13}$$

其中,由于 $\boldsymbol{C}$ 非退化,故 $\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{C}\boldsymbol{X})=\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{X})$,因为"$\boldsymbol{C}\boldsymbol{X}=\boldsymbol{Z}\Leftrightarrow\boldsymbol{X}=\boldsymbol{C}^{-1}\boldsymbol{Z}$"。而且,球型扰动项的假定也得到满足,因为

$$\begin{aligned}\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}}\mid\tilde{\boldsymbol{X}})&=\mathrm{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\boldsymbol{X})=\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}'|\boldsymbol{X})=\boldsymbol{C}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{C}'\\&=\sigma^2\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}'=\sigma^2\boldsymbol{C}(\boldsymbol{V}^{-1})^{-1}\boldsymbol{C}'=\sigma^2\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}'\\&=\sigma^2\boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{C}')^{-1}\boldsymbol{C}'=\sigma^2\boldsymbol{I}_n\end{aligned} \tag{7.14}$$

因此,高斯-马尔可夫定理仍然成立。对变换后的模型使用OLS即得到GLS估计量:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}&=(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}=[(\boldsymbol{C}\boldsymbol{X})'(\boldsymbol{C}\boldsymbol{X})]^{-1}(\boldsymbol{C}\boldsymbol{X})'\boldsymbol{C}\boldsymbol{y}\\&=(\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{y}=(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}\end{aligned} \tag{7.15}$$

因此，虽然 $\boldsymbol{C}$ 不唯一，但 $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 却是唯一的，因为 $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 不依赖于 $\boldsymbol{C}$。由于高斯－马尔可夫定理成立，故 $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 是 BLUE。因此，GLS 比 OLS 更有效。当然，前提是必须确切地知道协方差矩阵 $\boldsymbol{V}$，而这在实践中并没有保证。

3. 加权最小二乘法(WLS)

WLS 是 GLS 的特例，也提供了直观理解 GLS 的好机会。假设仅存在异方差，而没有自相关，即 $\boldsymbol{V}(\boldsymbol{X})$ 为对角矩阵，但主对角线上的元素不完全相等。显然，方差较小的数据提供的信息量较大，而方差较大的数据提供的信息量较小。WLS 正是根据信息量的大小对数据进行加权处理的。

假定 $\mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i) = \mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i) = \sigma^2 v_i(\boldsymbol{X})$，即

$$\boldsymbol{V} = \begin{pmatrix} v_1 & & 0 \\ & v_2 & \\ & & \ddots \\ 0 & & v_n \end{pmatrix}, \quad \boldsymbol{V}^{-1} = \begin{pmatrix} 1/v_1 & & 0 \\ & 1/v_2 & \\ & & \ddots \\ 0 & & 1/v_n \end{pmatrix} \tag{7.16}$$

由于 $\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$，可知

$$\boldsymbol{C} = \boldsymbol{C}' = \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & & \ddots \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \tag{7.17}$$

$$\tilde{\boldsymbol{y}} \equiv \boldsymbol{C}\boldsymbol{y} = \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & \ddots & \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_1/\sqrt{v_1} \\ y_2/\sqrt{v_2} \\ \vdots \\ y_n/\sqrt{v_n} \end{pmatrix} \tag{7.18}$$

$$\begin{aligned} \tilde{\boldsymbol{X}} \equiv \boldsymbol{C}\boldsymbol{X} &= \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & \ddots & \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ x_{21} & \cdots & x_{2K} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nK} \end{pmatrix} \\ &= \begin{pmatrix} x_{11}/\sqrt{v_1} & \cdots & x_{1K}/\sqrt{v_1} \\ x_{21}/\sqrt{v_2} & \cdots & x_{2K}/\sqrt{v_2} \\ \vdots & & \vdots \\ x_{n1}/\sqrt{v_n} & \cdots & x_{nK}/\sqrt{v_n} \end{pmatrix} \end{aligned} \tag{7.19}$$

所以，权重为 $1/\sqrt{v_i}$(标准差的倒数)。对于第 i 个观测值(个体)，其回归方程变为

$$\frac{y_i}{\sqrt{v_i}} = \beta_1 \frac{x_{i1}}{\sqrt{v_i}} + \beta_2 \frac{x_{i2}}{\sqrt{v_i}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{v_i}} + \frac{\varepsilon_i}{\sqrt{v_i}} \tag{7.20}$$

由于新的扰动项为 $\varepsilon_i/\sqrt{v_i}$，故可以将 WLS 视为最小化“加权的残差平方和”：

$$\min_{\tilde{\beta}} \text{SSR} = \sum_{i=1}^{n} \left(e_i / \sqrt{v_i} \right)^2 = \sum_{i=1}^{n} \frac{e_i^2}{v_i} \tag{7.21}$$

从这个角度来看，权重为 $1/v_i$（方差的倒数），在 Stata 中也是这样约定的。

4. 可行广义最小二乘法（Feasible GLS，简记 FGLS）[①]

GLS 与 WLS 的缺点是假设扰动项的协方差矩阵为已知。这常常是不现实的假定。为此，必须先用样本数据来一致地估计 $\boldsymbol{V}(\boldsymbol{X})$，然后才能使用 GLS。这种方法被称为 FGLS 或“可行加权最小二乘法”（Feasible WLS，简记 FWLS），即

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} \tag{7.22}$$

其中，$\hat{\boldsymbol{V}}$ 是 $\boldsymbol{V}$ 的一致估计。FGLS 的困难在于 $\boldsymbol{V}(\boldsymbol{X})$ 中包含太多的参数，待估计参数个数可能多达 $[n(n+1)/2]$（$\boldsymbol{V}(\boldsymbol{X})$ 是 n 阶对称矩阵），而样本容量仅为 n。在实际操作中，常考虑参数较少的情形，比如只有异方差，或只有一阶自相关（参见第 8 章）。

下面以 FWLS 为例，来说明 FGLS 的操作。在仅有异方差的情况下，在作 BP 检验的时候，通过辅助回归 $e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i$ 就可以获得对 σ_i^2 的估计值 $\hat{\sigma}_i^2$。在作此回归时，如果某些变量不显著（对 e_i^2 无解释力），则可以略去；也可以加上其他变量，如 $\hat{y}_i$ 及 $\hat{y}_i^2$。为了保证 $\hat{\sigma}_i^2$ 为正，实践上常假设此辅助回归为指数函数的形式：

$$e_i^2 = \sigma^2 \exp(\delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK}) v_i \tag{7.23}$$

其中，v_i 为乘积形式的扰动项。取对数后可得

$$\ln e_i^2 = (\ln \sigma^2 + \delta_1) + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + \ln v_i \tag{7.24}$$

通过回归上式，可以得到对 $\ln e_i^2$ 的预测值，记为 $\ln \hat{\sigma}_i^2$，进而得到拟合值 $\hat{\sigma}_i^2 = e^{\ln \hat{\sigma}_i^2}$，然后以 $1/\hat{\sigma}_i^2$ 为权重进行 WLS 估计。

5. 究竟使用“OLS + 稳健标准误”还是 FWLS

“OLS + 稳健标准误”与 FWLS 各有利弊。从理论上来说，GLS 是 BLUE，这是 GLS 吸引人的地方。然而 FGLS 还是 BLUE 吗？不是。事实上，FGLS 既非线性估计，也不是无偏估计，根本就没有资格参加 BLUE 的评选。根据方程（7.22），$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$，而 $\hat{\boldsymbol{V}}$ 是数据 $(\boldsymbol{y}, \boldsymbol{X})$ 的非线性函数，故 $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ 是 $\boldsymbol{y}$ 的非线性函数，一般来说是有偏的。FWLS 的优点主要体现在大样本理论中。如果 $\hat{\boldsymbol{V}}$ 是 $\boldsymbol{V}$ 的一致估计，则 FWLS 是一致的，而且在大样本下比 OLS 更有效。FWLS 的另一缺点是必须估计条件方差函数 $\text{Var}(\varepsilon_i | \boldsymbol{x}_i)$，而通常并不知道条件方差函数的具体形式[②]。如果该函数的形式设定不正确，则根据 FWLS 计算的标准误可能失效[③]，导致不正确的统计推断。

使用“OLS + 稳健标准误”的好处是，它对回归系数及标准误的估计都是一致的，并不需要知道条件方差函数的形式。而在 Stata 中的操作也十分简单，只要在回归命令 `reg` 之后加选择项“`robust`”即可。

总之，“OLS + 稳健标准误”更为稳健（即适用于更一般的情形），而 FWLS 更有效。因此，

---

① 也称为“Estimated GLS”，简记 EGLS。

② 这种形式不一定是线性的，也可以有平方项、对数等非线性形式。即使对于一元回归，条件方差函数也可能有很多种形式。对于多元回归，可能的函数形式就更多了。

③ 因为 GLS 估计量的方差依赖于待估计的 $\boldsymbol{V}$，参见习题。

必须在稳健性与有效性之间做一个选择。在某种意义上，前者相当于“万金油”[①]，后者则相当于“特效药”。由于“病情”通常难以诊断，不知道该用哪种特效药（哪种条件方差函数？），故特效药也可能失效或起反作用。具体来说，如果对 V 的估计不准，则 FGLS 的性能可能还不如 OLS。因此，Stock and Watson (2011) 推荐，在大多数情况下应该使用“OLS + 稳健标准误”。

7.5 处理异方差的 Stata 命令及实例

1. 画残差图

完成回归后，可使用以下命令得到残差图：

```
rvfplot              (residual-versus-fitted plot)
rvpplot varname      (residual-versus-predictor plot)
```

仍以 Nerlove(1963)数据为例：

```
.use nerlove.dta,clear
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs = 145
				F(4, 140) = 437.90
Model	269.524728	4	67.3811819	Prob > F = 0.0000
Residual	21.5420958	140	.153872113	R-squared = 0.9260
				Adj R-squared = 0.9239
Total	291.066823	144	2.02129738	Root MSE = .39227

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

```
. rvfplot  (画残差与拟合值的散点图，结果如图 7.2)
```

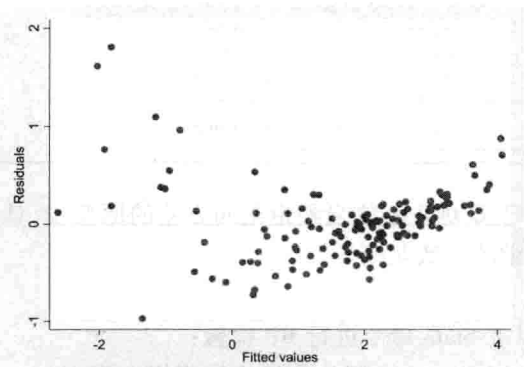


图 7.2 残差与拟合值的散点图

① 指谁都可以用。

从图7.2可以大致看出，当 $\hat{y}$（即lntc的拟合值）较小时，扰动项的方差较大。进一步考察残差与解释变量lnq的散点图。

```
. rvpplot lnq   (画残差与解释变量的散点图，结果如图7.3)
```

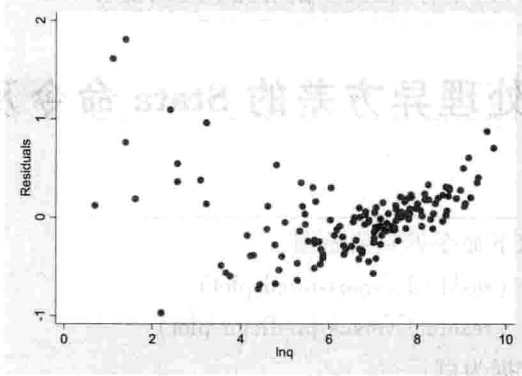


图7.3 残差与解释变量lnq的散点图

以上两图的大致轮廓基本一致，表明很可能存在异方差，即扰动项的方差随着观测值而变。

2. 怀特检验

完成回归后，使用Stata命令“estat imtest,white”即可以进行怀特检验。其中，“estat”指的是“post-estimation statistics”（估计后统计量），而“imtest”指的是“information matrix test”[①]。继续用Nerlove(1963)为例：

```
. estat imtest,white
```

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(14)      =     73.88
         Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
Heteroskedasticity |      73.88     14    0.0000
            Skewness |      22.79      4    0.0001
            Kurtosis |       2.62      1    0.1055
---------------------+-----------------------------
               Total |      99.29     19    0.0000
---------------------------------------------------
```

检验结果显示，p 值等于0.000 0，故强烈拒绝同方差的原假设，认为存在异方差。这个检验结果证实了根据残差图所做的大致判断。

3. BP检验

完成回归后，可使用以下Stata命令进行BP检验：

```
estat hettest              (默认设置为使用拟合值 ŷ)
estat hettest,rhs          (使用方程右边的解释变量，而不是 ŷ)
```

① 也可以输入命令“ssc install whitetst”，下载非官方命令whitetst。

```
estat hettest [varlist]  (指定使用某些解释变量)
```

最初的 BP 检验假设扰动项 ε_i 服从正态分布,有一定局限性。Koenker(1981)将此假定减弱为独立同分布(iid),在实际中较多采用,其对应的 Stata 命令为

```
estat hettest,iid
estat hettest,rhs iid
estat hettest [varlist],iid
```

回到 Nerlove(1963)的例子:

```
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lntc

         chi2(1)      =    29.13
         Prob > chi2  =   0.0000

. estat hettest,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lnq lnpl lnpk lnpf

         chi2(4)      =    36.16
         Prob > chi2  =   0.0000

. estat hettest lnq,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lnq

         chi2(1)      =    32.10
         Prob > chi2  =   0.0000
```

以上各种形式 BP 检验的 p 值都等于 0.000 0,故强烈拒绝同方差的原假设,这个结果与怀特检验相同。

4. WLS

在得到扰动项方差的估计值 $\{\hat{\sigma}_i^2\}_{i=1}^n$ 后,可以使用 WLS 来估计原回归方程。假设已把 $\{\hat{\sigma}_i^2\}_{i=1}^n$ 存储在变量"var"上,则可通过如下 Stata 命令来实现 WLS:

```
reg y x1 x2 x3 [aw=1/var]
```

其中"aw"表示"analytical weight",为扰动项方差(而不是标准差)的倒数。对于 Nerlove(1963)的数据,假设"$\ln \hat{\sigma}_i^2 = \delta \ln q_i + u_i$"(无截距项)。首先计算残差(记为 e1):

```
. quietly reg lntc lnq lnpl lnpk lnpf
. predict e1,res
```

其中,"quietly"表示不显示命令运行的结果。

其次,生成残差的平方(记为 e2):

```
. g e2=e1^2
```

然后取对数,再进行辅助回归:

```
. g lne2=log(e2)
```

```
. reg lne2 lnq,noc
```

Source	SS	df	MS	Number of obs = 145
				F(1, 144) = 419.95
Model	2065.53636	1	2065.53636	Prob > F = 0.0000
Residual	708.275258	144	4.91857818	R-squared = 0.7447
				Adj R-squared = 0.7429
Total	2773.81162	145	19.1297353	Root MSE = 2.2178

lne2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	-.5527533	.0269733	-20.49	0.000	-.6060681	-.4994384

上表显示，$R^2=0.7447$，即解释变量 lnq 可以解释 $\ln e_i^2$ 近75%的变动，残差平方的变动与 lnq 高度相关。然后计算以上辅助回归的拟合值（记为 lne2f）：

```
. predict lne2f
(option xb assumed;fitted values)
```

去掉对数，即得到 WLS 所要使用的权重之倒数：

```
. g e2f = exp(lne2f)
```

最后，进行 WLS 回归：

```
. reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f]
```

Source	SS	df	MS	Number of obs = 145
				F(4, 140) = 895.03
Model	173.069988	4	43.2674971	Prob > F = 0.0000
Residual	6.76790874	140	.048342205	R-squared = 0.9624
				Adj R-squared = 0.9613
Total	179.837897	144	1.24887428	Root MSE = .21987

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.8759035	.0153841	56.94	0.000	.8454883	.9063187
lnpl	.5603879	.1734141	3.23	0.002	.2175389	.9032369
lnpk	-.0929807	.1960402	-0.47	0.636	-.4805627	.2946014
lnpf	.4672438	.0616476	7.58	0.000	.3453632	.5891243
_cons	-5.522088	.9928472	-5.56	0.000	-7.485	-3.559176

WLS 回归的结果显示，lnpk 的系数估计值由“-0.22”（OLS 估计值）改进为“-0.09”（其理论值应为正数）。另一方面，使用 OLS 时，变量 lnpl 的 p 值为 0.13，即使在10%的水平上也不显著；而使用 WLS 后，该变量的 p 值变为 0.002，在1%的水平上显著不为0。由此可知，由于 Nerlove(1963)数据存在明显的异方差，故使用 WLS 后提高了估计效率。

也可使用外部命令 wls0 进行加权最小二乘法，下载方法为“net install wls0.pkg”（或输入命令“findit wls0”寻找下载地址）。

7.6 Stata 命令的批处理

在进行计量实证分析时，有时需要使用一系列命令对数据集进行处理。如果每次只在命令窗口输入一个命令，可能带来不便，因为有时需要对某些命令进行调整，然后把所有命令重新执

行一遍[①]。此时,可以把所有命令放入一个 Stata“do 文件”(即以“do”为扩展名的程序文件),进行批处理。

在 Stata 中,点击菜单 Window→Do-file Editor→New Do-file Editor(或直接点击 New Do-file Editor 图标,在 Data Editor 图标左边),即可打开一个“do 文件编辑器”(Do-file Editor),在其中写入需要执行的命令。

以上面的加权最小二乘法为例。假设已在电脑的 E 盘创建了一个叫“wls_nerlove.smcl”的日志文件(log file),并且数据文件“nerlove.dta”也在 E 盘的根目录。可在“do 文件编辑器”中输入如下命令:

```
* WLS for Nerlove(1963)
log using E:\\wls_nerlove.smcl,replace
set more off
use E:\\nerlove.dta,clear
reg lntc lnq lnpl lnpk lnpf
predict e1,res
g e2 = e1^2
g lne2 = log(e2)
reg lne2 lnq,noc
predict lne2f
g e2f = exp(lne2f)
* Weighted least square regression
reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f]
log close
exit
```

其中,“*”表示不执行其后的命令,常用来作为注释(便于自己或他人理解此程序)。“log using E:\\wls_nerlove.smcl,replace”表示将 Stata 运行结果记录于日志文件“wls_nerlove.smcl”(并可覆盖此文件的原有内容)。当 Stata 的输出结果超过一个屏幕时,通常需要点击“more”,才能继续。为了避免这种不便,“set more off”使得 Stata 输出结果可以自动地连贯显示。

输入以上命令后,即可点击“do 文件编辑器”窗口的菜单 File→Save(或 Save As)存储此程序文件,比如,存为“wls_nerlove.do”。如果要执行此“do 文件”,只要在其存储位置用鼠标双击文件“wls_nerlove.do”的图标即可。也可以在 Stata 中点击菜单“File”→“Do”来执行此文件。如果要编辑此文件,可以用鼠标右键点击“wls_nerlove.do”的图标,然后选择用“记事本”(Notepad)打开,编辑后直接存盘即可。

在上面的程序中,不同的命令是通过换行来区分的。然而,如果某个命令很长,一行排不下,这种格式可能会出现问题。此时,可以选择以分号“;”来表示一个命令的结束。比如,可以把上面的程序改写如下:

```
* WLS for Nerlove(1963)
# delimit;
```

① 或许你的导师或审稿人要求你修改模型,然后重新计算一遍。

```
log using E:\\wls_nerlove.smcl,replace;
set more off;
use E:\\nerlove.dta,clear;
reg lntc lnq lnpl lnpk lnpf;
predict e1,res;
g e2 = e1^2;
g lne2 = log(e2);
reg lne2 lnq,noc;
predict lne2f;
g e2f = exp(lne2f);
 * Weighted least square regression
reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f];
log close;
exit;
```

其中,"# delimit;"告诉 Stata 使用分号";"来表示一个命令的结束。为此,每一个命令结束后,都必须加上分号";";否则,Stata 将把下一行(直至下一个分号";"出现之前的内容)也作为命令的一部分,从而出错。

习 题

7.1 在条件异方差情况下,普通的非稳健标准误是真实标准误的一致估计吗?请具体说明为什么。

7.2 假设 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=\sigma^2\boldsymbol{V}\neq\sigma^2\boldsymbol{I}_n$,其中 $\boldsymbol{V}$ 为对称正定矩阵且已知。

(1) 计算 $\mathrm{Var}(\boldsymbol{b}\mid\boldsymbol{X})$,其中 $\boldsymbol{b}$ 为 OLS 估计量。

(2) 计算 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid\boldsymbol{X})$,其中 $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 为 GLS 估计量。

(3) 矩阵 $\mathrm{Var}(\boldsymbol{b}\mid\boldsymbol{X})$ 与 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid\boldsymbol{X})$ 有何关系?

7.3 数据集 hprice2a.dta 包含了美国波士顿 506 个社区的房屋中位数价格的横截面数据。考虑如下的"特征价格回归"(Hedonic Price Regression,即认为房价由房屋性能所决定):

$$\mathrm{lprice}_i=\beta_1+\beta_2\ \mathrm{lnox}_i+\beta_3\ \mathrm{ldist}_i+\beta_4\ \mathrm{rooms}_i+\beta_5\ \mathrm{stratio}_i+\varepsilon_i \tag{7.25}$$

其中,lprice 为房价的对数,lnox 为空气污染程度的对数,ldist 为社区到就业中心的距离,rooms 为房屋的平均房间数,stratio 为社区学校的学生-教师比例,下标 i 表示"第 i 个社区"。

(1) 使用普通标准误进行回归,并分别检验"$H_0:\beta_2=\beta_5$"与"$H_0:\beta_4=0.33$"。

(2) 使用稳健标准误进行回归,并分别检验"$H_0:\beta_2=\beta_5$"与"$H_0:\beta_4=0.33$"。

(3) 以 5% 的置信度,使用怀特检验,检验是否存在异方差。

(4) 以 5% 的置信度,使用 BP 检验,检验是否存在异方差(假设扰动项为独立同分布,分别以拟合值 $\hat{y}$ 以及所有解释变量进行检验)。

(5) 假设条件异方差仅依赖于拟合值 $\hat{y}$(不含常数项),进行 FWLS 估计。

附 录

A7.1 命题 **对于对称正定矩阵 $\boldsymbol{V}_{n\times n}$,存在非退化矩阵 $\boldsymbol{C}_{n\times n}$,使得 $\boldsymbol{V}^{-1}=\boldsymbol{C}'\boldsymbol{C}$。**

证明:由于 $\boldsymbol{V}$ 为实对称矩阵,故 $\boldsymbol{V}$ 可对角化,即存在正交矩阵 $\boldsymbol{B}$,满足 $\boldsymbol{B}^{-1}=\boldsymbol{B}'$,使得 $\boldsymbol{B}^{-1}\boldsymbol{V}\boldsymbol{B}=\boldsymbol{\Lambda}$,其中 $\boldsymbol{\Lambda}$ 为对

角矩阵。由于 $\boldsymbol{V}$ 正定，故 $\boldsymbol{\Lambda}$ 的主对角线元素（即 $\boldsymbol{V}$ 的特征值）均为正，而 $\boldsymbol{\Lambda}^{-1}$ 的主对角线元素也都为正。$\boldsymbol{B}$ 中的向量为 $\boldsymbol{V}$ 的特征向量（$\boldsymbol{B}$ 不唯一）：

$$\boldsymbol{B}^{-1}\boldsymbol{V}\boldsymbol{B}=\boldsymbol{\Lambda}\Rightarrow\boldsymbol{V}=\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{-1}\quad(\text{左乘}\boldsymbol{B},\text{右乘}\boldsymbol{B}^{-1})$$

$$\boldsymbol{V}^{-1}=(\boldsymbol{B}^{-1})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^{-1}=\boldsymbol{B}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^{-1}=(\boldsymbol{B}\boldsymbol{\Lambda}^{-1/2})(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{B}')\equiv\boldsymbol{C}\boldsymbol{C}'$$

其中，$\boldsymbol{\Lambda}^{-1/2}$ 为将对角矩阵 $\boldsymbol{\Lambda}^{-1}$ 中的每个元素开平方，而 $\boldsymbol{C}\equiv\boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}$。由于 $\boldsymbol{B}$ 与 $\boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}$ 都是非退化矩阵，故 $\boldsymbol{C}\equiv\boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}$ 也非退化，即 $\boldsymbol{C}^{-1}$ 存在。

由于 $\boldsymbol{V}^{-1}$ 为对称矩阵，故 $\boldsymbol{V}^{-1}=(\boldsymbol{V}^{-1})'=(\boldsymbol{C}\boldsymbol{C}')'=\boldsymbol{C}'\boldsymbol{C}$。

注意：由于 $\boldsymbol{B}$ 不唯一，故 $\boldsymbol{C}$ 也不唯一，但这并不影响 GLS 的最终结果。

第8章 自 相 关

8.1 自相关的后果

违反球型扰动项的另一情形是自相关。如果存在 $i \neq j$，使得 $E(\varepsilon_i \varepsilon_j | \boldsymbol{X}) \neq 0$，即扰动项的协方差阵 $Var(\varepsilon | \boldsymbol{X})$ 的非主对角线元素不全为 0，则称存在“自相关”(autocorrelation)或“序列相关”(serial correlation)。在有自相关的情况下：

(1) OLS 估计量依然是无偏且一致的，这是因为，在证明这些性质时，并未用到“无自相关”的假定；

(2) OLS 估计量依然服从渐近正态分布[①]；

(3) OLS 估计量方差 $Var(\boldsymbol{b} | \boldsymbol{X})$ 的表达式不再是 $\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$，因为 $Var(\boldsymbol{\varepsilon} | \boldsymbol{X}) \neq \sigma^2 \boldsymbol{I}$，因此，通常的 t 检验、F 检验也失效了；

(4) 高斯 - 马尔可夫定理不再成立，即 OLS 不再是 BLUE，为了直观地理解为何 OLS 不再是 BLUE，假设扰动项存在正自相关，即 $E(\varepsilon_i \varepsilon_j | \boldsymbol{X}) > 0$，参见图 8.1。

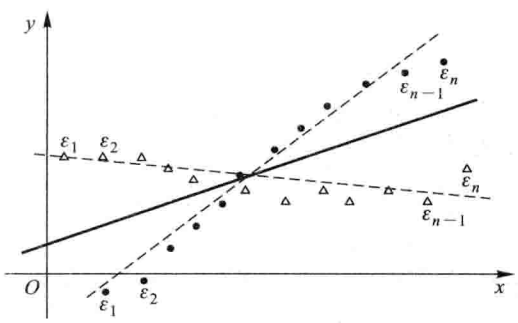


图 8.1 自相关的后果

在图 8.1 中，实线表示真实的总体回归线。如果 $\varepsilon_1 > 0$(图中左边小三角形)，由于扰动项存在正自相关，则 $\varepsilon_2 > 0$ 的可能性也很大。而如果 $\varepsilon_{n-1} < 0$(图中右边小三角形)，则 $\varepsilon_n < 0$ 的可能性也就很大。此时，样本回归线(虚线)很可能左侧翘起、右侧下垂，使得对回归线斜率的估计过小。

反之，如果 $\varepsilon_1 < 0$(图中左边小圆点)，由于扰动项存在正自相关，故 $\varepsilon_2 < 0$ 的可能性也很大。而如果 $\varepsilon_{n-1} > 0$(图中右边小圆点)，则 $\varepsilon_n > 0$ 的可能性也就很大。此时，样本回归线(虚线)很可能左侧下垂、右侧翘起，使得对回归线斜率的估计过大。

总之，由于自相关的存在，使得根据样本数据估计的回归线上下摆动幅度增大，导致参数估计变得不准确。而 OLS 估计忽略了扰动项自相关所包含的信息，故不是最有效的估计方法。

---

① 为了得到这个结论，需要使用不同于“鞅差分序列”的中心极限定理，因为在有常数项的情况下，第 5 章假定 5.5(即 $\boldsymbol{g}_i \equiv \boldsymbol{x}_i \varepsilon_i$ 为鞅差分序列)意味着扰动项“无自相关”(参见下文)。

8.2 自相关的例子

(1) 时间序列数据中的自相关:由于经济活动通常具有某种连续性或持久性,自相关现象在时间序列中比较常见。比如,相邻两年的 GDP 增长率、通货膨胀率。又比如,某个意外事件或新政策的效应需要逐步地随时间释放出来。再比如,最优资本存量需要通过若干年的投资才能逐渐达到(滞后的调整过程)。

(2) 截面数据中的自相关:一般来说,截面数据不容易出现自相关,但相邻的观测单位之间也可能存在"溢出效应"(spillover effect or neighborhood effect),这种自相关也称为"空间自相关"(spatial autocorrelation)。比如,相邻的省份、国家之间的经济活动相互影响(通过贸易、投资、劳动力流动等);相邻地区的农业产量受到类似天气变化的影响;同一社区内的房屋价格存在相关性。参见第 29 章"空间计量经济学"。

(3) 对数据的人为处理:如果数据中包含移动平均数(moving average)、内插值(interpolation)或季节调整(seasonal adjustment)时,则从理论上即可判断存在自相关。需要注意的是,统计局提供的某些数据可能已经事先经过了这些人为处理。

(4) 设定误差(misspecification):如果模型设定中遗漏了某个自相关的解释变量,并被纳入到扰动项中,则会引起扰动项的自相关。这种由于设定误差而导致的自相关,即便在截面数据中也可能存在。

8.3 自相关的检验

1. 画图

可以将残差 e_t 与滞后残差 e_{t-1} 画成散点图(命令 scatter),也可以画残差的"自相关图"(correlogram),显示各阶样本自相关系数(命令 ac)或偏自相关系数(命令 pac)[①]。此法虽直观,但不严格。

2. BG 检验

假设原模型为 $y_t=\beta_0+\beta_1 x_{t1}+\cdots+\beta_K x_{tK}+\varepsilon_t$,并假设存在一阶自相关,即 $\varepsilon_t=\rho\varepsilon_{t-1}+u_t$[②],其中 u_t 为白噪声,则很自然地可以考虑回归 $e_t \xrightarrow{\text{OLS}} e_{t-1}$,并检验 $H_0:\rho=0$。更一般地,由于可能存在高阶自相关,考虑扰动项的 p 阶自回归过程:

$$\varepsilon_t=\rho_1\varepsilon_{t-1}+\cdots+\rho_p\varepsilon_{t-p}+u_t \tag{8.1}$$

检验原假设 $H_0:\rho_1=\cdots=\rho_p=0$。由于扰动项 $\{\varepsilon_t\}$ 不可观测,故用残差 $\{e_t\}$ 来替代,并引入所

① 有关"偏自相关",参见第 20 章。

② 此处假定常数项为 0,因为 $E(\varepsilon_t)=0$。

有解释变量①，考虑以下辅助回归：

$$e_t \xrightarrow{\text{OLS}} x_{t1},\cdots,x_{tK},e_{t-1},\cdots,e_{t-p} \quad (t=p+1,\cdots,n) \tag{8.2}$$

由于使用了滞后残差值 e_{t-p}，损失了 p 个样本值②，故辅助回归的样本容量仅为 $(n-p)$。使用 nR^2 形式的 LM 统计量：

$$(n-p)R^2 \xrightarrow{d} \chi^2(p) \tag{8.3}$$

如果 $(n-p)R^2$ 超过了 $\chi^2(p)$ 的临界值，则拒绝“无自相关”的原假设。这个检验被称为“Breusch-Godfrey 检验”，简称 BG 检验（Breusch，1978；Godfrey，1978）。Davidson and MacKinnon（1993）建议，把残差向量 $\boldsymbol{e}$ 中因滞后而缺失的项用其期望值 $\mathrm{E}(\boldsymbol{e})=\mathbf{0}$ 来代替③，以保持样本容量仍为 n，然后使用统计量 $nR^2 \xrightarrow{d} \chi^2(p)$。Davidson-MacKinnon 方法为 Stata 的默认设置。

3. Box-Pierce Q 检验

定义残差的各阶样本自相关系数为

$$\hat{\rho}_j \equiv \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2} \quad (j=1,2,\cdots,p) \tag{8.4}$$

如果原假设 $H_0:\rho_1=\cdots=\rho_p=0$ 成立，则 $\hat{\rho}_j \xrightarrow{p} 0$，$\sqrt{n}\hat{\rho}_j \xrightarrow{d}$ 正态分布，$j=1,2,\cdots,p$。残差的各阶样本自相关系数平方和的 n 倍，就是“Box-Pierce Q 统计量”（Box and Pierce，1970），即

$$Q_{\mathrm{BP}} \equiv n\sum_{j=1}^{p}\hat{\rho}_j^2 \xrightarrow{d} \chi^2(p) \tag{8.5}$$

经过改进的“Ljung-Box Q 统计量”（Ljung and Box，1979）为

$$Q_{\mathrm{LB}} \equiv n(n+2)\sum_{j=1}^{p}\frac{\hat{\rho}_j^2}{n-j} \xrightarrow{d} \chi^2(p) \tag{8.6}$$

这两种 Q 统计量在大样本下是等价的（参见习题），但 Ljung-Box Q 统计量的小样本性质更好，故为 Stata 所采用。

如何确定自相关阶数 p 呢？没有确定的规则。如果 p 太小，则可能忽略了高阶自相关的存在；但如果 p 较大（与样本容量 n 相比），则 Q 统计量的小样本分布可能与 $\chi^2(p)$ 相差较远。Stata 默认的 p 值为 $p=\min\{\mathrm{floor}(n/2)-2,40\}$，其中 $\mathrm{floor}(n/2)$ 为不超过 $n/2$ 的最大整数。

4. DW 检验

“DW 检验”（Durbin and Watson，1950）是较早出现的自相关检验，但现已不常用。它的主要缺点是只能检验一阶自相关，而且必须在解释变量满足严格外生性的情况下才成立（BG 检验与 Q 检验都没有这些限制）④。DW 检验的统计量为

---

① 如果不在辅助回归中引入解释变量 $\{x_{t1},\cdots,x_{tK}\}$，则由于样本残差 e_t 是 $\{x_{t1},\cdots,x_{tK}\}$ 的函数，故可以将 $\{x_{t1},\cdots,x_{tK}\}$ 视为遗漏变量（omitted variables）。此时，除非假设严格外生性（strict exogeneity），则辅助回归中的解释变量 $\{e_{t-1},\cdots,e_{t-p}\}$ 将与辅助回归的扰动项相关，导致不一致的估计。如果在辅助回归中引入 $\{x_{t1},\cdots,x_{tK}\}$，则不需要严格外生性的假设。这使得 BG 检验更加稳健。

② 如果用 e_1 作被解释变量，需要知道 $e_0,e_{-1},\cdots,e_{-p+1}$，但我们并没有这些数据，故会损失 p 个样本值。

③ 即令 $e_0=e_{-1}=\cdots=e_{-p+1}=0$。

④ 因此，如果解释变量包括被解释变量的滞后值，则不能使用 DW 检验。

$$\mathrm{DW} \equiv d \equiv \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2} \tag{8.7}$$

$$\approx 2 - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \equiv 2(1 - \hat{\rho}_1)$$

其中，$\hat{\rho}_1$ 为残差的一阶自相关系数。因此，大致而言，当 $d=2$ 时，$\hat{\rho}_1 \approx 0$，无一阶自相关；当 $d=0$ 时，$\hat{\rho}_1 \approx 1$，一阶正自相关；当 $d=4$ 时，$\hat{\rho}_1 \approx -1$，一阶负自相关。

DW 检验的另一缺点是其 d 统计量依赖于数据矩阵 $\boldsymbol{X}$，无法制成统计表，而必须使用上限分布 d_U 与下限分布 d_L（$d_L < d < d_U$）来判断。尽管如此，得到 d_U 与 d_L 的临界值后，仍然存在无结论区域。从 DW 统计量的表达式来看，其本质也就是残差的一阶自相关系数，不能指望它提供太多的信息。

8.4 自相关的处理

如果经过检验确认存在自相关，则大致有以下四种处理方法。

1. 使用“OLS + 异方差自相关稳健的标准误”

仍然用 OLS 来估计回归系数，但使用“异方差自相关稳健的标准误”（Heteroskedasticity and Autocorrelation Consistent Standard Error，简记 HAC），即在存在异方差与自相关的情况下也成立的稳健标准误。这种方法被称为“Newey-West 估计法”（Newey and West，1987），它只改变标准误的估计值，并不改变回归系数的估计值。

为什么第 5 章介绍的“异方差稳健标准误”不适用于存在自相关的情形呢？回顾在推导异方差稳健标准误的过程中，什么地方引入了扰动项无自相关的假设，可知问题出在假定 5.5，即 $\boldsymbol{g}_i \equiv \boldsymbol{x}_i \varepsilon_i$ 为鞅差分序列的假定。

命题 如果回归模型含有截距项，则假定 5.5 意味着扰动项 ε_i 无自相关。

证明：根据假定 5.5，$\boldsymbol{g}_i$ 为鞅差分序列，故

$$\mathrm{E}(\boldsymbol{g}_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = \mathrm{E}(\boldsymbol{x}_i \varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = 0$$

因为模型含有截距项，故向量 $\boldsymbol{g}_i \equiv \boldsymbol{x}_i \varepsilon_i$ 的第一个元素为 ε_i。因此，$\mathrm{E}(\varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1) = 0$。由于 $\{\varepsilon_{i-1}, \cdots, \varepsilon_1\} \subset \{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1\}$（前者是后者的子集，故前者的信息完全包含于后者之中），根据迭代期望定律可得

$$\begin{aligned} \mathrm{E}(\varepsilon_i \mid \varepsilon_{i-1}, \cdots, \varepsilon_1) &= \mathrm{E}_{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1}[\mathrm{E}(\varepsilon_i \mid \varepsilon_{i-1}, \cdots, \varepsilon_1) \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1] \\ &= \mathrm{E}_{\boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1}\underbrace{[\mathrm{E}(\varepsilon_i \mid \boldsymbol{g}_{i-1}, \cdots, \boldsymbol{g}_1)]}_{=0} = 0 \end{aligned} \tag{8.8}$$

因此，ε_i 均值独立于 $(\varepsilon_{i-1}, \cdots, \varepsilon_1)$，故扰动项 ε_i 无自相关。

根据第 7 章，异方差稳健的协方差矩阵 $\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{S}}\boldsymbol{S}_{XX}^{-1}$ 为夹心估计量，其中 $\boldsymbol{S}_{XX} \equiv \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$，$\hat{\boldsymbol{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$。异方差自相关稳健的协方差矩阵也是夹心估计量，其形式为 $\boldsymbol{S}_{XX}^{-1}\hat{\boldsymbol{Q}}\boldsymbol{S}_{XX}^{-1}$（两侧“面包”仍为 $\boldsymbol{S}_{XX}^{-1}$，但中间的“菜”变为 $\hat{\boldsymbol{Q}}$），其中

$$\hat{\boldsymbol{Q}} = \hat{\boldsymbol{S}} + \frac{1}{n}\sum_{j=1}^{p}\sum_{t=j+1}^{n}\left(1-\frac{j}{p+1}\right)e_t e_{t-j}(\boldsymbol{x}_t\boldsymbol{x}_{t-j}' + \boldsymbol{x}_{t-j}\boldsymbol{x}_t') \tag{8.9}$$

其中，p 为自相关的阶数，也称为“截断参数”(truncation parameter)。一般建议取 $p = n^{1/4}$ 或 $p = 0.75n^{1/3}$，然后取整数。公式(8.9)看似十分复杂，为了直观地理解它，考虑最简单的一元回归情形

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n \tag{8.10}$$

则 β_1 的 OLS 估计量为

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})[\beta_1(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon})]}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{8.11}$$

其中，由于 $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}$，故 $y_i - \bar{y} = \beta_1(x_i-\bar{x}) + \varepsilon_i - \bar{\varepsilon}$。因此，抽样误差

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{8.12}$$

其中，$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i - \frac{1}{n}\bar{\varepsilon}\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{=0}$。记 $v_i \equiv (x_i-\bar{x})\varepsilon_i$，则在大样本中，$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n}v_i}{\sigma_x^2}$，其中 σ_x^2 为 x_i 的方差。故在大样本中

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}v_i\right)}{(\sigma_x^2)^2} \tag{8.13}$$

首先考虑 $n=2$ 的最简单情形，则上式的分子为

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}v_i\right) &= \mathrm{Var}\left[\frac{1}{2}(v_1+v_2)\right] = \frac{1}{4}[\mathrm{Var}(v_1) + \mathrm{Var}(v_2) + 2\mathrm{Cov}(v_1,v_2)] \\ &= \frac{1}{2}\sigma_v^2 + \frac{1}{2}\rho_1\sigma_v^2 = \frac{1}{2}\sigma_v^2(1+\rho_1) \equiv \frac{1}{2}\sigma_v^2 f_2\end{aligned} \tag{8.14}$$

其中，$\sigma_v^2 \equiv \mathrm{Var}(v_i)$，$\rho_1 \equiv \mathrm{corr}(v_1, v_2)$ 为一阶自相关系数，而 $f_2 \equiv (1+\rho_1)$ 是一个修正系数。如果不存在自相关，$\rho_1 = 0$，$f_2 = 1$，则 $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}v_i\right) = \frac{1}{2}\sigma_v^2$，就得到通常的方差公式。在存在自相关的数据中，$\rho_1 \neq 0$，故方差的公式有所不同。考虑样本容量为 n 的一般情况，则 $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}v_i\right) = \frac{1}{n}\sigma_v^2 f_n$，其中 $f_n \equiv 1 + 2\sum_{j=1}^{n-1}\left(\frac{n-j}{n}\right)\rho_j$ 为对应于样本容量为 n 的修正系数[①]，而 ρ_j 为 j 阶自相关系数；因此

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_v^2}{n(\sigma_x^2)^2}\cdot f_n \tag{8.15}$$

上式是普通方差公式的 f_n 倍(如果自相关较强，则 f_n 可能较大)。如果知道 f_n，则可以直接

① f_n 的这个表达式由 $\mathrm{Var}[(v_1+v_2+\cdots+v_n)/n]$ 展开整理而来。f_n 可以看成是各阶自相关系数的加权平均。

计算上式。但由于 f_n 包含未知的自相关系数 ρ_j，故需要对其进行估计，比如 $\hat{f}_n \equiv 1 + 2\sum_{j=1}^{n-1}\left(\frac{n-j}{n}\right)\hat{\rho}_j$，其中 $\hat{\rho}_j$ 为 j 阶样本自相关系数。但此式中待估计的参数 $(\rho_1, \cdots, \rho_{n-1})$ 太多了，而且其数目随着样本容量 n 同步增长，故估计误差也随着样本容量增长，导致此估计量不一致。另一个极端做法是，仅考虑前几阶自相关系数（比如，只考虑一阶自相关系数 ρ_1）；但这样的估计量也不一致，因为它忽略了高阶自相关。正确的做法是，包括足够多阶数的自相关系数，并让此阶数 p 随着样本容量的增长而增长。一般建议取 $p = n^{1/4}$ 或 $p = 0.75n^{1/3}$，即公式（8.9）中的截断参数。由于 HAC 标准误取决于截断参数 p，故在实践中，建议使用不同的截断参数，以考察 HAC 标准误是否对截断参数的取值敏感。

2. 使用“OLS + 聚类稳健的标准误”

如果样本观测值可以分为不同的“聚类”（clusters），在同一聚类里的观测值互相相关，而不同聚类之间的观测值不相关，这种样本称为“聚类样本”（cluster sample）。比如，在 Nerlove（1963）对美国电力企业的研究中，同一个州的电力企业可能受到相同州政策的影响而自相关，但不同州之间的电力企业可能不相关[①]。此时，“州”（state）被称为“聚类变量”（cluster variable）[②]。又比如，如果以全班同学为样本，则聚类变量可能是宿舍或专业。

如果将观测值按聚类的归属顺序排列，则扰动项的协方差矩阵为“块对角”（block diagonal）。此时，仍可用 OLS 来估计系数，但需使用“聚类稳健的标准误”（cluster robust standard error）。假设样本容量为 N，包括 M 个聚类，其中第 j 个聚类包含 M_j 位个体。记第 j 个聚类中第 i 位个体的解释变量为 $\boldsymbol{x}_{ij}$，残差为 e_{ij}，然后定义 $\boldsymbol{u}_j \equiv \sum_{i=1}^{M_j} e_{ij}\boldsymbol{x}_{ij}$，则聚类稳健的协方差矩阵可以写为

$$\frac{N-1}{N-K}\frac{M}{M-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{j=1}^{M}\boldsymbol{u}_j'\boldsymbol{u}_j\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{8.16}$$

其中，$\frac{N-1}{N-K}\frac{M}{M-1}$ 为对自由度的调整。由此可见，聚类稳健的标准误也是夹心估计量，只不过夹在两个“面包”$\boldsymbol{S}_{XX}^{-1}$ 之间的“菜”略为复杂。而且，在推导上式的过程中并未假定同方差，故聚类稳健的标准误也是异方差稳健的，即在异方差与组内自相关的情况下依然成立。使用聚类稳健标准误的前提是，聚类中的个体数 M_j 较少，而聚类数很多（$M \to \infty$）；此时，聚类稳健标准误是真实标准误的一致估计。

在处理面板数据时，经常使用聚类稳健的标准误（参见第 15 章）。

3. 使用可行广义最小二乘法（FGLS）

为了使用 FGLS，首先必须估计扰动项的协方差阵 $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})$。为了减少待估计的参数，通常假设扰动项为一阶自回归形式：

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad |\rho| < 1, \quad u_t \text{ 为白噪声} \tag{8.17}$$

记扰动项的 j 阶协方差 $\rho_j \equiv \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-j}|\boldsymbol{X})$，则

---

① 事实上，由于 Nerlove（1963）的部分数据以“州”变量来替代“企业”变量（比如，只有“州”这个层次的工资与燃料价格数据，而没有每个企业的相应数据），故同一州的数据必然存在很强的相关性。

② 但 Nerlove（1963）的数据并未提供每个电力企业属于哪个州的信息。

$$\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\begin{pmatrix}\rho_0 & \rho_1 & \cdots & \rho_{n-1}\\ \rho_1 & \rho_0 & \cdots & \rho_{n-2}\\ \vdots & \vdots & & \vdots\\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_0\end{pmatrix} \tag{8.18}$$

容易证明(参见第5章),$\rho_0=\sigma^2=\mathrm{Var}(\varepsilon_t)=\dfrac{\sigma_u^2}{1-\rho^2}$,其中 $\sigma_u^2\equiv\mathrm{Var}(u_t)$。

而 $\rho_1=\rho\sigma^2$,故$\dfrac{\rho_1}{\rho_0}=\dfrac{\rho\sigma^2}{\sigma^2}=\rho$ 为一阶自相关系数;$\rho_2=\rho^2\sigma^2,\cdots,\rho_{n-1}=\rho^{n-1}\sigma^2$,故

$$\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\begin{pmatrix}1 & \rho & \cdots & \rho^{n-1}\\ \rho & 1 & \cdots & \rho^{n-2}\\ \vdots & \vdots & & \vdots\\ \rho^{n-1} & \rho^{n-2} & \cdots & 1\end{pmatrix}\equiv\sigma^2\boldsymbol{V} \tag{8.19}$$

因此,只要估计唯一的参数 ρ,就可以使用 FGLS 了。Stata 默认的估计方法为使用 OLS 残差进行辅助回归,$e_t=\hat{\rho}e_{t-1}+error_t$。也可以通过残差一阶自相关系数 $\hat{\rho}=\dfrac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=1}^{n}e_t^2}$,或 $\hat{\rho}=1-\dfrac{\mathrm{DW}}{2}$来估计 ρ。知道 ρ 后,就可以得到 $\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})\equiv\sigma^2\boldsymbol{V}$,并将 $\boldsymbol{V}$ 的逆矩阵分解为$\boldsymbol{V}^{-1}=\boldsymbol{C}'\boldsymbol{C}$。可以证明:

$$\boldsymbol{C}=\frac{1}{\sqrt{1-\rho^2}}\begin{pmatrix}\sqrt{1-\rho^2} & 0 & \cdots & 0 & 0\\ -\rho & 1 & \cdots & 0 & 0\\ 0 & -\rho & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & -\rho & 1\end{pmatrix} \tag{8.20}$$

以 $\sqrt{1-\rho^2}\boldsymbol{C}$ 左乘原模型,并定义 $\tilde{\boldsymbol{y}}\equiv\sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{y},\tilde{\boldsymbol{X}}\equiv\sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{X},\tilde{\boldsymbol{\varepsilon}}\equiv\sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{\varepsilon}$,则变换后的扰动项 $\tilde{\boldsymbol{\varepsilon}}$ 满足球型扰动项的假设,故高斯-马尔可夫定理成立(因为这种变换是 GLS 的一个特例):

$$\tilde{\boldsymbol{y}}\equiv\sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{y}=\begin{pmatrix}\sqrt{1-\rho^2} & 0 & \cdots & 0 & 0\\ -\rho & 1 & \cdots & 0 & 0\\ 0 & -\rho & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & -\rho & 1\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix}\sqrt{1-\rho^2}y_1\\ y_2-\rho y_1\\ \vdots\\ y_n-\rho y_{n-1}\end{pmatrix} \tag{8.21}$$

$$\tilde{\boldsymbol{X}}\equiv\sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{X}=\begin{pmatrix}\sqrt{1-\rho^2} & 0 & \cdots & 0 & 0\\ -\rho & 1 & \cdots & 0 & 0\\ 0 & -\rho & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & -\rho & 1\end{pmatrix}\begin{pmatrix}x_{11} & \cdots & x_{1K}\\ x_{21} & \cdots & x_{2K}\\ \vdots & & \vdots\\ x_{n1} & \cdots & x_{nK}\end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{1-\rho^2}x_{11} & \cdots & \sqrt{1-\rho^2}x_{1K} \\ x_{21}-\rho x_{11} & \cdots & x_{2K}-\rho x_{1K} \\ \vdots & & \vdots \\ x_{n1}-\rho x_{n-1,1} & \cdots & x_{nK}-\rho x_{n-1,K} \end{pmatrix} \tag{8.22}$$

$$\tilde{\boldsymbol{\varepsilon}} \equiv \sqrt{1-\rho^2}\boldsymbol{C}\boldsymbol{\varepsilon} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ 0 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}\varepsilon_1 \\ \varepsilon_2-\rho\varepsilon_1 \\ \vdots \\ \varepsilon_n-\rho\varepsilon_{n-1} \end{pmatrix} \tag{8.23}$$

可以写出每个观测值(个体)的回归方程:

$$\begin{aligned} \sqrt{1-\rho^2}y_1 &= \sqrt{1-\rho^2}\beta_1 + \sqrt{1-\rho^2}\beta_2 x_{12} + \cdots + \sqrt{1-\rho^2}\beta_K x_{1K} + \tilde{\varepsilon}_1 \\ y_2-\rho y_1 &= (1-\rho)\beta_1 + \beta_2(x_{22}-\rho x_{12}) + \cdots + \beta_K(x_{2K}-\rho x_{1K}) + \tilde{\varepsilon}_2 \\ &\cdots\cdots\cdots\cdots \\ y_n-\rho y_{n-1} &= (1-\rho)\beta_1 + \beta_2(x_{n2}-\rho x_{n-1,2}) + \cdots + \beta_K(x_{nK}-\rho x_{n-1,K}) + \tilde{\varepsilon}_n \end{aligned} \tag{8.24}$$

注意到(8.24)中第一个方程的形式与后面$(n-1)$个方程的形式不同。用 OLS 估计这个变换后的模型,即为"Prais-Winsten 估计法"(Prais and Winsten,1954,简记 PW)。如果为了计算方便而将第一个方程(即第一个观测数据)删去,则称为"Cochrane-Orcutt 估计法"(Cochrane and Orcutt,1949,简记 CO)。该方法有一个更简洁的推导过程。原模型为

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t \tag{8.25}$$

将上式滞后一期,然后方程两边同时乘以 ρ 得

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{t-1,2} + \cdots + \rho\beta_K x_{t-1,K} + \rho\varepsilon_{t-1} \tag{8.26}$$

将方程(8.25)减去方程(8.26)可得

$$y_t-\rho y_{t-1} = (1-\rho)\beta_1 + \beta_2(x_{t2}-\rho x_{t-1,2}) + \cdots + \beta_K(x_{tK}-\rho x_{t-1,K}) + \varepsilon_t - \rho\varepsilon_{t-1} \tag{8.27}$$

其中,$t=2,\cdots,K$。显然,新扰动项 $\varepsilon_t-\rho\varepsilon_{t-1}=u_t$ 满足球型扰动项的古典假定(根据假定,u_t 为白噪声)。因此,这一方法也被称为"准差分法"(quasi differences)。在实际操作中,常使用迭代法,即首先用 OLS 估计原模型,然后作辅助回归得到 $\hat{\rho}^{(1)}$(对 ρ 的第一轮估计),再用 $\hat{\rho}^{(1)}$ 进行 FGLS 估计,然后使用新的残差估计 $\hat{\rho}^{(2)}$(对 ρ 的第二轮估计),再用 $\hat{\rho}^{(2)}$ 进行 FGLS 估计,以此类推,直至收敛(即相邻两轮的 ρ 与 $\boldsymbol{\beta}$ 估计值之差足够小)。Stata 会把迭代收敛的过程显示出来。在大样本中,是否使用迭代法并无差别,但在有限样本中,迭代法通常比两步法更好些。

使用 FGLS 对自相关进行处理,如果对自相关系数的估计较准确,而且满足严格外生性的假定,则 FGLS 比 OLS 更有效率。然而,如果不满足严格外生性,而仅仅满足前定解释变量的假定,则 FGLS 可能是不一致的,尽管 OLS 依然一致。比如,在使用准差分法时,变换后的新扰动项为$(\varepsilon_t-\rho\varepsilon_{t-1})$,而新解释变量为$(x_{tk}-\rho x_{t-1,k})$,二者可能存在相关性,从而导致不一致的估计。总之,FGLS 不如 OLS 稳健。

4. 修改模型设定

在许多情况下,存在自相关的深层原因是模型本身的设定有误,比如,遗漏了自相关的解释变量;或将动态模型(即解释变量中包含被解释变量的滞后值)误设为静态模型,而后者也可以视为遗漏了解释变量。

例如，假设真实模型为 $y_t = \rho y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_t$。由于 y_t 是 y_{t-1} 的函数，故 $\{y_t\}$ 存在自相关。假设这个模型被错误地估计成 $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \underbrace{[\rho y_{t-1} + \varepsilon_t]}_{=v_t}$，即被解释变量的滞后项 ρy_{t-1} 被纳入到扰动项 v_t 中，则会导致扰动项 $\{v_t\}$ 自相关，因为 $\{y_t\}$ 存在自相关。这个例子也说明，对于时间序列数据中存在的自相关，有时可以通过引入被解释变量的滞后值来消除。总之，对于由于模型设定误差而导致的自相关最好从改进模型设定着手解决，而不是单纯地使用 FGLS。

8.5 处理自相关的 Stata 命令及实例

1. 时间序列算子

为了在 Stata 中使用时间序列算子(time-series operator)，首先要定义时间变量(必须是时间序列数据或面板数据，才能定义时间变量)。假设时间变量为“year”，则可使用如下命令，

```
. tsset year
```

其中，“tsset”表示“time series set”，它告诉 Stata，该数据集为时间序列，其时间变量为“year”。Stata 提供四个不同的时间序列算子，即滞后(lag)、前移(lead, forward)、差分(difference)、季节差分(seasonal difference)，分别以“L.,F.,D.,S.”来表示(可以小写)。

最常用的为滞后算子。一阶滞后算子为“L.”，即 L.$x_t = x_{t-1}$；二阶滞后算子为“L2.”，即 L2.$x_t = x_{t-2}$，以此类推。如果要同时表示一阶至四阶滞后，可简写为“L(1/4).”，即 L(1/4).$x_t = (x_{t-1}x_{t-2}x_{t-3}x_{t-4})$；比如，命令“reg y L.x L2.x L3.x L4.x”可简写为“reg y L(1/4).x”。类似地，“L(0/1).(x y)”表示 L(0/1).$(x_t\ y_t) = (x_t\ x_{t-1}\ y_t\ y_{t-1})$，其中“0”表示零阶滞后，即当前值。

一阶前移算子为“F.”，即 F.$x_t = x_{t+1}$；二阶前移算子为“F2.”，即 F2.$x_t = x_{t+2}$，以此类推。一阶差分算子为“D.”，即 D.$x_t = \Delta x_t = x_t - x_{t-1}$；二阶差分算子为“D2.”，即 D2.$x_t = \Delta(\Delta x_t) = \Delta(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$(二阶差分为一阶差分的差分)。一阶季节差分算子为“S.”，即 S.$x_t = x_t - x_{t-1}$(等价于一阶差分算子“D.”)；二阶季节差分算子为“S2.”，即 S2.$x_t = x_t - x_{t-2}$；以此类推。对于季度数据，如果要计算本季度与去年同季度的同比变化，则可用“S4.”，即 S4.$x_t = x_t - x_{t-4}$；对于月度数据，则可用“S12.”来计算同比变化，即 S12.$x_t = x_t - x_{t-12}$。

以上时间序列算子可以混合使用。比如，“LD.”表示一阶差分的滞后值，“DL.”表示滞后值的一阶差分，二者实际上是等价的，因为 LD.x_t = L.$(x_t - x_{t-1}) = x_{t-1} - x_{t-2}$ = D.x_{t-1} = DL.x_t。更多有关时间序列算子的说明，参见“help tsvarlist”。

2. 画残差图

在作完回归后，假设将残差记为 e1：

```
scatter e1 L.e1   (L.e1 为滞后一期的残差)
ac e1             (看残差自相关图)
pac e1            (看残差偏自相关图①)
```

① 有关“偏自相关”，参见第 20 章。

3. BG 检验

```
estat bgodfrey              (默认 p = 1)
estat bgodfrey,lags(p)
estat bgodfrey,nomiss0      (使用不添加 0 的 BG 检验)
```

如何确定 p 呢？一个简单方法是，看自相关图（用 Stata 命令 ac，在 95% 的阴影置信区域以外的自相关阶数为显著地不等于 0），或偏自相关图。或者设定一个较大的 p 值，作回归 $e_t = \alpha_0 + \alpha_1 x_{t1} + \cdots + \alpha_K x_{tK} + \rho_1 e_{t-1} + \cdots + \rho_p e_{t-p} + error_t$，然后看最后一个系数 ρ_p 的显著性；如果 ρ_p 不显著，则考虑滞后 $(p-1)$ 期，以此类推。

4. Ljung-Box Q 检验

```
reg y x1 x2 x3
predict e1,resid     (将回归残差命名为 e1)
wntestq e1           (使用 Stata 提供的默认滞后期)
wntestq e1,lags(p)   (使用自己指定的滞后期)
```

其中，wntestq 指的是“white noise test Q”，因为白噪声是没有自相关的。

5. DW 检验

作完 OLS 回归后可使用命令“estat dwatson”显示 DW 统计量。

6. HAC 稳健标准误

```
newey y x1 x2 x3,lag(p)            (HAC 标准误，必须指定滞后阶数 p)
reg y x1 x2 x3,cluster(state)      (聚类稳健标准误，假设聚类变量为 state)
```

7. 处理一阶自相关的 FGLS

```
prais y x1 x2 x3          (使用默认的 PW 估计法)
prais y x1 x2 x3,corc     (使用 CO 估计法)
```

下面以 Hildreth and Lu(1960)对冰淇淋需求函数的研究为例[①]。数据集 icecream. dta 包含了下列变量的 30 个月度时间序列数据：consumption（人均冰淇淋消费量），income（平均家庭收入），price（冰淇淋价格），temp（平均华氏气温），time（时间）。

常识告诉我们，气温越高，则对冰淇淋的需求越大。为此，首先看一下冰淇淋的消费量与气温的时间序列趋势图（如图 8.2 所示）：

```
. use icecream.dta,clear
. tsset time
. graph twoway connect consumption temp100 time, msymbol (circle)
msymbol(triangle)
```

其中，变量 temp100 为 temp/100，选择项“msymbol(circle) msymbol(triangle)”表示“图标”（marker symbol）分别为圆圈与三角形。运行命令结果如图 8.2。

图 8.2 显示，冰淇淋消费量与温度正相关。考虑以下线性回归模型：

$$\text{consumption}_t = \beta_0 + \beta_1 \text{ temp}_t + \beta_2 \text{ price}_t + \beta_3 \text{ income}_t + \varepsilon_t \tag{8.28}$$

进行 OLS 回归，得

```
. reg consumption temp price income
```

① 此例来自 Verbeek(2004)。

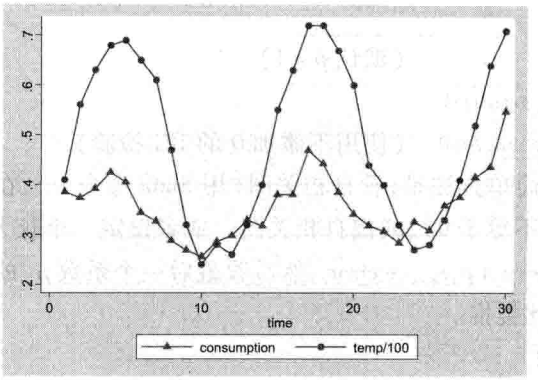


图 8.2 冰淇淋消费与气温的时间趋势

Source	SS	df	MS	Number of obs =	30
				F(3, 26) =	22.17
Model	.090250523	3	.030083508	Prob > F =	0.0000
Residual	.035272835	26	.001356647	R-squared =	0.7190
				Adj R-squared =	0.6866
Total	.125523358	29	.004328392	Root MSE =	.03683

consumption	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
temp	.0034584	.0004455	7.76	0.000	.0025426	.0043743
price	-1.044413	.834357	-1.25	0.222	-2.759458	.6706322
income	.0033078	.0011714	2.82	0.009	.0008999	.0057156
_cons	.1973149	.2702161	0.73	0.472	-.3581223	.752752

OLS 估计结果显示,“气温”(temp)与“收入”(income)均在 1% 的水平上显著地不等于 0,而“价格”(price)与“常数项”(_cons)则不显著。由于这是时间序列数据,我们怀疑其扰动项存在自相关。首先计算残差(记为 e1),及其滞后值(记为 e2),然后画残差与残差滞后的散点图(如图 8.3 所示):

```
. predict e1,res
. g e2 =L.e1
. twoway (scatter e1 e2)(lfit e1 e2)
```

其中,“lfit”表示“线性拟合”(linear fit),即画出 e1 与 e2 的拟合回归线,结果如图 8.3。

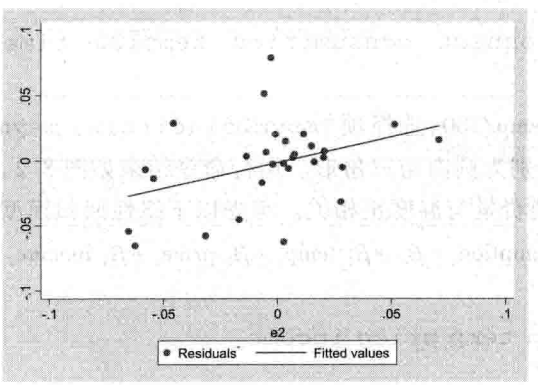


图 8.3 残差与残差滞后的散点图

图 8.3 显示，扰动项可能存在正的自相关，即正的扰动项后面更可能跟着正的扰动项，而负的扰动项后面更可能跟着负的扰动项。接着看自相关图（结果如图 8.4）与偏自相关图（结果如图 8.5）。

```
. ac e1
```

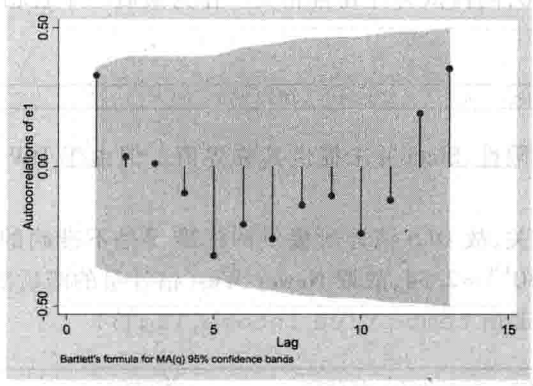


图 8.4 自相关图

```
. pac e1
```

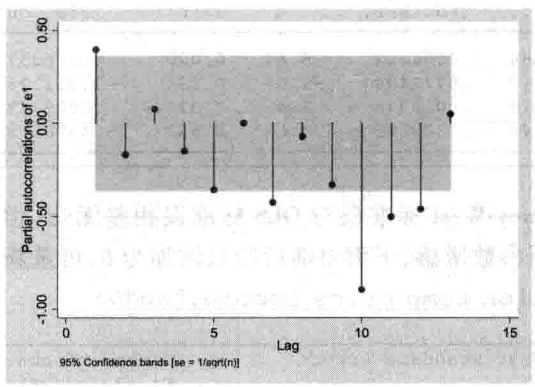


图 8.5 偏自相关图

以上两图显示，自相关的形式主要为一阶自相关（统计量落在 95% 的阴影置信区间之外或附近，表明一阶自相关显著不为 0），大致可以忽略高阶自相关。下面进行正式的 BG 检验：

```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	4.237	1	0.0396

H0: no serial correlation

BG 检验的 p 值为 0.039 6，即可以在 5% 的显著性水平上拒绝“无自相关”原假设，而认为存在自相关。再来看 Q 检验：

```
. wntestq e1
```

```
Portmanteau test for white noise
---------------------------------------
 Portmanteau (Q) statistic =     26.1974
 Prob > chi2(13)           =      0.0160
```

Q 检验的 p 值为 0.016，同样认为存在自相关。作为最后一个自相关检验，计算 DW 统计量：

```
. estat dwatson
```

```
Durbin-Watson d-statistic(  4,    30) =  1.021169
```

由于 DW 统计量的局限性，Stata 并未提供其临界值。但由于 DW = 1.02，离 2 很远，故可大致判断存在正自相关。

由于扰动项存在自相关，故 OLS 估计所提供的标准误是不准确的，应使用异方差自相关稳健的标准误。由于 $n^{1/4}=30^{1/4}\approx 2.34$，故取 Newey-West 估计量的滞后阶数为 $p=3$：

```
. newey consumption temp price income,lag(3)
```

```
Regression with Newey-West standard errors              Number of obs   =         30
maximum lag: 3                                          F(  3,    26)   =      27.63
                                                        Prob > F        =     0.0000

------------------------------------------------------------------------------
             |             Newey-West
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0034584   .0004002     8.64   0.000     .0026357    .0042811
       price |  -1.044413   .9772494    -1.07   0.295    -3.053178    .9643518
      income |   .0033078   .0013278     2.49   0.019     .0005783    .0060372
       _cons |   .1973149   .3378109     0.58   0.564    -.4970655    .8916952
------------------------------------------------------------------------------
```

从上表可以看出，Newey-West 标准误与 OLS 标准误相差无几（但略大）。为了考察 Newey-West 标准误是否对于截断参数敏感，下面将滞后阶数增加为 6，再重新估计。

```
. newey consumption temp price income,lag(6)
```

```
Regression with Newey-West standard errors              Number of obs   =         30
maximum lag: 6                                          F(  3,    26)   =      52.97
                                                        Prob > F        =     0.0000

------------------------------------------------------------------------------
             |             Newey-West
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0034584   .0003504     9.87   0.000     .0027382    .0041787
       price |  -1.044413   .9821798    -1.06   0.297    -3.063313    .9744864
      income |   .0033078     .00132     2.51   0.019     .0005945     .006021
       _cons |   .1973149   .3299533     0.60   0.555    -.4809139    .8755437
------------------------------------------------------------------------------
```

从上表可以看出，无论截断参数为 3 还是 6，Newey-West 标准误变化不大，比较稳健。如果使用“OLS + HAC 标准误”来处理自相关，那么做到这里就可以了。

由于存在自相关，OLS 不再是 BLUE，故可考虑使用可行广义最小二乘法（FGLS），对模型进行转换、重新估计。首先使用 CO 估计法：

```
. prais consumption temp price income,corc
```

```
Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.4006
Iteration 2:  rho = 0.4008
Iteration 3:  rho = 0.4009
Iteration 4:  rho = 0.4009
Iteration 5:  rho = 0.4009
Iteration 6:  rho = 0.4009
Iteration 7:  rho = 0.4009

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  3,    25) =   15.40
       Model |  .047040596     3  .015680199            Prob > F      =  0.0000
    Residual |  .025451894    25  .001018076            R-squared     =  0.6489
-------------+------------------------------           Adj R-squared =  0.6068
       Total |  .072492491    28  .002589018            Root MSE      =  .03191

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0035584   .0005547     6.42   0.000      .002416    .0047008
       price |  -.8923963   .8108501    -1.10   0.282    -2.562373    .7775807
      income |   .0032027   .0015461     2.07   0.049     .0000186    .0063869
       _cons |   .1571479   .2896292     0.54   0.592    -.4393546    .7536504
-------------+----------------------------------------------------------------
         rho |   .4009256
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.021169
Durbin-Watson statistic (transformed) 1.548837
```

使用 CO 估计法得到的系数估计值与 OLS 比较接近。上表的最后一行显示，经过模型转换后的 DW 值改进为 1.55。然后使用 PW 估计法：

```
. prais consumption temp price income,nolog
```

```
Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   14.35
       Model |  .04494596     3  .014981987            Prob > F      =  0.0000
    Residual |  .027154354    26  .001044398            R-squared     =  0.6234
-------------+------------------------------           Adj R-squared =  0.5799
       Total |  .072100315    29  .002486218            Root MSE      =  .03232

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0029541   .0007109     4.16   0.000     .0014929    .0044152
       price |  -1.048854    .759751    -1.38   0.179    -2.610545    .5128361
      income |  -.0008022   .0020458    -0.39   0.698    -.0050074    .0034029
       _cons |   .5870049   .2952699     1.99   0.057    -.0199311    1.193941
-------------+----------------------------------------------------------------
         rho |   .8002264
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.021169
Durbin-Watson statistic (transformed) 1.846795
```

其中，选择项“nolog”表示，不显示迭代过程。上表的结果显示，虽然使用 PW 估计法使得 DW 统计量进一步改进为 1.85，但对收入（income）的系数估计值却变为负数（-0.000 8）。尽管它只是绝对值很小的负数，且在统计上不显著，但由于 PW 估计法使得收入的系数估计值与理论预

期相反[①],似乎 PW 估计法反而不如 OLS 稳健。

前面提到,自相关的存在可能是由于模型设定不正确。为此,考虑在解释变量中加入气温(temp)的滞后值,然后进行 OLS 回归:

```
. reg consumption temp L.temp price income
```

Source	SS	df	MS	Number of obs =	29
				F(4, 24) =	28.98
Model	.103387183	4	.025846796	Prob > F =	0.0000
Residual	.021406049	24	.000891919	R-squared =	0.8285
				Adj R-squared =	0.7999
Total	.124793232	28	.004456901	Root MSE =	.02987

consumption	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
temp						
--.	.0053321	.0006704	7.95	0.000	.0039484	.0067158
L1.	-.0022039	.0007307	-3.02	0.006	-.0037119	-.0006959
price	-.8383021	.6880205	-1.22	0.235	-2.258307	.5817025
income	.0028673	.0010533	2.72	0.012	.0006934	.0050413
_cons	.1894822	.2323169	0.82	0.423	-.2899963	.6689607

结果显示,气温的滞后项(L.temp)在 1% 的水平上显著地不等于 0,但符号为负(-0.002 2);而当期气温仍然显著地为正(0.005 3)。这可能意味着,当气温上升时,对冰淇淋的需求上升,但不会在当月全部消费完,而是增加冰箱中的冰淇淋库存,导致下期对冰淇淋的开支下降[②]。使用 BG 检验来检验自相关:

```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	0.120	1	0.7292

H0: no serial correlation

由于 p 值为 0.73,故可以放心地接受"无自相关"的原假设:

```
. estat dwatson
```

```
Durbin-Watson d-statistic(  5,    29) =  1.582166
```

DW 值也改进为 1.58。因此,修改模型设定,加入气温的滞后项后,扰动项基本上不再存在自相关。

究竟应该使用以上哪种模型,在一定程度上取决于研究者的判断。一种较好的做法是,在研究文献中同时列出以上各种模型的结果,借以说明系数估计值与标准误的稳健性(不依估计方法的改变而剧烈变化),从而给文献的阅读者自己做出判断的机会。

① 一般来说,收入效应应该为正。

② 变量 consumption 指的是,用于购买冰淇淋的每月开支,而非消费者实际每月吃多少冰淇淋。

习　题

8.1 证明当 $n\to\infty$ 时，Box-Pierce Q 与 Ljung-Box Q 统计量之差依概率收敛于 0。[提示：定义 $x_n \equiv \left(\left(\sqrt{n}\hat{\rho}_1\right)^2 \left(\sqrt{n}\hat{\rho}_2\right)^2 \cdots \left(\sqrt{n}\hat{\rho}_p\right)^2\right)'$，并把两个统计量之差写成 $\boldsymbol{a}_n'\boldsymbol{x}_n$ 的形式。]

8.2 考虑在只有一阶自相关的情况下使用 GLS，对于变换后的新扰动项 $\tilde{\boldsymbol{\varepsilon}} \equiv (\sqrt{1-\rho^2}\,\varepsilon_1 (\varepsilon_2 - \rho\varepsilon_1) \cdots (\varepsilon_n - \rho\varepsilon_{n-1}))'$，直接证明 $\tilde{\boldsymbol{\varepsilon}}$ 满足同方差的假定。

8.3 PW 估计法比 CO 估计法更有效率吗？为什么？

8.4 假设扰动项存在二阶自相关，即 $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + u_t$，其中 u_t 为白噪声。此时，还可以使用 CO 估计法吗？若可以，如何进行？

8.5 使用数据集 ukrates. dta，考察英国政府如何根据长期利率（$r20$）的变化来调整短期利率（rs），即货币政策反应函数。

（1）作如下回归：

$$\Delta rs_t = \alpha + \beta\Delta r20_{t-1} + \varepsilon_t \tag{8.29}$$

其中，$\Delta rs_t = rs_t - rs_{t-1}$，$\Delta r20_{t-1} = r20_{t-1} - r20_{t-2}$。

（2）画残差散点图 (e_{t-1}, e_t)；

（3）画残差的自相关图与偏自相关图，以显示其各阶自相关与偏自相关系数；

（4）用 BG 检验，检验扰动项是否存在自相关；

（5）用 Q 检验，检验扰动项是否存在自相关；

（6）计算 DW 统计量；

（7）使用异方差与自相关稳健的标准误（HAC）重新进行估计，将滞后期数设为 $n^{1/4}$；

（8）使用迭代式 PW 估计法进行 FGLS 估计；

（9）使用迭代式 CO 估计法进行 FGLS 估计。

第9章　模型设定与数据问题

如果模型设定(model specification)不当,如解释变量选择不当、测量误差、函数形式不妥等,则会出现"设定误差"(specification error),即模型本身的设定所带来的误差。另外,数据本身也可能存在问题,如多重共线性、对回归结果影响很大的极端数据等。本章将讨论这些问题导致的后果及处理方法。

9.1　遗漏变量

由于某些数据难以获得,遗漏变量现象几乎难以避免。假设真实的模型为

$$y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + \varepsilon_i \tag{9.1}$$

其中,x_1,x_2 可以是向量,且与扰动项 ε 不相关。而实际估计的模型(estimated model)为

$$y_i = x_{i1}'\beta_1 + u_i \tag{9.2}$$

对比方程(9.1)与(9.2)可知,遗漏变量(omitted variables)$x_{i2}'\beta_2$ 被归入新扰动项 $u_i = x_{i2}'\beta_2 + \varepsilon_i$ 中去了。考虑以下两种情形。

(1) 遗漏变量 x_{i2} 与解释变量 x_{i1} 不相关,即 $\mathrm{Cov}(x_{i1}, x_{i2}) = 0$。在这种情况下,扰动项 u_i 与解释变量 x_{i1} 不相关,根据大样本理论,OLS 依然可以一致地估计 β_1。但由于遗漏变量 $x_{i2}'\beta_2$ 被归入扰动项 u_i 中,这可能增大扰动项的方差,从而影响 OLS 估计的精确度。

(2) 遗漏变量 x_{i2} 与解释变量 x_{i1} 相关,即 $\mathrm{Cov}(x_{i1}, x_{i2}) \neq 0$。在这种情况下,根据大样本理论,OLS 不再是一致估计,称其偏差为"遗漏变量偏差"(omitted variable bias)。这种偏差在计量实践中较常见,成为某些实证研究的致命伤。比如,在研究教育投资回报时,个人的先天能力因为无法观测而被遗漏,但能力与受教育年限很可能存在正相关。

总之,存在遗漏变量本身并不要紧,你甚至可以说"只对 $\mathrm{E}(y|x_1)$ 感兴趣,故不把 x_2 放入解释变量中"。问题的关键是,被遗漏的变量不能与包括在方程内的解释变量相关。解决遗漏变量偏差的方法主要有:

(i) 加入尽可能多的控制变量(control variable);

(ii) 使用"代理变量"(proxy variable);

(iii) 工具变量法(第 10 章);

(iv) 使用面板数据(第 15 ~ 17 章);

(v) 随机实验与自然实验(第 18 章)。

第(i)种方法"加入尽可能多的控制变量"着眼于直接解决遗漏变量问题,即把遗漏的变量补上去。具体来说,首先从理论出发,列出所有可能对被解释变量有影响的变量,然后尽可能地去收集数据。如果有些相关变量确实无法获得(在实践中经常出现),则需要从理论上说明,遗漏变量不会与解释变量相关,或相关性很弱。

例　李宏彬等(2012)通过就业调查数据,研究"官二代"大学毕业生的起薪是否高于非官

二代。由于可能存在遗漏变量,该文包括了尽可能多的控制变量,比如年龄、性别、城镇户口、父母收入、父母学历、高考成绩、大学成绩、文理科、党员、学生会干部、兼职实习经历、拥有技术等级证书等。李宏彬等(2012,p. 1012)认为,“尽管我们无法利用自然实验来解决线性回归中存在的遗漏变量问题,但是我们可以通过控制大量可能影响工资的变量来降低可能存在的估计偏差。其中,最为重要的是,我们将高考成绩作为学生能力或智商的代理变量”。

当控制变量不可得时,可以考虑使用第(ii)种方法,即代理变量法。比如,在教育投资回归中,可以使用智商(IQ)来作为个人能力的代理变量。一个理想的代理变量应满足以下两个条件。

(1) 多余性(redundancy):即代理变量仅通过影响遗漏变量而作用于被解释变量。比如,“智商”仅通过对“能力”的作用来影响工资收入。换言之,假如有“能力”的数据,那么再引入“智商”作为解释变量就是多余的。

(2) 剩余独立性:遗漏变量中不受代理变量影响的剩余部分与所有解释变量均不相关。

命题 如果上述两个条件满足,则使用代理变量就能获得一致的估计量。

证明:假设真实模型为

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_K x_K+\gamma q+\varepsilon \tag{9.3}$$

其中,q 为不可观测的遗漏变量。假定 $\mathrm{Cov}(x_k,\varepsilon)=0,\forall k$,但遗漏变量 q 与某解释变量 x_m 相关($1\leqslant m\leqslant K$),即 $\mathrm{Cov}(x_m,q)\neq 0$,故用 OLS 估计(9.3)是不一致的。

假设找到了一个代理变量 z,满足

$$q=\delta_0+\delta_1 z+v,\quad \mathrm{Cov}(z,v)=0 \tag{9.4}$$

根据第一个条件(多余性),代理变量 z 只通过 q 对 y 发生作用,故在回归方程已经包含 q 的情况下,z 与 y 的扰动项 ε 不相关,即 $\mathrm{Cov}(z,\varepsilon)=0$。根据第二个条件,$q$ 的扰动项 v 与所有解释变量均不相关,即 $\mathrm{Cov}(x_k,v)=0,\forall k$。将 q 的表达式代入原模型(9.3)可得

$$y=(\beta_0+\gamma\delta_0)+\beta_1 x_1+\cdots+\beta_K x_K+\gamma\delta_1 z+(\gamma v+\varepsilon) \tag{9.5}$$

其中,$\gamma v+\varepsilon$ 为新的扰动项。容易证明新扰动项与所有解释变量均不相关,

$$\mathrm{Cov}(x_k,\gamma v+\varepsilon)=\gamma\underbrace{\mathrm{Cov}(x_k,v)}_{\text{condition 2}}+\underbrace{\mathrm{Cov}(x_k,\varepsilon)}_{\text{assumption}}=0+0=0\quad(\forall k) \tag{9.6}$$

$$\mathrm{Cov}(z,\gamma v+\varepsilon)=\gamma\underbrace{\mathrm{Cov}(z,v)}_{\text{assumption}}+\underbrace{\mathrm{Cov}(z,\varepsilon)}_{\text{condition 1}}=0+0=0 \tag{9.7}$$

因此,将代理变量引入回归方程后使用 OLS,就可以得到一致估计量。

在实践中,对于代理变量是否满足以上两个条件,通常只能做定性讨论,无法严格检验。如果使用不满足这两个条件的“不完美代理变量”(imperfect proxy variable),则仍会导致不一致的估计。

关于解决遗漏变量偏差的第(iii)~(v)种方法将在以后各章介绍。

总之,由于影响被解释变量的因素往往很多,而局限于数据的可获得性,故在任何实证研究中几乎总是存在遗漏变量。因此,一篇专业水准的实证论文几乎总是需要说明,它是如何在存在遗漏变量的情况下避免遗漏变量偏差的。如果无法令人信服地说明这一点,则其结果就是可疑的。

9.2 无关变量

与遗漏变量相反的情形是在回归方程中加入了与被解释变量无关的变量。假设真实模型为

$$y_i = x_{i1}'\beta_1 + \varepsilon_i \tag{9.8}$$

其中，$\mathrm{Cov}(x_{i1}, \varepsilon_i) = 0$。而实际估计的模型为

$$y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + (\varepsilon_i - \underbrace{x_{i2}'\beta_2}_{=0}) \tag{9.9}$$

其中，加入了与被解释变量 y 无关的解释变量 x_{i2}'。由于真实参数 $\beta_2 = 0$，故可以将模型写为 $y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + \varepsilon_i$，即扰动项仍是原来的 ε_i。由于 x_2 与 y 无关，故根据“无关变量”的定义，x_2 也与 y 的扰动项 ε 无关，即 $\mathrm{Cov}(x_{i2}, \varepsilon_i) = 0$。因此，扰动项 ε 与所有解释变量均无关，故 OLS 仍然是一致的，即 $\operatorname{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1$，$\operatorname{plim}_{n\to\infty} \hat{\beta}_2 = \beta_2 = 0$。

另一方面，引入无关变量后，由于受到无关变量的干扰，估计量 $\hat{\beta}_1$ 的方差一般会增大。总之，对于解释变量的选择最好遵循经济理论的指导。

9.3 建模策略：“由小到大”还是“由大到小”

“由小到大”（specific to general）的建模方式首先从最简单的小模型开始，然后逐渐增加解释变量。从理论上来说，这种方法的缺点是，小模型很可能存在遗漏变量偏差，这样系数估计量就不一致，t 检验、F 检验都将失效，因此很难确定该如何取舍变量。

与此相反，“由大到小”（general to specific）的建模方式[①]从一个尽可能大的模型开始，收集所有可能的解释变量，然后再逐步剔除不显著的解释变量。这样做虽然冒着包含无关变量的危险，但其危害性毕竟没有遗漏变量严重。然而，在实际操作上，常常很难找到所有与被解释变量相关的解释变量。

因此，在实证研究中，常常采用以上两种策略的折中方案。

9.4 解释变量个数的选择

好的经济理论应该尽可能地用简洁的模型来很好地描述复杂的经济现实。但这两个目标常常是矛盾的。在计量模型的设定上，加入过多的解释变量虽然可以提高模型的解释力（比如，拟合优度 R^2），但也牺牲了模型的简洁性（parsimony）。故需要在模型的解释力与简洁性之间找到一个最佳的平衡。在时间序列模型里，常常需要选择解释变量滞后的期数（比如，确定自回归模型的阶数）。更一般地，则要确定解释变量的个数。可供选择的权衡标准如下。

（1）校正可决系数 $\bar{R}^2$：选择解释变量的个数 K 以最大化 $\bar{R}^2$。

（2）“赤池信息准则”（Akaike Information Criterion，简记 AIC）[②]：选择解释变量的个数 K，使得以下目标函数最小化：

$$\min_K \mathrm{AIC} \equiv \ln(\boldsymbol{e}'\boldsymbol{e}/n) + \frac{2}{n}K \tag{9.10}$$

其中，右边第一项为对模型拟合度的奖励（减少残差平方和），而第二项为对解释变量过多的惩

① 为 David Hendry 所提倡，也被称为“LSE methodology”（意指其发源于 London School of Economics）。

② 参见 Akaike(1974)。

罚(解释变量个数 K 的增函数)。当 K 上升时,第一项下降而第二项上升。

(3)“贝叶斯信息准则”(Bayesian Information Criterion,简记 BIC)[①]或“施瓦茨信息准则”(Schwarz Information Criterion,简记 SIC 或 SBIC)[②]:选择解释变量的个数 K,使得以下目标函数最小化:

$$\min_K \text{BIC} \equiv \ln(\boldsymbol{e}'\boldsymbol{e}/n) + \frac{\ln n}{n}K \tag{9.11}$$

BIC 准则与 AIC 准则只有第二项有差别。一般来说,$\ln n > 2$(除非样本容量很小),故 BIC 准则对于解释变量过多的惩罚比 AIC 准则更为严厉。也就是说,BIC 准则更强调模型的简洁性。

(4)“汉南 - 昆信息准则”(Hannan-Quinn Information Criterion,简记 HQIC)[③]:选择解释变量的个数 K,使得以下目标函数最小化:

$$\min_K \text{HQIC} \equiv \ln(\boldsymbol{e}'\boldsymbol{e}/n) + \frac{\ln[\ln(n)]}{n}K \tag{9.12}$$

在时间序列模型中,常用信息准则来确定滞后阶数。比如,考虑以下 p 阶自回归模型(详见第 20 章),

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1, \cdots, T \tag{9.13}$$

其中,滞后阶数 p 待确定。可以证明,根据 BIC 或 HQIC 计算的 $\hat{p}$ 是真实参数 p 的一致估计量,即当样本容量 $T \to \infty$ 时,$\Pr(\hat{p} < p) \to 0$,$\Pr(\hat{p} = p) \to 1$,$\Pr(\hat{p} > p) \to 0$。然而,根据 AIC 计算的 $\hat{p}$ 却不是一致估计量,即使在大样本中也可能高估 p,虽然 $\Pr(\hat{p} < p) \to 0$,但 $\Pr(\hat{p} > p) \to c > 0$。证明参见附录。

在实践中,比较常用 AIC 与 BIC,不常用 HQIC。而且,虽然在大样本中 BIC 是一致估计,而 AIC 不是一致估计,但现实样本通常是有限的,而 BIC 准则可能导致模型过小(对解释变量过多的惩罚太严厉),故 AIC 准则依然很常用。

计算信息准则的 Stata 命令为

```
reg y x1 x2 x3
estat ic  (ic 表示 information criterion)
```

以数据集 icecream.dta(参见第 8 章)为例,考虑应该引入气温(temp)的几阶滞后项。

```
. use icecream.dta,clear
. quietly reg consumption temp price income
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	30	39.57876	58.61944	4	-109.2389	-103.6341

在以上模型中加入气温的一阶滞后项(L.temp),重新进行估计。

```
. qui reg consumption temp L.temp price income
. estat ic
```

① 此准则根据“贝叶斯方法”(参见第 31 章)推导而来。

② 参见 Schwarz (1978)。

③ 参见 Hannan and Quinn (1979)。

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	29	37.85248	63.41576	5	-116.8315	-109.995

从以上结果可以看出,增加解释变量 L.temp 后,AIC 与 BIC 都下降了。进一步,引入气温的二阶滞后项(L2.temp),

```
. qui reg consumption temp L.temp L2.temp price income
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	28	36.08382	61.12451	6	-110.249	-102.2558

结果显示,加入气温的二阶滞后项后,AIC 与 BIC 反而比仅包括气温的滞后项上升了。因此,只包含气温的滞后项可以达到 AIC 与 BIC 的最小值。这意味着,从信息准则的角度,应包含气温的一阶滞后项,但不该引入更高阶的气温滞后项。

9.5 对函数形式的检验

显然,很多经济关系是非线性的。因此,多元线性回归只能被看做是非线性经济关系的一阶线性近似。但是,二阶乃至高阶的非线性部分真的不重要吗?首先,如果回归方程中存在非线性项,则解释变量对被解释变量的边际效应将不再是常数。来看一个具体例子:

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\gamma x_1^2+\delta x_2 x_3+\varepsilon \tag{9.14}$$

分别计算各变量的边际效应如下,

$$\frac{\mathrm{E}(y)}{\partial x_1}=\beta_1+2\gamma x_1,\quad \frac{\mathrm{E}(y)}{\partial x_2}=\beta_2+\delta x_3,\quad \frac{\mathrm{E}(y)}{\partial x_3}=\beta_3+\delta x_2 \tag{9.15}$$

从方程(9.15)可以看出,x_1 的边际效应依赖于 x_1;x_2 的边际效应依赖于 x_3;而 x_3 的边际效应依赖于 x_2;这些边际效应都不是常数。反过来,如果怀疑变量的边际效应并非常数,则应考虑在回归方程中引入非线性项。

"Ramsey's RESET 检验"(Regression Equation Specification Error Test)的基本思想是,如果怀疑非线性项被遗漏了,那么就把非线性项引入方程,并检验其系数是否显著。

假设线性回归模型为 $y=\boldsymbol{x}'\boldsymbol{\beta}+\varepsilon$。回归后可得拟合值 $\hat{y}=\boldsymbol{x}'\boldsymbol{b}$。既然 $\hat{y}$ 是解释变量 $\boldsymbol{x}$ 的一个线性组合,$\hat{y}^2$ 中就包含了各解释变量二次项(含平方项与交叉项)的信息,而 $\hat{y}^3$ 中就包含了各解释变量三次项的信息,以此类推。考虑以下回归方程:

$$y=\boldsymbol{x}'\boldsymbol{\beta}+\delta_2\hat{y}^2+\delta_3\hat{y}^3+\delta_4\hat{y}^4+\mu \tag{9.16}$$

对原假设"$H_0:\delta_2=\delta_3=\delta_4=0$"作 F 检验。如果拒绝 H_0,则说明模型中应该有高次项;如果接受 H_0,就说明可以使用线性模型。RESET 检验的缺点是在拒绝 H_0 的情况下,并不提供具体遗漏哪些高次项的信息。

RESET 检验的 Stata 命令为

```
reg y x1 x2 x3
```

```
estat ovtest        (使用 ŷ², ŷ³, ŷ⁴ 作为非线性项)
estat ovtest,rhs    (使用解释变量的幂作为非线性项)
```

其中,"ovtest"表示"omitted variable test",因为遗漏高次项的后果类似于遗漏解释变量。比如,假设真实模型为 $y=\alpha+\beta x+(\gamma x^2+\varepsilon)$,$\text{Cov}(x,\varepsilon)=0$,但 γx^2 被遗漏。显然,$\text{Cov}(x,\gamma x^2+\varepsilon)=\gamma\text{Cov}(x,x^2)+\text{Cov}(x,\varepsilon)=\gamma\text{Cov}(x,x^2)\neq 0$。因此,遗漏高次项也会导致遗漏变量偏差。

另一模型设定检验为"连接检验"(link test),最早由 Tukey (1949)与 Pregibon (1979,1980)提出。此处的"连接"指的是,将解释变量与被解释变量连接在一起的函数形式是否正确。连接检验的步骤是,首先进行以下回归:

$$y=\delta_0+\delta_1\hat{y}+\delta_2\hat{y}^2+error \tag{9.17}$$

然后检验"$H_0:\delta_2=0$",即拟合值平方 $\hat{y}^2$ 的系数是否为 0。如果模型设定正确,则 $\hat{y}^2$ 不应对被解释变量还有解释力。因此,如果拒绝 $H_0:\delta_2=0$,则认为模型设定有误,可考虑加入非线性项或改变回归的函数形式(比如,取对数)。

连接检验的 Stata 命令为

```
reg y x1 x2 x3
linktest
```

关于如何确定回归方程的函数形式,最好从经济理论出发,通过经济模型的推导得到回归方程的具体形式。比如,通过对人力资本的理论研究可知,教育投资回报方程应该采用单对数形式,即被解释变量取对数,而解释变量为线性(不取对数)。在缺乏理论指导的情况下,可以先从线性模型出发,然后进行 RESET 或连接检验,看是否应加入非线性项。

以数据集 nerlove. dta 为例:

```
. use nerlove.dta,clear
. qui reg lntc lnq lnpl lnpk lnpf
```

首先进行连接检验。

```
. linktest
```

Source	SS	df	MS	Number of obs = 145
				F(2, 142) = 1460.70
Model	277.574775	2	138.787388	Prob > F = 0.0000
Residual	13.4920481	142	.095014423	R-squared = 0.9536
				Adj R-squared = 0.9530
Total	291.066823	144	2.02129738	Root MSE = .30824

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	.791953	.0293837	26.95	0.000	.733867	.8500389
_hatsq	.0941454	.0102281	9.20	0.000	.0739264	.1143643
_cons	-.0962174	.0425807	-2.26	0.025	-.1803914	-.0120434

上表显示,拟合值平方项(_hatsq)依然十分显著,故存在模型设定误差。下面进行 RESET 检验。

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of lntc
       Ho:  model has no omitted variables
                  F(3, 137) =      32.72
                  Prob > F =      0.0000
```

```
. estat ovtest,rhs
```

```
Ramsey RESET test using powers of the independent variables
      Ho:  model has no omitted variables
                 F(12, 128) =      8.96
                  Prob > F =      0.0000
```

以上结果均强烈拒绝“无遗漏变量”的原假设,即认为遗漏了高阶非线性项。考虑引入$(\ln q)^2$作为解释变量(记为 lnq2):

```
. g lnq2 = lnq^2
. reg lntc lnq lnq2 lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs =	145
				F(5, 139) =	622.86
Model	278.630831	5	55.7261661	Prob > F =	0.0000
Residual	12.4359927	139	.089467573	R-squared =	0.9573
				Adj R-squared =	0.9557
Total	291.066823	144	2.02129738	Root MSE =	.29911

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.1166562	.0613522	1.90	0.059	-.004648	.2379605
lnq2	.0536124	.0053141	10.09	0.000	.0431055	.0641194
lnpl	.0206146	.2326431	0.09	0.930	-.4393621	.4805913
lnpk	-.568725	.2614871	-2.17	0.031	-1.085732	-.0517185
lnpf	.4804816	.0766894	6.27	0.000	.3288531	.6321101
_cons	-.1627064	1.398139	-0.12	0.908	-2.927075	2.601662

上表显示,$(\ln q)^2$的系数估计值为正(0.05),且在1%的水平上显著。这表明,lnq 对 lntc 的边际效应并非常数,而是“0.12 +0.1lnq”(对[0.12ln q +0.05 $(\ln q)^2$]求导数),随着 lnq 的增加而变大。再次进行连接检验:

```
. linktest
```

Source	SS	df	MS	Number of obs =	145
				F(2, 142) =	1591.85
Model	278.638903	2	139.319451	Prob > F =	0.0000
Residual	12.4279206	142	.087520568	R-squared =	0.9573
				Adj R-squared =	0.9567
Total	291.066823	144	2.02129738	Root MSE =	.29584

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	1.009721	.0365875	27.60	0.000	.9373943	1.082047
_hatsq	-.0031437	.0103516	-0.30	0.762	-.0236068	.0173193
_cons	-.0013733	.0394759	-0.03	0.972	-.0794096	.0766631

上表显示,拟合值的平方项已高度不显著(p 值为 0.762)。再次进行 RESET 检验:

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of lntc
      Ho:  model has no omitted variables
                 F(3, 136) =      1.19
                  Prob > F =      0.3165
```

这个结果表明,引入$(\ln q)^2$后,不再遗漏拟合值($\hat{y}$)的幂。

更一般地，可以考虑以下“超越对数模型”（translog model）：

$$\begin{aligned}\ln tc = {} & \beta_0 + \beta_1 \ln q + \beta_2 \ln pl + \beta_3 \ln pk + \beta_4 \ln pf + \delta_1 (\ln q)^2 + \delta_2 (\ln pl)^2 \\ & + \delta_3 (\ln pk)^2 + \delta_4 (\ln pf)^2 + \gamma_{12} \ln q \cdot \ln pl + \gamma_{13} \ln q \cdot \ln pk \\ & + \gamma_{14} \ln q \cdot \ln pf + \gamma_{23} \ln pl \cdot \ln pk + \gamma_{24} \ln pl \cdot \ln pf + \gamma_{34} \ln pk \cdot \ln pf + \varepsilon\end{aligned} \tag{9.18}$$

上式包含了解释变量 lnq，lnpl，lnpk，lnpf 的所有二次项（含交叉项），故可以视为对任意非线性函数的二阶泰勒近似（second order Taylor approximation），具有非常灵活的函数形式（flexible functional form）。另一方面，它仍然是线性回归模型（是所有参数的线性函数），可以直接进行 OLS 估计。这使得超越对数模型在研究生产函数、成本函数、需求函数等领域得到广泛的应用，详见 Christensen and Greene（1976）与 Greene（2012）。

9.6 多重共线性

如果数据矩阵 $\boldsymbol{X}$ 不满列秩（列秩小于 K），即某一解释变量可以由其他解释变量线性表出，则存在“严格多重共线性”。此时，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在，总体参数 $\boldsymbol{\beta}$ 不可识别，无法定义最小二乘估计量。严格多重共线性在现实数据中很少出现，即使出现，Stata 也会自动识别并删去多余的解释变量。

较常见的是近似（非严格）的多重共线性。其表现为，如果将第 k 个解释变量 x_k 对其余的解释变量 $\{x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_K\}$ 进行回归，所得到的可决系数（记为 R_k^2）较高[①]。在存在近似多重共线性的情况下，OLS 仍然是最佳线性无偏估计，即在所有线性无偏估计中仍具有最小的方差。但这并不意味着 OLS 估计量方差在绝对意义上小。由于存在多重共线性，矩阵 $(\boldsymbol{X}'\boldsymbol{X})$ 变得几乎不可逆，故从某种意义上来说，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 变得很“大”，致使方差 $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 增大，使得对系数的估计变得不准确。在这种情况下，只要数据矩阵 $\boldsymbol{X}$ 中的元素轻微变化，就可能引起 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 很大的变化，进而导致 OLS 估计值 $\boldsymbol{b}$ 发生很大变化。通常的“症状”是，虽然整个回归方程的 R^2 较大、F 检验也很显著，但单个系数的 t 检验却不显著，或者系数估计值不合理，甚至符号与理论预期相反。另一可能“症状”是，增减解释变量使得系数估计值发生较大变化（比如，最后加入的解释变量与已有解释变量构成多重共线性）。直观来看，如果两个（或多个）解释变量之间高度相关，则不容易区分它们各自对被解释变量的单独影响力。在极端情况下，一个变量刚好是另一变量的倍数，则完全无法区分。

可以证明，协方差矩阵主对角线上的第 k 个元素为

$$\mathrm{Var}(b_k|\boldsymbol{X}) = \frac{\sigma^2}{(1 - R_k^2) S_{kk}} \tag{9.19}$$

其中，$S_{kk} \equiv \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2$ 为 x_k 的离差平方和，反映 x_k 的变动幅度。如果 x_k 变动很少，则很难准确地估计 x_k 对 y 的作用。在极端情况下，x_k 完全不变，$S_{kk} = 0$，则完全无法估计 b_k。在方程（9.19）中，我们更多地关注 $(1 - R_k^2)$。为此，定义第 k 个解释变量 x_k 的“方差膨胀因子”（Variance Inflation Factor，简记 VIF）为

---

① 仅仅看解释变量的“两两相关系数”（pairwise correlations）是不够的。即使两两相关系数较低，仍然有可能“复合”的相关系数较高，即 R_k^2 较大。

$$\mathrm{VIF}_k \equiv \frac{1}{1-R_k^2} \tag{9.20}$$

则 $\mathrm{Var}(b_k \mid \boldsymbol{X}) = \mathrm{VIF}_k \cdot (\sigma^2/S_{kk})$。VIF 越大则说明多重共线性问题越严重。一个经验规则是，最大的 VIF，即 $\max\{\mathrm{VIF}_1, \cdots, \mathrm{VIF}_K\}$，不超过 10。

在作完回归后，可以使用 Stata 命令“estat vif”计算 VIF。

仍以数据集“nerlove.dta”为例：

```
. use nerlove.dta,clear
. qui reg lntc lnq lnpl lnpk lnpf
. estat vif
```

Variable	VIF	1/VIF
lnpf	1.21	0.824250
lnpl	1.21	0.829013
lnpk	1.09	0.918113
lnq	1.04	0.960914
Mean VIF	1.14	

由于最大的 VIF 为 1.21，远小于 10，故不必担心存在多重共线性。这是 Nerlove(1963)估计成本函数的好处之一，因为要素价格之间的相关性通常较弱；如果直接估计生产函数，则要素投入之间的相关性会高得多。

如果发现存在多重共线性，可以采取以下处理方法。

(1) 如果不关心具体的回归系数，而只关心整个方程预测被解释变量的能力，则通常可以不必理会多重共线性(假设你的整个方程是显著的)。这是因为，多重共线性的主要后果是使得对单个变量的贡献估计不准，但所有变量的整体效应仍可以较准确地估计。

(2) 如果关心具体的回归系数，但多重共线性并不影响所关心变量的显著性，那么也可以不必理会。即使在有方差膨胀的情况下，这些系数依然显著；如果没有多重共线性，则只会更加显著。

(3) 如果多重共线性影响到所关心变量的显著性，则需要增大样本容量，剔除导致严重共线性的变量，或对模型设定进行修改。

9.7 极端数据

如果样本数据中的少数观测值离大多数观测值很远[①]，它们可能对 OLS 的回归系数产生很大影响。这些数据被称为“极端观测值”(outliers or influential data)，参见图 9.1。

对于一元回归，可以通过画(x,y)的散点图来直观地考察是否存在极端观测值。但画图的方法对于多元回归则不可行。可以证明，第 i 个观测数据对回归系数的“影响力”或“杠杠作用”(leverage)可以通过投影矩阵 $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的第 i 个主对角线元素来表示：

$$\mathrm{lev}_i \equiv \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{9.21}$$

所有观测数据的影响力 lev_i 满足：(i) $0 \leqslant \mathrm{lev}_i \leqslant 1, (i=1,\cdots,n)$；(ii) $\sum_{i=1}^{n} \mathrm{lev}_i = K$(解释变

① 可以想象这是在$(K+1)$维空间$\{y_i, x_i\}$中的距离。

量个数)。因此,影响力 lev_i 的平均值为 (K/n)。记 $\boldsymbol{b}^{(i)}$ 为去掉第 i 个观测数据后的 OLS 估计值,可以证明:

$$\boldsymbol{b}-\boldsymbol{b}^{(i)}=\left(\frac{1}{1-\text{lev}_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i \tag{9.22}$$

因此,lev_i 越大则 $(\boldsymbol{b}-\boldsymbol{b}^{(i)})$ 的变化越大。如果某些数据的 lev_i 比平均值 (K/n) 高很多,则可能对回归系数有很大影响。

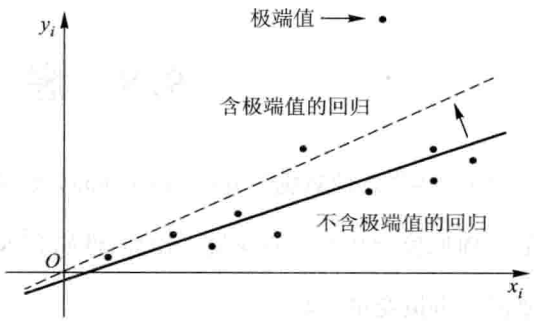


图 9.1 极端观测值对回归系数的影响

如何处理极端观测值呢?首先,应仔细检查是否因数据输入有误而导致极端观测值[①]。其次,对出现极端观测值的个体进行背景调查,看看是否由与研究课题无关的特殊现象所致,必要时可以删除极端数据。最后,比较稳健的做法是同时汇报"全样本"(full sample)与删除极端数据后的"子样本"(subsample)的回归结果,让读者自己做判断。在对跨国数据的回归中,就经常同时汇报所有国家(全样本)与非石油输出国家(子样本)的结果,比如 Mankiw et al (1992)。

计算影响力 lev_i 的 Stata 命令为:

```
reg y x1 x2 x3
predict lev,leverage   (列出所有解释变量的 lev 值)
gsort - lev            (将所有观测数据按 lev 的降序排列)
sum lev                (看到 lev 的最大值与平均值)
list lev in 1/3        (列出从第 1 到第 3 个数据的 lev 值)
```

注:如果使用命令 sort,则只能按升序排列。

回到数据集 nerlove. dta 的例子:

```
. use nerlove.dta,clear
. qui reg lntc lnq lnpl lnpk lnpf
. predict lev,leverage
. su lev
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lev	145	.0344828	.0202164	.009924	.1177335

```
. dis r(max)/r(mean)
3.4142728
```

lev 的最大值是其平均值的 3.41 倍,似乎并不大。下面来看 lev 最大的三个数值:

```
. gsort -lev
. list lev in 1/3
```

	lev
1.	.1177335
2.	.1001472
3.	.0983759

① 比如,多输入了一个 0,或漏了一位数。

9.8 虚拟变量

如果使用“定性数据”(qualitative data)或“分类数据”(categorical data),通常需要引入“虚拟变量”,即取值为 0 或 1 的变量。比如,性别分男女,可定义 $D=\begin{cases}1, & \text{男}\\ 0, & \text{女}\end{cases}$。对于全球的五大洲,则需要四个虚拟变量,即

$$D_1=\begin{cases}1, & \text{非洲},\\ 0, & \text{其他};\end{cases}\quad D_2=\begin{cases}1, & \text{美洲},\\ 0, & \text{其他};\end{cases}\quad D_3=\begin{cases}1, & \text{欧洲},\\ 0, & \text{其他};\end{cases}\quad D_4=\begin{cases}1, & \text{非洲},\\ 0, & \text{其他}。\end{cases}$$

如果 $D_1=D_2=D_3=D_4=0$,则表明为大洋洲。

在有常数项的模型中,如果定性指标共分 M 类,则最多只能有($M-1$)个虚拟变量。如果在回归方程中包含了 M 个虚拟变量,则会产生严格多重共线性,因为如果将这 M 个虚拟变量在数据矩阵 $\boldsymbol{X}$ 中对应的列向量相加,就会得到与常数项完全相同的向量,即$(1\cdots1)'$(因为 M 类中必居其一)。这种情况被称为“虚拟变量陷阱”(dummy variable trap)。由于 Stata 会自动识别严格多重共线性,这种担心已不重要。如果模型中没有常数项,则可以有 M 个虚拟变量。

在模型中引入虚拟变量,会带来什么影响呢?考虑一个有关中国经济的时间序列模型:

$$y_t=\alpha+\beta x_t+\varepsilon_t,\quad t=1950,\cdots,2000 \tag{9.23}$$

由于经济结构可能在 1978 年改革开放以后有变化,引入虚拟变量:

$$D=\begin{cases}1, & \text{若 } t\geqslant 1978\\ 0, & \text{其他}\end{cases} \tag{9.24}$$

考虑以下两种情况。

(1) 仅仅引入虚拟变量本身

$$y_t=\alpha+\beta x_t+\gamma D_t+\varepsilon_t \tag{9.25}$$

显然,该模型等价于

$$y_t=\begin{cases}\alpha+\beta x_t+\varepsilon_t, & \text{若 } t<1978\\ (\alpha+\gamma)+\beta x_t+\varepsilon_t, & \text{若 } t\geqslant 1978\end{cases} \tag{9.26}$$

因此,仅仅引入虚拟变量相当于在两个不同的时期使用不同的截距项,参见图 9.2。

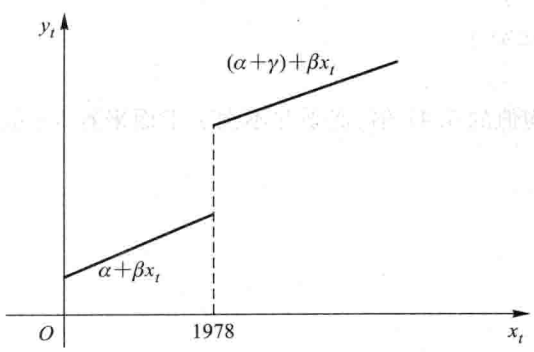


图 9.2 仅引入虚拟变量的效果

(2) 引入虚拟变量,以及虚拟变量与解释变量的"互动项"(interaction term)

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \tag{9.27}$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + (\beta + \delta) x_t + \varepsilon_t, & \text{若 } t \geqslant 1978 \end{cases} \tag{9.28}$$

因此,引入虚拟变量及其互动项相当于,在两个不同的时期使用不同的截距项与斜率,参见图 9.3。如果仅仅引入互动项,则仅改变斜率。

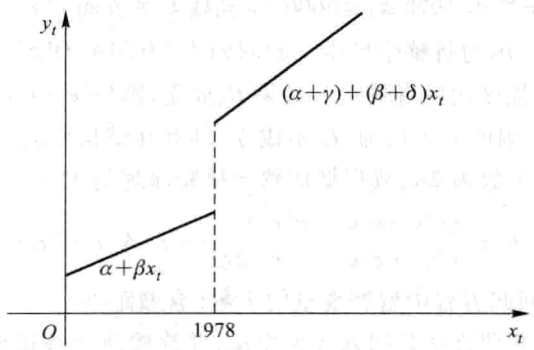


图 9.3 引入虚拟变量及其互动项的效果

假设时间变量为 year,则可用如下 Stata 命令生成此虚拟变量:

```
gen d = (year > =1978)
```

如果有 30 个省的名字储存于变量 province,希望为每省设立一个虚拟变量,记为"pr1, pr2,…,pr30",可使用如下 Stata 命令:

```
tabulate province,generate(pr)
```

注:这些虚拟变量的排序将依照变量 province 的字母顺序而定。

在回归时可以使用变量的简略写法,比如:

```
reg x1 x2 x3 pr2 -pr30
```

9.9 经济结构变动的检验

1. 结构变动日期已知

对于时间序列模型而言,模型系数的稳定性(stability)是一个很重要的问题。如果存在"结构变动"(structural break),但未加考虑,也是一种模型设定误差。首先考虑结构变动日期已知的情形。

继续 9.8 节的例子,假设要检验中国经济是否在 1978 年发生结构变动。定义第 1 个时期为 $1950 \leqslant t < 1978$,第 2 个时期为 $1978 \leqslant t \leqslant 2000$,则两个时期对应的回归方程可以分别记为

$$\boldsymbol{y}^1 = \boldsymbol{X}^1 \boldsymbol{\beta}^1 + \boldsymbol{\varepsilon}^1 \tag{9.29}$$

$$\boldsymbol{y}^2 = \boldsymbol{X}^2 \boldsymbol{\beta}^2 + \boldsymbol{\varepsilon}^2 \tag{9.30}$$

需要检验的原假设为,经济结构在这两个时期内没有变化,即"$H_0: \boldsymbol{\beta}^1 = \boldsymbol{\beta}^2$"。假设有 K 个解释变量,则 H_0 共有 K 个约束。在无约束的情况下,可对两个时期分别进行回归。而在有约束

(即 H_0 成立)的情况下,可以将模型合并为

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{9.31}$$

其中,$\boldsymbol{y}\equiv\begin{pmatrix}\boldsymbol{y}^1\\ \boldsymbol{y}^2\end{pmatrix}$,$\boldsymbol{X}\equiv\begin{pmatrix}\boldsymbol{X}^1\\ \boldsymbol{X}^2\end{pmatrix}$,$\boldsymbol{\varepsilon}\equiv\begin{pmatrix}\boldsymbol{\varepsilon}^1\\ \boldsymbol{\varepsilon}^2\end{pmatrix}$。因此,可以将所有样本数据合在一起回归。传统的"邹检验"(Chow,1960)通过作以下三个回归来检验"无结构变动"的原假设。

首先,回归整个样本 $1950\leqslant t\leqslant 2000$,得到残差平方和 $\boldsymbol{e}'\boldsymbol{e}$。

其次,回归第 1 部分子样本 $1950\leqslant t<1978$,得到残差平方和 $\boldsymbol{e}_1'\boldsymbol{e}_1$。

最后,回归第 2 部分子样本 $1978\leqslant t\leqslant 2000$,得到残差平方和 $\boldsymbol{e}_2'\boldsymbol{e}_2$。

显然,$\boldsymbol{e}'\boldsymbol{e}\geqslant\boldsymbol{e}_1'\boldsymbol{e}_1+\boldsymbol{e}_2'\boldsymbol{e}_2$,因为将整个样本一起回归为"有约束 OLS",而将样本一分为二为"无约束 OLS",故前者的拟合优度比后者更差。如果 H_0 成立,则($\boldsymbol{e}'\boldsymbol{e}-\boldsymbol{e}_1'\boldsymbol{e}_1-\boldsymbol{e}_2'\boldsymbol{e}_2$)应该比较小。如果($\boldsymbol{e}'\boldsymbol{e}-\boldsymbol{e}_1'\boldsymbol{e}_1-\boldsymbol{e}_2'\boldsymbol{e}_2$)很大,则倾向于认为 H_0 不成立,即存在结构变动。由于约束条件共有 K 个,而无约束回归的解释变量个数为 $2K$,故根据似然比检验原理的 F 统计量为

$$F=\frac{(\boldsymbol{e}'\boldsymbol{e}-\boldsymbol{e}_1'\boldsymbol{e}_1-\boldsymbol{e}_2'\boldsymbol{e}_2)/K}{(\boldsymbol{e}_1'\boldsymbol{e}_1+\boldsymbol{e}_2'\boldsymbol{e}_2)/(n-2K)}\sim F(K,n-2K) \tag{9.32}$$

其中,n 为样本容量,K 为回归方程中解释变量的个数(含截距项)。

检验结构变动的另一简便方法是引入虚拟变量,并检验所有虚拟变量以及其与解释变量交叉项的系数的联合显著性。比如,在前面的例子中,进行如下回归:

$$y_t=\alpha+\beta x_t+\gamma D_t+\delta D_t x_t+\varepsilon_t \tag{9.33}$$

然后检验"$H_0:\gamma=\delta=0$"。这个检验所得到的 F 统计量与传统的邹检验完全相同。因此,虚拟变量法与邹检验是等价的。

与传统的邹检验相比,虚拟变量法的优点包括:(1)只需生成虚拟变量即可检验,十分方便;(2)邹检验是在"扰动项同方差"的假设下得到的,并不适用于条件异方差的情形。在条件异方差的情况下,仍可使用虚拟变量法,只要估计方程(9.33)使用异方差稳健的标准误即可。(3)如果发现存在结构变动,邹检验并不提供究竟是截距项还是斜率变动的信息(至少需要再作一个邹检验),而虚拟变量法则可以同时提供这些信息。

2. 结构变动日期未知

更一般地,可能不知道结构变动的具体时间。比如,也许不能肯定结构变动一定发生在 1978 年。此时,选择一个区间$[\tau_0,\tau_1]\subseteq[1,T]$(无法检验过于靠近端点的位置,故只能取其中的一个区间),其中 T 为样本容量,而 1950 年对应于第 1 年,可以按照以上方法计算在此区间中的每一年份 $t(\tau_0\leqslant t\leqslant\tau_1)$ 所对应的 F 统计量,然后取其最大者。这个统计量被称为"匡特似然比"(Quandt Likelihood Ratio,简记 QLR)[①],是邹统计量的推广。

由于 QLR 统计量为许多 F 统计量之最大者,它不再服从 F 分布。QLR 统计量的分布取决于约束条件的个数(即有多少个变量的系数可能发生变动),以及(τ_0/T)与(τ_1/T)。如果 τ_0 太接近于 1,或 τ_1 太接近于 T,则 QLR 统计量的渐近分布对有限样本分布的近似将变得不准确。通常选择 $\tau_0=0.15T$,$\tau_1=0.85T$(选择最接近的整数),称之为"15% 修边"(15% trimming),即只对样本中间的 70% 观测值计算 F 统计量,然后取其最大者得到 QLR 统计量。QLR 统计量的 10%,5% 与 1% 显著性水平的临界值见表 9.1。

① 也称为"sup-Wald"统计量。

表 9.1 QLR 统计量临界值表(15%修边)

约束条件个数	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97
13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

资料来源:Stock and Watson (2011),p. 559,Table 14.6。

如果 QLR 统计量小于临界值,则接受“无结构变动”的原假设。反之,则认为发生了结构变动,而 F 统计量取最大值的那个日期 $\hat{\tau}$ 就是对结构变动日期(break date)τ 的一致估计。当然,当 QLR 统计量超过临界值而拒绝原假设时,也可能存在“多个结构变动”(multiple breaks)。

下面以数据集 consumption_china.dta 为例,考察中国的消费函数是否在 1992 年发生了结构变化。先看一下中国 1978—2006 年“居民人均消费”(c)与“人均国内总产值”(y)的年度(year)时间趋势图(如图 9.4),以当年价格计。数据来自国家统计局网站:

```
. use consumption_china.dta,clear
. graph twoway connect c y year,msymbol(circle) msymbol(triangle)
```

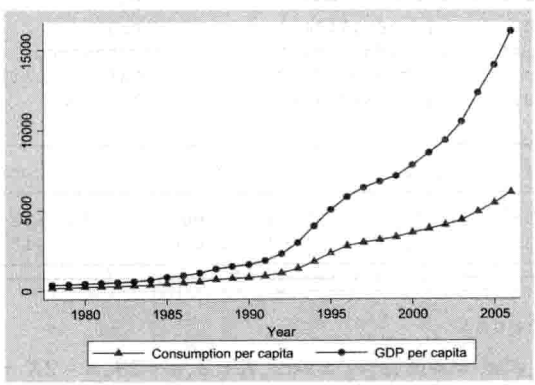


图 9.4 居民人均消费与人均国内总产值时间趋势

显然,二者的走势具有较强的相关性。考察一个简单(粗糙)的消费函数:

$$c_t = \alpha + \beta y_t + \varepsilon_t \tag{9.34}$$

首先,使用传统的邹检验(F 检验)来检验消费函数是否在1992年发生结构变动。分别对整个样本、1992年之前及之后的子样本进行回归,以获得其残差平方和:

```
. reg c y
```

Source	SS	df	MS	Number of obs = 29
				F(1, 27) = 2441.51
Model	92575330.6	1	92575330.6	Prob > F = 0.0000
Residual	1023766.1	27	37917.2631	R-squared = 0.9891
				Adj R-squared = 0.9887
Total	93599096.7	28	3342824.88	Root MSE = 194.72

c	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y	.3977205	.0080491	49.41	0.000	.3812051	.414236
_cons	188.588	51.57697	3.66	0.001	82.76078	294.4152

```
. scalar ssr = e(rss)
```

上述命令将回归的残差平方和(e(rss))记为标量(scalar) ssr。

```
. reg c y if year<1992
```

Source	SS	df	MS	Number of obs = 14
				F(1, 12) = 4381.66
Model	829144.945	1	829144.945	Prob > F = 0.0000
Residual	2270.7688	12	189.230733	R-squared = 0.9973
				Adj R-squared = 0.9970
Total	831415.714	13	63955.0549	Root MSE = 13.756

c	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y	.4995544	.0075468	66.19	0.000	.4831113	.5159975
_cons	12.98254	7.87353	1.65	0.125	-4.172404	30.13749

```
. scalar ssr1 = e(rss)
. reg c y if year>=1992
```

Source	SS	df	MS	Number of obs = 15
				F(1, 13) = 764.58
Model	28749474.3	1	28749474.3	Prob > F = 0.0000
Residual	488822.141	13	37601.7031	R-squared = 0.9833
				Adj R-squared = 0.9820
Total	29238296.4	14	2088449.74	Root MSE = 193.91

c	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y	.359886	.0130153	27.65	0.000	.3317682	.3880038
_cons	566.4531	115.2172	4.92	0.000	317.5415	815.3646

```
. scalar ssr2 = e(rss)
```

由上述结果可知,$\boldsymbol{e}'\boldsymbol{e}$ = ssr,$\boldsymbol{e}_1'\boldsymbol{e}_1$ = ssr1,$\boldsymbol{e}_2'\boldsymbol{e}_2$ = ssr2,$K=2$,$n=29$,$n-2K=25$,故可以计算 F 统计量如下:

```
. di ((ssr - ssr1 - ssr2)/2)/((ssr1 + ssr2)/25)
13.558361
```

F 统计量等于 13.56。

其次,使用虚拟变量法进行结构变动的检验。生成虚拟变量 d(对于 1992 年及以后,d = 1;反之,d = 0);以及虚拟变量 d 与人均收入 y 的互动项 yd:

```
. gen d = (year >1991)
. gen yd = y * d
```

引入 d 与 yd,进行 OLS 回归:

```
. reg c y d yd
```

Source	SS	df	MS	Number of obs = 29
Model	93108003.8	3	31036001.3	F(3, 25) = 1579.95
Residual	491092.91	25	19643.7164	Prob > F = 0.0000
				R-squared = 0.9948
Total	93599096.7	28	3342824.88	Adj R-squared = 0.9941
				Root MSE = 140.16

c	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y	.4995544	.0768917	6.50	0.000	.341193	.6579158
d	553.4705	115.6304	4.79	0.000	315.3252	791.6159
yd	-.1396684	.077465	-1.80	0.083	-.2992106	.0198738
_cons	12.98254	80.22052	0.16	0.873	-152.2347	178.1998

然后检验 d 与 yd 的联合显著性:

```
. test d yd
```

```
( 1)  d = 0
( 2)  yd = 0

       F(  2,    25) =   13.56
            Prob > F =    0.0001
```

这个结果表明,使用虚拟变量法所得到的 F 统计量也等于 13.56,与传统邹检验完全相同。该检验的 p 值为 0.000 1,故可以在 1% 的显著性水平上强烈拒绝“没有结构变动”的原假设,即认为中国的消费函数在 1992 年发生了结构变动。

然而,上述结构变化检验仅在扰动项同方差的情况下才成立。事实上,如果使用第 7 章介绍的方法进行异方差检验,将拒绝“同方差”的原假设(读者可自行验证,在此从略)。下面,使用稳健标准误进行虚拟变量法的检验。

```
. reg c y d yd,r
```

Linear regression					Number of obs = 29
					F(3, 25) = 2290.56
					Prob > F = 0.0000
					R-squared = 0.9948
					Root MSE = 140.16

c	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
y	.4995544	.0104615	47.75	0.000	.4780085	.5211003
d	553.4705	138.6738	3.99	0.001	267.8665	839.0746
yd	-.1396684	.0186167	-7.50	0.000	-.1780103	-.1013265
_cons	12.98254	8.966059	1.45	0.160	-5.483399	31.44849

```
. test d yd
```

```
( 1)  d = 0
( 2)  yd = 0

      F(  2,    25) =   34.01
           Prob > F =    0.0000
```

上表显示,无论是否存在异方差(稳健标准误在同方差的情况下依然成立),都可以强烈拒绝"没有结构变动"的原假设。

9.10 缺失数据与线性插值

在现实数据中,有时会出现某些时期数据缺失(missing data)的情形,尤其是历史比较久远的数据。缺失的观测值在 Stata 中以"."来表示,在运行 Stata 命令时(比如 reg),会自动将缺失观测值从样本中去掉,导致样本容量损失。在数据缺失不严重的情况下,为了保持样本容量,可采用"线性插值"(linear interpolation)的方法来补上缺失数据。

考虑最简单的情形。已知 x_{t-1} 与 x_{t+1},但缺失 x_t 的数据,则 x_t 对时间 t 的线性插值为

$$\hat{x}_t = \frac{x_{t-1} + x_{t+1}}{2} \tag{9.35}$$

更一般地,假设与 x(通常为时间)对应的 y 缺失,而最临近的两个点分别为 (x_0, y_0) 与 (x_1, y_1),且 $x_0 < x < x_1$,则 y 对 x 的线性插值为

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0 \tag{9.36}$$

公式(9.36)的示意图参见图 9.5。线性插值的基本假设是变量以线性的速度均匀地变化。因此,如果变量 y 有指数增长趋势(比如 GDP),则应先取对数,再用 $\ln y$ 进行线性插值,以避免偏差。如果需要以原变量 y 进行回归,可将线性插值的对数值 $\ln \hat{y}$ 再取反对数(antilog),即计算 $\exp(\ln \hat{y})$。

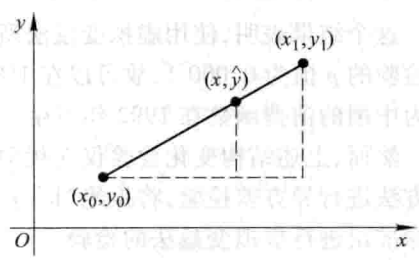


图 9.5 线性插值示意图

线性插值的 Stata 命令为

```
ipolate y x,gen(newvar)
```

其中,"ipolate"表示 interpolate,即将变量 y 对变量 x 进行线性插值,并将插值的结果记为变量 newvar。

继续以数据集 consumption_china.dta 为例。

```
. use consumption_china.dta,clear
```

为了演示的目的,假设 1980 年、1990 年与 2000 年的人均 GDP 数据缺失。首先,生成缺失这些年份数据的人均 GDP 变量,将其记为 y1。

```
. gen y1 =y
. replace y1 =. if year = =1980 |year = =1990 |year = =2000
(3 real changes made,3 to missing)
```

直接用 y1 对 year 进行线性插值,并将结果记为 y2。

```
. ipolate y1 year,gen(y2)
```

由于人均 GDP 有指数增长趋势，故更好的做法是，先对 y1 取对数，进行线性插值，再取反对数，并将结果记为 y3。

```
. gen lny1 = log(y1)
(3 missing values generated)
. ipolate lny1 year,gen(lny3)
. gen y3 = exp(lny3)
```

下面，对比这两种方法的效果。

```
. list year y y2 y3 if year = =1980 |year = =1990 |year = =2000
```

	year	y	y2	y3
3.	1980	463	455.5	454.0352
13.	1990	1644	1706	1695.72
23.	2000	7858	7890.5	7856.52

从上表可知，直接插值的结果 y2 倾向于高估真实值 y，而且整体估计效果不如先取对数再插值的结果 y3（1980 年的结果是个例外）。

9.11　变量单位的选择

在选择变量单位时，应尽量避免变量间的数量级差别过于悬殊，以免出现计算机运算的较大误差。比如，通货膨胀率通常小于 1，而如果模型中有 GDP 这个变量，则 GDP 应该使用亿或万亿作为单位。否则，变量 GDP 的取值将是通货膨胀率的很多倍，即数据矩阵 $\boldsymbol{X}$ 中某列的数值是另一列的很多倍，这可能使计算机在对 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 进行数值计算时出现较大误差。这是因为，计算机的存储空间有限，实际上只能作近似计算，即精确到小数点后若干位。

习　题

9.1　使用数据集 hprice2a. dta，对习题 7.3 中的回归方程，进行如下操作。

（1）进行 RESET 检验（分别使用拟合值 $\hat{y}$ 与 *rhs* 解释变量）；

（2）计算信息准则 AIC 与 BIC；

（3）计算所有解释变量的方差膨胀因子（VIF）。存在严重多重共线性吗？

（4）计算观测数据的 lev 的最大值与平均值。最大值是平均值的几倍？列出影响力最大的前 3 个数据。

9.2　使用数据集 ukrates. dta（参见第 8 章），检验英国的货币政策反应函数是否在 1973 年 10 月石油危机后发生结构变动。提示：定义虚拟变量“g d = (month > tm(1973m10))”，其中“tm”表示月度数据格式。

附　录

A9.1　**对于自回归模型，BIC 准则是一致估计。**

证明：首先考虑最简单的情形。假设真实滞后阶数为 $p=1$，然后使用 BIC 准则从 $\hat{p}=0,1,2$ 中进行选择。下面将证明：(1) $\Pr(\hat{p}<p)\to 0$；(2) $\Pr(\hat{p}>p)\to 0$；故 $\Pr(\hat{p}=p)\to 1$。对于更一般的情形 $0\leqslant p\leqslant p_{\max}$，可以类似地证明。

(1) 由于 $p=1$,故 $\hat{p}<p \Leftrightarrow \hat{p}=0$。而要选择 $\hat{p}=0$,必须有 $\mathrm{BIC}(0)<\mathrm{BIC}(1)$,即 $\mathrm{BIC}(0)-\mathrm{BIC}(1)<0$。记相应的残差平方和为 $\mathrm{SSR}(0)$ 与 $\mathrm{SSR}(1)$,则

$$\begin{aligned}\mathrm{BIC}(0)-\mathrm{BIC}(1) &= \left[\ln\left(\frac{\mathrm{SSR}(0)}{T}\right)+\frac{\ln T}{T}\right]-\left[\ln\left(\frac{\mathrm{SSR}(1)}{T}\right)+\frac{2\ln T}{T}\right] \\ &= \ln\left(\frac{\mathrm{SSR}(0)}{T}\right)-\ln\left(\frac{\mathrm{SSR}(1)}{T}\right)-\frac{\ln T}{T}\end{aligned} \tag{9.37}$$

当 $\hat{p}=0$ 时,$y_t=\hat{\beta}_0+e_t$,故 $\dfrac{\mathrm{SSR}(0)}{T}=\dfrac{T-1}{T}\cdot s_y^2 \xrightarrow{p} \sigma_y^2$;另一方面,当 $\hat{p}=1$ 时,$y_t=\hat{\beta}_0+\hat{\beta}_1 y_{t-1}+e_t$,故 $\dfrac{\mathrm{SSR}(1)}{T}=\dfrac{T-1}{T}\cdot s^2 \xrightarrow{p} \sigma_e^2$。根据微积分的洛必达法则可知,$\dfrac{\ln T}{T}\to 0$。因此,

$$\mathrm{BIC}(0)-\mathrm{BIC}(1) \xrightarrow{p} (\ln \sigma_y^2-\ln \sigma_e^2)>0 \tag{9.38}$$

其中,$\sigma_y^2>\sigma_e^2$(参见第5章有关 AR(1)为平稳过程的证明)。因此,$\Pr[\mathrm{BIC}(0)-\mathrm{BIC}(1)]<0$,故 $\Pr(\hat{p}=0)\to 0$。

(2) 要选择 $\hat{p}=2$,必须有 $\mathrm{BIC}(2)<\mathrm{BIC}(1)$,即 $\mathrm{BIC}(2)-\mathrm{BIC}(1)<0$。

$$\begin{aligned}T[\mathrm{BIC}(2)-\mathrm{BIC}(1)] &= T\left\{\left[\ln\left(\frac{\mathrm{SSR}(2)}{T}\right)+\frac{3\ln T}{T}\right]-\left[\ln\left(\frac{\mathrm{SSR}(1)}{T}\right)+\frac{2\ln T}{T}\right]\right\} \\ &= T\ln\left(\frac{\mathrm{SSR}(2)}{\mathrm{SSR}(1)}\right)+\ln T \\ &= -T\ln\left(1+\frac{F}{T-2}\right)+\ln T\end{aligned} \tag{9.39}$$

其中,$F\equiv\dfrac{\mathrm{SSR}(1)-\mathrm{SSR}(2)}{\mathrm{SSR}(2)/(T-2)}$ 为检验方程"$y_t=\beta_0+\beta_1 y_{t-1}+\beta_2 y_{t-2}+\varepsilon_t$"中"$H_0:\beta_2=0$"的 F 统计量(适用于同方差的情形)。如果扰动项 ε_t 同方差,则 F 的渐近分布为 $\chi^2(1)$;其他情况下,则为另外的渐近分布。因此,

$$\begin{aligned}\Pr[\mathrm{BIC}(2)-\mathrm{BIC}(1)<0] &= \Pr[T(\mathrm{BIC}(2)-\mathrm{BIC}(1))<0] \\ &= \Pr\left[-T\ln\left(1+\frac{F}{T-2}\right)+\ln T<0\right] \\ &= \Pr\left[T\ln\left(1+\frac{F}{T-2}\right)>\ln T\right]\end{aligned} \tag{9.40}$$

由于 $\ln(1+a)\approx a$(当 $a\to 0$),故 $\ln\left(1+\dfrac{F}{T-2}\right)\approx\dfrac{F}{T-2}$(由于 F 的渐近分布给定,故 F 的取值有限,而 T 不断增大,故 $\dfrac{F}{T-2}$ 很小,此近似成立),因此,$T\ln\left(1+\dfrac{F}{T-2}\right)\approx T\cdot\dfrac{F}{T-2}\approx F$。综合上面的结果,有

$$\Pr[\mathrm{BIC}(2)-\mathrm{BIC}(1)<0]\longrightarrow \Pr[F>\ln T]\to 0 \tag{9.41}$$

因此,$\Pr(\hat{p}=2)\to 0$。

A9.2 对于自回归模型,HQIC 准则是一致估计。

证明:在 A9.1(1)的证明中,将 $\ln T$ 替换为 $\ln(\ln T)$,结论依然成立,即 $\Pr(\hat{p}=0)\to 0$。

在 A9.1(2)的证明中,将 $\ln T$ 替换为 $\ln(\ln T)$,则

$$\Pr[\mathrm{BIC}(2)-\mathrm{BIC}(1)<0]\longrightarrow \Pr[F>\ln(\ln T)]\to 0 \tag{9.42}$$

因此,$\Pr(\hat{p}=2)\to 0$。

A9.3 对于自回归模型,AIC 准则不是一致估计。

证明:在 A9.1(1)的证明中,将 $\ln T$ 替换为 2,结论依然成立,即 $\Pr(\hat{p}=0)\to 0$。

在 A9.1(2)的证明中,将 $\ln T$ 替换为 2,则

$$\Pr[\mathrm{BIC}(2)-\mathrm{BIC}(1)<0]\longrightarrow \Pr[F>2]>0 \tag{9.43}$$

如果扰动项 ε_t 同方差,则 F 的渐近分布为 $\chi^2(1)$,故 $\Pr[F>2]\to\Pr[\chi^2(1)>2]=0.16$。在一般情况下,$\Pr(\hat{p}=2)\to c>0$。因此,$\Pr(\hat{p}=p)\to d<1$。

第 10 章　工具变量,2SLS 与 GMM

OLS 能够成立的最重要条件是解释变量与扰动项不相关(即前定变量的假设)。否则,OLS 估计量将是不一致的,即无论样本容量多大,OLS 估计量也不会收敛到真实的总体参数。一般来说,这是无法接受的。然而,解释变量与扰动项相关的例子却比比皆是。主要解决方法之一为本章介绍的工具变量法,它对于实证研究有着重要的价值。

10.1　解释变量与扰动项相关的例子

例　农产品市场均衡模型

$$\begin{cases} q_t^d = \alpha_0 + \alpha_1 p_t + u_t & \text{(需求)} \\ q_t^s = \beta_0 + \beta_1 p_t + v_t & \text{(供给)} \\ q_t^d = q_t^s & \text{(均衡)} \end{cases} \tag{10.1}$$

令 $q_t \equiv q_t^d = q_t^s$,可得

$$\begin{cases} q_t = \alpha_0 + \alpha_1 p_t + u_t \\ q_t = \beta_0 + \beta_1 p_t + v_t \end{cases} \tag{10.2}$$

显然,这两个方程中的被解释变量与解释变量完全一样。如果直接作回归 $q_t \xrightarrow{\text{OLS}} p_t$,那么估计的究竟是需求函数还是供给函数呢?两者都不是!参见图 10.1。

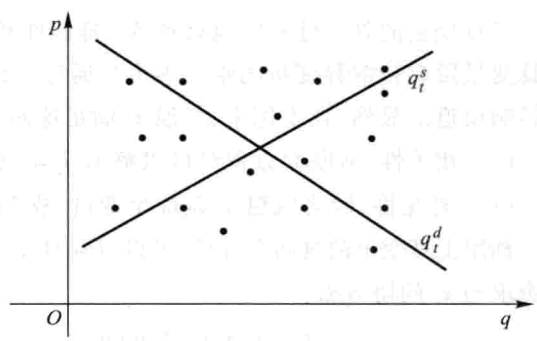


图 10.1　需求与供给决定市场均衡

如果把线性方程组(10.2)中的(p_t, q_t)看成是未知数(内生变量),而把(u_t, v_t)看做已知,则可以求解(p_t, q_t)为(u_t, v_t)的函数[①]:

$$\begin{cases} p_t = p_t(u_t, v_t) = \dfrac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \dfrac{v_t - u_t}{\alpha_1 - \beta_1} \\ q_t = q_t(u_t, v_t) = \dfrac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \dfrac{\alpha_1 v_t - \beta_1 u_t}{\alpha_1 - \beta_1} \end{cases} \tag{10.3}$$

显然,由于 p_t 为(u_t, v_t)的函数,故 $\text{Cov}(p_t, u_t) \neq 0$, $\text{Cov}(p_t, v_t) \neq 0$。因此,OLS 估计值 $\hat{\alpha}_1, \hat{\beta}_1$ 不是 α_1, β_1 的一致估计量。称这种偏差为"联立方程偏差"(simultaneity bias)或"内生变量偏差"(endogeneity bias)。在这个例子中,我们无法从价格变化的信息中得知,究竟这种变化是由于需求还是供给引起的。

---

① 这也符合数据生成过程(Data Generating Process)的视角,即(p_t, q_t)由(u_t, v_t)所决定。

既然 OLS 的不一致性是由于“内生变量”(endogenous variables)[①]与扰动项相关而引起,如果我们能够将内生变量分成两部分,即一部分与扰动项相关,而另一部分与扰动项不相关,那么就有希望用与扰动项不相关的那一部分得到一致估计。对内生变量的这种分离可以借助于对内生变量的深入认识来完成[②],而更常见的方法则借助另外一个“工具变量”来实现。

假设在图 10.1 中,存在某个因素(变量)使得供给曲线经常移动,而需求曲线基本不动。此时,就可以估计需求曲线,参见图 10.2。这个使得供给曲线移动的变量就是工具变量。假设影响方程组(10.1)中供给方程扰动项的因素可以分解为两部分,即可观测的气温 x_t 与不可观测的其他因素:

$$q_t^s = \beta_0 + \beta_1 p_t + \beta_2 x_t + v_t \qquad (10.4)$$

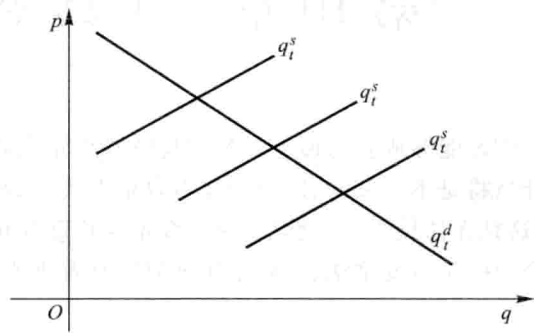


图 10.2 稳定的需求与变动的供给

假定气温 x_t 是个前定变量[③],与两个扰动项都不相关,即 $\mathrm{Cov}(x_t, u_t) = 0, \mathrm{Cov}(x_t, v_t) = 0$。由于气温 x_t 的变化使得供给函数 q_t^s 沿着需求函数 q_t^d 移动,这使得我们可以估计出需求函数 q_t^d。

在这种情况下,称 x_t 为“工具变量”(Instrumental Variable,简记 IV)。在回归方程中(此处为需求方程),一个有效(valid)的工具变量应满足以下两个条件。

(i) 相关性:工具变量与内生解释变量相关,即 $\mathrm{Cov}(x_t, p_t) \neq 0$。

(ii) 外生性:工具变量与扰动项不相关,即 $\mathrm{Cov}(x_t, u_t) = 0$。

工具变量的外生性有时也被称为“排他性约束”(exclusion restriction),因为外生性意味着,工具变量影响被解释变量的唯一渠道是通过与其相关的内生解释变量,它排除了所有其他的可能影响渠道。显然,在本例中,气温 x_t 满足这两个条件。

(i) 相关性:从联立方程组可以解出 $p_t = p_t(x_t, u_t, v_t)$,故 $\mathrm{Cov}(x_t, p_t) \neq 0$。

(ii) 外生性:因为气温 x_t 是前定变量,故 $\mathrm{Cov}(x_t, u_t) = 0$。

利用工具变量的这两个性质,可以得到对 α_1 的一致估计。同时对需求方程 $q_t = \alpha_0 + \alpha_1 p_t + u_t$ 两边求与 x_t 的协方差:

$$\begin{aligned} \mathrm{Cov}(q_t, x_t) &= \mathrm{Cov}(\alpha_0 + \alpha_1 p_t + u_t, x_t) \\ &= \alpha_1 \mathrm{Cov}(p_t, x_t) + \underbrace{\mathrm{Cov}(u_t, x_t)}_{=0} = \alpha_1 \mathrm{Cov}(p_t, x_t) \end{aligned} \qquad (10.5)$$

根据工具变量的相关性,$\mathrm{Cov}(p_t, x_t) \neq 0$,可以把上式两边同除以 $\mathrm{Cov}(p_t, x_t)$,得

$$\alpha_1 = \frac{\mathrm{Cov}(q_t, x_t)}{\mathrm{Cov}(p_t, x_t)} \qquad (10.6)$$

使用表达式(10.6)对应的样本值,就可以得到一致的“工具变量估计量”(Instrumental Variable Estimator):

---

① 在计量经济学中,把所有与扰动项相关的解释变量都称为“内生变量”,这与经济学中的定义有所不同。

② 比如,考虑货币政策对宏观经济的影响。由于货币政策的制定者会根据宏观经济的运行情况来调整货币政策,故货币政策是个内生变量。Romer and Romer(2004)通过阅读有关美联储的历史文献将货币政策的变动分为“内生”(对经济的反应)与“外生”(货币当局的自主调整)两部分。

③ 一般而言,人类活动对气温的影响可以忽略。

$$\hat{\alpha}_{1,\mathrm{IV}}=\frac{\widehat{\mathrm{Cov}(q_t,x_t)}}{\widehat{\mathrm{Cov}(p_t,x_t)}}\xrightarrow{p}\frac{\mathrm{Cov}(q_t,x_t)}{\mathrm{Cov}(p_t,x_t)}=\alpha_1 \tag{10.7}$$

从上式也可以看出，如果工具变量与内生变量无关，即 $\mathrm{Cov}(x_t,p_t)=0$，则无法定义工具变量法[①]。如果工具变量与内生变量的相关性很弱，即 $\mathrm{Cov}(x_t,p_t)\approx 0$，则会导致估计量 $\hat{\alpha}_{1,\mathrm{IV}}$ 的方差变得很大，这被称为“弱工具变量问题”（详见下文）。传统的工具变量法一般通过“二阶段最小二乘法”（Two Stage Least Square，简记 2SLS 或 TSLS）来实现（Theil，1953；Basmann，1957），顾名思义，即通过作两个回归来完成。

第一阶段回归：用内生解释变量对工具变量回归，即 $p_t\xrightarrow{\mathrm{OLS}}x_t$，得到拟合值 $\hat{p}_t$。

第二阶段回归：用被解释变量对第一阶段回归的拟合值进行回归，即 $q_t\xrightarrow{\mathrm{OLS}}\hat{p}_t$。

为什么这样做能得到好结果呢？把需求方程 $q_t=\alpha_0+\alpha_1 p_t+u_t$ 分解为

$$q_t=\alpha_0+\alpha_1\hat{p}_t+\underbrace{[u_t+\alpha_1(p_t-\hat{p}_t)]}_{\equiv\varepsilon_t} \tag{10.8}$$

命题 在第二阶段回归中，$\hat{p}_t$ 与新扰动项 $\varepsilon_t\equiv u_t+\alpha_1(p_t-\hat{p}_t)$ 不相关。

证明：由于 $\varepsilon_t\equiv u_t+\alpha_1(p_t-\hat{p}_t)$，故

$$\mathrm{Cov}(\hat{p}_t,\varepsilon_t)=\mathrm{Cov}(\hat{p}_t,u_t)+\alpha_1\mathrm{Cov}(\hat{p}_t,p_t-\hat{p}_t) \tag{10.9}$$

首先，由于 $\hat{p}_t$ 是 x_t 的线性函数（$\hat{p}_t$ 为第一阶段回归的拟合值），而 $\mathrm{Cov}(x_t,u_t)=0$（工具变量的外生性），故上式右边的第一项 $\mathrm{Cov}(\hat{p}_t,u_t)=0$。

其次，由于在第一阶段回归中，拟合值 $\hat{p}_t$ 与残差 $p_t-\hat{p}_t$ 正交（OLS 估计量的正交性），故上式右边的第二项 $\mathrm{Cov}(\hat{p}_t,p_t-\hat{p}_t)=0$。

由于第二阶段回归方程的解释变量 $\hat{p}_t$ 与扰动项 ε_t 不相关，故 2SLS 能得到一致估计。从这个例子可以看出，2SLS 的实质是把内生解释变量 p_t 分成两部分，即由工具变量 x_t 所造成的外生部分（$\hat{p}_t$）以及与扰动项相关的其余部分（$p_t-\hat{p}_t$）；然后，把被解释变量 q_t 对 p_t 中的这个外生部分（$\hat{p}_t$）进行回归，从而满足 OLS 对前定变量的要求而得到一致估计。下面的章节将对工具变量法与 2SLS 法进行更正式的推导。

例 宏观经济模型中的消费函数

$$\begin{cases}C_t=\alpha_0+\alpha_1 Y_t+\varepsilon_t\\ Y_t=C_t+I_t+G_t+X_t\end{cases} \tag{10.10}$$

其中，Y_t,C_t,I_t,G_t,X_t 分别代表国民收入、总消费、总投资、政府净支出与净出口。第一个方程为消费方程，而第二个方程为国民收入恒等式。可以证明，如果单独对消费方程进行 OLS 估计，将得到不一致的估计（参见习题）。

例 解释变量测量误差（measurement error or errors-in-variables）。假设真实模型为

$$y=\alpha+\beta x^*+\varepsilon,\quad \mathrm{Cov}(x^*,\varepsilon)=0,\quad \beta\neq 0 \tag{10.11}$$

但 x^* 无法精确观测，而只能观测到 x，二者满足如下关系：

$$x=x^*+u,\quad \mathrm{Cov}(x^*,u)=0,\quad \mathrm{Cov}(u,\varepsilon)=0 \tag{10.12}$$

其中，测量误差 u 与被测量变量 x^* 不相关，也与扰动项 ε 不相关。将表达式（10.12）代入（10.11）可得

---

① 确切地说，依然可以根据（10.7）来计算 $\hat{\alpha}_{1,\mathrm{IV}}$，但（10.6）式无法定义，而且 $\hat{\alpha}_{1,\mathrm{IV}}$ 也不是 α_1 的一致估计。

$$y=\alpha+\beta x+(\varepsilon-\beta u) \tag{10.13}$$

可以证明,新扰动项$(\varepsilon-\beta u)$与解释变量 x 存在相关性:

$$\begin{aligned}\mathrm{Cov}(x,\varepsilon-\beta u)&=\mathrm{Cov}(x^*+u,\varepsilon-\beta u)\\&=\underbrace{\mathrm{Cov}(x^*,\varepsilon)}_{=0}-\beta\underbrace{\mathrm{Cov}(x^*,u)}_{=0}+\underbrace{\mathrm{Cov}(u,\varepsilon)}_{=0}-\beta\mathrm{Cov}(u,u)\\&=-\beta\mathrm{Var}(u)\neq 0\end{aligned} \tag{10.14}$$

因此,对方程(10.13)进行 OLS 估计是不一致的,称之为"测量误差偏差"(measurement error bias)。事实上,还可以进一步确定此偏差的方向,因为

$$\begin{aligned}\hat{\beta}&=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\\&\xrightarrow{p}\frac{\mathrm{Cov}(x_i,y_i)}{\mathrm{Var}(x_i)}=\frac{\mathrm{Cov}(x_i^*+u,\alpha+\beta x^*+\varepsilon)}{\mathrm{Var}(x_i^*+u)}\\&=\frac{\beta\mathrm{Var}(x_i^*)}{\mathrm{Var}(x_i^*)+\mathrm{Var}(u)}=\beta\cdot\frac{1}{1+(\sigma_u^2/\sigma_{x^*}^2)}\end{aligned} \tag{10.15}$$

在上式中,由于 σ_u^2 与 $\sigma_{x^*}^2$ 一定为正,故 $0<\dfrac{1}{1+(\sigma_u^2/\sigma_{x^*}^2)}<1$。因此,无论真实参数 β 大于或小于 0,此偏差总是使得 $\hat{\beta}$ 的绝对值变小而趋向于 0,故也被称为"衰减偏差"(attenuation bias)或"向 0 衰减"(attenuation toward zero)。相对于 x_i^* 的方差 $\sigma_{x^*}^2$,如果测量误差 u_i 的方差 σ_u^2 越大,则$(\sigma_u^2/\sigma_{x^*}^2)$越大,$\dfrac{1}{1+(\sigma_u^2/\sigma_{x^*}^2)}\to 0$,则向 0 衰减的偏差越严重。

另一方面,如果被解释变量存在测量误差,后果却不那么严重。假设真实模型为

$$y^*=\beta x+\varepsilon,\quad \mathrm{Cov}(x,\varepsilon)=0,\quad \beta\neq 0 \tag{10.16}$$

但 y^* 无法精确观测,而只能观测到 y,二者满足如下关系:

$$y=y^*+v \tag{10.17}$$

其中,v 为测量误差。将方程(10.17)代入(10.16)可得

$$y=\beta x+(\varepsilon+v) \tag{10.18}$$

此时,只要被解释变量的测量误差 v 与解释变量 x 没有系统相关,即 $\mathrm{Cov}(x,v)=0$,则 OLS 估计量仍然是一致的,但可能会增大扰动项的方差。

10.2 工具变量法作为一种矩估计

1. 矩估计

首先以一个简单的例子来说明传统的"矩估计"(Method of Moments)方法,简记 MM。

假设随机变量 $x\sim N(\mu,\sigma^2)$,其中 μ,σ^2 为待估计的参数。因为有两个待估参数,故需要使用以下两个总体矩条件(population moment conditions):

一阶原点矩:$\mathrm{E}(x)=\mu$

二阶原点矩:$\mathrm{E}(x^2)=\mathrm{Var}(x)+[\mathrm{E}(x)]^2=\sigma^2+\mu^2$

用对应的样本矩(sample moments)来替代总体矩条件可得以下联立方程组,求解后即得到对期

望与方差的矩估计：

$$\begin{cases}\dfrac{1}{n}\sum_{i=1}^{n}x_i=\hat{\mu}\\ \dfrac{1}{n}\sum_{i=1}^{n}x_i^2=\hat{\mu}^2+\hat{\sigma}^2\end{cases}\Rightarrow\begin{cases}\hat{\mu}=\bar{x}\\ \hat{\sigma}^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\end{cases}\tag{10.19}$$

其中，$\bar{x}\equiv\dfrac{1}{n}\sum_{i=1}^{n}x_i$ 为样本均值。在上式的推导中用到了等式，$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$。

推而广之，任何随机向量 $\boldsymbol{x}$ 的函数 $f(\boldsymbol{x})$ 的期望 $\mathrm{E}[f(\boldsymbol{x})]$ 都被称为“总体矩”。事实上，OLS 也是一种矩估计。利用解释变量与扰动项的正交性，可以得到以下总体矩条件：

$$\begin{aligned}\mathrm{E}[\boldsymbol{x}_i\varepsilon_i]=\boldsymbol{0}&\Rightarrow\mathrm{E}[\boldsymbol{x}_i(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})]=\boldsymbol{0}&&(\text{代入 }\varepsilon_i=y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\\&\Rightarrow\mathrm{E}(\boldsymbol{x}_iy_i)=\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\boldsymbol{\beta}&&(\text{展开、移项})\\&\Rightarrow\boldsymbol{\beta}=[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\mathrm{E}(\boldsymbol{x}_iy_i)&&(\text{假设}[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}\text{存在，求解}\boldsymbol{\beta})\end{aligned}$$

以样本矩替代上式中的总体矩，即可得到矩估计：

$$\hat{\boldsymbol{\beta}}_{\mathrm{MM}}=\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_iy_i\right)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\tag{10.20}$$

显然，这就是 OLS 估计量。从以上推导也可以看出，解释变量与扰动项的正交性是 OLS 能够成立的重要前提。

2. 工具变量法作为一种矩估计

假设回归模型为

$$y_i=\beta_1x_{i1}+\cdots+\beta_{K-1}x_{i,K-1}+\beta_Kx_{iK}+\varepsilon_i\tag{10.21}$$

其中，只有最后一个解释变量 x_{iK} 为内生变量，即 $\mathrm{Cov}(x_{iK},\varepsilon_i)\neq0$，因此 OLS 是不一致的。假设有一个有效工具变量 w 满足 $\mathrm{Cov}(x_{iK},w_i)\neq0$（相关性），以及 $\mathrm{Cov}(w_i,\varepsilon_i)=0$（外生性）。由于 $x_1,\cdots,x_{K-1}$ 不是内生变量，故可以把自身作为自己的工具变量（因为满足工具变量法的两个条件）。

记解释向量 $\boldsymbol{x}_i\equiv(x_{i1}\cdots x_{i,K-1}x_{iK})'$，则原模型为 $y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i$。记工具向量 $\boldsymbol{z}_i\equiv(z_{i1}\cdots z_{i,K-1}z_{iK})'\equiv(x_{i1}\cdots x_{i,K-1}w_i)'$。定义 $\boldsymbol{g}_i\equiv\boldsymbol{z}_i\varepsilon_i$。由于工具向量与扰动项正交，故 $\mathrm{E}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{z}_i\varepsilon_i)=\boldsymbol{0}$ 为“总体矩条件”或“正交条件”（orthogonality condition）。由此可得

$$\begin{aligned}\mathrm{E}(\boldsymbol{z}_i\varepsilon_i)=\boldsymbol{0}&\Rightarrow\mathrm{E}[\boldsymbol{z}_i(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})]=\boldsymbol{0}&&(\text{代入 }\varepsilon_i=y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\\&\Rightarrow\mathrm{E}(\boldsymbol{z}_iy_i)=[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]\boldsymbol{\beta}&&(\text{展开、移项})\\&\Rightarrow\boldsymbol{\beta}=[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}\mathrm{E}(\boldsymbol{z}_iy_i)&&(\text{假设}[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}\text{存在})\end{aligned}$$

以样本矩代替上式中的总体矩，即可得到工具变量估计量，

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}=\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_iy_i\right)=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}\tag{10.22}$$

其中，$\boldsymbol{Z}\equiv(\boldsymbol{z}_1\cdots\boldsymbol{z}_{n-1}\boldsymbol{z}_n)'$。显然，OLS 也是一种工具变量法。这是因为，如果 $\boldsymbol{x}_i$ 全部是前定变量，则可以将自己作为工具变量，即 $\boldsymbol{z}_i=\boldsymbol{x}_i$，$\boldsymbol{Z}=\boldsymbol{X}$。因此，$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}=\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$。

工具变量法具有怎样的大样本性质呢？

命题 如果秩条件“$\mathrm{rank}[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]=K$”成立（即方阵 $\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$ 满秩），则在一定的正则条件下，$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 是 $\boldsymbol{\beta}$ 的一致估计，且 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 服从渐近正态分布。

证明：因为 $\mathrm{rank}[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]=K$，故 $[\mathrm{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$ 存在。因此，抽样误差

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}-\boldsymbol{\beta}=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta}=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{x}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i\right)$$

$$= \boldsymbol{S}_{ZX}^{-1}\,\bar{\boldsymbol{g}} \xrightarrow{p} [\mathrm{E}(z_i \boldsymbol{x}_i')]^{-1}\underbrace{\mathrm{E}(\boldsymbol{g}_i)}_{=\mathbf{0}} = \mathbf{0} \tag{10.23}$$

其中,$\boldsymbol{S}_{ZX} \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{x}_i'$,$\bar{\boldsymbol{g}} \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i$。根据与“大样本 OLS 理论”(第 5 章)类似的假定与推导,可以证明,$\sqrt{n}\,\bar{\boldsymbol{g}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{S})$,其中 $\boldsymbol{S} \equiv \mathrm{E}(\boldsymbol{g}_i \boldsymbol{g}_i') = \mathrm{E}(\varepsilon_i^2 z_i z_i')$。进一步,工具变量估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 渐近地服从正态分布,即 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathrm{Avar}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}))$,其中渐近方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) = [\mathrm{E}(z_i \boldsymbol{x}_i')]^{-1}\boldsymbol{S}[\mathrm{E}(\boldsymbol{x}_i z_i')]^{-1}$。

秩条件 $\mathrm{rank}[\mathrm{E}(z_i \boldsymbol{x}_i')] = K$ 意味着,工具变量 w_i 与解释变量 x_i 相关。如果不相关,则秩条件无法满足。以一元回归为例,此时,$K = 2$,$\boldsymbol{x}_i = (1 \quad x_i)'$,$z_i = (1 \quad w_i)'$,则 $\mathrm{E}(z_i \boldsymbol{x}_i') = \mathrm{E}\left[\begin{pmatrix} 1 \\ w_i \end{pmatrix}(1 \quad x_i)\right] = \mathrm{E}\begin{bmatrix} 1 & x_i \\ w_i & w_i x_i \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{E}(x_i) \\ \mathrm{E}(w_i) & \mathrm{E}(w_i x_i) \end{bmatrix}$,因此,

$$\mathrm{rank}[\mathrm{E}(z_i \boldsymbol{x}_i')] = K = 2 \quad \Leftrightarrow \text{行列式} \begin{vmatrix} 1 & \mathrm{E}(x_i) \\ \mathrm{E}(w_i) & \mathrm{E}(w_i x_i) \end{vmatrix} \neq 0$$

$$\Leftrightarrow \mathrm{E}(w_i x_i) - \mathrm{E}(w_i)\mathrm{E}(x_i) \neq 0$$

$$\Leftrightarrow \mathrm{Cov}(w_i, x_i) \neq 0,\text{即 } w_i \text{ 与 } x_i \text{ 相关。}$$

阶条件:显然,满足秩条件的必要条件是 z_i 中至少包含 K 个变量,即不在方程中出现的工具变量个数不能少于方程中内生解释变量的个数。称此条件为“阶条件”(order condition)。

根据是否满足阶条件可分为三种情况:

(1) 不可识别(unidentified):工具变量个数小于内生解释变量个数;

(2) 恰好识别(just or exactly identified):工具变量个数等于内生解释变量个数;

(3) 过度识别(overidentified):工具变量个数大于内生解释变量个数。

以上介绍的工具变量法仅适用于“恰好识别”的情形。在“过度识别”的情况下,$\boldsymbol{Z}'\boldsymbol{X}$ 不是方阵,$(\boldsymbol{Z}'\boldsymbol{X})^{-1}$ 不存在,也就无法定义工具变量估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$。解决方法之一是扔掉“多余”的工具变量。但这显然不是有效的做法,因为被扔掉的工具变量包含有用的信息。有效的做法是二阶段最小二乘法。

10.3 二阶段最小二乘法

显然,多个工具变量的线性组合仍然是工具变量,因为仍满足工具变量的两个条件(相关性与外生性)。如果生成工具变量的 K 个线性组合,则又回到恰好识别的情形。然而,什么样的线性组合才最有效率呢?可以证明,在球型扰动项的假定下,由二阶段最小二乘法(2SLS)所提供的工具变量线性组合是所有线性组合中最渐近有效的[①]。这个结论类似于小样本理论中的高斯 - 马尔可夫定理。

① 因为在条件同方差的情况下,最优 GMM 还原为 2SLS,而最优 GMM 是渐近有效的,参见下文。

第一阶段(分离出内生变量的外生部分)

将每个解释变量 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_K$ 分别对所有 L 个工具变量 $\{z_1,z_2,\cdots,z_L\}$ 作 OLS 回归,其中第 k 个解释变量 $\boldsymbol{x}_k\equiv(x_{1k}\cdots x_{nk})'_{n\times 1}$,$k=1,\cdots,K$(不同于第 3 章对第 i 个观测数据 $\boldsymbol{x}_i$ 的定义)。得到拟合值

$$\hat{\boldsymbol{x}}_1=\boldsymbol{P}\boldsymbol{x}_1,\quad \hat{\boldsymbol{x}}_2=\boldsymbol{P}\boldsymbol{x}_2,\quad \cdots,\quad \hat{\boldsymbol{x}}_K=\boldsymbol{P}\boldsymbol{x}_K \tag{10.24}$$

其中,$\boldsymbol{P}\equiv\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$为对应于 $\boldsymbol{Z}$ 的投影矩阵。写成矩阵形式,可以定义

$$\hat{\boldsymbol{X}}\equiv(\hat{\boldsymbol{x}}_1\ \hat{\boldsymbol{x}}_2\cdots\hat{\boldsymbol{x}}_K)=\boldsymbol{P}(\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_K)=\boldsymbol{P}\boldsymbol{X}=\boldsymbol{Z}[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}] \tag{10.25}$$

第二阶段(使用此外生部分进行回归)

由于 $\hat{\boldsymbol{X}}$ 是 $\{z_1,z_2,\cdots,z_L\}$ 的线性组合(参见第一阶段回归),故 $\hat{\boldsymbol{X}}$ 恰好包含 K 个工具变量。使用 $\hat{\boldsymbol{X}}$ 为工具变量对原模型 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 进行工具变量法估计:

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}=(\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}=(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \tag{10.26}$$

上式的最后一个等号能成立,是由于 $\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}=(\boldsymbol{P}\boldsymbol{X})'(\boldsymbol{P}\boldsymbol{X})=\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{X}=\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{X}=\hat{\boldsymbol{X}}'\boldsymbol{X}$,其中,投影矩阵 $\boldsymbol{P}$ 为对称幂等矩阵,即 $\boldsymbol{P}'=\boldsymbol{P},\boldsymbol{P}^2=\boldsymbol{P}$。因此,可以将 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 视为把 $\boldsymbol{y}$ 对 $\hat{\boldsymbol{X}}$ 进行 OLS 回归而得到,故名"二阶段最小二乘法"。需要注意的是,第二阶段回归所得到的残差为 $\boldsymbol{e}_2\equiv\boldsymbol{y}-\hat{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$,而原方程残差却是 $\boldsymbol{e}\equiv\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$。因此,执行 2SLS 最好不要自己进行两次手工回归,而是直接使用 Stata 的命令(相信 Stata 会计算正确的残差!)。

由于 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 的表达式在形式上完全类似于 OLS 估计量,故在条件同方差的假设下,$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ 的协方差矩阵估计量为 $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})=s^2(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}$,其中 $s^2\equiv\dfrac{\boldsymbol{e}'\boldsymbol{e}}{n-K}$,这是 Stata 的默认方法。在存在异方差的情况下,则应该使用稳健的协方差矩阵估计量,即 $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})=(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\left(\sum_{i=1}^{n}e_i^2\,\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i'\right)(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}$。

将 $\hat{\boldsymbol{X}}=\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$ 代入方程(10.26),可得到 2SLS 的最终表达式:

$$\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}=(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y}=[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} \tag{10.27}$$

2SLS 的 Stata 命令为

```
ivregress 2sls depvar [varlist1] (varlist2 = instlist)
```

其中,"depvar"为被解释变量,"varlist1"为外生解释变量,"varlist2"为内生解释变量,而"instlist"为工具变量。举个具体例子,假设回归模型为 $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$,其中 x_1 为外生变量,x_2 为内生变量,而工具变量为 z_1,z_2,则 2SLS 的 Stata 命令应为

```
ivregress 2sls y x1 (x2 = z1 z2)  (使用普通标准误)
ivregress 2sls y x1 (x2 = z1 z2),r first
```

其中,选择项"r"表示使用异方差稳健的标准误,选择项"first"表示显示第一阶段的回归。

10.4 有关工具变量的检验

在使用工具变量法时,必须对工具变量的有效性进行检验。如果工具变量不有效,则可能导致估计不一致,或估计量的方差过大。为此,需要进行一系列的检验。

1. 不可识别检验

使用工具变量法的前提之一是秩条件成立,即 $\text{rank}[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]=K$(满列秩),其中 $\boldsymbol{z}_i\equiv(z_{i1}\cdots z_{iL})'$($L$ 个工具变量),$\boldsymbol{x}_i\equiv(x_{i1}\cdots x_{iK})'$($K$ 个解释变量),$\boldsymbol{z}_i$ 与 $\boldsymbol{x}_i$ 可有重叠元素,且 $L\geqslant K$(阶条件)。如果矩阵 $\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$ 的列秩小于 K,则不可识别。对于秩条件是否成立,可进行"不可识别检验"(underidentification test)。其原假设为"H_0 : $\text{rank}[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]=K-1$",而替代假设为"$H_1$: $\text{rank}[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]=K$"。

在扰动项为 iid 的假设下(这要求不存在异方差),可以使用"Anderson LM 统计量"(Anderson,1951),其渐近分布为 $\chi^2(L-K+1)$。如果不作 iid 扰动项的假设(允许存在异方差),则应使用"Kleibergen-Paap rk LM 统计量(Kleibergen-Paap,2006),其渐近分布也是 $\chi^2(L-K+1)$。秩条件成立的直观意义是工具变量与解释变量相关。因此,针对秩条件的不可识别检验也可在一定程度上验证是否存在弱工具变量,但不能取代对弱工具变量的检验(参见下文)。

为了进行不可识别检验,可通过命令"`ssc install ivreg2`"下载非官方命令 `ivreg2`。

2. 弱工具变量检验

如果工具变量 $\boldsymbol{z}$ 与内生解释变量 $\boldsymbol{x}$ 完全不相关,则无法使用工具变量法,因为 $[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$不存在①。如果 $\boldsymbol{z}$ 与 $\boldsymbol{x}$ 仅仅微弱地相关,则可大致认为 $[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$很大,导致工具变量法估计量的渐近方差 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{IV}})=[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}\boldsymbol{S}[\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')]^{-1}$变得很大。直观上,由于 $\boldsymbol{z}$ 中仅包含很少与 $\boldsymbol{x}$ 有关的信息,利用这部分信息进行的工具变量法估计就不准确,即使样本容量很大也很难收敛到真实的参数值。这种工具变量称为"弱工具变量"(weak instruments)。弱工具变量的后果类似于样本容量过小,会导致 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的小样本性质变得很差,而 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的大样本分布也可能离正态分布相去甚远,致使基于大样本理论的统计推断失效。

判断弱工具变量的方法主要有以下四种。方法之一为使用"偏 R^2"。假设回归模型为

$$y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+x_2\beta_2+\varepsilon \tag{10.28}$$

其中,只有 x_2 为内生解释变量。记工具变量为$(\boldsymbol{x}_1\boldsymbol{z}_2)$,其中 $\boldsymbol{z}_2$ 为方程外的工具变量。在 2SLS 的第一阶段回归中,$x_2\xrightarrow{\text{OLS}}\boldsymbol{x}_1,\boldsymbol{z}_2$,其 R^2 包含了内生变量 x_2 与工具变量 $\boldsymbol{z}_2$ 相关性的信息,但也可能由于 x_2 与 $\boldsymbol{x}_1$ 的相关性所造成。为此,应该使用滤去 $\boldsymbol{x}_1$ 影响的"偏 R^2"(partial R^2),记为 R_p^2。具体操作步骤如下:首先把 x_2 对 $\boldsymbol{x}_1$ 回归,$x_2\xrightarrow{\text{OLS}}\boldsymbol{x}_1$,记其残差为 e_{x_2},代表 x_2 中不能由 $\boldsymbol{x}_1$ 解释的部分;其次,把 $\boldsymbol{z}_2$ 对 $\boldsymbol{x}_1$ 回归,$\boldsymbol{z}_2\xrightarrow{\text{OLS}}\boldsymbol{x}_1$,记其残差为 $\boldsymbol{e}_{z_2}$,代表 $\boldsymbol{z}_2$ 中不能由 $\boldsymbol{x}_1$ 解释的部分;最后,对两个残差进行回归,即 $e_{x_2}\xrightarrow{\text{OLS}}\boldsymbol{e}_{z_2}$,所得可决系数就是 R_p^2。然而,具体 R_p^2 多低才构成弱工具变量,目前尚无共识。Shea (1997)将 R_p^2 推广到多个内生解释变量的情形,Stata 称之为"Shea's partial R^2"。

判断弱工具变量的方法之二为,在第一阶段回归中,$x_2=\boldsymbol{x}_1'\boldsymbol{\gamma}_1+\boldsymbol{z}_2'\boldsymbol{\gamma}_2+error$,检验原假设"$H_0$: $\boldsymbol{\gamma}_2=\boldsymbol{0}$"(即工具变量 $\boldsymbol{z}_2$ 的系数为 0)。一个经验规则(rule of thumb)是,如果此检验的 F 统计量大于 10,则可拒绝"存在弱工具变量"的原假设,不必担心弱工具变量问题。在多个内生解释变量的情况下,将有多个第一阶段回归,故有多个 F 统计量。此时,可以使用 Stock and Yogo (2005)提出的"最小特征值统计量"(minimum eigenvalue statistic)。如果只有一个内生解释变量,则该统计量还原为 F 统计量。Stata 提供了最小特征值统计量的临界值。

---

① 假设 $\boldsymbol{z}$ 与 $\boldsymbol{x}$ 的维度相同。

判断弱工具变量的方法之三为，如果假设扰动项为 iid，则可以使用“Cragg-Donald Wald F 统计量”（Cragg and Donald，1993），其临界值由 Stock and Yogo（2005）提供。

判断弱工具变量的方法之四为，如果不作 iid 扰动项的假设，则应使用“Kleibergen-Paap Wald rk F 统计量”，其临界值也来自 Stock and Yogo（2005）。

检验弱工具变量的 Stata 命令为

```
estat firststage,all forcenonrobust
```

该命令将显示与弱工具变量有关的第一个阶段回归统计量及临界值。选择项“all”表示显示每个内生变量的统计量，而非仅仅是所有内生变量综合的统计量。选择项“forcenonrobust”表示，即使在进行工具变量法时用了稳健标准误（选择项robust），也仍然允许计算“estat firststage”中的统计量（这些统计量基于同方差的假设）。

如果要计算“Cragg-Donald Wald F 统计量”（方法之三）或“Kleibergen-Paap Wald rk F 统计量”（方法之四），则应使用非官方命令“ivreg2”。

解决弱工具变量问题的方法包括：

（i）寻找更强的工具变量；

（ii）使用对弱工具变量更不敏感的“有限信息最大似然估计法”（Limited Information Maximum Likelihood Estimation，简记 LIML，参见第 24 章）；在大样本下，LIML 与 2SLS 是渐近等价的，但在存在弱工具变量的情况下，LIML 的小样本性质可能优于 2SLS。

LIML 的 Stata 命令为

```
ivregress liml depvar [varlist 1] (varlist2 = instlist)
```

（iii）如果有较多工具变量，可舍弃弱工具变量。在选择舍弃哪个工具变量时，可以进行“冗余检验”（redundancy test）。冗余工具变量（redundant instruments）的含义是，使用这些工具变量不会提高估计量的渐近效率（asymptotic efficiency）。其具体表现包括，在第一阶段回归中，这些工具变量的系数不显著；或者这些工具变量与内生解释变量的“偏相关系数”（partial correlations）为 0 或接近于 0（偏相关系数的平方即为偏 R^2）。命令“ivreg2”提供了一个选择项“redundant (varlist)”来进行此冗余检验。基于偏相关的思想，该检验考察内生变量与可能的冗余工具变量之间在过滤掉（partial out）其他工具变量影响之后的矩阵交叉乘积（matrix cross product）的秩是否为 0。该冗余检验的原假设是，指定的工具变量为多余的。该检验统计量的渐近分布为 χ^2 分布，自由度为“内生变量个数”乘以“冗余工具变量个数”。

3. 过度识别检验

在恰好识别的情况下，目前公认无法检验工具变量的外生性，即工具变量与扰动项不相关。在这种情况下，只能进行定性讨论或依赖于专家的意见。定性讨论通常基于以下逻辑：如果工具变量是外生的，则其对被解释变量发生影响的唯一渠道就是通过内生变量，除此以外别无其他渠道。由于此唯一渠道（内生变量）已被包括在回归方程中，故工具变量不会再出现在被解释变量的扰动项中，或对此扰动项有影响。此条件被称为“排他性约束”（exclusion restriction），因为它排除了工具变量除了通过内生变量而影响被解释变量的所有其他渠道。在实际操作中，则需要找出工具变量影响被解释变量的所有其他可能渠道，然后一一排除，才能比较信服地说明工具变量的外生性。

在过度识别的情况下，则可进行“过度识别检验”（overidentification test）。此检验的大前提（maintained hypothesis）是该模型至少是恰好识别的，即有效工具变量至少与内生解释变量一样多。在此大前提下，过度识别检验的原假设为“H_0：所有工具变量都是外生的”。如果拒绝该原

假设,则认为至少某个变量不是外生的,与扰动项相关。

不失一般性,假设前$(K-r)$个解释变量$\{x_1,\cdots,x_{K-r}\}$为外生解释变量,而后 r 个解释变量$\{x_{K-r+1},\cdots,x_K\}$为内生解释变量。假设共有 m 个方程外的工具变量$\{z_1,\cdots,z_m\}$,其中 $m>r$。把工具变量法的残差对所有外生变量(即所有外生解释变量与工具变量)进行以下辅助回归:

$$e_{i,\mathrm{IV}}=\gamma_1 x_{i1}+\cdots+\gamma_{K-r}x_{i,K-r}+\delta_1 z_{i1}+\cdots+\delta_m z_{im}+error_i \tag{10.29}$$

将工具变量法的残差 $e_{i,\mathrm{IV}}$视为对扰动项 ε_i 的估计,则"扰动项 ε 与工具变量$\{z_1,\cdots,z_m\}$无关"的原假设可以写为"$H_0:\delta_1=\cdots=\delta_m=0$"。记此辅助回归的可决系数为 R^2,则 Sargan 统计量为

$$nR^2\xrightarrow{d}\chi^2(m-r) \tag{10.30}$$

其中,χ^2 分布的自由度$(m-r)$为过度识别约束的个数(即"多余"的工具变量个数)[①]。显然,如果恰好识别,则 $m-r=0$(自由度为 0),$\chi^2(0)$无定义,故无法使用这个"过度识别检验"。这个检验背后的直观思想是,在过度识别的情况下,可以使用不同的工具变量组合来进行工具变量法估计;而如果所有工具变量都有效,则这些工具变量估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$都将收敛到相同的真实参数 $\boldsymbol{\beta}$。为此,可以检验不同的工具变量估计量之间的差是否收敛于 0;如果不是,则说明这些工具变量不全是有效的。在恰好识别的情况下,只有唯一的工具变量估计量,无法进行这种比较,故过度识别检验失效。另一方面,如果拒绝原假设,过度识别检验并不能告诉我们,哪些工具变量是无效的。

需要注意的是,即使接受了过度识别的原假设,也不能证明这些工具变量的外生性。事实上,过度识别检验的成立有一个大前提,即至少该模型是恰好识别的。此大前提无法检验,只能假定其成立。比如,如果只有一个内生变量,则在进行过度识别检验时,我们隐含地假定至少有一个工具变量是外生的,然后检验所有其他工具变量的外生性。上文提到,过度识别检验的直观思想是检验由不同工具变量组合而生成的 IV 估计量是否收敛到同一值。如果所有工具变量都不是外生的,则即使由它们生成的 IV 估计量差别很小,也只能说明这些 IV 估计量都收敛到了一个错误的值。反之,如果至少有一个工具变量是外生的,而且这些 IV 估计量都收敛到同一值,则可以认为所有工具变量都是外生的。

过度识别检验的 Stata 命令为"`estat overid`"。

4. 究竟该用 OLS 还是工具变量法:对解释变量内生性的检验

使用工具变量法的前提是存在内生解释变量,这也需要检验。如何从统计上检验解释变量是否为内生呢?由于扰动项不可观测,故无法直接检验解释变量与扰动项的相关性。但如果找到有效的工具变量,则可以借助工具变量来检验解释变量的内生性。

假设存在方程外的工具变量。如果所有解释变量都是外生变量,则 OLS 比工具变量法更有效[②]。在这种情况下使用工具变量法,虽然估计量仍然是一致的,但相当于"无病用药",反而会增大估计量的方差。反之,如果存在内生解释变量,则 OLS 是不一致的,而工具变量法是一致的。

"豪斯曼检验"(Hausman specification test)(Hausman,1978)的原假设为"H_0:所有解释变量均为外生变量"。如果 H_0 成立,则 OLS 与工具变量法都是一致的,即在大样本下 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$与 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$都收

① 虽然工具变量共有 m 个,但其中的 r 个被用于"模型识别"(model identification),故损失了 r 个自由度。

② 此时,如果满足球型扰动项的假定,则 OLS 是 BLUE,而工具变量法不是。另外,当所有解释变量均为外生时,2SLS 就是 OLS,而 2SLS 是渐近有效的。

敛于真实的参数值 $\boldsymbol{\beta}$，因此 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$（称为“对比向量”，vector of contrast）依概率收敛于 $\mathbf{0}$。反之，如果 H_0 不成立，则工具变量法一致而 OLS 不一致，故 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$ 不会收敛于 $\mathbf{0}$。豪斯曼检验正是基于这一思想进行的。如果 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$ 的距离很大，则倾向于拒绝原假设。根据沃尔德检验原理，以二次型来度量此距离可得

$$(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})'\boldsymbol{D}^{-}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})\xrightarrow{d}\chi^2(r) \tag{10.31}$$

其中，$\boldsymbol{D}\equiv\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})}$，而 $\boldsymbol{D}^{-}$ 为 $\boldsymbol{D}$ 的广义逆矩阵（因为 $\boldsymbol{D}$ 不一定可逆；但如果 $\boldsymbol{D}$ 正定，则可逆）[①]。当 $\boldsymbol{D}$ 可逆时，广义逆矩阵就是一般的逆矩阵，即 $\boldsymbol{D}^{-}=\boldsymbol{D}^{-1}$。$r$ 为内生解释变量的个数（不包括外生解释变量）。容易证明

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})=\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV})+\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})-2\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{IV},\hat{\boldsymbol{\beta}}_{OLS}) \tag{10.32}$$

但上式中的 $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{IV},\hat{\boldsymbol{\beta}}_{OLS})$ 不易计算。如果在 H_0 成立的情况下，OLS 是最有效率的，则可以证明 $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{IV},\hat{\boldsymbol{\beta}}_{OLS})=\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})$（参见附录），故

$$\boldsymbol{D}=\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV})}-\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})} \tag{10.33}$$

如果拒绝 H_0，则认为存在内生解释变量，应该使用工具变量法；反之，如果接受 H_0，则认为不存在内生解释变量，应该使用 OLS。

豪斯曼检验的 Stata 命令为

```
reg y x1 x2
estimates store ols                    (存储 OLS 的结果)
ivregress 2sls y x1 (x2 = z1 z2)       (假设怀疑 x2 为内生变量)
estimates store iv                     (存储 2SLS 的结果)
hausman iv ols,constant sigmamore      (根据存储的结果进行豪斯曼检验)
```

其中，选择项“sigmamore”表示统一使用更有效的估计量（即 OLS）所对应的残差来计算 $\hat{\sigma}^2$。这样有助于保证根据样本数据计算的 $[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV})}-\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})}]$ 为正定矩阵。选择项“constant”表示 $\hat{\boldsymbol{\beta}}_{IV}$ 与 $\hat{\boldsymbol{\beta}}_{OLS}$ 中都包括常数项（默认不包含常数项）。

上述检验的缺点是，它假设在 H_0 成立的情况下，OLS 是最有效率的。然而，如果存在异方差，OLS 并不是最有效率的（不是 BLUE）。故传统的豪斯曼检验不适用于异方差的情形。

解决方法之一为，通过“自助法”（bootstrap），即计算机模拟“再抽样”（resampling）的方法来计算 $\boldsymbol{D}\equiv\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})}$，参见第 19 章。

解决方法之二为，使用“杜宾 - 吴 - 豪斯曼检验”（Durbin - Wu - Hausman Test，简记 DWH）[②]，该检验在异方差的情况下也适用，更为稳健。首先考虑只有一个内生解释变量的情形。假设回归模型为，$y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+x_2\beta_2+\varepsilon$，其中 x_2 为唯一的内生解释变量。记工具变量为 $\boldsymbol{z}\equiv(\boldsymbol{x}_1\ \ z_2)$，其中 z_2 为方程外的工具变量。考虑 2SLS 的第一阶段回归，即 $x_2=\boldsymbol{x}_1'\boldsymbol{\gamma}+\boldsymbol{z}'\boldsymbol{\delta}+v$。由于工具变量 $\boldsymbol{z}$ 与 ε 不相关，故

---

① $\boldsymbol{D}$（不一定为方阵）的广义逆矩阵 $\boldsymbol{D}^{-}$ 需要满足四个条件，即 $\boldsymbol{D}\boldsymbol{D}^{-}\boldsymbol{D}=\boldsymbol{D}$，$\boldsymbol{D}^{-}\boldsymbol{D}\boldsymbol{D}^{-}=\boldsymbol{D}^{-}$，$\boldsymbol{D}\boldsymbol{D}^{-}$ 为对称矩阵，$\boldsymbol{D}^{-}\boldsymbol{D}$ 为对称矩阵，参见 Poirier(1995, p. 630)。

② 参见 Durbin(1954)，Wu(1973) 与 Hausman(1978)。

$$\mathrm{E}(x_2\varepsilon)=\mathrm{E}[(\boldsymbol{x}_1'\boldsymbol{\gamma}+\boldsymbol{z}'\boldsymbol{\delta}+v)\varepsilon]=\underbrace{\mathrm{E}(\boldsymbol{x}'_1\boldsymbol{\gamma}\varepsilon)}_{=0}\underbrace{\mathrm{E}(\boldsymbol{z}'\boldsymbol{\delta}\varepsilon)}_{=0}+\mathrm{E}(v\varepsilon)=\mathrm{E}(v\varepsilon) \tag{10.34}$$

这意味着

$$\text{“}x_2\text{ 为内生变量”}\Leftrightarrow\text{“}\mathrm{E}(x_2\varepsilon)\neq 0\text{”}\Leftrightarrow\text{“}\mathrm{E}(v\varepsilon)\neq 0\text{”}$$

故只需要检验第一阶段回归的扰动项 v 是否与原模型的扰动项 ε 相关即可。因此,考虑以下回归模型:

$$\varepsilon=\rho v+\xi \tag{10.35}$$

其中,ρ 为 ε 对 v 的回归系数(没有常数项,因为扰动项的期望值为 0)。如果 ε 与 v 不相关,则 $\rho=0$。将方程(10.35)代入原模型可得

$$y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+x_2\beta_2+\rho v+\xi \tag{10.36}$$

由于 v 不可观测,故使用第一阶段回归的残差 $\hat{v}$ 来代替 v,进行以下辅助回归:

$$y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+x_2\beta_2+\rho\hat{v}+error \tag{10.37}$$

然后对原假设“$H_0:\rho=0$”进行 t 检验。如果拒绝“$H_0:\rho=0$”,则认为存在内生解释变量;否则,认为所有解释变量均为外生。考虑到可能存在异方差,则需要在作 t 检验时使用稳健标准误。

如果存在多个内生解释变量,即 $y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+\boldsymbol{x}_2'\boldsymbol{\beta}_2+\varepsilon$,则在第一阶段回归中可以得到与内生解释变量 $\boldsymbol{x}_2$ 相对应的多个残差 $\hat{\boldsymbol{v}}$,进行以下辅助回归:

$$y=\boldsymbol{x}_1'\boldsymbol{\beta}_1+\boldsymbol{x}_2'\boldsymbol{\beta}_2+\hat{\boldsymbol{v}}'\boldsymbol{\rho}+error \tag{10.38}$$

然后对原假设“$H_0:\boldsymbol{\rho}=\mathbf{0}$”进行 F 检验即可。同样地,如果担心存在异方差,则可在作 F 检验时使用稳健标准误。对于 DWH 检验,可以在 Stata 中依以上步骤手工进行,也可以使用命令“`estat endogenous`”来直接进行。

10.5 GMM 的假定

在球型扰动项的假定下,2SLS 是最有效率的。但如果扰动项存在异方差或自相关,则存在更有效的方法,即“广义矩估计”(Generalized Method of Moments,简记 GMM)。在某种意义上,GMM 之于 2SLS,正如 GLS 之于 OLS。

首先,引入以下关于 GMM 的假定(类似于第 5 章“大样本 OLS”的系列假定)。

假定 10.1 线性假定(linearity)

$$y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\quad(i=1,\cdots,n) \tag{10.39}$$

其中,$\boldsymbol{x}_i\equiv(x_{i1}x_{i2}\cdots x_{iK})'$为第 i 个观测数据。

假定 10.2 渐近独立的平稳过程

记 L 维工具变量为 $\boldsymbol{z}_i$(可能与 $\boldsymbol{x}_i$ 有重叠部分),$\boldsymbol{w}_i$ 由 $\{y_i,\boldsymbol{x}_i,\boldsymbol{z}_i\}$ 中不重复的变量构成且不含常数项。随机过程 $\{\boldsymbol{w}_i\}$ 为渐近独立的平稳过程。

假定 10.3 工具变量的正交性

所有工具变量 $\boldsymbol{z}_i$ 均为“前定”,即与同期扰动项正交。定义 L 维列向量 $\boldsymbol{g}_i\equiv\boldsymbol{z}_i\varepsilon_i$[①],则 $\mathrm{E}(\boldsymbol{g}_i)=\mathrm{E}(\boldsymbol{z}_i\varepsilon_i)=\mathbf{0}$。

假定 10.4 秩条件

① 此处 $\boldsymbol{g}_i$ 的定义与第 5 章“大样本 OLS”中不同。

$L \times K$ 维矩阵 $\mathrm{E}(\boldsymbol{z}_i \boldsymbol{x}_i')$ 满列秩，即 $\mathrm{rank}[\mathrm{E}(\boldsymbol{z}_i \boldsymbol{x}_i')] = K$。记 $\boldsymbol{\Sigma}_{ZX} \equiv \mathrm{E}(\boldsymbol{z}_i \boldsymbol{x}_i')$。

假定 10.5 $\{\boldsymbol{g}_i\}$ 为鞅差分序列，其协方差矩阵 $\boldsymbol{S} \equiv \mathrm{E}(\boldsymbol{g}_i \boldsymbol{g}_i') = \mathrm{E}(\varepsilon_i^2 \boldsymbol{z}_i \boldsymbol{z}_i')$ 为非退化矩阵。

假定 10.6 四阶矩 $\mathrm{E}[(x_{ik} z_{ij})^2]$ 存在且有限，$\forall\, i, j, k$（finite fourth moments）。

10.6 GMM 的推导

与总体矩条件 $\mathrm{E}(\boldsymbol{g}_i) = \mathrm{E}(\boldsymbol{z}_i \varepsilon_i) = \boldsymbol{0}$ 相对应的样本矩条件为

$$\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}) \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}) = \boldsymbol{0} \tag{10.40}$$

将上式看成一个联立方程组，则未知数 $\hat{\boldsymbol{\beta}}$ 共有 K 个，而方程个数为 L 个（$\boldsymbol{z}_i$ 的维度）。如果 $L < K$，为不可识别，则 $\hat{\boldsymbol{\beta}}$ 有无穷多解。如果 $L = K$，为恰好识别，则 $\hat{\boldsymbol{\beta}}$ 有唯一解，即 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$。如果 $L > K$，为过度识别，则 $\hat{\boldsymbol{\beta}}$ 无解。此时传统的矩估计法行不通。既然无法找到 $\hat{\boldsymbol{\beta}}$ 使得 $\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$，脑筋急转弯一下，总可以找到 $\hat{\boldsymbol{\beta}}$，使得向量 $\boldsymbol{g}_n(\hat{\boldsymbol{\beta}})$ 尽可能地接近 $\boldsymbol{0}$，比如，使二次型 $(\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}))'(\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}))$ 最小。更一般地，可以用一个“权重矩阵”（weighting matrix）$\boldsymbol{W}$ 来构成二次型。假设 $\hat{\boldsymbol{W}}$ 为一个 $L \times L$ 维对称正定矩阵（可以是依赖于样本的随机矩阵，故用 $\hat{\boldsymbol{W}}$ 来表示），而且 $\operatorname{plim}_{n \to \infty} \hat{\boldsymbol{W}} = \boldsymbol{W}$，其中 $\boldsymbol{W}$ 为非随机的对称正定矩阵。定义最小化的目标函数为

$$\min_{\hat{\boldsymbol{\beta}}} J(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{W}}) \equiv n(\boldsymbol{g}_n(\hat{\boldsymbol{\beta}}))' \hat{\boldsymbol{W}} (\boldsymbol{g}_n(\hat{\boldsymbol{\beta}})) \tag{10.41}$$

其中，因子 n 只是为了统计量计算方便而加上的，不影响最小化。定义“GMM 估计量”为此无约束二次型最小化问题的解（Hansen，1982）：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\hat{\boldsymbol{W}}) \equiv \operatorname*{argmin}_{\hat{\boldsymbol{\beta}}} J(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{W}}) \tag{10.42}$$

其中，“argmin”（argument of the minimum）表示能使 $J(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{W}})$ 最小化的 $\hat{\boldsymbol{\beta}}$ 的取值。显然，GMM 估计量取决于权重矩阵 $\hat{\boldsymbol{W}}$。对 $\hat{\boldsymbol{W}}$ 的自由选择是 GMM 的最大优点之一，因为可以通过最优地选择 $\hat{\boldsymbol{W}}$ 使得 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 最有效。不同矩条件的强弱程度一般不同，一个强的矩条件意味着其对应的方差较小（矩阵 $\boldsymbol{S} = \mathrm{E}(\boldsymbol{g}_i \boldsymbol{g}_i')$ 的主对角线元素），是一个比较紧的约束，故会通过 $\hat{\boldsymbol{W}}$ 得到较大的权重。因此，在某种意义上，GMM 的思想与 GLS 有相通之处（参见习题）。

根据方程（10.40），$\boldsymbol{g}_n(\hat{\boldsymbol{\beta}})$ 是 $\hat{\boldsymbol{\beta}}$ 的一次函数，故 $J(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{W}})$ 是 $\hat{\boldsymbol{\beta}}$ 的二次（型）函数，通过向量微分可以得到其最小化问题的解（推导方法类似于 OLS 估计量，参见附录）：

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\hat{\boldsymbol{W}}) = (\boldsymbol{S}_{ZX}' \hat{\boldsymbol{W}} \boldsymbol{S}_{ZX})^{-1} \boldsymbol{S}_{ZX}' \hat{\boldsymbol{W}} \boldsymbol{S}_{Zy} \tag{10.43}$$

其中，$\boldsymbol{S}_{ZX} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{x}_i'$，$\boldsymbol{S}_{Zy} \equiv \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i y_i$。秩条件 $\mathrm{rank}[\mathrm{E}(\boldsymbol{z}_i \boldsymbol{x}_i')] = K$ 及 $\hat{\boldsymbol{W}}$ 为正定矩阵保证了在大样本下，$(\boldsymbol{S}_{ZX}' \hat{\boldsymbol{W}} \boldsymbol{S}_{ZX})^{-1}$ 存在。

在恰好识别的情况下，$\boldsymbol{S}_{ZX}$ 为方阵，则 GMM 还原为普通的工具变量法，因为

$$\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\hat{\boldsymbol{W}}) = \boldsymbol{S}_{ZX}^{-1} \underbrace{\hat{\boldsymbol{W}}^{-1} \boldsymbol{S}_{ZX}'^{-1} \boldsymbol{S}_{ZX}' \hat{\boldsymbol{W}}}_{=\boldsymbol{I}} \boldsymbol{S}_{Zy} = \boldsymbol{S}_{ZX}^{-1} \boldsymbol{S}_{Zy} = \hat{\boldsymbol{\beta}}_{\mathrm{IV}} \tag{10.44}$$

所以,GMM 确实是矩估计的推广。由此可知,只有在过度识别的情况下,才有必要使用 GMM。

10.7 GMM 的大样本性质

假如已经掌握了第 5 章"OLS 大样本性质"的证明思想,对于 GMM 的大样本性质则几乎可如法炮制。

定理(GMM 估计量的大样本性质)

(1) ($\hat{\boldsymbol{\beta}}_{\text{GMM}}$为一致估计) 在假定 10.1 ~ 10.4 之下,$\underset{n\to\infty}{\text{plim}}\,\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})=\boldsymbol{\beta}$。

(2) ($\hat{\boldsymbol{\beta}}_{\text{GMM}}$为渐近正态) 如果假定 10.3 (即 $\text{E}(\boldsymbol{g}_i)=\boldsymbol{0}$)强化为假定 10.5(即$\{\boldsymbol{g}_i\}$为鞅差分序列),则$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}}-\boldsymbol{\beta})\xrightarrow{d}N(\boldsymbol{0},\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}}))$,其中渐近协方差矩阵

$$\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})=(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1}\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{\Sigma}_{ZX}(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1},$$
$$\boldsymbol{S}=\text{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\text{E}(\varepsilon_i^2\boldsymbol{z}_i\boldsymbol{z}_i'),\boldsymbol{\Sigma}_{ZX}\equiv\text{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$$

(3) ($\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$的一致估计量)如果 $\hat{\boldsymbol{S}}$ 是 $\boldsymbol{S}$ 的一致估计量,则在假定 10.2 下,$\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$的一致估计量为

$$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})}=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{S}}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}$$

证明:(1) 抽样误差可以写为

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta}&=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{z}_i y_i\right)-\boldsymbol{\beta}\\&=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{z}_i(\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i)\right)-\boldsymbol{\beta}\\&=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\left(\boldsymbol{S}_{ZX}\boldsymbol{\beta}+\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{z}_i\varepsilon_i\right)-\boldsymbol{\beta}\\&=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\bar{\boldsymbol{g}}\end{aligned}\tag{10.45}$$

其中,$\bar{\boldsymbol{g}}\equiv\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{g}_i,\boldsymbol{g}_i\equiv\boldsymbol{z}_i\varepsilon_i$。由于$(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\xrightarrow{p}(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1},\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\xrightarrow{p}\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}$,而$\bar{\boldsymbol{g}}\xrightarrow{p}\text{E}(\boldsymbol{g}_i)=\text{E}(\boldsymbol{z}_i\varepsilon_i)=\boldsymbol{0}$。因此,$\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta}\xrightarrow{p}\boldsymbol{0}$。从这个证明可以看出,保证 GMM 一致性的最重要条件仍是 $\text{E}(\boldsymbol{z}_i\varepsilon_i)=\boldsymbol{0}$,即工具变量与扰动项正交。

(2) 由于抽样误差 $\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta}=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\bar{\boldsymbol{g}}$,故$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta})=(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}(\sqrt{n}\bar{\boldsymbol{g}})$。根据假定 10.5 及鞅差分序列的中心极限定理,$\sqrt{n}\bar{\boldsymbol{g}}\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{S})$,其中 $\boldsymbol{S}\equiv\text{E}(\boldsymbol{g}_i\boldsymbol{g}_i')=\text{E}(\varepsilon_i^2\boldsymbol{z}_i\boldsymbol{z}_i')$。由于$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta})$是$\sqrt{n}\bar{\boldsymbol{g}}$的线性组合,故$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\boldsymbol{W}})-\boldsymbol{\beta})\xrightarrow{d}N(\boldsymbol{0},\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}}))$。由于$(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\xrightarrow{p}(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1},\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\xrightarrow{p}\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}$,故 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})=(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1}\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{\Sigma}_{ZX}(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1}$,其中$(\boldsymbol{\Sigma}_{ZX}'\boldsymbol{W}\boldsymbol{\Sigma}_{ZX})^{-1}$为对称矩阵。

(3) 由于 $\hat{\boldsymbol{S}}\xrightarrow{p}\boldsymbol{S}$,而且 $\boldsymbol{S}_{ZX}\xrightarrow{p}\boldsymbol{\Sigma}_{ZX},\hat{\boldsymbol{W}}\xrightarrow{p}\boldsymbol{W}$,故估计量$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})}=\underbrace{(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}}_{\text{面包}}\underbrace{\hat{\boldsymbol{S}}}_{\text{菜}}\underbrace{\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}(\boldsymbol{S}_{ZX}'\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}}_{\text{面包}}$是 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$的一致估计量。从形式上看,这也是一个夹心估计量(只是

“面包”成分更加丰富)。

命题 在假定 10.1、假定 10.2 与假定 10.6 下(四阶矩存在),对于 $\boldsymbol{\beta}$ 的任何一致估计量 $\hat{\boldsymbol{\beta}}$,定义残差 $e_i \equiv y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$,则 $s^2 \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2$ 是 $\sigma^2 \equiv \mathrm{E}(\varepsilon_i^2)$ 的一致估计,而且 $\hat{\boldsymbol{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{z}_i \boldsymbol{z}_i'$ 是 $\boldsymbol{S} \equiv \mathrm{E}(\varepsilon_i^2 \boldsymbol{z}_i \boldsymbol{z}_i')$ 的一致估计。

命题 使 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})$ 最小化的“最优权重矩阵”(optimal weighting matrix)为 $\hat{\boldsymbol{W}} = \hat{\boldsymbol{S}}^{-1}$。

这意味着,使用任何其他权重矩阵进行 GMM 估计,其估计量的渐近方差矩阵都将(在矩阵意义上)大于或等于使用 $\hat{\boldsymbol{S}}^{-1}$ 作为权重的渐近方差矩阵,即前者与后者之差为半正定矩阵。

定义 使用 $\hat{\boldsymbol{S}}^{-1}$ 为权重矩阵的 GMM 估计量被称为“效率 GMM”(efficient GMM)或“最优 GMM”(optimal GMM)。

为了使用最优权重矩阵 $\hat{\boldsymbol{S}}^{-1}$,首先必须估计 $\hat{\boldsymbol{S}}$。由于 2SLS 也是一致的,故用 2SLS 的残差来计算 $\hat{\boldsymbol{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{z}_i \boldsymbol{z}_i'$ 也是一致的。因此,可以进行以下“两步最优 GMM 估计”。

第一步:使用 2SLS,得到残差,计算 $\hat{\boldsymbol{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{z}_i \boldsymbol{z}_i'$。

第二步:最小化 $J(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{S}}^{-1})$,得到 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\hat{\boldsymbol{S}}^{-1})$。

在实际操作中,常使用“迭代法”(iterative GMM)直至估计值收敛,即用第二步所获的残差再来计算 $\hat{\boldsymbol{S}}$,然后再求 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\hat{\boldsymbol{S}}^{-1})$,以此类推。

在条件同方差的假定下,最优 GMM 的表达式可以大大简化。

命题 在条件同方差的情况下,最优 GMM 就是 2SLS。

证明:假设 $\mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{z}_i) = \sigma^2 > 0$(条件同方差①),则根据迭代期望定律

$$\boldsymbol{S} \equiv \mathrm{E}(\boldsymbol{z}_i \boldsymbol{z}_i' \varepsilon_i^2) = \mathrm{E}_{\boldsymbol{z}_i}\mathrm{E}(\boldsymbol{z}_i \boldsymbol{z}_i' \varepsilon_i^2 \mid \boldsymbol{z}_i) = \mathrm{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i \boldsymbol{z}_i' \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{z}_i)] = \sigma^2 \mathrm{E}(\boldsymbol{z}_i \boldsymbol{z}_i') \tag{10.46}$$

因此,$\tilde{\boldsymbol{S}} \equiv s^2 \boldsymbol{S}_{ZZ}$ 是 $\boldsymbol{S}$ 的一致估计量,其中 $\boldsymbol{S}_{ZZ} \equiv \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$。使用 $\tilde{\boldsymbol{S}}^{-1} = (s^2 \boldsymbol{S}_{ZZ})^{-1}$ 为最优权重矩阵,则最优 GMM 估计量为

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\tilde{\boldsymbol{S}}^{-1}) &= (\boldsymbol{S}_{ZX}'(s^2 \boldsymbol{S}_{ZZ})^{-1}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'(s^2 \boldsymbol{S}_{ZZ})^{-1}\boldsymbol{S}_{Zy} \\ &= (\boldsymbol{S}_{ZX}'\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}_{ZX}'\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{Zy} \end{aligned} \tag{10.47}$$

注意到在上式中,s^2 被消去了,即最优权重矩阵的常数倍并不影响 $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ 的取值。

由于 $\boldsymbol{S}_{ZX} \equiv \frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}, \boldsymbol{S}_{ZZ} \equiv \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}, \boldsymbol{S}_{Zy} \equiv \frac{1}{n}\boldsymbol{Z}'\boldsymbol{y}$,故

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{GMM}}(\tilde{\boldsymbol{S}}^{-1}) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z} \cdot n(\boldsymbol{Z}'\boldsymbol{Z})^{-1} \cdot \frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}'\boldsymbol{Z} \cdot n(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\frac{1}{n}\boldsymbol{Z}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} \equiv \hat{\boldsymbol{\beta}}_{\mathrm{2SLS}} \end{aligned} \tag{10.48}$$

由此可知,在条件同方差的情况下,两步最优 GMM 可以省略为一步。这是因为,两步最优 GMM 中第一步的目的只是得到 $\hat{\boldsymbol{S}}^{-1}$,而在条件同方差假定下,可以直接令 $\hat{\boldsymbol{S}}^{-1} = \boldsymbol{S}_{ZZ}^{-1}$。因此,2SLS 有时也被称为“一步 GMM”。

---

① 此处的“条件同方差”指的是,在给定工具变量 $\boldsymbol{z}_i$(而非解释变量 $\boldsymbol{x}_i$)情况下的条件方差相同。

GMM 的过度识别检验(Overidentification Test or Hansen's *J* Test)

在恰好识别的情况下,GMM 最小化的目标函数 $J(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{S}}^{-1}) = 0$。在过度识别的情况下,如果所有的过度识别约束都成立,则目标函数 $J(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{S}}^{-1})$ 应该离 0 不远。如果 $J(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{S}}^{-1})$ 大于 0 很多,则可倾向于认为某些过度识别约束不成立。在原假设"H_0:所有矩条件均成立"的情况下,目标函数本身就是检验统计量

$$J(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{S}}^{-1}) \xrightarrow{d} \chi^2(L-K) \tag{10.49}$$

其中,$(L-K)$ 为过度识别约束的个数,因为在估计 $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ 的过程中失去了 K 个自由度。可以证明,在条件同方差的情况下,J 统计量与 Sargan 统计量相等。

有关 GMM 的 Stata 命令为

```
ivregress gmm y x1 (x2 = z1 z2)          (两步 GMM)
ivregress gmm y x1 (x2 = z1 z2),igmm     (迭代 GMM)
estat overid                             (过度识别检验)
```

检验部分工具变量的正交性(Testing Subsets of Orthogonality Condition)

如果过度识别检验拒绝"所有工具变量均为外生"的原假设,则可怀疑部分工具变量不满足正交性(外生性)。不失一般性,假设在 L 个工具变量 $\boldsymbol{z}_i$ 中,已知前 L_1 个工具变量 $\boldsymbol{z}_{i1}$ 满足正交性,而怀疑后 $(L-L_1)$ 个工具变量 $\boldsymbol{z}_{i2}$ 不满足正交性,即要检验原假设"H_0: $\mathrm{E}(\boldsymbol{z}_{i2}\varepsilon_i) = 0$"。进行此检验的前提条件是 $L_1 \geqslant K$,保证即使仅用前 L_1 个工具变量 $\boldsymbol{z}_{i1}$ 进行估计,该模型也至少为恰好识别。如果 $L_1 \geqslant K$,则可分别用所有 L 个工具变量 $\boldsymbol{z}_i$ 或前 L_1 个工具变量 $\boldsymbol{z}_{i1}$ 进行 GMM 估计(假设仅使前 L_1 个工具变量 $\boldsymbol{z}_{i1}$ 的秩条件也成立),并分别记相应的 J 统计量为 J 与 J_1。显然,如果将后 $(L-L_1)$ 个工具变量 $\boldsymbol{z}_{i2}$ 也用于 GMM 估计,使得 J 统计量大大增加,则倾向于拒绝原假设"H_0: $\mathrm{E}(\boldsymbol{z}_{i2}\varepsilon_i) = 0$"。可以证明[①],

$$C \equiv J - J_1 \xrightarrow{d} \chi^2(L-L_1) \tag{10.50}$$

其中,C 统计量也称为"GMM 距离"(GMM distance)统计量或"Sargan 差"(difference-in-Sargan)统计量,因为它其实是两个 GMM 估计的 Sargan - Hansen 统计量之差。C 统计量服从渐近 χ^2 分布,自由度为 $(L-L_1)$,即怀疑正交性不成立的工具变量个数。C 统计量可通过命令"`ivreg2`"的选择项"`orthog(varlist)`"来获得。

在存在自相关的情况下使用 GMM

我们知道,在存在异方差的情况下,GMM 依然是稳健与最优的。在时间序列数据中,即使存在自相关,也仍然可以使用 GMM,只要采用异方差自相关稳健的标准误来进行统计推断就行。此时,GMM 估计量依然满足一致性、渐近正态性与渐近有效性,只是最优权重矩阵 $\hat{\boldsymbol{S}}^{-1}$ 的表达式不同。

在 Stata 中,使用异方差自相关稳健的标准误的 GMM 命令为

```
ivregress gmm y x1 (x2 = z1 z2),vce(hac nwest[#])
```

其中,选择项"`vce`"指的是"Variance - Covariance Matrix Estimated",而"`nwest`"指的是"Newey - West 标准误"(即异方差自相关稳健的标准误),`[#]` 指的是滞后阶数,其默认值为 $T-2$(T 为时

---

① 一个技术细节是,为了保证 C 统计量为非负,使用 $\boldsymbol{z}_{i2}$ 进行 GMM 估计时,最优权重矩阵 $\hat{\boldsymbol{S}}_{11}^{-1}$ 应来自于使用 $\boldsymbol{z}_i$ 进行 GMM 估计时最优权重矩阵 $\hat{\boldsymbol{S}}^{-1}$ 的相应部分。

间序列数据的样本容量)。

10.8 如何获得工具变量

使用工具变量法的前提是存在有效的工具变量。因此,如何寻找工具变量十分重要。然而,工具变量的两个要求(相关性与外生性)常常自相矛盾,即与内生解释变量相关的变量往往与被解释变量的扰动项也相关。故在实践上,寻找合适的工具变量通常比较困难,需要一定的创造性与想象力。寻找工具变量的步骤大致可以分为两步:

(i) 列出与内生解释变量(x)相关的尽可能多的变量清单(这一步较容易);

(ii) 从这一清单中剔除与扰动项相关的变量(这一步较难)。

(ii) 的操作有一定难度,因为扰动项不可观测。既然扰动项不可观测,那么又如何能判断某候选变量(z)是否与不可观测的扰动项(ε)相关呢?由于扰动项是被解释变量(y)的扰动项,故可以从该候选变量与被解释变量的相关性着手。显然 z 与 y 相关,因为 z 与内生解释变量 x 相关。重要的是,z 对 y 的影响仅仅通过 x 来起作用(类似于代理变量的"多余性"要求),因为如果 z 与 ε 相关,则 z 对 y 的影响必然还有除 x 以外的渠道,参见图 10.3。至于是否"z 对 y 的影响仅仅通过 x 来起作用",有时可以通过定性的讨论来确定。这就是上文提到的"排他性约束"(exclusion restriction)。

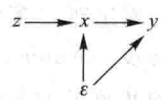


图 10.3 工具变量示意图
(箭头表示相关,无箭头表示不相关)

下面举几个工具变量法的经典实例,其中有些不乏争议。

例 滞后变量。对于时间序列或面板数据,常常使用内生解释变量的滞后变量作为工具变量。显然,一方面,内生解释变量与其滞后变量相关。另一方面,由于滞后变量已经发生,故为"前定"(从当期的角度看,其取值已经固定),可能与当期的扰动项不相关。比如,在对"动态面板"的估计中,大量地使用滞后变量作为工具变量(参见第 16 章)。比如,Groves et al (1994)考察国企改革(员工奖金激励制度)对企业生产率的作用。一般来说,奖金占员工中报酬比重越高,则越能促进生产率的提高。但生产率越高的企业越有能力给员工发奖金,故存在双向因果关系。为此,Groves et al (1994)使用奖金比重的滞后值作为当期奖金比重的工具变量。二者的相关性是显然的。但当期的生产率不可能影响过去的奖金比重,故奖金比重的滞后值具有外生性。

例 警察人数与犯罪率。一般认为,警察人数越多,执法力度越大,则犯罪率应该越低。然而,如果直接把犯罪率对警察人数进行回归,以此度量警察人数对犯罪率的作用,就会出现内生变量偏差。这是因为,警察人数其实是内生变量,比如,某城市的犯罪率很高,则市政府通常会增加警察人数。为此,必须找到与警察人数相关,但对犯罪率没有其他影响渠道的工具变量。Levitt (1997)创造性地使用"市长选举的政治周期"作为犯罪率(包括 7 种类型的犯罪)的工具变量。通常,在任市长在竞选连任时,为了拉选票,会增加警察人数,故满足相关性。另一方面,选举周期一般以机械的方式确定,除了对警察人数有影响外,不会单独地对犯罪率起作用,故满足外生性。然而,McCrary (2002)发现 Levitt (1997)的电脑程序有误(在将 7 种类型的犯罪放在一起作 2SLS 时,由于不同类型犯罪率的方差不同而使用了加权回归,但将权重误设为扰动项的标准差,而非扰动项标准差的倒数,参见第 7 章),而更正此错误后,主要结论不再成立。Levitt (2002)承认了此错误,并提出以消防队员人数或其他市政人员人数作为警察人数的工具变量。

例 国际贸易与经济增长。国际贸易会带来财富是一个古老的学说。但要实证地研究国际

贸易对经济增长的促进作用却面临着内生解释变量的问题,因为经济增长也可以反作用于国际贸易,即随着经济增长,国际贸易也跟着增加了。Frankel and Romer (1999)使用地理因素作为工具变量。首先,国际贸易受地理因素的影响(比如,距离较近的国家之间的贸易量较大),故满足相关性。其次,地理因素对经济增长的影响可能仅仅通过国际贸易这个渠道来实现。

例 制度对经济增长的影响。好的制度能促进经济增长,但制度变迁常常也依赖于经济增长。因此,制度本身是内生变量。Acemoglu et al (2001)使用"殖民者死亡率"(settler mortality)作为制度的工具变量。当近代欧洲的殖民者在全世界进行殖民时,由于各地的气候及疾病环境(disease environment)不同,欧洲殖民者的死亡率差异较大。在死亡率高的地方(比如,非洲),殖民者难以定居,故在当地建立掠夺性制度(extractive institutions)。而在死亡率低的地方(比如,北美),则建立有利于经济增长的制度(比如,较好的产权保护)。这种初始制度上的差异一直延续到今天。因此,一方面,殖民者死亡率与今天的制度相关,满足相关性。另一方面,殖民者死亡率除了对制度有影响外,不再对当前的经济增长有任何直接影响,故满足外生性。然而,历史上的殖民者死亡率并不易度量。Albouy (2012)质疑 Acemoglu et al (2001)构建殖民者死亡率数据的方法与实证结果,Acemoglu (2012)作了回应,引起较大争议。

例 看电视过多引发小儿自闭症? 在美国,电视的普及与小儿自闭症(autism)发生率的攀升几乎同步。Waldman et al (2006,2008)研究过多观看电视是否引发小儿自闭症。然而,有自闭倾向的儿童可能更经常看电视,而不喜欢户外活动或与人交往;故存在双向因果关系。为此,Waldman et al (2006,2008)使用降雨量作为电视观看时间的工具变量。二者存在相关性,即降雨越多的地区,人们待在室内的时间越长,故看电视时间也越长;而降雨量很可能是外生的(只通过看电视时间而影响被解释变量)。研究结果支持过多观看电视为小儿自闭症的诱因。此项研究引起很多媒体关注与医学界的争议,因为 Waldman 等人均为经济学家而非医生,而他们使用的工具变量法与主流的医学实验法相差甚远(如果进行随机实验,成本将很高,包括道德成本)。

例 学区竞争与教育质量。一个城市的学区是否越多,学区间竞争越激烈,则越有利于提高教育质量(给定教育支出,更好的学生成绩)? 如果直接将二者进行回归,将面临内生性问题,因为在学区形成的过程中,效率高的学区会变得更大,或许兼并相邻的学区。为此,Hoxby (2000)使用一个城市河流的数目作为该城市学区个数的工具变量。历史上,如果一个城市的河流越多,一方面妨碍交通的自然障碍越多,导致城市设立更多的学区;故二者满足相关性。另一方面,河流数目很可能不会直接影响教育质量,故满足外生性。然而,计算河流数目并非易事(比如,在沼泽地区)。Rothstein (2007)质疑 Hoxby (2000)计算河流数目的方法以及实证结果,Hoxby (2007)作了回应,引起较大争议。

10.9 MLE 也是 GMM

第 6 章曾提及,即使似然函数不正确,"准最大似然估计"(QMLE)也可能是一致估计量。这是因为,MLE 也可以被视为 GMM。

第 6 章附录证明了,在似然函数正确的情况下,得分函数的期望值 $\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})]=\mathbf{0}$,其中 $\boldsymbol{s}(\boldsymbol{\theta}_0;\boldsymbol{y})\equiv\dfrac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial \boldsymbol{\theta}}$ 为对数似然函数的一阶偏导。根据同样的逻辑,第 i 个数据对得分函数的

贡献 $\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)\equiv\dfrac{\partial \ln L(\boldsymbol{\theta}_0;\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}}$ 也满足这个性质，即

$$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathbf{0} \tag{10.51}$$

将上式视为总体矩条件，则其对应的样本矩条件为

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{s}_i(\hat{\boldsymbol{\theta}};\boldsymbol{y}_i)=\mathbf{0} \tag{10.52}$$

去掉上式中的 $\dfrac{1}{n}$，就是 MLE 的一阶条件。因此，可以将 MLE 看作是恰好识别的 GMM 估计量（上式中方程个数等于未知数个数）。根据 GMM 的性质，只要矩条件“$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathbf{0}$”正确，则 GMM 估计量（也就是 MLE）就是一致的。显然，矩条件“$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathbf{0}$”比“似然函数正确”更弱。因此，即使在扰动项为非正态分布的情况下，只要“$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0;\boldsymbol{y}_i)]=\mathbf{0}$”成立，则 QMLE 依然是一致的。

比如，对于线性回归模型，如果其扰动项服从正态分布，则第 i 个观测值的似然函数为

$$L(\boldsymbol{\beta},\sigma^2;y_i)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2/2\sigma^2\} \tag{10.53}$$

其对数似然函数及偏导数为

$$\ln L(\boldsymbol{\beta},\sigma^2;y_i)=-\frac{1}{2}\ln 2\pi-\frac{1}{2}\ln\sigma^2-\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2} \tag{10.54}$$

$$\boldsymbol{s}_i(\boldsymbol{\beta})=\frac{\partial \ln L(\boldsymbol{\beta},\sigma^2;y_i)}{\partial\boldsymbol{\beta}}=\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i}{\sigma^2} \tag{10.55}$$

$$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\beta})]=\frac{\mathrm{E}[(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i]}{\sigma^2}=\mathbf{0} \tag{10.56}$$

而这正是 OLS 所对应的矩条件，它的成立与否并不依赖于似然函数是否正确。总之，许多计量方法都可以归结为 GMM。Hayashi（2000）[①]正是以 GMM 为核心方法来介绍计量经济学的。

10.10 工具变量法的 Stata 命令及实例[②]

Mincer（1958）最早研究了工资与受教育年限的正相关关系[③]，但遗漏了“能力”这个变量，导致遗漏变量偏差。针对美国面板调查数据中的年轻男子组群（Young Men's Cohort of the National Longitudinal Survey，简记 NLS - Y），Griliches（1976）采用工具变量法对遗漏变量问题进行了校正。Blackburn and Neumark（1992）更新了 Griliches（1976）的数据，即这个例子中将要使用的数据集“grilic. dta”。

该数据集中包括以下变量：lw（工资对数），s（受教育年限），age（年龄），expr（工龄），tenure（在现单位的工作年数），iq（智商），med（母亲的受教育年限），kww（在“knowledge of the World of Work”测试中的成绩），mrt（婚姻虚拟变量，已婚 = 1），rns（美国南方虚拟变量，住在南方 = 1），smsa（大城市虚拟变量，住在大城市 = 1），year（有数据的最早年份，1966—1973 年中的某一年）。

① 中译本为林文夫（2005）。

② 此例来自 Hayashi（2000）。

③ 把工资的对数对受教育年限及其他控制变量进行的这类回归，称为“Mincerian regression”。

这是一个两期面板数据,初始期为当以上变量有数据的最早年份,结束期为 1980 年。不带“80”字样的变量名为初始期,带“80”字样的变量名为 1980 年数据。比如,iq 指的是初始期的智商,而 lw80 指的是 1980 年的工资对数。

(1) 先看一下数据的统计特征。

```
. use grilic.dta,clear
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rns	758	.2691293	.4438001	0	1
rns80	758	.292876	.4553825	0	1
mrt	758	.5145119	.5001194	0	1
mrt80	758	.8984169	.3022988	0	1
smsa	758	.7044855	.456575	0	1
smsa80	758	.7124011	.452942	0	1
med	758	10.91029	2.74112	0	18
iq	758	103.8562	13.61867	54	145
kww	758	36.57388	7.302247	12	56
year	758	69.03166	2.631794	66	73
age	758	21.83509	2.981756	16	30
age80	758	33.01187	3.085504	28	38
s	758	13.40501	2.231828	9	18
s80	758	13.70712	2.214693	9	18
expr	758	1.735429	2.105542	0	11.444
expr80	758	11.39426	4.210745	.692	22.045
tenure	758	1.831135	1.67363	0	10
tenure80	758	7.362797	5.05024	0	22
lw	758	5.686739	.4289494	4.605	7.051
lw80	758	6.826555	.4099268	4.749	8.032

(2) 考察智商与受教育年限的相关关系。

```
. pwcorr iq s,sig
```

	iq	s
iq	1.0000	
s	0.5131	1.0000
	0.0000	

上表显示,智商(在一定程度上可视为“能力”的代理变量)与受教育年限具有较强的正相关关系(相关系数为 0.51,且在 1% 水平上显著)。

(3) 作为一个参照系,先进行 OLS 回归,并使用稳健标准误。

```
. reg lw s expr tenure rns smsa,r
```

其中 expr,tenure,rns,smsa 均为控制变量[①],而我们主要感兴趣的变量为 s(受教育年限)。

① 也可以在控制变量中包括“年度虚拟变量”(year dummies)。使用命令“tab year,gen(year)”即可生成虚拟变量 year1-year8,分别代表 1966—1973 年。

```
Linear regression                               Number of obs =       758
                                                F(  5,   752) =     84.05
                                                Prob > F      =    0.0000
                                                R-squared     =    0.3521
                                                Root MSE      =    .34641
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |    .102643   .0062099    16.53   0.000     .0904523    .1148338
        expr |   .0381189   .0066144     5.76   0.000      .025134    .0511038
      tenure |   .0356146   .0079988     4.45   0.000     .0199118    .0513173
         rns |  -.0840797    .029533    -2.85   0.005    -.1420566   -.0261029
        smsa |   .1396666    .028056     4.98   0.000     .0845893     .194744
       _cons |   4.103675   .0876665    46.81   0.000     3.931575    4.275775
------------------------------------------------------------------------------
```

回归结果显示,教育投资的年回报率为 10.26%,而且在 1% 的水平上显著性不为 0。这意味着,多受一年教育,则未来的工资将高出 10.26%,这个教育投资回报率似乎太高了。可能的原因是,由于遗漏变量"能力"与受教育年限正相关,故"能力"对工资的贡献也被纳入教育的贡献,因此高估了教育的回报率。

(4) 引入智商(iq)作为"能力"的代理变量[①],再进行 OLS 回归。

```
. reg lw s iq expr tenure rns smsa,r
```

```
Linear regression                               Number of obs =       758
                                                F(  6,   751) =     71.89
                                                Prob > F      =    0.0000
                                                R-squared     =    0.3600
                                                Root MSE      =    .34454
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .0927874   .0069763    13.30   0.000     .0790921    .1064826
          iq |   .0032792   .0011321     2.90   0.004     .0010567    .0055016
        expr |   .0393443   .0066603     5.91   0.000     .0262692    .0524193
      tenure |    .034209   .0078957     4.33   0.000     .0187088    .0497092
         rns |  -.0745325   .0299772    -2.49   0.013    -.1333815   -.0156834
        smsa |   .1367369   .0277712     4.92   0.000     .0822186    .1912553
       _cons |   3.895172   .1159286    33.60   0.000     3.667589    4.122754
------------------------------------------------------------------------------
```

加入"能力"的代理变量 iq 后,教育投资的回报率下降为 9.28%,变得更为合理,但仍然显得过高。

(5) 由于用 iq 来度量能力存在"测量误差",故 iq 是内生变量,考虑使用变量(med,kww,mrt,age)作为 iq 的工具变量,进行 2SLS 回归,并使用稳健标准误。

```
. ivregress 2sls lw s expr tenure rns smsa (iq = med kww mrt age),r
```

① 文献中还曾使用高中考试成绩(Brewer et al,1999;李宏彬等,2012)、美国参军资格考试(Armed Forces Qualification Test,简记 AFQT)(Griliches and Mason,1972; Leigh and Gill,1997)作为能力的代理变量。

```
Instrumental variables (2SLS) regression          Number of obs =     758
                                                  Wald chi2(6)  =  355.73
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.2002
                                                  Root MSE      =  .38336
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0115468   .0056376    -2.05   0.041    -.0225962   -.0004974
           s |   .1373477   .0174989     7.85   0.000     .1030506    .1716449
        expr |   .0338041   .0074844     4.52   0.000      .019135    .0484732
      tenure |    .040564   .0095848     4.23   0.000     .0217781     .05935
         rns |  -.1176984   .0359582    -3.27   0.001    -.1881751   -.0472216
        smsa |    .149983   .0322276     4.65   0.000     .0868182    .2131479
       _cons |   4.837875   .3799432    12.73   0.000       4.0932     5.58255
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa med kww mrt age
```

在此 2SLS 回归中,教育回报率反而上升到 13.73%,而智商(iq)对工资的贡献居然为负,似乎并不可信。使用工具变量法的前提是工具变量的有效性。为此,进行过度识别检验,考察是否所有工具变量均外生,即与扰动项不相关。

```
. estat overid
```

```
Test of overidentifying restrictions:

Score chi2(3)          =  51.5449  (p = 0.0000)
```

结果强烈拒绝"所有工具变量均外生"的原假设(p 值为 0.000 0),即认为某些(或某个)工具变量不合格(invalid),与扰动项相关。我们怀疑(mrt,age)不满足外生性,故使用 C 统计量检验这两个工具变量的外生性。由于"ivregress"不提供 C 统计量,故下载非官方命令"ivreg2"。

```
. ssc install ivreg2
. ivreg2 lw s expr tenure rns smsa (iq = med kww mrt age), r orthog(mrt age)
```

命令"ivreg2"的默认估计量为 2SLS(如果加上选择项"gmm2s robust",则为两步最优 GMM 估计量),选择项"r"表示异方差稳健的标准误,选择项"orthog(mrt age)"表示检验(mrt,age)是否满足外生性(正交性)。有关"ivreg2"的更多说明,参见"help ivreg2"以及 Baum et al (2007)。

```
IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                  Number of obs =       758
                                                  F(  6,   751) =     58.74
                                                  Prob > F      =    0.0000
Total (centered) SS     =  139.2861498            Centered R2   =    0.2002
Total (uncentered) SS   =  24652.24662            Uncentered R2 =    0.9955
Residual SS             =   111.39959             Root MSE      =     .3834

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0115468   .0056376    -2.05   0.041    -.0225962   -.0004974
           s |   .1373477   .0174989     7.85   0.000     .1030506    .1716449
        expr |   .0338041   .0074844     4.52   0.000      .019135    .0484732
      tenure |    .040564   .0095848     4.23   0.000     .0217781     .05935
         rns |  -.1176984   .0359582    -3.27   0.001    -.1881751   -.0472216
        smsa |    .149983   .0322276     4.65   0.000     .0868182    .2131479
       _cons |   4.837875   .3799432    12.73   0.000       4.0932     5.58255
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):              33.294
                                                   Chi-sq(4) P-val =     0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                10.538
                         (Kleibergen-Paap rk Wald F statistic):           9.585
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias     16.85
                                         10% maximal IV relative bias     10.27
                                         20% maximal IV relative bias      6.71
                                         30% maximal IV relative bias      5.34
                                         10% maximal IV size              24.58
                                         15% maximal IV size              13.96
                                         20% maximal IV size              10.26
                                         25% maximal IV size               8.31
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         51.545
                                                   Chi-sq(3) P-val =     0.0000
-orthog- option:
Hansen J statistic (eqn. excluding suspect orthog. conditions):           0.116
                                                   Chi-sq(1) P-val =     0.7333
C statistic (exogeneity/orthogonality of suspect instruments):           51.429
                                                   Chi-sq(2) P-val =     0.0000
Instruments tested:   mrt age
------------------------------------------------------------------------------
Instrumented:         iq
Included instruments: s expr tenure rns smsa
Excluded instruments: med kww mrt age
------------------------------------------------------------------------------
```

从上表可以看出，使用“ivreg2”得到的回归系数和稳健标准误差与“ivregress 2sls”完全相同。不可识别检验显示，Kleibergen-Paap rk LM 统计量的 p 值为 0.000 0，强烈拒绝不可识别的原假设。服从 $\chi^2(2)$ 分布的 C 统计量为 51.43，对应的 p 值为 0.000 0，故强烈拒绝“(mrt, age)满足外生性”的原假设。

(6) 考虑仅使用变量(med,kww)作为 iq 的工具变量，再次进行 2SLS 回归，同时显示第一阶段的回归结果。

```
. ivregress 2sls lw s expr tenure rns smsa (iq = med kww),r first
```

```
First-stage regressions
                                                Number of obs   =       758
                                                F(   7,    750) =     47.74
                                                Prob > F        =    0.0000
                                                R-squared       =    0.3066
                                                Adj R-squared   =    0.3001
                                                Root MSE        =   11.3931
------------------------------------------------------------------------------
             |               Robust
          iq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |    2.467021   .2327755    10.60   0.000     2.010052    2.92399
        expr |   -.4501353   .2391647    -1.88   0.060    -.9196471   .0193766
      tenure |    .2059531    .269562     0.76   0.445    -.3232327   .7351388
         rns |   -2.689831   .8921335    -3.02   0.003    -4.441207   -.938455
        smsa |    .2627416   .9465309     0.28   0.781    -1.595424   2.120907
         med |    .3470133   .1681356     2.06   0.039     .0169409   .6770857
         kww |    .3081811   .0646794     4.76   0.000     .1812068   .4351553
       _cons |    56.67122   3.076955    18.42   0.000     50.63075   62.71169
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression        Number of obs =       758
                                                Wald chi2(6)  =    370.04
                                                Prob > chi2   =    0.0000
                                                R-squared     =    0.2775
                                                Root MSE      =    .36436
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |    .0139284   .0060393     2.31   0.021     .0020916    .0257653
           s |    .0607803   .0189505     3.21   0.001      .023638    .0979227
        expr |    .0433237   .0074118     5.85   0.000     .0287968    .0578505
      tenure |    .0296442    .008317     3.56   0.000     .0133432    .0459452
         rns |   -.0435271   .0344779    -1.26   0.207    -.1111026    .0240483
        smsa |    .1272224   .0297414     4.28   0.000     .0689303    .1855146
       _cons |    3.218043   .3983683     8.08   0.000     2.437256    3.998831
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa med kww
```

上表显示,教育投资回报率降为 6.08%,比较合理;而且 iq 的贡献也重新变为正。再次进行过度识别检验:

```
. estat overid
```

```
Test of overidentifying restrictions:

Score chi2(1)          =  .151451  (p = 0.6972)
```

由于 p 值为 0.70,故接受原假设,认为(med,kww)外生,与扰动项不相关。

(7) 进一步考察有效工具变量的第二个条件,即工具变量与内生变量的相关性。从第一阶段的回归结果可以看出,工具变量(med,kww)对内生变量 iq 均有较好的解释力,p 值都小于 0.05。更正式的检验可以通过如下命令进行:

```
. estat firststage,all forcenonrobust
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(2,750)	Prob > F
iq	0.3066	0.3001	0.0382	13.4028	0.0000

Shea's partial R-squared

Variable	Shea's Partial R-sq.	Shea's Adj. Partial R-sq.
iq	0.0382	0.0305

Minimum eigenvalue statistic = 14.9058

Critical Values # of endogenous regressors: 1
Ho: Instruments are weak # of excluded instruments: 2

	5%	10%	20%	30%
2SLS relative bias	(not available)			

	10%	15%	20%	25%
2SLS Size of nominal 5% Wald test	19.93	11.59	8.75	7.25
LIML Size of nominal 5% Wald test	8.68	5.33	4.42	3.92

从以上结果可以看出，虽然 Shea's partial R^2 不到 0.04，但 F 统计量为 13.40（超过 10），而且 F 统计量的 p 值为 0.000 0[①]。

我们知道，虽然 2SLS 是一致的，但是有偏的，故使用 2SLS 会带来“显著性水平扭曲”（size distortion），而且这种扭曲随着弱工具变量而增大。上表的最后部分显示，如果在结构方程中对内生解释变量的显著性进行“名义显著性水平”（nominal size）为 5% 的沃尔德检验，假如可以接受“真实显著性水平”（true size）不超过 15%，则可以拒绝“弱工具变量”的原假设，因为最小特征值统计量为 14.91[②]，大于对应的临界值 11.59。

总之，我们有理由认为不存在弱工具变量。但为了稳健起见，下面使用对弱工具变量更不敏感的有限信息最大似然法（LIML）：

```
. ivregress liml lw s expr tenure rns smsa (iq = med kww),r
```

Instrumental variables (LIML) regression

Number of obs = 758
Wald chi2(6) = 369.62
Prob > chi2 = 0.0000
R-squared = 0.2768
Root MSE = .36454

lw	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
iq	.0139764	.0060681	2.30	0.021	.0020831	.0258697
s	.0606362	.019034	3.19	0.001	.0233303	.0979421
expr	.0433416	.0074185	5.84	0.000	.0288016	.0578816
tenure	.0296237	.008323	3.56	0.000	.0133109	.0459364
rns	-.0433875	.034529	-1.26	0.209	-.1110631	.0242881
smsa	.1271796	.0297599	4.27	0.000	.0688512	.185508
_cons	3.214994	.4001492	8.03	0.000	2.430716	3.999272

Instrumented: iq
Instruments: s expr tenure rns smsa med kww

① 此检验的原假设是，工具变量（med，kww）在第一阶段回归中的系数都为 0。

② 在只有一个内生解释变量的情况下，F 统计量应该与“最小特征值统计量”相等。但由于此处使用了稳健标准误，故“稳健 F 统计量”（Robust F）与“最小特征值统计量”略有差别。

结果发现,LIML 的系数估计值与 2SLS 非常接近,这也从侧面印证了“不存在弱工具变量”。

(8) 下面使用非官方命令“ivreg2”进一步考察弱工具变量问题,

```
. ivreg2 lw s expr tenure rns smsa (iq = med kww),r redundant(kww)
```

其中,选择项“redundant(kww)”表示对工具变量 kww 进行冗余检验(为了演示的目的)。

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =       758
                                                      F(  6,   751) =     61.10
                                                      Prob > F      =    0.0000
Total (centered) SS     =  139.2861498                Centered R2   =    0.2775
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9959
Residual SS             =  100.6291971                Root MSE      =     .3644

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0139284   .0060393     2.31   0.021     .0020916    .0257653
           s |   .0607803   .0189505     3.21   0.001      .023638    .0979227
        expr |   .0433237   .0074118     5.85   0.000     .0287968    .0578505
      tenure |   .0296442    .008317     3.56   0.000     .0133432    .0459452
         rns |  -.0435271   .0344779    -1.26   0.207    -.1111026    .0240483
        smsa |   .1272224   .0297414     4.28   0.000     .0689303    .1855146
       _cons |   3.218043   .3983683     8.08   0.000     2.437256    3.998831
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):             24.223
                                                   Chi-sq(2) P-val =    0.0000
-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    22.222
                                                   Chi-sq(1) P-val =    0.0000
Instruments tested:  kww
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               14.906
                         (Kleibergen-Paap rk Wald F statistic):         13.403
Stock-Yogo weak ID test critical values: 10% maximal IV size            19.93
                                         15% maximal IV size            11.59
                                         20% maximal IV size             8.75
                                         25% maximal IV size             7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.151
                                                   Chi-sq(1) P-val =    0.6972
------------------------------------------------------------------------------
Instrumented:         iq
Included instruments: s expr tenure rns smsa
Excluded instruments: med kww
------------------------------------------------------------------------------
```

从上表可知,弱工具变量检验的两个统计量均显示,对于名义显著性水平为 5% 的检验,其真实显著性水平不会超过 15%(此结论与上文通过 minimum eigenvalue 统计量而进行的检验结果相一致)。冗余检验的结果表明,强烈拒绝“kww 为冗余工具变量”的原假设。

如果认为扰动项为 iid,则可以去掉选择项“robust”(通常不建议这样做,这里仅为演示的目的)。

```
. ivreg2 lw s expr tenure rns smsa (iq = med kww)
```

```
IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =       758
                                                      F(  6,   751) =     61.94
                                                      Prob > F      =    0.0000
Total (centered) SS     =  139.2861498                Centered R2   =    0.2775
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9959
Residual SS             =  100.6291971                Root MSE      =     .3644

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0139284   .0058572     2.38   0.017     .0024485    .0254084
           s |   .0607803   .0186481     3.26   0.001     .0242306    .0973301
        expr |   .0433237   .0070053     6.18   0.000     .0295935    .0570539
      tenure |   .0296442   .0085218     3.48   0.001     .0129418    .0463466
         rns |  -.0435271   .0347602    -1.25   0.210    -.1116558    .0246016
        smsa |   .1272224   .0299973     4.24   0.000     .0684288    .1860161
       _cons |   3.218043   .3830327     8.40   0.000     2.467313    3.968774
------------------------------------------------------------------------------

Underidentification test (Anderson canon. corr. LM statistic):          28.978
                                                   Chi-sq(2) P-val =    0.0000

Weak identification test (Cragg-Donald Wald F statistic):               14.906
Stock-Yogo weak ID test critical values: 10% maximal IV size            19.93
                                         15% maximal IV size            11.59
                                         20% maximal IV size             8.75
                                         25% maximal IV size             7.25
Source: Stock-Yogo (2005).  Reproduced by permission.

Sargan statistic (overidentification test of all instruments):          0.130
                                                   Chi-sq(1) P-val =    0.7185

Instrumented:         iq
Included instruments: s expr tenure rns smsa
Excluded instruments: med kww
```

上表弱工具变量检验的 Cragg-Donald Wald F 统计量显示，对于名义显著性水平为 5% 的检验，其真实显著性水平不会超过 15%。在 iid 扰动项的假设下，过度识别检验提供的是 Sargan 统计量，而非 Hansen J 统计量。

（9）使用工具变量法的前提是存在内生解释变量。为此须进行豪斯曼检验，其原假设为“所有解释变量均为外生”，即不存在内生变量。

```
. qui reg lw iq s expr tenure rns smsa
. estimates store ols
. qui ivregress 2sls lw s expr tenure rns smsa (iq = med kww)
. estimates store iv
. hausman iv ols,constant sigmamore
```

	Coefficients (b) iv	(B) ols	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
iq	.0139284	.0032792	.0106493	.0054318
s	.0607803	.0927874	-.032007	.0163254
expr	.0433237	.0393443	.0039794	.0020297
tenure	.0296442	.034209	-.0045648	.0023283
rns	-.0435271	-.0745325	.0310054	.0158145
smsa	.1272224	.1367369	-.0095145	.0048529
_cons	3.218043	3.895172	-.6771285	.3453751

```
            b = consistent under Ho and Ha; obtained from ivregress
 B = inconsistent under Ha, efficient under Ho; obtained from regress

Test:  Ho:  difference in coefficients not systematic

              chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =         3.84
            Prob>chi2 =       0.0499
            (V_b-V_B is not positive definite)
```

上表显示,可以在 5% 的显著性水平上拒绝“所有解释变量均为外生”的原假设,即认为 iq 为内生变量。由于传统的豪斯曼检验建立在同方差的前提下,故在上述回归中均没有使用稳健标准误(没有用选择项“r”)。

由于传统的豪斯曼检验在异方差的情形下不成立,下面进行异方差稳健的 DWH 检验:

```
. estat endogenous①
```

```
Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1)          =  3.87962  (p = 0.0489)
Wu-Hausman F(1,750)             =  3.85842  (p = 0.0499)
```

由于 DWH 检验的 p 值小于 0.05,故可认为 iq 为内生解释变量。下面使用“ivreg2”来进行稳健的内生性检验。

```
. ivreg2 lw s expr tenure rns smsa (iq = med kww),r endog(iq)
```

其中,选择项“endog(iq)”表示检验变量 iq 是否为内生变量。

① 如果在你的 Stata 版本中无法执行命令“estat endogenous”,可能是程序未更新。可以在命令窗口输入命令“update all”,更新全部程序后,再运行“estat endogenous”。

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                Number of obs =       758
                                                F(  6,   751) =     61.10
                                                Prob > F      =    0.0000
Total (centered) SS     =  139.2861498          Centered R2   =    0.2775
Total (uncentered) SS   =  24652.24662          Uncentered R2 =    0.9959
Residual SS             =  100.6291971          Root MSE      =     .3644

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0139284   .0060393     2.31   0.021     .0020916    .0257653
           s |   .0607803   .0189505     3.21   0.001      .023638    .0979227
        expr |   .0433237   .0074118     5.85   0.000     .0287968    .0578505
      tenure |   .0296442    .008317     3.56   0.000     .0133432    .0459452
         rns |  -.0435271   .0344779    -1.26   0.207    -.1111026    .0240483
        smsa |   .1272224   .0297414     4.28   0.000     .0689303    .1855146
       _cons |   3.218043   .3983683     8.08   0.000     2.437256    3.998831
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):              24.223
                                                   Chi-sq(2) P-val =     0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                14.906
                         (Kleibergen-Paap rk Wald F statistic):          13.403
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.151
                                                   Chi-sq(1) P-val =     0.6972
-endog- option:
Endogeneity test of endogenous regressors:                               3.615
                                                   Chi-sq(1) P-val =     0.0573
Regressors tested:   iq
------------------------------------------------------------------------------
Instrumented:        iq
Included instruments: s expr tenure rns smsa
Excluded instruments: med kww
------------------------------------------------------------------------------
```

上表显示，内生性检验的 $\chi^2(1)$ 统计量为 3.615，其 p 值为 5.73%，接近于上文“ivregress” Wu-Hausman F 检验的结果。

（10）如果存在异方差，则 GMM 比 2SLS 更有效率。为此，进行如下最优 GMM 估计。

```
. ivregress gmm lw s expr tenure rns smsa (iq = med kww)
```

```
Instrumental variables (GMM) regression                 Number of obs =       758
                                                        Wald chi2(6)  =    372.75
                                                        Prob > chi2   =    0.0000
                                                        R-squared     =    0.2750
GMM weight matrix: Robust                               Root MSE      =    .36499

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0140888   .0060357     2.33   0.020     .0022591    .0259185
           s |   .0603672   .0189545     3.18   0.001     .0232171    .0975174
        expr |   .0431117   .0074112     5.82   0.000     .0285861    .0576373
      tenure |   .0299764   .0082728     3.62   0.000      .013762    .0461908
         rns |   -.044516   .0344404    -1.29   0.196    -.1120179    .0229859
        smsa |   .1267368   .0297633     4.26   0.000     .0684018    .1850718
       _cons |   3.207298    .398083     8.06   0.000     2.427069    3.987526
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa med kww
```

上表显示,两步最优 GMM 的系数估计值与 2SLS 很接近。

进行过度识别检验:

```
. estat overid
```

```
  Test of overidentifying restriction:

  Hansen's J chi2(1) = .151451 (p = 0.6972)
```

由于 p 值为 0.70,故认为所有工具变量均为外生。考虑迭代 GMM:

```
. ivregress gmm lw s expr tenure rns smsa (iq = med kww),igmm
```

```
Iteration 1: change in beta =  1.753e-05   change in W =  1.100e-02
Iteration 2: change in beta =  4.872e-08   change in W =  7.880e-05
Iteration 3: change in beta =  2.501e-10   change in W =  2.304e-07

Instrumental variables (GMM) regression                 Number of obs =       758
                                                        Wald chi2(6)  =    372.73
                                                        Prob > chi2   =    0.0000
                                                        R-squared     =    0.2750
GMM weight matrix: Robust                               Root MSE      =    .36499

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0140901   .0060357     2.33   0.020     .0022603     .02592
           s |   .0603629   .0189548     3.18   0.001     .0232122    .0975135
        expr |   .0431101   .0074113     5.82   0.000     .0285841     .057636
      tenure |   .0299752   .0082729     3.62   0.000     .0137606    .0461898
         rns |  -.0445114   .0344408    -1.29   0.196    -.1120142    .0229913
        smsa |   .1267399   .0297637     4.26   0.000     .0684041    .1850757
       _cons |   3.207224   .3980878     8.06   0.000     2.426986    3.987462
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa med kww
```

容易看出,迭代 GMM 与两步 GMM 的系数估计值相差无几。

如果希望将以上各种估计法的系数估计值及其标准误列在同一张表中[①],可使用以下命令:

```
. qui reg lw s expr tenure rns smsa,r
. est sto ols_no_iq
. qui reg lw iq s expr tenure rns smsa,r
. est sto ols_with_iq
. qui ivregress 2sls lw s expr tenure rns smsa (iq = med kww),r
. est sto tsls
. qui ivregress liml lw s expr tenure rns smsa (iq = med kww),r
. est sto liml
. qui ivregress gmm lw s expr tenure rns smsa (iq = med kww)
. est sto gmm
. qui ivregress gmm lw s expr tenure rns smsa (iq = med kww),igmm
. est sto igmm
. estimates table ols_no_iq ols_with_iq tsls liml gmm igmm,b se
```

其中,选择项"b"表示显示回归系数,而选择项"se"表示显示标准误差。

Variable	ols_no_iq	ols_with~q	tsls	liml	gmm	igmm
s	.10264304	.09278735	.06078035	.06063623	.06036723	.06036285
	.00620988	.00697626	.01895051	.01903397	.01895452	.01895478
expr	.0381189	.03934425	.04332367	.04334159	.04311171	.04311006
	.00661439	.00666033	.00741179	.0074185	.00741117	.00741133
tenure	.03561456	.03420896	.02964421	.02962365	.02997643	.02997521
	.00799884	.00789567	.00831697	.00832297	.00827281	.00827289
rns	-.08407974	-.07453249	-.04352713	-.04338751	-.04451599	-.04451145
	.02953295	.02997719	.03447789	.03452902	.03444039	.03444082
smsa	.13966664	.13673691	.12722244	.1271796	.12673682	.12673991
	.02805598	.02777116	.02974144	.02975994	.0297633	.02976369
iq		.00327916	.01392844	.01397639	.01408883	.01409011
		.00113212	.00603931	.00606812	.00603567	.00603575
_cons	4.103675	3.8951718	3.2180433	3.2149943	3.2072978	3.2072239
	.08766646	.11592863	.39836829	.40014925	.39808304	.39808779

legend: b/se

如果希望用一颗星表示 10% 的显著性水平,两颗星表示 5% 的显著性水平,三颗星表示 1% 的显著性水平,则可以使用以下命令,

```
. estimates table ols_no_iq ols_with_iq tsls liml gmm igmm,star(0.1 0.05 0.01)
```

Variable	ols_no_iq	ols_with_iq	tsls	liml	gmm	igmm
s	.10264304***	.09278735***	.06078035***	.06063623***	.06036723***	.06036285***
expr	.0381189***	.03934425***	.04332367***	.04334159***	.04311171***	.04311006***
tenure	.03561456***	.03420896***	.02964421***	.02962365***	.02997643***	.02997521***
rns	-.08407974***	-.07453249**	-.04352713	-.04338751	-.04451599	-.04451145
smsa	.13966664***	.13673691***	.12722244***	.1271796***	.12673682***	.12673991***
iq		.00327916***	.01392844**	.01397639**	.01408883**	.01409011**
_cons	4.103675***	3.8951718***	3.2180433***	3.2149943***	3.2072978***	3.2072239***

legend: * p<.1; ** p<.05; *** p<.01

① 有时,在论文中也采用类似的表格,便于对各种估计法的结果进行比较。

遗憾的是,官方 Stata 命令“estimates table”无法同时显示回归系数、标准误差与表示显著性的星号(在正式的论文中通常需要同时显示)。为此,下载非官方命令 estout。

```
. ssc install estout
. esttab ols_no_iq ols_with_iq tsls liml gmm igmm,se r2 mtitle star(* 0.1 * * 0.05 * * * 0.01)
```

其中,选择项“se”表示在括弧中显示标准误差(默认值为显示 t 统计量,如果使用选择项“p”则显示 p 值),选择项“r2”表示显示 R^2(如果使用选择项“pr2”则显示准 R^2),选择项“mtitle”表示使用模型名字(model name)作为表中每一列的标题(默认使用被解释变量作为标题),选择项“star(* 0.1 * * 0.05 * * * 0.01)”表示以星号表示显著性水平。更多说明,参见“help estout”。

	(1) ols_no_iq	(2) ols_with_iq	(3) tsls	(4) liml	(5) gmm	(6) igmm
s	0.103***	0.0928***	0.0608***	0.0606***	0.0604***	0.0604***
	(0.00621)	(0.00698)	(0.0190)	(0.0190)	(0.0190)	(0.0190)
expr	0.0381***	0.0393***	0.0433***	0.0433***	0.0431***	0.0431***
	(0.00661)	(0.00666)	(0.00741)	(0.00742)	(0.00741)	(0.00741)
tenure	0.0356***	0.0342***	0.0296***	0.0296***	0.0300***	0.0300***
	(0.00800)	(0.00790)	(0.00832)	(0.00832)	(0.00827)	(0.00827)
rns	-0.0841***	-0.0745**	-0.0435	-0.0434	-0.0445	-0.0445
	(0.0295)	(0.0300)	(0.0345)	(0.0345)	(0.0344)	(0.0344)
smsa	0.140***	0.137***	0.127***	0.127***	0.127***	0.127***
	(0.0281)	(0.0278)	(0.0297)	(0.0298)	(0.0298)	(0.0298)
iq		0.00328***	0.0139**	0.0140**	0.0141**	0.0141**
		(0.00113)	(0.00604)	(0.00607)	(0.00604)	(0.00604)
_cons	4.104***	3.895***	3.218***	3.215***	3.207***	3.207***
	(0.0877)	(0.116)	(0.398)	(0.400)	(0.398)	(0.398)
N	758	758	758	758	758	758
R-sq	0.352	0.360	0.278	0.277	0.275	0.275

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

如果要将上表输出到一个 Microsoft Word 文档,并以文件名 iv 来命名此文档,则可运行如下命令:

```
. esttab ols_no_iq ols_with_iq tsls liml gmm igmm using iv.rtf,se r2 mtitle star(* 0.1 * * 0.05 * * * 0.01)
```

```
(output written to iv.rtf)
```

其中,“iv.rtf”的扩展名“rtf”表示“rich text format”。点击输出结果中的“iv.rtf”链接,即可打开此文件。

(11) 后续研究。由于教育投资回报率是劳动经济学的核心课题,研究者对此问题的兴趣长盛不衰。始自 Behrman et al (1980),不少经济学家通过比较教育年限不同的同卵双胞胎(identical twins)来控制遗传基因与家庭背景等因素。另一影响力深远的研究为 Angrist and Krueger (1991),使用出生的季度(quarter of birth)作为教育年限的工具变量。在美国,小学的入学时间为每年的 1 月 1 日,因此一季度出生小孩的入学年龄比四季度出生小孩更晚。另一方面,美国的强制教育法(compulsory schooling laws)要求青少年不得在 16 岁或 17 岁生日前离开学校

(不同州的法律略有不同)。这导致一季度出生小孩的平均教育年限低于二、三、四季度出生的小孩,因为一季度出生小孩更早达到法定退学年龄(legal dropout age);经验数据也证实这一点,故出生季度与教育年限相关。另一方面,出生季度很可能与家庭背景等因素无关,只可能通过教育年限来影响未来收入,故满足外生性。Angrist and Krueger (1991)发现,以出生季度作为工具变量进行 2SLS 估计,得到的结果与 OLS 相似。在此之后,出生季度被广泛地作为工具变量使用。然而,Bound et al (1995)发现,出生季度为弱工具变量,在第一阶段回归中检验工具变量联合显著性的 F 统计量远小于 10,故即使在大样本中仍是有偏估计。另外,根据 Buckles and Hungerman (2012)的最新研究,出生季度也并非与家庭背景无关。冬天出生的孩子的母亲更可能是单身、未满二十岁(teenagers)、非白人(nonwhite)或没有高中文凭(出现此规律的原因尚不清楚)。比如,一月份出生的孩子的母亲没有高中文凭的比例比五月份出生的孩子的母亲高 10%。因此,出生季度其实也不满足外生性,以之作为工具变量仍会导致不一致的估计。经济学家寻找教育年限的有效工具变量的努力仍将继续。

习　题

10.1　如果"工具变量"不满足外生性,则 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 还是 $\boldsymbol{\beta}$ 的一致估计量吗?为什么?

10.2　考虑以下简单宏观经济模型中的消费函数

$$\begin{cases} C_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases} \tag{10.57}$$

其中,Y_t, C_t, I_t, G_t, X_t 分别为国民收入、总消费、总投资、政府净支出与净出口。证明:如果单独对第一个消费方程进行 OLS 估计,将得到不一致的估计。

10.3　在上题中,假设消费函数没有常数项,即 $C_t = \alpha_1 Y_t + \varepsilon_t$(弗里德曼的消费函数)。证明:(1)可以用常数项作为工具变量(放宽对工具变量相关性的要求为"不与解释变量正交");(2)工具变量估计量为 $\hat{\alpha}_{1,\text{IV}} = \bar{C}/\bar{Y}$,其中 $\bar{C} \equiv \frac{1}{n}\sum_{t=1}^{n} C_t$,$\bar{Y} \equiv \frac{1}{n}\sum_{t=1}^{n} Y_t$(提示:对消费方程两边同时取期望)。

10.4　(对效率 GMM 的 GLS 解释)考虑回归模型 $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times K}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$,而 $\boldsymbol{Z}_{n\times L}$ 为由工具变量构成的矩阵(每列都是一个工具变量),$L \geqslant K$。将原模型两边同时左乘 $\boldsymbol{Z}'$ 可得,$\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{\varepsilon}$。记 $\boldsymbol{S} \equiv \text{Var}(\boldsymbol{Z}'\boldsymbol{\varepsilon})$,而 $\hat{\boldsymbol{S}}$ 为 $\boldsymbol{S}$ 的一致估计,对此转换后的模型使用 FGLS 估计,并证明它是最优 GMM 估计量。

10.5　参照本章的 Stata 实例,使用数据集 grilic. dta 中 1980 年的变量重新进行 OLS,2SLS,GMM 估计以及相应的检验,比如"`reg lw80 s80 iq expr80 tenure80 rns80 smsa80,r`"。

附　录

A10.1(豪斯曼检验)　在 H_0 成立的情况下,如果 OLS 是最有效率的(fully efficient),则 $\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{IV}}, \hat{\boldsymbol{\beta}}_{\text{OLS}}) = \text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$。

证明:为了简单起见,只考虑一维的情形,即 $\hat{\beta}_{\text{IV}}$ 与 $\hat{\beta}_{\text{OLS}}$ 都是一维的。多维情形可类似地证明,参见 Hausman (1978)。首先,定义这两个估计量之差为 $\hat{q} \equiv \hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}$。由于 $\text{Cov}(\hat{\beta}_{\text{IV}}, \hat{\beta}_{\text{OLS}}) = \text{Cov}(\hat{\beta}_{\text{OLS}} + \hat{q}, \hat{\beta}_{\text{OLS}}) = \text{Var}(\hat{\beta}_{\text{OLS}}) + \text{Cov}(\hat{q}, \hat{\beta}_{\text{OLS}})$,故只需证明 $\text{Cov}(\hat{q}, \hat{\beta}_{\text{OLS}}) = 0$ 即可。

其次,在 H_0 成立的情况下,$\hat{\beta}_{\text{IV}}$ 与 $\hat{\beta}_{\text{OLS}}$ 都是一致估计量,故

$\text{plim}\hat{q} \equiv \text{plim}\hat{\beta}_{IV} - \text{plim}\hat{\beta}_{OLS} = \beta - \beta = 0$。定义新估计量 $\hat{\beta} = \hat{\beta}_{OLS} + \lambda\hat{q}$,其中 λ 为任意常数。显然,$\hat{\beta}$ 也是一致估计量,即 $\text{plim}\hat{\beta} = \beta$。估计量 $\hat{\beta}$ 的方差为

$$\text{Var}(\hat{\beta}) = \text{Var}(\hat{\beta}_{OLS}) + \lambda^2\text{Var}(\hat{q}) + 2\lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) \geqslant \text{Var}(\hat{\beta}_{OLS}) \tag{10.58}$$

由于OLS估计量是最有效率的,故在所有一致估计量的方差中,$\text{Var}(\hat{\beta}_{OLS})$ 最小,上式中的不等式成立。上式两边消去 $\text{Var}(\hat{\beta}_{OLS})$ 可得,

$$\lambda^2\text{Var}(\hat{q}) + 2\lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) \geqslant 0 \tag{10.59}$$

此不等式对于任意 λ 都成立。下面将说明,除非 $\text{Cov}(\hat{q},\hat{\beta}_{OLS}) = 0$,此不等式不可能对任意 λ 都成立。

首先,如果 $\text{Cov}(\hat{q},\hat{\beta}_{OLS}) > 0$,则可令 $\lambda = -\dfrac{\text{Cov}(\hat{q},\hat{\beta}_{OLS})}{\text{Var}(\hat{q})} < 0$,使得 $\lambda^2\text{Var}(\hat{q}) + 2\lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) = \lambda[\lambda\text{Var}(\hat{q}) + 2\text{Cov}(\hat{q},\hat{\beta}_{OLS})] = \lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) < 0$,得到矛盾。

其次,如果 $\text{Cov}(\hat{q},\hat{\beta}_{OLS}) < 0$,则可令 $\lambda = -\dfrac{\text{Cov}(\hat{q},\hat{\beta}_{OLS})}{\text{Var}(\hat{q})} > 0$,故 $\lambda^2\text{Var}(\hat{q}) + 2\lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) = \lambda\text{Cov}(\hat{q},\hat{\beta}_{OLS}) < 0$,同样得到矛盾。因此,$\text{Cov}(\hat{q},\hat{\beta}_{OLS}) = 0$,得证。

A10.2(GMM估计量) GMM估计量为 $\hat{\boldsymbol{\beta}}_{GMM}(\hat{\boldsymbol{W}}) = (\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy}$

证明:GMM估计量的最小化目标函数为

$$\begin{aligned} J(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{W}}) &= n(\boldsymbol{S}_{Zy} - \boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{W}}(\boldsymbol{S}_{Zy} - \boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) = n(\boldsymbol{S}'_{Zy} - \hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX})\hat{\boldsymbol{W}}(\boldsymbol{S}_{Zy} - \boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) \\ &= n(\boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}} - \hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}})(\boldsymbol{S}_{Zy} - \boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) \\ &= n(\boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} - \hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} - \boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) \\ &= n(\boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} + \hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{10.60}$$

其中,$(\hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy})' = \boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}$,且都是一维标量(因为目标函数是一维标量),故 $\hat{\boldsymbol{\beta}}'\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} = \boldsymbol{S}'_{Zy}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}$,可以合并在一起。

使用向量微分法则(参见第3章),对 $\hat{\boldsymbol{\beta}}$ 求导,可以得到最小化的一阶条件,

$$\frac{\partial J(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{W}})}{\partial\hat{\boldsymbol{\beta}}} = n(-2\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} + 2\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}}) = 0 \tag{10.61}$$

方程两边同时消去 $2n$,移项可得,

$$\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX}\hat{\boldsymbol{\beta}} = \boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy} \tag{10.62}$$

如果 $(\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}$ 存在,则 $\hat{\boldsymbol{\beta}}_{GMM}(\hat{\boldsymbol{W}}) = (\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{ZX})^{-1}\boldsymbol{S}'_{ZX}\hat{\boldsymbol{W}}\boldsymbol{S}_{Zy}$,得证。

第 11 章　二值选择模型

11.1　离散被解释变量的例子

如果解释变量是离散的(比如,虚拟变量),这并不影响回归。但有时被解释变量是离散的,而非连续的。比如,个体的如下选择行为(人生充满了选择)。

二值选择(binary choices):考研或不考研;就业或待业;买房或不买房;买保险或不买保险;贷款申请被批准或拒绝;出国或不出国;回国或不回国;战争或和平;生或死。

多值选择(multiple choices):对不同交通方式的选择(走路、骑车、坐车);对不同职业的选择。

这类模型被称为"离散选择模型"(discrete choice model)或"定性反应模型"(qualitative response model)。另外,有时被解释变量只能取非负整数。比如,企业在某段时间内获得的专利数;某人在一定时间内去医院看病的次数;某省在一年内发生煤矿事故的次数。这类数据被称为"计数数据"(count data),其被解释变量也是离散的。考虑到离散被解释变量的特点,通常不宜用 OLS 进行回归。

本章考虑二值选择模型,第 12 章考虑多值选择模型,而第 13 章考虑排序与计数模型。

11.2　二值选择模型

假设个体只有两种选择,比如 $y=1$(考研)或 $y=0$(不考研)。是否考研,取决于研究生毕业后的预期收入、个人兴趣、本科毕业后直接就业的收入前景等。假设这些解释变量都包括在向量 $\boldsymbol{x}$ 中。

最简单的模型为"线性概率模型"(Linear Probability Model,简记 LPM):

$$y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i \quad (i=1,\cdots,n) \tag{11.1}$$

对 $\boldsymbol{\beta}$ 的一致估计要求 $\mathrm{Cov}(\boldsymbol{x}_i,\varepsilon_i)=0$。然而,由于 $\varepsilon_i=y_i-\boldsymbol{x}_i'\boldsymbol{\beta}$,故 $\varepsilon_i=1-\boldsymbol{x}_i'\boldsymbol{\beta}$ 或 $\varepsilon_i=-\boldsymbol{x}_i'\boldsymbol{\beta}$,因此 ε_i 必然与 $\boldsymbol{x}_i$ 相关,导致估计不一致。显然,ε_i 服从两点分布,而非正态分布。而且,由于 $\mathrm{Var}(\varepsilon_i)=\mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta})$,故扰动项 ε_i 的方差依赖于 $\boldsymbol{x}_i$,存在异方差(故应使用稳健标准误)。另一困难是,虽然明知被解释变量 y 的取值非 0 即 1,但根据这个线性概率模型所作的预测值却可能出现 $\hat{y}>1$ 或 $\hat{y}<0$ 的不现实情形,参见图 11.1。由于线性概率模型的以上缺点,故一般只将其作为粗略的参考。尽管如此,线性概率模型的优点是,计算方便,而且容易得到边际效应。

为了使 y 的预测值总是介于[0,1]之间,在给定 $\boldsymbol{x}$ 的情况下,考虑 y 的两点分布概率:

$$\begin{cases} \mathrm{P}(y=1\mid \boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) \\ \mathrm{P}(y=0\mid \boldsymbol{x}) = 1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases} \tag{11.2}$$

此函数 $F(\boldsymbol{x},\boldsymbol{\beta})$ 也称为“连接函数”(link function),因为它将解释变量 $\boldsymbol{x}$ 与被解释变量 y 连接起来。由于 y 的取值要么为 0,要么为 1,故 y 肯定服从两点分布。连接函数的选择具有一定的灵活性。通过选择合适的连接函数 $F(\boldsymbol{x},\boldsymbol{\beta})$(比如,某随机变量的累积分布函数),可以保证 $0\leqslant\hat{y}\leqslant 1$,并将 $\hat{y}$ 理解为“$y=1$”发生的概率,因为

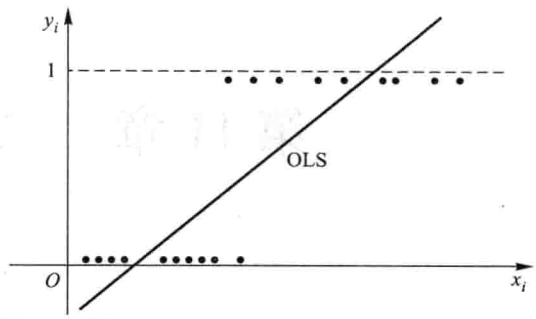


图 11.1 OLS 与二值选择模型

$$\mathrm{E}(y\mid \boldsymbol{x}) = 1\cdot \mathrm{P}(y=1\mid \boldsymbol{x}) + 0\cdot \mathrm{P}(y=0\mid \boldsymbol{x}) = \mathrm{P}(y=1\mid \boldsymbol{x}) \tag{11.3}$$

如果 $F(\boldsymbol{x},\boldsymbol{\beta})$ 为标准正态的累积分布函数(cdf),则

$$\mathrm{P}(y=1\mid \boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(t)\,\mathrm{d}t \tag{11.4}$$

该模型被称为“Probit”。如果 $F(\boldsymbol{x},\boldsymbol{\beta})$ 为“逻辑分布”(logistic distribution)的累积分布函数,则

$$\mathrm{P}(y=1\mid \boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \tag{11.5}$$

该模型称为“Logit”。逻辑分布的累积分布函数的图形与标准正态分布的图形比较相似,其密度函数关于原点对称,期望为 0,方差为 $\pi^2/3$(大于标准正态的方差)。与标准正态相比,逻辑分布具有厚尾(fat tails),更接近于自由度为 7 的 t 分布。

由于逻辑分布的累积分布函数有解析表达式(而标准正态分布没有),故计算 Logit 通常比 Probit 更为方便。显然,这是一个非线性模型,可使用最大似然法(MLE)进行估计。以 Logit 模型为例。第 i 个观测数据的概率密度为

$$f(y_i\mid \boldsymbol{x}_i,\boldsymbol{\beta}) = \begin{cases} \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), \text{若 } y_i=1 \\ 1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), \text{若 } y_i=0 \end{cases} \tag{11.6}$$

将其更紧凑地写为

$$f(y_i\mid \boldsymbol{x}_i,\boldsymbol{\beta}) = [\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]^{1-y_i} \tag{11.7}$$

取对数可得

$$\ln f(y_i\mid \boldsymbol{x}_i,\boldsymbol{\beta}) = y_i\ln[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] + (1-y_i)\ln[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] \tag{11.8}$$

假设样本中的个体相互独立,则整个样本的对数似然函数为

$$\ln L(\boldsymbol{\beta}\mid \boldsymbol{y},\boldsymbol{x}) = \sum_{i=1}^{n} y_i\ln[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] + \sum_{i=1}^{n}(1-y_i)\ln[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] \tag{11.9}$$

可以使用数值计算的方法来求解这个非线性最大化问题。

需要注意的是,在这个非线性模型中,估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ 并非边际效应(marginal effects)。以 Probit 为例,

$$\frac{\partial \mathrm{P}(y=1\mid \boldsymbol{x})}{\partial x_k} = \frac{\partial \mathrm{P}(y=1\mid \boldsymbol{x})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\cdot\frac{\partial(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k} = \phi(\boldsymbol{x}'\boldsymbol{\beta})\cdot\beta_k \tag{11.10}$$

在上式中,使用了微分的链锁法则(chain rule),而且假设 x_k 为连续变量。由于 Probit 与 Logit 所使用的分布函数不同,其参数估计值并不直接可比。需要分别计算二者的边际效应,然后进行比较。然而,对于非线性模型而言,边际效应本身也不是常数,它随着解释变量而变。常

用的边际效应概念包括：

（1）平均边际效应（average marginal effect），即分别计算在每个样本观测值上的边际效应，然后进行简单算术平均。

（2）样本均值处的边际效应（marginal effect at mean），即在 $\boldsymbol{x}=\bar{\boldsymbol{x}}$ 处的边际效应。

（3）在某代表值处的边际效应（marginal effect at a representative value），即给定 $\boldsymbol{x}^*$，在 $\boldsymbol{x}=\boldsymbol{x}^*$ 处的边际效应。

以上三种边际效应的计算结果可能有较大差异。传统上，常计算样本均值处的边际效应（计算方便）。但在非线性模型中，样本均值处的个体行为并不等于样本中个体的平均行为（average behavior of individuals differs from behavior of the average individual）。对于政策分析而言，使用平均边际效应（Stata 的默认方法），或在某代表值处的边际效应通常更有意义。

既然 $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ 并非边际效应，那么它究竟有什么含义呢？对于 Logit 模型，记 $p\equiv \text{P}(y=1\mid \boldsymbol{x})$，则 $1-p=\text{P}(y=0\mid \boldsymbol{x})$。由于 $p=\dfrac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$，$1-p=\dfrac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$，故

$$\frac{p}{1-p}=\exp(\boldsymbol{x}'\boldsymbol{\beta}) \tag{11.11}$$

$$\ln\left(\frac{p}{1-p}\right)=\boldsymbol{x}'\boldsymbol{\beta} \tag{11.12}$$

其中，“$p/(1-p)$”被称为“几率比”（odds ratio）或“相对风险”（relative risk）。假设在一个检验药物疗效的随机实验中，“$y=1$”表示“生”，而“$y=0$”表示“死”。如果几率比为 2，则意味着存活的概率是死亡概率的两倍。对方程（11.12）右边求导数可知，$\hat{\beta}_j$ 表示解释变量 x_j 增加一个微小量引起“对数几率比”（log-odds ratio）的边际变化。从另一角度，可把 $\hat{\beta}_j$ 视为半弹性，即 x_j 增加一单位引起几率比的变化百分比。比如，$\hat{\beta}_j=0.12$，意味着 x_j 增加一单位引起几率比增加 12%。

另一解释方法为：假设 x_j 增加一单位，从 x_j 变为 x_j+1，记 p 的新值为 p^*，则新几率比与原先几率比的比率可以写为（此处不使用微积分）

$$\frac{\dfrac{p^*}{1-p^*}}{\dfrac{p}{1-p}}=\frac{\exp[\beta_1+\beta_2x_2+\cdots+\beta_j(x_j+1)+\cdots+\beta_Kx_K]}{\exp(\beta_1+\beta_2x_2+\cdots+\beta_jx_j+\cdots+\beta_Kx_K)}=\exp(\beta_j) \tag{11.13}$$

为此，有些研究者（特别在生物统计领域）偏好计算 $\exp(\hat{\beta}_j)$，它表示解释变量 x_j 增加一单位引起几率比的变化倍数。比如，$\hat{\beta}_j=0.12$，则 $\exp(\hat{\beta}_j)=e^{0.12}=1.13$，故当 x_j 增加一单位时，新几率比是原先几率比的 1.13 倍，或增加 13%，因为 $\exp(\hat{\beta}_j)-1=1.13-1=0.13$。基于此，Stata 称 $\exp(\hat{\beta}_j)$ 为几率比（odds ratio）。事实上，如果 $\hat{\beta}_j$ 较小，则 $\exp(\hat{\beta}_j)-1\approx\hat{\beta}_j$（将 $\exp(\hat{\beta}_j)$ 泰勒展开），此时以上两种方法是等价的。然而，如果 x_j 至少必须变化一个单位（比如性别、婚否等虚拟变量，年龄，子女个数），则应使用 $\exp(\hat{\beta}_j)$。需要注意的是，对于 Probit 模型，无法对其系数 $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ 进行类似的解释；这是 Probit 模型的劣势。

如何衡量（非线性）二值模型的拟合优度呢？由于不存在平方和分解公式，故无法计算 R^2。Stata 仍然汇报一个“准 R^2”（Pseudo R^2），由 McFadden（1974）所提出，其定义为

$$\text{准 } R^2\equiv\frac{\ln L_0-\ln L_1}{\ln L_0} \tag{11.14}$$

其中，$\ln L_1$ 为原模型的对数似然函数之最大值，而 $\ln L_0$ 为以常数项为唯一解释变量的对数似然函数之最大值。由于 y 为离散的两点分布，似然函数的最大可能值为 1，故对数似然函数的最大可能值为 0，记为 $\ln L_{max}$。显然，$0 \geqslant \ln L_1 \geqslant \ln L_0$，而 $0 \leqslant$ 准 $R^2 \leqslant 1$，参见图 11.2。由于“准 R^2”可以写为 $\dfrac{\ln L_1 - \ln L_0}{\ln L_{max} - \ln L_0}$，故可以将其视为对数似然函数的实际增加值（$\ln L_1 - \ln L_0$）占最大可能增加值（$\ln L_{max} - \ln L_0$）的比重。

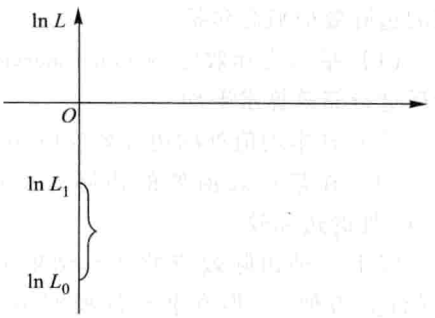


图 11.2 准 R^2 的计算

判断拟合优度的另一方法是计算“正确预测的百分比”（precent correctly predicted）。如果发生概率的预测值 $\hat{y} \geqslant 0.5$，则认为其预测 $y=1$；反之，则认为其预测 $y=0$。将预测值与实际值（样本数据）进行比较，就能计算正确预测的百分比。另外，Stata 还会汇报一个似然比检验统计量（LR），检验除常数项以外所有其他系数的显著性。

对于 Probit 与 Logit 模型，如果分布函数设定不正确，则为准最大似然估计（QMLE）。由于二值选择模型的分布必然为两点分布（属于线性指数分布族），故只要条件期望函数 $E(y|\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta})$ 正确，则 MLE 估计就是一致的（参见第 6 章第 8 节）。由于两点分布的特殊性，在 iid 的情况下，只要 $E(y|\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta})$ 成立，稳健标准误就等于 MLE 的普通标准误。因此，如果认为模型设定正确，就没有必要使用稳健标准误（但使用稳健标准误也没有错）。

反之，如果模型设定不正确（即 $E(y|\boldsymbol{x}) \neq F(\boldsymbol{x},\boldsymbol{\beta})$），则 Probit 与 Logit 模型并不能得到对系数 $\boldsymbol{\beta}$ 的一致估计，使用稳健标准误也就没有太大意义（只是更精确地估计了错误参数的标准误）；首先应解决参数估计的一致性问题。另一方面，如果稳健标准误与普通标准误相差甚远，则可大致诊断模型的设定不正确。然而，如果数据并非 iid，比如可将样本分为若干组（聚类），而每组内的个体存在组内自相关，则应使用聚类稳健的标准误。

二值模型的 Stata 命令为

```
probit y x1 x2 x3,r                                  (probit 模型)
logit y x1 x2 x3,or vce(cluster clustvar)            (logit 模型)
```

其中，选择项“r”表示使用稳健标准误，选择项“or”表示显示几率比（而不显示系数），选择项“vce(cluster *clustvar*)”表示使用以“*clustvar*”为聚类变量的聚类稳健标准误。

完成估计后，可用以下命令进行预测，并计算准确预测的百分比：

```
predict yhat                    (计算发生概率的预测值 ŷ，并记为“yhat”)
estat clas                      (计算预测准确的百分比，clas 表示 classification)
```

在 Stata 12 中，计算边际效应的命令为[①]，

```
margins,dydx(*)                 (计算所有解释变量的平均边际效应)
margins,dydx(*) atmeans         (计算所有解释变量在样本均值处的边际效应)
margins,dydx(*) at(x1=0)        (计算所有解释变量在 x1=0 处的边际效应)
margins,dydx(x1)                (计算解释变量 x1 的平均边际效应)
margins,eyex(*)                 (计算平均弹性，其中两个“e”均指 elasticity)
```

① Stata 12 的“margins”命令取代了 Stata 10 的“mfx”命令，前者比后者功能更为强大（比如，后者无法计算平均边际效应），详见帮助文件。

```
margins,eydx(*)          (计算平均半弹性,x 变化 1 单位引起 y 变化百分之几)
margins,dyex(*)          (计算平均半弹性,x 变化 1% 引起 y 变化几个单位)
```

其中,“*”代表所有解释变量。

下面以数据集 womenwk.dta 为例[①],估计决定美国妇女就业与否的二值选择模型。该数据集包括以下变量:work(是否就业),age(年龄),married(婚否),children(子女数),education(教育年限)。考虑以下模型:

$$\text{work}_i = \beta_0 + \beta_1 \text{ age}_i + \beta_2 \text{ married}_i + \beta_3 \text{ children}_i + \beta_4 \text{ education}_i + \varepsilon_i \tag{11.15}$$

作为对照,首先使用 OLS 进行线性概率模型(LPM)估计:

```
. use womenwk.dta,clear
. reg work age married children education,r
```

```
Linear regression                               Number of obs =    2000
                                                F(  4,  1995) =  192.58
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2026
                                                Root MSE      =  .41992

------------------------------------------------------------------------------
             |               Robust
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0102552   .0012236     8.38   0.000     .0078556    .0126548
     married |   .1111116   .0226719     4.90   0.000     .0666485    .1555748
    children |   .1153084   .0056978    20.24   0.000     .1041342    .1264827
   education |   .0186011   .0033006     5.64   0.000     .0121282     .025074
       _cons |  -.2073227   .0534581    -3.88   0.000    -.3121622   -.1024832
------------------------------------------------------------------------------
```

其次,使用 Logit 进行估计:

```
. logit work age married children education,nolog
```

```
Logistic regression                             Number of obs   =       2000
                                                LR chi2(4)      =     476.62
                                                Prob > chi2     =     0.0000
Log likelihood = -1027.9144                     Pseudo R2       =     0.1882

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0579303    .007221     8.02   0.000     .0437773    .0720833
     married |   .7417775   .1264705     5.87   0.000     .4938998    .9896552
    children |   .7644882   .0515289    14.84   0.000     .6634935     .865483
   education |   .0982513   .0186522     5.27   0.000     .0616936     .134809
       _cons |  -4.159247   .3320401   -12.53   0.000    -4.810034   -3.508461
------------------------------------------------------------------------------
```

上表显示,准 R^2 为 0.19。LR 统计量为 476.62,对应的 p 值为 0.00,故整个方程所有系数(除常数项外)的联合显著性很高。下面,使用稳健标准误进行 Logit 估计。

```
. logit work age married children education,r nolog
```

① 此例来自 Baum(2006)。

```
Logistic regression                                     Number of obs   =       2000
                                                        Wald chi2(4)    =     344.54
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1027.9144                       Pseudo R2       =     0.1882
```

work	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0579303	.0072054	8.04	0.000	.0438079	.0720527
married	.7417775	.1272191	5.83	0.000	.4924326	.9911224
children	.7644882	.0497584	15.36	0.000	.6669635	.8620129
education	.0982513	.019011	5.17	0.000	.0609904	.1355121
_cons	-4.159247	.327398	-12.70	0.000	-4.800936	-3.517559

对比以上两个表格可知，稳健标准误与普通标准误非常接近，故大致可以不必担心模型设定问题。由于各解释变量(age，married，children，education)的最小变化量至少为一单位，为了便于解释回归结果，下面让 Stata 汇报几率比而非系数。

```
. logit work age married children education,or nolog
```

```
Logistic regression                                     Number of obs   =       2000
                                                        LR chi2(4)      =     476.62
                                                        Prob > chi2     =     0.0000
Log likelihood = -1027.9144                             Pseudo R2       =     0.1882
```

work	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.059641	.0076517	8.02	0.000	1.04475	1.074745
married	2.099664	.2655457	5.87	0.000	1.638694	2.690307
children	2.147895	.1106786	14.84	0.000	1.941563	2.376153
education	1.10324	.0205779	5.27	0.000	1.063636	1.144318
_cons	.0156193	.0051862	-12.53	0.000	.0081476	.029943

上表显示，在给定其他变量的情况下，已婚妇女参加工作的几率比是未婚妇女的 2.10 倍(即高出 110%)；而年龄每增加一岁，参加工作的几率比就会增加 6%；其他变量对应的几率比可以类似地解释。为了与 OLS 估计的回归系数比较，计算 Probit 模型的平均边际效应：

```
. margins,dydx(*)
```

```
Average marginal effects                                Number of obs   =       2000
Model VCE    : OIM

Expression   : Pr(work), predict()
dy/dx w.r.t. : age married children education
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0099674	.0011682	8.53	0.000	.0076778	.0122569
married	.127629	.021152	6.03	0.000	.0861717	.1690862
children	.1315365	.007073	18.60	0.000	.1176736	.1453994
education	.0169049	.0031243	5.41	0.000	.0107814	.0230285

简单目测可知，Logit 模型的平均边际效应与 OLS 回归系数相差不大。为了演示的目的，下面计算在样本均值处的边际效应。

```
. margins,dydx(*) atmeans
```

```
Conditional marginal effects                    Number of obs   =       2000
Model VCE    : OIM

Expression   : Pr(work), predict()
dy/dx w.r.t. : age married children education
at           : age             =      36.208 (mean)
               married         =       .6705 (mean)
               children        =      1.6445 (mean)
               education       =      13.084 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0115031   .0014236     8.08   0.000     .0087129    .0142934
     married |   .1472934   .0248209     5.93   0.000     .0986453    .1959415
    children |    .151803   .0093768    16.19   0.000     .1334249    .1701812
   education |   .0195096   .0036991     5.27   0.000     .0122596    .0267596
------------------------------------------------------------------------------
```

对比以上两个表格的输出结果可知，在样本均值处的边际效应与平均边际效应有所不同。如果只关心变量 age 在“age = 30”处的边际效应，可输入命令，

```
. margins,dydx(age) at(age = 30)
```

```
Average marginal effects                        Number of obs   =       2000
Model VCE    : OIM

Expression   : Pr(work), predict()
dy/dx w.r.t. : age
at           : age             =          30

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .011179   .0014719     7.59   0.000      .008294    .0140639
------------------------------------------------------------------------------
```

从上表可知，对于此非线性模型，变量 age 在“age = 30”处的边际效应并不等于该变量的平均边际效应，也不等于在样本均值处的边际效应。下面计算 Logit 模型准确预测的比率：

```
. estat clas
```

```
Logistic model for work

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |      1177            361  |       1538
     -     |       166            296  |        462
-----------+--------------------------+-----------
   Total   |      1343            657  |       2000

Classified + if predicted Pr(D) >= .5
True D defined as work != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   87.64%
Specificity                     Pr( -|~D)   45.05%
Positive predictive value       Pr( D| +)   76.53%
Negative predictive value       Pr(~D| -)   64.07%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   54.95%
False - rate for true D         Pr( -| D)   12.36%
False + rate for classified +   Pr(~D| +)   23.47%
False - rate for classified -   Pr( D| -)   35.93%
--------------------------------------------------
Correctly classified                        73.65%
--------------------------------------------------
```

上表显示,正确预测的比率为(1 177 + 296)/2 000 = 73.65%。为了演示的目的,假设年龄相同的个体存在组内相关,故使用 age 为聚类变量来计算聚类稳健的标准误。

```
. logit work age married children education,nolog vce(cluster age)
```

```
Logistic regression                                     Number of obs   =       2000
                                                        Wald chi2(4)    =     576.81
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1027.9144                       Pseudo R2       =     0.1882

                                        (Std. Err. adjusted for 40 clusters in age)
------------------------------------------------------------------------------
             |               Robust
        work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0579303   .0055907    10.36   0.000     .0469728    .0688879
     married |   .7417775   .1084937     6.84   0.000     .5291337    .9544213
    children |   .7644882   .0540759    14.14   0.000     .6585014     .870475
   education |   .0982513   .0148423     6.62   0.000     .0691609    .1273416
       _cons |  -4.159247   .2494119   -16.68   0.000    -4.648086   -3.670409
------------------------------------------------------------------------------
```

类似地,可以对此模型进行 Probit 估计:

```
. probit work age married children education,nolog
```

```
Probit regression                                       Number of obs   =       2000
                                                        LR chi2(4)      =     478.32
                                                        Prob > chi2     =     0.0000
Log likelihood = -1027.0616                             Pseudo R2       =     0.1889

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0347211   .0042293     8.21   0.000     .0264318    .0430105
     married |   .4308575    .074208     5.81   0.000     .2854125    .5763025
    children |   .4473249   .0287417    15.56   0.000     .3909922    .5036576
   education |   .0583645   .0109742     5.32   0.000     .0368555    .0798735
       _cons |  -2.467365   .1925635   -12.81   0.000    -2.844782   -2.089948
------------------------------------------------------------------------------
```

下面考察 Probit 模型的边际效应及预测准确度。

```
. margins,dydx(*)
```

```
Average marginal effects                                Number of obs   =       2000
Model VCE    : OIM

Expression   : Pr(work), predict()
dy/dx w.r.t. : age married children education

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0100768   .0011647     8.65   0.000     .0077941    .0123595
     married |   .1250441   .0210541     5.94   0.000     .0837788    .1663094
    children |   .1298233   .0068418    18.98   0.000     .1164137    .1432329
   education |   .0169386   .0031183     5.43   0.000     .0108269    .0230504
------------------------------------------------------------------------------
```

```
. estat clas
```

```
Probit model for work

                    True
Classified |      D          ~D      |     Total
-----------+------------------------+----------
     +     |     1177          361   |      1538
     -     |      166          296   |       462
-----------+------------------------+----------
   Total   |     1343          657   |      2000

Classified + if predicted Pr(D) >= .5
True D defined as work != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   87.64%
Specificity                     Pr( -|~D)   45.05%
Positive predictive value       Pr( D| +)   76.53%
Negative predictive value       Pr(~D| -)   64.07%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   54.95%
False - rate for true D         Pr( -| D)   12.36%
False + rate for classified +   Pr(~D| +)   23.47%
False - rate for classified -   Pr( D| -)   35.93%
--------------------------------------------------
Correctly classified                        73.65%
--------------------------------------------------
```

从以上各表可以看出，Logit 模型的边际效应、准 R^2 以及正确预测比率与 Probit 模型几乎完全相同，故可视为基本等价（两者的估计系数虽有差距，但估计系数没有可比性）。

11.3 二值选择模型的微观基础

在上节的 Probit 或 Logit 模型中，似乎看不到扰动项的存在。为此，本节考察二值选择模型的微观基础。对于二值选择行为，通常可通过一个“潜变量”（latent variable）来概括该行为的净收益（收益减去成本）。如果净收益大于 0，则选择做；否则，选择不做。假设净收益为

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \tag{11.16}$$

其中，净收益 y^* 为潜变量，不可观测。上式也被称为“指数函数”（index function）。个体的选择规则为

$$y = \begin{cases} 1, & \text{若 } y^* > 0 \\ 0, & \text{若 } y^* \leqslant 0 \end{cases} \tag{11.17}$$

因此，

$$\mathrm{P}(y=1 \mid \boldsymbol{x}) = \mathrm{P}(y^* > 0 \mid \boldsymbol{x}) = \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0 \mid \boldsymbol{x}) = \mathrm{P}(\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}) \tag{11.18}$$

假设 $\varepsilon \sim N(0, \sigma^2)$ 或服从逻辑分布，则

$$\mathrm{P}(y=1 \mid \boldsymbol{x}) = \mathrm{P}(\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}) = \mathrm{P}(\varepsilon < \boldsymbol{x}'\boldsymbol{\beta}) = F_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta}) \tag{11.19}$$

其中，$F_\varepsilon(\cdot)$ 为 ε 的累积分布函数，方程(11.19)的第二个等号用到了密度函数关于原点对称的性质。这个结果与方程(13.2)相同。如果 ε 服从正态分布，则为 Probit；如果 ε 服从逻辑分布，则为 Logit。

应该注意的是，对于任意常数 $k > 0$，$\mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0) = \mathrm{P}(k\boldsymbol{x}'\boldsymbol{\beta} + k\varepsilon > 0)$。记扰动项方差 $\sigma^2 \equiv \mathrm{Var}(\varepsilon)$，则 $\mathrm{Var}(k\varepsilon) = k^2\sigma^2$，故 $(k\boldsymbol{\beta}, k^2\sigma^2)$ 对模型的拟合与 $(\boldsymbol{\beta}, \sigma^2)$ 完全一样，故无法同时“识别”（identify）$\boldsymbol{\beta}$ 与 σ^2。因此，对于 Probit 模型，令扰动项之方差 σ^2 为 1，即 $\varepsilon \sim N(0,1)$；而对于 Logit 模型，则令扰动项之方差为 $\pi^2/3$。

另外一种可能的微观基础为“随机效用最大化”模型(Random Utility Maximization,简记 RUM)。假设选择 a,可带来效用 U_a;选择 b,可带来效用 U_b。如果 $U_a > U_b$,则选 a,记 $y=1$;如果 $U_a \leqslant U_b$,则选 b,记 $y=0$。由于存在很多决定效用的未知因素以及未来的不确定性,效用方程中包含一个扰动项,故名“随机效用”。假定 $U_a = \boldsymbol{x}'\boldsymbol{\beta}_a + \varepsilon_a$,$U_b = \boldsymbol{x}'\boldsymbol{\beta}_b + \varepsilon_b$,则

$$\begin{aligned} \mathrm{P}(y=1 \mid \boldsymbol{x}) &= \mathrm{P}(U_a > U_b \mid \boldsymbol{x}) \\ &= \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta}_a + \varepsilon_a > \boldsymbol{x}'\boldsymbol{\beta}_b + \varepsilon_b \mid \boldsymbol{x}) \\ &= \mathrm{P}[\boldsymbol{x}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + (\varepsilon_a - \varepsilon_b) > 0 \mid \boldsymbol{x}] \end{aligned} \tag{11.20}$$

定义 $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_a - \boldsymbol{\beta}_b$,$\varepsilon \equiv \varepsilon_a - \varepsilon_b$,则又回到与前面潜变量法相同的表达式,即方程(11.19)。如果 ε_a 与 ε_b 均为正态且相互独立,则 $(\varepsilon_a - \varepsilon_b)$ 也服从正态分布。只要将 $\mathrm{Var}(\varepsilon_a - \varepsilon_b)$ 标准化为 1,即得到 Probit 模型。另一方面,如果 ε_a 与 ε_b 相互独立,且均服从从非对称的“Ⅰ型极值分布”(Type Ⅰ extreme value distribution)①,即累积分布函数为 $F(\varepsilon) = \exp\{-\mathrm{e}^{-\varepsilon}\}$,则 $(\varepsilon_a - \varepsilon_b)$ 服从逻辑分布,证明参见 Cameron and Trivedi (2005, p. 486)。随机效用法的优点是,它可以比较容易地推广到多值选择的情形。第 17 章将介绍面板数据的二值选择模型。

11.4 二值选择模型中的异方差问题

标准的 Probit 或 Logit 模型假设扰动项为同方差,并据此写出似然函数。对于这个同方差假设,可以进行似然比检验(LR)。对于 Probit 模型,同方差的原假设 H_0 为

$$\mathrm{P}(y_i=1 \mid \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) \tag{11.21}$$

其中,扰动项的标准差 $\sigma=1$。而“异方差”的替代假设 H_1 为

$$\mathrm{P}(y_i=1 \mid \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma_i) \tag{11.22}$$

其中,$\sigma_i^2 \equiv \mathrm{Var}(\varepsilon_i)$。进一步,假设 σ_i^2 依赖于外生变量 $\boldsymbol{z} \equiv (z_1, \cdots, z_m)$

$$\sigma_i^2 = \exp(\boldsymbol{z}_i'\boldsymbol{\delta}) \tag{11.23}$$

其中,$\boldsymbol{z}$ 可以与解释变量 $\boldsymbol{x}$ 有重叠部分,或者包括 $\boldsymbol{x}$,但不包括常数项②。对方程(11.23)两边取对数可得

$$\ln \sigma_i^2 = \boldsymbol{z}_i'\boldsymbol{\delta} \tag{11.24}$$

在 Stata 中,$\ln \sigma_i^2$ 被称为“lnsigma2”。在异方差的替代假设下,同样可写出似然函数,同时估计原方程(11.22)与条件方差方程(11.24)。

在异方差情况下进行 Probit 估计的 Stata 命令为

```
hetprob y x1 x2 x3,het(varlist)
```

其中,选择项“het(varlist)”指定对扰动项方差有影响的所有变量,即 $\boldsymbol{z} \equiv (z_1, \cdots, z_m)$。在 Stata 的输出结果中,将汇报对“$H_0: \boldsymbol{\delta} = \boldsymbol{0}$”进行似然比检验(LR)的结果,即检验条件方差方程(11.24)的联合显著性。如果接受“$H_0: \boldsymbol{\delta} = \boldsymbol{0}$”,则可以使用同方差的 Probit 模型;否则,应该使用异方差的 Probit 模型。

---

① 考虑从某总体随机抽取 n 个观测数据 $\{x_1, \cdots, x_n\}$,则其最大值 $\max\{x_1, \cdots, x_n\}$ 可视为一个顺序统计量(order statistic)。当 $n \to \infty$ 时,$\max\{x_1, \cdots, x_n\}$ 的渐近分布就是Ⅰ型极值分布,故名。该分布也被称为“对数威布尔分布”(log Weibull distribution),其密度函数的形状与“对数正态分布”(log normal distribution)相似。

② 如果包括常数项,将使得各扰动项的方差同比例变化,故无法识别此常数项。

继续以数据集 womenwk. dta 为例：

```
. hetprob work age married children education,het(age married children education) nolog
```

```
Heteroskedastic probit model                      Number of obs     =       2000
                                                  Zero outcomes     =        657
                                                  Nonzero outcomes  =       1343

                                                  Wald chi2(4)      =      12.80
Log likelihood =  -1026.19                        Prob > chi2       =     0.0123
```

work	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
work						
age	.0356192	.0132539	2.69	0.007	.009642	.0615964
married	.5068326	.1629512	3.11	0.002	.1874542	.826211
children	.5013477	.1510223	3.32	0.001	.2053494	.797346
education	.0766883	.0355709	2.16	0.031	.0069706	.146406
_cons	-2.799595	.9868774	-2.84	0.005	-4.733839	-.8653511
lnsigma2						
age	-.0056187	.006938	-0.81	0.418	-.0192169	.0079795
married	.0761414	.121846	0.62	0.532	-.1626724	.3149551
children	.0035041	.041605	0.08	0.933	-.0780401	.0850483
education	.0199841	.0217614	0.92	0.358	-.0226675	.0626356

```
Likelihood-ratio test of lnsigma2=0: chi2(4) =     1.74   Prob > chi2 = 0.7828
```

上表的中间部分为对原方程(work)的估计结果，而下部为对方差方程(lnsigma2)的估计结果。上表最后一行显示，似然比检验的 p 值为 0.78，故可以接受“同方差”的原假设。

11.5 稀有事件偏差(选读)

对于二值选择模型，有时“$y=1$”发生的频率非常小，称为“稀有事件”(rare events)。比如，战争、政变、革命、流行病、经济危机、百年一遇的灾害等，其发生几率一般很小。此时，在二值选择模型的被解释变量中，可看到大量的 0，却只有很少的 1。

虽然使用 MLE(比如 Probit 或 Logit)来估计二值选择模型是一致的，但在有限样本下(样本容量小于 200)，Probit 或 Logit 估计依然存在偏差。而且，如果存在稀有事件，则该偏差将进一步放大；导致即使样本容量达到数千，而偏差依然存在，称为“稀有事件偏差”(rare event bias)。

为了直观地理解稀有事件偏差，考虑只有一个解释变量 x 而被解释变量为 y 的 Logit 模型：

$$\mathrm{P}(y=1\mid x)=\Lambda(\beta_0+\beta_1 x)\equiv\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \tag{11.25}$$

假设 $\beta_1>0$，则 x 与 y 正相关，即 x 越大，则“$y=1$”的概率越大。由此可知，给定“$y=1$”情况下 x 的条件分布位于给定“$y=0$”情况下 x 的条件分布的右侧，参见图 11.3。

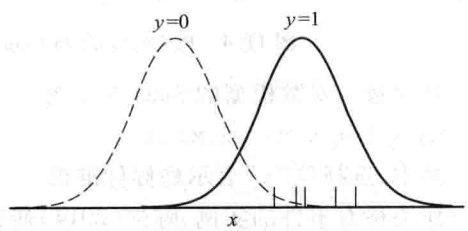


图 11.3 稀有事件偏差示意图

在图 11.3 中，将观测数据按 x 的取值从小到大排列，虚线表示条件分布 $x\mid y=0$ 的密度函数，而实线表示条件分布 $x\mid y=1$ 的密度函数。由于样本中

有很多“$y=0$”的观测数据(未在图中显示),故很容易估计$x \mid y=0$ 的密度。然而,由于样本中“$y=1$”的观测数据很少(在图中以五条竖线表示),故很难估计$x \mid y=1$ 的密度。由于 x 与 y 正相关,且“$y=1$”为稀有事件,故绝大多数“$y=0$”的观测数据将分布在 x 轴的左边,而“$y=1$”的观测数据分布在 x 轴的右边,二者很少重叠。考虑寻找 x 的一个分界点(cutting point),能够最好地区分“$y=0$”与“$y=1$”的观测数据(即犯最少的错误,比如把“$y=1$”的点放在右边,而把“$y=0$”的点放在左边)。此分界点与 β_1 的 MLE 估计量有关,而且它很可能落在“$y=1$”观测数据中 x 最大者或次大者(图中最右边的两条竖线)的左边。由于$x \mid y=0$ 的右尾(right tail)可以得到很好的估计,而$x \mid y=1$ 的左尾(left tail)很难估计,故此分界点将被系统地高估(为了将更多的 0 置于其左边),导致“$y=1$”的概率被系统地低估。

解决稀有事件偏差的方法通常有两种。方法之一,继续使用 Logit 模型,但对由于稀有事件而造成的小样本偏差 bias($\hat{\boldsymbol{\beta}}$)进行估计,然后对原 Logit 模型的估计系数进行修正,以得到“偏差修正估计”(bias-corrected estimates),即[$\hat{\boldsymbol{\beta}}$ - bias($\hat{\boldsymbol{\beta}}$)]。与此同时,估计量的标准误差也得到改善。此方法由 King and Zeng (2001a, 2001b) 提出,并可在作者 Gary King 的网页(http://gking. harvard. edu/relogit)下载 Stata 命令“relogit,norobust”来实现;其中,“re”表示“rare events”;选择项“nor”表示不使用稳健标准误(默认使用稳健标准误),安装方法参见下载文件中的“read me”文档。

方法之二,在方程(11.2)中,使用非对称的“极值分布”(extreme value distribution),则可以得到“补对数 - 对数模型”(complementary log-log model),其事件发生概率为

$$p = \mathrm{P}(y=1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = 1 - \exp\{-\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}\} \tag{11.26}$$

根据方程(11.26)可写出似然函数,然后进行 MLE 估计。之所以称为“补对数 - 对数模型”,是因为在上式中,$\boldsymbol{x}'\boldsymbol{\beta} = \ln[-\ln(1-p)]$,即如果取发生概率 p 的补数(complement,即 $1-p$),再取两次对数(其中一次须加负号,因为 $\log(1-p) < 0$),则得到 $\boldsymbol{x}'\boldsymbol{\beta}$。由于正态分布与逻辑分布都关于原点对称,故在 Probit 与 Logit 模型中,事件发生概率 p 趋于 1 的速度与趋于 0 的速度相等。另一方面,由于极值分布左偏(left-skewed),故在补对数 - 对数模型中,事件发生概率 p 趋于 1 的速度快于趋于 0 的速度;此性质正好对应于稀有事件的情形,参见图 11.4。

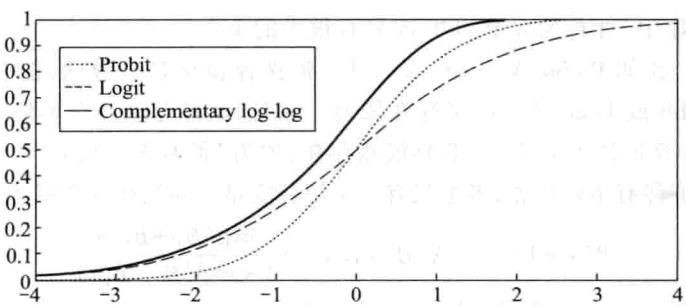


图 11.4 Probit, Logit 与 Complementary log-log 模型的累积分布函数比较

补对数 - 对数模型的 Stata 命令为

```
cloglog y x1 x2 x3,r
```

其中,选择项“r”表示稳健标准误。

作为稀有事件的实例,陈强(2014)研究了中原王朝被游牧民族征服的决定因素。以每十年作为观测单位建立时间序列数据,从公元前 221 年秦朝建立至 1911 年清朝灭亡,共有 213 个观

测值;而中原王朝被征服(conquered)仅发生 7 次。主要解释变量包括[①]:中原王朝早于游牧政权建立的年数(diff),中原王朝的绝对年龄(age),中原是否在长城的有效保护之下(wall),中国北方在十年中发生旱灾的年数比例的一阶滞后(drought1)。

```
. use nomadic_conquest.dta,clear
. sum conquered diff age wall drought1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
conquered	213	.0328638	.1787001	0	1
diff	213	15.99061	55.10522	-116	248
age	213	9.873239	7.502697	1	29
wall	213	.6619718	.4741525	0	1
drought1	212	.440566	.2937834	0	1

由上表可知,"conquered = 1"的发生频率为 3.29%,或可视为稀有事件。首先,进行普通的 Logit 回归。

```
. logit conquered diff age wall drought1,r nolog
```

```
Logistic regression                                     Number of obs   =        212
                                                        Wald chi2(4)    =      33.12
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -21.437566                       Pseudo R2       =     0.3030
```

conquered	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
diff	.0386435	.0091584	4.22	0.000	.0206934	.0565936
age	-.2634332	.0962616	-2.74	0.006	-.4521025	-.074764
wall	-2.486576	1.28163	-1.94	0.052	-4.998525	.0253721
drought1	3.915505	1.37519	2.85	0.004	1.220183	6.610827
_cons	-3.361682	1.038785	-3.24	0.001	-5.397663	-1.325702

上表显示,中原王朝越早于游牧政权建立(根据王朝周期假说,则中原王朝相对较弱)、干旱越严重(游牧民族为了生存而进攻农耕汉族),则中原王朝被征服的概率越高。另外,在控制相对年龄(diff)的情况下,中原王朝的绝对年龄(age)越长,被征服的概率反而降低。长城的保护作用也几乎在 5% 水平上显著(p 值为 5.2%)。为了便于比较不同模型,下面计算平均边际效应。

```
. margins,dydx(*)
```

```
Average marginal effects                                Number of obs   =        212
Model VCE    : Robust

Expression   : Pr(conquered), predict()
dy/dx w.r.t. : diff age wall drought1
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
diff	.0010269	.0004282	2.40	0.016	.0001877	.0018661
age	-.0070005	.0040678	-1.72	0.085	-.0149732	.0009722
wall	-.0660784	.0330503	-2.00	0.046	-.1308557	-.001301
drought1	.1040508	.0459542	2.26	0.024	.0139821	.1941194

① 原文还有其他控制变量,在此从略。

下面使用 King and Zeng (2001a,b)的方法对 Logit 模型进行偏差修正。

```
. relogit conquered diff age wall drought1
```

```
Corrected logit estimates                                   Number of obs =       212

------------------------------------------------------------------------------
  conquered |      Coef.   Robust Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
       diff |   .0311768   .0089471     3.48   0.000     .0136407    .0487129
        age |  -.2010725   .0940423    -2.14   0.033    -.3853921   -.0167529
       wall |  -2.137077   1.252087    -1.71   0.088    -4.591123    .3169691
   drought1 |   3.432347   1.343493     2.55   0.011     .7991486    6.065545
      _cons |  -3.041812   1.014843    -3.00   0.003    -5.030869   -1.052755
------------------------------------------------------------------------------
```

从上表可以看出,进行偏差修正后,Logit 模型的估计系数有些变化,而标准误差均有所下降,故变量的显著性基本没变。命令“relogit”并不提供对于边际效应的计算。下面估计补对数 - 对数模型。

```
. cloglog conquered diff age wall drought1,r nolog
```

```
Complementary log-log regression                Number of obs     =        212
                                                Zero outcomes     =        205
                                                Nonzero outcomes  =          7

                                                Wald chi2(4)      =      43.06
Log pseudolikelihood = -21.549974               Prob > chi2       =     0.0000

------------------------------------------------------------------------------
  conquered |      Coef.   Robust Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
       diff |   .0342854   .0080802     4.24   0.000     .0184485    .0501222
        age |  -.2257135   .0984111    -2.29   0.022    -.4185957   -.0328314
       wall |   -2.28261   1.136038    -2.01   0.045    -4.509203    -.056016
   drought1 |   3.508806   1.314683     2.67   0.008     .9320757    6.085537
      _cons |  -3.412064   1.069532    -3.19   0.001    -5.508308   -1.315821
------------------------------------------------------------------------------
```

从上表可知,使用补对数 - 对数也并不改变系数的显著性。

```
. margins,dydx(*)
```

```
Average marginal effects                        Number of obs     =        212
Model VCE    : Robust

Expression   : Pr(conquered), predict()
dy/dx w.r.t. : diff age wall drought1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |   .0009932    .000422     2.35   0.019     .0001661    .0018204
         age |  -.0065388   .0040806    -1.60   0.109    -.0145366    .0014591
        wall |  -.0661257   .0342384    -1.93   0.053    -.1332318    .0009804
    drought1 |   .1016478   .0478811     2.12   0.034     .0078025    .1954931
------------------------------------------------------------------------------
```

从上表可知,使用补对数 - 对数模型估计出来的边际效应与 Logit 模型十分接近。这说明,在此例中,发生频率为 3.29% 的事件可能还不够稀有①,故稀有事件偏差并不明显。

① 在 King and Zeng (2001a)研究第二次世界大战以来国际关系的样本中,国与国之间发生战争的比例仅占 0.3%,绝大多数国家之间相安无事。

11.6 含内生变量的 Probit 模型(选读)

对于二值选择模型,有时会遇到解释变量为内生变量的情形。比如,是否买保险取决于收入,但收入或许为内生变量,因为可能存在遗漏变量同时影响收入与买保险的决定。此时,由于扰动项与内生解释变量相关,使用通常的 Probit 或 Logit 模型将得不到一致估计。为此,考虑以下模型,

$$y_{1i}^{*} = \boldsymbol{x}_i'\boldsymbol{\alpha} + \beta y_{2i} + u_i \tag{11.27}$$

$$y_{2i} = \boldsymbol{x}_i'\boldsymbol{\gamma}_1 + \boldsymbol{z}_i'\boldsymbol{\gamma}_2 + v_i \tag{11.28}$$

$$y_{1i} = \mathbf{1}(y_{1i}^{*} > 0) \tag{11.29}$$

其中,y_{1i}为可观测的虚拟变量,y_{1i}^{*}为不可观测的潜变量,y_{2i}是模型中唯一的内生解释变量(以下方法可推广到多个内生解释变量的情形)。方程(11.27)称为"结构方程"(该方程右边含内生变量),而方程(11.28)称为"第一阶段方程"(first-stage equation)或"简化式方程"(reduced-form equation)(该方程右边不含内生变量)。

假设扰动项(u_i, v_i)服从期望值为 0 的二维正态分布,即

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_v \\ \rho\sigma_v & \sigma_v^2 \end{pmatrix}\right] \tag{11.30}$$

其中,u_i 的方差被标准化为 1,而 ρ 为(u_i, v_i)的相关系数。显然,由于 v_i 服从正态分布,故 y_{2i}也服从正态分布,因此 y_{2i}必须为连续变量。进一步,假设(u_i, v_i)独立于$\boldsymbol{x}_i$ 与$\boldsymbol{z}_i$,故在方程(11.27)中,$\boldsymbol{x}_i$ 为外生解释变量。而且,$\boldsymbol{z}_i$ 可作为方程(11.27)中内生变量 y_{2i}的工具变量,因为$\boldsymbol{z}_i$ 与内生变量 y_{2i}相关(参见方程(11.28)),且$\boldsymbol{z}_i$ 与 $\boldsymbol{u}_i$ 无关。在此模型中,y_{2i}的内生性完全来自于 u_i 与 v_i 的相关性;如果二者的相关系数 $\rho = 0$,则 y_{2i}为外生变量。因此,对于 y_{2i}内生性的检验可通过检验"$H_0: \rho = 0$"来进行。

由方程(11.27)—(11.30)所构成的模型,在给定$\boldsymbol{x}_i$ 与 $\boldsymbol{z}_i$ 的情况下,(y_{1i}, y_{2i})的条件概率分布已完全确定(fully specified)。将联合概率密度 $f(y_{1i}, y_{2i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$ 分解为 $f(y_{1i} \mid y_{2i}, \boldsymbol{x}_i, \boldsymbol{z}_i)$ $f(y_{2i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$,可写出样本数据$(y_{1i}, y_{2i})$的似然函数,然后进行最有效率的 MLE 估计。此法称为"工具变量 Probit"(Instrumental Variable Probit,简记 IV Probit)。

尽管 MLE 最有效率,但在数值计算时,可能不易收敛,特别在多个内生解释变量的情形下。为此,Newey (1987)与 Rivers and Vuong (1988)提出"两步法"(two-step method),其基本思想如下。在方程(11.27)中,既然 y_{2i}的内生性是由于遗漏了变量 v_i 所造成,那么如果能把 v_i 作为控制变量加入方程(11.27)即可得到一致估计。显然,方程(11.28)的扰动项 v_i 不可观测,但可用 OLS 残差 $\hat{v}_i$ 作为 v_i 的一致估计。由于 $\hat{v}_i$ 在两步法中被作为控制变量使用,故两步法也被称为"控制函数法"(control function approach)。

由于(u_i, v_i)服从二维正态分布,故根据多元统计知识,u_i 对于 v_i 的总体回归方程(population regression equation)可写为

$$u_i = \delta v_i + \varepsilon_i \tag{11.31}$$

其中,$\delta \equiv \dfrac{\mathrm{Cov}(u_i, v_i)}{\mathrm{Var}(v_i)}$,而扰动项 ε_i 独立于 v_i(故也独立于 y_{2i}),也独立于$\boldsymbol{x}_{1i}$。而且,ε_i 也服从正态

分布，期望为$E(\varepsilon_i)=0$，而方差为

$$\begin{aligned}\mathrm{Var}(\varepsilon_i)&=\mathrm{Var}(u_i-\delta v_i)=1-\frac{\mathrm{Cov}^2(u_i,v_i)}{\mathrm{Var}^2(v_i)}\cdot\mathrm{Var}(v_i)\\&=1-\frac{\mathrm{Cov}^2(u_i,v_i)}{\mathrm{Var}(v_i)}=1-\rho^2\end{aligned}\tag{11.32}$$

其中，$\mathrm{Var}(u_i)=1$。将方程(11.31)代入方程(11.27)可得

$$y_{1i}^{*}=\boldsymbol{x}_i'\boldsymbol{\alpha}+\beta y_{2i}+\delta v_i+\varepsilon_i\tag{11.33}$$

其中，$\varepsilon_i\sim N(0,1-\rho^2)$，而且独立于$\boldsymbol{x}_i$，$y_{2i}$与$v_i$。为了把$\varepsilon_i$的方差标准化为1，将方程(11.33)两边同除以$\sqrt{1-\rho^2}$可得

$$\frac{y_{1i}^{*}}{\sqrt{1-\rho^2}}=\boldsymbol{x}_i'\frac{\boldsymbol{\alpha}}{\sqrt{1-\rho^2}}+\frac{\beta}{\sqrt{1-\rho^2}}y_{2i}+\frac{\delta}{\sqrt{1-\rho^2}}v_i+\frac{\varepsilon_i}{\sqrt{1-\rho^2}}\tag{11.34}$$

由于方程(11.34)中的v_i不可观测，故两步法由以下两步构成。

第一步：对简化式方程(11.28)进行OLS回归，得到残差$\hat{v}_i$。

第二步：以残差$\hat{v}_i$替代方程(11.34)中的v_i，进行Probit估计，得到对变换后系数$\left(\frac{\boldsymbol{\alpha}}{\sqrt{1-\rho^2}},\frac{\beta}{\sqrt{1-\rho^2}},\frac{\delta}{\sqrt{1-\rho^2}}\right)$的估计。由此可见，两步法估计系数与MLE估计系数并不直接可比，二者相差$\sqrt{1-\rho^2}$倍。由于$0\leqslant\sqrt{1-\rho^2}\leqslant 1$，故两步法估计系数的绝对值一般比MLE估计系数的绝对值更大。在使用两步法的情况下，对y_{2i}内生性的检验可通过检验原假设"$H_0:\delta=0$"来进行，因为如果$\delta=0$，则u_i与v_i不相关(参见方程(11.31))。使用两步法时，由于第一步的误差被带入第二步中，故两步法不如MLE更有效率；两步法的优势主要在计算方便上。另一估计方法是，直接使用线性概率模型(LPM)，然后用2SLS或GMM进行估计，但这样做将无视y_{1i}为虚拟变量的特征，只具有参考价值。

含内生变量Probit模型的Stata命令为

```
. ivprobit y1 x1 x2 (y2 = z1 z2),r
. ivprobit y1 x1 x2 (y2 = z1 z2),first twostep
```

其中，"y1"为被解释变量，"y2"为内生解释变量，"x1 x2"为外生解释变量，而"z1 z2"为工具变量；选择项"r"表示稳健标准误。选择项"twostep"表示使用两步法，默认进行MLE估计；选择项"first"表示显示第一步回归的结果。

下面以数据集mus14data.dta为例[①]。该数据集包括2002年3 206名美国老年人。被解释变量为ins(是否购买私人医疗保险)，解释变量包括个人健康及社会经济变量：linc(家庭收入的对数)，hstatusg(自我评估的健康状况虚拟变量)，adl(日常生活中受限活动数目，number of limitations on activities of daily living)，chronic(慢性病数目)，age(年龄)，age2(年龄平方)，female(是否女性)，educyear(教育年限)，married(是否结婚)，hisp(是否拉丁裔)，white(是否白人)。

首先进行一般的Probit回归。

```
. probit ins linc female age age2 educyear married hisp white chronic
adl hstatusg,r nolog
```

① 此例来自Cameron and Trivedi (2009)。

```
Probit regression                                 Number of obs   =       3197
                                                  Wald chi2(11)   =     366.94
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1933.4275                 Pseudo R2       =     0.0946
```

ins	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	.3466893	.0402173	8.62	0.000	.2678648	.4255137
female	-.0815374	.0508549	-1.60	0.109	-.1812112	.0181364
age	.1162879	.1151924	1.01	0.313	-.109485	.3420608
age2	-.0009395	.0008568	-1.10	0.273	-.0026187	.0007397
educyear	.0464387	.0089917	5.16	0.000	.0288153	.0640622
married	.1044152	.0636879	1.64	0.101	-.0204108	.2292412
hisp	-.3977334	.1080935	-3.68	0.000	-.6095927	-.1858741
white	-.0418296	.0644391	-0.65	0.516	-.168128	.0844687
chronic	.0472903	.0186231	2.54	0.011	.0107897	.0837909
adl	-.0945039	.0353534	-2.67	0.008	-.1637953	-.0252125
hstatusg	.1138708	.0629071	1.81	0.070	-.0094248	.2371664
_cons	-5.744548	3.871615	-1.48	0.138	-13.33277	1.843677

上表显示,linc 的系数为 0.35,且在 1% 水平上显著,即家庭收入越高,越可能购买私人健康保险。但怀疑 linc 为内生变量,因为可能存在同时影响 ins 与 linc 的遗漏变量。为此,使用数据集中另一变量 retire(本人是否退休)作为工具变量。是否退休显然与收入 linc 相关,满足工具变量的相关性;另一方面,假设是否退休不直接影响购买私人健康保险的决策,故满足工具变量的外生性。下面使用此工具变量进行 IV Probit 估计。

```
. ivprobit ins female age age2 educyear married hisp white chronic adl
hstatusg (linc = retire),r nolog
```

```
Probit model with endogenous regressors           Number of obs   =       3197
                                                  Wald chi2(11)   =     417.68
Log pseudolikelihood = -5408.3856                 Prob > chi2     =     0.0000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	-.6199245	.358139	-1.73	0.083	-1.321864	.0820151
female	-.1424789	.0476242	-2.99	0.003	-.2358205	-.0491373
age	.298707	.1237695	2.41	0.016	.0561234	.5412907
age2	-.002233	.0009006	-2.48	0.013	-.0039981	-.000468
educyear	.1185139	.0208859	5.67	0.000	.0775783	.1594495
married	.7571106	.2162909	3.50	0.000	.3331881	1.181033
hisp	-.5103837	.1032417	-4.94	0.000	-.7127337	-.3080337
white	.174971	.097886	1.79	0.074	-.016882	.3668241
chronic	.0014631	.026898	0.05	0.957	-.0512561	.0541823
adl	-.1362883	.03404	-4.00	0.000	-.2030055	-.069571
hstatusg	.2419553	.066904	3.62	0.000	.1108258	.3730847
_cons	-10.25689	3.977975	-2.58	0.010	-18.05358	-2.460204
/athrho	.7576717	.3626733	2.09	0.037	.0468452	1.468498
/lnsigma	-.3313108	.0233359	-14.20	0.000	-.3770484	-.2855732
rho	.6397035	.2142599			.0468109	.8992906
sigma	.717982	.0167548			.6858829	.7515833

```
Instrumented:  linc
Instruments:   female age age2 educyear married hisp white chronic adl
               hstatusg retire
```

```
Wald test of exogeneity (/athrho = 0): chi2(1) =     4.36 Prob > chi2 = 0.0367
```

上表底部提供了对外生性原假设“H_0：$\rho=0$”的沃尔德检验结果，其 p 值为 0.036 7，故可在 5% 的水平上认为 linc 为内生变量。根据模型估计结果，u_i 与 v_i 的相关系数为 $\hat{\rho}=0.64$，这表明，未度量的遗漏变量在增加家庭收入的同时，也会提高购买保险的倾向。相关系数高达 0.64 也意味着 ivprobit 的估计结果会与 probit 大不相同。事实上，使用 ivprobit 时，linc 的估计系数为 −0.62，不仅改变了符号，而且在 10% 水平上显著（p 值为 0.083）。这表明，如果使用一般的 Probit 模型进行估计，由于忽略了 linc 的内生性，将高估家庭收入对于保险购买倾向的正作用。

下面进行两步法估计。

```
. ivprobit ins female age age2 educyear married hisp white chronic adl
hstatusg (linc = retire),first twostep
```

```
Checking reduced-form model...
First-stage regression
```

Source	SS	df	MS	Number of obs =	3197
				F(11, 3185) =	205.93
Model	1172.11143	11	106.555584	Prob > F =	0.0000
Residual	1648.04737	3185	.517440303	R-squared =	0.4156
				Adj R-squared =	0.4136
Total	2820.15879	3196	.882402626	Root MSE =	.71933

linc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
retire	-.0955818	.0286254	-3.34	0.001	-.1517078	-.0394557
female	-.1065852	.0282552	-3.77	0.000	-.1619853	-.051185
age	.262005	.0626898	4.18	0.000	.1390887	.3849214
age2	-.0018737	.0004643	-4.04	0.000	-.002784	-.0009635
educyear	.0945358	.00435	21.73	0.000	.0860068	.1030649
married	.7647985	.0312142	24.50	0.000	.7035966	.8260004
hisp	-.2340063	.0523453	-4.47	0.000	-.3366401	-.1313725
white	.2329204	.0347781	6.70	0.000	.1647307	.30111
chronic	-.0386546	.0100859	-3.83	0.000	-.0584301	-.0188791
adl	-.0741595	.0173479	-4.27	0.000	-.1081738	-.0401453
hstatusg	.1739715	.0338516	5.14	0.000	.1075984	.2403446
_cons	-7.505571	2.114278	-3.55	0.000	-11.65106	-3.360086

```
Two-step probit with endogenous regressors          Number of obs   =      3197
                                                    Wald chi2(11)   =    207.35
                                                    Prob > chi2     =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
linc	-.8064908	.6526256	-1.24	0.217	-2.085613	.4726318
female	-.1853647	.084243	-2.20	0.028	-.3504779	-.0202515
age	.3886148	.2094068	1.86	0.063	-.0218149	.7990446
age2	-.0029052	.0015308	-1.90	0.058	-.0059056	.0000952
educyear	.1541854	.0617338	2.50	0.013	.0331895	.2751814
married	.9849852	.503158	1.96	0.050	-.0011864	1.971157
hisp	-.6640124	.1983783	-3.35	0.001	-1.052827	-.2751979
white	.2276309	.1709828	1.33	0.183	-.1074892	.562751
chronic	.0019054	.0339542	0.06	0.955	-.0646436	.0684544
adl	-.1773114	.0621765	-2.85	0.004	-.299175	-.0554478
hstatusg	.3147822	.1360959	2.31	0.021	.0480392	.5815251
_cons	-13.34419	6.427051	-2.08	0.038	-25.94097	-.7473972

```
Instrumented:  linc
Instruments:   female age age2 educyear married hisp white chronic adl
               hstatusg retire
Wald test of exogeneity:     chi2(1) =     4.37          Prob > chi2 = 0.0366
```

上表显示,两步法的估计系数与 MLE 类似,但前者的绝对值比后者约大 25%,这与 MLE 对 ρ 的估计相一致,因为 $\sqrt{1-\rho^2}=\sqrt{1-0.639\ 703\ 5^2}\approx 0.77$。上表底部对于 linc 外生性原假设的 Wald 检验结果表明,p 值为 0.036 6,故可在 5% 水平上认为 linc 为内生变量。另外,第一步回归的结果显示,工具变量 retire 对于内生变量 linc 具有较强的解释力。

11.7 双变量 Probit 模型(选读)

有时候需要对两个虚拟变量同时考虑其发生的概率。比如,被解释变量之一是病人是否去看医生,而被解释变量之二是病人是否住院。如果对这两个被解释变量分别进行 Probit 建模,则可能损失效率(但依然为一致估计),因为"看医生"与"住院"这两件事通常是相关的,即两个 Probit 方程的扰动项之间可能存在相关性。具体考察以下模型:

$$\begin{cases} y_1^* = \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* = \boldsymbol{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{cases} \tag{11.35}$$

其中,y_1^* 与 y_2^* 为不可观测的潜变量,扰动项 $(\varepsilon_1,\varepsilon_2)$ 服从二维联合正态分布,期望为 0,方差为 1,而相关系数为 ρ,即

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right\} \tag{11.36}$$

可观测变量 y_1 与 y_2 由以下方程决定:

$$y_1 = \begin{cases} 1 \text{ 若 } y_1^* > 0 \\ 0 \text{ 若 } y_1^* \leqslant 0 \end{cases} \tag{11.37}$$

$$y_2 = \begin{cases} 1 \text{ 若 } y_2^* > 0 \\ 0 \text{ 若 } y_2^* \leqslant 0 \end{cases} \tag{11.38}$$

如果 $\boldsymbol{x}_1=\boldsymbol{x}_2$(两个方程的解释变量完全相同),则方程(11.35)—(11.38)称为"双变量 Probit";反之,如果 $\boldsymbol{x}_1\neq\boldsymbol{x}_2$(两个方程的解释变量不完全相同),则称为"似不相关双变量 Probit"(seemingly unrelated bivariate probit),因为这两个方程的唯一联系是扰动项的相关性[①]。显然,如果 $\rho=0$,则此模型等价于两个单独的 Probit 模型。当 $\rho\neq 0$ 时,可写下 (y_1,y_2) 的取值概率,然后进行最大似然估计。比如:

$$\begin{aligned} p_{11} &\equiv \mathrm{P}(y_1=1,y_2=1) = \mathrm{P}(y_1^*>0,y_2^*>0) \\ &= \mathrm{P}(\varepsilon_1 > -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \varepsilon_2 > \boldsymbol{x}_2'\boldsymbol{\beta}_2) \\ &= \mathrm{P}(\varepsilon_1 < \boldsymbol{x}_1'\boldsymbol{\beta}_1, \varepsilon_2 < \boldsymbol{x}_2'\boldsymbol{\beta}_2) \\ &= \int_{-\infty}^{\boldsymbol{x}_1'\boldsymbol{\beta}_1}\int_{-\infty}^{\boldsymbol{x}_2'\boldsymbol{\beta}_2} \phi(z_1,z_2,\rho)\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \Phi(\boldsymbol{x}_1'\boldsymbol{\beta}_1, \boldsymbol{x}_2'\boldsymbol{\beta}_2, \rho) \end{aligned} \tag{11.39}$$

其中,$\phi(z_1,z_2,\rho)$ 与 $\Phi(z_1,z_2,\rho)$ 分别为标准化的二维正态分布(standardized bivariate normal)的概率密度函数与累积分布函数,期望为 0,方差为 1,而相关系数为 ρ。类似地,可计算 p_{10},p_{01},

① 有关"似不相关回归"(seemingly unrelated regression),参见第 23 章。

p_{00}，将这些概率取对数后加总，即得到对数似然函数。最后，对原假设“$H_0: \rho = 0$”进行检验，可判断有无必要使用双变量 Probit 模型，或估计两个单独的 Probit 模型。在理论上，双变量 Probit 模型可推广到多变量 Probit(multivariate probit)，但实际操作上较难计算高维积分，故 Stata 尚未提供多变量 Probit 的命令。

双变量 Probit 的 Stata 命令为

```
. biprobit y1 y2 x1 x2 x3,r                    (解释变量完全相同的情形)
. biprobit (y1 = x1 x2) (y2 = x1 x3),r         (解释变量不完全相同的情形)
```

其中，选择项“r”表示使用稳健标准误。

以 Pindyck and Rubin (1998)所提供的数据集 school. dta 为例。该数据集包括了 1973 年美国密西根州特洛伊市(Troy, Michigan)增加房产税以支持公立学校议案的 95 名投票人信息。主要变量包括：vote(是否投票赞成)，private(是否有孩子在私立学校上学)，logptax(房产税对数)，loginc(家庭年收入对数)，school(投票人是否教师)，years(在特洛伊市居住的年数)。假设同时关心是否送孩子上私立学校以及是否投票赞成，而这两件事可能相关(如果孩子上私立学校，则可能不愿增加房产税以支持公立学校)，故使用双变量 Probit 模型进行估计。

```
. biprobit (private = loginc logptax school) (vote = loginc logptax
years school),r nolog
```

```
Seemingly unrelated bivariate probit              Number of obs   =         95
                                                  Wald chi2(7)    =     611.22
Log pseudolikelihood = -84.695294                 Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
private      |
      loginc |   .3911055   .5826198     0.67   0.502    -.7508083    1.533019
     logptax |  -.1877972    .758957    -0.25   0.805    -1.675326    1.299731
      school |  -4.889164   .2553132   -19.15   0.000    -5.389568   -4.388759
       _cons |  -3.795514   6.355104    -0.60   0.550    -16.25129    8.660261
-------------+----------------------------------------------------------------
vote         |
      loginc |   1.262054   .4367689     2.89   0.004     .4060021    2.118105
     logptax |  -1.254895    .535599    -2.34   0.019    -2.30465   -.2051403
       years |  -.0138503   .0177383    -0.78   0.435    -.0486168    .0209162
      school |   1.724164   .4377039     3.94   0.000     .8662806    2.582048
       _cons |  -3.543213   4.549227    -0.78   0.436    -12.45953    5.373108
-------------+----------------------------------------------------------------
     /athrho |  -.193618    .2324816    -0.83   0.405    -.6492736    .2620376
-------------+----------------------------------------------------------------
         rho |  -.1912343   .2239796                    -.5711808    .2562004
------------------------------------------------------------------------------
Wald test of rho=0:                chi2(1) =  .693608   Prob > chi2 = 0.4049
```

上表显示，送孩子上私立学校(private)与教师身份(school)显著负相关；而投票赞成(vote)与家庭收入(loginc)、教师身份(school)显著正相关，与房地产税(logptax)显著负相关，而与居住年数(years)无关。上表最后一行对原假设“$H_0: \rho = 0$”的沃尔德检验显示，p 值高达 0.40，故可认为 $\rho = 0$，没必要使用双变量 Probit 模型。

为了比较，下面对变量 vote 进行单独的 Probit 估计。

```
. probit vote loginc logptax years school,r nolog
```

```
Probit regression                                 Number of obs   =         95
                                                  Wald chi2(4)    =      22.07
                                                  Prob > chi2     =     0.0002
Log pseudolikelihood = -54.693088                 Pseudo R2       =     0.1324
```

vote	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
loginc	1.260286	.4369633	2.88	0.004	.403854	2.116719
logptax	-1.23849	.5281871	-2.34	0.019	-2.273717	-.2032619
years	-.0130624	.0177873	-0.73	0.463	-.0479248	.0217999
school	1.72503	.4363312	3.95	0.000	.8698368	2.580224
_cons	-3.646993	4.479862	-0.81	0.416	-12.42736	5.133376

从上表可知,单独的 Probit 估计结果(系数与标准误)与双变量 Probit 很接近,这也印证了没必要使用双变量 Probit。

11.8 部分可观测的双变量 Probit 模型(选读)

对于双变量 Probit 模型,有时无法同时观测到 y_1 与 y_2,而只能看到 y_1 与 y_2 都取值为 1 的情形。比如,估计实习生在某公司实习后留任的概率。显然,"留任"既取决于该公司是否向该实习生发聘书(公司的二值选择),也取决于该实习生是否愿意在此公司工作(实习生的二值选择)。而我们只能观测到该实习生是否留任了。如果留任,则 y_1 与 y_2 都取值为 1;如果未留任,则只知道 y_1 与 y_2 至少有一个为 0。在此情况下,公司的二值选择与实习生的二值选择可能相关,故适用于双变量 Probit 模型的框架。但对于 y_1 与 y_2,只有其乘积 y_1y_2 可观测,要么 $y_1y_2=1$,要么 $y_1y_2=0$;称为"部分可观测的双变量 Probit"(bivariate probit with partial observability),最早由 Poirier (1980)提出。作为另一例子,由两名委员组成的委员会匿名对一个议案投票,而该议案只有全票通过才生效。对于外人来说,只能看到该议案通过与否;如果没有通过,则无法知道谁投了否决票。

定义 $z\equiv y_1y_2$,可计算 z 的取值概率,并写出似然函数进行 MLE 估计。

$$P(z=1)=P(y_1=1,y_2=1)=\Phi(\boldsymbol{x}_1'\boldsymbol{\beta}_1,\boldsymbol{x}_2'\boldsymbol{\beta}_2,\rho) \tag{11.40}$$

$$P(z=0)=1-P(y_1=1,y_2=1)=1-\Phi(\boldsymbol{x}_1'\boldsymbol{\beta}_1,\boldsymbol{x}_2'\boldsymbol{\beta}_2,\rho) \tag{11.41}$$

部分可观测的双变量 Probit 模型的 Stata 命令为

```
. biprobit (y1 = x1 x2) (y2 = x1 x3), r partial difficult
```

其中,选择项"partial"表示估计部分可观测的双变量 Probit,但加上此选择项后可能使得数值计算不收敛。如果出现不收敛的情形,可使用选择项"difficult",它表示在最大化的目标函数存在非凹区域(nonconcave regions)的情况下,使用另一迭代方法。

在使用命令"biprobit,partial"进行估计时,只用到 y_1y_2 的信息,与 y_1,y_2 的分别取值无关。对于现实中的部分可观测的双变量 Probit 模型,其实只知道 y_1y_2,记 $z_1=y_1y_2$。为此,可先定义另一个被解释变量 $z_2=z_1$,然后再进行估计。显然,由于 $z_1=z_2$ 都是虚拟变量,故 $z_1z_2=z_1$。需要注意的是,我们无法令 $z_2\equiv 1$(尽管也满足 $z_1z_2=z_1$),因为 Stata 不允许 Probit 模型中出现所有取值都是 1 的被解释变量。具体操作命令如下:

```
. gen z2 = z1
. biprobit(z1 = x1 x2) (z2 = x1 x3),r partial difficult
```

为了演示的目的,对上节的 school. dta 数据集进行部分可观测的双变量 Probit 估计(假设只能观测到 vote 与 private 的乘积)。

```
. biprobit (private = loginc logptax school) (vote = loginc logptax years school),r partial difficult nolog
```

```
Partial observability bivariate probit          Number of obs     =        95
                                                Wald chi2(6)      =         .
Log pseudolikelihood = -14.126282               Prob > chi2       =         .
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
private						
loginc	.3416941	.9233022	0.37	0.711	-1.467945	2.151333
logptax	-.8713551	1.01297	-0.86	0.390	-2.85674	1.11403
school	-6.106526	.5756618	-10.61	0.000	-7.234802	-4.978249
_cons	1.587501	6.879705	0.23	0.818	-11.89647	15.07148
vote						
loginc	21.08716	1.287221	16.38	0.000	18.56425	23.61006
logptax	-30.21983	2.239179	-13.50	0.000	-34.60854	-25.83112
years	-14.56743	.8514695	-17.11	0.000	-16.23628	-12.89858
school	-4.1137	.7031372	-5.85	0.000	-5.491824	-2.735577
_cons	56.3888	11.57035	4.87	0.000	33.71134	79.06626
/athrho	-3.430044	.0173811	-197.34	0.000	-3.464111	-3.395978
rho	-.9979046	.0000728			-.9980424	-.997757

```
Wald test of rho=0:                     chi2(1) =  38944.3    Prob > chi2 = 0.0000
```

上表显示,部分可观测的双变量 Probit 估计结果与(全信息)双变量 Probit 在估计系数上有较大差别。对于前者而言,由于信息较少,故损失了效率。上表最后一行的 Wald 检验 p 值为 0.000 0,故强烈拒绝"$\rho=0$"的原假设(当然,这是基于信息不完全情况下作出的统计推断;根据上节的结果,如果同时知道 private 与 vote,则可接受"$\rho=0$")。

习 题

11.1 对于二值变量 y_i,考虑线性概率模型 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$,其中 $P(y_i=1\mid x_i)=\beta_0+\beta_1 x_i$。

(1) 证明 $E(\varepsilon_i\mid x_i)=0$。

(2) 求 $Var(\varepsilon_i\mid x_i)$的表达式。

(3) 扰动项 ε_i 是否存在异方差?为什么?

11.2 假定第 i 个家庭选择消费水平 c_i 与捐赠水平 q_i,来求解以下预算约束条件下的效用最大化问题:

$$\max_{c_i,q_i} U_i=c_i+a_i\ln(1+q_i)$$

$$s.t.\ c_i+p_i q_i\leqslant m_i, c_i, q_i\geqslant 0$$

其中,m_i 为收入,p_i 为捐赠的"价格"且 $p_i<1$(因为捐赠可以抵税),a_i 表示捐赠的边际效用。将 m_i 与 p_i 均视为外生变量。可以证明,如果 $a_i\leqslant p_i$,则最优捐赠 $q_i=0$;反之,如果 $a_i>p_i$,则最优捐赠 $q_i=(a_i/p_i)-1$。定义二值变量 $y_i=\begin{cases}1,若\ q_i>0\\0,若\ q_i=0\end{cases}$,假设家庭的捐赠偏好 a_i 取决于 $a_i=\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i)$,其中 $\boldsymbol{x}_i$ 为可观测的家庭特征,而

ε_i 为不可观测的扰动项。假定 ε_i 独立于 $(\boldsymbol{x}_i, m_i, p_i)$，且 $\varepsilon_i \sim N(0, \sigma^2)$。求 $P(y_i = 1 \mid \boldsymbol{x}_i, m_i, p_i)$ 的表达式。这是一个 Probit 模型吗？

11.3 数据集 hmda_short.dta① 包含了美国波士顿有关房贷申请的横截面数据，包括以下变量：虚拟变量 deny（房贷申请被拒），piratio（payment to income ratio，支出占收入比重），score（信用分数，1 最高，6 最低），black（是否为黑人）。考虑以下模型

$$\text{deny}_i = \beta_0 + \beta_1 \text{ piratio}_i + \beta_2 \text{ score}_i + \beta_3 \text{ black}_i + \varepsilon_i \tag{11.42}$$

（1）用 OLS 估计线性概率模型；

（2）估计 Probit 模型，计算边际效应及正确预测的比率；

（3）估计 Logit 模型，计算边际效应及正确预测的比率；

（4）估计条件异方差 Probit 模型，并确定是否存在条件异方差。

11.4 数据集 gss_cdvm.dta 来自 2000 年与 2002 年的 General Social Survey，包含以下变量：虚拟变量 trust（是否信任大多数人），www（是否用互联网），educate（教育年限），income（家庭收入），age（年龄），male（是否男性）。

（1）估计 Probit 模型，考察 educate，income，age，male 对 trust 的影响。

（2）估计 Probit 模型，考察 educate，income，age，male 对 www 的影响。

（3）估计双变量 Probit 模型，同时考察 trust 与 www 的决定。是否有必要使用双变量 Probit 模型？

① 此数据集来自 Stock and Watson (2011)。

第 12 章　多值选择模型

12.1　多项 Logit 与多项 Probit

个体面临的选择有时是多值的，而不仅仅是二值的。比如，交通工具的选择，职业的选择，手机消费品牌的选择。假设可供个体选择的方案为 $y=1,2,\cdots,J$，其中 J 为正整数，即共有 J 种互相排斥的选择。如果 $J=2$，即为二值选择模型（参见第 11 章）。

使用随机效用法，假设个体 i 选择方案 j 所能带来的随机效用为

$$U_{ij}=\boldsymbol{x}_i'\boldsymbol{\beta}_j+\varepsilon_{ij}\quad(i=1,\cdots,n;j=1,\cdots,J) \tag{12.1}$$

其中，解释变量 $\boldsymbol{x}_i$ 只随个体 i 而变，不随方案 j 而变。比如，个体的性别、年龄、收入等特征。这种解释变量被称为“只随个体而变”（case-specific）或“不随方案而变”（alternative-invariant）。系数 $\boldsymbol{\beta}_j$ 表明，$\boldsymbol{x}_i$ 对随机效用 U_{ij} 的作用取决于方案 j。

显然，个体 i 选择方案 j，当且仅当方案 j 带来的效用高于所有其他方案，故个体 i 选择方案 j 的概率可写为

$$\begin{aligned}\mathrm{P}(y_i=j\mid\boldsymbol{x}_i)&=\mathrm{P}(U_{ij}\geqslant U_{ik},\forall\,k\neq j)\\&=\mathrm{P}(U_{ik}-U_{ij}\leqslant 0,\forall\,k\neq j)\\&=\mathrm{P}(\varepsilon_{ik}-\varepsilon_{ij}\leqslant\boldsymbol{x}_i'\boldsymbol{\beta}_j-\boldsymbol{x}_i'\boldsymbol{\beta}_k,\forall\,k\neq j)\end{aligned} \tag{12.2}$$

假设 $\{\varepsilon_{ij}\}$ 为 iid 且服从 I 型极值分布（type I extreme value distribution），则可证明：

$$\mathrm{P}(y_i=j\mid\boldsymbol{x}_i)=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)} \tag{12.3}$$

显然，选择各项方案的概率之和为 1，即 $\sum_{j=1}^{J}\mathrm{P}(y_i=j\mid\boldsymbol{x}_i)=1$。方程（12.3）是二值选择 Logit 模型向多值选择模型的自然推广。需要注意的是，无法同时识别所有的系数 $\boldsymbol{\beta}_k,k=1,\cdots,J$。这是因为，如果将 $\boldsymbol{\beta}_k$ 变为 $\boldsymbol{\beta}_k^*=\boldsymbol{\beta}_k+\boldsymbol{\alpha}$（$\boldsymbol{\alpha}$ 为某常数向量），完全不会影响模型的拟合。为此，通常将某方案（比如，方案 1）作为“参照方案”（base category），然后令其相应系数 $\boldsymbol{\beta}_1=\mathbf{0}$。由此，个体 i 选择方案 j 的概率为

$$\mathrm{P}(y_i=j\mid\boldsymbol{x}_i)=\begin{cases}\dfrac{1}{1+\sum_{k=2}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)} & (j=1)\\[2ex]\dfrac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{1+\sum_{k=2}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)} & (j=2,\cdots,J)\end{cases} \tag{12.4}$$

其中，“$j=1$”所对应的方案为参照方案。此模型称为“多项 Logit”（Multinomial Logit），可用 MLE 进行估计。个体 i 的似然函数为

$$L_i(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_J)=\prod_{j=1}^{J}[\mathrm{P}(y_i=j\mid\boldsymbol{x}_i)]^{\mathbf{1}(y_i=j)} \tag{12.5}$$

其对数似然函数为

$$\ln L_i(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_J) = \sum_{j=1}^{J} \mathbf{1}(y_i=j)\cdot \ln \mathrm{P}(y_i=j\mid \boldsymbol{x}_i) \tag{12.6}$$

其中,$\mathbf{1}(\cdot)$为示性函数(indicator function),即如果括号中的表达式成立,则取值为1;反之,取值为0。将所有个体的对数似然函数加总,即得到整个样本的对数似然函数,将其最大化则得到系数估计值$\hat{\boldsymbol{\beta}}_1,\cdots,\hat{\boldsymbol{\beta}}_J$。

另外,在方程(12.2)中,如果假设$\{\varepsilon_{i1},\cdots,\varepsilon_{iJ}\}$服从$J$维正态分布,则可得到"多项 Probit"(Multinomial Probit)模型,但该模型的选择概率涉及高维积分,不易计算。在扰动项$\{\varepsilon_{i1},\cdots,\varepsilon_{iJ}\}$相互独立的假定下,Stata 首先把此高维积分转化为一维积分,然后进行数值计算(但此一维积分也没有解析解)。

12.2 条件 Logit 模型

多项 Logit 模型仅考虑不随方案而变的解释变量(比如,个体收入),但有些解释变量可能既随个体而变,也随方案而变。比如,在选择交通工具时,乘车时间既因个体而异(不同个体的行车路线可能不同),也因交通工具而异(坐公交的时间与乘出租车的时间不同)。这种解释变量称为"随方案而变"(alternative-specific),既包括同时随方案与个体而变的变量,也包括随方案而变但不随个体而变的变量(比如,选择加入某协会时,所有人交的会费都一样)。

此时,个体i选择方案j所带来的随机效用为

$$U_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \quad (i=1,\cdots,n;\ j=1,\cdots,J) \tag{12.7}$$

其中,解释变量$\boldsymbol{x}_{ij}$表明,它既随个体i而变,也随方案j而变。系数$\boldsymbol{\beta}$表明,$\boldsymbol{x}_{ij}$对随机效用U_{ij}的作用不依赖于方案j,比如乘车时间依个体与方案而变,但乘车时间太长所带来的负效用是一致的。根据与多项 Logit 类似的推导可知,个体i选择方案j的概率为

$$\mathrm{P}(y_i=j\mid \boldsymbol{x}_{ij}) = \frac{\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})}{\sum_{k=1}^{J}\exp(\boldsymbol{x}_{ik}'\boldsymbol{\beta})} \tag{12.8}$$

此模型称为"条件 Logit"(Conditional Logit),也称为"McFadden 选择模型"(McFadden's Choice Model)(McFadden,1974)①。条件 Logit 模型的估计方法与多项 Logit 模型类似,即通过 MLE 估计以得到系数估计值$\hat{\boldsymbol{\beta}}$。与多项 Logit 不同的是,在条件 Logit 模型中,由于系数$\boldsymbol{\beta}$不依赖于方案,故不需要选择参照方案,也不需要将$\boldsymbol{\beta}$的某部分标准化为0。

12.3 混合 Logit 模型

上面分别考虑了解释变量不随方案而变的多项 Logit 模型,以及解释变量随方案而变的条件 Logit 模型。很自然地,可考虑这两种情况同时发生的混合情形。假设个体i选择方案j所能带来的随机效用为

$$U_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\gamma}_j + \varepsilon_{ij} \quad (i=1,\cdots,n;j=1,\cdots,J) \tag{12.9}$$

① Daniel McFadden 由于在离散选择领域的开创性贡献而与 James Heckman 分享 2000 年诺贝尔经济学奖。

其中，解释变量 $\boldsymbol{x}_{ij}$ 既随个体 i 而变，也随方案 j 而变；而解释变量 $\boldsymbol{z}_i$ 只随个体 i 而变。经过类似的推导可知，个体 i 选择方案 j 的概率为

$$\mathrm{P}(y_i=j\mid \boldsymbol{x}_{ij})=\frac{\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\beta}_j)}{\sum_{k=1}^{J}\exp(\boldsymbol{x}_{ik}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\beta}_k)} \tag{12.10}$$

此模型在文献中称为“混合 Logit”(Mixed Logit)[①]，但 Stata 仍称其为条件 Logit。为了识别该模型，在方程(12.10)中，也需要选择一个参照方案(比如方案 1)，然后令 $\boldsymbol{\beta}_1=\mathbf{0}$。

对于以上三种模型，当各方案本身的特征不重要，或缺乏有关方案特征的数据时，常使用多项 Logit 模型。如果需要考虑不同方案的特征，则应使用条件 Logit 模型或混合 Logit 模型。另外，在这些多值选择模型中，由于被解释变量的分布必然为“多项分布”(multinomial distribution)，故一般不必使用稳健标准误，使用普通标准误即可；这一点类似于二值选择模型。然而，如果数据为聚类样本，则仍应使用聚类稳健的标准误(使用 Stata 命令选择项 `vce(cluster clustvar)`)。

需要注意的是，在多项 Logit 与混合 Logit 模型中，对参数估计值 $\hat{\boldsymbol{\beta}}_j$ 的解释是以参照方案(base category)为转移的(可根据理论或方便来选择参照方案)。以多项 Logit 为例，假设“方案 1”或“方案 j”($j\neq 1$)必然发生(二者必居其一)，则在此条件下，“方案 j”发生的条件概率为

$$\mathrm{P}(y=j\mid y=1\ \text{or}\ j)=\frac{\mathrm{P}(y=j)}{\mathrm{P}(y=1)+\mathrm{P}(y=j)}=\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)} \tag{12.11}$$

上式与二值选择的 Logit 模型具有完全相同的形式。而“几率比”(odds ratio)或“相对风险”(relative risk)为

$$\frac{\mathrm{P}(y=j)}{\mathrm{P}(y=1)}=\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j) \tag{12.12}$$

故对数几率比(log-odds ratio)为

$$\ln\left[\frac{\mathrm{P}(y=j)}{\mathrm{P}(y=1)}\right]=\boldsymbol{x}_i'\boldsymbol{\beta}_j \tag{12.13}$$

从条件概率 $\mathrm{P}(y=j\mid y=1\ \text{or}\ j)$ 的表达式可以看出，该条件概率并不依赖于任何其他方案。换言之，如果将多值选择模型中的任何两个方案单独挑出来，都是二值 Logit 模型。此假定称为“无关方案的独立性”(Independence of Irrelevant Alternatives，简记 IIA)。根据类似的推导可知，条件 Logit 模型也服从 IIA 假定。然而，在实践中，如果不同方案之间很类似，则 IIA 假设不一定满足。这是多项 Logit、条件 Logit 与混合 Logit 模型的共同缺点。

比如，假设共有三个备选的交通方式，即自驾车、乘红色公共汽车(red bus)、乘蓝色公共汽车(blue bus)。根据“无关选择的独立性”假定，在给定选择“自驾车”或“乘红色公共汽车”条件下，选择“自驾车”的条件概率与是否存在“乘蓝色公共汽车”的选择无关。然而，由于“蓝色公共汽车”与“红色公共汽车”仅仅颜色不同，故在加上“乘蓝色公共汽车”的选择之后，对选择“自驾车”的概率没有什么影响，但将使选择“乘红色公共汽车”的概率降低一半。因此，引入“蓝色公共汽车”后，上述条件概率将增大，与“无关选择的独立性”假定相矛盾。这个问题也被称为“红车蓝车问题”(red bus-blue bus problem)。

对于 IIA 假定，检验方法之一为豪斯曼检验。其基本思想是，如果 IIA 假定成立，则去掉某个方

---

① “Mixed logit”也被用来指另一完全不同的模型，即“随机系数 Logit”(random parameters logit)，参见 Cameron and Trivedi (2005)，在此不作介绍。

案不影响对其他方案参数的一致估计，只是降低了效率。这意味着，在 IIA 原假设成立的情况下，去掉某个方案后子样本的系数估计值（记为 $\hat{\boldsymbol{\beta}}_R$）与全样本的系数估计值（记为 $\hat{\boldsymbol{\beta}}_F$，不含被去掉方案的对应系数）没有系统差别。为此，Hausman and McFadden (1984) 提出以下统计量：

$$(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F)'[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_R)} - \widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_F)}]^{-1}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_F) \xrightarrow{d} \chi^2(m) \tag{12.14}$$

其中，m 等于 $\hat{\boldsymbol{\beta}}_R$ 的维度。检验 IIA 假定的方法之二由 Small and Hsiao (1985) 提出。然而，Cheng and Long (2007) 通过蒙特卡罗方法发现，这两个检验的小样本性质都不好（即有限样本下统计量的真实分布偏离方程 (12.14) 的渐近分布较远），故其结论只具有参考价值（两者的结论可能矛盾）。

在 Stata 中，多项选择模型所使用的数据格式依赖于解释变量的类型。对于所有解释变量都只随个体而变（case-specific）的多项 Logit 或多项 Probit 模型，应使用“宽形格式”（wide form），即同一个体的所有数据均出现在 Stata 数据表格的同一行。反之，在条件 Logit 模型与混合 Logit 模型中，由于存在随方案而变（alternative-specific）的解释变量，故应使用“长形格式”（long form），即同一个体的数据出现在 Stata 数据表格的几行，对应于备选的几个方案（一个方案占一行）。在 Stata 中，可通过命令 reshape 在宽形与长形格式之间转换，参见第 23 章。

多项 Logit 模型的 Stata 命令为

```
mlogit y x1 x2 x3,rrr base(#)
```

其中，选择项“base(#)”用于指定参照组（base category）；如果不指定，则默认使用观测值最多的方案为参照方案。选择项“rrr”表示汇报“相对风险比率”（Relative Risk Ratio，简记 RRR），即汇报 $\exp(\boldsymbol{\beta}_j)$，而非 $\boldsymbol{\beta}_j$。

在估计多项 Logit 模型后，可通过非官方命令 mlogtest 对 IIA 假定进行检验：

```
. net install sg155  (下载安装命令 mlogtest)①
. mlogtest,hausman base
. mlogtest,smhsiao base
```

其中，选择项“hausman”表示进行豪斯曼检验，选择项“smhsiao”表示进行 Small-Hsiao 检验，而选择项“base”表示在检验中包括“去掉参照方案，而以剩余方案中观测值最多的方案为参照方案”的检验。

多项 Probit 模型的 Stata 命令为

```
mprobit y x1 x2 x3,rrr base(#)
```

条件 Logit 模型的 Stata 命令为

```
clogit y x1 x2 x3,group(varname) or
```

其中，必选项“group(varname)”用来指定由归属同一个体的观测值所构成的一组（因为数据按长形排列）；选择项“or”表示汇报“几率比”（odds ratio）。命令 clogit 不能直接处理只随个体而变（case-specific）的解释变量。

混合 Logit 模型的 Stata 命令为

```
asclogit  y x1 x2 x3,case(varname) alternatives(varname) casevars
(varname) base(#) or
```

其中，命令“asclogit”表示“alternative-specific conditional logit”，必选项“case(varname)”用

① 或输入命令“findit sg155”寻找下载地址。

来指定个体(case),必选项"alternatives(varname)"用来指定方案(数据也按长形排列),选择项"casevars(varname)"表示只随个体而变的解释变量,选择项"base(#)"用来指定参照方案,选择项"or"表示汇报"几率比"(odds ratio)。

下面以数据集 nomocc2. dta 为例[①],进行多项 Logit 与多项 Probit 估计。此数据集来自 1982 年美国 General Social Survey,该问卷调查将受访者职业分为五类(occ),即服务人员(menial),蓝领(blue collar),工匠(craft),白领(white collar)以及专业人士(professional)。解释变量包括:是否白人(white),受教育年限(ed)以及工龄(exper)。显然,这三个解释变量都只依赖于个体,而不依赖于方案,故使用多项 Logit 或多项 Probit 回归。

```
. use nomocc2.dta,clear
```

先看一下数据的基本特征。

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
occ	337	3.397626	1.367913	1	5
white	337	.9169139	.2764227	0	1
ed	337	13.09496	2.946427	3	20
exper	337	20.50148	13.95936	2	66

其次,通过前 6 个数据来看数据格式。

```
. list in 1/6
```

	occ	white	ed	exper
1.	Menial	1	11	3
2.	Menial	1	12	14
3.	Menial	1	12	44
4.	Menial	1	12	18
5.	Menial	0	14	24
6.	Menial	1	13	38

由于数据按职业排序,故前 6 名都从事服务业。其中,5 名为白人,受教育年限从 11 ~ 14 年不等,而工龄从 3 ~ 44 年不等。由于数据表中每行对应于一名个体,故数据格式为宽形(wide form)。

下面通过列表粗略地考察受教育年限(ed)与职业(occ)的关系:

```
. table occ,contents(N ed mean ed sd ed)
```

其中,命令"table"表示将变量的统计特征列表(tables of summary statistics),而选择项"contents()"则用来指定最多 5 个统计量,"N ed mean ed sd ed"表示罗列变量 ed 的样本容量、均值与标准差(按变量 occ 分为 5 个子样本)。

Occupation	N(ed)	mean(ed)	sd(ed)
Menial	31	11.774194	2.186469
BlueCol	69	11.217391	2.571733
Craft	84	11.964286	2.119596
WhiteCol	41	13.170732	2.096455
Prof	112	15.4375	2.608998

① 此例来自 Long and Feese (2006)。

从上表可知,在样本中专业人士最多,且平均受教育年限最高;而蓝领的受教育程度最低。下面进行多项 Logit 回归。

```
. mlogit occ white ed exper,nolog
```

```
Multinomial logistic regression                 Number of obs   =        337
                                                LR chi2(12)     =     166.09
                                                Prob > chi2     =     0.0000
Log likelihood = -426.80048                     Pseudo R2       =     0.1629

------------------------------------------------------------------------------
         occ |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Menial       |
       white |  -1.774306   .7550543    -2.35   0.019    -3.254186   -.2944273
          ed |  -.7788519   .1146293    -6.79   0.000    -1.003521   -.5541826
       exper |  -.0356509    .018037    -1.98   0.048    -.0710028    -.000299
       _cons |   11.51833   1.849356     6.23   0.000     7.893659      15.143
-------------+----------------------------------------------------------------
BlueCol      |
       white |  -.5378027   .7996033    -0.67   0.501    -2.104996    1.029391
          ed |  -.8782767   .1005446    -8.74   0.000     -1.07534   -.6812128
       exper |  -.0309296   .0144086    -2.15   0.032      -.05917   -.0026893
       _cons |   12.25956   1.668144     7.35   0.000     8.990061    15.52907
-------------+----------------------------------------------------------------
Craft        |
       white |  -1.301963    .647416    -2.01   0.044    -2.570875   -.0330509
          ed |  -.6850365   .0892996    -7.67   0.000    -.8600605   -.5100126
       exper |  -.0079671   .0127055    -0.63   0.531    -.0328693    .0169351
       _cons |   10.42698   1.517943     6.87   0.000     7.451864    13.40209
-------------+----------------------------------------------------------------
WhiteCol     |
       white |  -.2029212   .8693072    -0.23   0.815    -1.906732     1.50089
          ed |  -.4256943   .0922192    -4.62   0.000    -.6064407   -.2449479
       exper |   -.001055   .0143582    -0.07   0.941    -.0291967    .0270866
       _cons |   5.279722   1.684006     3.14   0.002     1.979132    8.580313
-------------+----------------------------------------------------------------
Prof         |  (base outcome)
------------------------------------------------------------------------------
```

由于没有指定参照方案(base outcome),故命令 mlogit 自动选择观测值最多的方案(即专业人士)为参照方案。上表显示,在 5% 的显著性水平上,给定其他变量,白人(white)更不可能选择服务业或工匠;但是否白人对于选择蓝领或白领没有显著影响。受教育程度(ed)越高,越不可能选择除专业人士以外的职业。工龄越长(exper),越不可能选择服务业或蓝领;工龄对于选择工匠或白领无显著影响。

由于 IIA 假定是多项 Logit 模型的前提,下面检验 IIA 假定是否满足。

```
. mlogtest,hausman base
```

```
**** Hausman tests of IIA assumption

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.
You used the old syntax of hausman. Click here to learn about the new syntax.

(storing estimation results as _HAUSMAN)

 Omitted |      chi2     df    P>chi2   evidence
---------+--------------------------------------
  Menial |       7.324   12     0.835   for Ho
 BlueCol |       0.320   12     1.000   for Ho
   Craft |     -14.436   12     1.000   for Ho
WhiteCol |      -5.541   11     1.000   for Ho
    Prof |      15.566   12     0.212   for Ho
-----------------------------------------------
```

上表的前四行豪斯曼检验结果显示，去掉四个非参照方案（nonbase alternatives）中的任何一个方案，都不会拒绝 IIA 的原假设。由于使用了选择项“base”，故上表第五行计算去掉参照方案（Prof），而以剩余方案中观测值最多的方案作为参照方案的检验结果，同样也不拒绝 IIA 假设。但由于某种原因，却无法进行 Small – Hsiao 检验。

```
. mlogtest,smhsiao base
```

```
**** Small-Hsiao tests of IIA assumption

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.
equation 5 not found
r(303);
```

上文提及，这两个检验的小样本性质均不好，只具有参考价值；但至少没有发现违背 IIA 假定的迹象（可能这五类职业太不相同了）。

如要显示相对风险比率，可输入命令：

```
. mlogit occ white ed exper,nolog rrr
```

```
Multinomial logistic regression                 Number of obs   =        337
                                                LR chi2(12)     =     166.09
                                                Prob > chi2     =     0.0000
Log likelihood = -426.80048                     Pseudo R2       =     0.1629
```

occ	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Menial						
white	.169601	.128058	-2.35	0.019	.0386123	.7449581
ed	.4589326	.0526071	-6.79	0.000	.3665863	.5745417
exper	.9649771	.0174053	-1.98	0.048	.9314593	.999701
_cons	100542	185937.9	6.23	0.000	2680.234	3771571
BlueCol						
white	.5840301	.4669924	-0.67	0.501	.1218461	2.79936
ed	.4154983	.0417761	-8.74	0.000	.3411816	.5060029
exper	.9695438	.0139698	-2.15	0.032	.9425465	.9973143
_cons	210989.6	351961.2	7.35	0.000	8022.948	5548662
Craft						
white	.2719974	.1760954	-2.01	0.044	.0764686	.9674893
ed	.5040718	.0450134	-7.67	0.000	.4231365	.600488
exper	.9920646	.0126046	-0.63	0.531	.967665	1.017079
_cons	33758.18	51243	6.87	0.000	1723.071	661385.6
WhiteCol						
white	.8163426	.7096525	-0.23	0.815	.1485651	4.485678
ed	.653316	.0602483	-4.62	0.000	.5452883	.7827453
exper	.9989455	.0143431	-0.07	0.941	.9712254	1.027457
_cons	196.3154	330.5962	3.14	0.002	7.236457	5325.773
Prof	(base outcome)					

下面，根据模型预测个体选择各种职业的可能性，分别记为 occ1，occ2，occ3，occ4，occ5，并显示对前 5 个观测值的预测结果。

```
. predict occ1 occ2 occ3 occ4 occ5
(option pr assumed; predicted probabilities)
. list occ1 - occ5 in 1/5
```

	occ1	occ2	occ3	occ4	occ5
1.	.1681295	.4128002	.2760952	.085288	.0576871
2.	.1257816	.2945018	.3076293	.1328948	.1391926
3.	.0644456	.1738508	.3616529	.1922331	.2078175
4.	.1161744	.2771936	.3174044	.1409616	.148266
5.	.1691383	.0988214	.410424	.1063373	.215279

也可以选择其他职业作为参照方案,比如服务业。

```
. mlogit occ white ed exper,base(1) nolog
```

Multinomial logistic regression — Number of obs = 337; LR chi2(12) = 166.09; Prob > chi2 = 0.0000; Log likelihood = -426.80048; Pseudo R2 = 0.1629

occ	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Menial	(base outcome)					
BlueCol						
white	1.236504	.7244352	1.71	0.088	-.1833631	2.656371
ed	-.0994247	.1022812	-0.97	0.331	-.2998922	.1010428
exper	.0047212	.0173984	0.27	0.786	-.0293789	.0388214
_cons	.7412336	1.51954	0.49	0.626	-2.23701	3.719477
Craft						
white	.4723436	.6043097	0.78	0.434	-.7120817	1.656769
ed	.0938154	.097555	0.96	0.336	-.0973888	.2850197
exper	.0276838	.0166737	1.66	0.097	-.004996	.0603636
_cons	-1.091353	1.450218	-0.75	0.452	-3.933728	1.751022
WhiteCol						
white	1.571385	.9027216	1.74	0.082	-.1979166	3.340687
ed	.3531577	.1172786	3.01	0.003	.1232959	.5830194
exper	.0345959	.0188294	1.84	0.066	-.002309	.0715007
_cons	-6.238608	1.899094	-3.29	0.001	-9.960764	-2.516453
Prof						
white	1.774306	.7550543	2.35	0.019	.2944273	3.254186
ed	.7788519	.1146293	6.79	0.000	.5541826	1.003521
exper	.0356509	.018037	1.98	0.048	.000299	.0710028
_cons	-11.51833	1.849356	-6.23	0.000	-15.143	-7.893659

从上表可知,系数估计值随参照方案的不同而变化。下面,使用 multinomial probit 来估计此模型。

```
. mprobit occ white ed exper,nolog
```

Multinomial probit regression — Number of obs = 337
Log likelihood = -429.31856 — Wald chi2(12) = 105.61
Prob > chi2 = 0.0000

occ	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Menial						
white	-1.144907	.5027501	-2.28	0.023	-2.130279	-.1595352
ed	-.5094985	.0698816	-7.29	0.000	-.6464639	-.3725331
exper	-.0234636	.0109546	-2.14	0.032	-.0449343	-.0019929
_cons	7.46242	1.145854	6.51	0.000	5.216587	9.708253
BlueCol						
white	-.392222	.5182974	-0.76	0.449	-1.408066	.6236222
ed	-.5845723	.063011	-9.28	0.000	-.7080715	-.461073
exper	-.0225903	.00984	-2.30	0.022	-.0418764	-.0033042
_cons	8.188586	1.069264	7.66	0.000	6.092867	10.28431
Craft						
white	-.8903573	.457069	-1.95	0.051	-1.786196	.0054814
ed	-.4718874	.0579237	-8.15	0.000	-.5854157	-.3583591
exper	-.0077824	.0090923	-0.86	0.392	-.025603	.0100382
_cons	7.140264	.9896954	7.21	0.000	5.200496	9.080031
WhiteCol						
white	-.1434167	.5530156	-0.26	0.795	-1.227307	.9404739
ed	-.3038566	.0576254	-5.27	0.000	-.4168003	-.1909129
exper	-.0039043	.0095574	-0.41	0.683	-.0226365	.0148279
_cons	3.76544	1.036649	3.63	0.000	1.733645	5.797234
Prof	(base outcome)					

从上表可知，多项 Probit 的系数估计值与多项 Logit 并不相同，但二者的系数不具可比性；具有可比性的是两个模型的预测概率。为此，计算多项 Probit 模型所预测的各种职业的选择概率，分别记为 occ1p，occ2p，occ3p，occ4p，occ5p。

```
. predict occ1p occ2p occ3p occ4p occ5p
(option pr assumed; predicted probabilities)
```

下面计算两个模型职业预测概率的相关性。

```
. corr occ1 occ1p
```

	occ1	occ1p
occ1	1.0000	
occ1p	0.9979	1.0000

```
. corr occ2 occ2p
```

	occ2	occ2p
occ2	1.0000	
occ2p	0.9985	1.0000

```
. corr occ3 occ3p
```

	occ3	occ3p
occ3	1.0000	
occ3p	0.9935	1.0000

```
. corr occ4 occ4p
```

	occ4	occ4p
occ4	1.0000	
occ4p	0.9929	1.0000

```
. corr occ5 occ5p
```

	occ5	occ5p
occ5	1.0000	
occ5p	0.9989	1.0000

由以上结果可知,两个模型所预测的职业选择概率高度一致,相关系数均在 99% 以上。这意味着,使用多项 Logit 或多项 Probit 在实际上并无多少区别;只是多项 Probit 的计算时间更长,且无法从几率比角度解释系数估计值,故实践中常使用多项 Logit。

下面以数据集 travel2. dta 为例[①],进行条件 Logit 与混合 Logit 的估计。该数据集的观测单位为 152 组人群(每组可视为一个小型旅行团),每个人群从以下三种度假旅行方式中选择一种,即火车,长途大巴,或自驾车。随方案而变的解释变量包括 time(总旅行时间)与 invc(乘车成本,in-vehicle cost)。不随方案而变的解释变量包括 hinc(家庭收入,household income)与 psize(旅行团人数,party size)。

首先,通过前 6 个观测值来考察数据格式。

```
. use travel2.dta,clear
. list id mode train bus time invc choice hinc psize in 1/6,sepby(id)
```

	id	mode	train	bus	time	invc	choice	hinc	psize
1.	1	Train	1	0	406	31	0	35	1
2.	1	Bus	0	1	452	25	0	35	1
3.	1	Car	0	0	180	10	1	35	1
4.	2	Train	1	0	398	31	0	30	2
5.	2	Bus	0	1	452	25	0	30	2
6.	2	Car	0	0	255	11	1	30	2

其中,选择项"sepby(id)"表示根据变量 id 的取值来画上表中的横线(默认每隔 5 个观测值画一条横线)。上表显示,每个旅行团(由变量 id 指定)对应于 3 行数据,每行对应于一种旅行方式(以变量 mode 表示)。虚拟变量 Train = 1 表示,该行数据对应于乘火车的方案;虚拟变量 Bus = 1 表示,该行数据对应于乘长途大巴的方案;如果 Train = 0 且 Bus = 0,则该行数据对应于自驾车的方案。即使对于同一旅行团,每个方案的总旅行时间、乘车成本也是不同的。另一方面,对于同一旅行团,其家庭收入与旅行团规模则不变。被解释变量 choice 为虚拟变量,表示选择哪一种方案(比如,第 1 个旅行团选择自驾车,故 car 所对应的那一行 choice = 1,而其他两行 choice = 0)。

尽管数据集中只有 152 个旅行团,但由于每个旅行团占据 3 行数据,故实际的样本容量为 456(即 152 ×3)。下面看一下数据的统计特征。

```
. sum id mode train bus time invc choice hinc psize
```

① 此数据集取自 Long and Freese (2006),最早由 Greene and Hensher (1995)所使用。

Variable	Obs	Mean	Std. Dev.	Min	Max
id	456	103.8882	61.03044	1	210
mode	456	2	.8173933	1	3
train	456	.3333333	.4719223	0	1
bus	456	.3333333	.4719223	0	1
time	456	632.1096	270.2547	180	1440
invc	456	33.95175	21.795	2	109
choice	456	.3333333	.4719223	0	1
hinc	456	31.80921	19.25813	2	72
psize	456	1.809211	1.069457	1	6

上表显示，变量 id 的最大值为 210，但 id 的编码并不连续，实际的旅行团数为 152。

下面，进行条件 Logit 估计。由于命令 clogit 只接受随方案而变的解释变量，故仅使用总旅行时间（单位为分钟）、乘车成本、虚拟变量 train 以及虚拟变量 bus 为解释变量（以旅行方式 car 为参照方案）。

```
. clogit choice train bus time invc,group(id) nolog
```

```
Conditional (fixed-effects) logistic regression   Number of obs    =        456
                                                  LR chi2(4)       =     172.06
                                                  Prob > chi2      =     0.0000
Log likelihood = -80.961135                       Pseudo R2        =     0.5152

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |   2.671238   .4531611     5.89   0.000     1.783058    3.559417
         bus |   1.472335   .4007152     3.67   0.000     .6869474    2.257722
        time |  -.0191453   .0024509    -7.81   0.000    -.0239489   -.0143417
        invc |  -.0481658   .0119516    -4.03   0.000    -.0715905   -.0247411
------------------------------------------------------------------------------
```

上表显示，如果其他解释变量（time，invc）的取值相同，则旅行团最有可能选择火车，其次为长途大巴。另外，一个方案的总旅行时间越长，乘车成本越高，则选择该方案的概率越低。然而，由于这是非线性模型，故不易通过系数估计值来评价边际效应。为此，在上述命令中加上选择项“or”来计算风险比率。

```
. clogit choice train bus time invc,group(id) nolog or
```

```
Conditional (fixed-effects) logistic regression   Number of obs    =        456
                                                  LR chi2(4)       =     172.06
                                                  Prob > chi2      =     0.0000
Log likelihood = -80.961135                       Pseudo R2        =     0.5152

------------------------------------------------------------------------------
      choice | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |   14.45786   6.551738     5.89   0.000     5.94802    35.14272
         bus |   4.359401   1.746879     3.67   0.000     1.987639    9.561286
        time |   .9810368   .0024044    -7.81   0.000     .9763356    .9857607
        invc |   .9529758   .0113896    -4.03   0.000      .930912    .9755624
------------------------------------------------------------------------------
```

从上表可知，变量 time 的风险比率为 0.98，这意味着在给定其他变量的情况下，一个方案的总旅行时间每增加 1 分钟，则选择此方案的概率将乘以 0.98，即下降 2%。变量 invc 的风险比率可类似地解释。另一方面，虚拟变量 bus 的风险比率为 4.36，这意味着，如果各方案的时间与成本均相等，则旅行团选择长途大巴的概率是选择自驾车概率的 4.36 倍；虚拟变量 train 的风险比

率也可类似地解释。

下面计算条件 Logit 模型的预测概率。

```
. predict prob
(option pc1 assumed; probability of success given one success within group)
```

看一下第 1 个旅行团的预测概率。

```
. list id mode prob choice time invc in 1/3
```

	id	mode	prob	choice	time	invc
1.	1	Train	.0642477	0	406	31
2.	1	Bus	.0107205	0	452	25
3.	1	Car	.9250318	1	180	10

上表显示,第 1 个旅行团实际选择自驾车(Car),而模型预测选择自驾车的概率高达 0.925,而且正好是旅行时间与成本最低的方案。

对于条件 Logit 模型,也可用命令 asclogit 来估计。

```
. asclogit choice time invc,case(id) alternatives(mode) base(3) nolog
```

```
Alternative-specific conditional logit          Number of obs      =       456
Case variable: id                               Number of cases    =       152

Alternative variable: mode                      Alts per case: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                Wald chi2(2)       =     70.53
Log likelihood = -80.961135                     Prob > chi2        =    0.0000
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mode						
time	-.0191453	.0024509	-7.81	0.000	-.0239489	-.0143417
invc	-.0481658	.0119516	-4.03	0.000	-.0715905	-.0247411
Train						
_cons	2.671238	.4531611	5.89	0.000	1.783058	3.559417
Bus						
_cons	1.472335	.4007152	3.67	0.000	.6869474	2.257722
Car	(base alternative)					

对比上表与命令 clogit 的估计结果可知,二者的系数估计值与标准误均完全相同。使用命令 asclogit 的主要好处在于,它可以通过选择项“casevars(varname)”将只随个体而变的解释变量也包括在模型中。下面将解释变量“家庭收入”(hinc)与“旅行团人数”(psize)也包括进来,进行混合 Logit 模型估计。

```
. asclogit choice time invc, case(id) alternatives(mode) base(3)
casevars(hinc psize) nolog
```

```
Alternative-specific conditional logit          Number of obs      =       456
Case variable: id                               Number of cases    =       152

Alternative variable: mode                      Alts per case: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                Wald chi2(6)       =     69.09
Log likelihood = -77.504846                     Prob > chi2        =    0.0000
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mode						
time	-.0185035	.0025035	-7.39	0.000	-.0234103	-.0135966
invc	-.0402791	.0134851	-2.99	0.003	-.0667095	-.0138488
Train						
hinc	-.0342841	.0158471	-2.16	0.031	-.0653438	-.0032243
psize	-.0038421	.3098075	-0.01	0.990	-.6110537	.6033695
_cons	3.499641	.7579665	4.62	0.000	2.014054	4.985228
Bus						
hinc	-.0080174	.0200322	-0.40	0.689	-.0472798	.031245
psize	-.5141037	.4007015	-1.28	0.199	-1.299464	.2712569
_cons	2.486465	.8803649	2.82	0.005	.7609815	4.211949
Car	(base alternative)					

上表显示，家庭收入(hinc)越高，越不倾向于选择火车；但对于选择长途大巴则无显著影响。旅行团规模(psize)没有显著影响。旅行时间(time)与乘车成本(invc)的估计系数依然显著为负，且与前面条件 Logit 模型的估计系数很接近。上表并不汇报准 R^2，但可以很容易地手工计算。从上表可知，该模型的对数似然函数为 -77.504 846，下面估计一个只含常数项的模型。

```
. asclogit choice,case(id) alternatives(mode) base(3) nolog
```

```
Alternative-specific conditional logit          Number of obs      =       456
Case variable: id                               Number of cases    =       152

Alternative variable: mode                      Alts per case: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                Wald chi2(0)       =         .
Log likelihood = -160.00172                     Prob > chi2        =         .
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Train						
_cons	.0655973	.1811689	0.36	0.717	-.2894872	.4206818
Bus						
_cons	-.67634	.2242376	-3.02	0.003	-1.115838	-.2368425
Car	(base alternative)					

上表显示，只含常数项模型的对数似然函数为 -160.001 72，故可计算准 R^2 如下。

```
. dis (160.00172 -77.504846)/160.00172
.51559992
```

12.4 嵌套 Logit

上文提及，多项 Logit、条件 Logit 与混合 Logit 都须满足“无关方案的独立性”假定（IIA）。但在实践中，如果方案之间比较类似，则 IIA 假设可能不满足；限制了这些模型的应用。解决方法之一是，把相似的方案归入一组，允许同组内的方案相关，但不同组的方案相互独立。比如，在交通工具的选择上，可把“公交车、地铁”归入“公共交通组”，而把“自驾车、出租车”归入“私人交通组”，形成嵌套式（nested）的树形结构，参见图 12.1。

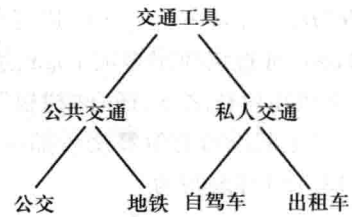


图 12.1 嵌套 Logit 的树形结构

图 12.1 中的“交通工具”可视为此树形结构的“根部”（root），第一层（level 1）的“公共交通、私人交通”为“树干”（limb），而第二层（level 2）的“公交、地铁、自驾车、出租车”为“树枝”（branch）。类似的例子包括消费者对居住地（residential location）的选择，即首先选择住在哪个社区，然后再选择住在该社区的哪套房子。

也可将此树形结构视为“决策树”（decision tree），但关键在于它的统计性质，即在第二层的每组内部允许扰动项相关，比如（$\varepsilon_{i,\text{bus}}$，$\varepsilon_{i,\text{subway}}$）相关，（$\varepsilon_{i,\text{car}}$，$\varepsilon_{i,\text{taxi}}$）相关；但这两组之间不相关，即（$\varepsilon_{i,\text{bus}}$，$\varepsilon_{i,\text{subway}}$）与（$\varepsilon_{i,\text{car}}$，$\varepsilon_{i,\text{taxi}}$）不相关。如果所有方案的扰动项都互不相关，则回到多项 Logit 或条件 Logit 的情形。

更一般地，假设在第一层共有 J 个树干备选，而树干 j 包含 K_j 个树枝（不同树干包含的树枝数可以不同），分别以 $j1,\cdots,jK_j$ 来指代。个体 i 选择树干 j 树枝 k 方案所能带来的效用为（为了简洁，略去个体 i 的下标）

$$U_{jk} = \boldsymbol{x}'_{jk}\boldsymbol{\beta} + \boldsymbol{z}'_j\boldsymbol{\gamma}_j + \varepsilon_{jk} \quad (j=1,\cdots,J;k=1,\cdots,K) \tag{12.15}$$

其中，解释变量 $\boldsymbol{z}_j$ 只随树干而不随树枝方案而变（故其系数 $\boldsymbol{\gamma}_j$ 可随 j 而变），而 $\boldsymbol{x}_{jk}$ 同时随树干与树枝方案而变（故其系数 $\boldsymbol{\beta}$ 没有下标，不随 j,k 而变）。McFadden（1978）假定扰动项 ε_{jk} 服从“广义极值分布”（Generalized Extreme Value distribution，简记 GEV），其累积分布函数为

$$F(\boldsymbol{\varepsilon}) = \exp[-G(\mathrm{e}^{-\varepsilon_{11}},\cdots,\mathrm{e}^{-\varepsilon_{1K_1}};\cdots;\mathrm{e}^{-\varepsilon_{J1}},\cdots,\mathrm{e}^{-\varepsilon_{JK_J}})] \tag{12.16}$$

其中，函数 $G(\cdot)$ 的形式为

$$G(\boldsymbol{w}) = G(w_{11},\cdots,w_{1K_1};\cdots;w_{J1},\cdots,w_{JK_J}) = \sum_{j=1}^{J}\left(\sum_{k=1}^{K_j} w_{jk}^{1/\tau_j}\right)^{\tau_j} \tag{12.17}$$

上式右边的每一项 $\left(\sum_{k=1}^{K_j} w_{jk}^{1/\tau_j}\right)^{\tau_j}$，在形式上为微观经济学中的不变替代弹性函数（Constant Elasticity of Substitution，简记 CES），其参数 $\tau_j = \sqrt{1-\mathrm{Corr}(\varepsilon_{jk},\varepsilon_{jl})}$，与（$\varepsilon_{jk}$，$\varepsilon_{jl}$）的相关系数呈反向变动的关系，故称为“不相似参数”（dissimilarity parameter），在 Stata 中记为 tau。显然，$0 \leqslant \tau_j \leqslant 1$。如果 $\tau_j = 1(j=1,\cdots,J)$，则所有方案的扰动项均不相关，满足 IIA 假定，回到多项 Logit、条件 Logit 或混合 Logit 的情形。基于以上假设，可以证明，选择树干 j 树枝 k 方案的概率为

$$p_{jk} = p_j \times p_{k|j} = \frac{\exp(\boldsymbol{z}'_j\boldsymbol{\gamma} + \tau_j I_j)}{\sum_{m=1}^{J}\exp(\boldsymbol{z}'_m\boldsymbol{\gamma} + \tau_m I_m)} \times \frac{\exp(\boldsymbol{x}'_{jk}\boldsymbol{\beta}_j/\tau_j)}{\sum_{l=1}^{K_j}\exp(\boldsymbol{x}'_{jl}\boldsymbol{\beta}_j/\tau_j)} \tag{12.18}$$

其中，p_j 为选择树干 j 的概率，$p_{k|j}$ 为在选择树干 j 的情况下，选择树枝 k 的条件概率；而 I_j 的定义为

$$I_j \equiv \ln\left[\sum_{l=1}^{K_j} \exp(\boldsymbol{x}_{jl}'\boldsymbol{\beta}_j/\tau_j)\right] \tag{12.19}$$

其中，I_j 称为“包含价值”(inclusive value)或“对数和”(log-sum)。根据公式(12.18)，可写出整个样本的对数似然函数，称为“全信息最大似然估计”(Full Information Maximum Likelihood，简记 FIML)。另一方法为分步对 p_j 与 $p_{k|j}$ 进行 MLE 估计，称为“有限信息最大似然估计”(Limited Information Maximum Likelihood，简记 LIML)，但不如 FIML 有效率。进行 FIML 估计后，可对联合假设“$H_0: \tau_1 = \cdots = \tau_J = 1$”进行似然比检验。如果接受原假设，则 IIA 假定成立，不需要使用嵌套 Logit，可直接使用多项 Logit、条件 Logit 或混合 Logit 模型。另外，也可以考虑更多层的嵌套结构，比如在树枝之下，还有“树杈”(twig)，估计方法类似。

以上假设所有解释变量都随方案而变。如果解释变量包括不随方案而变的个体特征，则方程(12.15)可改写为

$$U_{jk} = \boldsymbol{x}_{jk}'\boldsymbol{\beta} + \boldsymbol{z}_j'\boldsymbol{\gamma}_j + \boldsymbol{w}'\boldsymbol{\delta}_{jk} + \varepsilon_{jk} \quad (j = 1, \cdots, J; k = 1, \cdots, K) \tag{12.20}$$

在 Stata 中，如果系数只随 j 而变(比如 $\boldsymbol{\gamma}_j$)，则将其归属于“level 1 equation”，简记 lev1；如果系数既随 j 也随 k 而变(比如 $\boldsymbol{\delta}_{jk}$)，则将其归属于“level 2 equation”，简记 lev2；既不随 j 也不随 k 而变的系数(比如 $\boldsymbol{\beta}$)，则直接归属于模型本身(需要注意的是，以上只是 Stata 的称呼，整个模型是一起估计的，实际上只有一个方程)。

在 Stata 中，估计嵌套 Logit 之前，首先要定义树形的嵌套结构，其命令为

```
. nlogitgen newaltvar = altvar (branchlist)
```

其中，“altvar”为原先定义方案的变量，“newaltvar”为新定义的方案分组变量，而“branchlist”定义树枝的分组归类。

进行嵌套 Logit 估计的命令为

```
. nlogit y x1 x2 x3 ||lev1_equation ||lev2_equation,case(varname) base
(#) notree
```

其中，“x1 x2 x3”为随方案而变(alternative-specific)的随机变量(其系数相当于 $\boldsymbol{\beta}$，不随 j, k 而变)，“||”为分隔符(delimiter)，“lev1_equation”用来指定系数只随树干(level 1)而不随树枝(level 2)而变的解释变量(其系数相当于 $\boldsymbol{\gamma}_j$)，“lev2_equation”用来指定系数既随树干(level 1)又随树枝(level 2)而变的解释变量(其系数相当于 $\boldsymbol{\delta}_{jk}$)，必选项“case(varname)”用来指定个体，选择项“base(#)”用来指定参照方案，选择项“notree”表示不显示嵌套的树形结构(默认显示树形结构)。

继续以数据集 travel2.dta 为例。假设 Train 与 Bus 可归入一组(称为 public)，而 Car 自成一组(称为 private)，将定义嵌套结构的变量记为 type，可输入以下命令来定义树形结构：

```
. nlogitgen type =mode(public: Train |Bus,private: Car)
```

```
new variable type is generated with 2 groups
label list lb_type
lb_type:
           1 public
           2 private
```

如果要显示此树形结构，可使用以下命令：

```
. nlogittree mode type
```

```
tree structure specified for the nested logit model
 type      N       mode    N
-----------------------------
 public   304 ┬─ Train  152
              └─ Bus    152
 private  152 ── Car    152
-----------------------------
                total  456
N = number of observations at each level
```

更直观地,通过前 6 个观测值来看新生成的变量 type:

```
. list id mode train bus time invc choice hinc type in 1/6,sepby(id)
```

	id	mode	train	bus	time	invc	choice	hinc	type
1.	1	Train	1	0	406	31	0	35	public
2.	1	Bus	0	1	452	25	0	35	public
3.	1	Car	0	0	180	10	1	35	private
4.	2	Train	1	0	398	31	0	30	public
5.	2	Bus	0	1	452	25	0	30	public
6.	2	Car	0	0	255	11	1	30	private

从上表可知,对于方案 Train 与 Bus,变量 Type 的取值为 public;而对于方案 Car,变量 Type 的取值为 private。在数据集中,有两个不随方案而变的解释变量,即 hinc 与 psize。为了演示的目的,假设 psize 的系数只随树干方案而变(仅取决于是公共交通还是私人交通)①,而 hinc 的系数可随树枝方案而变(对于 Train,Bus,Car 各不相同)。

```
. nlogit choice time invc || type: psize,base(private) || mode: hinc,
base(3) case(id) nolog notree
```

```
RUM-consistent nested logit regression          Number of obs      =       456
Case variable: id                               Number of cases    =       152
Alternative variable: mode                      Alts per case: min =         3
                                                               avg =       3.0
                                                               max =         3
                                                Wald chi2(5)       =     40.43
Log likelihood = -71.596852                     Prob > chi2        =    0.0000
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mode						
time	-.0131104	.0025116	-5.22	0.000	-.0180331	-.0081877
invc	-.023953	.0092628	-2.59	0.010	-.0421077	-.0057983
type equations						
public						
psize	-.2313463	.2475569	-0.93	0.350	-.7165488	.2538563
private						
psize	0	(base)				
mode equations						
Train						
hinc	-.0357403	.0138418	-2.58	0.010	-.0628696	-.0086109
_cons	3.301306	.6493802	5.08	0.000	2.028544	4.574068
Bus						
hinc	-.0192732	.0150936	-1.28	0.202	-.0488562	.0103097
_cons	2.431383	.6584803	3.69	0.000	1.140785	3.721981
Car						
hinc	0	(base)				
_cons	0	(base)				
dissimilarity parameters						
type						
/public_tau	.3271257	.1057812			.1197983	.5344531
/private_tau	1	595784.1			-1167714	1167716

LR test for IIA (tau = 1): chi2(2) = 13.95 Prob > chi2 = 0.0009

① 也可指定 psize 的系数随树枝方案而变。

上表最后一行的似然比检验强烈拒绝 IIA 假定，故应使用嵌套 Logit 模型。另外，旅行团规模(psize)对于选择公共交通(public)或私人交通(private)没有显著影响。家庭收入(hinc)越高，越不倾向于选择火车，但对选择长途大巴无显著影响。变量 time 与 invc 的系数依然显著地为负，但估计值与前面的条件 Logit 或混合 Logit 模型的估计值有一定差距(由于后者没有考虑 Train 与 Bus 两个公共交通方案间的相关性，故可能高估了变量 time 与 invc 的作用)。最后，上表下部还显示，public 组的不相似参数为 0.327；而对于 private 组，由于只有一个方案，故不相似参数标准化为 1。

需要注意的是，使用嵌套 Logit 的前提是，各备选方案之间应有一个清晰的嵌套结构；而有些多值选择模型不存在清晰的嵌套结构。如果 IIA 假定不成立，但又没有清晰的嵌套结构，可考虑随机系数 Logit 模型(参见 Cameron and Trivedi，2005，在此从略)。

习　题

12.1 (此例来自 Cameron and Trivedi，2005)考虑潜变量模型 $y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$，其中 $\varepsilon_i \sim N(0,1)$。假设可观测变量 $y_i = 2$，如果 $y_i^* < \alpha$；$y_i = 1$，如果 $\alpha \leqslant y_i^* < U_i$；$y_i = 0$，如果 $y_i^* \geqslant U_i$。其中，U_i 已知但可随个体 i 而变，而 α 为未知。

(1) 给定 $\boldsymbol{x}_i$，计算 $y_i = 0$，$y_i = 1$，及 $y_i = 2$ 的条件概率。

(2) 具体说明，如何一致地估计 α 与 $\boldsymbol{\beta}$。

12.2 数据集 mlogit.dta① 包含 200 名高中学生的如下变量：定性变量 ice_cream(取值为 chocolate，vanilla 与 strawberry)，female(是否女性)，video(电子游戏成绩)，puzzle(猜谜成绩)。

(1) 估计一个多项 Logit 模型，以变量 female，video 与 puzzle 解释这些学生对冰淇淋口味的偏好。

(2) 解释此模型估计结果的经济意义。

(3) 对 ⅡA 假定，进行豪斯曼检验。

(4) 根据多项 Logit 模型，预测每位学生选择冰淇淋口味的可能性。

(5) 汇报此模型的相对风险比率(RRR)。

(6) 使用同样的变量，估计一个多项 Probit 模型。

(7) 根据多项 Probit 模型，预测每位学生选择冰淇淋口味的可能性。

(8) 比较这两种模型预测结果的相关性。

12.3 数据集 travel.dta② 包含有关交通工具选择的以下变量：subject(个体 ID)，mode(1 = Air，2 = Train，3 = Bus，4 = Car)，choice(1 = 选中的交通工具)，time(等待时间)，cost(成本)，income(家庭收入)，air(1 = 乘飞机)，train(坐火车)，bus(1 = 坐公交车)，car(1 = 自驾车)，air_inc(air 与 income 的互动项)

(1) 使用解释变量 time，cost，air_inc，估计条件 Logit 模型。

(2) 将交通工具分为两组，即空中交通(fly)与地面交通(ground)，进行嵌套 Logit 估计。

---

① 此数据集的下载地址为 http://www.ats.ucla.edu/stat/stata/output/mlogit.dta。

② 此数据集来自 http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm。

第 13 章　排序与计数模型

13.1　排 序 模 型

有时，离散数据有着天然的排序。比如，公司债券的评级（AAA，AA，A，B，C 级），对“春节联欢晚会”的满意度（很满意、满意、不满意、很不满意）。又比如，Li and Zhou（2005）研究经济增长绩效对地方官员仕途的影响，以 0 表示“卸任”，1 表示“留任或平级调动”，2 表示“提拔”。这种数据被称为“排序数据”（ordered data）。如果使用 multinomial logit，将无视数据内在的排序，而 OLS 又把排序视为基数来处理。对于排序数据，仍然可以使用潜变量法来推导出 MLE 估计量。

假设 $y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$（$y^*$ 不可观测），而选择规则为

$$y = \begin{cases} 0, & \text{若 } y^* \leqslant r_0 \\ 1, & \text{若 } r_0 < y^* \leqslant r_1 \\ 2, & \text{若 } r_1 < y^* \leqslant r_2 \\ \cdots\cdots & \\ J, & \text{若 } r_{J-1} \leqslant y^* \end{cases} \tag{13.1}$$

其中，$r_0 < r_1 < r_2 < \cdots < r_{J-1}$ 为待估参数，称为“切点”（cutoff points，在 Stata 输出结果中简称“cut”），参见图 13.1。

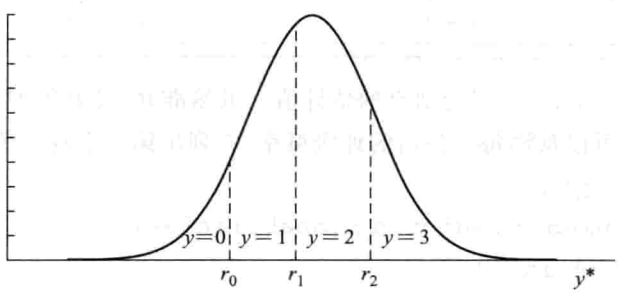


图 13.1　ordered logit 示意图

假设 $\varepsilon \sim N(0,1)$（将扰动项 ε 的方差标准化为 1），则

$$\begin{aligned} \mathrm{P}(y=0 \mid \boldsymbol{x}) &= \mathrm{P}(y^* \leqslant r_0 \mid \boldsymbol{x}) = \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \leqslant r_0 \mid \boldsymbol{x}) \\ &= \mathrm{P}(\varepsilon \leqslant r_0 - \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}) = \Phi(r_0 - \boldsymbol{x}'\boldsymbol{\beta}) \\ \mathrm{P}(y=1 \mid \boldsymbol{x}) &= \mathrm{P}(r_0 < y^* \leqslant r_1 \mid \boldsymbol{x}) \\ &= \mathrm{P}(y^* \leqslant r_1 \mid \boldsymbol{x}) - \mathrm{P}(y^* < r_0 \mid \boldsymbol{x}) \\ &= \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \leqslant r_1 \mid \boldsymbol{x}) - \Phi(r_0 - \boldsymbol{x}'\boldsymbol{\beta}) \end{aligned} \tag{13.2}$$

$$= P(\varepsilon \leqslant r_1 - \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}) - \Phi(r_0 - \boldsymbol{x}'\boldsymbol{\beta}) \tag{13.3}$$
$$= \Phi(r_1 - \boldsymbol{x}'\boldsymbol{\beta}) - \Phi(r_0 - \boldsymbol{x}'\boldsymbol{\beta})$$

$$P(y=2 \mid \boldsymbol{x}) = \Phi(r_2 - \boldsymbol{x}'\boldsymbol{\beta}) - \Phi(r_1 - \boldsymbol{x}'\boldsymbol{\beta}) \tag{13.4}$$
$$\cdots\cdots$$
$$P(y=J \mid \boldsymbol{x}) = 1 - \Phi(r_{J-1} - \boldsymbol{x}'\boldsymbol{\beta})$$

这样，可写出样本似然函数，并得到 MLE 估计量，即 ordered probit 模型。如果假设扰动项服从逻辑分布，则可得到 ordered logit 模型。

排序模型的 Stata 命令为

```
oprobit y x1 x2 x3        (ordered probit 模型)
ologit y x1 x2 x3         (ordered logit 模型)
```

下面以数据集“panel84extract. dta”为例[①]，估计决定公司债券评级的排序数据模型。被解释变量为 rating83c（1983 年公司债券评级，取值 2 - 5，其中 5 为最优评级），而解释变量为 ia83（1983 年 income-to-asset 比率），dia（ia 比率从 1982 年到 1983 年的变化，即 ia83 减 ia82）。

首先进行 ordered probit 估计：

```
. use panel84extract.dta,clear
. oprobit rating83c ia83 dia,nolog
```

```
Ordered probit regression                         Number of obs   =         98
                                                  LR chi2(2)      =      10.33
                                                  Prob > chi2     =     0.0057
Log likelihood = -127.87756                       Pseudo R2       =     0.0388

------------------------------------------------------------------------------
   rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ia83 |   .0512509   .0167257     3.06   0.002      .018469    .0840327
         dia |  -.0496082   .0254406    -1.95   0.051    -.0994709    .0002545
-------------+----------------------------------------------------------------
       /cut1 |  -.1877442   .2048422                     -.5892276    .2137392
       /cut2 |   .6344613   .2156938                      .2117092    1.057213
       /cut3 |   1.064523   .2231881                      .6270819    1.501963
------------------------------------------------------------------------------
```

上表中，“cut1，cut2，cut3”为切点的估计值。虽然准 R^2 仅为 0.04，但两个解释变量都比较显著。根据该模型，可以预测每个公司的评级概率，并列出第 1 个观测值的预测结果：

```
. predict p2 p3 p4 p5
(option pr assumed; predicted probabilities)
. list p2 p3 p4 p5 in 1/1
```

	p2	p3	p4	p5
1.	.2530799	.309446	.1590192	.2784549

下面进行 ordered logit 估计，然后预测每个公司的债券评级概率，并列出第 1 个观测值的预测结果：

```
. ologit rating83c ia83 dia,nolog
```

① 此例来自 Baum (2006)。

```
Ordered logistic regression                       Number of obs   =         98
                                                  LR chi2(2)      =      11.54
                                                  Prob > chi2     =     0.0031
Log likelihood = -127.27146                       Pseudo R2       =     0.0434

------------------------------------------------------------------------------
    rating83c |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         ia83 |   .0939166   .0296196     3.17   0.002     .0358633    .1519699
          dia |  -.0866925   .0449789    -1.93   0.054    -.1748496    .0014646
--------------+---------------------------------------------------------------
        /cut1 |  -.1853053   .3571432                     -.8852932    .5146826
        /cut2 |   1.185726   .3882099                      .4248488    1.946604
        /cut3 |   1.908412   .4164896                      1.092108    2.724717
------------------------------------------------------------------------------
```

```
. predict r2 r3 r4 r5
(option pr assumed; predicted probabilities)
. list r2 r3 r4 r5 in 1/1
```

	r2	r3	r4	r5
1.	.2532052	.3186541	.1615769	.2665637

根据上述结果可知，虽然 ordered probit 与 ordered logit 模型的系数与切点估计有所差别，但预测债券评级概率的结果却很接近。

13.2 泊松回归

有些被解释变量只能取非负整数，即 0,1,2,…，比如，专利个数、奥运金牌个数、子女人数、看病次数。对于这一类计数数据，常使用“泊松回归”(Poisson regression)。首先回顾泊松分布。假设在一次实验中某事件的发生概率为 p，共进行了 n 次相互独立的随机实验，记该事件的发生次数为 Y，则 $Y=y$ 的概率为

$$\mathrm{P}(Y=y)=C_n^y p^y(1-p)^{n-y} \quad (y=0,1,\cdots,n) \tag{13.5}$$

可以证明(参见标准的概率统计教材)，当 $p\to 0$，$n\to\infty$，而 $np=\lambda>0$ 时，此概率的极限为泊松分布：

$$\lim_{n\to\infty}\mathrm{P}(Y=y)=\lim_{n\to\infty}C_n^y p^y(1-p)^{n-y}=\frac{\mathrm{e}^{-\lambda}\lambda^y}{y!} \quad (y=0,1,\cdots,n) \tag{13.6}$$

由于此分布在 $p\to 0$ 时成立，故泊松分布也称为“稀有事件定律”(Law of rare events)，即如果在一次实验中事件发生概率很小(p 很小)，并进行了大量的实验(n 很大)，则事件发生次数大致服从泊松分布。回到计数数据的情形，对于个体 i，记被解释变量为 Y_i，假设 $Y_i=y_i$ 的概率由参数为 λ_i 的泊松分布决定：

$$\mathrm{P}(Y_i=y_i\mid \boldsymbol{x}_i)=\frac{\mathrm{e}^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad (y_i=0,1,2,\cdots) \tag{13.7}$$

其中，$\lambda_i>0$ 为“泊松到达率”(Poisson arrival rate)，表示事件发生的平均次数，由解释变量 $\boldsymbol{x}_i$ 所决定。泊松分布的期望与方差都等于泊松到达率，即 $\mathrm{E}(Y_i\mid \boldsymbol{x}_i)=\mathrm{Var}(Y_i\mid \boldsymbol{x}_i)=\lambda_i$。为了保证 λ_i 非负，假设 Y_i 的“条件期望函数”(conditional mean function)为

$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{13.8}$$

因此，$\ln \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$，为对数线性模型。假定样本独立同分布，则样本的似然函数为

$$L(\boldsymbol{\beta}) = \frac{\exp\left(-\sum_{i=1}^{n}\lambda_i\right) \cdot \prod_{i=1}^{n}\lambda_i^{y_i}}{\prod_{i=1}^{n} y_i!} \tag{13.9}$$

其对数似然函数为

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[-\lambda_i + y_i \ln \lambda_i - \ln(y_i!)\right] \\ &= \sum_{i=1}^{n}\left[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) + y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \ln(y_i!)\right] \end{aligned} \tag{13.10}$$

最大化的一阶条件为

$$\sum_{i=1}^{n}\left[y_i - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]\boldsymbol{x}_i = \mathbf{0} \tag{13.11}$$

通过数值计算可得 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$。根据 MLE 理论，如果似然函数正确，则 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ 为一致估计量。事实上，即使似然函数不正确，由于泊松分布属于线性指数分布族，故只要条件期望函数 $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ 正确，则准最大似然估计（QMLE）$\hat{\boldsymbol{\beta}}_{\mathrm{QMLE}}$ 就是一致的。直观来看，如果 $\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$，则方程（13.11）左边的条件期望就为 $\mathbf{0}$：

$$\mathrm{E}\left\{\sum_{i=1}^{n}\left[y_i - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]\boldsymbol{x}_i \mid \boldsymbol{x}_i\right\} = \sum_{i=1}^{n}\left[\mathrm{E}(y_i \mid \boldsymbol{x}_i) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]\boldsymbol{x}_i = \mathbf{0} \tag{13.12}$$

然而，如果似然函数不正确，则常规的标准误（比如，OIM 或 BHHH 法）就不是真实标准误的一致估计量，基于常规标准误的统计推断（比如，t 检验或 F 检验）均失效。此时，必须使用在 QMLE 基础上计算的稳健标准误（参见第 6 章），它对于似然函数是否正确比较稳健[①]。

显然，$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ 并不表示边际效应。由于 $\ln \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$，故 $\dfrac{\partial \ln \lambda_i}{\partial x_k} = \beta_k$。因此，可以将 β_k 解释为“半弹性”（semi-elasticity），即当解释变量 x_k 增加微小量时，事件的平均发生次数将增加多少百分点。另外，由于泊松到达率 $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$，故也可计算 $\exp(\beta_k)$，称为“发生率比”（Incidence Rate Ratio，简记 IRR），表示当 x_k 增加一单位时（从 x_k 增加到 $x_k + 1$），事件的平均发生次数将是原来的多少倍，因为 $\exp[(x_k + 1)\beta_k] / \exp(x_k\beta_k) = \exp(\beta_k)$。对模型拟合优度的衡量可以通过准 R^2 来进行。

泊松分布描述的是，在给定时间内某观测单位的事件发生次数。因此，如果观测时间变长，或观测单位的空间规模变大，则事件发生的平均次数也应同比例增多。记个体 i 在单位时间内事件发生的平均次数为 ϕ_i。如果不同个体的时间或空间规模不同，记为 T_i，称为“暴露期”（exposure），则该事件发生的平均次数也须相应调整为 $\phi_i T_i$。比如，考察某疾病在不同城市的发病人数，而各城市的人口基数不同。此时，可将方程（13.7）改写为

$$\mathrm{P}(Y_i = y_i \mid \boldsymbol{x}_i, T_i) = \frac{\mathrm{e}^{-\phi_i T_i}(\phi_i T_i)^{y_i}}{y_i!} \quad (y_i = 0, 1, 2, \cdots) \tag{13.13}$$

依然假设 $\phi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$。在上式中，令 $\lambda_i \equiv \phi_i T_i$，则 $\ln \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \ln T_i$。因此，如果暴露期 T_i 随 i 而变，则应把 $\ln T_i$ 作为解释变量放入泊松回归，并且令其系数为 1。如果暴露期不随个体而变，则自动归入常数项，无须作此特别处理。另外，关于“零断尾泊松回归”，参见第 14 章。

---

① 由于 $\mathrm{Var}(y_i \mid \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$，故此模型本来就允许异方差，因此其协方差矩阵是关于异方差稳健的。

泊松回归的 Stata 命令为

```
poisson y x1 x2 x3,r irr
poisson y x1 x2 x3,r exposure(x1)
poisson y x1 x2 x3,r offset(x1)
```

其中,选择项"r"表示使用稳健标准误;选择项"irr"表示显示发生率比;选择项"exposure(x1)"表示把 ln(x1)作为解释变量,并令其系数为 1;选择项"offset(x1)"表示把 x1 作为解释变量,并令其系数为 1。

13.3 负二项回归

泊松回归的局限是泊松分布的期望与方差一定相等,被称为"均等分散"(equidispersion);但这个特征常与实际数据不符。如果被解释变量的方差明显大于期望,即存在"过度分散"(overdispersion)。通常的处理方法是,在条件期望函数的对数表达式中加入一项:

$$\ln \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{13.14}$$

其中,随机变量 ε_i 表示条件期望函数中的不可观测部分或个体的异质性。由方程(13.14)可得:

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \cdot \exp(\varepsilon_i) \equiv u_i v_i \tag{13.15}$$

其中,$u_i \equiv \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ 为 $\boldsymbol{x}_i$ 的确定性函数,而 $v_i \equiv \exp(\varepsilon_i) > 0$ 仍为随机变量。给定 $\boldsymbol{x}_i$ 与 v_i,则 y_i 依然服从泊松分布:

$$P(Y_i = y_i \mid \boldsymbol{x}_i, v_i) = \frac{e^{-u_i v_i}(u_i v_i)^{y_i}}{y_i!} \quad (y_i = 0,1,2,\cdots) \tag{13.16}$$

但由于 v_i 不可观测,故无法对方程(13.16)进行估计。为此,记 v_i 的概率密度函数为 $g(v_i)$,则可以将 v_i 积分掉,以计算 y_i 的边缘密度:

$$P(Y_i = y_i \mid \boldsymbol{x}_i) = \int_0^{\infty} \frac{e^{-u_i v_i}(u_i v_i)^{y_i}}{y_i!} g(v_i)\, dv_i \tag{13.17}$$

由于 $v_i > 0$,故通常选择 v_i 服从 Gamma 分布(指数分布与卡方分布都是 Gamma 分布的特例)。特别地,假设 $v_i \sim \text{Gamma}(1/\alpha, \alpha)$,其中 $\alpha > 0$。由数理统计知识可知,对于 Gamma(a,b),其期望为 ab,而方差为 ab^2。故 $E(v_i) = 1$,而 $\text{Var}(v_i) = \alpha$(参数 α 即 v_i 的方差)。将 Gamma$(1/\alpha, \alpha)$的概率密度代入方程(13.17),可得到负二项分布的概率密度(证明参见 Cameron and Trivedi, 2005, p. 675),由此可以写出样本数据的似然函数,然后进行 MLE 估计。这称为"负二项回归"(negative binomial regression),因为它事实上假设样本来自"负二项分布"(negative binomial distribution)。

有关负二项分布,简单介绍如下。假设某事件在一次实验中成功的概率为 $\theta(0 < \theta < 1)$。记 Y 为在第 J 次成功前失败的总次数[①],则离散随机变量 Y 的分布律为

$$P(Y = y \mid \theta, J) = C_{y+J-1}^{J-1}\theta^J (1-\theta)^y \quad (y = 0,1,2,\cdots) \tag{13.18}$$

由于第$(y+J)$次一定为成功,故只要在前面的$(y+J-1)$次中找出成功的$(J-1)$次的组合次数即可。如果 $J = 1$,则称为"几何分布"(geometric distribution)。

可以证明,负二项回归模型的条件期望仍为 $E(Y_i \mid \boldsymbol{x}_i) = u_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$(负二项回归不影响

① 大概因为计算的是失败次数,故称为"负二项分布"。

条件期望),而条件方差为

$$\mathrm{Var}(Y_i \mid \boldsymbol{x}_i) = u_i + \alpha u_i^2 > u_i = \mathrm{E}(Y_i \mid \boldsymbol{x}_i) \tag{13.19}$$

这表明,在负二项回归中,条件方差大于条件期望(在泊松回归中,二者相等)。由于条件方差的表达式包含条件期望 u_i 的平方项,故也称为"NB2 模型"。从方程(13.19)可以看出,条件方差是 α 的增函数,故 α 称为"过度分散参数"(overdispersion parameter);特别地,当 $\alpha \to 0$ 时,泊松回归实际上是负二项回归的特例。因此,在进行负二项回归后,只要对原假设"$H_0: \alpha = 0$"进行检验,即可确定应使用负二项回归还是泊松回归。为此,Stata 命令提供一个 LR 检验,其原假设为"不存在过度分散,应使用泊松回归"。在完成负二项回归估计后,Stata 将自动输出检验结果"LR test of alpha = 0",其中过度分散参数"alpha = 0"对应于"泊松回归";即如果原假设"alpha = 0"成立,则可使用泊松回归。如果使用选择项"r"(稳健标准误),则 Stata 将只输出 alpha 的 95% 置信区间。根据这个置信区间,同样可以检验原假设"alpha = 0"。

另外,如果将 Gamma$(1/\alpha, \alpha)$ 中的 α 换为 δ/u_i,即假设 $v_i \sim$ Gamma$(u_i/\delta, \delta/u_i)$,其中 $\delta > 0$,则条件方差函数(13.19)变为

$$\mathrm{Var}(Y_i \mid x_i) = u_i + (\delta/u_i)u_i^2 = u_i + \delta u_i > u_i \tag{13.20}$$

此时,条件方差变为条件期望 u_i 的一次函数,故称为"NB1 模型",是负二项回归的另一形式。如果 $\delta \to 0$,则回到泊松回归的特例。此时,Stata 会汇报对原假设"delta = 0"的检验结果。在实践中,常使用 NB2 模型。在多数情况下,NB2 模型的二次条件方差函数更符合数据的特点。使用 NB2 模型的另一好处是,即使似然函数不正确,只要条件期望函数正确,则 $\hat{\beta}_{\mathrm{QMLE}}$ 依然是一致估计(但 $\hat{\alpha}_{\mathrm{QMLE}}$ 不是一致估计),此时应使用稳健标准误;而 NB1 模型则无此优点,参见 Cameron and Trivedi (1998)。

对于负二项回归,如果个体的暴露期 T_i 不同,同样应将 $\ln T_i$ 作为解释变量,并令其系数为 1。另外,关于"零断尾负二项回归",参见第 14 章。

在 NB2 负二项回归中,条件方差函数主要由参数 α 来刻画,不随个体 i 而变。作为一个推广,也可让此参数依个体 i 而变,记为 α_i,并让 $\ln \alpha_i$ 依赖于变量 $\boldsymbol{z}_i$($\boldsymbol{z}_i$ 可以与解释变量 $\boldsymbol{x}_i$ 有重叠),正如 $\ln \lambda_i$ 依赖于解释变量 $\boldsymbol{x}_i$,然后使用 MLE 对条件均值方程与条件方差方程同时进行估计,这称为"广义负二项回归"(generalized negative binomial regression);但在实践中较少使用。

负二项回归的 Stata 命令为

```
nbreg y x1 x2 x3,r exposure(x1)
nbreg y x1 x2 x3,r dispersion(constant) offset(x1)
```

其中,选择项"r"表示使用稳健标准误;选择项"dispersion(constant)"表示使用 NB1 模型(默认为 NB2 模型);选择项"exposure(x1)"表示把 $\ln x_1$ 作为解释变量,并令其系数为 1;选择项"offset(x1)"表示把 x_1 作为解释变量,并令其系数为 1。

广义负二项回归的 Stata 命令为

```
gnbreg y x1 x2 x3,r lnalpha(z1 z2)
```

其中,选择项"r"表示使用稳健标准误;选择项"lnalpha(z1 z2)"表示 $\ln \alpha_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2}$。如果不使用选择项"lnalpha(varlist)",则默认为 $\ln \alpha_i = \gamma_0$,即 α 为常数,将得到与负二项回归 nbreg 同样的结果。

究竟何时使用泊松回归或负二项回归?在某种意义上,泊松回归与负二项回归的关系有如线性模型中 OLS 与 WLS 的关系。即使数据中存在过度分散,"泊松回归 + 稳健标准误"依

然提供了对参数及标准误的一致估计；这类似于在异方差的情况下使用“OLS + 稳健标准误”。另一方面，如果比较了解条件方差函数，则“负二项回归 + 稳健标准误”可提供更有效率的估计；正如在异方差的情况下使用 WLS 可提高估计效率。因此，如果研究者只关心参数 $\boldsymbol{\beta}$ 的估计值，或许泊松回归就足够了；但如果希望预测“$Y=y$”的发生概率，则可考虑负二项回归。另外，对“$H_0: \alpha=0$”的 LR 检验结果也提供了在泊松回归与负二项回归之间选择的参考依据。

13.4 零膨胀泊松回归与负二项回归

如果计数数据中含有大量的“0”值(zero outcome)，则可考虑使用“零膨胀泊松回归”(Zero-inflated Poisson Regression，简记 ZIP)或“零膨胀负二项回归”(Zero-inflated Negative Binomial Regression，简记 ZINB)。从理论上来说，决策可能分两阶段进行。首先，决定“取零”(无)或“取正整数”(有)，这相当于二值选择。其次，如果决定“取正整数”，则进一步确定具体选择哪个正整数。为此，假定被解释变量 y_i 服从以下“混合分布”(mixed distribution)：

$$\begin{cases} \mathrm{P}(y_i=0\mid \boldsymbol{x}_i)=\theta \\ \mathrm{P}(y_i=j\mid \boldsymbol{x}_i)=\dfrac{(1-\theta)\mathrm{e}^{-\lambda_i}\lambda_i^j}{j!(1-\mathrm{e}^{-\lambda_i})} \quad (j=1,2,\cdots) \end{cases} \tag{13.21}$$

其中，$\lambda_i=\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$，而 $\theta>0$ 与 $\boldsymbol{\beta}$ 为待估参数。可以证明，$\sum_{j=0}^{\infty}\mathrm{P}(y_i=j\mid \boldsymbol{x}_i)=1$(参见习题)。因此，这是一个离散随机变量的分布律。进一步，可以让 θ 依赖于解释变量 $\boldsymbol{z}_i$($\boldsymbol{z}_i$ 可以等于 $\boldsymbol{x}_i$，或与 $\boldsymbol{x}_i$ 有重叠部分)，并用 Logit 模型来估计此二值选择问题，即 $y_i=0$ 或 $y_i>0$。使用 MLE 估计以上模型，即得到“零膨胀泊松回归”。类似地，可以定义“零膨胀负二项回归”。

究竟应该使用标准的泊松回归(standard Poisson)还是零膨胀泊松回归(ZIP)？Stata 提供了一个“Vuong 统计量”(Vuong，1989)，其渐近分布为标准正态。如果 Vuong 统计量很大(为正数)，则应选择零膨胀泊松回归(或零膨胀负二项回归)；反之，如果 Vuong 统计量很小(为负数)，则应选择标准的泊松回归(或“标准的负二项回归”)。

零膨胀回归的 Stata 命令为

```
zip y x1 x2 x3,inflate(varlist) vuong        (零膨胀泊松回归)
zinb y x1 x2 x3,inflate(varlist) vuong       (零膨胀负二项回归)
```

其中，选择项“inflate(varlist)”不可缺少，它列出 $\boldsymbol{z}_i$ 中的所有变量。如果 $\boldsymbol{z}_i$ 中只有常数项，则使用选择项“inflate(_cons)”。选择项“vuong”表示显示 Vuong 统计量。

第 17 章将介绍计数数据的面板模型。

13.5 计数模型的 Stata 实例

下面以有关犯罪率的数据集“CRIME1.dta”为例[①]，使用 Stata 来处理计数数据。该数据集包含以下变量：narr86(# of times arrested in 1986，1986 年被捕次数)，pcnv(proportion of prior

① 此例来自 Wooldridge (2003)。

convictions,有前科的比例),avgsen(average sentence length in months,平均判刑月数),tottime(total time in prison since 18 in months,18 岁以来入狱月数),ptime86(months in prison during 1986,1986 年入狱月数),qemp86(# of quarters employed in 1986,1986 年就业季度数),inc86(legal income in 1986 in $100s,1986 年合法收入),black(是否黑人),hispan(是否拉丁裔),born60(是否生于 1960 年)。考虑以下模型:

$$\begin{aligned} narr86_i = {} & \beta_0 + \beta_1\, pcnv_i + \beta_2\, avgsen_i + \beta_3\, tottime_i + \beta_4\, ptime86_i \\ & + \beta_5\, qemp86_i + \beta_6\, inc86_i + \beta_7\, black_i + \beta_8\, hispan_i + \beta_9\, born60_i + \varepsilon_i \end{aligned} \tag{13.22}$$

首先看一下被解释变量 narr86(1986 年被逮捕的次数)的分布:

```
. use CRIME1.dta,clear
. tab narr86
```

# times arrested, 1986	Freq.	Percent	Cum.
0	1,970	72.29	72.29
1	559	20.51	92.81
2	121	4.44	97.25
3	42	1.54	98.79
4	12	0.44	99.23
5	13	0.48	99.71
6	4	0.15	99.85
7	1	0.04	99.89
9	1	0.04	99.93
10	1	0.04	99.96
12	1	0.04	100.00
Total	2,725	100.00	

从上表可知,在样本容量 2 725 人中,有 1 970 人在 1986 年从未被逮捕。最高被逮捕纪录为 12 次。更直观地,可以看直方图。

```
. histogram narr86,discrete frequency
```

其中,选择项"discrete"表示 narr86 为离散变量,而"frequency"表示纵坐标为频数,结果参见图 13.2。

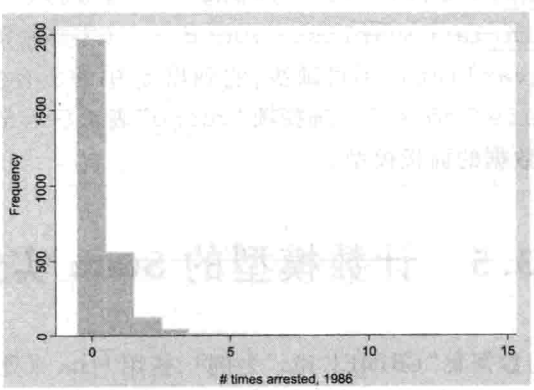


图 13.2 被捕次数 narr86 的直方图

显然,被解释变量为计数数据。尽管如此,作为对照,首先进行 OLS 回归[①]:

```
. reg narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r
```

```
Linear regression                               Number of obs =    2725
                                                F(  9,  2715) =   25.93
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0725
                                                Root MSE      =  .82873

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |   -.131886   .0335876    -3.93   0.000    -.1977458   -.0660262
      avgsen |  -.0113316   .0141409    -0.80   0.423    -.0390595    .0163963
     tottime |   .0120693   .0131776     0.92   0.360    -.0137699    .0379084
    ptime86 |  -.0408735   .0067985    -6.01   0.000    -.0542043   -.0275426
     qemp86 |  -.0513099    .014205    -3.61   0.000    -.0791636   -.0234562
      inc86 |  -.0014617   .0002289    -6.38   0.000    -.0019106   -.0010128
      black |   .3270097   .0584381     5.60   0.000     .2124221    .4415973
     hispan |   .1938094   .0401625     4.83   0.000     .1150572    .2725616
     born60 |   -.022465    .032094    -0.70   0.484    -.0853961    .0404661
      _cons |    .576566   .0426021    13.53   0.000     .4930302    .6601019
------------------------------------------------------------------------------
```

虽然 R^2 仅为 0.07,但多数解释变量都很显著。下面进行泊松回归,并使用稳健标准误:

```
. poisson narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r nolog
```

```
Poisson regression                              Number of obs   =       2725
                                                Wald chi2(9)    =     246.22
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -2248.7611               Pseudo R2       =     0.0791

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4015713   .1011619    -3.97   0.000    -.5998449   -.2032976
      avgsen |  -.0237723   .0236078    -1.01   0.314    -.0700427    .0224981
     tottime |   .0244904   .0205023     1.19   0.232    -.0156934    .0646741
    ptime86 |  -.0985584   .0223035    -4.42   0.000    -.1422724   -.0548445
     qemp86 |  -.0380187   .0341509    -1.11   0.266    -.1049532    .0289158
      inc86 |  -.0080807   .0012276    -6.58   0.000    -.0104867   -.0056747
      black |   .6608376   .0994572     6.64   0.000     .4659051     .85577
     hispan |   .4998133   .0923874     5.41   0.000     .3187374    .6808892
     born60 |  -.0510286   .0811403    -0.63   0.529    -.2100606    .1080034
      _cons |  -.5995888   .0893463    -6.71   0.000    -.7747044   -.4244732
------------------------------------------------------------------------------
```

尽管泊松回归的估计系数与 OLS 差别很大,但二者并不具可比性。为便于比较,计算泊松回归的平均边际效应:

---

① 对计数数据进行 OLS 回归时,常把被解释变量取对数,即把 $\ln y$ 对 x 进行 OLS 回归。但由于此处 narr86 可以等于 0,故无法直接取对数。解决方法包括把 $\ln(y+1)$ 对 x 进行 OLS 回归,或以 $\ln 0.5$ 来替代 $\ln 0$。一般来说,用 OLS 处理计数数据的主要目的为"探索性数据分析"(exploratory data analysis);比如,计数模型的某个解释变量为内生变量,而一般计量软件尚无法直接处理计数模型中的内生变量问题。

```
. margins,dydx(*)
```

```
Average marginal effects                        Number of obs   =       2725
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : pcnv avgsen tottime ptime86 qemp86 inc86 black hispan born60
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	-.1623969	.0404199	-4.02	0.000	-.2416184	-.0831754
avgsen	-.0096136	.0095756	-1.00	0.315	-.0283815	.0091543
tottime	.009904	.0083122	1.19	0.233	-.0063877	.0261957
ptime86	-.0398574	.0091649	-4.35	0.000	-.0578202	-.0218946
qemp86	-.0153749	.0139299	-1.10	0.270	-.0426771	.0119272
inc86	-.0032679	.0005148	-6.35	0.000	-.0042769	-.0022588
black	.2672451	.0420561	6.35	0.000	.1848168	.3496735
hispan	.2021263	.0382924	5.28	0.000	.1270745	.2771781
born60	-.0206361	.0328797	-0.63	0.530	-.0850793	.043807

从上表可知，泊松模型的平均边际效应与 OLS 的回归系数很接近。为便于解释系数，下面计算发生率比。

```
. poisson narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r irr nolog
```

```
Poisson regression                              Number of obs   =       2725
                                                Wald chi2(9)    =     246.22
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -2248.7611               Pseudo R2       =     0.0791
```

narr86	IRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	.6692676	.0677044	-3.97	0.000	.5488968	.8160353
avgsen	.976508	.0230532	-1.01	0.314	.932354	1.022753
tottime	1.024793	.0210106	1.19	0.232	.9844291	1.066811
ptime86	.9061427	.0202101	-4.42	0.000	.8673849	.9466324
qemp86	.9626949	.0328769	-1.11	0.266	.9003667	1.029338
inc86	.9919519	.0012177	-6.58	0.000	.9895681	.9943414
black	1.936414	.1925902	6.64	0.000	1.593456	2.353186
hispan	1.648413	.1522926	5.41	0.000	1.37539	1.975634
born60	.9502515	.0771037	-0.63	0.529	.8105351	1.114052
_cons	.5490374	.0490545	-6.71	0.000	.46084	.6541143

上表显示，给定其他变量，黑人被逮捕的平均次数比白人多 93.6%；其他变量的效应也可类似地解释。使用泊松回归的前提之一是，被解释变量的期望与方差相等。为此，考察 narr86 的统计特征。

```
. sum narr86,detail
```

times arrested, 1986

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	2725
25%	0	0	Sum of Wgt.	2725
50%	0		Mean	.4044037
		Largest	Std. Dev.	.8590768
75%	1	7		
90%	1	9	Variance	.7380129
95%	2	10	Skewness	4.111134
99%	4	12	Kurtosis	32.66587

结果显示,样本方差(Variance)几乎是样本均值(Mean)的两倍。为放松此假定,下面进行负二项回归(NB2):

```
. nbreg narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r nolog
```

Negative binomial regression — Number of obs = 2725
Dispersion = mean — Wald chi2(9) = 246.07
Log pseudolikelihood = -2157.628 — Prob > chi2 = 0.0000

narr86	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	-.4770963	.1013889	-4.71	0.000	-.6758148	-.2783778
avgsen	-.0173385	.0243654	-0.71	0.477	-.0650939	.0304168
tottime	.0197394	.0201359	0.98	0.327	-.0197263	.0592051
ptime86	-.1073997	.0230838	-4.65	0.000	-.1526432	-.0621562
qemp86	-.0504884	.0355401	-1.42	0.155	-.1201457	.0191689
inc86	-.0077126	.0012147	-6.35	0.000	-.0100934	-.0053318
black	.6560406	.0980021	6.69	0.000	.4639601	.8481211
hispan	.5048465	.0923508	5.47	0.000	.3238422	.6858508
born60	-.046412	.0798343	-0.58	0.561	-.2028843	.1100603
_cons	-.5637368	.0929054	-6.07	0.000	-.7458281	-.3816455
/lnalpha	-.0738912	.1459399			-.3599282	.2121458
alpha	.9287728	.135545			.6977264	1.236328

上表显示,alpha 的 95% 置信区间为(0.70,1.24),故可在 5% 的显著性水平上拒绝过度分散参数"alpha = 0"的原假设(对应于泊松回归),即认为应使用负二项回归。为演示目的,下面估计 NB1 负二项回归(尽管 NB1 不如 NB2 稳健)。

```
. nbreg narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r d(c) nolog
```

```
Negative binomial regression                    Number of obs   =       2725
Dispersion          = constant                  Wald chi2(9)    =     252.36
Log pseudolikelihood = -2163.7034               Prob > chi2     =     0.0000
```

narr86	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	-.5581407	.0938184	-5.95	0.000	-.7420214	-.3742601
avgsen	-.0134707	.0271602	-0.50	0.620	-.0667038	.0397624
tottime	.0173537	.0245949	0.71	0.480	-.0308514	.0655588
ptime86	-.1034884	.0213284	-4.85	0.000	-.1452912	-.0616856
qemp86	-.0030067	.0315804	-0.10	0.924	-.0649031	.0588898
inc86	-.0075688	.0011525	-6.57	0.000	-.0098276	-.00531
black	.623694	.0859821	7.25	0.000	.4551722	.7922159
hispan	.4491803	.0851075	5.28	0.000	.2823727	.6159879
born60	-.0155624	.0726358	-0.21	0.830	-.157926	.1268012
_cons	-.6198049	.0794918	-7.80	0.000	-.7756059	-.464004
/lndelta	-.7792728	.1637161			-1.10015	-.4583952
delta	.4587395	.075103			.332821	.6322976

上表显示，detla 的 95% 置信区间为(0.33,0.63)，故可在 5% 的显著性水平上拒绝“delta = 0”的原假设(对应于泊松回归)，认为应使用负二项回归。假设条件方差参数 α 依赖于 pcnv 与 ptime86，可进行以下广义负二项回归。

```
. gnbreg narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,r lna(pcnv ptime86) nolog
```

```
Generalized negative binomial regression        Number of obs   =       2725
                                                Wald chi2(9)    =     201.09
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -2130.1596               Pseudo R2       =     0.0505
```

narr86	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
narr86						
pcnv	-.3950848	.1103196	-3.58	0.000	-.6113072	-.1788624
avgsen	-.0144596	.0271741	-0.53	0.595	-.06772	.0388007
tottime	.0175512	.0211157	0.83	0.406	-.0238348	.0589371
ptime86	.1705355	.0731268	2.33	0.020	.0272095	.3138615
qemp86	-.0330864	.0342524	-0.97	0.334	-.1002198	.034047
inc86	-.0074203	.0011977	-6.20	0.000	-.0097677	-.0050729
black	.6395796	.0982466	6.51	0.000	.4470198	.8321395
hispan	.4099333	.0908442	4.51	0.000	.231882	.5879846
born60	-.0945714	.0797743	-1.19	0.236	-.2509261	.0617833
_cons	-.6190885	.091377	-6.78	0.000	-.7981842	-.4399928
lnalpha						
pcnv	2.057551	.4733131	4.35	0.000	1.129874	2.985228
ptime86	.3532886	.0428277	8.25	0.000	.2693479	.4372293
_cons	-.9914996	.3291731	-3.01	0.003	-1.636667	-.3463322

上表下部显示，条件方差方程(lnalpha)的两个变量均高度显著。而且，广义负二项回归的对数似然函数(-2 130.159 6)也大于相应的负二项回归(-2 163.703 4)，表明模型的拟合优度提高了。由于被解释变量的大多数取值为 0，下面考虑零膨胀泊松回归：

```
. zip narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born60,inf(_cons) vuong nolog
```

```
Zero-inflated Poisson regression                Number of obs   =      2725
                                                Nonzero obs     =       755
                                                Zero obs        =      1970

Inflation model = logit                         LR chi2(9)      =    290.06
Log likelihood  =  -2204.81                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
narr86       |
        pcnv |  -.3599746   .0967869    -3.72   0.000    -.5496734   -.1702758
      avgsen |   -.031854   .0241106    -1.32   0.186      -.07911    .015402
     tottime |   .0323113   .0180048     1.79   0.073    -.0029774    .0676001
     ptime86 |  -.1030596   .0235883    -4.37   0.000    -.1492918   -.0568274
      qemp86 |  -.0698728    .032354    -2.16   0.031    -.1332855   -.0064601
       inc86 |  -.0074046   .0010912    -6.79   0.000    -.0095433    -.005266
       black |   .6028718   .0834132     7.23   0.000     .4393849    .7663588
      hispan |   .4899381   .0827269     5.92   0.000     .3277963    .6520798
      born60 |  -.0607905   .0709329    -0.86   0.391    -.1998165    .0782356
       _cons |  -.1201744    .087349    -1.38   0.169    -.2913753    .0510264
-------------+----------------------------------------------------------------
inflate      |
       _cons |  -.5970509   .1258435    -4.74   0.000    -.8436996   -.3504022
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:           z =     3.23  Pr>z = 0.0006
```

上表显示，Vuong 统计量为 3.23，远远大于 1.96（比该统计量更大的概率仅为 0.000 6），故拒绝“标准泊松回归”，认为应使用“零膨胀泊松回归”。

下面进行零膨胀负二项回归：

```
. zinb narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan
born,inf(_cons) vuong nocon nolog
```

```
Zero-inflated negative binomial regression      Number of obs   =      2725
                                                Nonzero obs     =       755
                                                Zero obs        =      1970

Inflation model = logit                         Wald chi2(9)    =    296.49
Log likelihood  = -2174.656                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
narr86       |
        pcnv |  -.6181529   .1047333    -5.90   0.000    -.8234264   -.4128794
      avgsen |   -.022769   .0268294    -0.85   0.396    -.0753536    .0298157
     tottime |   .0231646   .0198022     1.17   0.242    -.0156469    .0619761
     ptime86 |  -.1210453   .0253992    -4.77   0.000    -.1708269   -.0712638
      qemp86 |  -.1349078    .032077    -4.21   0.000    -.1977776   -.0720381
       inc86 |  -.0070007   .0011356    -6.16   0.000    -.0092264   -.0047749
       black |   .4710273   .0861432     5.47   0.000     .3021898    .6398648
      hispan |   .3884124   .0879129     4.42   0.000     .2161063    .5607185
      born60 |  -.1547468   .0758095    -2.04   0.041    -.3033307    -.006163
-------------+----------------------------------------------------------------
inflate      |
       _cons |  -1.442806   .3069071    -4.70   0.000    -2.044333   -.8412795
-------------+----------------------------------------------------------------
    /lnalpha |  -.4957576   .2157331    -2.30   0.022    -.9185868   -.0729284
-------------+----------------------------------------------------------------
       alpha |   .6091093   .1314051                      .3990826    .9296674
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =    -3.66  Pr>z = 0.9999
```

其中，选择项“nocon”表示不带常数项[①]。上表显示，alpha 的 95% 置信区间为(0.40,0.93)，故可在 5% 的显著性水平上拒绝“alpha = 0”的原假设(对应于泊松回归)，即认为应使用负二项回归。另一方面，Vuong 统计量为 -3.66，远远小于 -1.96(比该统计量更小的概率仅为 0.000 1)，故拒绝“零膨胀负二项回归”，认为应使用“标准负二项回归”。总之，在以上各种模型设定中，更倾向于选择负二项回归。

习　题

13.1 对于零膨胀泊松回归，证明 $\sum_{j=0}^{\infty} \mathrm{P}(y_i = j \mid \boldsymbol{x}_i) = 1$。

13.2 使用排序数据集“mus18data1. dta”[②]，估计以下模型，

$$\text{health}_i = \beta_0 + \beta_1 \text{ age}_i + \beta_2 \text{ linc}_i + \beta_3 \text{ ndisease}_i + \varepsilon_i \tag{13.23}$$

其中，health 为健康状况(1 = 差，2 = 好，3 = 非常好)，age 为年龄，linc 为收入的对数，ndisease 为患慢性疾病的个数。

(1) 进行 OLS 回归；

(2) 进行 ordered probit 回归；

(3) 进行 ordered logit 回归。

13.3 使用数据集“poissonreg. dta”，估计决定“初中生旷课天数”(daysabs)的计数模型，解释变量为“语言艺术课成绩”(langarts)与“是否男性”(male)。

(1) 估计泊松回归模型；

(2) 估计负二项回归模型；

(3) 计算被解释变量的均值与方差，是否方差比均值大很多？

(4) 根据 LR 检验，应该使用泊松回归，还是负二项回归？

(5) 估计零膨胀泊松回归模型；

(6) 估计零膨胀负二项回归模型；

(7) 应该使用零膨胀模型吗？

---

① 这样做的原因有二：其一，在前面的零膨胀泊松回归中，常数项并不显著；其二，如果包括常数项，在进行零膨胀负二项回归的数值计算时，很难收敛。

② 此数据集截取自 Cameron and Trivedi (2010)。

第 14 章　受限被解释变量

在某些情况下，被解释变量的取值范围可能受限制，称为“受限被解释变量”（limited dependent variable）。本章考虑断尾回归（truncated regression）、归并回归（censored regression）与样本选择模型（sample selection）。

14.1　断 尾 回 归

对于线性模型 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i (i = 1, 2, \cdots, n)$，假设由于某种原因，只有满足 $y_i \geqslant c$（c 为常数）的数据才能观测到。因此，当 $y_i < c$ 时，没有任何有关 $\{y_i, \boldsymbol{x}_i\}$ 的数据。例如，y_i 的总体为某地区所有企业的年销售收入，而统计局只收集规模以上企业的数据，比如 $y_i \geqslant 100\ 000$。这样，被解释变量在 100 000 处就存在“左边断尾”（left truncation）。

断尾随机变量的概率分布

一个随机变量 y 断尾后，其概率密度函数也随之发生变化。记 y 原来的概率密度为 $f(y)$，则在 c 处左边断尾后的条件密度函数为

$$f(y \mid y > c) = \begin{cases} \dfrac{f(y)}{\mathrm{P}(y > c)}, & \text{若 } y > c \\ 0, & \text{若 } y \leqslant c \end{cases} \tag{14.1}$$

由于概率密度函数曲线下的面积必须为 1，故“断尾随机变量”（truncated random variable）的密度函数均乘以因子 $\dfrac{1}{\mathrm{P}(y > c)}$。断尾的效果参见图 14.1。

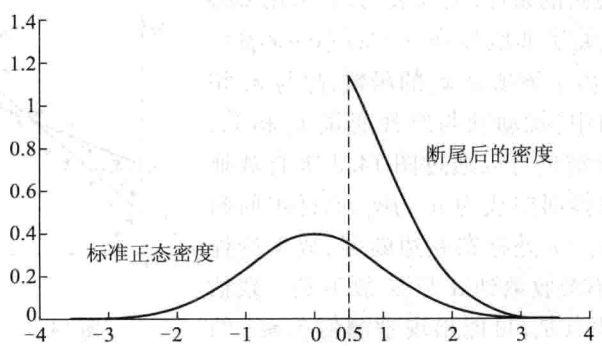


图 14.1　断尾的效果

显然，断尾分布的期望也发生了变化。由于回归的实质就是估计条件期望函数，下面以左边

断尾(left truncation)[①]为例来计算断尾分布的条件期望。首先,对于最简单的情形,$y \sim N(0,1)$,可以证明(参见本章附录)

$$\mathrm{E}(y \mid y > c) = \frac{\phi(c)}{1 - \Phi(c)} \tag{14.2}$$

对于任意实数 c,定义"反米尔斯比率"(Inverse Mill's Ratio,简记 IMR)[②]为 $\lambda(c) \equiv \frac{\phi(c)}{1-\Phi(c)}$,则 $\mathrm{E}(y \mid y > c) = \lambda(c)$。反米尔斯比率在 c 点的取值,等于标准正态的密度函数 $\phi(c)$ 除以密度函数曲线下比 c 大的阴影部分面积 $[1 - \Phi(c)]$,参见图 14.2。

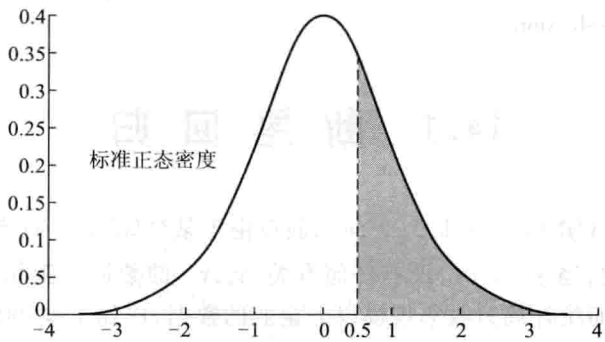


图 14.2 反米尔斯比率

其次,对于一般的正态分布 $y \sim N(\mu, \sigma^2)$,定义 $z \equiv \frac{y-\mu}{\sigma} \sim N(0,1)$,则 $y = \mu + \sigma z$。故

$$\begin{aligned}\mathrm{E}(y \mid y > c) &= \mathrm{E}(\mu + \sigma z \mid \mu + \sigma z > c) = \mathrm{E}[\mu + \sigma z \mid z > (c-\mu)/\sigma] \\ &= \mu + \sigma \mathrm{E}[z \mid z > (c-\mu)/\sigma] = \mu + \sigma \cdot \lambda[(c-\mu)/\sigma]\end{aligned} \tag{14.3}$$

其中,最后一个等号使用了标准正态的断尾期望公式(14.2)。对于回归模型 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$,假设 $\varepsilon_i \mid \boldsymbol{x}_i \sim N(0, \sigma^2)$。因此,$y_i \mid \boldsymbol{x}_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$。套用方程(14.3)可得

$$\mathrm{E}(y_i \mid y_i > c) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \lambda[(c - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma] \tag{14.4}$$

由于"$y_i > c$"是样本可观测的条件,上式表明,如果用 OLS 估计 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$,则遗漏了非线性项 $\sigma \cdot \lambda[(c - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma]$,被纳入扰动项中。由于该项是 $\boldsymbol{x}_i$ 的函数,故与 $\boldsymbol{x}_i$ 相关。因此,在 OLS 回归中,扰动项与解释变量 $\boldsymbol{x}_i$ 相关,导致不一致的估计。此结论可以通过图 14.3 来直观地展示。在图 14.3 中,总体回归线为 $\alpha + \beta x_i$,而样本回归线为 $\hat{\alpha} + \hat{\beta} x_i$。由于在 $y_i = c$ 处存在左边断尾,故无论样本容量多大,$\hat{\alpha}$ 与 $\hat{\beta}$ 都不会收敛到 α 与 β,故不是一致估计。OLS 回归的另一缺点是,可能出现预测值 $\hat{y}_i \leqslant c$ 的不可能情形。

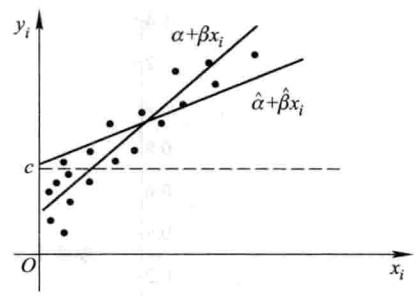


图 14.3 断尾回归示意图

使用 MLE 可以得到一致的估计。断尾前 y_i 的概率密度函数为

① 右边断尾(right truncation)的情形可类似地计算。

② 也称为"风险函数"(hazard function)。

$$f(y_i)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)^2\right\}=\frac{1}{\sigma}\phi\left(\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) \tag{14.5}$$

样本被观测到的概率为

$$\begin{aligned}\mathrm{P}(y_i>c\,|\,\boldsymbol{x}_i)&=1-\mathrm{P}(y_i\leqslant c\,|\,\boldsymbol{x}_i)\\&=1-\mathrm{P}\left(\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\leqslant\frac{c-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\,\middle|\,\boldsymbol{x}_i\right)\\&=1-\mathrm{P}\left(\frac{\varepsilon_i}{\sigma}\leqslant\frac{c-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\,\middle|\,\boldsymbol{x}_i\right)\\&=1-\Phi\left(\frac{c-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\end{aligned} \tag{14.6}$$

其中，ε_i/σ 服从标准正态分布。因此，根据公式(14.1)，断尾后的条件密度为

$$f(y_i\,|\,y_i>c,\boldsymbol{x}_i)=\frac{\frac{1}{\sigma}\phi[(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})/\sigma]}{1-\Phi[(c-\boldsymbol{x}_i'\boldsymbol{\beta})/\sigma]} \tag{14.7}$$

由此，可以计算出整个样本的似然函数，然后使用 MLE。

断尾回归的 Stata 命令为

```
truncreg y x1 x2 x3,ll(#)ul(#)
```

其中，选择项“ll(#)”表示 lower limit，即左边断尾；选择项“ul(#)”表示 upper limit，即右边断尾；如果同时使用这两个选择项，则表示双边断尾。

以数据集 laborsub.dta 为例[①]，估计一个决定妇女劳动时间的模型。该数据集包含以下变量：whrs(妻子的劳动时间，wife's hours of work)，kl6(6 岁以下儿童数，# of kids less than 6)，k618(6～18 岁少年数，# of kids between 6 and 18)，wa(妻子年龄，wife's age)，we(妻子教育年限，wife's years of education)，以及 lfp(是否就业，labor force participation)。

显然，如果 lfp=1(有工作)，则 whrs>0。为此，先看一下 lfp 的分布：

```
. use laborsub.dta,clear
. tab lfp
```

1 if woman worked in 1975	Freq.	Percent	Cum.
0	100	40.00	40.00
1	150	60.00	100.00
Total	250	100.00	

上表显示，在样本容量 250 人中，有 100 人不工作，即工作时间为 0。作为对照，先对有工作的 150 人的子样本进行 OLS 回归：

```
. reg whrs kl6 k618 wa we if whrs>0
```

① 此例来自 Baum(2006)。

Source	SS	df	MS	Number of obs = 150
				F(4, 145) = 2.80
Model	7326995.15	4	1831748.79	Prob > F = 0.0281
Residual	94793104.2	145	653745.546	R-squared = 0.0717
				Adj R-squared = 0.0461
Total	102120099	149	685369.794	Root MSE = 808.55

whrs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
kl6	-421.4822	167.9734	-2.51	0.013	-753.4748	-89.48953
k618	-104.4571	54.18616	-1.93	0.056	-211.5538	2.639668
wa	-4.784917	9.690502	-0.49	0.622	-23.9378	14.36797
we	9.353195	31.23793	0.30	0.765	-52.38731	71.0937
_cons	1629.817	615.1301	2.65	0.009	414.0371	2845.597

下面进行断尾回归,假设在“whrs = 0”处存在左边断尾:

```
. truncreg whrs kl6 k618 wa we,ll(0)nolog
```

```
Truncated regression
Limit:   lower =          0                    Number of obs =     150
         upper =       +inf                    Wald chi2(4)  =   10.05
Log likelihood = -1200.9157                    Prob > chi2   =  0.0395
```

whrs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
kl6	-803.0042	321.3614	-2.50	0.012	-1432.861	-173.1474
k618	-172.875	88.72898	-1.95	0.051	-346.7806	1.030578
wa	-8.821123	14.36848	-0.61	0.539	-36.98283	19.34059
we	16.52873	46.50375	0.36	0.722	-74.61695	107.6744
_cons	1586.26	912.355	1.74	0.082	-201.9233	3374.442
/sigma	983.7262	94.44303	10.42	0.000	798.6213	1168.831

比较以上两表可知,断尾回归的结果与对子样本进行 OLS 回归有较大不同。

14.2 零断尾泊松回归与负二项回归

对于离散数据,也可能出现断尾的情形。特别地,对于计数数据(参见第 13 章),有时样本中仅包括正整数,而不包括取值为 0 的观测值。比如,在商场发放问卷调查,研究消费者每周去商场的次数;在公交车上发放问卷调查,研究乘车者每周坐公交车的次数。在这两个例子中,虽然被解释变量为计数数据,但取值都不为 0,故在 0 处存在左边断尾,称为“零断尾”(zero-truncated)。

对于零断尾的计数模型,无论使用泊松回归还是负二项回归,如果不对似然函数进行适当调整,将得不到一致估计。记 $f(y)$ 为离散被解释变量 y 的概率函数,而 $F(y) \equiv \mathrm{P}(Y \leqslant y)$ 为其累积分布函数。如果存在零断尾,则断尾后的概率函数为

$$f(y \mid y \geqslant 1) = \frac{f(y)}{1 - F(0)}, \quad y = 1, 2, \cdots \tag{14.8}$$

如果 y 服从泊松分布,则

$$f(y|y\geq 1)=\frac{e^{-\lambda}\lambda^{y}}{y!\ (1-e^{-\lambda})},\quad y=1,2,\cdots \tag{14.9}$$

由此可写出样本数据的似然函数,进行 MLE 估计,即得到“零断尾泊松回归”(zero-truncated Poisson regression)。类似地,如果 y 服从负二项分布(可使用 NB1 或 NB2 模型),可进行“零断尾负二项回归”(zero-truncated negative binomial regression)。相应的 Stata 命令为:

```
ztp y x1 x2 x3,r  (零断尾泊松回归)
ztnb y x1 x2 x3,r  (零断尾负二项回归,默认为 NB2 模型)
ztnb y x1 x2 x3,r dispersion(constant)  (零断尾负二项回归,NB1 模型)
```

其中,选择项“r”表示使用稳健标准误。

下面以第 13 章有关犯罪率的数据集 CRIME1.dta 为例。为了使用零断尾回归,首先去掉样本中没有犯罪记录的观测值。

```
. use CRIME1.dta,clear
. drop if  narr86 = =0
(1970 observations deleted)
```

首先,进行零断尾泊松回归。

```
. ztp narr86 pcnv avgsen tottime ptime86 qemp86 inc86 black hispan born60,r nolog
```

```
Zero-truncated Poisson regression               Number of obs   =        755
                                                Wald chi2(9)    =     145.98
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -667.14707               Pseudo R2       =     0.1942
```

narr86	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	.6609906	.1852481	3.57	0.000	.297911	1.02407
avgsen	-.0680206	.0271185	-2.51	0.012	-.1211718	-.0148694
tottime	.0539391	.0143497	3.76	0.000	.0258143	.0820639
ptime86	.0322599	.0475728	0.68	0.498	-.060981	.1255009
qemp86	-.1939824	.0706526	-2.75	0.006	-.3324589	-.0555059
inc86	-.0070621	.003688	-1.91	0.056	-.0142905	.0001663
black	.385323	.1837578	2.10	0.036	.0251644	.7454817
hispan	.3182917	.1715818	1.86	0.064	-.0180025	.6545858
born60	-.2538259	.1493695	-1.70	0.089	-.5465847	.0389328
_cons	-.1183123	.1836823	-0.64	0.520	-.478323	.2416983

然而,如果直接用所有解释变量进行零断尾负二项回归,无论 NB1 还是 NB2,都不会收敛。为了演示目的,在解释变量中去掉在断尾回归中最不显著的两个变量 ptime86 与 born60,再进行 NB2 回归。

```
. ztnb narr86 pcnv avgsen tottime qemp86 inc86 black hispan,r nolog
```

```
Zero-truncated negative binomial regression     Number of obs   =        755
Dispersion     = mean                           Wald chi2(7)    =     131.10
Log likelihood = -624.91294                     Prob > chi2     =     0.0000
```

narr86	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pcnv	.868443	.242886	3.58	0.000	.3923953	1.344491
avgsen	-.0754736	.0378035	-2.00	0.046	-.1495671	-.0013801
tottime	.0623184	.0217107	2.87	0.004	.0197663	.1048706
qemp86	-.2788742	.0862416	-3.23	0.001	-.4479047	-.1098436
inc86	-.0069477	.0038483	-1.81	0.071	-.0144903	.0005949
black	.53225	.2279557	2.33	0.020	.085465	.9790349
hispan	.5326132	.2184151	2.44	0.015	.1045274	.9606989
_cons	-15.45257	4.655327	-3.32	0.001	-24.57684	-6.328298
/lnalpha	15.50324	4.785254			6.124316	24.88217
alpha	5407201	2.59e+07			456.8322	6.40e+10

14.3 随机前沿模型(选读)

在经济学中,常需估计生产函数或成本函数。教科书对生产函数 $f(\boldsymbol{x})$ 的标准定义为在给定投入 $\boldsymbol{x}$ 情况下的最大产出。但现实中的厂商可能达不到此最大产出的前沿(frontier)。为此,假设厂商 i 的产量为

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})\xi_i \tag{14.10}$$

其中,$\boldsymbol{\beta}$ 为待估参数;ξ_i 为厂商 i 的效率水平,满足 $0 < \xi_i \leqslant 1$。如果 $\xi_i = 1$,则厂商 i 正好位于效率前沿。考虑到生产函数还会受到随机冲击,故将方程(14.10)改写为

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})\xi_i e^{v_i} \tag{14.11}$$

其中,$e^{v_i} > 0$ 为随机冲击。方程(14.11)意味着生产函数的前沿 $f(\boldsymbol{x}_i, \boldsymbol{\beta})e^{v_i}$ 是随机的,故此类模型称为"随机前沿模型"(stochastic frontier model),最早由 Aigner, Lovell and Schmidt(1977)提出。随机前沿模型在实证研究中应用广泛,Kumbhakar and Lovell(2000)为此领域的专著。

假设 $f(\boldsymbol{x}_i, \boldsymbol{\beta}) = e^{\beta_0} x_{1i}^{\beta_1} \cdots x_{Ki}^{\beta_K}$(Cobb-Douglas 生产函数,且共有 K 个投入品),则对方程(14.11)两边取对数可得

$$\ln y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{ki} + \ln \xi_i + v_i \tag{14.12}$$

由于 $0 < \xi_i \leqslant 1$,故 $\ln \xi_i \leqslant 0$。定义 $u_i \equiv -\ln \xi_i \geqslant 0$,则方程(14.12)可写为

$$\ln y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{ki} + \underbrace{v_i - u_i}_{\varepsilon_i}, \quad u_i \geqslant 0 \tag{14.13}$$

其中,$u_i \geqslant 0$ 为"无效率项"(inefficiency term),反映厂商 i 离效率前沿的距离;而 v_i 称为"idiosyncratic error"。假设 v_i 为 iid,对称分布且独立于 u_i,则方程(14.13)的复合扰动项 $\varepsilon_i \equiv v_i - u_i$ 为非对称分布(因为 $u_i \geqslant 0$)。如果假设 u_i, v_i 都与 $\boldsymbol{x}_i$ 不相关,则使用 OLS 可以一致地估计 $\beta_1, \cdots, \beta_K$;但 OLS 无法一致地估计 β_0,因为复合扰动项的期望值 $E(\varepsilon_i) = E(v_i - u_i) = E(-u_i) \leqslant 0$。除了参数估计外,使用随机前沿模型的另一目的在于估计每个厂商的生产效率 u_i;但 OLS 无

法提供此信息。为了估计 u_i,必须对 u_i, v_i 的分布作假设;在此基础上,则可进行更有效率的 MLE 估计。首先,假设 u_i, v_i 均为 iid 且相互独立,而且独立于解释变量 $\boldsymbol{x}_i$。其次,假设 $v_i \sim N(0, \sigma_v^2)$,故其密度函数为 $\frac{1}{\sqrt{2\pi}\sigma_v}\exp(-v_i^2/2\sigma_v^2)$。最后,有关 u_i 的分布,常假定 $u_i \sim N^+(0, \sigma_u^2)$,即期望为 0 的正态分布,但在原点左边断尾,称为"半正态分布"(half-normal distribution);此模型也称为"正态 - 半正态模型"(normal-half normal model)(因为 v_i 服从正态分布),简称"半正态模型"。根据断尾分布的密度公式(14.1)可知,u_i 的密度函数为 $\frac{2}{\sqrt{2\pi}\sigma_u}\exp(-u_i^2/2\sigma_u^2)$。由于 u_i, v_i 相互独立,故其联合分布密度为二者边缘密度的乘积:

$$f(u_i, v_i) = \frac{1}{\pi\sigma_v\sigma_u}\exp\left\{-\frac{u_i^2}{2\sigma_u^2} - \frac{v_i^2}{2\sigma_v^2}\right\} \tag{14.14}$$

由此,可证明 $\varepsilon_i \equiv v_i - u_i$ 的密度函数为(Kumbhakar and Lovell,2000,p.75)

$$f(\varepsilon_i) = \frac{2}{\sigma}\phi(\varepsilon_i/\sigma)\Phi(-\varepsilon_i\lambda/\sigma) \tag{14.15}$$

其中,$\sigma \equiv \sqrt{\sigma_u^2 + \sigma_v^2}$,$\lambda \equiv \sigma_u/\sigma_v$,而 $\phi(\cdot)$,$\Phi(\cdot)$ 分别为标准正态的密度函数与累积分布函数。λ 度量的是 u 与 v 对 ε 的相对贡献。由于 $\phi(\varepsilon_i/\sigma) = \frac{1}{\sqrt{2\pi}}\exp[-(\varepsilon_i/\sigma)^2/2]$,故厂商 i 的对数似然函数为(忽略常数项)

$$\ln f(\varepsilon_i) = -\ln\sigma - \frac{1}{2}(\varepsilon_i/\sigma)^2 + \ln\Phi(-\varepsilon_i\lambda/\sigma) \tag{14.16}$$

假设样本为 iid,则样本容量为 n 的样本似然函数为

$$\ln L = -n\ln\sigma - \frac{1}{2}\sum_{i=1}^{n}(\varepsilon_i/\sigma)^2 + \sum_{i=1}^{n}\ln\Phi(-\varepsilon_i\lambda/\sigma) \tag{14.17}$$

最大化表达式(14.17)即得到对半正态模型的 MLE 估计。

由于半正态分布的众数(mode)为 0,故半正态模型意味着位于效率前沿的厂商多于位于任何其他位置的厂商。此假定在现实中未必成立。一种可能的情形是,只有少数厂商达到效率前沿或处于最无效率的状态,而多数厂商处于中间状态。为此,Stevenson(1980)引入如下"正态 - 断尾正态模型"(normal-truncated normal model),简称"断尾正态模型",即假设 $u_i \sim N^+(\mu, \sigma_u^2)$,即期望为 μ 的正态分布,但在原点左边断尾。显然,半正态模型是断尾正态模型的特例,即 $\mu = 0$ 的情形。由于断尾正态模型比半正态模型多了一个待估计参数 μ(Stata 记为 mu),故可更灵活地拟合数据。可以证明,断尾正态模型的样本对数似然函数为

$$\ln L = -n\ln\sigma - n\ln\Phi(-\mu/\sigma_u) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\varepsilon_i + \mu}{\sigma}\right)^2 + \sum_{i=1}^{n}\ln\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon_i\lambda}{\sigma}\right) \tag{14.18}$$

显然,在方程(14.18)中,如果令 $\mu = 0$,则简化为半正态模型的方程(14.17)。另一常用模型假设 u_i 服从指数分布,其密度函数为 $f(u_i) = \frac{1}{\sigma_u}\exp(-u_i/\sigma_u)$,$u_i \geqslant 0$,称为"正态 - 指数模型"(normal-exponential model),简称"指数模型"[①]。类似地,可以证明,指数模型的样本对数似然函

① 指数模型还可推广为 Gamma 模型,因为单一参数的指数分布是双参数 Gamma 分布的特例,参见 Kumbhakar and Lovell (2000)以及 Greene(2012)。但 Stata 不提供对 Gamma 模型的估计。

数为(Kumbhakar and Lovell,2000,p. 81)

$$\ln L = -n\ln \sigma_u - n\left(\frac{\sigma_v^2}{2\sigma_u^2}\right) + \sum_{i=1}^{n} \ln \Phi\left(-\frac{\varepsilon_i}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right) + \sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma_u}\right) \tag{14.19}$$

对以上模型进行 MLE 估计后,可计算残差 $\hat{\varepsilon}_i \equiv \ln y_i - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k \ln x_{ki}$,但这依然不是对 u_i 的估计。由于 $\varepsilon_i \equiv v_i - u_i$ 包含了 u_i 的信息①,故可通过条件分布 $f(u_i \mid \varepsilon_i)$ 来对 u_i 进行点估计。最常用的方法是以条件期望 $\mathrm{E}(u_i \mid \varepsilon_i)$ 作为 u_i 的估计,有时也以条件众数 $\mathrm{M}(u_i \mid \varepsilon_i)$ 作为 u_i 的估计(即条件分布 $f(u_i \mid \varepsilon_i)$ 的众数)。

在以上推导中均假设样本为 iid,排除了异方差的可能性。事实上,由于厂商规模不同,u_i 与 v_i 都可能存在异方差。对于半正态或指数模型,可假设 σ_i^2 依赖于变量 $\boldsymbol{z} \equiv (z_1, \cdots, z_m)$②:

$$\sigma_i^2 = \exp(\boldsymbol{z}_i'\boldsymbol{\delta}) \tag{14.20}$$

其中,σ_i^2 可以是 $\mathrm{Var}(u_i)$ 或 $\mathrm{Var}(v_i)$,而 $\boldsymbol{z}_i$ 为条件方差方程(14.20)的解释变量(可与 $\boldsymbol{x}_i$ 重叠)。将方程(14.20)代入样本对数似然函数,即可进行 MLE 估计。通过检验系数 $\boldsymbol{\delta}$ 的联合显著性即可判断是否存在异方差。

对于断尾正态模型,Stata 可以估计条件期望模型(conditional mean model),即让期望 μ_i 依赖于变量 $\boldsymbol{w}$:

$$\mu_i = \boldsymbol{w}_i'\boldsymbol{\gamma} \tag{14.21}$$

其中,$\boldsymbol{w}_i$ 为条件期望方程(14.21)的解释变量(可与 $\boldsymbol{x}_i$ 重叠)。

随机前沿模型可以很容易地用于估计成本函数。根据微观经济学的知识,给定成本情况下的产出最大化问题,其对偶问题就是给定产出情况下的成本最小化问题。对于 Cobb-Douglas 生产函数,容易证明其成本函数在形式上也是 Cobb-Douglas 函数。经过与生产函数的随机前沿模型类似的推导可得

$$\ln c_i = \beta_0 + \beta_y \ln y_i + \sum_{k=1}^{K} \beta_k \ln p_{ki} + \underbrace{v_i + u_i}_{\varepsilon_i}, \quad u_i \geqslant 0 \tag{14.22}$$

其中,c_i 为厂商 i 的成本,y_i 为产出,p_{ki} 为要素 k 的价格,u_i 为无效率项(inefficiency term),而 v_i 为对成本函数的随机冲击。对于成本函数而言,如果 $u_i = 0$,则达到最低成本的效率前沿;反之,如果 $u_i > 0$,则厂商需付出更高成本。由方程(14.22)可知,对于成本函数的随机前沿模型,可类似地进行 MLE 估计(只需略微调整对数似然函数即可)。

使用随机前沿模型的前提是无效率项 u_i 的存在。此假定可通过检验"$H_0: \sigma_u^2 = 0$ vs $H_1: \sigma_u^2 > 0$"来判断是否成立。如果接受原假设,则无须使用随机前沿模型。但困难在于,原假设 $\sigma_u^2 = 0$ 正好落在 σ_u^2 的参数空间($\sigma_u^2 \geqslant 0$)的边界上,导致标准的似然比检验不成立;此时,应使用单边的广义似然比检验(one-sided generalized likelihood ratio test),参见 Gutierrez et al (2001)。

检验无效率项 u_i 是否存在的另一方法通过检验 OLS 残差的偏度来进行,其基本思想如下。如果不存在无效率项 u_i,则复合扰动项 ε_i 简化为 v_i,为对称分布,偏度为 0;反之,如果存在无效率项 u_i,则复合扰动项 ε_i 的偏度将不为零。比如,对于生产函数模型,$\varepsilon_i \equiv v_i - u_i$;由于 u_i 与 v_i 相互独立,故 $\mathrm{E}(\varepsilon_i^3) \equiv \mathrm{E}(v_i^3) - \mathrm{E}(u_i^3) = -\mathrm{E}(u_i^3) \leqslant 0$(由于 v_i 为对称分布,故 $\mathrm{E}(v_i^3) = 0$),ε_i 的偏度为

① 比如,如果 ε_i 为正,则可大致推断 u_i 较小;反之,如果 ε_i 为负,则可大致推断 u_i 较大。

② 此处理方法类似于含异方差的 Probit 模型,参见第 11 章。

负数,即 ε_i 更多地分布于原点左边。因此,可通过 OLS 残差来检验复合扰动项的分布是否有偏,其原假设为"H_0 : $E(\varepsilon_i^3)=0$",而替代假设为"H_0 : $E(\varepsilon_i^3)<0$",故为单边左侧检验;反之,对于成本函数模型,$\varepsilon_i \equiv v_i + u_i$;故 ε_i 的偏度为正数,即 ε_i 更多地分布于原点右边;因此替代假设为"H_0 : $E(\varepsilon_i^3)>0$",为单边右侧检验。在原假设成立的条件下,Coelli(1995)证明检验统计量

$$\frac{m_3}{(6m_2^3/n)^{1/2}} \xrightarrow{d} N(0,1) \tag{14.23}$$

其中,m_2 与 m_3 分别为 OLS 残差的样本二阶矩与三阶矩①。根据此统计量进行上述的单边检验。如果接受原假设,则认为不存在无效率项 u_i,不必使用随机前沿模型;反之,则认为存在无效率项 u_i。此检验方法的优点在于它是基于 OLS 残差的,计算十分方便。

使用横截面数据的随机前沿模型一般假设 u_i,v_i 与 $\boldsymbol{x}_i$ 不相关。然而,由于厂商很可能知道自己的 u_i,并据此调整最优投入品 $\boldsymbol{x}_i$,故此假定可能不成立,导致不一致的估计。但这是横截面数据所无法克服的困难。一般来说,需要使用面板数据,才能在 u_i 与 $\boldsymbol{x}_i$ 相关的情况下得到一致估计。有关面板随机前沿模型,参见第 17 章。

随机前沿模型的 Stata 命令格式为

```
frontier lny lnx1 lnx2 lnx3, cost distribution(exponential)
distribution(tnormal) uhet(varlist) vhet(varlist) cm(varlist)
```

其中,"lny"为对数形式的被解释变量,"lnx1 lnx2 lnx3"为对数形式的解释变量。选择项"cost"表示估计成本函数,默认估计生产函数。选择项"distribution(exponential)"与"distribution(tnormal)"分别表示估计指数模型与断尾正态模型,默认估计半正态模型,即"distribution(hnormal)"。选择项"uhet(varlist)"用于指定决定异方差 Var(u_i) 的解释变量,选择项"vhet(varlist)"用于指定影响异方差 Var(v_i) 的解释变量;这两个选择项仅适用于半正态模型与指数模型(不适用于断尾正态模型)。选择项"cm(varlist)"用于指定影响条件期望 μ_i 的解释变量("cm"表示 conditional mean),此选择项仅适用于断尾正态模型。

下面以数据集 greene9.dta 为例。该数据集来自 Greene(2008),最早为 Zellner and Revankar (1969)所使用。它涉及美国 25 个州的运输设备制造行业,主要变量包括:lnv(增加值对数)、lnk(资本对数)、lnl(劳动力对数),state(州),以及 nfirm(该州的企业个数)。作为对照,首先以 OLS 估计生产函数。

```
. use greene9.dta,clear
. reg lnv lnk lnl,r
```

```
Linear regression                               Number of obs =      25
                                                F(  2,    22) =  625.05
                                                Prob > F      =  0.0000
                                                R-squared     =  0.9731
                                                Root MSE      =  .23571

------------------------------------------------------------------------------
             |               Robust
         lnv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnk |   .2454281   .0877269     2.80   0.010     .0634935    .4273626
         lnl |    .805183   .1230483     6.54   0.000     .5499965    1.060369
       _cons |   1.844416   .2829646     6.52   0.000     1.257583    2.431248
------------------------------------------------------------------------------
```

① 由于 OLS 残差的均值为 0,故残差的中心矩与原点矩相等。

其次,估计随机前沿的半正态模型。

```
. frontier lnv lnk lnl,nolog
```

```
Stoc. frontier normal/half-normal model          Number of obs   =        25
                                                 Wald chi2(2)    =    743.71
Log likelihood =  2.4695222                      Prob > chi2     =    0.0000

------------------------------------------------------------------------------
         lnv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnk |   .2585478    .098764     2.62   0.009     .0649738    .4521218
         lnl |   .7802451   .1199399     6.51   0.000     .5451672    1.015323
       _cons |   2.081135    .281641     7.39   0.000     1.529128    2.633141
-------------+----------------------------------------------------------------
    /lnsig2v |   -3.48401   .6195353    -5.62   0.000    -4.698277   -2.269743
    /lnsig2u |  -3.014599    1.11694    -2.70   0.007    -5.203761   -.8254368
-------------+----------------------------------------------------------------
     sigma_v |   .1751688   .0542616                      .0954514    .3214633
     sigma_u |   .2215073   .1237052                       .074134    .6618486
      sigma2 |   .0797496   .0426989                     -.0039388     .163438
      lambda |   1.264536   .1678684                      .9355204    1.593552
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01) =  0.43   Prob>=chibar2 = 0.256
```

上表中的 sigma_v,sigma_u,sigma2 与 lambda 分别表示 σ_v,σ_u,$\sigma^2 \equiv \sigma_u^2 + \sigma_v^2$ 以及 $\lambda \equiv \sigma_u / \sigma_v$。由于 $\hat{\lambda} = 1.26$,故无效率项 u 在复合扰动项 $\varepsilon \equiv v - u$ 占据了主导地位。然而,上表最后一行的似然比检验却接受了原假设“$H_0 : \sigma_u^2 = 0$”,即认为不存在无效率项,不必使用随机前沿模型;这很可能是由于样本容量过小所致(仅为 25)。下面,估计每个州的无效率项 u_i,并将此估计值记为 u_h。

```
. predict u_h,u
```

其中,选择项“u”表示使用 $E(u_i | \varepsilon_i)$ 预测 u_i,而默认预测 $\ln \hat{y}_i$。

再次,估计随机前沿的指数模型。

```
. frontier lnv lnk lnl,d(e) nolog
```

```
Stoc. frontier normal/exponential model          Number of obs   =        25
                                                 Wald chi2(2)    =    845.68
Log likelihood =  2.8604897                      Prob > chi2     =    0.0000

------------------------------------------------------------------------------
         lnv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnk |   .2624859   .0919988     2.85   0.004     .0821717    .4428002
         lnl |   .7703795   .1109569     6.94   0.000     .5529079    .9878511
       _cons |   2.069242   .2356159     8.78   0.000     1.607444    2.531041
-------------+----------------------------------------------------------------
    /lnsig2v |  -3.527598   .4486176    -7.86   0.000    -4.406873   -2.648324
    /lnsig2u |  -4.002457   .9274575    -4.32   0.000    -5.820241   -2.184674
-------------+----------------------------------------------------------------
     sigma_v |   .1713925   .0384448                      .1104231    .2660258
     sigma_u |   .1351691   .0626818                      .0544692    .3354317
      sigma2 |   .0476461   .0157921                       .016694    .0785981
      lambda |   .7886525    .087684                      .616795    .9605101
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01) =  1.21   Prob>=chibar2 = 0.135
```

上表最后一行依然接受了原假设“$H_0 : \sigma_u^2 = 0$”,尽管检验的 p 值缩小为 0.135。下面,估计每个州的无效率项 u_i,并将此估计值记为 u_e。

```
. predict u_e,u
```

由于样本容量过小,如果估计断尾正态模型,将无法收敛。下面罗列半正态模型与指数模型对无效率项的预测。

```
. list state u_h u_e
```

	state	u_h	u_e
1.	Alabama	.2011338	.1459287
2.	California	.1448097	.0972165
3.	Connecticut	.1903485	.134788
4.	Florida	.5175314	.5903303
5.	Georgia	.1039791	.0714099
6.	Illinois	.121267	.0830415
7.	Indiana	.2112821	.1545066
8.	Iowa	.2493315	.2007308
9.	Kansas	.1009952	.0685763
10.	Kentucky	.0562692	.0415244
11.	Louisiana	.2033273	.150664
12.	Maine	.2226316	.1724579
13.	Maryland	.1353406	.092455
14.	Massachusetts	.15637	.1093292
15.	Michigan	.1580957	.1075692
16.	Missouri	.1028805	.0704146
17.	NewJersey	.0958434	.0658799
18.	NewYork	.2778779	.2224942
19.	Ohio	.2291423	.1698186
20.	Pennsylvania	.1500667	.103029
21.	Texas	.2029787	.1455222
22.	Virginia	.1400013	.0967608
23.	Washington	.1104758	.0753325
24.	WestVirginia	.1556139	.1123615
25.	Wisconsin	.1406707	.0970861

然后计算这两个模型预测结果的相关系数。

```
. corr u_h u_e
```

	u_h	u_e
u_h	1.0000	
u_e	0.9773	1.0000

由上表可知,半正态模型与指数模型对于 u_i 预测值的相关系数高达 0.98。这表明,从预测无效率项 u_i 的角度来看,无论使用何种模型,结果几乎相同,并不依赖于对 u_i 具体分布的假定。

虽然在样本容量较小的情况下,一般不宜估计含异方差的随机前沿模型[①];但为了演示目的,下面估计允许 u_i 与 v_i 都有异方差的半正态模型,并假定异方差取决于变量 nfirm(企业个数)。

```
. frontier lnv lnk lnl,uhet(nfirm) vhet(nfirm) nolog
```

① 否则,可能导致对条件方差方程的参数估计不准确,参见 Stata 手册有关命令 frontier 的说明。

```
Stoc. frontier normal/half-normal model          Number of obs   =        25
                                                 Wald chi2(2)    =   2135.66
Log likelihood =  8.9688909                      Prob > chi2     =    0.0000
```

lnv	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnv						
lnk	.2429007	.0540886	4.49	0.000	.136889	.3489125
lnl	.9309447	.0667146	13.95	0.000	.8001864	1.061703
_cons	1.537479	.1417207	10.85	0.000	1.259711	1.815246
lnsig2v						
nfirm	-.1017466	.0362518	-2.81	0.005	-.1727988	-.0306944
_cons	1.62307	1.891768	0.86	0.391	-2.084727	5.330867
lnsig2u						
nfirm	.0055647	.0028622	1.94	0.052	-.0000451	.0111745
_cons	-4.001382	.8248283	-4.85	0.000	-5.618016	-2.384748

上表中部的结果表明，变量 nfirm 对 v_i 的异方差有显著的负影响，p 值为 0.005；而上表底部的结果表明，变量 nfirm 对 u_i 的异方差的影响也几乎在 5% 水平上显著（p 值为 0.052）。当然，由于样本容量太小，这些结果未必稳健或可信。

下面使用 Stata 提供的数据集 frontier2.dta 来演示断尾正态模型。该数据用于估计 1231 家企业的成本函数，主要变量包括 lncost（成本对数）、lnout（产出对数）、lnp_k（资本价格对数）、lnp_l（劳动价格对数）、quality（企业管理质量评级，取值为 1,2,3,4,5）。

```
. use frontier2,clear
. frontier lncost lnp_k lnp_l lnout,cost d(t) nolog
```

```
Stoc. frontier normal/truncated-normal model     Number of obs   =      1231
                                                 Wald chi2(3)    =      8.82
Log likelihood = -2386.2493                      Prob > chi2     =    0.0318
```

lncost	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnp_k	.3410717	.2363861	1.44	0.149	-.1222366	.80438
lnp_l	.6608628	.4951499	1.33	0.182	-.3096131	1.631339
lnout	.7528653	.3468968	2.17	0.030	.0729601	1.432771
_cons	2.602609	1.083004	2.40	0.016	.4799595	4.725259
/mu	1.095705	.881517	1.24	0.214	-.632037	2.823446
/lnsigma2	1.5534	.1873464	8.29	0.000	1.186208	1.920592
/ilgtgamma	1.257862	.2589522	4.86	0.000	.7503255	1.765399
sigma2	4.727518	.8856833			3.274641	6.825001
gamma	.7786579	.0446303			.6792496	.8538846
sigma_u2	3.681119	.7503408			2.210478	5.15176
sigma_v2	1.046399	.2660035			.5250413	1.567756

```
H0: No inefficiency component:                   z =   5.595          Prob>=z = 0.000
```

上表最后一行提供了 Coelli(1995) 基于 OLS 残差偏度的检验。检验结果表明，可以在 1% 水平上拒绝“没有无效率项”的原假设，即认为存在无效率项，应使用随机前沿模型。

对于断尾正态模型，也可通过选择项“cm(varlist)”来指定影响条件期望 μ_i 的解释变量。假设 μ_i 取决于企业 i 的管理质量（quality）。由于变量 quality 分为五级，故可通过“i.quality”来

表示由 quality 所生产的四个虚拟变量[①]。相应的命令为

```
. frontier lncost lnp_k lnp_l lnout,cost d(t) cm (i.quality) nolog
```

```
Stoc. frontier normal/truncated-normal model      Number of obs   =      1231
                                                  Wald chi2(3)    =      9.31
Log likelihood = -2382.3233                       Prob > chi2     =    0.0254
```

lncost	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lncost						
lnp_k	.3611204	.2359749	1.53	0.126	-.1013819	.8236227
lnp_l	.680446	.4934935	1.38	0.168	-.2867835	1.647675
lnout	.7605533	.3466102	2.19	0.028	.0812098	1.439897
_cons	2.550769	1.078911	2.36	0.018	.4361417	4.665396
mu						
quality						
2	.5056067	.3382907	1.49	0.135	-.1574309	1.168644
3	.783223	.376807	2.08	0.038	.0446947	1.521751
4	.5577511	.3355061	1.66	0.096	-.0998288	1.215331
5	.6792882	.3428073	1.98	0.048	.0073981	1.351178
_cons	.6014025	.990167	0.61	0.544	-1.339289	2.542094
/lnsigma2	1.541784	.1790926	8.61	0.000	1.190769	1.892799
/ilgtgamma	1.242302	.2588968	4.80	0.000	.734874	1.749731
sigma2	4.67292	.8368852			3.289611	6.637923
gamma	.7759645	.0450075			.6758739	.8519189
sigma_u2	3.62602	.7139576			2.226689	5.025351
sigma_v2	1.0469	.2583469			.5405491	1.553251

上表中部的 mu 表示条件期望 μ，其中“quality = 3”与“quality = 5”对应的虚拟变量均在 5% 水平上对条件期望有显著影响。

14.4 偶然断尾与样本选择

有时，被解释变量 y_i 的断尾与另一变量 z_i 有关。这被称为“偶然断尾”(incidental truncation)或“样本选择”(sample selection)[②]，称 z_i 为选择变量。

例 在美国的亚裔移民给人的整体印象是聪明能干。但在美国的亚裔并非亚洲人口的代表性样本。通常，只有受过高等教育或具有吃苦冒险精神的亚裔才会“自我选择”(self selection)移民。此时，决定移民与否的变量便对被解释变量产生了断尾作用。因此，“样本选择”将导致“选择性偏差”(selection bias)。

例 考察妇女的劳动力供给模型：

劳动时间方程 $\text{hours} = \alpha_0 + \alpha_1 \text{wage} + \alpha_2 \text{children} + \alpha_3 \text{marriage} + u$

工资方程 $\text{w}^o - \text{w}^r = \beta_0 + \beta_1 \text{age} + \beta_2 \text{education} + \beta_3 \text{children} + \beta_0 \text{location} + v$

其中，w^o 表示市场提供的工资(offered wage)，w^r 表示保留工资(reservation wage)。如果 $\text{w}^o -$

① 这是 Stata 11 新增的功能。在 Stata 10 或更早的版本，需手工生成这些虚拟变量。

② 也称为“Tobit Ⅱ”模型。

$w^r<0$，则该妇女将选择不工作，也就无法观测到她的劳动时间（hours），造成劳动时间方程的偶然断尾与样本选择问题。

为了纠正选择性偏差（selection bias），首先必须搞清样本选择的具体形式，即它究竟是怎么发生的，然后再进行针对性的校正。

考虑二维正态随机向量(y,z)，记期望为(μ_y,μ_z)，标准差为(σ_y,σ_z)，相关系数为ρ，联合密度函数为$f(y,z)$。假设决定个体是否有可能进入样本的“选择机制”（selection mechanism）为“选择变量z大于某常数c”。比如，在妇女劳动力供给的例子中，$z=w^o-w^r$，而$c=0$。显然，断尾后的联合分布为

$$f(y,z\mid z>c)=\frac{f(y,z)}{P(z>c)} \tag{14.24}$$

将上式中的z积分掉，可以证明偶然断尾y的条件期望为

$$E(y\mid z>c)=\mu_y+\rho\sigma_y\lambda[(c-\mu_z)/\sigma_z] \tag{14.25}$$

这个公式类似于普通断尾的公式(14.3)，其中$\lambda(\cdot)$为反米尔斯比率（IMR）函数。显然，如果$\rho=0$（y与z相互独立），则z的选择过程并不对y产生影响。如果$\rho>0$（即y与z正相关），则“$z>c$”偶然断尾的结果是把y的整个分布推向右边（因为$\lambda(\cdot)>0$），从而使得条件期望$E(y\mid z>c)$大于无条件期望$E(y)$。类似地，在“$z<c$”条件下，偶然断尾y的条件期望为

$$E(y\mid z<c)=\mu_y-\rho\sigma_y\lambda[(\mu_z-c)/\sigma_z] \tag{14.26}$$

假设回归模型为$y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i(i=1,2,\cdots,n)$。其中，被解释变量$y_i$是否可以观测取决于二值选择变量$z_i$（取值为0或1）

$$y_i=\begin{cases}\text{可观测，若} & z_i=1\\ \text{不可观测，若} & z_i=0\end{cases} \tag{14.27}$$

而决定二值变量z_i的方程为

$$z_i=\begin{cases}1, & \text{若} \quad z_i^*>0\\ 0, & \text{若} \quad z_i^*\leqslant 0\end{cases} \tag{14.28}$$

$$z_i^*=\boldsymbol{w}_i'\boldsymbol{\gamma}+u_i \tag{14.29}$$

其中，z_i^*为不可观测的潜变量。假设u_i服从正态分布，则z_i为Probit模型，故$P(z_i=1\mid\boldsymbol{w}_i)=\Phi(\boldsymbol{w}_i'\boldsymbol{\gamma})$。可观测样本的条件期望为

$$\begin{aligned}E(y_i\mid y_i\text{可观测})&=E(y_i\mid z_i^*>0)=E(\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\mid\boldsymbol{w}_i'\boldsymbol{\gamma}+u_i>0)\\&=E(\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\mid u_i>-\boldsymbol{w}_i'\boldsymbol{\gamma})=\boldsymbol{x}_i'\boldsymbol{\beta}+E(\varepsilon_i\mid u_i>-\boldsymbol{w}_i'\boldsymbol{\gamma})\\&=\boldsymbol{x}_i'\boldsymbol{\beta}+\rho\sigma_\varepsilon\lambda(-\boldsymbol{w}_i'\boldsymbol{\gamma})\quad\text{（使用偶然断尾的期望公式）}\end{aligned}$$

其中，$E(\varepsilon_i)=E(u_i)=0$，并将Probit扰动项的标准差$\sigma_u$标准化为1。显然，如果直接用OLS来估计样本数据，将遗漏非线性项$\rho\sigma_\varepsilon\lambda(-\boldsymbol{w}_i'\boldsymbol{\gamma})$。一般来说，$\boldsymbol{w}_i$与$\boldsymbol{x}_i$相关（比如，包含某个相同的变量），故OLS将导致不一致的估计，除非“$\rho=0$”（即y与z不相关）。考察解释变量x_{ik}变动的边际效应可知

$$\frac{\partial E(y_i\mid z_i^*>0)}{\partial x_{ik}}=\beta_k+\rho\sigma_\varepsilon\frac{\partial\lambda(-\boldsymbol{w}_i'\boldsymbol{\gamma})}{\partial x_{ik}} \tag{14.30}$$

其中，右边第一项为x_{ik}对y_i的直接影响，而第二项则是通过改变个体被选入样本的可能性而产生的间接影响（即选择性偏差）。

如果知道$\boldsymbol{\gamma}$，就知道$\lambda(-\boldsymbol{w}_i'\boldsymbol{\gamma})$，从而可以把它作为解释变量引入回归方程中。基于此，

Heckman(1979)提出“两步估计法”(two-step estimation),也称为“Heckit”①。

第一步:用 Probit 估计方程 $P(z_i=1|\boldsymbol{w})=\Phi(\boldsymbol{w}_i'\boldsymbol{\gamma})$,得到估计值 $\hat{\boldsymbol{\gamma}}$,计算 $\hat{\lambda}(-\boldsymbol{w}_i'\hat{\boldsymbol{\gamma}})$。

第二步:用 OLS 回归 $y_i \xrightarrow{\text{OLS}} \boldsymbol{x}_i, \hat{\lambda}_i$,得到估计值 $\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\sigma}_\varepsilon$。

另外一种更有效率的方法则是用 MLE 来估计该模型。在 Heckit 的两步估计法中,由于第一步的误差被带入第二步,故其效率不如 MLE 的整体估计。两步法的优点在于,操作简便;对于分布的假设也更弱(即使不假设二维正态分布,也可能成立)。

为了检验是否存在样本选择偏差(sample selection bias),可以使用似然比检验来检验原假设“$H_0:\rho=0$”。如果使用 Heckit,则无法进行此似然比检验。另外,样本选择模型也被称为“Type Ⅱ Tobit”。

样本选择模型的 Stata 命令为

```
heckman y x1 x2 x3,select(z1 z2)(默认使用 MLE,选择方程的被解释变量为 y)
heckman y x1 x2 x3,select(z1 z2) twostep(两步法,选择方程的被解释变量为 y)
heckman y x1 x2 x3,select(w=z1 z2)(默认使用 MLE,选择方程的被解释变量为 w)
```

需要注意的是,对于没有观测值的 y_i,在 Stata 中应将其设为“缺失”(missing)。Stata 将缺失的观测值记为“.”。执行以上命令后,Stata 将汇报原回归方程与样本选择方程的系数估计值,并检验原假设“$H_0:\rho=0$”。

下面再次以数据集 womenwk.dta 为例,但定义被解释变量为 lw(如果工作,则等于工资的对数;如果不工作,则为缺失值),进行如下“样本选择”MLE 回归:

```
. use womenwk.dta,clear
. heckman lw education age children, select (age married children education) nolog
```

Heckman selection model (regression model with sample selection)

Number of obs	=	2000			
Censored obs	=	657			
Uncensored obs	=	1343			
Wald chi2(3)	=	454.78			
Log likelihood = -1052.857		Prob > chi2	=	0.0000	

lw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lw						
education	.0397189	.0024525	16.20	0.000	.0349121	.0445256
age	.0075872	.0009748	7.78	0.000	.0056767	.0094977
children	-.0180477	.0064544	-2.80	0.005	-.0306981	-.0053973
_cons	2.305499	.0653024	35.30	0.000	2.177509	2.43349
select						
age	.0350233	.0042344	8.27	0.000	.0267241	.0433225
married	.4547724	.0735876	6.18	0.000	.3105434	.5990014
children	.4538372	.0288398	15.74	0.000	.3973122	.5103621
education	.0565136	.0110025	5.14	0.000	.0349492	.0780781
_cons	-2.478055	.1927823	-12.85	0.000	-2.855901	-2.100208
/athrho	.3377674	.1152251	2.93	0.003	.1119304	.5636045
/lnsigma	-1.375543	.0246873	-55.72	0.000	-1.423929	-1.327156
rho	.3254828	.1030183			.1114653	.5106469
sigma	.2527024	.0062385			.2407662	.2652304
lambda	.0822503	.0273475			.0286501	.1358505

LR test of indep. eqns. (rho = 0): chi2(1) = 5.53 Prob > chi2 = 0.0187

① Heckman 因此于 2000 年获诺贝尔经济学奖。

上表底部的似然比检验显示，可以拒绝原假设“H_0 ：$\rho=0$”，即应该使用样本选择模型。表中的“/athrho”指的是 ρ 的双曲反正切（hyperbolic arctangent）函数，即 $\text{arctanh}(\rho)\equiv\frac{1}{2}\ln[(1+\rho)/(1-\rho)]$。由于相关系数 ρ 必须满足参数约束“$|\rho|<1$”，这给 MLE 的最大化问题带来不便。对其进行参数变换后（re-parameterization），新参数 $\text{arctanh}(\rho)$ 不再受任何约束。因此，可以先估计 $\text{arctanh}(\rho)$，然后根据 MLE 的不变性，反推出 $\hat{\rho}$。同理，“/lnsigma”指的是 $\ln\sigma_\varepsilon$。由于标准差 σ_ε 必须满足 $\sigma_\varepsilon>0$ 的参数约束，故在参数变换后，首先估计 $\ln\sigma_\varepsilon$，再反推出 $\hat{\sigma}_\varepsilon$。

也可以使用 Heckit 的两步估计法：

```
. heckman lw education age children, select (age married children education) twostep nolog
```

Heckman selection model -- two-step estimates (regression model with sample selection)

Number of obs	=	2000				
Censored obs	=	657				
Uncensored obs	=	1343				
Wald chi2(3)	=	405.68				
Prob > chi2	=	0.0000				

lw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lw						
education	.0427067	.003106	13.75	0.000	.0366191	.0487944
age	.009322	.0014343	6.50	0.000	.0065108	.0121333
children	-.0019549	.0115202	-0.17	0.865	-.0245341	.0206242
_cons	2.124787	.1249789	17.00	0.000	1.879833	2.369741
select						
age	.0347211	.0042293	8.21	0.000	.0264318	.0430105
married	.4308575	.074208	5.81	0.000	.2854125	.5763025
children	.4473249	.0287417	15.56	0.000	.3909922	.5036576
education	.0583645	.0109742	5.32	0.000	.0368555	.0798735
_cons	-2.467365	.1925635	-12.81	0.000	-2.844782	-2.089948
mills						
lambda	.1822815	.0638285	2.86	0.004	.05718	.307383
rho	0.66698					
sigma	.27329216					

对比以上两表可知，这两种方法的估计结果很接近。

14.5 归并回归

受限被解释变量的另一种情形是，对于线性模型 $y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i$，可能当 $y_i\geqslant c$（或 $y_i\leqslant c$）时，所有 y_i 都被归并为 c。这种数据被称为“归并数据”（censored data）。

例 （上不封顶的数据，top coding）在问卷调查中，常有诸如“收入在 ¥50 000 及以上”这样的选项。对于这部分数据，我们只知道 $y_i\geqslant 50\,000$，而不知道 y_i 的具体取值。

例 （边角解，corner solution）考虑买车的决定，并考察“买车开支”这个变量。如果不买车，则“买车开支”的最优解为边角解，即买车开支为 0；反之，如果买车，则买车开支一定为正数。

例 （边角解）考察“劳动时间”这个变量。对于失业或待业者而言，“劳动时间”的最优解

为边角解，即劳动时间为0；而就业者的劳动时间一定为正数。

归并回归（censored regression）与断尾回归不同的是，虽然有全部的观测数据，但对于某些观测数据，被解释变量 y_i 被压缩在一个点上了。此时，y_i 的概率分布就变成由一个离散点与一个连续分布所组成的混合分布（mixed distribution）。下面将说明，在这种情况下，如果用 OLS 来估计，无论使用的是整个样本，还是去掉离散点后的子样本，都不能得到一致的估计。

假设 $y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$（$y_i^*$ 不可观测），扰动项 $\varepsilon_i | \boldsymbol{x}_i \sim N(0, \sigma^2)$。不失一般性，为简便起见，假定归并点为 $c = 0$。假设可以观测到 $y_i = \begin{cases} y_i^*, & \text{若 } y_i^* > 0, \\ 0, & \text{若 } y_i^* \leqslant 0 \end{cases}$ 下面分别计算子样本的条件期望 $\mathrm{E}(y_i | \boldsymbol{x}_i; y_i > 0)$，以及整个样本的条件期望 $\mathrm{E}(y_i | \boldsymbol{x}_i)$。

对于满足条件“$y_i > 0$”的子样本，

$$
\begin{aligned}
\mathrm{E}(y_i | \boldsymbol{x}_i; y_i > 0) &= \mathrm{E}(y_i^* | \boldsymbol{x}_i; y_i > 0) \quad \text{（给定 } y_i > 0, \text{必然 } y_i = y_i^*) \\
&= \mathrm{E}(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i | \boldsymbol{x}_i; y_i^* > 0) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i | \boldsymbol{x}_i; \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i | \boldsymbol{x}_i; \varepsilon_i > -\boldsymbol{x}_i'\boldsymbol{\beta}) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \lambda(-\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) \quad \text{（使用正态分布的断尾公式，且 } \mathrm{E}(\varepsilon_i) = 0)
\end{aligned}
$$

因此，在使用子样本进行回归时，由于忽略了非线性项 $\sigma \cdot \lambda(-\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$，被纳入扰动项中，导致扰动项与解释变量 $\boldsymbol{x}_i$ 相关，故 OLS 估计是不一致的。对于整个样本，

$$
\begin{aligned}
\mathrm{E}(y_i | \boldsymbol{x}_i) &= 0 \cdot \mathrm{P}(y_i = 0 | \boldsymbol{x}_i) + \mathrm{E}(y_i | \boldsymbol{x}_i; y_i > 0) \cdot \mathrm{P}(y_i > 0 | \boldsymbol{x}_i) \\
&= \mathrm{E}(y_i | \boldsymbol{x}_i; y_i > 0) \cdot \mathrm{P}(y_i > 0 | \boldsymbol{x}_i)
\end{aligned} \tag{14.31}
$$

其中，$\mathrm{P}(y_i > 0 | \boldsymbol{x}_i) = \mathrm{P}(y_i^* > 0 | \boldsymbol{x}_i) = \mathrm{P}(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0 | \boldsymbol{x}_i)$

$$
\begin{aligned}
&= \mathrm{P}(\varepsilon_i > -\boldsymbol{x}_i'\boldsymbol{\beta} | \boldsymbol{x}_i) = \mathrm{P}\left(\frac{\varepsilon_i}{\sigma} > \frac{-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \middle| \boldsymbol{x}_i \right) \\
&= 1 - \Phi(-\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) \quad \text{（标准正态以原点为对称）}
\end{aligned}
$$

因此，$\mathrm{E}(y_i | \boldsymbol{x}_i) = \mathrm{E}(y_i | \boldsymbol{x}_i, y_i > 0) \cdot \mathrm{P}(y_i > 0 | \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)[\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma \cdot \lambda(-\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)]$ 是解释变量 $\boldsymbol{x}_i$ 的非线性函数。如果使用 OLS 对整个样本进行线性回归，其非线性项将被纳入扰动项中，导致不一致的估计。

Tobin（1958）提出用 MLE 估计这个模型，该方法因此被称为“Tobit”，也称为“Type Ⅰ Tobit”或“归并回归”（censored regression）。在归并数据的情况下，$y_i > 0$ 时的概率密度依然不变，仍为 $\frac{1}{\sigma}\phi[(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma]$，$\forall\, y_i > 0$。而 $y_i \leqslant 0$ 时的分布却被挤到一个点“$y_i = 0$”上了，即 $\mathrm{P}(y_i = 0 | \boldsymbol{x}) = 1 - \mathrm{P}(y_i > 0 | \boldsymbol{x}) = 1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)$。因此，该混合分布的概率密度函数可以写为

$$
f(y_i | \boldsymbol{x}) = [1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)]^{\mathbf{1}(y_i = 0)} \left[\frac{1}{\sigma}\phi((y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma) \right]^{\mathbf{1}(y_i > 0)} \tag{14.32}
$$

其中，$\mathbf{1}(\cdot)$ 为示性函数（indicator function），即如果括号里的表达式为真，取值为1；反之，取值为0。因此，可以写出整个样本的似然函数，然后使用 MLE 来估计。

归并回归的 Stata 命令为

```
tobit y x1 x2 x3,ll(#) ul(#)
```

其中，选择项“ll(#)”表示 lower limit，即左归并；选择项“ul(#)”表示 upper limit，即右归并；如果同时使用这两个选择项，则表示双边归并。如果把选择项“ll(#)”写为“ll”，则 Stata 将把样本中 y 的最小值作为左归并点。类似地，如果把选择项“ul(#)”写为“ul”，则 Stata 将把样本中

y 的最大值作为右归并点。

Tobit 模型的一个缺陷是对分布的依赖性很强，不够稳健。如果似然函数不正确，比如扰动项不服从正态分布或存在异方差，则 QMLE 估计就不一致。直观来看，这是因为，从方程(14.32)导出的 MLE 一阶条件是复杂的非线性函数。为此，在使用 Tobit 模型时，需要检验其正态性与同方差性。然而，由于这是一个非线性模型，故不能直接使用第 6 章的 JB 检验或第 7 章的异方差检验来检验正态性与同方差性。

为了检验正态性，可将 Tobit 模型的 MLE 一阶条件视为某种矩条件，并基于此进行“条件矩检验”(conditional moment test)(Skeels and Vella,1999)。但 Drukker(2002)发现，条件矩统计量的真实分布与渐近分布有相当差距，存在较严重的显著性水平扭曲(size distortion)，故使用“参数自助法”(parametric bootstrap)来获得校正的临界值(有关显著性水平扭曲与自助法，参见第 19 章)。条件矩检验可通过非官方命令 tobcm 来实现；但该命令仅适用于左归并(left-censored)，且归并点为 0 的情形(这也是最常见的情形)。

```
net install tobcm.pkg①  (下载安装命令 tobcm)
tobcm,pbs  (对正态性假设进行条件矩检验)
```

其中，选择项“pbs”表示使用参数自助法获得临界值。

为了检验同方差的原假设“$H_0:\sigma_i^2=\sigma^2$”，考虑替代假设“$H_1:\sigma_i^2=\exp(\boldsymbol{z}_i'\boldsymbol{\alpha})$”，其中 $\boldsymbol{z}_i$ 通常等于解释变量 $\boldsymbol{x}_i$(但也可以不同)。然后，可以通过辅助回归构建 LM 统计量来检验 $\boldsymbol{\alpha}=\mathbf{0}$(类似于第 6 章第 7 节介绍的方法)。具体步骤及 Stata 操作，参见 Cameron and Trivedi(2010,p.550)。

如果发现扰动项并不服从正态分布或存在异方差，解决方法之一为使用更稳健的“归并最小绝对离差法”(Censored Least Absolute Deviations，简记 CLAD)(Powell,1984)。其他解决方法包括上节的“样本选择模型”，以及下节的“两部分模型”。

CLAD 法仅要求扰动项为 iid，即使在非正态与异方差的情况下也能得到一致估计。而且，在一定的正则条件下，估计量服从渐近正态分布。首先，将归并数据模型简洁地写为

$$y_i=\max(0,\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i) \tag{14.33}$$

即如果 $\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i\geqslant 0$，则 $y_i=\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i$；反之，$y_i=0$。CLAD 法的目标函数为离差绝对值之和：

$$\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}|y_i-\max(0,\boldsymbol{x}_i'\boldsymbol{\beta})| \tag{14.34}$$

选择 $\boldsymbol{\beta}$ 使得离差绝对值之和最小化，即可得到 CLAD 估计量 $\hat{\boldsymbol{\beta}}_{CLAD}$，可通过非官方 Stata 命令 clad 来实现。这实际上是中位数回归，为分位数回归的特例(参见第 26 章)。$\hat{\boldsymbol{\beta}}_{CLAD}$ 的协方差矩阵可通过自助法得到(参见第 19 章)。

```
net install sg153.pkg②  (下载安装命令 clad)
clad y x1 x2 x3,ll(#)ul(#)  (进行 CLAD 估计)
```

其中，选择项“ll(#)”与“ul(#)”分别表示左归并点与右归并点。

下面以美国妇女就业数据集 womenwk.dta 为例③，估计一个工资决定模型。被解释变量为 lwf(equals log of wage if working and 0 if not working，如果工作，则等于工资的对数；如果不工作，则等于 0)。解释变量为 age(年龄)，married(婚否)，children(子女数)，与 education(教育年限)。

作为对照，首先对整个样本进行 OLS 回归：

---

① 或输入命令“findit tobcm”寻找下载地址。

② 或输入命令“findit clad”寻找下载地址。

③ 此数据集曾用于第 11 章二值模型的估计。

```
. use womenwk.dta,clear
. reg lwf age married children education,r
. estimates store ols
```

```
Linear regression                               Number of obs =    2000
                                                F(  4,  1995) =  197.56
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2120
                                                Root MSE      =  1.3218
------------------------------------------------------------------------------
             |               Robust
         lwf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0363624   .0038582     9.42   0.000     .0287959    .0439288
     married |   .3188214   .0710831     4.49   0.000     .1794164    .4582263
    children |   .3305009   .0181054    18.25   0.000     .2949935    .3660084
   education |   .0843345    .010691     7.89   0.000     .0633677    .1053012
       _cons |  -1.077738   .1659689    -6.49   0.000    -1.403228   -.7522473
------------------------------------------------------------------------------
```

其次，画被解释变量 lwf 的直方图，结果如图 14.4。

```
. hist lwf
```

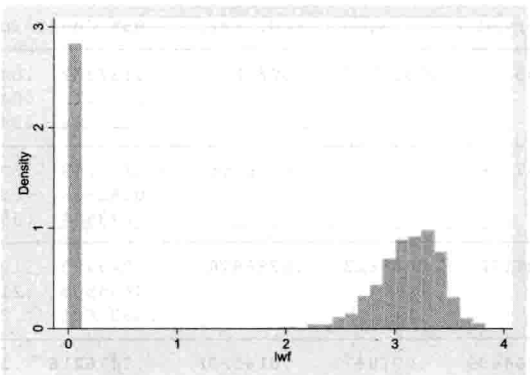


图 14.4 被解释变量 lwf 的直方图

图 14.4 显示，在相当多的观测点上，被解释变量 lwf 的取值为 0。为此，进行归并回归，假设左归并点为 0：

```
. tobit lwf age married children education,ll(0)
. estimates store tobit
```

```
Tobit regression                                Number of obs     =       2000
                                                LR chi2(4)        =     461.85
                                                Prob > chi2       =     0.0000
Log likelihood = -3349.9685                     Pseudo R2         =     0.0645
------------------------------------------------------------------------------
         lwf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .052157   .0057457     9.08   0.000     .0408888    .0634252
     married |   .4841801   .1035188     4.68   0.000     .2811639    .6871964
    children |   .4860021   .0317054    15.33   0.000     .4238229    .5481812
   education |   .1149492   .0150913     7.62   0.000     .0853529    .1445454
       _cons |  -2.807696   .2632565   -10.67   0.000    -3.323982   -2.291409
-------------+----------------------------------------------------------------
      /sigma |   1.872811    .040014                      1.794337    1.951285
------------------------------------------------------------------------------
  Obs. summary:        657  left-censored observations at lwf<=0
                      1343     uncensored observations
                         0 right-censored observations
```

下面,使用命令 tobcm 检验扰动项是否服从正态分布。

```
. tobcm,p
```

```
Conditional moment test against the null of normal errors

                    critical values
   CM         %10         %5          %1
 1370.2    5.16182   7.4033384  16.871058
```

结果显示,条件矩统计量高达 1 370.2,故强烈拒绝“扰动项服从正态分布”的原假设。为此,考虑在非正态情况下也成立的 CLAD 估计法。

```
. clad lwf age married children education,ll(0)
. estimates store clad
```

```
Initial sample size = 2000
Final sample size = 2000
Pseudo R2 = .06322683

Bootstrap statistics

Variable |  Reps   Observed      Bias    Std. Err.   [95% Conf. Interval]
---------+--------------------------------------------------------------
     age |   100   .0268424   .0012875   .0066561    .0136352   .0400496  (N)
         |                                .017413    .0448288  (P)
         |                                .016788    .0444136  (BC)
---------+--------------------------------------------------------------
 married |   100   .1567166   .0158871   .0876595   -.0172189   .3306521  (N)
         |                               .0442563    .355791  (P)
         |                                .043674   .3516315  (BC)
---------+--------------------------------------------------------------
children |   100   .1206152   .0045623   .0388678     .043493   .1977374  (N)
         |                               .0668565   .2165551  (P)
         |                               .0631661    .215001  (BC)
---------+--------------------------------------------------------------
education|   100   .0764499   .0018672   .0145288    .0476216   .1052782  (N)
         |                               .0554966   .1169837  (P)
         |                               .0605122   .1173066  (BC)
---------+--------------------------------------------------------------
   const |   100   .4908203  -.1189834   .5931536   -.6861252   1.667766  (N)
         |                              -1.09028    1.235906  (P)
         |                               -.978726   1.336596  (BC)
---------+--------------------------------------------------------------
                          N = normal, P = percentile, BC = bias-corrected
```

下面将 OLS,Tobit 与 CLAD 的估计结果列表比较。

```
. esttab ols tobit clad,mtitles se star(*0.1 **0.05 ***0.01)
```

	(1) ols	(2) tobit	(3) clad
main			
age	0.0364***	0.0522***	0.0356***
	(0.00386)	(0.00575)	(0.00364)
married	0.319***	0.484***	0.175***
	(0.0711)	(0.104)	(0.0650)
children	0.331***	0.486***	0.139***
	(0.0181)	(0.0317)	(0.0199)
education	0.0843***	0.115***	0.0879***
	(0.0107)	(0.0151)	(0.00959)
_cons	-1.078***	-2.808***	-0.122
	(0.166)	(0.263)	(0.160)
sigma			
_cons		1.873***	
		(0.0400)	
N	2000	2000	2000

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

从上表可知,对于变量 married 与 children 而言,CLAD 估计值与 Tobit 估计值相差较大,尽管这两种估计法在变量显著性上没有质的不同。Wooldridge(2010,p. 689)指出,如果 Tobit 模型的设定正确(比如,满足正态性与同方差性),则 CLAD 与 Tobit 的估计结果应相差不多。从这个角度来看,也可将 CLAD 的估计结果大致视为对 Tobit 模型的设定检验(specification test)。在本例中,由于 CLAD 估计值与 Tobit 相差较大,故可认为 Tobit 模型的设定有误,并倾向于使用 CLAD 模型。

14.6 归并数据的两部分模型

上节提到,如果归并数据的扰动项不服从正态分布或存在异方差,则 Tobit MLE 估计是不一致的。Tobit 模型的特征是,被解释变量为 0 或正数的产生机制是一样的;但此假定未必成立。正如第 13 章零膨胀泊松回归的情形,有时可将决策分为两个阶段,并允许两个阶段拥有不同的决定机制。比如,在第一阶段决定是否参加工作(或购买耐用消费品),称为"参与决策"(participation decision);而在第二阶段决定工作时间(或购买耐用消费品的开支),称为"数量决策"(amount decision)。

为此,作为 Tobit 模型的推广,Cragg(1971)提出了以下"两部分模型"(two-part model)。首先,定义虚拟变量 d,将样本中的活动参与者(比如,工作者或购买者)记为 $d=1$,非参与者记为 $d=0$。对于参与者,可以观测到 $y>0$;而对于非参与者,只能观测到 $y=0$。对于非参与者而言,只能估计 $P(d=0)$,即"$d=0$"的概率。对于参与者而言,给定 $d=1$ 的条件密度可写为 $f(y|d=1)=P(d=1|\boldsymbol{x})f(y|d=1,\boldsymbol{x})$。因此,关于 y 的两部分模型为:

$$f(y|x)=\begin{cases}\mathrm{P}(d=0|\boldsymbol{x}) & \text{若} \quad y=0\\ \mathrm{P}(d=1|\boldsymbol{x})f(y|d=1,\boldsymbol{x}) & \text{若} \quad y>0\end{cases} \tag{14.35}$$

其中，为了保证参与者的 y 为正数，$f(y|d=1,\boldsymbol{x})$ 应为某取值为正的随机变量的密度函数，比如对数正态分布(log-normal)。为此，在实践中常把 $\ln y$ 作为第二部分的被解释变量。显然，对于第一阶段的参与决策 d，可以用 Probit 或 Logit 来建模。从潜变量的角度来看，如果 $\boldsymbol{x}_i'\boldsymbol{\beta}+\varepsilon_i$ 超过了 0，则 $d=1$。因此，这个模型也称为“跨栏模型”(hurdle model)，因为只有跨过这个栏(hurdle)或门槛，才能参与活动。

对于两部分模型，通常假设其两个部分相互独立，故可分别进行估计(如果不相互独立，则应使用本章第 4 节的样本选择模型)。对于第一部分的二值选择模型，可以使用全样本进行 Probit 或 Logit 估计；对于第二部分的模型，则可使用由参与者组成的子样本进行 OLS 估计。通常，第一部分与第二部分模型所使用的解释变量完全相同；但也可以不同(这正是两部分模型的灵活性)。

继续以上节的美国妇女就业数据集为例。

```
. use womenwk.dta,clear
```

对于第一部分模型，首先生成一个被解释的虚拟变量，然后使用全样本数据进行 Probit 回归。

```
. gen lwd = (lwf >0)
. probit lwd age married children education,nolog
```

```
Probit regression                                 Number of obs   =       2000
                                                  LR chi2(4)      =     478.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -1027.0616                       Pseudo R2       =     0.1889

------------------------------------------------------------------------------
         lwd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0347211   .0042293     8.21   0.000     .0264318    .0430105
     married |   .4308575    .074208     5.81   0.000     .2854125    .5763025
    children |   .4473249   .0287417    15.56   0.000     .3909922    .5036576
   education |   .0583645   .0109742     5.32   0.000     .0368555    .0798735
       _cons |  -2.467365   .1925635   -12.81   0.000    -2.844782   -2.089948
------------------------------------------------------------------------------
```

上表显示，各解释变量对于参加工作的概率都有显著的正影响。对于第二部分模型，则仅使用参加工作者的子样本进行 OLS 线性回归。

```
. reg lwf age married children education if lwd = =1,r
```

```
Linear regression                                 Number of obs   =       1343
                                                  F(  4,  1338)   =     123.87
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.2570
                                                  Root MSE        =     .24733

------------------------------------------------------------------------------
             |               Robust
         lwf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0066617    .000873     7.63   0.000     .0049492    .0083743
     married |  -.0298994    .016201    -1.85   0.065    -.0616815    .0018826
    children |  -.0321762   .0048209    -6.67   0.000    -.0416336   -.0227188
   education |   .0380189   .0021444    17.73   0.000     .0338121    .0422257
       _cons |   2.448392   .0426429    57.42   0.000     2.364738    2.532046
------------------------------------------------------------------------------
```

上表显示，两部分模型与 Tobit 模型估计结果的最大不同是，变量 married 与 children 的系数

显著地为负(前者在10%水平上显著)。这表明,虽然已婚与孩子数量能提高进入就业市场的概率,但却会降低进入就业市场后的工资水平。在两部分模型中,是否依然存在非正态扰动项或异方差?由于这是线性回归模型,故可进行常规的异方差检验。

```
. qui reg lwf age married children education if lwd = =1
. estat imtest,white
```

```
White's test for Ho: homoskedasticity
        against Ha: unrestricted heteroskedasticity

        chi2(13)      =     45.78
        Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	45.78	13	0.0000
Skewness	37.56	4	0.0000
Kurtosis	6.75	1	0.0094
Total	90.09	18	0.0000

上表显示,BP检验的结果强烈拒绝同方差的原假设,即存在异方差。为了检验扰动项的正态性,首先生成残差。

```
. predict lwf_resid,resid
. sktest lwf_resid
```

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
lwf_resid	2.0e+03	0.0000	.	.	.

结果也强烈地拒绝了正态性的原假设。更直观地,可以画残差的核密度图,结果见图14.5。

```
. kdensity lwf_resid
```

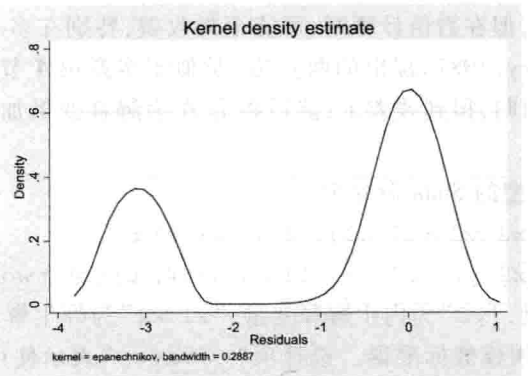


图14.5 两部分模型残差的核密度图

从图14.5可以看出,扰动项的分布明显地存在双峰,与正态分布相去甚远。然而,对于线性回归模型,非正态性与异方差并不影响估计量的一致性与渐近正态性,只要使用异方差稳健的标

准误即可。

14.7 含内生解释变量的 Tobit 模型(选读)

对于归并数据使用 Tobit 模型,有时会遇到解释变量为内生变量的情形。不失一般性,假设在 0 处存在左边归并。考虑以下模型:

$$y_{1i}^{*} = \boldsymbol{x}_i'\boldsymbol{\delta} + \beta y_{2i} + u_i \tag{14.36}$$

$$y_{2i} = \boldsymbol{x}_i'\boldsymbol{\gamma}_1 + \boldsymbol{z}_i'\boldsymbol{\gamma}_2 + v_i \tag{14.37}$$

$$y_{1i} = \begin{cases} y_1^{*} & \text{若 } y_1^{*} > 0 \\ 0 & \text{若 } y_1^{*} = 0 \end{cases} \tag{14.38}$$

其中,y_{1i}为可观测的被解释变量,y_{1i}^{*}为不可观测的潜变量,y_{2i}是模型中唯一的内生解释变量(以下方法可推广到多个内生解释变量的情形)。含内生变量 Tobit 模型的估计方法类似于含内生变量的 Probit 模型(Ⅳ Probit,参见第 11 章)。假设扰动项(u_i, v_i)服从期望为 0 的二维正态分布:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha\sigma_v \\ \alpha\sigma_v & \sigma_v^2 \end{pmatrix}\right] \tag{14.39}$$

其中,u_i 的方差被标准化为 1,而 α 为(u_i, v_i)的相关系数。显然,由于 v_i 服从正态分布,故 y_{2i}也服从正态分布,因此 y_{2i}必须是一个连续变量。进一步,假设(u_i, v_i)独立于$\boldsymbol{x}_i$ 与$\boldsymbol{z}_i$,故在方程(14.36)中,$\boldsymbol{x}_i$ 为外生解释变量。而且,$\boldsymbol{z}_i$ 可以作为方程(14.36)中内生变量 y_{2i}的工具变量,因为 $\boldsymbol{z}_i$ 与内生变量 y_{2i}相关(参见方程(14.37)),并且$\boldsymbol{z}_i$ 与 u_i 无关。在此模型中,y_{2i}的内生性完全来自于 u_i 与 v_i 的相关性;如果二者的相关系数 $\alpha = 0$(Stata 记为 alpha),则 y_{2i}为外生变量。因此,对于 y_{2i}内生性的检验可通过检验"$H_0 : \alpha = 0$"来进行。

对于由方程(14.36) ~ (14.39)构成的模型,在给定$\boldsymbol{x}_i$ 与$\boldsymbol{z}_i$ 的情况下,(y_{1i}, y_{2i})的条件概率分布已完全确定。将联合概率密度 $f(y_{1i}, y_{2i} | \boldsymbol{x}_i, \boldsymbol{z}_i)$分解为 $f(y_{1i} | y_{2i}, \boldsymbol{x}_i, \boldsymbol{z}_i) f(y_{2i} | \boldsymbol{x}_i, \boldsymbol{z}_i)$,则可写出样本数据$(y_{1i}, y_{2i})$的似然函数,进行最有效率的 MLE 估计。这种方法称为"工具变量 Tobit"(Instrumental Variable Tobit,简记Ⅳ Tobit)。

尽管 MLE 最有效率,但在数值计算时,可能不易收敛,特别在多个内生解释变量的情形下。为此,可以考虑使用 Newey(1987)提出的两步法(类似于本章第 4 节介绍的两步法),即首先对方程(14.37)进行 OLS 回归,得到残差 $\hat{v}_i$;然后将 $\hat{v}_i$ 作为解释变量加入到方程(14.36)中,进行 Tobit 估计。

含内生变量 Tobit 模型的 Stata 命令为

```
. ivtobit y1 x1 x2 (y2 = z1 z2), ll(#) ul(#) r
. ivtobit y1 x1 x2 (y2 = z1 z2), ll(#) ul(#) first twostep
```

其中,"y1"为被解释变量,"y2"为内生解释变量,"x1 x2"为外生解释变量,而"z1 z2"为工具变量;选择项"r"表示使用稳健标准误。选择项"twostep"表示使用两步法,而默认进行 MLE 估计;选择项"first"表示显示第一步回归的结果。如果把选择项"ll(#)"写为"ll",则 Stata 将把样本中"y1"的最小值作为左归并点。类似地,如果把选择项"ul(#)"写为"ul",则 Stata 将把样本中"y1"的最大值作为右归并点。

下面,以 Stata 提供的数据集 laborsup.dta 为例。被解释变量为 fem_inc(妇女收入),解释变

量包括 fem_educ(妇女教育水平),kids(孩子数量),以及 other_inc(其他家庭收入)。假设如果妇女不工作,则可领取 $10 000 福利与儿童赡养费。因此,妇女收入肯定大于或等于 $10 000,故在 $10 000 处存在左边归并。另一方面,other_inc(其他家庭收入)可能为内生变量,我们使用 male_inc(丈夫收入)作为工具变量。

```
. use laborsup.dta
```

首先,进行Ⅳ Tobit 估计。

```
. ivtobit fem_inc fem_educ kids (other_inc = male_educ),ll nolog
```

```
Tobit model with endogenous regressors          Number of obs   =        500
                                                Wald chi2(3)    =     117.42
Log likelihood = -3226.0845                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   other_inc |  -.9045399   .1329762    -6.80   0.000    -1.165168   -.6439114
    fem_educ |   3.272391   .3968708     8.25   0.000     2.494538    4.050243
        kids |  -3.312357   .7218628    -4.59   0.000    -4.727182   -1.897532
       _cons |   19.24735   7.372391     2.61   0.009     4.797725    33.69697
-------------+----------------------------------------------------------------
      /alpha |   .2907654   .1379965     2.11   0.035     .0202972    .5612336
        /lns |   2.874031   .0506672    56.72   0.000     2.774725    2.973337
        /lnv |   2.813383   .0316228    88.97   0.000     2.751404    2.875363
-------------+----------------------------------------------------------------
           s |   17.70826    .897228                      16.03422    19.55707
           v |   16.66621   .5270318                      15.66461    17.73186
------------------------------------------------------------------------------
Instrumented:  other_inc
Instruments:   fem_educ kids male_educ
------------------------------------------------------------------------------
Wald test of exogeneity (/alpha = 0): chi2(1) =     4.44  Prob > chi2 = 0.0351

  Obs. summary:        272  left-censored observations at fem_inc<=10
                       228     uncensored observations
                         0 right-censored observations
```

上表底部的沃尔德检验结果表明,可以拒绝"$\alpha=0$"的外生性原假设,即认为存在内生变量。其次,进行两步法估计,并显示第一步回归的结果。

```
. ivtobit fem_inc fem_educ kids (other_inc = male_educ), ll
first twostep
```

```
First-stage regression

------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  3,   496) =   34.36
       Model |  28864.2732     3  9621.42439           Prob > F      =  0.0000
    Residual |  138881.269   496  280.002558           R-squared     =  0.1721
-------------+------------------------------           Adj R-squared =  0.1671
       Total |  167745.542   499  336.163411           Root MSE      =  16.733

------------------------------------------------------------------------------
   other_inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  male_educ |   2.845253   .2838838    10.02   0.000      2.28749    3.403016
   fem_educ |   .3351866   .2837344     1.18   0.238    -.2222829    .8926562
       kids |   .8329056   .5497701     1.52   0.130    -.2472597    1.913071
      _cons |   9.872562   5.049432     1.96   0.051    -.0483506    19.79347
------------------------------------------------------------------------------
```

在第一步回归中，变量 male_educ 的系数显著为正，整个方程的 F 值为 34.36，故不存在弱工具变量。

```
Two-step tobit with endogenous regressors          Number of obs   =       500
                                                   Wald chi2(3)    =    117.38
                                                   Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   other_inc |  -.9045397   .1330015    -6.80   0.000    -1.165218   -.6438616
    fem_educ |    3.27239   .3969399     8.24   0.000     2.494402    4.050378
        kids |  -3.312356   .7220066    -4.59   0.000    -4.727463   -1.897249
       _cons |   19.24735    7.37392     2.61   0.009     4.794728    33.69997
------------------------------------------------------------------------------
Instrumented:  other_inc
Instruments:   fem_educ kids male_educ
------------------------------------------------------------------------------
Wald test of exogeneity:     chi2(1) =     4.64           Prob > chi2 = 0.0312

  Obs. summary:         272  left-censored observations at fem_inc<=10
                        228     uncensored observations
                          0 right-censored observations
```

上表底部的 Wald 检验结果再次拒绝了外生性的原假设。经目测可知，两步法的估计值与 IV Tobit 十分接近。

习　题

14.1 数据集 tobacco.dta① 包含了 1995 年 2 724 个比利时家庭的以下变量：share2（budget share of tobacco，烟草占预算比重），age（年龄），ln x（家庭开支的对数），nadults（number of adults in household，家庭中成年人数），nkids（number of kids > 2 year old，两岁以上儿童数），nkids2（number of kids ≤ 2 year old，两岁及以下儿童数）。考虑以下恩格尔函数（用于烟草的开支比重随着总开支而变化）：

$$\text{share2}_i = \beta_0 + \beta_1 \ln x_i + \beta_2\ \text{age}_i + \beta_3\ \text{nadults}_i + \beta_4\ \text{nkids}_i + \beta_5 \text{nkids2}_i + \varepsilon_i$$

（1）进行 OLS 回归；

（2）假设在被解释变量“share2 = 0”处存在左边断尾，进行断尾回归；

（3）假设在被解释变量“share2 = 0”处存在左边归并，进行归并回归（Tobit）；

（4）将被解释变量“share2 = 0”的取值定义为“缺失”，然后用 MLE 进行样本选择回归，选择方程的解释变量为“age”，“nadults”，“nkids”与“nkids2”（选择是否抽烟，可能不取决于收入，而取决于对自己及家人健康的考虑）。

（5）恩格尔定律适用于烟草吗（即用于烟草的开支比重随着总开支的增加而下降）？

14.2 重复上题，但把被解释变量改为“share1”（budget share of alcohol，酒精占预算比重）。

附　录

标准正态的断尾期望公式

证明：假设 $y \sim N(0,1)$，c 为任意常数，则

① 此数据集来自 Verbeek（2004）。

$$
\begin{aligned}
E(y \mid y > c) &= \int_c^{\infty} y \frac{\phi(y)}{1-\Phi(c)} dy \\
&= \frac{1}{1-\Phi(c)} \int_c^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&= \frac{1}{1-\Phi(c)} \frac{1}{\sqrt{2\pi}} \int_c^{\infty} e^{-y^2/2} d(y^2/2) \\
&= \frac{1}{1-\Phi(c)} \frac{1}{\sqrt{2\pi}} \int_{c^2/2}^{\infty} e^{-z} dz \\
&= \frac{1}{1-\Phi(c)} \left[\frac{1}{\sqrt{2\pi}} e^{-c^2/2} \right] \\
&= \frac{\phi(c)}{1-\Phi(c)}
\end{aligned}
$$

在上面的推导中，使用了积分变换 $z \equiv y^2/2$。

第15章 短 面 板

15.1 面板数据的特点

面板数据(panel data 或 longitudinal data,也译为“平行数据”),指的是在一段时间内跟踪同一组个体(individual)的数据。它既有横截面的维度(n 位个体),又有时间维度(T 个时期)。一个 $T=3$ 的面板数据结构如表 15.1[①]。

表 15.1 面板数据的结构(长形排列)

	y	x_1	x_2	x_3
Individual 1: $t=1$				
Individual 1: $t=2$				
Individual 1: $t=3$				
……				
Individual$n$: $t=1$				
Individual$n$: $t=2$				
Individual$n$: $t=3$				

通常的面板数据 T 较小,而 n 较大,在使用大样本理论时让 n 趋于无穷大。这种面板数据被称为“短面板”(short panel)。反之,如果 T 较大,而 n 较小,则被称为“长面板”(long panel)。

在面板模型中,如果解释变量包含被解释变量的滞后值,则称为“动态面板”(dynamic panel);反之,则称为“静态面板”(static panel)。本章关注静态的短面板,第 16 章讨论长面板与动态面板,第 17 章考虑非线性面板。

如果在面板数据中,每个时期在样本中的个体完全一样,则称为“平衡面板数据”(balanced panel);反之,则称为“非平衡面板数据”(unbalanced panel)。我们主要关注平衡面板,但将在本章第 10 节讨论非平衡面板。

面板数据的主要优点如下。

(1) 可以解决遗漏变量问题:遗漏变量偏差是一个普遍存在的问题。虽然可以用工具变量法解决,但有效的工具变量常常很难找。遗漏变量常常是由于不可观测的个体差异或“异质性”(heterogeneity)造成的,如果这种个体差异“不随时间而改变”(time invariant),则面板数据提供了解决遗漏变量问题的又一利器。

(2) 提供更多个体动态行为的信息:由于面板数据同时有横截面与时间两个维度,有时它可以解决单独的截面数据或时间序列数据所不能解决的问题。比如,考虑如何区分规模效应与

① 也可将数据排成“宽形”(wide form),参见第 23 章。一般以“长形”(long form)排列最为常见。

技术进步对企业生产效率的影响。对于截面数据来说,由于没有时间维度,故无法观测到技术进步。然而,对于单个企业的时间序列数据来说,我们无法区分其生产效率的提高究竟有多少是由于规模扩大,有多少是由于技术进步。又比如,对于失业问题,截面数据能告诉我们在某个时点上哪些人失业,而时间序列数据能告诉我们某个人就业与失业的历史,但这两种数据均无法告诉我们是否失业的总是同一批人(意味着低流转率,low turnover rate),还是失业的人群总在变动(意味着高流转率,high turnover rate)。如果有面板数据,就可能解决上述问题。

(3) 样本容量较大:由于同时有截面维度与时间维度,通常面板数据的样本容量更大,从而可以提高估计的精确度。

当然,面板数据也会带来一些问题,比如,样本数据通常不满足独立同分布的假定,因为同一个体在不同期的扰动项一般存在自相关。另外,面板数据的收集成本通常较高,不易获得。

15.2 面板数据的估计策略

估计面板数据的一个极端策略是将其看成截面数据而进行混合回归(pooled regression),即要求样本中每个个体都拥有完全相同的回归方程。另一极端策略则是,为每个个体估计一个单独的回归方程。前者忽略了个体间不可观测或被遗漏的异质性,而该异质性可能与解释变量相关从而导致估计不一致。后者则忽略了个体间的共性,也可能没有足够大的样本容量。因此,在实践中常采用折中的估计策略,即假定个体的回归方程拥有相同的斜率,但可以有不同的截距项,以此来捕捉异质性,参见图 15.1。这种模型被称为"个体效应模型"(individual-specific effects model),即

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + u_i + \varepsilon_{it} \quad (i = 1, \cdots, n; t = 1, \cdots, T) \tag{15.1}$$

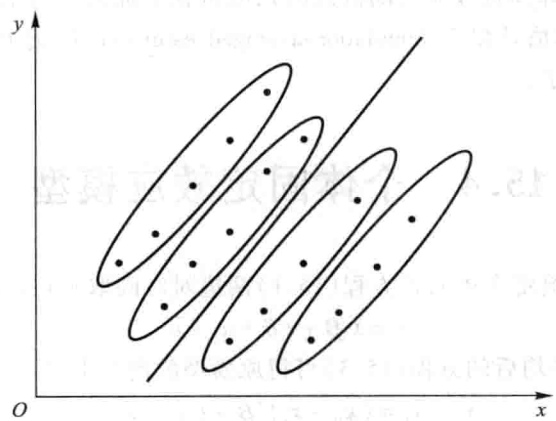


图 15.1 面板数据中不同个体的截距项可以不同

其中,$\boldsymbol{z}_i$ 为不随时间而变(time invariant)的个体特征(即 $\boldsymbol{z}_{it} = \boldsymbol{z}_i, \forall t$),比如性别;而 $\boldsymbol{x}_{it}$ 可以随个体及时间而变(time-varying)。扰动项由 $(u_i + \varepsilon_{it})$ 两部分构成,称为"复合扰动项"(composite error term),而方程(15.1)也称为"复合扰动项模型"(error components model)。其中,不可观测的随机变量 u_i 是代表个体异质性的截距项,故方程(15.1)也称为"不可观测效应模型"(unobserved effects model)。在较早的文献中有时将 u_i 视为常数,但这也只是随机变量的特例,即退化的随机

变量。ε_{it}为随个体与时间而改变的扰动项。假设$\{\varepsilon_{it}\}$为独立同分布(在下一章的长面板中将放松此假定),且与 u_i 不相关。

如果 u_i 与某个解释变量相关,则进一步称之为"固定效应模型"(Fixed Effects Model,简记FE)①。在这种情况下,OLS 是不一致的。解决的方法是将模型转换,消去 u_i 后获得一致估计量。

如果 u_i 与所有解释变量$(\boldsymbol{x}_{it},\boldsymbol{z}_i)$均不相关,则进一步称之为"随机效应模型"(Random Effects Model,简记 RE)。从经济理论的角度来看,随机效应模型比较少见②,但仍须通过数据来检验究竟该用随机效应还是固定效应模型。

显然,与截面数据相比,面板数据提供了更为丰富的模型与估计方法。

15.3 混 合 回 归

如果所有个体都拥有完全一样的回归方程,则方程(15.1)可以写为

$$y_{it}=\alpha+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\varepsilon_{it} \tag{15.2}$$

其中,$\boldsymbol{x}_{it}$不包括常数项。这样,就可以把所有的数据放在一起,像对待横截面数据那样进行 OLS 回归,故被称为"混合回归"(pooled regression)。由于面板数据的特点,虽然通常可以假设不同个体之间的扰动项相互独立,但同一个体在不同时期的扰动项之间往往存在自相关。此时,对标准误的估计应该使用聚类稳健的标准误(cluster-robust standard error),而所谓聚类(cluster),就是由每个个体不同时期的所有观测值所组成。同一聚类(个体)的观测值允许存在相关性,而不同聚类(个体)的观测值则不相关。

混合回归的基本假设是不存在个体效应。对于这个假设必须进行统计检验。由于个体效应以两种不同的形态存在(即固定效应与随机效应),故将在下面两节分别介绍其检验方法。混合回归也被称为"总体平均估计量"(Population-averaged estimator,简记 PA),因为可以把它理解为,将个体效应都平均掉了。

15.4 个体固定效应模型

对于固定效应模型,给定个体 i,将方程(15.1)两边对时间取平均可得

$$\bar{y}_i=\bar{\boldsymbol{x}}_i'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\bar{\varepsilon}_i \tag{15.3}$$

将方程(15.1)减去平均后的方程(15.3)可得原模型的离差形式③

$$y_{it}-\bar{y}_i=(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta}+(\varepsilon_{it}-\bar{\varepsilon}_i) \tag{15.4}$$

定义 $\tilde{y}_{it}\equiv y_{it}-\bar{y}_i$,$\tilde{\boldsymbol{x}}_{it}\equiv\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i$,$\tilde{\varepsilon}_{it}\equiv\varepsilon_{it}-\bar{\varepsilon}_i$,则

$$\tilde{y}_{it}=\tilde{\boldsymbol{x}}_{it}'\boldsymbol{\beta}+\tilde{\varepsilon}_{it} \tag{15.5}$$

---

① 固定效应(fixed effects)这个名词有时易引起误解。即使在固定效应模型中,个体效应 u_i 也是随机的(尽管其取值不随时间而变),而非固定(fixed)的常数。但这已是大家都接受的传统称呼。

② 一般来说,不可观测的异质性通常会对解释变量有影响。

③ 这种变换被称为"mean-differencing"或"time demeaning"。

由于上式中已将 u_i 消去，故只要 $\tilde{\varepsilon}_{it}$ 与 $\tilde{\boldsymbol{x}}_{it}$ 不相关，则可以用 OLS 一致地估计 $\boldsymbol{\beta}$，称为“固定效应估计量”(Fixed Effects Estimator)，记为 $\hat{\boldsymbol{\beta}}_{\text{FE}}$。由于 $\hat{\boldsymbol{\beta}}_{\text{FE}}$ 主要使用了每位个体的组内离差信息，故也称为“组内估计量”(within estimator)。即使个体特征 u_i 与解释变量 $\boldsymbol{x}_{it}$ 相关，只要使用组内估计量，就可以得到一致估计，这是面板数据的一大优势。然而，在作离差转换的过程中，$\boldsymbol{z}_i'\boldsymbol{\delta}$ 也被消掉了，故无法估计 $\boldsymbol{\delta}$。也就是说，$\hat{\boldsymbol{\beta}}_{\text{FE}}$ 无法估计不随时间而变的变量之影响，这是 FE 的一大缺点。另外，为了保证 $(\varepsilon_{it}-\bar{\varepsilon}_i)$ 与 $(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)$ 不相关，则要求第 i 个观测值满足严格外生性，即 $E(\varepsilon_{it}|\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT})=0$，因为 $\bar{\boldsymbol{x}}_i$ 中包含了所有 $(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT})$ 的信息。换言之，扰动项必须与各期的解释变量均不相关(而不仅仅是当期的解释变量)，这是一个比较强的假定。

如果在原方程中引入 $(n-1)$ 个虚拟变量(如果没有截距项，则引入 n 个虚拟变量)来代表不同的个体，则可以得到与上述离差模型同样的结果。因此，FE 也被称为“最小二乘虚拟变量模型”(Least Square Dummy Variable Model，简记 LSDV)。为什么二者的结果相同？直观来说，这正如线性回归与离差形式的回归在某种意义上是等价的[①]。比如，

$$y_i=\alpha+\beta x_i+\varepsilon_i \quad \Leftrightarrow \quad y_i-\bar{y}=\beta(x_i-\bar{x})+(\varepsilon_i-\bar{\varepsilon}) \tag{15.6}$$

但是，如果作完 LSDV 后发现某些个体的虚拟变量不显著而将其删去，那么 LSDV 的结果就不会与 FE 相同。使用 LSDV 的好处是可以得到对个体异质性 u_i 的估计。LSDV 法的缺点是，如果 n 很大，则须在回归方程中引入很多虚拟变量，可能超出计量软件(比如 Stata)所允许的解释变量个数。

15.5 时间固定效应

以上个体固定效应模型解决了不随时间而变(time invariant)但随个体而异的遗漏变量问题。类似地，引入时间固定效应，则可以解决不随个体而变(individual invariant)但随时间而变(time varying)的遗漏变量问题。假设模型为

$$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\gamma S_t+u_i+\varepsilon_{it} \tag{15.7}$$

其中，S_t 不可观测。定义 $\lambda_t\equiv\gamma S_t$，则上式可以写为

$$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\lambda_t+u_i+\varepsilon_{it} \tag{15.8}$$

在上式中，可将 λ_t 视为第 t 期独有的截距项，并将其解释为“第 t 期”对被解释变量 y 的效应。因此，这些 $\lambda_1,\cdots,\lambda_T$ 称为“时间固定效应”(time fixed effects)。显然，对于方程(15.8)，可使用 LSDV 法来估计，即对每个时期定义一个虚拟变量，然后把 $(T-1)$ 个时间虚拟变量包括在回归方程中(未包括的时间虚拟变量即为基期)，比如

$$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\gamma_2 D2_t+\cdots+\gamma_T DT_t+u_i+\varepsilon_{it} \tag{15.9}$$

其中，时间虚拟变量 $D2_t=1$，如果 $t=2$；$D2_t=0$，如果 $t\neq 2$；以此类推。由于方程(15.9)既考虑了个体固定效应，又考虑了时间固定效应，故称为“双向固定效应”(Two-way FE)。相应地，如果仅考虑个体固定效应(即上节的模型)，则称为“单向固定效应”(One-way FE)。

在有些情况下，为了节省参数，可以引入一个时间趋势项，以替代上述 $(T-1)$ 个时间虚拟变量)：

$$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\gamma t+u_i+\varepsilon_{it} \tag{15.10}$$

---

① 严格证明参见 Cameron and Trivedi(2005)，第 732 页。

显然，上式隐含着一个较强的假定，即每个时期的时间效应相等，每期均增加 γ。如果此假定不太可能成立，则应使用方程(15.9)的时间虚拟变量法(可通过该法的估计结果判断每期的时间效应是否大致相等)。

15.6 一阶差分法

对于固定效应模型，可对方程(15.1)两边进行一阶差分，以消去个体效应 u_i(但同时也把 $\boldsymbol{z}_i'\boldsymbol{\delta}$ 消掉了)，

$$y_{it}-y_{i,t-1}=(\boldsymbol{x}_{it}-\boldsymbol{x}_{i,t-1})'\boldsymbol{\beta}+(\varepsilon_{it}-\varepsilon_{i,t-1}) \tag{15.11}$$

对此差分形式的方程使用 OLS 即得到“一阶差分估计量”(First Differencing Estimator)，记为 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$。由于 u_i 不再出现在差分方程中，只要扰动项的一阶差分($\varepsilon_{it}-\varepsilon_{i,t-1}$)与解释变量的一阶差分($\boldsymbol{x}_{it}-\boldsymbol{x}_{i,t-1}$)不相关，则 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$是一致的。此一致性条件比保证 $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$一致的严格外生性假定更弱，这是 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$的主要优点。可以证明(参见习题)，如果 $T=2$，则 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}=\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$。但对于 $T>2$，如果$\{\varepsilon_{it}\}$为独立同分布的，则组内估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$比一阶差分估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$更有效率。因此，在实践上，主要使用 $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$，而较少用 $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$。但对于动态面板(参见第16章)，严格外生性假定无法满足，故转而使用差分法。

15.7 随机效应模型

对于回归方程 $y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\varepsilon_{it}$，随机效应模型假设 u_i 与解释变量$\{\boldsymbol{x}_{it},\boldsymbol{z}_i\}$均不相关，故 OLS 是一致的。然而，由于扰动项由($u_i+\varepsilon_{it}$)组成，不是球型扰动项，因此 OLS 不是最有效率的。假设不同个体之间的扰动项互不相关。由于 u_i 的存在，同一个体不同时期的扰动项之间存在自相关，

$$\mathrm{Cov}(u_i+\varepsilon_{it},u_i+\varepsilon_{is})=\begin{cases}\sigma_u^2, & \text{若 } t\neq s\\ \sigma_u^2+\sigma_\varepsilon^2, & \text{若 } t=s\end{cases} \tag{15.12}$$

其中，σ_u^2 为 u_i 的方差(不随 i 变化)，而 σ_ε^2 为 ε_{it}的方差(不随 i,t 变化)。当 $t\neq s$ 时，其自相关系数为

$$\rho\equiv\mathrm{Corr}(u_i+\varepsilon_{it},u_i+\varepsilon_{is})=\frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2} \tag{15.13}$$

显然，同一个体不同时期的扰动项之间的自相关系数 ρ 不随时间距离($t-s$)而改变，故随机效应模型也被称为“等相关模型”(equicorrelated model)或“可交换扰动项模型”(exchangeable errors model)[①]。ρ 越大，则复合扰动项($u_i+\varepsilon_{it}$)中个体效应的部分(u_i)越重要。Stata 记 ρ 为“rho”。

同一个体扰动项的协方差阵为

① 对于自回归模型，一般来说，扰动项之间的自相关系数 ρ(或其绝对值)随时间而递减。

$$\Sigma = \begin{pmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\varepsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\varepsilon^2 \end{pmatrix}_{T\times T} \tag{15.14}$$

由表达式(15.14)可知,同一个体的扰动项具有相同的方差,但存在组内自相关。整个样本扰动项的协方差阵为块对角矩阵(block diagonal matrix),

$$\Omega = \begin{pmatrix} \Sigma & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \Sigma \end{pmatrix}_{nT\times nT} \tag{15.15}$$

由于 OLS 是一致的,且其扰动项为($u_i + \varepsilon_{it}$),故可以用 OLS 的残差来估计($\sigma_u^2 + \sigma_\varepsilon^2$)。另一方面,FE 也是一致的,且其扰动项为($\varepsilon_{it} - \bar{\varepsilon}_i$),故可以用 FE 的残差来估计 σ_ε^2(Stata 记 $\hat{\sigma}_u$ 为"sigma_u",记 $\hat{\sigma}_\varepsilon$ 为"sigma_e")。然后,就可以使用可行广义最小二乘法(FGLS)来估计原模型,得到"随机效应估计量"(Random Effects Estimator),记为 $\hat{\boldsymbol{\beta}}_{RE}$。具体来说,用 OLS 来估计以下"广义离差"(quasi-demeaned)模型,

$$y_{it} - \hat{\theta}\bar{y}_i = (\boldsymbol{x}_{it} - \hat{\theta}\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (1-\hat{\theta})\boldsymbol{z}_i'\boldsymbol{\delta} + \underbrace{[(1-\hat{\theta})u_i + (\varepsilon_{it} - \hat{\theta}\bar{\varepsilon}_i)]}_{\text{误差项}} \tag{15.16}$$

其中,$\hat{\theta}$ 是 $\theta \equiv 1 - \dfrac{\sigma_\varepsilon}{(T\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}}$的一致估计量(Stata 记 $\hat{\theta}$ 为"theta")。可以证明(参见习题),方程(15.16)的扰动项不再有自相关。显然,$0 \leqslant \theta \leqslant 1$。进一步,如果 $\hat{\theta} = 0$,则为混合回归。而如果 $\hat{\theta} = 1$,则为组内估计量。

对于随机效应模型,如果进一步假设扰动项服从正态分布,则可写出样本的对数似然函数,然后进行最大似然估计(MLE)。

15.8 组间估计量

对于随机效应模型,还可以使用"组间估计量"。如果每位个体的时间序列数据较不准确或噪音较大,可对每位个体取时间平均值,然后用平均值来回归。

$$\bar{y}_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i \quad (i = 1, \cdots, n) \tag{15.17}$$

对上式使用 OLS,就可以得到"组间估计量"(Between Estimator),记为 $\hat{\boldsymbol{\beta}}_{BE}$。由于$\{\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{z}}_i\}$中包含了$\{\boldsymbol{x}_{it}, \boldsymbol{z}_i\}$的信息,如果 u_i 与解释变量$\{\boldsymbol{x}_{it}, \boldsymbol{z}_i\}$相关,则 $\hat{\boldsymbol{\beta}}_{BE}$是不一致的。因此,不能在固定效应模型下使用组间估计法。即使在随机效应模型下,由于面板数据被压缩为截面数据,损失了较多信息量,故组间估计法也不常用。

15.9 拟合优度的度量

对于线性回归模型,一般用 R^2 衡量拟合优度。在有常数项的情况下,线性回归模型的 R^2 就等于被解释变量 y 与 OLS 预测值 $\hat{y}$ 之间相关系数的平方,即 $R^2 = [\text{corr}(y, \hat{y})]^2$(参见第 3 章习

题)。对于面板模型,如果使用混合回归,则可直接用混合回归的 R^2 衡量拟合优度。但如果使用固定效应、随机效应或组间回归,拟合优度的度量略为复杂。任意给定估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$,Stata 提供了以下三种 R^2(不一定具有线性回归 R^2 的全部性质)。

首先,对应于原模型(15.1),称$[\text{Corr}(y_{it},\boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}+\boldsymbol{z}_i'\hat{\boldsymbol{\delta}})]^2$为"整体 R^2"(R^2overall),衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对原模型的拟合优度。

其次,对应于组内模型(15.5),称$[\text{Corr}(\tilde{y}_{it},\tilde{\boldsymbol{x}}_{it}'\hat{\boldsymbol{\beta}})]^2$为"组内 R^2"(R^2within),衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组内模型的拟合优度。

再次,对应于组间模型(15.17),称$[\text{Corr}(\bar{y}_i,\bar{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}+\boldsymbol{z}_i'\hat{\boldsymbol{\delta}})]^2$为"组间 R^2"(R^2between),衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组间模型的拟合优度。

显然,无论固定效应、随机效应还是组间回归,都可以计算以上三种 R^2。对于固定效应模型,建议使用组内 R^2,也就是组内方程(15.5)的 OLS R^2;而整体 R^2 与组间 R^2 只是相应的相关系数平方而已。对于组间回归模型,建议使用组间 R^2,也就是组间方程(15.17)的 OLS R^2;而整体 R^2 与组内 R^2 只是相应的相关系数平方而已。然而,对于随机效应模型,这三种 R^2 都只是相应的相关系数平方,而非方程(15.16)的 OLS R^2。

15.10 非平衡面板

在面板数据中,如果每个时期在样本中的个体完全一样,则称为"平衡面板数据"(balanced panel)。然而,有时某些个体的数据可能缺失(比如,个体死亡、企业倒闭或被兼并、个体不再参与调查),或者新的个体在后来才加入到调查中来。在这种情况下,每个时期观测到的个体不完全相同,被称为"非平衡面板数据"(unbalanced panel)或"不完全面板"(incomplete panel)。

显然,非平衡面板数据并不影响计算离差形式的组内估计量(within estimator),因此,固定效应模型的估计可以照样进行。对于随机效应模型而言,非平衡面板数据也没有实质性影响,只要在做广义离差变换时让 $\theta_i \equiv 1-\dfrac{\sigma_\varepsilon}{(T_i\sigma_u^2+\sigma_\varepsilon^2)^{1/2}}$($T_i$ 为第 i 位个体的时间维度),就可以照常进行 FGLS 估计。当然,非平衡面板数据使得估计量及其协方差矩阵的数学表达式更加复杂,但这些都由 Stata 在幕后进行。

非平衡面板可能出现的最大问题是,那些原来在样本中但后来丢掉的个体,如果其"丢掉"的原因是内生的(即与扰动项相关),则会导致样本不具有代表性(不再是随机样本),从而导致估计量不一致。比如,低收入的人群更容易从面板数据中丢掉。此时,需要用"样本选择模型"(sample selection model)来进行处理,参见第 14 章。

如果从非平衡面板数据中提取一个平衡的面板数据子集[①],然后进行数据处理,则必然会损失样本容量,降低估计效率。更进一步,如果人为"丢掉"的个体并非完全随机,则同样会破坏样本的随机性。

① 可通过连玉君老师提供的 Stata 命令"xtbalance"来实现,下载方法为"ssc install xtbalance",使用方法详见"help xtbalance"。

15.11 究竟该用固定效应还是随机效应模型

在处理面板数据时，究竟应该使用固定效应模型还是随机效应模型是一个根本问题。为此，希望检验原假设“H_0：u_i 与 $\boldsymbol{x}_{it}$，$\boldsymbol{z}_i$ 不相关”（即随机效应模型为正确模型）。无论原假设成立与否，FE 都是一致的。然而，如果原假设成立，则 RE 比 FE 更有效。但如果原假设不成立，则 RE 不一致。因此，如果 H_0 成立，则 FE 与 RE 估计量将共同收敛于真实的参数值，故 $(\hat{\boldsymbol{\beta}}_{\text{FE}}-\hat{\boldsymbol{\beta}}_{\text{RE}})\xrightarrow{p}\mathbf{0}$。反之，如果二者的差距过大，则倾向于拒绝原假设。豪斯曼检验（Hausman，1978）的统计量为

$$(\hat{\boldsymbol{\beta}}_{\text{FE}}-\hat{\boldsymbol{\beta}}_{\text{RE}})'\left[\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{FE}})}-\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{RE}})}\right]^{-1}(\hat{\boldsymbol{\beta}}_{\text{FE}}-\hat{\boldsymbol{\beta}}_{\text{RE}})\xrightarrow{d}\chi^2(K) \tag{15.18}$$

其中，K 为 $\hat{\boldsymbol{\beta}}_{\text{FE}}$ 的维度，即 x_{it} 中所包含的随时间而变的解释变量个数（因为 $\hat{\boldsymbol{\beta}}_{\text{FE}}$ 无法估计不随时间而变的解释变量系数）。如果该统计量大于临界值，则拒绝 H_0。

上述检验的缺点是，它假设在 H_0 成立的情况下，$\hat{\boldsymbol{\beta}}_{\text{RE}}$ 是最有效率的（fully efficient）。然而，如果扰动项存在异方差，则 $\hat{\boldsymbol{\beta}}_{\text{RE}}$ 并非最有效率的估计量。因此，上面的检验不适用于异方差的情形。

解决方法之一，通过自助法（bootstrap），即计算机模拟再抽样（resampling）的方法来计算 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{FE}}-\hat{\boldsymbol{\beta}}_{\text{RE}})$，参见第 19 章。

解决方法之二，进行以下辅助回归（Wooldridge，2010），

$$y_{it}-\hat{\theta}\bar{y}_i=(\boldsymbol{x}_{it}-\hat{\theta}\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta}+(1-\hat{\theta})\boldsymbol{z}_i'\boldsymbol{\delta}+(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)'\boldsymbol{\gamma}+[(1-\hat{\theta})u_i+(\varepsilon_{it}-\hat{\theta}\bar{\varepsilon}_i)] \tag{15.19}$$

然后，使用聚类稳健标准误来检验原假设“H_0：$\boldsymbol{\gamma}=\mathbf{0}$”。这个检验方法在异方差的情况下也适用。由于总可以把原模型变换为方程（15.16），故在上式中，$\boldsymbol{\gamma}=\mathbf{0}$。如果随机效应模型成立，则 OLS 是一致的，故 $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}=\mathbf{0}$。反之，如果固定效应模型成立，扰动项 $[(1-\hat{\theta})u_i+(\varepsilon_{it}-\hat{\theta}\bar{\varepsilon}_i)]$ 与 $(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)$ 相关（因为 u_i 与 $\boldsymbol{x}_{it}$ 相关），故 OLS 是不一致的，即 $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}^*\neq\boldsymbol{\gamma}=\mathbf{0}$。因此，拒绝“$H_0$：$\boldsymbol{\gamma}=\mathbf{0}$”，则意味着拒绝随机效应，接受固定效应。对于非平衡面板，则以 $\hat{\theta}_i$ 替代方程（15.19）中的 $\hat{\theta}$ 即可。

15.12 个体时间趋势

迄今为止，本章所考虑的个体效应都体现为个体截距项 u_i。事实上，个体异质性还可能表现为个体的不同时间趋势。比如，在一个跨国面板中，各国的经济增长率可能不同。为此，考虑以下模型：

$$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+\gamma_i t+u_i+\varepsilon_{it} \tag{15.20}$$

其中，$\gamma_i t$ 为个体时间趋势。类似于 u_i 的性质，一般将 γ_i 视为来自某分布的随机变量（但从该分布随机抽出一个观测值后，就不再随时间而变）。因此，此模型称为“随机趋势模型”（random trend model）。另外，如果 y_{it} 取对数形式（比如 $\ln GDP_{it}$），则 γ_i 可解释为在给定 $(\boldsymbol{x}_{it},\boldsymbol{z}_i)$ 条件下的

平均增长率(即$\partial \mathrm{E}(\ln GDP_{it})/\partial t$),故此模型也称为"随机增长模型"(random growth model)。

为了估计此模型,首先对方程(15.20)两边做差分,以去掉u_i:

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \gamma_i + \Delta \varepsilon_{it} \tag{15.21}$$

其中,$\gamma_i t - \gamma_i (t-1) = \gamma_i$。在形式上,方程(15.21)与标准的个体效应模型(15.1)一模一样。如果γ_i与解释变量$\Delta \boldsymbol{x}_{it}$不相关,则可用随机效应模型(RE)估计方程(15.21)。更一般地,如果γ_i与解释变量$\Delta \boldsymbol{x}_{it}$相关,则可用固定效应模型(FE)或一阶差分法(FD)估计方程(15.21)。

15.13 短面板的 Stata 命令及实例

1. 面板数据的设定

设定面板数据的 Stata 命令为

```
xtset panelvar timevar
```

命令"xtset"告诉 Stata 你的数据为面板数据,其中面板(个体)变量"panelvar"的取值必须为整数且不重复,相当于把样本中每个个体进行编号;而"timevar"为时间变量。假如"panelvar"本来是字符串(比如,国家名字 country),则可以使用以下命令将其转换为数字型变量:

```
encode country,gen(cntry)
```

这样,变量 cntry 就以"1,2,3,…"来指代不同的国家。

显示面板数据统计特性的 Stata 命令为

```
xtdes              (显示面板数据的结构,是否为平衡面板)
xtsum              (显示组内、组间与整体的统计指标)
xttab varname      (显示组内、组间与整体的分布频率,tab 指的是 tabulate)
xtline varname     (对每位个体分别显示该变量的时间序列图;如果希望将所有个体的时间序列图叠放在一起,可加上选择项 overlay)
```

以数据集"traffic.dta"为例[①]。该面板数据集包含了美国 48 个州 1982—1988 年"交通死亡率"(traffic fatality rates)的相关变量:fatal(交通死亡率),beertax(啤酒税),spircons(酒精消费量),unrate(失业率),perinck(人均收入,以千元记)[②],state(州),year(年)。

首先,设定 state 与 year 为面板(个体)变量及时间变量:

```
. use traffic.dta,clear
. xtset state year
```

```
      panel variable:  state (strongly balanced)
       time variable:  year, 1982 to 1988
               delta:  1 unit
```

上表显示,这是一个平衡的面板数据。其次,显示数据集的结构:

```
. xtdes
```

① 此例来自 Baum(2006)。

② Per capita income in thousands.

```
   state:  1, 4, ..., 56                                   n =         48
    year:  1982, 1983, ..., 1988                           T =          7
           Delta(year) = 1 unit
           Span(year)  = 7 periods
           (state*year uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%      95%      max
                         7       7        7         7         7        7        7

     Freq.   Percent     Cum.  |  Pattern
------------------------------+-----------
        48    100.00   100.00  |  1111111
------------------------------+-----------
        48    100.00           |  XXXXXXX
```

上表清晰地显示，$n=48$，而 $T=7$。由于 n 大而 T 小，故这是一个短面板。

再次，显示数据集中以上变量的统计特征：

```
. xtsum fatal beertax spircons unrate perinck state year
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
fatal	overall	2.040444	.5701938	.82121	4.21784	N =	336
	between		.5461407	1.110077	3.653197	n =	48
	within		.1794253	1.45556	2.962664	T =	7
beertax	overall	.513256	.4778442	.0433109	2.720764	N =	336
	between		.4789513	.0481679	2.440507	n =	48
	within		.0552203	.1415352	.7935126	T =	7
spircons	overall	1.75369	.6835745	.79	4.9	N =	336
	between		.6734649	.8614286	4.388572	n =	48
	within		.147792	1.255119	2.265119	T =	7
unrate	overall	7.346726	2.533405	2.4	18	N =	336
	between		1.953377	4.1	13.2	n =	48
	within		1.634257	4.046726	12.14673	T =	7
perinck	overall	13.88018	2.253046	9.513762	22.19345	N =	336
	between		2.122712	9.95087	19.51582	n =	48
	within		.8068546	11.43261	16.55782	T =	7
state	overall	30.1875	15.30985	1	56	N =	336
	between		15.44883	1	56	n =	48
	within		0	30.1875	30.1875	T =	7
year	overall	1985	2.002983	1982	1988	N =	336
	between		0	1985	1985	n =	48
	within		2.002983	1982	1988	T =	7

上表显示，变量 state 的组内（within）标准差为 0，因为分在同一组的数据属于同一个州。另一方面，变量 year 的组间（between）标准差为 0，因为不同组的这一变量取值完全相同。

下面，看一下被解释变量 fatal 在 48 个州的时间趋势图，结果如图 15.2。

```
. xtline fatal
```

从上图可以看出，不同州的交通死亡率（fatal）的时间趋势不尽相同，有些州很平稳（比如，佛罗里达州，FL），有些州呈下降趋势（比如，俄克拉荷马州，OK），而有些州则为上升趋势（比如，南卡罗来纳州，SC）。在一定程度上，交通死亡率的州际差异有助于估计决定交通死亡率的因素。

2. 混合回归

作为参照系，首先进行混合回归。其 Stata 命令的基本格式为

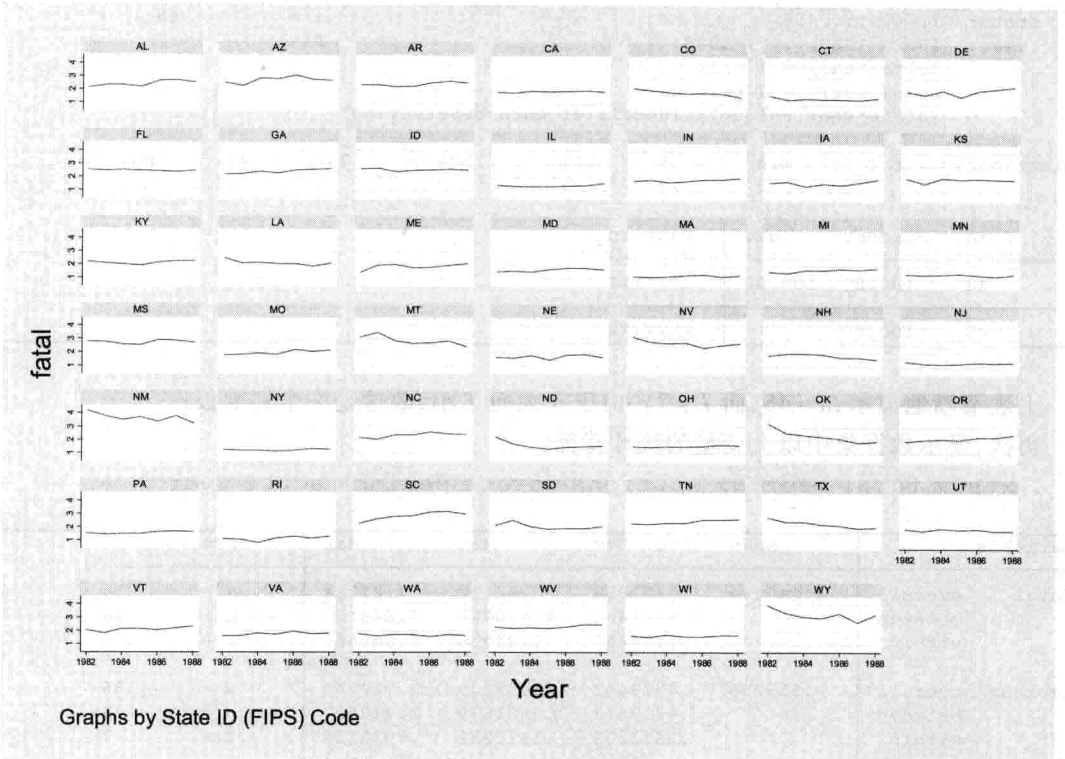


图 15.2 48 个州的时间趋势图

```
reg y x1 x2 x3,vce(cluster id)
```

其中,“id”指用来确定每位个体的变量。

```
. reg fatal beertax spircons unrate perinck,vce(cluster state)
. estimates store OLS
```

```
Linear regression                               Number of obs =       336
                                                F(  4,    47) =     17.43
                                                Prob > F      =    0.0000
                                                R-squared     =    0.3019
                                                Root MSE      =    .47929

                                    (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
       fatal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   .0971997    .116884     0.83   0.410    -.1379406     .33234
    spircons |   .1623471    .1070988    1.52   0.136     -.053108    .3778021
      unrate |  -.0291014    .0208687   -1.39   0.170    -.0710838     .012881
    perinck  |  -.1584291    .0371404   -4.27   0.000     -.233146   -.0837122
       _cons |   4.118674    .6765746    6.09   0.000     2.757581    5.479766
------------------------------------------------------------------------------
```

其中,选择项“vce(cluster state)”表示,使用以“state”为聚类变量的聚类稳健标准误。如果使用普通标准误,则为

```
. reg fatal beertax spircons unrate perinck
```

Source	SS	df	MS	Number of obs =	336
				F(4, 331) =	35.78
Model	32.878445	4	8.21961124	Prob > F =	0.0000
Residual	76.037065	331	.22971923	R-squared =	0.3019
				Adj R-squared =	0.2934
Total	108.91551	335	.325120925	Root MSE =	.47929

fatal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	.0971997	.0615518	1.58	0.115	-.0238823	.2182817
spircons	.1623471	.0432466	3.75	0.000	.0772743	.2474198
unrate	-.0291014	.0127156	-2.29	0.023	-.0541151	-.0040877
perinck	-.1584291	.0169869	-9.33	0.000	-.1918449	-.1250132
_cons	4.118674	.296995	13.87	0.000	3.534438	4.702909

对比普通标准误与聚类稳健标准误可知，前者大约仅是后者的一半。由于同一个州不同期之间的扰动项一般存在自相关，而默认的普通标准误计算方法假设扰动项为独立同分布的，故普通标准误的估计并不准确。

3. 固定效应

由于每个州的“州情”不同，可能存在不随时间而变的遗漏变量，故考虑使用固定效应模型（FE）。固定效应模型（组内估计量）的 Stata 命令基本格式为

```
xtreg y x1 x2 x3,fe r
```

LSDV 法的 Stata 命令为

```
reg y x1 x2 x3 i.id,r
```

其中，选择项“r”表示使用聚类稳健标准误；如果使用选择项“vce(cluster id)”也能达到完全相同的效果①。“id”表示用来确定个体的变量，“i.id”则表示根据变量“id”而生成的虚拟变量。

首先使用组内估计量：

```
. xtreg fatal beertax spircons unrate perinck,fe r
. estimates store FE_robust
```

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.3526                         Obs per group: min =         7
       between = 0.1146                                        avg =       7.0
       overall = 0.0863                                        max =         7

                                                F(4,47)            =     21.27
corr(u_i, Xb)  = -0.8804                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 48 clusters in state)
```

fatal	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.4840728	.2218754	-2.18	0.034	-.9304285	-.037717
spircons	.8169652	.1272627	6.42	0.000	.5609456	1.072985
unrate	-.0290499	.0094581	-3.07	0.004	-.0480772	-.0100227
perinck	.1047103	.0341455	3.07	0.004	.0360184	.1734022
_cons	-.383783	.7091738	-0.54	0.591	-1.810457	1.042891
sigma_u	1.1181913					
sigma_e	.15678965					
rho	.98071823	(fraction of variance due to u_i)				

① 在使用命令 xtreg 时，Stata 已经知道这是面板数据，故使用选择项“r”或“vce(robust)”，也能得到完全相同的聚类稳健标准误。

上表的Stata输出结果中包括一个常数项(_cons),这是所有个体效应 u_i 的平均值。上表最后一行显示,"rho=0.98",故复合扰动项($u_i+\varepsilon_{it}$)的方差主要来自个体效应 u_i 的变动。

究竟应该使用混合回归还是固定效应模型呢?

在使用命令"xtreg,fe"时,如果不加选择项"r",则输出结果还包含一个 F 检验,其原假设为"H_0 : $all\ u_i=0$",即混合回归是可以接受的:

```
. xtreg fatal beertax spircons unrate perinck,fe
. estimates store FE
```

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.3526                         Obs per group: min =         7
       between = 0.1146                                        avg =       7.0
       overall = 0.0863                                        max =         7

                                                F(4,284)           =     38.68
corr(u_i, Xb)  = -0.8804                        Prob > F           =    0.0000
```

fatal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.4840728	.1625106	-2.98	0.003	-.8039508	-.1641948
spircons	.8169652	.0792118	10.31	0.000	.6610484	.9728819
unrate	-.0290499	.0090274	-3.22	0.001	-.0468191	-.0112808
perinck	.1047103	.0205986	5.08	0.000	.064165	.1452555
_cons	-.383783	.4201781	-0.91	0.362	-1.210841	.4432754
sigma_u	1.1181913					
sigma_e	.15678965					
rho	.98071823	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(47, 284) =    59.77                 Prob > F = 0.0000
```

对于原假设"H_0 : $all\ u_i=0$",由于上表最后一行 F 检验的 p 值为0.000 0,故强烈拒绝原假设,即认为FE明显优于混合回归,应该允许每位个体拥有自己的截距项。然而,由于未使用聚类稳健标准误,故这个 F 检验并不有效,因为普通标准误大约只是聚类稳健标准误的一半[①]。为此,进一步通过LSDV法来考察:

```
. reg fatal beertax spircons unrate perinck i.state,vce(cluster state)
. estimates store LSDV
```

① 尽管如此,由于 p 值很小,为0.000 0,故即使按聚类标准误来计算 F 值,大致也能拒绝原假设。

Linear regression

Number of obs	= 336
F(3, 47)	= .
Prob > F	= .
R-squared	= 0.9359
Root MSE	= .15679

(Std. Err. adjusted for 48 clusters in state)

fatal	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.4840728	.2395323	-2.02	0.049	-.9659495	-.002196
spircons	.8169652	.1373903	5.95	0.000	.5405716	1.093359
unrate	-.0290499	.0102108	-2.85	0.007	-.0495914	-.0085085
perinck	.1047103	.0368628	2.84	0.007	.0305519	.1788687
state						
4	-1.146089	.3464123	-3.31	0.002	-1.842981	-.4491975
5	-.3422802	.2516739	-1.36	0.180	-.8485829	.1640224
6	-2.478906	.4335366	-5.72	0.000	-3.35107	-1.606743
8	-2.359016	.3971192	-5.94	0.000	-3.157917	-1.560115
9	-3.517465	.4888933	-7.19	0.000	-4.500992	-2.533939
10	-2.621819	.4507483	-5.82	0.000	-3.528608	-1.715031
12	-1.484055	.2576804	-5.76	0.000	-2.002441	-.9656691
13	-.4193311	.222352	-1.89	0.065	-.8666457	.0279835
16	-.475246	.3043054	-1.56	0.125	-1.08743	.1369375
17	-2.66262	.3935873	-6.77	0.000	-3.454415	-1.870824
18	-1.551228	.3208617	-4.83	0.000	-2.196718	-.9057372
19	-1.512083	.2850596	-5.30	0.000	-2.085549	-.9386167
20	-1.425295	.2782851	-5.12	0.000	-1.985132	-.8654573
21	-1.008736	.3429057	-2.94	0.005	-1.698573	-.3188985
22	-.9904429	.2155297	-4.60	0.000	-1.424033	-.556853
23	-1.728957	.2458348	-7.03	0.000	-2.223513	-1.234401
24	-2.971806	.4333382	-6.86	0.000	-3.84357	-2.100042
25	-3.542156	.4444227	-7.97	0.000	-4.43622	-2.648093
26	-2.039022	.3147281	-6.48	0.000	-2.672174	-1.405871
27	-2.726702	.3617594	-7.54	0.000	-3.454468	-1.998936
28	.1623338	.156378	1.04	0.305	-.152258	.4769257
29	-1.473051	.312394	-4.72	0.000	-2.101507	-.8445953
30	-.6003982	.327793	-1.83	0.073	-1.259833	.0590363
31	-1.812884	.2951366	-6.14	0.000	-2.406622	-1.219146
32	-3.433805	.6642254	-5.17	0.000	-4.770054	-2.097556
33	-4.226736	.5945373	-7.11	0.000	-5.422791	-3.030682
34	-3.425003	.4771024	-7.18	0.000	-4.384809	-2.465196
35	.4537402	.3027855	1.50	0.141	-.1553857	1.062866
36	-3.146312	.4211584	-7.47	0.000	-3.993574	-2.299051
37	-.6387593	.1008379	-6.33	0.000	-.8416191	-.4358996
38	-2.124115	.3257623	-6.52	0.000	-2.779464	-1.468765
39	-1.689753	.2868615	-5.89	0.000	-2.266843	-1.112662
40	-.6842853	.1722774	-3.97	0.000	-1.030863	-.3377078
41	-1.388066	.347058	-4.00	0.000	-2.086257	-.6898754
42	-1.92044	.323107	-5.94	0.000	-2.570447	-1.270432
44	-3.097237	.3984525	-7.77	0.000	-3.89882	-2.295654
45	-.0251951	.0789182	-0.32	0.751	-.1839581	.1335679
46	-1.345773	.2595865	-5.18	0.000	-1.867994	-.8235526
47	-.780764	.3155413	-2.47	0.017	-1.415551	-.1459769
48	-1.117104	.2845582	-3.93	0.000	-1.689561	-.544647
49	-.7663286	.2268098	-3.38	0.001	-1.222611	-.3100462
50	-1.844306	.29671	-6.22	0.000	-2.441209	-1.247402
51	-1.891904	.2537621	-7.46	0.000	-2.402408	-1.381401
53	-2.176212	.3644988	-5.97	0.000	-2.909489	-1.442935
54	-.2142207	.2815892	-0.76	0.451	-.780705	.3522637
55	-2.41626	.3894621	-6.20	0.000	-3.199757	-1.632763
56	-.7183534	.402788	-1.78	0.081	-1.528658	.0919516
_cons	1.269226	.7728348	1.64	0.107	-.2855167	2.823969

从上表可知，大多数个体虚拟变量均很显著（p 值为 0.000），故可放心地拒绝“所有个体虚拟变量都为 0”的原假设，即认为存在个体效应，不应使用混合回归。

对于固定效应模型，也可使用一阶差分法（FD）。Stata 没有专门执行一阶差分法的命令，但在使用命令“xtserial，output”对组内自相关进行检验时（参见第 16 章），可以附带提供一阶差分法的结果：

```
. xtserial fatal beertax spircons unrate perinck,output
. estimates store FD
```

```
Linear regression                                   Number of obs =       288
                                                    F(  4,    47) =      3.66
                                                    Prob > F      =    0.0113
                                                    R-squared     =    0.0814
                                                    Root MSE      =    .19047

                                  (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
    D.fatal  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    beertax  |
         D1. |   .1187701   .2276875     0.52   0.604    -.339278    .5768183
             |
    spircons |
         D1. |    .523584   .1633135     3.21   0.002    .1950396    .8521285
             |
      unrate |
         D1. |   .003399    .0118903     0.29   0.776    -.0205212   .0273192
             |
    perinck  |
         D1. |   .1417981   .0468699     3.03   0.004     .047508    .2360882
------------------------------------------------------------------------------

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
       F(  1,      47) =      6.246
            Prob > F =      0.0160
```

从上表可以看出，一阶差分估计量（FD）的估计系数与组内估计量（FE）有较大差别。比如，如果使用 FD 估计法，啤酒税（beertax）对交通死亡率（fatal）的影响系数竟然为正（0.12），尽管在统计上不显著（p 值为 0.60）。一般认为，FE 比 FD 更有效率，故较少使用 FD。

在使用命令 xtserial 计算一阶差分估计量时，同时得到了对“面板一阶自相关”的检验。该检验的 p 值为 0.016，即在 5% 的显著性水平上存在面板自相关。但这是一个只有 7 期的短面板（$T=7$），我们可以不考虑“面板自相关”的问题。

也可以在固定效应模型中考虑时间效应，即双向固定效应（Two-way FE）。为此，定义年度虚拟变量：

```
. tab year,gen(year)
```

Year	Freq.	Percent	Cum.
1982	48	14.29	14.29
1983	48	14.29	28.57
1984	48	14.29	42.86
1985	48	14.29	57.14
1986	48	14.29	71.43
1987	48	14.29	85.71
1988	48	14.29	100.00
Total	336	100.00	

此命令将在 Stata 的变量窗口生成时间虚拟变量 year1,year2,…,year7。

```
. xtreg fatal beertax spircons unrate perinck year2 -year7,fe r
. estimates store FE_TW
```

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.4528                         Obs per group: min =         7
       between = 0.1090                                        avg =       7.0
       overall = 0.0770                                        max =         7

                                                F(10,47)           =     14.13
corr(u_i, Xb)  = -0.8728                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
       fatal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.4347195   .2442775    -1.78   0.082    -.9261425    .0567036
    spircons |    .805857   .1161087     6.94   0.000     .5722764    1.039438
      unrate |  -.0549084    .011763    -4.67   0.000    -.0785725   -.0312443
     perinck |   .0882636   .0322971     2.73   0.009     .0232901     .153237
       year2 |  -.0533713   .0312438    -1.71   0.094    -.1162256    .0094831
       year3 |  -.1649828   .0439375    -3.75   0.000    -.2533737    -.076592
       year4 |  -.1997376   .0496167    -4.03   0.000    -.2995535   -.0999218
       year5 |  -.0508034   .0661756    -0.77   0.447    -.1839315    .0823248
       year6 |  -.1000728   .0756768    -1.32   0.192    -.2523149    .0521693
       year7 |   -.134057   .0863572    -1.55   0.127    -.3077854    .0396714
       _cons |   .1290568   .6238303     0.21   0.837    -1.125928    1.384041
-------------+----------------------------------------------------------------
     sigma_u |  1.0987683
     sigma_e |  .14570531
         rho |  .98271904   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

其中,year1(即 1982 年)被作为基期(对应于常数项_cons),而不包括在上述回归命令中[①]。时间效应的符号均为负,但有些年度虚拟变量显著,而有些则不显著。下面检验所有年度虚拟变量的联合显著性:

```
. test year2 year3 year4 year5 year6 year7
```

```
 ( 1)  year2 = 0
 ( 2)  year3 = 0
 ( 3)  year4 = 0
 ( 4)  year5 = 0
 ( 5)  year6 = 0
 ( 6)  year7 = 0

       F(  6,    47) =    8.21
            Prob > F =    0.0000
```

结果强烈拒绝"无时间效应"的原假设,认为应在模型中包括时间效应。在 Stata 12 中,还可以直接用以下命令来估计双向固定效应模型(不必生成时间虚拟变量):

```
.xtreg fatal beertax spircons unrate perinck i.year,fe r
```

① 否则,将出现虚拟变量陷阱,导致完全多重共线性。

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.4528                         Obs per group: min =         7
       between = 0.1090                                        avg =       7.0
       overall = 0.0770                                        max =         7

                                                F(10,47)           =     14.13
corr(u_i, Xb)  = -0.8728                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
       fatal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   -.4347195   .2442775    -1.78   0.082    -.9261425    .0567036
    spircons |    .805857   .1161087     6.94   0.000     .5722764    1.039438
      unrate |   -.0549084    .011763    -4.67   0.000    -.0785725   -.0312443
     perinck |    .0882636   .0322971     2.73   0.009     .0232901     .153237
             |
        year |
        1983 |   -.0533713   .0312438    -1.71   0.094    -.1162256    .0094831
        1984 |   -.1649828   .0439375    -3.75   0.000    -.2533737    -.076592
        1985 |   -.1997376   .0496167    -4.03   0.000    -.2995535   -.0999218
        1986 |   -.0508034   .0661756    -0.77   0.447    -.1839315    .0823248
        1987 |   -.1000728   .0756768    -1.32   0.192    -.2523149    .0521693
        1988 |    -.134057   .0863572    -1.55   0.127    -.3077854    .0396714
             |
       _cons |    .1290568   .6238303     0.21   0.837    -1.125928    1.384041
-------------+----------------------------------------------------------------
     sigma_u |  1.0987683
     sigma_e |  .14570531
         rho |  .98271904   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

4. 随机效应

以上结果已基本确认了个体效应的存在,但个体效应仍可能以随机效应(RE)的形式存在。随机效应的 Stata 命令为

```
xtreg y x1 x2 x3,re r theta   (随机效应 FGLS)
xtreg y x1 x2 x3,mle          (随机效应 MLE)
```

其中,选择项"r"表示使用聚类稳健标准误;如果使用选择项"vce(cluster id)"也能达到完全相同的效果。选择项"theta"表示显示用于随机效应估计的 θ 值。Breusch and Pagan(1980)提供了一个检验个体效应的 LM 检验(LM test for individual-specific effects),其原假设为"$H_0: \sigma_u^2 = 0$",而备择假设为"$H_1: \sigma_u^2 \neq 0$"。如果拒绝 H_0,则说明原模型中应该有一个反映个体特性的随机扰动项 u_i,而不应该使用混合回归。该 LM 检验的 Stata 命令为"xttest0"(在执行命令"xtreg,re"后才能进行)。

```
. xtreg fatal beertax spircons unrate perinck,re r theta
. estimates store RE
```

```
Random-effects GLS regression                   Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.2263                         Obs per group: min =         7
       between = 0.0123                                        avg =       7.0
       overall = 0.0042                                        max =         7

                                                Wald chi2(4)       =     42.30
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
theta          = .85922063

                                (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
       fatal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    beertax |   .0442768   .1283189     0.35   0.730    -.2072237    .2957772
   spircons |   .3024711   .1080475     2.80   0.005     .0907019    .5142404
     unrate |  -.0491381    .009052    -5.43   0.000    -.0668798   -.0313964
    perinck |  -.0110727    .024949    -0.44   0.657    -.0599718    .0378263
      _cons |   2.001973   .5226507     3.83   0.000      .977597     3.02635
-------------+----------------------------------------------------------------
    sigma_u |  .41675665
    sigma_e |  .15678965
        rho |  .87601197   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

上表最后三行显示，$\sigma_u = 0.416\ 756\ 65$，$\sigma_\varepsilon = 0.416\ 756\ 65$，而 $\rho \equiv \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = 0.876\ 011\ 97$。

```
. xttest0
```

```
Breusch and Pagan Lagrangian multiplier test for random effects

        fatal[state,t] = Xb + u[state] + e[state,t]

        Estimated results:
                         |       Var      sd = sqrt(Var)
                ---------+------------------------------
                   fatal |    .3251209        .5701938
                       e |     .024583        .1567897
                       u |    .1736861        .4167567

        Test:   Var(u) = 0
                                 chibar2(01) =    550.22
                              Prob > chibar2 =    0.0000
```

上表显示，LM 检验强烈拒绝“不存在个体随机效应”的原假设，即认为在“随机效应”与“混合回归”二者之间，应该选择“随机效应”。

作为对照，也可以对随机效应模型进行 MLE 估计：

```
. xtreg fatal beertax spircons unrate perinck,mle nolog
. estimates store MLE
```

```
Random-effects ML regression                    Number of obs      =       336
Group variable: state                           Number of groups   =        48
Random effects u_i ~ Gaussian                   Obs per group: min =         7
                                                               avg =       7.0
                                                               max =         7
                                                LR chi2(4)         =     67.39
Log likelihood  =  12.784035                    Prob > chi2        =    0.0000
```

fatal	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
beertax	-.1982558	.1505869	-1.32	0.188	-.4934007	.0968891
spircons	.5815301	.0871644	6.67	0.000	.4106909	.7523692
unrate	-.0381467	.0091776	-4.16	0.000	-.0561344	-.0201589
perinck	.0543694	.0223621	2.43	0.015	.0105405	.0981984
_cons	.6479714	.4682047	1.38	0.166	-.2696929	1.565636
/sigma_u	.8559797	.1147673			.6581681	1.113243
/sigma_e	.159379	.0070215			.1461946	.1737525
rho	.9664931	.0096949			.9425673	.9815318

```
Likelihood-ratio test of sigma_u=0: chibar2(01)=  479.84 Prob>=chibar2 = 0.000
```

上表显示，随机效应 MLE 的系数估计值与随机效应 FGLS 有所不同，但在性质上依然类似。另外，上表最后一行强烈拒绝原假设“H_0 ：$\sigma_u = 0$”，即认为存在个体随机效应，不应进行混合回归。

5. 组间估计量

组间估计量(BE)的 Stata 命令为

```
xtreg y x1 x2 x3,be
```

如果数据质量不好，可考虑使用组间估计量，但会损失较多的信息量。在此，也把它作为一种对照的估计方法：

```
. xtreg fatal beertax spircons unrate perinck,be
. estimates store BE
```

```
Between regression (regression on group means)  Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.0479                         Obs per group: min =         7
       between = 0.4565                                        avg =       7.0
       overall = 0.2583                                        max =         7

                                                F(4,43)            =      9.03
sd(u_i + avg(e_i.))=  .4209489                  Prob > F           =    0.0000
```

fatal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	.0740362	.1456333	0.51	0.614	-.2196614	.3677338
spircons	.2997517	.1128135	2.66	0.011	.0722417	.5272618
unrate	.0322333	.038005	0.85	0.401	-.0444111	.1088776
perinck	-.1841747	.0422241	-4.36	0.000	-.2693277	-.0990218
_cons	3.796343	.7502025	5.06	0.000	2.283415	5.309271

6. 固定效应还是随机效应：豪斯曼检验

在处理面板数据时，究竟使用固定效应还是随机效应模型，这是一个基本问题。为此，进行豪斯曼检验。豪斯曼检验的 Stata 命令如下：

```
xtreg y x1 x2 x3,fe  (固定效应估计)
```

```
estimates store FE  (存储结果)
xtreg y x1 x2 x3,re  (随机效应估计)
estimates store RE  (存储结果)
hausman FE RE,constant sigmamore  (豪斯曼检验)
```

其中,选择项"constant"表示在比较系数估计值时包括常数项(默认设置不包括常数项),而选择项"sigmamore"表示统一使用更有效率的那个估计量(即随机效应估计量)的方差估计。由于在前面我们已经存储了相关的结果,故可以直接进行豪斯曼检验:

```
. hausman FE RE,constant sigmamore
```

	Coefficients			
	(b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
beertax	-.4840728	.0442768	-.5283495	.1462601
spircons	.8169652	.3024711	.514494	.0663028
unrate	-.0290499	-.0491381	.0200882	.0037898
perinck	.1047103	-.0110727	.115783	.0140556
_cons	-.383783	2.001973	-2.385756	.3078253

```
                b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

              chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =          90.46
            Prob>chi2 =         0.0000
            (V_b-V_B is not positive definite)
```

由于 p 值为 0.000 0,故强烈拒绝原假设"H_0 : u_i 与 $\boldsymbol{x}_{it}$,$\boldsymbol{z}_i$ 不相关",认为应该使用固定效应模型,而非随机效应模型。

然而,传统的豪斯曼检验假定,在 H_0 成立的情况下,随机效应模型是最有效率的(fully efficient)。这意味着,u_i 与 ε_{it}都必须是独立同分布的①。因此,如果聚类稳健标准误与普通标准误相差较大(在本例中,二者约相差一倍),则传统的豪斯曼检验不适用。事实上,在 Stata 中进行以上豪斯曼检验时,如果使用聚类稳健标准误,比如"xtreg y x1 x2 x3,fe vce(cluster id)",则 Stata 将无法执行"hausman FE RE"命令。

解决方法之一是使用自助法(参见第 19 章)。解决方法之二是进行以下辅助回归:

$$y_{it}-\hat{\theta}\bar{y}_i=(\boldsymbol{x}_{it}-\hat{\theta}\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta}+(1-\hat{\theta})\boldsymbol{z}_i'\boldsymbol{\delta}+(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)'\boldsymbol{\gamma}+[(1-\hat{\theta})u_i+(\varepsilon_{it}-\hat{\theta}\bar{\varepsilon}_i)] \tag{15.22}$$

然后,使用聚类稳健标准误来检验原假设"H_0 : $\boldsymbol{\gamma}=\mathbf{0}$"。为了演示的目的,首先介绍如何在 Stata 中手工进行此辅助回归(也可直接使用非官方命令 xtoverid,参见下文)②。首先,进行 RE 估计,得到 $\hat{\theta}$(使用命令"xtreg,re"时,加上选择项"theta",即会显示 $\hat{\theta}$ 的估计值;或者使用命令"scalar theta=e(theta)"从 RE 回归结果中直接获得 $\hat{\theta}$ 的取值);然后用 $\hat{\theta}$ 对模型进行如上变换(生成相应的广义离差变量),最后进行辅助回归与检验:

```
. quietly xtreg fatal beertax spircons unrate perinck,re
```

① 否则,即使 FGLS 也不是最有效率的,因为 FGLS 必须首先估计一些参数,再进行 GLS 估计。

② 此程序仅适用于平衡面板,因为非平衡面板会有很多 $\hat{\theta}_i$,程序将更为复杂。对于非平衡面板,可直接使用下文介绍的非官方命令 xtoverid。

```
. scalar theta = e(theta)
. global yandxforhausman fatal beertax spircons unrate perinck
. sort state
. foreach x of varlist $yandxforhausman{
2. by state:egen mean`x´=mean(`x´)
3. gen md`x´=`x´-mean`x´
4. gen red`x´=`x´-theta*mean`x´
5. }
. quietly reg redfatal redbeertax redspircons redunrate redperinck
mdbeertax mdspircons mdunrate mdperinck,vce(cluster state)
. test mdbeertax mdspircons mdunrate mdperinck
```

```
( 1)  mdbeertax = 0
( 2)  mdspircons = 0
( 3)  mdunrate = 0
( 4)  mdperinck = 0

      F(  4,    47) =   15.86
           Prob > F =    0.0000
```

其中，“global”为“全局宏”(global macro)，即用“yandxforhausman”来表示“fatal beertax spircons unrate perinck”。如此定义之后，但凡看见“ $yandxforhausman”，就视同“fatal beertax spircons unrate perinck”。“foreach”为循环语句，表示对变量名单“varlist”上的每个变量“x”都进行同样的操作。“egen”为命令“gen”的扩展(extension)，使用“egen”可以调用更多的函数。新变量名的前缀“md”表示原变量的“mean-differences”，即($x_{it}-\bar{x}_i$)；而新变量名的前缀“red”表示原变量的“RE-differences”，即($x_{it}-\hat{\theta}\bar{x}_i$)。由于该检验的 p 值为0.000 0，故强烈拒绝“H_0 ：$\gamma=0$”，即拒绝随机效应，认为应该使用固定效应模型。

除了使用以上手工程序，更直接的方法是下载非官方命令 xtoverid 来进行此辅助回归。此处“overid”指的是过度识别检验(overidentification test)，因为随机效应模型与固定效应模型相比，前者多了“个体异质性 u_i 与解释变量不相关”的约束条件，也可视为过度识别条件。

```
. ssc install xtoverid  (下载安装命令 xtoverid)
```

在使用命令 xtoverid 之前，须先以稳健标准误来执行命令“xtreg,re”。

```
. quietly xtreg fatal beertax spircons unrate perinck,re r
. xtoverid
```

```
Test of overidentifying restrictions: fixed vs random effects
Cross-section time-series model: xtreg re  robust cluster(state)
Sargan-Hansen statistic  63.427  Chi-sq(4)     P-value = 0.0000
```

命令 xtoverid 汇报的 $\chi^2(4)$ 统计量为 63.427，p 值为 0.000 0。根据第 5 章附录，F 分布与 χ^2 分布在大样本下是等价的。根据上文的手工程序，F 统计量为 $F(4,47)=15.86$，则相应的 χ^2 统计量为 $\chi^2(4)=4\times F(4,47)=4\times 15.86=63.44$。因此，使用命令 xtoverid 的结果与手工程序完全一致(只是小数点的精确位数不同)。使用命令 xtoverid 的好处在于，它也适用于非平衡面板。

为了便于比较，将以上各主要方法的系数估计值及标准误列表：

```
. estimates table OLS FE_robust FE_TW RE BE,b se
```

Variable	OLS	FE_robust	FE_TW	RE	BE
beertax	.0971997	-.48407277	-.43471948	.04427676	.07403619
	.11688401	.22187542	.24427755	.12046126	.14563334
spircons	.16234707	.81696515	.805857	.30247114	.29975174
	.10709883	.12726271	.11610873	.06429535	.11281346
unrate	-.02910139	-.02904993	-.05490839	-.0491381	.03223326
	.02086869	.00945812	.01176299	.00981968	.03800499
perinck	-.15842906	.10471027	.08826358	-.01107273	-.18417475
	.03714042	.0341455	.03229713	.01947461	.04222407
year2			-.05337126		
			.03124376		
year3			-.16498284		
			.04393752		
year4			-.19973763		
			.04961666		
year5			-.05080338		
			.0661756		
year6			-.10007279		
			.0756768		
year7			-.13405703		
			.08635725		
_cons	4.1186735	-.38378301	.12905677	2.0019735	3.796343
	.67657457	.70917383	.62383029	.38112466	.75020251

legend: b/se

从上表可知，不同方法的系数估计值差别较大。其中，固定效应（FE_robust，FE_TW）的系数估计值与其他估计法（OLS，RE，BE）的差别最大。另一方面，在固定效应模型中加入时间效应（FE_TW），对估计系数的影响不大。

习　题

15.1　证明当 $T=2$ 时，$\hat{\boldsymbol{\beta}}_{\mathrm{FD}}=\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$。

15.2　证明经过广义离差变换后的随机效应模型，其扰动项不再存在自相关，即

$$\mathrm{Cov}[(1-\theta)u_i+(\varepsilon_{it}-\theta\bar{\varepsilon}_i),(1-\theta)u_i+(\varepsilon_{is}-\theta\bar{\varepsilon}_i)]=0,\quad t\neq s \tag{15.23}$$

其中，$\theta\equiv 1-\dfrac{\sigma_\varepsilon}{(T\sigma_u^2+\sigma_\varepsilon^2)^{1/2}}$。

15.3　面板数据集 mus08psidextract.dta 包含 595 名美国工人 1976—1982 年的以下变量：lwage（工资对数），ed（教育年限），exp（工龄），exp2（exp 的平方），wks（工作周数，weeks worked）。考虑以下模型：

$$\mathrm{lwage}_{it}=\alpha+\beta_1\,\mathrm{exp}_{it}+\beta_2\mathrm{exp2}_{it}+\beta_3\,\mathrm{wks}_{it}+\beta_4\,\mathrm{ed}_{it}+u_i+\varepsilon_{it} \tag{15.24}$$

（1）进行混合回归；

（2）对随机效应模型进行 FGLS 估计，检验是否存在个体效应；

（3）对随机效应模型进行 MLE 估计；

（4）对固定效应模型计算组内估计量；

（5）对固定效应模型进行 LSDV 估计，检验是否存在个体效应；

（6）估计带时间效应的固定效应模型；

（7）计算一阶差分估计量；

（8）计算组间估计量；

（9）进行传统的豪斯曼检验；

（10）进行稳健的豪斯曼检验。

第 16 章　长面板与动态面板

16.1　长面板的估计策略

上一章关注短面板。对于短面板，由于时间维度 T 较小，每位个体的信息较少，无法探讨扰动项 $\{\varepsilon_{it}\}$ 是否存在自相关，故一般假设 $\{\varepsilon_{it}\}$ 独立同分布。对于长面板（long panel），由于 T 较大，信息较多，故可以放松这个假定，考虑 $\{\varepsilon_{it}\}$ 可能存在的异方差与自相关。在长面板中，由于 n 相对于 T 较小，对于可能存在的固定效应，只要加入个体虚拟变量即可（即 LSDV 法）。对于时间效应，可以通过加上时间趋势项或其平方项来控制（由于 T 较大，如果加上时间虚拟变量，则可能损失较多的自由度）。

为此，考虑以下模型：

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{16.1}$$

其中，$\boldsymbol{x}_{it}$ 可以包括常数项、时间趋势项（或其平方项）、个体虚拟变量以及不随时间变化的解释变量 z_i。下面考虑扰动项 $\{\varepsilon_{it}\}$ 存在异方差或自相关的几种情形。

（1）记个体 i 的扰动项方差为 $\sigma_i^2 \equiv \mathrm{Var}(\varepsilon_{it})$。如果存在 $\sigma_i^2 \neq \sigma_j^2 (i \neq j)$，则称扰动项 $\{\varepsilon_{it}\}$ 存在“组间异方差”（groupwise heteroskedasticity）。

（2）如果存在 $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0 (t \neq s, \forall i)$，则称扰动项 $\{\varepsilon_{it}\}$ 存在“组内自相关”（autocorrelation within panel）。

（3）如果存在 $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt}) \neq 0 (i \neq j, \forall t)$，则称扰动项 $\{\varepsilon_{it}\}$ 存在“组间同期相关”（contemporaneous correlation）或“截面相关”（cross-sectional correlation）。比如，对于省际面板数据（provincial panel data），相邻省份之间的同期经济活动可能通过贸易或投资相互影响。这种相关也称为“空间相关”（spatial correlation）。

对于扰动项 $\{\varepsilon_{it}\}$ 可能存在的组间异方差、组内自相关或组间同期相关，主要有两类处理方法。方法一，继续使用 OLS（即 LSDV）来估计系数，只对标准误差进行校正（即下文的面板校正标准误差）。方法二，对异方差或自相关的具体形式进行假设，然后使用可行广义最小二乘法（FGLS）进行估计。

16.2　面板校正标准误

对于模型（16.1），即使扰动项 $\{\varepsilon_{it}\}$ 存在组间异方差或组间同期相关，OLS（即 LSDV）也依然是一致的。此时，只要使用“组间异方差、组间同期相关”稳健的标准误差即可，即“面板校正标准误差”（Panel-Corrected Standard Error，简记 PCSE）。相应的 Stata 命令为

```
xtpcse y x1 x2 x3,hetonly
```

其中，选择项“hetonly”表示存在组间异方差，但不存在组间同期相关；默认既存在组间异方差，也存在组间同期相关。

下面以数据集 mus08cigar.dta 为例[①]。该面板数据集包含了美国 10 个州 1963—1992 年有关香烟消费量的相关变量：lnc（人均香烟销售量的对数），lnp（实际香烟价格的对数）[②]，lnpmin（相邻州最低香烟价格的对数），lny（人均可支配收入的对数），state（州），year（年）。由于 $n=10$，而 $T=30$，故这是一个长面板。

考虑估计如下香烟需求函数：

$$\ln c_{it}=\beta_0+\beta_1 \ln p_{it}+\beta_2 \ln pmin_{it}+\beta_3 \ln y_{it}+u_i+\varepsilon_{it} \tag{16.2}$$

其中，变量 lnp 代表价格效应，lnpmin 代表替代效应（可以购买价格更低的邻州香烟），而 lny 则代表收入效应。

```
. use mus08cigar.dta,clear
```

为了考虑时间效应，生成时间趋势变量 t[③]，然后用 LSDV 法估计双向固定效应模型（作为对比，先不考虑组间异方差或同期相关）：

```
. gen t =year -62
. reg lnc lnp lnpmin lny i.state t,vce(cluster state)
. estimates store OLS
```

```
Linear regression                               Number of obs =       300
                                                F(  3,     9) =         .
                                                Prob > F      =         .
                                                R-squared     =    0.7203
                                                Root MSE      =    .11203

                                  (Std. Err. adjusted for 10 clusters in state)
------------------------------------------------------------------------------
             |               Robust
         lnc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -1.027181   .4412156    -2.33   0.045    -2.02528   -.0290819
      lnpmin |   .5100582   .2627025     1.94   0.084   -.0842161    1.104332
         lny |   .4975365   .4139086     1.20   0.260   -.4387899    1.433863
             |
       state |
           2 |  -.0773908   .0594636    -1.30   0.225   -.2119069    .0571252
           3 |    .088557   .0090317     9.81   0.000    .0681259    .1089881
           4 |  -.1809375   .1695839    -1.07   0.314   -.5645629     .202688
           5 |  -.1066138   .2409767    -0.44   0.669    -.651741    .4385135
           6 |   .2177434   .1063333     2.05   0.071   -.0227993     .458286
           7 |    .115543   .1959351     0.59   0.570    -.327693    .5587791
           8 |   .1068277    .124702     0.86   0.414   -.1752679    .3889233
           9 |   .0433207     .04618     0.94   0.373   -.0611458    .1477872
          10 |   -.133583   .0333562    -4.00   0.003   -.2090399   -.0581261
             |
           t |  -.0429824   .0305275    -1.41   0.193   -.1120405    .0260757
       _cons |   3.488747    4.03674     0.86   0.410   -5.642993    12.62049
------------------------------------------------------------------------------
```

上表显示，有些州虚拟变量很显著，即存在固定效应，应允许每个州拥有自己的截距项；而时间效应则不太显著，p 值高达 0.19。然而，上表中的稳健标准误未考虑可能存在的组间异方差与

① 此例来自 Cameron and Trivedi (2009)。

② 不同州的香烟税不同，故香烟的实际价格不同。

③ 为了减少样本容量的损失，此处不用年度虚拟变量。

组间同期相关。为此,使用面板校正标准误进行估计。

```
. xtpcse lnc lnp lnpmin lny i.state t
. estimates store PCSE
```

```
Linear regression, correlated panels corrected standard errors (PCSEs)

Group variable:    state                          Number of obs      =       300
Time variable:     year                           Number of groups   =        10
Panels:            correlated (balanced)          Obs per group: min =        30
Autocorrelation:   no autocorrelation                               avg =        30
                                                                    max =        30
Estimated covariances      =        55             R-squared          =    0.7203
Estimated autocorrelations =         0             Wald chi2(13)      =   2147.36
Estimated coefficients     =        14             Prob > chi2        =    0.0000
```

lnc	Panel-corrected Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnp	-1.027181	.1332425	-7.71	0.000	-1.288332	-.7660305
lnpmin	.5100582	.1393134	3.66	0.000	.237009	.7831074
lny	.4975365	.1804528	2.76	0.006	.1438554	.8512176
state						
2	-.0773908	.0389125	-1.99	0.047	-.153658	-.0011237
3	.088557	.011143	7.95	0.000	.0667171	.1103968
4	-.1809375	.0829463	-2.18	0.029	-.3435093	-.0183657
5	-.1066138	.1018461	-1.05	0.295	-.3062284	.0930008
6	.2177434	.049543	4.40	0.000	.1206408	.3148459
7	.115543	.0980092	1.18	0.238	-.0765516	.3076376
8	.1068277	.0483694	2.21	0.027	.0120255	.20163
9	.0433207	.0202198	2.14	0.032	.0036907	.0829508
10	-.133583	.0221487	-6.03	0.000	-.1769937	-.0901722
t	-.0429824	.014234	-3.02	0.003	-.0708806	-.0150843
_cons	3.488747	1.637622	2.13	0.033	.2790665	6.698428

从上表可知,命令 xtpcse 的估计系数与命令 reg 的估计系数完全一样,只是标准误不同。在本例中,面板校正标准误差似乎反而更小。

16.3 仅解决组内自相关的 FGLS

对于模型(16.1),假设扰动项 ε_{it} 服从 AR(1)过程,即

$$\varepsilon_{it}=\rho_i\varepsilon_{i,t-1}+v_{it} \tag{16.3}$$

其中,$|\rho_i|<1$,$\{v_{it}\}$ 为独立同分布的,且期望为 0。如果 $\rho_i=\rho(i=1,\cdots,n)$,则所有个体的扰动项都服从自回归系数相同的 AR(1)过程。使用 Prais-Winsten 估计法对原模型进行广义差分变换(参见第 8 章),即可得到 FGLS 估计量。相应的 Stata 命令为

```
xtpcse y x1 x2 x3,corr(ar1) corr(psar1)
```

其中,选择项"corr(ar1)"对应于 $\rho_i=\rho$ 的情形;而选择项"corr(psar1)"则允许每个面板有自己的 ρ_i,psar1 指 panel-specific ar1。如果 T 并不比 n 大很多,则建议使用选择项"corr(ar1)",即约束每个面板(个体)的自回归系数 ρ 均相等,因为时间维度 T 可能无法提供足够的信息来分别估计每个面板自己的 ρ_i。

如果既不指定 corr(ar1),也不指定 corr(psar1),则进行 OLS 回归。无论使用 OLS 还是 Prais-Winsten 估计法,命令 xtpcse 均提供面板校正标准误(在组间异方差与同期相关的情况下依然成立)。

继续上面的实例。考虑存在组内自相关的情形,但要求各组的自回归系数相同:

```
. xtpcse lnc lnp lnpmin lny i.state t,corr(ar1)
. estimates store AR1
```

```
Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

Group variable:    state                          Number of obs      =       300
Time variable:     year                           Number of groups   =        10
Panels:            correlated (balanced)          Obs per group: min =        30
Autocorrelation:   common AR(1)                                  avg =        30
                                                                 max =        30
Estimated covariances      =        55           R-squared          =    0.9824
Estimated autocorrelations =         1           Wald chi2(13)      =    504.30
Estimated coefficients     =        14           Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |             Panel-corrected
         lnc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -.3440621   .0571589    -6.02   0.000    -.4560914   -.2320328
      lnpmin |   .1000875   .0736082     1.36   0.174     -.044182    .2443569
         lny |   .5661073   .1566725     3.61   0.000     .2590348    .8731798
             |
       state |
           2 |  -.0784486   .0519567    -1.51   0.131    -.1802818    .0233846
           3 |   .0954506   .0261747     3.65   0.000     .0441491     .146752
           4 |   -.207494   .0810219    -2.56   0.010     -.366294   -.048694
           5 |  -.1811666   .0977461    -1.85   0.064    -.3727454    .0104123
           6 |    .185644   .0554283     3.35   0.001     .0770065    .2942816
           7 |   .0701358    .120881     0.58   0.562    -.1667866    .3070582
           8 |   .0397148   .0507738     0.78   0.434    -.0598001    .1392297
           9 |   .0231754    .025083     0.92   0.356    -.0259863    .0723371
          10 |  -.1289763   .0387423    -3.33   0.001    -.2049099   -.0530428
             |
           t |  -.0498643   .0121104    -4.12   0.000    -.0736003   -.0261283
       _cons |   1.746348   1.305801     1.34   0.181    -.8129742    4.305671
-------------+----------------------------------------------------------------
         rho |   .7936188
------------------------------------------------------------------------------
```

从上表可知,是否考虑组内自相关,对估计系数有较大影响。下面考虑允许各组自回归系数不同的组内自相关情形:

```
. xtpcse lnc lnp lnpmin lny i.state t,corr(psar1)
. estimates store PSAR1
```

```
Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

Group variable:    state                          Number of obs      =       300
Time variable:     year                           Number of groups   =        10
Panels:            correlated (balanced)          Obs per group: min =        30
Autocorrelation:   panel-specific AR(1)                           avg =        30
                                                                  max =        30
Estimated covariances      =        55            R-squared          =    0.9954
Estimated autocorrelations =        10            Wald chi2(12)      =    855.24
Estimated coefficients     =        13            Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |             Panel-corrected
         lnc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -.2960784   .0539584    -5.49   0.000    -.401835   -.1903218
      lnpmin |   .0507118   .0673426     0.75   0.451   -.0812772    .1827009
         lny |   .5326598   .1350574     3.94   0.000    .2679522    .7973674
             |
       state |
           2 |  -.0639957   .0802776    -0.80   0.425   -.2213369    .0933456
           3 |          0  (omitted)
           4 |   -.187228   .0981734    -1.91   0.057   -.3796444    .0051885
           5 |  -.163645    .1128024    -1.45   0.147   -.3847336    .0574436
           6 |   .1980726   .0809248     2.45   0.014    .0394629    .3566822
           7 |   .0796166   .1709107     0.47   0.641   -.2553622    .4145954
           8 |   .0492827   .0817016     0.60   0.546   -.1108495    .2094149
           9 |   .0263335   .0630703     0.42   0.676    -.097282     .149949
          10 |  -.1229701    .065134    -1.89   0.059   -.2506303    .0046901
             |
           t |  -.0488194   .0103959    -4.70   0.000   -.0691949   -.0284439
       _cons |   2.015998   1.142949     1.76   0.078   -.2241421    4.256138
-------------+----------------------------------------------------------------
        rhos =  .9414582    .66151         1  .7045066  .7351521 ...  .6905643
------------------------------------------------------------------------------
```

其中，第 3 个州的州虚拟变量系数估计值及其标准误，由于过于接近于 0 而被 Stata 自动删除(omitted)。为了便于比较，将以上各种方法的系数估计值及标准误列表：

```
.esttab OLS PCSE AR1 PSAR1,r2 se mtitles star(* 0.1 ** 0.05 *** 0.01)
```

	(1) OLS	(2) PCSE	(3) AR1	(4) PSAR1
lnp	-1.027**	-1.027***	-0.344***	-0.296***
	(0.441)	(0.133)	(0.0572)	(0.0540)
lnpmin	0.510*	0.510***	0.100	0.0507
	(0.263)	(0.139)	(0.0736)	(0.0673)
lny	0.498	0.498***	0.566***	0.533***
	(0.414)	(0.180)	(0.157)	(0.135)
1b.state	0	0	0	0
	(.)	(.)	(.)	(.)
2.state	-0.0774	-0.0774**	-0.0784	-0.0640
	(0.0595)	(0.0389)	(0.0520)	(0.0803)
3.state	0.0886***	0.0886***	0.0955***	
	(0.00903)	(0.0111)	(0.0262)	
4.state	-0.181	-0.181**	-0.207**	-0.187*
	(0.170)	(0.0829)	(0.0810)	(0.0982)
5.state	-0.107	-0.107	-0.181*	-0.164
	(0.241)	(0.102)	(0.0977)	(0.113)
6.state	0.218*	0.218***	0.186***	0.198**
	(0.106)	(0.0495)	(0.0554)	(0.0809)
7.state	0.116	0.116	0.0701	0.0796
	(0.196)	(0.0980)	(0.121)	(0.171)
8.state	0.107	0.107**	0.0397	0.0493
	(0.125)	(0.0484)	(0.0508)	(0.0817)
9.state	0.0433	0.0433**	0.0232	0.0263
	(0.0462)	(0.0202)	(0.0251)	(0.0631)
10.state	-0.134***	-0.134***	-0.129***	-0.123*
	(0.0334)	(0.0221)	(0.0387)	(0.0651)
t	-0.0430	-0.0430***	-0.0499***	-0.0488***
	(0.0305)	(0.0142)	(0.0121)	(0.0104)
3o.state				0
				(.)
_cons	3.489	3.489**	1.746	2.016*
	(4.037)	(1.638)	(1.306)	(1.143)
N	300	300	300	300
R-sq	0.720	0.720	0.982	0.995

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

从上表可知，AR1 与 PSAR1 的系数估计值比较接近，但与 OLS 的结果相去甚远。究竟选择 AR1 还是 OLS，需要对组内自相关进行检验，参见下文。

16.4 全面 FGLS

虽然命令 xtpcse 提供了组间异方差与同期相关稳健的面板校正标准误差，但在进行 FGLS 估计时仅针对组内自相关，并未考虑组间异方差或同期相关。更为全面的 FGLS 估计则同时考虑这三个因素。为此，可先对方程(16.1)进行 OLS 估计，然而使用残差 $\{e_{it}\}$ 来估计 ε_{it} 的协方差矩阵，以此进行 FGLS 估计。还可进行迭代 FGLS 估计，即使用 FGLS 的残差再进行 FGLS 估计，不断迭代，直至收敛。相应的 Stata 命令为

```
xtgls y x1 x2 x3,panels(option) corr(option) igls
```

其中，选择项“panels(iid)”假定不同个体的扰动项为独立同分布的，“panels(het)”假定不同个体的扰动项相互独立但可以有不同的方差，而“panels(cor)”假定不同个体的扰动项同期相关且有不同的方差。选择项“corr(ar1)”对应于“$\rho_i=\rho$”的组内自相关情形，而“corr(psar1)”则允许每个面板有自己的自回归系数。选择项“igls”表示使用迭代式 FGLS。

继续以数据集 mus08cigar. dta 为例。首先，考虑同时存在组间异方差、同期相关以及组内自相关(自回归系数相同)的情形。

```
. xtgls lnc lnp lnpmin lny i.state t,panels(cor) cor(ar1)
```

```
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   common AR(1) coefficient for all panels  (0.7967)

Estimated covariances      =        55          Number of obs      =       300
Estimated autocorrelations =         1          Number of groups   =        10
Estimated coefficients     =        14          Time periods       =        30
                                                Wald chi2(13)      =    797.11
                                                Prob > chi2        =    0.0000
```

lnc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnp	-.3629823	.0239353	-15.17	0.000	-.4098947	-.3160699
lnpmin	.0258594	.0302743	0.85	0.393	-.033477	.0851959
lny	.5116673	.0733079	6.98	0.000	.3679865	.655348
state						
2	-.0719573	.0471313	-1.53	0.127	-.164333	.0204185
3	.0913241	.0260485	3.51	0.000	.04027	.1423782
4	-.1844595	.056133	-3.29	0.001	-.2944782	-.0744407
5	-.1466676	.0636223	-2.31	0.021	-.271365	-.0219701
6	.198174	.0424493	4.67	0.000	.114975	.2813731
7	.0861722	.1025173	0.84	0.401	-.1147579	.2871024
8	.0548621	.0386937	1.42	0.156	-.0209761	.1307003
9	.0227095	.0204346	1.11	0.266	-.0173416	.0627607
10	-.1279293	.0359818	-3.56	0.000	-.1984523	-.0574063
t	-.0448556	.0057821	-7.76	0.000	-.0561884	-.0335228
_cons	2.550302	.5964829	4.28	0.000	1.381217	3.719388

其次，考虑各组自回归系数不同的情形。

```
. xtgls lnc lnp lnpmin lny i.state t,panels(cor) cor(psar1)
```

```
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   panel-specific AR(1)

Estimated covariances      =        55          Number of obs      =       300
Estimated autocorrelations =        10          Number of groups   =        10
Estimated coefficients     =        14          Time periods       =        30
                                                Wald chi2(13)      =   1246.78
                                                Prob > chi2        =    0.0000
```

lnc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnp	-.3541126	.024395	-14.52	0.000	-.4019259	-.3062994
lnpmin	.0185409	.030352	0.61	0.541	-.040948	.0780297
lny	.5459469	.0712941	7.66	0.000	.406213	.6856808
state						
2	-.1229944	.0660784	-1.86	0.063	-.2525056	.0065169
3	-.3649205	.2134793	-1.71	0.087	-.7833323	.0534913
4	-.2405925	.0717201	-3.35	0.001	-.3811613	-.1000236
5	-.2086117	.0793916	-2.63	0.009	-.3642163	-.0530071
6	.1417767	.0643671	2.20	0.028	.0156194	.2679339
7	-.0337763	.1435772	-0.24	0.814	-.3151824	.2476297
8	-.0008845	.0647305	-0.01	0.989	-.1277539	.1259849
9	-.0350394	.0544958	-0.64	0.520	-.1418491	.0717703
10	-.1888756	.0569821	-3.31	0.001	-.3005584	-.0771928
t	-.0487174	.0055724	-8.74	0.000	-.059639	-.0377957
_cons	2.356635	.5905045	3.99	0.000	1.199267	3.514002

由以上两表可知，各种 FGLS 估计结果不尽相同。究竟使用何种 FGLS 估计，则取决于对组间异方差、组内自相关与组间同期相关的检验。总的来说，“OLS + 面板校正标准误差”最为稳健，全面的 FGLS 估计最有效率，而仅解决组内自相关的 FGLS 则介于二者之间。有关如何在稳健性与有效性之间进行权衡，参见第 7 章第 4 节。

16.5 组间异方差的检验

Greene (2000, p. 598) 提供了一个对组间异方差的沃尔德检验。考虑原假设“不同个体的扰动项方差均相等”，即 H_0: $\sigma_i^2 = \sigma^2 (i = 1, \cdots, n)$。在原假设成立的前提下，

$$\frac{\hat{\sigma}_i^2 - \sigma^2}{\sqrt{\mathrm{Var}(\hat{\sigma}_i^2)}} \xrightarrow{d} N(0,1) \tag{16.4}$$

其中，$\hat{\sigma}_i^2 = \sum_{t=1}^{T} e_{it}^2 / T$ 为 σ^2 的一致估计量，e_{it} 为对应于 ε_{it} 的残差。将上式平方可得

$$\frac{(\hat{\sigma}_i^2 - \sigma^2)^2}{\mathrm{Var}(\hat{\sigma}_i^2)} \xrightarrow{d} \chi^2(1) \tag{16.5}$$

另一方面，$\mathrm{Var}(\hat{\sigma}_i^2)$ 的一致估计量为

$$\widehat{\mathrm{Var}(\hat{\sigma}_i^2)} = \frac{1}{T} \frac{1}{T-1} \sum_{t=1}^{T} (e_{it}^2 - \hat{\sigma}_i^2)^2 \tag{16.6}$$

进一步假设每位个体的扰动项相互独立，可得沃尔德统计量：

$$W \equiv \sum_{i=1}^{n} \frac{(\hat{\sigma}_i^2 - \sigma^2)^2}{\widehat{\mathrm{Var}}(\hat{\sigma}_i^2)} \xrightarrow{d} \chi^2(n) \tag{16.7}$$

此沃尔德检验可通过非官方命令 xttest3 来实现，该命令只能在运行命令“xtreg,fe”或“xtgls”之后才能使用。

```
. ssc install xttest3     (下载安装命令 xttest3)
. quietly xtreg lnc lnp lnpmin lny i.state t,r fe
. xttest3
```

```
Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (10)  =     378.90
Prob>chi2 =      0.0000
```

```
. quietly xtgls lnc lnp lnpmin lny i.state t
. xttest3
```

```
Modified Wald test for groupwise heteroskedasticity
in cross-sectional time-series FGLS regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (10)  =     378.90
Prob>chi2 =      0.0000
```

由以上结果可知，无论在命令“xtreg,fe”或“xtgls”之后使用命令 xttest3，其统计量完全一样，均强烈拒绝同方差的原假设，认为存在组间异方差。

16.6 组内自相关的检验

Wooldridge (2002, p. 282 - 283)提供了一个对组内自相关的沃尔德检验。考虑原假设“不存在组内自相关”，即 H_0: $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) = 0 (t \neq s, \forall i)$。给定个体 i，对方程(16.1)进行一阶差分可得

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \Delta \varepsilon_{it} \tag{16.8}$$

在 $\{\varepsilon_{it}\}$ 不存在组内自相关的原假设下，方程(16.8)扰动项 $\Delta\varepsilon_{it}$ 的方差与自协方差(autocovariance)分别为

$$\mathrm{Var}(\Delta\varepsilon_{it}) = \mathrm{Var}(\varepsilon_{it} - \varepsilon_{i,t-1}) = \mathrm{Var}(\varepsilon_{it}) + \mathrm{Var}(\varepsilon_{i,t-1}) = 2\sigma_\varepsilon^2 \tag{16.9}$$

$$\begin{aligned} \mathrm{Cov}(\Delta\varepsilon_{it}, \Delta\varepsilon_{i,t-1}) &= \mathrm{Cov}(\varepsilon_{it} - \varepsilon_{i,t-1}, \varepsilon_{i,t-1} - \varepsilon_{i,t-2}) \\ &= -\mathrm{Cov}(\varepsilon_{i,t-1}, \varepsilon_{i,t-1}) = -\mathrm{Var}(\varepsilon_{i,t-1}) = -\sigma_\varepsilon^2 \end{aligned} \tag{16.10}$$

故自相关系数为

$$\mathrm{Corr}(\Delta\varepsilon_{it}, \Delta\varepsilon_{i,t-1}) = \frac{\mathrm{Cov}(\Delta\varepsilon_{it}, \Delta\varepsilon_{i,t-1})}{\mathrm{Var}(\Delta\varepsilon_{it})} = \frac{-\sigma_\varepsilon^2}{2\sigma_\varepsilon^2} = -0.5 \tag{16.11}$$

记 $\Delta\varepsilon_{it}$ 的样本值为 e_{it}(即一阶差分回归的残差)，对 e_{it} 进行一阶自回归：

$$e_{it} = \rho e_{i,t-1} + error_{it} \quad (i = 1, \cdots, n; t = 3, \cdots, T) \tag{16.12}$$

然后对原假设“H_0: $\rho = -0.5$”进行沃尔德检验(t 或 F 检验)。在 Stata 中,可以通过下载命令 xtserial 来实现这个检验:

```
net install st0039    (下载安装命令 xtserial)
xtserial y x1 x2 x3,output
```

其中,选择项“output”表示显示一阶差分回归的结果。

继续上面的例子。由于该命令不接受以“i.state”表示州虚拟变量,故首先生成州虚拟变量,以 state1-state10 来表示。

```
. tab state,gen(state)
```

U.S. state	Freq.	Percent	Cum.
1	30	10.00	10.00
2	30	10.00	20.00
3	30	10.00	30.00
4	30	10.00	40.00
5	30	10.00	50.00
6	30	10.00	60.00
7	30	10.00	70.00
8	30	10.00	80.00
9	30	10.00	90.00
10	30	10.00	100.00
Total	300	100.00	

```
. xtserial lnc lnp lnpmin lny state2 - state10 t
```

```
Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,        9) =      89.304
          Prob > F =       0.0000
```

结果强烈拒绝“不存在一阶组内自相关”的原假设。由于命令 xtserial 不允许使用“i.state”,故只能使用上文生成的州虚拟变量“state2-state10”。

16.7 组间同期相关的检验

考虑原假设“不存在组间同期相关”,即 H_0: $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = 0 (i \neq j, \forall t)$。如果此原假设成立,则根据残差计算的不同个体扰动项的相关系数应接近于 0。如果将这些相关系数排成一个矩阵,即“残差相关系数矩阵”(correlation matrix of residuals),则该矩阵非主对角线元素应离 0 不远。根据残差相关系数矩阵,可以设计以下几种检验。

Greene (2000, p. 601)提供了一个对组间同期相关的 Breusch-Pagan LM 检验,可由非官方命令 xttest2 来实现。此命令只能在运行命令“xtreg,fe”,“xtgls”或“ivreg2”之后使用。

```
. ssc install xttest2①    (下载安装命令 xttest2)
. quietly xtreg lnc lnp lnpmin lny t,fe
. xttest2
```

```
Correlation matrix of residuals:

            __e1      __e2      __e3      __e4      __e5      __e6      __e7      __e8      __e9     __e10
__e1     1.0000
__e2    -0.0937    1.0000
__e3     0.9592   -0.0621    1.0000
__e4    -0.4242    0.3875   -0.4670    1.0000
__e5    -0.5426    0.3441   -0.5872    0.5519    1.0000
__e6     0.0245    0.5696   -0.0405    0.5177    0.5805    1.0000
__e7    -0.7434    0.4153   -0.7509    0.5701    0.8446    0.4893    1.0000
__e8     0.5650    0.5380    0.5281    0.1007   -0.2150    0.4899   -0.3263    1.0000
__e9     0.8337    0.2859    0.8507   -0.2972   -0.3914    0.1548   -0.5800    0.7129    1.0000
__e10    0.7510    0.3314    0.7628    0.0002   -0.1575    0.2508   -0.4293    0.6318    0.8345    1.0000

Breusch-Pagan LM test of independence: chi2(45) =   376.963, Pr = 0.0000
Based on 30 complete observations over panel units
```

上表显示，Breusch-Pagan LM 统计量的 p 值为 0.000 0，故强烈拒绝“无同期相关”的原假设，认为存在同期相关。

以上 LM 检验的缺陷是，它仅适用于长面板。而检验组间同期相关的另一非官方命令 xtcsd 则也适用于 n 大 T 小的短面板，其中“csd”代表“cross-sectional dependence”（截面相关）。其原假设也是“不存在组间同期相关”。它包括三种检验方法，分别由 Friedman (1937)，Frees (1995，2004)，以及 Pesaran (2004) 所提出；前两种为半参数检验（参见第 27 章），而后一种为通常的参数检验。命令 xtcsd 须在运行命令 xtreg 之后才能使用。

```
ssc install xtcsd              (下载安装命令 xtcsd)
xtcsd,pesaran abs show         (Pesaran (2004)的检验，统计量服从标准正态分布)
xtcsd,friedman abs show        (Friedman (1937)的检验，统计量服从χ² 分布)
xtcsd,frees abs show           (Frees (1995,2004)的检验)
```

其中，选择项“show”表示显示残差的相关系数矩阵，而选择项“abs”表示显示该矩阵非主对角线元素的绝对值之平均。回到数据集 mus08cigar. dta 的例子：

```
. quietly xtreg lnc lnp lnpmin lny t,fe
. xtcsd,pes
```

```
Pesaran's test of cross sectional independence =      7.159, Pr = 0.0000
```

```
. xtcsd,fri
```

```
Friedman's test of cross sectional independence =     81.481, Pr = 0.0000
```

```
. xtcsd,fre abs show
```

① 如果以较早版本的命令 xttest2 来检验此数据集，可能会出现出错信息，诸如“Correlation matrix of residuals is singular，not possible with test”。此时，应使用命令“ssc install xttest2，replace”来下载最新版的命令 xttest2（选择项“replace”表示替换旧版命令）。

```
Correlation matrix of residuals:

          c1       c2       c3       c4       c5       c6       c7       c8       c9      c10
 r1   1.0000
 r2  -0.0937   1.0000
 r3   0.9592  -0.0621   1.0000
 r4  -0.4242   0.3875  -0.4670   1.0000
 r5  -0.5426   0.3441  -0.5872   0.5519   1.0000
 r6   0.0245   0.5696  -0.0405   0.5177   0.5805   1.0000
 r7  -0.7434   0.4153  -0.7509   0.5701   0.8446   0.4893   1.0000
 r8   0.5650   0.5380   0.5281   0.1007  -0.2150   0.4899  -0.3263   1.0000
 r9   0.8337   0.2859   0.8507  -0.2972  -0.3914   0.1548  -0.5800   0.7129   1.0000
r10   0.7510   0.3314   0.7628   0.0002  -0.1575   0.2508  -0.4293   0.6318   0.8345   1.0000

  Frees' test of cross sectional independence =     2.306
|---------------------------------------------------------|
  Critical values from Frees' Q distribution
                    alpha = 0.10 :   0.0861
                    alpha = 0.05 :   0.1119
                    alpha = 0.01 :   0.1598

Average absolute value of the off-diagonal elements =     0.466
```

以上三种检验的 p 值均小于 0.01，故强烈拒绝“无组间同期相关”的原假设，认为存在组间同期相关。上表最后一行显示，非主对角线元素绝对值的平均值高达 0.466。命令“xtcsd, fre abs show”所显示的残差相关系数矩阵与命令“xttest2”所显示完全相同。另外，还可以把面板数据排成“宽形”(wide form)，然后使用“似不相关回归”来检验组间截面相关(参见第 23 章)，但略为繁琐。

16.8 变系数模型

对于长面板数据，由于样本容量大，除了可以让每位个体拥有自己的截距项或时间趋势项外，还可以允许每位个体的回归方程斜率也不同，这被称为“变系数模型”。“变系数模型”可分为两大类，取决于将“可变系数”视为常数还是随机变量。

1. 将可变系数视为常数

假设 $y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}_i+\varepsilon_{it}$，其中 $\boldsymbol{\beta}_i$ 为个体 i 对应的系数。此时，可以对每个个体方程进行“分别回归”(separate regressions)。但如果不同个体的扰动项相关，则分别回归效率不高，因为它忽略了不同方程扰动项相关性的可用信息。有效率的做法是，把所有个体回归方程叠放(stack)，然后使用“似不相关回归”(SUR，参见第 23 章)对整个方程系统进行“系统估计”(system estimation)。

使用这个方法的缺点是，可能需要估计较多参数，从而损失自由度。作为一种折中，可以考虑“部分变系数模型”，即允许 $\boldsymbol{\beta}_i$ 中的部分系数(比如，研究者感兴趣的系数)依个体而变，而其余系数则不变。在这种情况下，不再适用 SUR，因为各个体方程除了扰动项相关外，还拥有部分相同的系数(跨方程约束)。此时，可以使用 LSDV 法，即在回归方程中，引入个体虚拟变量，以及虚拟变量与 $\boldsymbol{x}_{it}$ 中可变系数之解释变量的互动项。

继续前面的例子。假设人均收入(lny)的回归系数在每个州各不相同，而其他解释变量的系数相同。

```
. reg lnc lnp lnpmin lny i.state i.state#c.lny t,vce(cluster state)
```

```
Linear regression                               Number of obs =       300
                                                F(  2,     9) =         .
                                                Prob > F      =         .
                                                R-squared     =    0.9394
                                                Root MSE      =    .05301

                                   (Std. Err. adjusted for 10 clusters in state)
------------------------------------------------------------------------------
             |               Robust
         lnc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -.4153795   .1225736    -3.39   0.008    -.6926602   -.1380989
      lnpmin |   .0765648   .1063314     0.72   0.490    -.1639735    .3171032
         lny |   .6781285   .0888178     7.64   0.000     .4772086    .8790484
             |
       state |
           2 |   1.445143   .0460256    31.40   0.000     1.341026     1.54926
           3 |   -.109939   .0385508    -2.85   0.019     -.197147   -.0227311
           4 |   2.127437   .1178778    18.05   0.000     1.860779    2.394095
           5 |   1.894687   .0803023    23.59   0.000      1.71303    2.076343
           6 |   1.714489   .0646474    26.52   0.000     1.568247    1.860732
           7 |   4.817635   .1558296    30.92   0.000     4.465124    5.170146
           8 |   1.212741   .0803827    15.09   0.000     1.030903    1.394579
           9 |   .3182395   .0088129    36.11   0.000     .2983034    .3381756
          10 |   .5024975   .0815632     6.16   0.000     .3179888    .6870062
             |
 state#c.lny |
           2 |  -.1768319   .0043996   -40.19   0.000    -.1867846   -.1668792
           3 |   .0227827   .0045784     4.98   0.001     .0124256    .0331398
           4 |  -.2656664   .0102793   -25.84   0.000    -.2889197    -.242413
           5 |  -.2357472   .0084874   -27.78   0.000     -.254947   -.2165473
           6 |  -.1770753   .0054496   -32.49   0.000    -.1894032   -.1647474
           7 |  -.5347635   .0182173   -29.35   0.000    -.5759738   -.4935531
           8 |  -.1357293   .0098855   -13.73   0.000    -.1580919   -.1133667
           9 |  -.0355302   .0004363   -81.44   0.000    -.0365171   -.0345433
          10 |  -.0742162   .0084699    -8.76   0.000    -.0933764   -.0550559
             |
           t |  -.0463632    .007114    -6.52   0.000    -.0624563   -.0302702
       _cons |   1.168383   .8387717     1.39   0.197    -.7290501    3.065817
------------------------------------------------------------------------------
```

其中，“i.state”表示州虚拟变量，“i.state#c.lny”表示州虚拟变量与 lny 的互动项；“c”表示“continuous”，用以指明 lny 为连续型变量。这些变量在 Stata 中称为“因子变量”（factor variables），是变量窗口所列变量的扩展，使用方法参见“help fvvarlist”。

上表结果显示，州虚拟变量与 lny 的互动项均十分显著，故应使用变系数模型。

2. 随机系数模型（Random Coefficient Model）

在上一章考虑个体效应 u_i 或个体时间趋势 $\gamma_i t$ 时，曾分别将 u_i 或 γ_i 视为随机变量。类似地，可将系数（斜率）$\boldsymbol{\beta}_i$ 视为随机变量（仿佛从某个总体中获取随机样本，抽样后则取值不再随时间而变），并假设

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{v}_i \tag{16.13}$$

其中，$\boldsymbol{\beta}$ 为常数向量，而 $\boldsymbol{v}_i$ 为随机向量，且满足条件期望 $\mathrm{E}(\boldsymbol{v}_i \mid \boldsymbol{x}_i) = \mathbf{0}$（故影响斜率的随机因素 $\boldsymbol{v}_i$ 与解释变量 $\boldsymbol{x}_i$ 不相关），条件协方差矩阵 $\mathrm{Var}(\boldsymbol{v}_i \mid \boldsymbol{x}_i) = \boldsymbol{\Sigma}$（对角矩阵）。因此，

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it} = \boldsymbol{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{v}_i) + \varepsilon_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + (\boldsymbol{x}_{it}'\boldsymbol{v}_i + \varepsilon_{it}) \tag{16.14}$$

由于 $\mathrm{E}(\boldsymbol{v}_i \mid \boldsymbol{x}_i) = \mathbf{0}$，通过使用迭代期望定律，可以证明复合扰动项 $(\boldsymbol{x}_{it}'\boldsymbol{v}_i + \varepsilon_{it})$ 与解释变量 $\boldsymbol{x}_{it}$ 不相关（参见习题）。因此，OLS 是一致的。然而，复合扰动项 $(\boldsymbol{x}_{it}'\boldsymbol{v}_i + \varepsilon_{it})$ 的协方差矩阵为分块

对角矩阵，并非单位矩阵（类似于随机效应模型的情形，参见第 15 章）。Swamy（1970）提出用 FGLS 来估计此模型，即利用 OLS 残差来估计协方差矩阵中的参数，然后再使用 GLS。

随机系数模型的 Stata 命令为

```
xtrc y x1 x2 x3,betas
```

其中，选择项“betas”表示，显示对每组系数的估计。该命令提供了一个检验参数稳定性（test of parameter constancy）的 χ^2 统计量及其 p 值，其原假设为“H_0: $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n$”。如果拒绝此原假设，则认为应使用变系数模型。

继续前面的例子：

```
. xtrc lnc lnp lnpmin lny,beta
```

```
Random-coefficients regression                  Number of obs      =       300
Group variable: state                           Number of groups   =        10

                                                Obs per group: min =        30
                                                               avg =      30.0
                                                               max =        30

                                                Wald chi2(3)       =     99.29
                                                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
         lnc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |  -.4523322   .1322364    -3.42   0.001    -.7115108   -.1931536
      lnpmin |  -.0824106   .1196998    -0.69   0.491    -.3170178    .1521967
         lny |  -.0696917   .0531879    -1.31   0.190     -.173938    .0345545
       _cons |   7.833393   .5180469    15.12   0.000      6.81804    8.848746
------------------------------------------------------------------------------
Test of parameter constancy:     chi2(36) =  3611.15       Prob > chi2 = 0.0000
```

上表给出了对于共同的 $\boldsymbol{\beta}$ 系数向量估计值。最后一行的参数稳定性检验的 p 值为 0.000 0，故强烈拒绝“参数不变”的原假设，认为应使用变系数模型。对每组系数 $\boldsymbol{\beta}_i$ 的估计结果如下。

Group-specific coefficients

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Group 1						
lnp	-.6233233	.1659849	-3.76	0.000	-.9486476	-.2979989
lnpmin	.0943415	.1701196	0.55	0.579	-.2390868	.4277698
lny	.0980417	.0116712	8.40	0.000	.0751665	.1209168
_cons	6.242134	.2118045	29.47	0.000	5.827005	6.657264
Group 2						
lnp	-.2608889	.1671412	-1.56	0.119	-.5884796	.0667018
lnpmin	-.432141	.1654173	-2.61	0.009	-.7563529	-.1079291
lny	-.0678897	.0166591	-4.08	0.000	-.1005409	-.0352384
_cons	8.400657	.2507142	33.51	0.000	7.909266	8.892047
Group 3						
lnp	-.1361494	.1333002	-1.02	0.307	-.397413	.1251142
lnpmin	-.2787844	.1422594	-1.96	0.050	-.5576077	.0000388
lny	.1271675	.0093934	13.54	0.000	.1087568	.1455781
_cons	5.513523	.1886807	29.22	0.000	5.143716	5.883331
Group 4						
lnp	-.5621989	.1049334	-5.36	0.000	-.7678645	-.3565332
lnpmin	-.1648781	.1120782	-1.47	0.141	-.3845473	.0547912
lny	-.1911483	.0139909	-13.66	0.000	-.21857	-.1637266
_cons	9.707348	.2044615	47.48	0.000	9.306611	10.10809
Group 5						
lnp	-.7750139	.1123276	-6.90	0.000	-.9951719	-.5548559
lnpmin	.269899	.1201674	2.25	0.025	.0343751	.5054228
lny	-.1288682	.0146083	-8.82	0.000	-.1575	-.1002364
_cons	8.285053	.2041314	40.59	0.000	7.884963	8.685144
Group 6						
lnp	-.1837786	.0896636	-2.05	0.040	-.359516	-.0080413
lnpmin	-.1098901	.0926034	-1.19	0.235	-.2913895	.0716093
lny	-.1160046	.0100373	-11.56	0.000	-.1356772	-.0963319
_cons	7.343327	.1871318	39.24	0.000	6.976555	7.710098
Group 7						
lnp	-.5065479	.1197776	-4.23	0.000	-.7413077	-.2717881
lnpmin	.1068615	.1397386	0.76	0.444	-.1670212	.3807442
lny	-.4296878	.025671	-16.74	0.000	-.480002	-.3793736
_cons	10.69059	.3418134	31.28	0.000	10.02065	11.36054
Group 8						
lnp	-.2531896	.110291	-2.30	0.022	-.469356	-.0370233
lnpmin	-.2451675	.1257458	-1.95	0.051	-.4916248	.0012898
lny	-.0547215	.0133811	-4.09	0.000	-.080948	-.0284949
_cons	7.574233	.2314107	32.73	0.000	7.120677	8.02779
Group 9						
lnp	-.1528873	.1217504	-1.26	0.209	-.3915137	.0857392
lnpmin	-.3632993	.1235095	-2.94	0.003	-.6053735	-.121225
lny	.0726175	.0103202	7.04	0.000	.0523903	.0928447
_cons	6.425592	.2026637	31.71	0.000	6.028378	6.822805
Group 10						
lnp	-1.069344	.1829536	-5.84	0.000	-1.427927	-.7107616
lnpmin	.2989525	.1885074	1.59	0.113	-.0705152	.6684202
lny	-.006424	.0195543	-0.33	0.743	-.0447497	.0319016
_cons	8.151466	.2968129	27.46	0.000	7.569723	8.733208

16.9 面板工具变量法

虽然面板数据能在一定程度上解决遗漏变量(个体异质性)问题,但如果回归模型本身包含内生解释变量(endogenous regressors),则仍需使用工具变量法(参见第 10 章)。实际操作通常分为两步,即首先对模型进行变换以解决遗漏变量问题(比如,使用固定效应模型 FE 或一阶差分法 FD),然后对变换后的模型使用二阶段最小二乘法(2SLS)。当然,如果认为不存在遗漏变量问题,则可省略第一步。考虑以下几种情形。

1. 对固定效应模型先进行离差变换,再使用工具变量法

对于固定效应模型,组内估计量把$(y_{it}-\bar{y}_i)$对$(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)$进行 OLS 回归。现在假设$\boldsymbol{x}_{it}$包含内生解释变量,而$\boldsymbol{z}_{it}$为有效工具变量($\boldsymbol{x}_{it}$中的外生解释变量也被包括在$\boldsymbol{z}_{it}$中)。此时,可以使用工具变量$(\boldsymbol{z}_{it}-\bar{\boldsymbol{z}}_i)$,把$(y_{it}-\bar{y}_i)$对$(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)$进行 2SLS 回归。

对应的 Stata 命令为

```
xtivreg y [varlist1](varlist2 = varlist_iv),fe
```

其中,"varlist1"为外生解释变量,"varlist2"为内生解释变量,而"varlist_iv"为工具变量,选择项"fe"表示组内估计法。

2. 对固定效应模型先进行一阶差分,再使用工具变量法

对于固定效应模型,也可以先进行一阶差分,然后使用工具变量$(\boldsymbol{z}_{it}-\boldsymbol{z}_{i,t-1})$,把$(y_{it}-y_{i,t-1})$对$(\boldsymbol{x}_{it}-\boldsymbol{x}_{i,t-1})$进行 2SLS 回归。

对应的 Stata 命令为

```
xtivreg y [varlist1](varlist2 = varlist_iv),fd
```

其中,选择项"fd"表示一阶差分法。

3. 对随机效应模型使用工具变量法

此时,可以先对随机效应模型进行 FGLS 变换,然后对变换后的模型[①]进行 2SLS 回归。对应的 Stata 命令为

```
xtivreg y [varlist1](varlist2 = varlist_iv),re
```

其中,选择项"re"表示"随机效应"模型。

需要注意的是,Stata 命令 xtivreg 并不提供选择项"vce(robust)"来保证使用聚类稳健标准误。但可以使用选择项"vce(bootstrap)"以得到稳健的"自助标准误"(参见第 19 章)。

当工具变量个数多于内生解释变量个数时,对面板数据进行 GMM 估计会更有效率,但需要下载非官方 Stata 命令 xtivreg2 来执行。该命令只能处理固定效应模型,即先对模型进行 FE 或 FD 变换后,再对变换后的模型使用 GMM:

```
ssc install xtivreg2  (下载安装命令 xtivreg2)
xtivreg2 y [varlist1](varlist2 = varlist_iv),fe gmm (先进行 FE 变换)
xtivreg2 y [varlist1](varlist2 = varlist_iv),fd gmm (先进行 FD 变换)
```

---

① 对于工具变量$\boldsymbol{z}_{it}$也要做相应的变换。

4. 面板工具变量法的过度识别检验

对于面板工具变量法的过度识别检验(有关过度识别检验的概念,参见第 10 章),可通过非官方命令 xtoverid 来实现。该命令须在运行命令 xtreg, xtivreg, xtivreg2 或 xthtaylor(参见下文)之后才能使用,下载安装方法为"ssc install xtoverid",安装后可参考"help xtoverid"。

16.10 豪斯曼 - 泰勒估计量(选读)

固定效应模型的主要缺点是无法估计不随时间而变的变量系数。而这些不随时间变化的变量可能恰恰是我们感兴趣的,比如性别、受教育程度对工资的作用。如果有足够的工具变量,可以直接用工具变量法进行估计,即需要找到与内生解释变量相关,但与个体效应 u_i 无关的有效工具变量。

在上一章介绍的固定效应或随机效应模型中,要么全部解释变量与 u_i 相关(固定效应),要么全部解释变量都与 u_i 不相关(随机效应)。介于二者之间的混合情形是,某些解释变量与 u_i 相关,而其他解释变量与 u_i 不相关。在这种情况下,有可能使用工具变量法得到对不随时间变化的变量系数的一致估计。考虑以下面板模型:

$$y_{it} = \boldsymbol{x}'_{1,it}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2,it}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{1i}\boldsymbol{\delta}_1 + \boldsymbol{z}'_{2i}\boldsymbol{\delta}_2 + u_i + \varepsilon_{it} \tag{16.15}$$

其中,解释变量 $\boldsymbol{x}$ 随时间变化,而 $\boldsymbol{z}$ 不随时间变化。带下标 1 的变量(即 $\boldsymbol{x}_1$ 与 $\boldsymbol{z}_1$)为外生变量(与 u_i 不相关),而带下标 2 的变量(即 $\boldsymbol{x}_2$ 与 $\boldsymbol{z}_2$)为内生变量(与 u_i 相关)。所有解释变量均与 ε_{it} 不相关。由于 $\boldsymbol{x}_{1,it}$ 与 $\boldsymbol{z}_{1i}$ 外生,故可以用自己作为自己的工具变量。对于 $\boldsymbol{x}_{2,it}$,可以使用 $(\boldsymbol{x}_{2,it} - \bar{\boldsymbol{x}}_{2,i})$ 作为工具变量。显然,$\boldsymbol{x}_{2,it}$ 与 $(\boldsymbol{x}_{2,it} - \bar{\boldsymbol{x}}_{2,i})$ 相关;另一方面,根据迭代期望定律,

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{x}_{2,it} - \bar{\boldsymbol{x}}_{2,i})u_i] &= \mathrm{E}_{u_i}\mathrm{E}[(\boldsymbol{x}_{2,it} - \bar{\boldsymbol{x}}_{2,i})u_i \mid u_i] \\
&= \mathrm{E}_{u_i}\{u_i\mathrm{E}[(\boldsymbol{x}_{2,it} - \bar{\boldsymbol{x}}_{2,i}) \mid u_i]\} \\
&= \mathrm{E}_{u_i}\{u_i[\mathrm{E}(\boldsymbol{x}_{2,it} \mid u_i) - \mathrm{E}(\bar{\boldsymbol{x}}_{2,i} \mid u_i)]\} \\
&= \mathrm{E}_{u_i}\{u_i \cdot \boldsymbol{0}\} = \boldsymbol{0}
\end{aligned} \tag{16.16}$$

对于 $\boldsymbol{z}_{2i}$,则可以使用 $\bar{\boldsymbol{x}}_1$ 作为工具变量,即用随时间变化的外生变量的平均值作为不随时间变化的内生变量的工具变量。一般来说,$\boldsymbol{z}_{2i}$ 会与 $\bar{\boldsymbol{x}}_1$ 相关(二者同为 y_i 的解释变量);另一方面,根据定义,$\mathrm{Cov}(\bar{\boldsymbol{x}}_{1i}, u_i) = 0$。因此,$\bar{\boldsymbol{x}}_1$ 为有效工具变量。

显然,为了使用工具变量法,必须要求 $\boldsymbol{x}_{1i}$ 中所包含的外生变量个数比 $\boldsymbol{z}_{2i}$ 中所包含的内生变量个数更多。使用以上工具变量进行 2SLS 估计,就得到"豪斯曼 - 泰勒估计量"(Hausman and Taylor,1981)。Amemiya and MaCurdy(1986)则提出将 $(\boldsymbol{x}_{1,i1} - \bar{\boldsymbol{x}}_{1,i})$,$(\boldsymbol{x}_{1,i2} - \bar{\boldsymbol{x}}_{1,i})$,…,$(\boldsymbol{x}_{1,iT} - \bar{\boldsymbol{x}}_{1,i})$ 也作为工具变量使用,以增加估计的效率。

使用豪斯曼 - 泰勒估计量的重要前提是,所有解释变量均与 ε_{it} 不相关,而且部分解释变量与 u_i 不相关。在实证研究中,需要说明这些条件为何能够满足;否则,将导致不一致的估计。

豪斯曼 - 泰勒估计量的 Stata 命令为

```
xthtaylor depvar indepvars,endog(varlist) amacurdy
```

其中,选择项"endog(varlist)"用来指定内生解释变量。如果加上选择项"amacurdy",则使用 Amemiya-MaCurdy 估计量;否则,即为默认的豪斯曼 - 泰勒估计量。

16.11 动态面板

面板数据的一个优点是可以对个体的动态行为进行建模。有些经济理论认为，由于惯性或部分调整，个体的当前行为取决于过去行为，比如资本存量的调整。如果在面板模型中，解释变量包含了被解释变量的滞后值，则称之为“动态面板数据”（Dynamic Panel Data，简记 DPD）。对于动态面板数据，即使组内估计量（FE）也是不一致的（Nickell，1981）。比如，假设

$$y_{it}=\alpha+\rho y_{i,t-1}+\beta x_{it}+u_i+\varepsilon_{it}\quad(t=2,\cdots,T) \tag{16.17}$$

则其离差形式为

$$y_{it}-\bar{y}_i=\rho(y_{i,t-1}-\overline{Ly_i})+\beta(x_{it}-\bar{x}_i)+(\varepsilon_{it}-\bar{\varepsilon}_i)\quad(t=2,\cdots,T) \tag{16.18}$$

其中，$\bar{y}_i\equiv\frac{1}{T-1}\sum_{t=2}^{T}y_{it}$，$\overline{Ly_i}\equiv\frac{1}{T-1}\sum_{t=2}^{T}y_{i,t-1}$，$\bar{x}_i\equiv\frac{1}{T-1}\sum_{t=2}^{T}x_{it}$，$\bar{\varepsilon}_i\equiv\frac{1}{T-1}\sum_{t=2}^{T}\varepsilon_{it}$为时间平均值。显然，由于$\overline{Ly_i}$中包含$\{y_{i1},\cdots,y_{i,T-1}\}$的信息，而$\{y_{i1},\cdots,y_{i,T-1}\}$与$(\varepsilon_{it}-\bar{\varepsilon}_i)$相关，故$\overline{Ly_i}$肯定与$(\varepsilon_{it}-\bar{\varepsilon}_i)$相关。因此，FE 是不一致的，称之为“动态面板偏差”（dynamic panel bias）。显然，对于长面板而言，n 较小而 T 较大，故动态面板偏差较小（如果 $T\to\infty$，则$\overline{Ly_i}\xrightarrow{p}\mathrm{E}(y_i)$，$\bar{\varepsilon}_i\xrightarrow{p}0$），可通过校正偏差的方法得到一致估计，参见本章第 13 节。本节主要针对 n 较大而 T 较小的短动态面板（short dynamic panel）[①]。

1. 差分 GMM

更一般地，考虑以下动态面板模型：

$$y_{it}=\alpha+\rho y_{i,t-1}+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\varepsilon_{it}\quad(t=2,\cdots,T) \tag{16.19}$$

先作一阶差分以消去个体效应 u_i，可得

$$\Delta y_{it}=\rho\Delta y_{i,t-1}+\Delta\boldsymbol{x}_{it}'\boldsymbol{\beta}+\Delta\varepsilon_{it}\quad(t=2,\cdots,T) \tag{16.20}$$

然而，$\Delta y_{i,t-1}\equiv y_{i,t-1}-y_{i,t-2}$依然与 $\Delta\varepsilon_{it}\equiv\varepsilon_{it}-\varepsilon_{i,t-1}$相关，因为 $y_{i,t-1}$与 $\varepsilon_{i,t-1}$相关。因此，$\Delta y_{i,t-1}$为内生变量，需要寻找适当的工具变量才能得到一致估计。为此，Anderson and Hsiao（1981）提出使用 $y_{i,t-2}$作为 $\Delta y_{i,t-1}$的工具变量，然后进行 2SLS 估计，称为“Anderson-Hsiao 估计量”。显然，$y_{i,t-2}$与 $\Delta y_{i,t-1}=y_{i,t-1}-y_{i,t-2}$相关。如果$\{\varepsilon_{it}\}$不存在自相关（对此假设须进行检验，参见下节），则 $y_{i,t-2}$与 $\Delta\varepsilon_{it}=\varepsilon_{it}-\varepsilon_{i,t-1}$不相关（虽然 $y_{i,t-2}$依赖于 $\varepsilon_{i,t-2}$，但 ε_{it}，$\varepsilon_{i,t-1}$均与 $\varepsilon_{i,t-2}$不相关）。因此，在$\{\varepsilon_{it}\}$不存在自相关的前提下，$y_{i,t-2}$是有效工具变量。然而，根据同样的逻辑，更高阶的滞后变量$\{y_{i,t-3},y_{i,t-4},\cdots\}$也是有效工具变量，而 Anderson-Hsiao 估计量未加以利用，故不是最有效率的。Arellano and Bond（1991）使用所有可能的滞后变量作为工具变量（显然工具变量个数多于内生变量个数），进行 GMM 估计[②]。这就是“Arellano-Bond 估计量”，也被称为“差分 GMM”（Difference GMM）（因为是对差分后的方程进行 GMM 估计）。使用差分 GMM 的前提为扰动项$\{\varepsilon_{it}\}$不存在自相关，即 $\mathrm{Cov}(\varepsilon_{it},\varepsilon_{is})=0,t\neq s,\forall i$。

在使用高阶滞后变量进行 GMM 估计时，存在一个技术细节。记由工具变量组成的矩阵为 $\boldsymbol{Z}$，则工具变量 $y_{i,t-2}$将为 $\boldsymbol{Z}$ 贡献以下一个列向量：

---

① 本节的差分 GMM 与系统 GMM 方法仅适用于短动态面板，因为它们的大样本性质建立于在给定时间维度 T 的情况下，让横截面维度 n 趋于无穷。对于长动态面板，可考虑使用本章第 13 节的偏差校正 LSDV 法。

② 也可用 2SLS 来估计，但不如 GMM 有效率。

$$(\cdot\ y_{i1}\cdots y_{i,T-2})' \tag{16.21}$$

其中，第一个点“·”表示缺失值，这意味着将丢失第一行观测数据，损失样本容量。类似地，使用工具变量 $y_{i,t-3}$ 将损失前两行数据；以此类推，使用越高阶滞后作为工具变量，则损失的样本容量越多。为此，Holtz-Eakin et al (1988) 提出使用一系列的工具变量来表示 $y_{i,t-2}$，其中每个工具变量对应于一个时期，而将缺失值用0来代替：

$$\begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_{i1} & 0 & \cdots & 0 \\ 0 & y_{i2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & y_{i,T-2} \end{pmatrix} \tag{16.22}$$

这种形式的工具变量被称为“GMM式”(GMM-type，GMM-style)或“展开式”(uncollapsed)工具变量[①]，形如(16.22)。传统形式的工具变量则称为“标准式”(standard)、“IV式”(IV-style)或“collapsed”(折叠式)工具变量，形如(16.21)。在差分GMM以及下文的系统GMM估计中，为了尽量不损失样本容量，一般默认使用GMM式工具变量。但如果工具变量并非滞后变量(比如，额外的工具变量，或外生变量作为自己的工具变量)，则仍可使用标准式工具变量。由表达式(16.22)可知，如果使用GMM式工具变量，则工具变量的总数是时间维度 T 的二次函数，可能导致很多工具变量[②]。

然而，差分GMM在作差分时也会带来以下四个问题。

(1) 如果 $\boldsymbol{x}_{it}$ 仅为前定变量(predetermined)而非“严格外生”(strictly exogenous)，即虽然 $\boldsymbol{x}_{it}$ 与当期 $\boldsymbol{\varepsilon}_{it}$ 不相关，但与 $\boldsymbol{\varepsilon}_{i,t-1}$ 相关，则经过差分后，$\Delta\boldsymbol{x}_{it}=\boldsymbol{x}_{it}-\boldsymbol{x}_{i,t-1}$ 就可能与 $\Delta\boldsymbol{\varepsilon}_{it}=\boldsymbol{\varepsilon}_{it}-\boldsymbol{\varepsilon}_{i,t-1}$ 相关，致使 $\Delta\boldsymbol{x}_{it}$ 成为内生变量。此时，可以使用 $\{\boldsymbol{x}_{i,t-1},\boldsymbol{x}_{i,t-2},\cdots\}$ 作为 $\Delta\boldsymbol{x}_{it}$ 的工具变量。

(2) 如果 T 很大，则会有很多工具变量，容易出现弱工具变量(weak instruments)问题(通常滞后越多期则相关性越弱)，产生偏差。Windmeijer (2005) 使用差分GMM估计一个 $n=100$，$T=8$ 的动态面板，经模拟发现，将工具变量个数由28减为13可使系数估计值的平均偏差(average bias)下降40%。工具变量过多的另一后果是，它会弱化Hansen统计量(用于工具变量的过度识别检验)，甚至出现 p 值等于1的不可信结果。解决方法之一是在使用Stata命令 `xtabond` 时，限制最多使用 q 阶滞后变量作为工具变量，文献中对于如何选择工具变量个数并无明确指南。为了稳健起见，可使用两种(或多种)不同的工具变量个数，并比较其结果；如果二者结果相差很大，则应选择工具变量较少的那个模型。解决方法之二为使用折叠的IV式工具变量，而不使用展开的GMM式工具变量。

(3) 不随时间变化的变量 $\boldsymbol{z}_i$ 被消掉了，故差分GMM无法估计 $\boldsymbol{z}_i$ 的系数。

(4) 如果序列 $\{y_{it}\}$ 具有很强的持续性(highly persistent)，即一阶自回归系数接近于1，则 $y_{i,t-2}$ 与 $\Delta y_{i,t-1}\equiv y_{i,t-1}-y_{i,t-2}$ 的相关性可能很弱，导致弱工具变量问题。在极端情况下，如果 $\{y_{it}\}$ 为随机游走(random walk)，则 $\Delta y_{i,t-1}$ 为白噪声，完全与 $y_{i,t-2}$ 不相关，导致 $y_{i,t-2}$ 不再是 $\Delta y_{i,t-1}$ 的有效工具变量。因此，如果被解释变量的持续性很强，可能不适用差分GMM。

---

① 这个名词可能引起误解，因为使用标准式工具变量也可以进行GMM估计。

② 由于工具变量个数约为时间维度 T 的二次函数，故矩条件协方差矩阵中的元素个数为 T 的四次函数，在使用有限样本估计如此大的矩阵时可能有困难，有时导致所估计矩条件协方差矩阵不可逆(singular)，此时可使用广义逆矩阵(generalized inverse)，不影响GMM的估计。另一方面，如果Stata出现“singular”的提示，则意味着可能工具变量太多。

2. 水平 GMM

为了解决上述问题(3)与(4),Arellano and Bover(1995)重新回到了差分之前的水平方程(level equation)(16.19),并使用$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$作为$y_{i,t-1}$的工具变量①。显然,二者是相关的。另一方面,如果$\{\varepsilon_{it}\}$不存在自相关,则$E(\Delta y_{i,t-s},\varepsilon_{it})=E(y_{i,t-s},\varepsilon_{it})-E(y_{i,t-s-1}\varepsilon_{it})=0-0=0,s\geqslant 1$;但必须假设$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与个体效应$u_i$不相关,才能保证这些工具变量与水平方程的复合扰动项$(u_i+\varepsilon_{it})$不相关。假定$|\rho|<1$,则$\{y_{it}\}$将趋于某均衡点y_i^*(y_i^*取决于u_i),而"$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与$u_i$不相关"的假设意味着趋于均衡点的速度$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与固定效应$u_i$无关。因此,在整个样本期间,$\{y_{it}\}$应该在均衡点$y_i^*$附近,才能保证这个假定成立②。在以上条件都满足的情况下,可以使用$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$作为工具变量对水平方程进行 GMM 估计。这称为"水平 GMM"(Level GMM)。使用水平 GMM 的前提包括:扰动项$\{\varepsilon_{it}\}$不存在自相关;而且$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与个体效应$u_i$不相关。

3. 系统 GMM

Blundell and Bond(1998)则将差分 GMM 与水平 GMM 结合在一起,将差分方程与水平方程作为一个方程系统进行 GMM 估计③,称为"系统 GMM"(System GMM)。与差分 GMM 相比,系统 GMM 的优点是可以提高估计的效率,并且可以估计不随时间变化的变量$\boldsymbol{z}_i$的系数(因为系统 GMM 包含对水平方程的估计)。其缺点是,必须额外地假定$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与$u_i$无关;如果这个条件无法满足,则不能使用系统 GMM。使用系统 GMM 的前提包括(与水平 GMM 相同):扰动项$\{\varepsilon_{it}\}$不存在自相关;而且$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$与个体效应$u_i$不相关④。

在更一般的动态面板模型中,解释变量可以包括被解释变量的多阶滞后值,

$$y_{it}=\alpha+\rho_1 y_{i,t-1}+\rho_2 y_{i,t-2}+\cdots+\rho_p y_{i,t-p}+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\varepsilon_{it} \tag{16.23}$$

对此,可以基本沿用以上的方法。在以上的所有 GMM 估计中,都可以指定额外的工具变量(在滞后变量及差分滞后变量以外的工具变量)。另外,在水平方程中,如果$\boldsymbol{x}_{it}$也包括内生变量,则同样可以用其滞后值作为工具变量。

例 Shen and Yao (2008)使用动态面板数据来分析村庄选举对该村居民收入分配的影响。

例 Acemoglu et al (2008) 使用差分 GMM 估计跨国面板数据,发现人均收入(per capita income)对民主(democracy)的作用不显著。然而,Che et al (2013)使用系统 GMM 估计同样的数据集,却发现人均收入对民主有显著的正作用。Che et al (2013)认为差分 GMM 存在弱工具变量问题,而系统 GMM 比差分 GMM 更有效率。

16.12 动态面板的 Stata 命令及实例

差分 GMM 的 Stata 命令为

---

① 这种做法正好与差分 GMM 相反。差分 GMM 以水平变量$\{y_{i,t-2},y_{i,t-3},\cdots\}$作为差分变量$\Delta y_{i,t-1}$的工具变量;而水平 GMM 则以差分变量$\{\Delta y_{i,t-1},\Delta y_{i,t-2},\cdots\}$作为水平变量$y_{i,t-1}$的工具变量。

② 经济理论(比如,Solow 增长模型)通常认为,趋于均衡点(steady state)的速度取决于均衡点y_i^*的位置,而y_i^*又依赖于u_i,故收敛速度与个体特征u_i相关。只有在均衡点y_i^*附近,收敛速度才与y_i^*及u_i无关。

③ 有关方程系统,参见第 23 - 24 章。

④ 使用系统 GMM 的两个前提(尤其是第二个前提)常被遗忘,故在此重复强调一下。然而,第二个前提目前尚无法进行严格的统计检验,只能根据经济常识判断经济变量是否在稳态(steady state)附近;比如,发达国家的经济体很可能在稳态附近,而转型经济体或发展中国家则可能离稳态较远(由于频繁的制度变迁等原因)。

```
xtabond depvar indepvars,lags(p) maxldep(q) twostep vce(robust) pre(varlist) endogenous(varlist) inst(varlist)
```

其中,"depvar"为被解释变量,"indepvars"为满足严格外生性的解释变量。选择项"lags(p)"表示使用被解释变量的 p 阶滞后值作为解释变量,默认值为 $p=1$;"maxldep(q)"表示最多使用 q 阶被解释变量的滞后值作为工具变量,默认值为使用所有可能的滞后值。选择项"twostep"表示使用 GMM,默认方法为 2SLS[①];选择项"pre(varlist)","endogenous(varlist)","inst(varlist)"分别用来指定前定变量、内生变量与额外的工具变量。选择项"vce(robust)"表示使用稳健标准误,该稳健标准误允许 ε_{it} 存在异方差,而且针对两步 GMM 的估计进行了调整,在 Stata 中称为"WC-Robust Standard Error"(Windmeijer, 2005;其中 WC 表示"Windmeijer bias-corrected")。

系统 GMM 的 Stata 命令为[②]

```
xtdpdsys depvar [indepvars],lags(p) maxldep(q) twostep vce(robust) pre(varlist) endogenous(varlist) inst(varlist)
```

下面以数据集 mus08psidextract.dta 为例[③]。该面板数据集来自 Panel Study of Income Dynamics (PSID)[④],包含 595 名美国工人 1976—1982 年有关工资的以下变量($n=595, T=7$ 的短面板):lwage(工资对数),ed(教育年限),exp(工龄),exp2(工龄平方),wks(weeks worked,工作周数),occ(是否蓝领工人),ind(是否在制造业工作),ms(婚否),south(是否在美国南方),smsa(是否住大城市),union(是否由工会合同确定工资),fem(是否女性),blk(是否黑人)。

首先以差分 GMM 估计以下动态面板模型:

$$\begin{aligned}\text{lwage}_{it} = {} & \alpha + \rho_1 \text{lwage}_{i,t-1} + \rho_2 \text{lwage}_{i,t-2} + \beta_1 \text{occ}_{it} + \beta_2 \text{south}_{it} + \beta_3 \text{smsa}_{it} + \\ & \beta_4 \text{ind}_{it} + \beta_5 \text{wks}_{it} + \beta_6 \text{wks}_{i,t-1} + \beta_7 \text{ms}_{it} + \beta_8 \text{union}_{it} + u_i + \varepsilon_{it}\end{aligned} \tag{16.24}$$

其中,occ,south,smsa,ind 被视为外生解释变量,wks 及其一阶滞后被视为前定解释变量,而 ms,union 被视为内生解释变量[⑤]。由于差分 GMM 首先对原方程进行差分,所有不随时间变化的变量都将无法估计(比如 ed,fem,blk),故不包括在模型中。

Arellano-Bond 估计量的 Stata 命令为

```
. use mus08psidextract.dta,clear
. xtabond lwage occ south smsa ind,lags(2) maxldep(3) pre(wks,lag(1,2)) endogenous(ms,lag(0,2)) endogenous(union,lag(0,2)) twostep vce(robust)
. estimates store DiffGMM
```

① 可以将 2SLS 视为"one-step GMM"。

② 也可使用非官方命令 xtabond2 进行差分 GMM 或系统 GMM 估计,下载方法为"ssc install xtabond2",参见下文。

③ 此例来自 Cameron and Trivedi (2009)。

④ 该数据集最早为 Baltagi and Khanti-Akom (1990)所使用。

⑤ 为了演示的目的,解释变量中未包括 exp,因为对于这个数据集而言,很难区分工龄与工资滞后值对工资的影响。如果在解释变量中同时包括工龄与工资滞后值,将导致估计不准确。

```
Arellano-Bond dynamic panel-data estimation  Number of obs      =      2380
Group variable: id                           Number of groups   =       595
Time variable: t
                                             Obs per group:  min =         4
                                                             avg =         4
                                                             max =         4

Number of instruments =     40               Wald chi2(10)      =   1287.77
                                             Prob > chi2        =    0.0000
Two-step results
                                   (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |              WC-Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L1. |    .611753   .0373491    16.38   0.000     .5385501    .6849559
         L2. |   .2409058   .0319939     7.53   0.000     .1781989    .3036127
             |
         wks |
         --. |  -.0159751   .0082523    -1.94   0.053    -.0321493     .000199
         L1. |   .0039944   .0027425     1.46   0.145    -.0013807    .0093695
             |
          ms |   .1859324    .144458     1.29   0.198       -.0972    .4690649
       union |  -.1531329   .1677842    -0.91   0.361    -.4819839    .1757181
         occ |  -.0357509   .0347705    -1.03   0.304    -.1038999     .032398
       south |  -.0250368   .2150806    -0.12   0.907     -.446587    .3965134
        smsa |  -.0848223   .0525243    -1.61   0.106     -.187768    .0181235
         ind |   .0227008   .0424207     0.54   0.593    -.0604422    .1058437
       _cons |   1.639999   .4981019     3.29   0.001     .6637377    2.616261
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/4).lwage L(1/2).L.wks L(2/3).ms L(2/3).union
        Standard: D.occ D.south D.smsa D.ind
Instruments for level equation
        Standard: _cons
```

其中，选择项“lags(2)”表示解释变量中包含被解释变量的一阶与二阶滞后；选择项“maxldep(3)”表示最多使用被解释变量的三个滞后值作为工具变量①；选择项“pre(wks,lag(1,2))”②表示变量 wks 及其一阶滞后 L. wks 为前定解释变量，而使用其两个更高阶滞后值（即二阶与三阶）为工具变量；选择项“endogenous(ms,lag(0,2))”③指定变量 ms 为内生解释变量（其中的“0”表示内生解释变量不包含 ms 的任何滞后），而最多使用其两个更高阶滞后值（即二阶与三阶）为工具变量④。

上表的下部显示了所使用的全部工具变量，比如，“L(2/4).lwage”表示 lwage 的二至四阶滞后（如果存在），即$\text{lwage}_{i,t-2}$，$\text{lwage}_{i,t-3}$，$\text{lwage}_{i,t-4}$，以此类推。此表的左上部显示，该差分 GMM 估计共使用了 40 个工具变量（矩条件）。

---

① 比如，当 $T=7$ 时，使用 $\text{lwage}_{i,5}$，$\text{lwage}_{i,4}$，$\text{lwage}_{i,3}$ 为工具变量。

② 该选择项的一般格式为“pre(varlist [,lagstruct(prelags,premaxlags)])”，其中“prelags”指定把直至“prelags”阶的滞后值也作为解释变量（默认值 prelags=0），而“premaxlags”则指定最多使用“premaxlags”个更高阶滞后值作为工具变量。

③ 该选择项的一般格式为“endogenous(varlist [,lagstruct(endlags,endmaxlags)])”，其中“endlags”指定把直至“endlags”阶的滞后值也作为解释变量（默认值 endlags=0），而“endmaxlags”则指定最多使用“endmaxlags”个更高阶滞后值作为工具变量。

④ 虽然原模型仅包括 ms 的当期变量，但在差分方程中则包含 ms 的一阶滞后值，故两个更高阶滞后值为二阶与三阶滞后值。

表中结果显示,被解释变量的一阶与二阶滞后值都很显著,而其他解释变量则不太显著。

作为一致估计,差分 GMM 能够成立的前提是,扰动项$\{\varepsilon_{it}\}$不存在自相关。对此要进行检验。即使原假设"扰动项$\{\varepsilon_{it}\}$无自相关"成立,"扰动项的一阶差分"(first-differenced errors)仍将存在一阶自相关,因为

$$\mathrm{Cov}(\Delta\varepsilon_{it},\Delta\varepsilon_{i,t-1})=\mathrm{Cov}(\varepsilon_{it}-\varepsilon_{i,t-1},\varepsilon_{i,t-1}-\varepsilon_{i,t-2})=-\mathrm{Cov}(\varepsilon_{i,t-1},\varepsilon_{i,t-1})=-\mathrm{Var}(\varepsilon_{i,t-1})\neq 0 \tag{16.25}$$

但扰动项的差分将不存在二阶或更高阶的自相关,即$\mathrm{Cov}(\Delta\varepsilon_{it},\Delta\varepsilon_{i,t-k})=0,k\geqslant 2$。为此,可以通过检验扰动项的差分是否存在一阶与二阶自相关,来检验原假设。其 Stata 命令为

```
. estat abond
```

```
Arellano-Bond test for zero autocorrelation in first-differenced errors
+----------------------------+
| Order |     z    Prob > z  |
|-------+--------------------|
|     1 | -4.5244   0.0000   |
|     2 | -1.6041   0.1087   |
+----------------------------+
H0: no autocorrelation
```

结果显示,扰动项的差分存在一阶自相关,但不存在二阶自相关,故接受原假设"扰动项$\{\varepsilon_{it}\}$无自相关",可以使用差分 GMM。

也可以检验扰动项的差分是否存在更高阶的自相关:

```
. estat abond,artests(3)
```

```
Arellano-Bond test for zero autocorrelation in first-differenced errors
+----------------------------+
| Order |     z    Prob > z  |
|-------+--------------------|
|     1 | -4.5244   0.0000   |
|     2 | -1.6041   0.1087   |
|     3 |  .35729   0.7209   |
+----------------------------+
H0: no autocorrelation
```

其中,选择项"artests(3)"表示,检验扰动项的差分$\Delta\varepsilon_{it}$是否存在一阶至三阶自相关;默认值为检验至二阶自相关,即"artests(2)"。结果显示,也不存在三阶自相关,故可以接受原假设。

其次,由于此差分 GMM 使用了 40 个工具变量,故需要进行过度识别检验。为此,我们重新运行以上"xtabond"命令,但略去选择项"vce(robust)"①。

```
. quietly xtabond lwage occ south smsa ind,lags(2) maxldep(3) pre(wks,
lag(1,2)) endogenous(ms,lag(0,2)) endogenous(union,lag(0,2)) twostep
. estat sargan
```

```
Sargan test of overidentifying restrictions
        H0: overidentifying restrictions are valid

        chi2(29)     =  39.87571
        Prob > chi2  =    0.0860
```

① 如果使用选择项"vce(robust)",则无法运用下面的"estat sargan"命令,因为该命令假定扰动项ε_{it}为独立同分布(在独立同分布的情况下,不需要使用稳健标准误)。

上表显示，在5%的显著性水平上，无法拒绝“所有工具变量均有效”的原假设，因为 p 值 = 0.086 > 0.05。

下面使用系统 GMM 来估计以上模型：

```
. xtdpdsys lwage occ south smsa ind,lags(2) maxldep(3) pre(wks,lag(1,
2)) endogenous(ms,lag(0,2)) endogenous(union,lag(0,2)) twostep vce
(robust)
. estimates store SysGMM
```

```
System dynamic panel-data estimation          Number of obs      =      2975
Group variable: id                            Number of groups   =       595
Time variable: t
                                              Obs per group:  min =         5
                                                              avg =         5
                                                              max =         5

Number of instruments =     60                Wald chi2(10)      =   2270.88
                                              Prob > chi2        =    0.0000
Two-step results
```

lwage	Coef.	WC-Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
L1.	.6017533	.0291502	20.64	0.000	.5446199	.6588866
L2.	.2880537	.0285319	10.10	0.000	.2321322	.3439752
wks						
--.	-.0014979	.0056143	-0.27	0.790	-.0125017	.009506
L1.	.0006786	.0015694	0.43	0.665	-.0023973	.0037545
ms	.0395337	.0558543	0.71	0.479	-.0699386	.1490061
union	-.0422409	.0719919	-0.59	0.557	-.1833423	.0988606
occ	-.0508803	.0331149	-1.54	0.124	-.1157843	.0140237
south	-.1062817	.083753	-1.27	0.204	-.2704346	.0578713
smsa	-.0483567	.0479016	-1.01	0.313	-.1422422	.0455288
ind	.0144749	.031448	0.46	0.645	-.0471621	.0761118
_cons	.9584113	.3632287	2.64	0.008	.2464961	1.670327

```
Instruments for differenced equation
        GMM-type: L(2/4).lwage L(1/2).L.wks L(2/3).ms L(2/3).union
        Standard: D.occ D.south D.smsa D.ind
Instruments for level equation
        GMM-type: LD.lwage LD.wks LD.ms LD.union
        Standard: _cons
```

对比差分 GMM 与系统 GMM 的系数估计值与标准误：

```
. estimates table DiffGMM SysGMM,b se
```

Variable	DiffGMM	SysGMM
lwage		
L1.	.61175299	.60175325
	.03734912	.0291502
L2.	.24090583	.28805368
	.0319939	.02853189
wks		
--.	-.01597512	-.00149786
	.00825228	.00561431
L1.	.00399441	.0006786
	.00274247	.00156937
ms	.18593243	.03953373
	.14445799	.05585427
union	-.15313291	-.04224089
	.16778419	.07199186
occ	-.03575093	-.05088031
	.03477051	.03311489
south	-.02503681	-.10628169
	.2150806	.08375304
smsa	-.08482225	-.04835667
	.05252431	.04790164
ind	.02270076	.01447487
	.04242066	.031448
_cons	1.6399994	.95841132
	.49810188	.36322872

legend: b/se

从上表可知，系统 GMM 的系数估计值与差分 GMM 很接近，但前者的标准误比后者更小。也许因为使用了更多的工具变量（共 60 个），系统 GMM 估计得更准确些。下面对系统 GMM 扰动项的自相关性进行检验：

```
. estat abond
```

Arellano-Bond test for zero autocorrelation in first-differenced errors

Order	z	Prob > z
1	-4.1072	0.0000
2	-2.149	0.0316

H0: no autocorrelation

上表显示，可以在 5% 的显著性水平上拒绝“扰动项差分的二阶自相关系数为 0 的假设”，因为 p 值 $=0.03<0.05$。因此，可以拒绝“扰动项 $\{\varepsilon_{it}\}$ 无自相关”的原假设。这使得我们怀疑系统 GMM 是否适用。

继续进行过度识别检验：

```
. quietly xtdpdsys lwage occ south smsa ind,lags(2) maxldep(3) pre(wks,
lag(1,2)) endogenous(ms,lag(0,2)) endogenous(union,lag(0,2)) twostep
. estat sargan
```

```
Sargan test of overidentifying restrictions
        H0: overidentifying restrictions are valid

        chi2(49)     =  84.02056
        Prob > chi2  =    0.0014
```

结果强烈拒绝“所有工具变量都有效”的原假设，这意味着某些新增工具变量与扰动项相关，不是有效工具变量。

以上两个检验结果表明，对于这个数据集而言，虽然系统 GMM 可能更有效率，但其成立的前提似乎并不满足。然而，这也可能因为模型设定不当。为了解决扰动项自相关问题，我们在解释变量中引入被解释变量的三阶滞后[①]；为了改善工具变量的有效性，对于内生解释变量 ms，union 仅使用一个更高阶滞后值（即二阶滞后）为工具变量，但最多使用被解释变量的五个滞后值作为工具变量，重新进行系统 GMM 估计与检验：

```
. xtdpdsys lwage occ south smsa ind,lags(3) maxldep(5) pre(wks,lag (1,2))
endogenous(ms,lag(0,1)) endogenous(union,lag(0,1)) twostep vce(robust)
```

```
System dynamic panel-data estimation          Number of obs      =      2380
Group variable: id                            Number of groups   =       595
Time variable: t
                                              Obs per group:  min =         4
                                                              avg =         4
                                                              max =         4

Number of instruments =     45                Wald chi2(11)      =   2236.60
                                              Prob > chi2        =    0.0000
Two-step results
```

lwage	Coef.	WC-Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
L1.	.5851347	.044352	13.19	0.000	.4982063	.6720631
L2.	.2115442	.0347287	6.09	0.000	.1434771	.2796113
L3.	.094682	.0329862	2.87	0.004	.0300304	.1593337
wks						
--.	-.0034873	.0052114	-0.67	0.503	-.0137015	.006727
L1.	.0030127	.0017871	1.69	0.092	-.0004899	.0065154
ms	.0666652	.0658209	1.01	0.311	-.0623414	.1956718
union	.0167684	.0848163	0.20	0.843	-.1494686	.1830053
occ	-.0764385	.0334374	-2.29	0.022	-.1419746	-.0109024
south	-.1013125	.0765927	-1.32	0.186	-.2514313	.0488064
smsa	-.0706818	.0611874	-1.16	0.248	-.190607	.0492434
ind	.0439888	.0386347	1.14	0.255	-.0317339	.1197114
_cons	.9141559	.3503796	2.61	0.009	.2274245	1.600887

```
Instruments for differenced equation
        GMM-type: L(2/6).lwage L(1/2).L.wks L(2/2).ms L(2/2).union
        Standard: D.occ D.south D.smsa D.ind
Instruments for level equation
        GMM-type: LD.lwage LD.wks LD.ms LD.union
        Standard: _cons
```

从上表可知，被解释变量的三阶滞后均很显著。另外，工具变量的个数由 60 减少为 45。

```
. estat abond
```

① 解决“扰动项自相关”问题的另一方法是使用命令 xtdpd。该命令允许动态面板模型的扰动项 ε_{it} 服从较低阶的移动平均过程（a moving average process of low order），但句型更为复杂，参见“help xtdpd”。

```
Arellano-Bond test for zero autocorrelation in first-differenced errors
+----------------------------------+
| Order |     z      Prob > z      |
|-------+--------------------------|
|   1   |  -7.1406    0.0000       |
|   2   |  -1.0119    0.3116       |
+----------------------------------+
  H0: no autocorrelation
```

上表显示,可以接受扰动项无自相关的原假设。

```
. quietly xtdpdsys lwage occ south smsa ind,lags(3) maxldep(5) pre(wks,
lag(1,2)) endogenous(ms,lag(0,1)) endogenous(union,lag(0,1)) twostep
. estat sargan
```

```
Sargan test of overidentifying restrictions
        H0: overidentifying restrictions are valid

        chi2(33)     =  46.83873
        Prob > chi2  =    0.0559
```

结果显示,可以在 5% 的显著性水平上接受“所有工具变量都有效”的原假设。这表明,对修正的模型设定,可以进行系统 GMM 估计。

由于官方命令 xtabond 与 xtdpdsys 均不提供异方差稳健的 Hansen 统计量(仅提供基于 iid 假设的 Sargan 统计量),故下面转而使用非官方命令 xtabond2。该命令的格式与官方命令 xtabond 有所不同:

```
xtabond2 y l.y l2.y x1 x2 x3, gmmstyle(varlist) ivstyle(varlist)
nolevel twostep vce(robust)
```

其中,“y”为被解释变量,而“l.y l2.y x1 x2 x3”为解释变量(这里假设被解释变量的两阶滞后为解释变量)。选择项“gmmstyle(varlist)”指定 GMM 式工具变量,而选择项“ivstyle(varlist)”指定 IV 式工具变量。选择项“nolevel”表示不估计水平方程,即为差分 GMM;默认为系统 GMM。选择项“twostep”与“vce(robust)”的含义与官方命令 xtabond 相同。更多说明,参见“help xtabond2”以及 Roodman (2009)。

为了演示的目的,下面使用 xtabond2 来复制本节的差分 GMM 估计结果。

```
. ssc install xtabond2  (下载安装命令 xtabond2)
. xtabond2 lwage L(1/2).lwage L(0/1).wks ms union occ south smsa ind,
gmm(lwage, lag(2 4)) gmm(wks ms union, lag(2 3)) iv(occ south smsa ind)
nolevel twostep robust
```

其中,选择项“gmm(lwage,lag(2 4))”表示使用变量 lwage 的 2—4 阶滞后作为 GMM 式工具变量,选择项“gmm(wks ms union,lag(2 3))”表示使用变量 wks,ms 与 union 的 2—3 阶滞后作为 GMM 式工具变量,而选择项“iv(occ south smsa ind)”表示使用变量 occ,south,smsa 与 ind 作为自身的 IV 式工具变量(假设它们为外生变量)。

```
Dynamic panel-data estimation, two-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =      2380
Time variable : t                               Number of groups   =       595
Number of instruments = 39                      Obs per group: min =         4
Wald chi2(10) =   1287.77                                      avg =      4.00
Prob > chi2   =     0.000                                      max =         4
------------------------------------------------------------------------------
             |              Corrected
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L1. |    .611753   .0373491    16.38   0.000     .5385501    .6849559
         L2. |   .2409058   .0319939     7.53   0.000     .1781989    .3036127
             |
         wks |
         --. |  -.0159751   .0082523    -1.94   0.053    -.0321493     .000199
         L1. |   .0039944   .0027425     1.46   0.145    -.0013807    .0093695
             |
          ms |   .1859324    .144458     1.29   0.198       -.0972    .4690649
       union |  -.1531329   .1677842    -0.91   0.361    -.4819839    .1757181
         occ |  -.0357509   .0347705    -1.03   0.304    -.1038999     .032398
       south |  -.0250368   .2150806    -0.12   0.907     -.446587    .3965134
        smsa |  -.0848223   .0525243    -1.61   0.106     -.187768    .0181235
         ind |   .0227008   .0424207     0.54   0.593    -.0604422    .1058437
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(occ south smsa ind)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(2/3).(wks ms union)
    L(2/4).lwage
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -4.52  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -1.60  Pr > z =  0.109
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(29)   =  59.55  Prob > chi2 =  0.001
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(29)   =  39.88  Prob > chi2 =  0.086
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  gmm(lwage, lag(2 4))
    Hansen test excluding group:     chi2(18)   =  23.59  Prob > chi2 =  0.169
    Difference (null H = exogenous): chi2(11)   =  16.29  Prob > chi2 =  0.131
  gmm(wks ms union, lag(2 3))
    Hansen test excluding group:     chi2(5)    =   6.43  Prob > chi2 =  0.266
    Difference (null H = exogenous): chi2(24)   =  33.44  Prob > chi2 =  0.095
  iv(occ south smsa ind)
    Hansen test excluding group:     chi2(25)   =  28.00  Prob > chi2 =  0.308
    Difference (null H = exogenous): chi2(4)    =  11.87  Prob > chi2 =  0.018
```

从上表可知，使用 xtabond2 得到的系数与标准误估计值与上文使用 xtabond 的结果完全相同。然而，上表中的 Sargan 统计量为 59.55，Hansen 统计量为 39.88；却与 xtabond 的 Sargan 统计量相左（碰巧也是 39.88）；尽管不影响检验结果。一种可能性是，xtabond 汇报的 Sargan 统计量其实就是 Hansen 统计量。另外，正如上表所示，Sargan 统计量与 Hansen 统计量各有千秋。尽管前者不稳健，但不受工具变量过多的影响；而后者虽然稳健，却可能因工具变量过多而失效。xtabond2 的另一优势是，它还可检验部分工具变量的外生性，即上表中的“Difference-in-Hansen tests”（参见第 10 章）。

16.13 偏差校正 LSDV 法

上节介绍的差分 GMM 与系统 GMM 方法主要适用于短动态面板。这是因为，虽然基于 IV 或 GMM 的估计方法是一致估计量（即当 $n\to\infty$ 时，没有偏差），但对于 n 较小而 T 较大的长面板则可能存在较严重的偏差。Nickell（1981）证明，动态面板偏差在数量级上与 T^{-1} 相当（通常记为 $O(T^{-1})$），故当 $T\to\infty$ 时，动态面板偏差趋向于 0。因此，对于长面板，可考虑使用以下“偏差校正 LSDV 法”（Biased - corrected LSDV，简记 LSDVC）。蒙特卡罗模拟（参见第 19 章）结果显示，对于 n 较小的长面板，无论在偏差大小还是均方误差（RMSE）方面，LSDVC 法都明显优于差分 GMM 或系统 GMM（Kiviet，1995；Judson and Owen，1999）。

LSDVC 法的基本思想是，首先使用 LSDV 法估计动态面板模型，记估计系数为 $\hat{\boldsymbol{\beta}}_{\text{LSDV}}$；其次，估计 LSDV 法的偏差，记为 $\widehat{\boldsymbol{Bias}}$；最后，将 LSDV 系数估计值减去此偏差，即得到偏差校正后的一致估计：

$$\hat{\boldsymbol{\beta}}_{\text{LSDVC}}=\hat{\boldsymbol{\beta}}_{\text{LSDV}}-\widehat{\boldsymbol{Bias}} \tag{16.26}$$

其中，估计量 $\hat{\boldsymbol{\beta}}_{\text{LSDVC}}$ 的标准误差可通过自助法得到（参见第 19 章）。LSDVC 法的局限是，它要求所有解释变量都是严格外生的，而差分 GMM 或系统 GMM 可通过引入工具变量来解决内生变量或前定变量的问题。

LSDVC 法可通过非官方命令 xtlsdvc 来实现（Bruno，2005），其基本格式为

```
xtlsdvc y x1 x2 x3,initial(estimator) bias(#) vcov(#)
```

此命令仅允许以被解释变量的一阶滞后作为解释变量。其中，选择项“initial(estimator)”用来指定偏差校正的初始值；比如，“i(ah)”表示使用 Anderson-Hsiao 估计量作为初始值，“i(ab)”表示使用 Arellano-Bond 差分 GMM 估计量作为初始值，而“i(bb)”表示使用 Blundell-Bond 系统 GMM 估计量作为初始值。选择项“bias(#)”用来指定偏差校正的精确度；比如，“bias(1)”的精确度为 $O(T^{-1})$，“bias(2)”的精确度为 $O(n^{-1}T^{-1})$，而“bias(3)”的精确度为 $O(n^{-1}T^{-2})$（这类似于在进行泰勒展开时，使用更高阶的近似）。选择项“vcov(#)”指定用来估计方差 - 协方差矩阵（variance-covariance matrix）的自助法重复抽样的次数，一般可设为“vcov(50)”。

下面以数据集 mus08cigar.dta 为例（纯粹为了演示的目的）。

```
. use mus08cigar.dta,clear
. gen t=year-62
. ssc install xtlsdvc  （下载安装命令 xtlsdvc）
. xtlsdvc lnc lnp lnpmin lny t,initial(ab) vcov(50) bias(3)①
```

① 运行此命令可能费时较长。

LSDVC dynamic regression
(bootstrapped SE)

lnc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnc						
L1.	1.058233	.0580338	18.23	0.000	.9444888	1.171977
lnp	-.0299203	.050643	-0.59	0.555	-.1291788	.0693381
lnpmin	.0056437	.0522719	0.11	0.914	-.0968075	.1080948
lny	.0611414	.0988631	0.62	0.536	-.1326266	.2549095
t	-.0051682	.0076777	-0.67	0.501	-.0202161	.0098798

16.14 重复截面数据与组群分析

面板数据的最大优点是能够跟踪个体随时间的变化。但有时，虽然可以得到不同时点上的数据，但每个时点上样本所包含的个体却完全不同。比如，英国家庭支出调查（U. K. Family Expenditure Survey）每年均调查不同的家庭[①]。这类数据，虽然也有横截面与时间两个维度，但却不是真正的面板数据，被称为“重复截面”（repeated cross-sections）数据[②]。

对于随机效应模型，即个体效应不与解释变量相关的情形，可以直接用 OLS 混合回归估计重复截面数据。由于每位个体在样本中只出现一次，故可以认为样本为独立同分布的，即在截面与时间维度上都是独立同分布的。因此，并不需要如一般面板数据那样使用聚类稳健标准误，只需使用普通标准误即可（但仍可能需要使用异方差稳健的标准误）。

对于更常见的固定效应模型，即个体效应与解释变量相关的情形，混合回归将导致不一致的估计。另一方面，由于每位个体在样本中只出现一次，组内估计法（FE）与一阶差分法（FD）均不可行。为此，Browning，Deaton and Irish（1985）与 Deaton（1985）提出，将样本中的个体分成几个固定的“组群”（cohort），构建“准面板”（pseudo panel）或“合成面板”（synthetic panel），以控制固定效应，进行“组群分析”（cohort analysis）。

一个组群由拥有共同特征（比如性别、出生年份、教育水平）的一组个体所组成，其成员是固定不变的。比如，可以按照出生年份来区分组群，“80 后组群”指的是在 1980 年至 1989 年出生的人，而“90 后组群”指的是在 1990 年至 1999 年出生的人，等等。

考虑如下固定效应模型：

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it} \quad (t = 1, \cdots, T) \tag{16.27}$$

其中，u_i 表示“个体固定效应”（individual fixed effect），即 u_i 与 $\boldsymbol{x}_{it}$ 相关。由于每位个体仅出现一次，故无法使用组内估计法或一阶差分法消去 u_i。假设样本中所有个体可以按照某种特征分为 C 个组群，并记组群的下标为 $c = 1, \cdots, C$。设 g 为决定个体的组群归属的变量，即第 i 个个体属于组群 c，如果 $g_i \in I_c$，其中 I_c 为一个集合。

给定组群的归属，对方程（16.27）两边求条件期望可得

$$\mathrm{E}(y_{it} \mid g_i \in I_c) = \mathrm{E}(\boldsymbol{x}_{it}' \mid g_i \in I_c)\boldsymbol{\beta} + \mathrm{E}(u_i \mid g_i \in I_c) + \mathrm{E}(\varepsilon_{it} \mid g_i \in I_c) \tag{16.28}$$

以“ * ”来表示以上的条件期望，即“组群的总体平均值”（population cohort average）：

① 如果每年都调查同样的家庭，则会对这些家庭造成较大负担，不利于对家庭真实信息的收集。

② 也称为“Time Series of Independent Cross-Sectional Data”（TSICS）。

$$y_{ct}^{*}=\boldsymbol{x}_{ct}^{*}{}'\boldsymbol{\beta}+u_{c}^{*}+\varepsilon_{ct}^{*}\quad(c=1,\cdots,C;t=1,\cdots,T)\tag{16.29}$$

其中，$u_{c}^{*}\equiv\mathrm{E}(u_{i}\mid g_{i}\in I_{c})$被称为“组群固定效应”(cohort fixed effects)，它衡量的是不同组群之间的异质性差异①。假设组群固定效应 u_{c}^{*} 不随时间而改变，且 ε_{ct}^{*} 不与$\boldsymbol{x}_{ct}^{*}$相关。然而，由于 u_{i} 与 $\boldsymbol{x}_{it}$相关，故 u_{c}^{*} 也与$\boldsymbol{x}_{ct}^{*}$相关，因此仍然可以视为固定效应模型。

由于组群的总体平均值不可观测，故使用“组群时点平均值”(cohort-time average)$\bar{y}_{ct}$与$\bar{\boldsymbol{x}}_{ct}$来替代，即给定时间，对组群中的个体进行平均。因此，

$$\bar{y}_{ct}=\bar{\boldsymbol{x}}_{ct}'\boldsymbol{\beta}+\bar{u}_{c}+\bar{\varepsilon}_{ct}\quad(c=1,\cdots,C;t=1,\cdots,T)\tag{16.30}$$

其中，$\bar{y}_{ct}\equiv\dfrac{1}{n_{ct}}\sum_{g_{it}\in I_{c}}y_{it}$，$n_{c}$ 为在第 t 期第 c 个组群所包含的个体数；$\bar{\boldsymbol{x}}_{ct}'$，$\bar{u}_{c}$，$\bar{\varepsilon}_{ct}$的定义类似。显然，在这个替代过程中引入了新的误差，比如 $\bar{y}_{ct}=y_{ct}^{*}+\xi_{ct}$，$\bar{\boldsymbol{x}}_{ct}=\boldsymbol{x}_{ct}^{*}+\boldsymbol{v}_{ct}$。如果每个组群在每个时点上的观测值足够多，则可以忽略“测量误差”(measurement error)。在这种情况下，对方程(16.30)使用组内估计法，即可消去 $\bar{u}_{c}$，得到参数 $\boldsymbol{\beta}$ 的一致估计。具体来说，就是把$(\bar{y}_{ct}-\bar{y}_{c})$对$(\bar{\boldsymbol{x}}_{ct}-\bar{\boldsymbol{x}}_{c})$进行 OLS 回归，其中 $\bar{y}_{c}\equiv\dfrac{1}{T}\sum_{t=1}^{T}\bar{y}_{ct}$，$\bar{\boldsymbol{x}}_{c}\equiv\dfrac{1}{T}\sum_{t=1}^{T}\bar{\boldsymbol{x}}_{ct}$为给定组群对时间进行平均。反之，如果每个组群在每个时点上的观测值较少，则需要对测量误差进行调整，参见 Cameron and Trivedi(2005, p. 772)。

组群分析的主要缺点：(1)由于组群的界定有一定的任意性，故对组群内个体的同质性假定(intracohort homogeneity)可能过强；(2)将个体数据加总平均为组群数据，将导致信息的散失，也可能出现异方差问题(因为组群的规模往往不同)；(3)如果组群的个数 C 不大，建立于 $C\to\infty$ 基础上的大样本理论可能失效。

例 周绍杰、张俊森、李宏彬(2009)使用组群分析研究中国城市家庭的消费与储蓄行为。

习 题

16.1 对于随机系数模型，证明复合扰动项$(\boldsymbol{x}_{it}'\boldsymbol{v}_{i}+\varepsilon_{it})$与解释变量$\boldsymbol{x}_{it}$不相关。

16.2 面板数据集 grunfeld.dta 包含了 10 个公司 1935—1954 年的以下变量：invest(投资额)，mvalue(公司市场市值)，与 kstock(公司资本存量)。考虑估计以下投资函数：

$$\text{invest}_{it}=\beta_{0}+\beta_{1}\,\text{mvalue}_{it}+\beta_{2}\,\text{kstock}_{it}+u_{i}+\varepsilon_{it}\tag{16.31}$$

参照本章实例，将其视为“长面板数据”进行各种估计与检验。

16.3 尝试对数据集 traffic.dta 进行动态面板估计。动态面板模型适用于该数据集吗？

① 正如个体固定效应(individual fixed effects)衡量的是不同个体之间的差异。

第 17 章　非线性面板

第 15、16 章介绍了线性面板。对于面板数据，如果被解释变量为虚拟变量、计数变量、受限变量等，则为非线性面板。一般来说，非线性面板不再适用线性面板的方法。

17.1　面板二值选择模型

对于面板数据，如果被解释变量为虚拟变量，则称为“面板二值选择模型”（binary choice model for panel data）。对于二值选择行为，通常可通过一个“潜变量”（latent variable）来概括该行为的净收益（收益减去成本）。如果净收益大于 0，则选择做；否则，选择不做。假设净收益为

$$y_{it}^{*} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it} \quad (i = 1, \cdots, n; t = 1, \cdots, T) \tag{17.1}$$

其中，净收益 y_{it}^{*} 为不可观测的潜变量，u_i 为个体效应（individual effects），而解释变量 $\boldsymbol{x}_{it}$ 不含常数项。个体的选择规则为

$$y_{it} = \begin{cases} 1 & \text{若 } y_{it}^{*} > 0 \\ 0 & \text{若 } y_{it}^{*} \leqslant 0 \end{cases} \tag{17.2}$$

给定 $\boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i$，则有

$$\begin{aligned} \mathrm{P}(y_{it} = 1 \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) &= \mathrm{P}(y_{it}^{*} > 0 \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) \\ &= \mathrm{P}(\boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it} > 0 \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) \\ &= \mathrm{P}(\varepsilon_{it} > -u_i - \boldsymbol{x}_{it}'\boldsymbol{\beta} \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) \\ &= \mathrm{P}(\varepsilon_{it} < u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) \\ &= F(u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}) \end{aligned} \tag{17.3}$$

其中，$F(\cdot)$ 为 ε_{it} 的累积分布函数（cdf），并假设 ε_{it} 的密度函数关于原点对称。如果 $\varepsilon_{it} \sim N(0,1)$（$\varepsilon_{it}$ 的方差被标准化为 1），则为 Probit 模型

$$\mathrm{P}(y_{it} = 1 \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) = \Phi(u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}) \tag{17.4}$$

如果 ε_{it} 服从逻辑分布，则为 Logit 模型

$$\mathrm{P}(y_{it} = 1 \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}, u_i) = \Lambda(u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}) = \frac{\mathrm{e}^{u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}}}{1 + \mathrm{e}^{u_i + \boldsymbol{x}_{it}'\boldsymbol{\beta}}} \tag{17.5}$$

面板二值选择模型的主要估计方法包括混合回归、随机效应估计与固定效应估计。在方程（17.1）中，如果 $u_1 = u_2 = \cdots = u_n$，则为混合回归（pooled probit 或 pooled logit），可将此面板数据作为截面数据处理（因为不存在个体效应）。此时，只需使用截面数据的相关 Stata 命令即可进行混合回归。由于同一个体不同时期的扰动项可能存在自相关，故应使用以面板为聚类的聚类稳健标准误（cluster-robust standard error）。

混合回归的 Stata 命令为

```
probit y x1 x2 x3,vce(cluster id)     (混合 probit 回归)
logit y x1 x2 x3,vce(cluster id)      (混合 logit 回归)
```

其中,id 为用来确定面板单位的变量。

17.2 面板二值选择模型的随机效应估计

更一般地,我们允许个体效应的存在,即不同的个体拥有不同的 u_i。如果 u_i 与所有解释变量 $\boldsymbol{x}_{it}$ 均不相关,则称为"随机效应模型"(Random Effects Model,简记 RE)。如果 u_i 与某个解释变量相关,则称为"固定效应模型"(Fixed Effects Model,简记 FE)。

首先考虑随机效应模型。对于线性面板的随机效应模型,一般使用广义最小二乘法(GLS)进行估计。但非线性面板不便使用 GLS,故转而进行最大似然估计(MLE)。假设 $u_i \sim N(0,\sigma_u^2)$,记其密度函数为 $g(u_i)$。下面以 Logit 模型为例。给定 u_i,则个体 i 的条件分布为(推导过程可参考第 11 章的相关内容)

$$f(y_{i1},y_{i2},\cdots,y_{iT}\mid \boldsymbol{x}_{it},\boldsymbol{\beta},u_i)=\prod_{t=1}^{T}[\Lambda(u_i+\boldsymbol{x}_{it}'\boldsymbol{\beta})]^{y_i}[1-\Lambda(u_i+\boldsymbol{x}_{it}'\boldsymbol{\beta})]^{1-y_i} \tag{17.6}$$

但上式中的 u_i 不可观测。为此,记 $(y_{i1},y_{i2},\cdots,y_{iT},u_i)$ 的联合密度为 $f(y_{i1},y_{i2},\cdots,y_{iT},u_i)$,并进行如下分解

$$f(y_{i1},y_{i2},\cdots,y_{iT},u_i)=f(y_{i1},y_{i2},\cdots,y_{iT}\mid u_i)\cdot g(u_i) \tag{17.7}$$

在 $(y_{i1},y_{i2},\cdots,y_{iT},u_i)$ 的联合密度中,将 u_i 积分积掉,即可得到 $(y_{i1},y_{i2},\cdots,y_{iT})$ 的边缘密度(这是根据联合密度求边缘密度的一般方法),

$$\begin{aligned}f(y_{i1},y_{i2},\cdots,y_{iT})&=\int_{-\infty}^{+\infty}f(y_{i1},y_{i2},\cdots,y_{iT},u_i)\,\mathrm{d}u_i\\&=\int_{-\infty}^{+\infty}f(y_{i1},y_{i2},\cdots,y_{iT}\mid u_i)\cdot g(u_i)\,\mathrm{d}u_i\\&=\int_{-\infty}^{+\infty}\left\{\prod_{t=1}^{T}[\Lambda(u_i+\boldsymbol{x}_{it}'\boldsymbol{\beta})]^{y_i}[1-\Lambda(u_i+\boldsymbol{x}_{it}'\boldsymbol{\beta})]^{1-y_i}\right\}\cdot g(u_i)\,\mathrm{d}u_i\end{aligned} \tag{17.8}$$

上式的积分没有解析解,Butler and Moffitt (1982)提出使用"Gauss-Hermite quadrature"方法进行数值积分(numerical integration)。数值积分的基本思想就是以一系列矩形或梯形的面积之和来近似曲线下的面积。数学家已对这种方法进行了不少改进以提高计算速度与精度。Stata 的默认方法为在 12 个点上进行"adaptive Gauss-Hermite quadrature"计算。由于此积分的精确度依赖于数值计算的点数,故 Stata 提供命令 quadchk 来检验此数值积分的精确度,即使用有别于 12 的计算点数,以比较结果的稳定性。如果结果不稳健,则应增加计算点数,直至得到稳定结果。

假设不同个体的观测值相互独立,则可以写出整个样本的似然函数。最大化此似然函数即得到对 $\boldsymbol{\beta}$ 的"随机效应 Logit 估计量"(Random Effect Logit)。如果将方程(17.8)中的逻辑分布 $\Lambda(\cdot)$ 改为正态分布 $\Phi(\cdot)$,则为"随机效应 Probit 估计量"(Random Effect Probit)。由于在进行随机效应估计时已将 u_i 积分积掉,故得不到对个体效应 u_i 的估计,也因此无法预测 y_i 的发生概率或计算解释变量对 y_i 的边际效应。解决方法之一是假设 $u_i=0$。

由于不同个体的观测值相互独立，故不同个体之间的扰动项也互不相关。但由于 u_i 的存在，同一个体不同时期的扰动项之间仍存在自相关

$$\mathrm{Cov}(u_i+\varepsilon_{it},u_i+\varepsilon_{is})=\begin{cases}\sigma_u^2 & \text{若 } t\neq s\\ \sigma_u^2+\sigma_\varepsilon^2 & \text{若 } t=s\end{cases} \tag{17.9}$$

其中，σ_u^2 为 u_i 的方差，而 σ_ε^2 为 ε_{it} 的方差。当 $t\neq s$ 时，其自相关系数为

$$\rho\equiv\mathrm{Corr}(u_i+\varepsilon_{it},u_i+\varepsilon_{is})=\frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2} \tag{17.10}$$

ρ 越大，则复合扰动项 $(u_i+\varepsilon_{it})$ 中个体效应的部分 (u_i) 越重要。特别地，如果 $\rho=0$，则说明 $\sigma_u^2=0$，即不存在个体随机效应，应选择混合回归。

对于面板二值选择模型，其随机效应估计的 Stata 命令为

```
xtprobit y x1 x2 x3      (默认为随机效应 probit)
xtlogit y x1 x2 x3       (默认为随机效应 logit)
```

在 Stata 的输出结果中，包含了对原假设“H_0: $\rho=0$”的 LR 检验结果（Stata 将 ρ 记为“rho”）。如果拒绝“H_0: $\rho=0$”，则认为应采用随机效应模型；反之，则支持混合回归。

17.3 面板二值选择模型的固定效应估计

在面板二值选择模型中，如果个体效应 u_i 与解释变量 $\boldsymbol{x}_{it}$ 相关，则为固定效应模型，

$$\mathrm{P}(y_{it}=1\,|\,\boldsymbol{x}_{it},\boldsymbol{\beta},u_i)=F(u_i+\boldsymbol{x}_{it}'\boldsymbol{\beta}) \tag{17.11}$$

其中，$F(\cdot)$ 为 $\Phi(\cdot)$ 或 $\Lambda(\cdot)$。此时，如果使用随机效应模型或混合回归将得不到一致估计。对于线性面板数据，一般通过做组内变换 $(y_{it}-\bar{y}_i)$ 或一阶差分 $(y_{it}-y_{i,t-1})$ 来消去固定效应 u_i。但对于非线性面板数据，这些变换一般不起作用，因为无法建立可观测的 $(y_{it}-\bar{y}_i)$ 与不可观测的 $(y_{it}^*-\bar{y}_i^*)$（即潜变量 y_{it}^* 的组内离差）之间的对应关系。

对于二值选择的固定效应模型，即使在模型中增加个体虚拟变量（相当于 LSDV 法），也仍然得不到一致估计（除非 $T\to\infty$）。这是因为，当 $n\to\infty$ 时，待估计的个体效应 $\{u_i\}_{i=1}^n$ 的个数也随之增加，这些 $\{u_i\}_{i=1}^n$ 被称为“伴生参数”（incidental parameters）①。另一方面，每个 u_i 只能由个体 i 的 T 个观测值来估计，而 T 并不增加。对于现实中的数据，T 为有限值，而且常常很小（短面板）。因此，当 $n\to\infty$ 而 T 为有限值，对 $\{u_i\}_{i=1}^n$ 的估计将是不一致的（不是因为 $\hat{u}_i$ 收敛到不等于 u_i 的常数，而是 $\hat{u}_i$ 根本就不收敛）。更为严重的问题是，在非线性面板中，对于 u_i 的不一致估计还会“污染”$\boldsymbol{\beta}$ 的估计，导致对 $\boldsymbol{\beta}$ 的估计也不一致，这被称为“伴生参数问题”（incidental parameters problem）。

在线性面板中，可以通过做组内变换或差分变换来解决伴生参数问题。对于固定效应的面板 Probit 模型，目前尚无法解决此伴生参数问题。对于固定效应的面板 Logit 模型，可以通过寻找 u_i 的“充分统计量”（sufficient statistic），然后在给定此充分统计量的条件下进行“条件最大似然估计”（conditional MLE）。

---

① 虽然 u_i 为随机变量，但抽样后即为确定值，对于 y_{it} 而言可以视为参数。而且，我们主要对估计 $\boldsymbol{\beta}$ 感兴趣，故称 u_i 为“伴生参数”。有关“伴生参数问题”，参见 Neyman and Scott (1948) 以及 Lancaster (2000)。

充分统计量是统计学中的一个概念。考虑总体参数 θ 与统计量 W。如果统计量 W 包含了样本中所有可用来估计 θ 的信息，则称 W 为参数 θ 的充分统计量。换言之，给定 W 之后，任何根据样本计算的其他统计量都不可能提供关于 θ 的额外信息。因此，样本数据在给定充分统计量 W 后的条件分布将不再依赖于 θ（这正是充分统计量的严格定义）；否则，控制了 W 之后的条件分布仍将包含 θ 的信息。

对于 Logit 模型，Chamberlain（1980）提出使用 $n_i \equiv \sum_{t=1}^{T} y_{it}$ 作为 u_i 的充分统计量，并计算在给定 n_i 情况下的条件似然函数（根据充分统计量的性质，此条件似然函数不再依赖于 u_i），然后进行条件最大似然估计。然而，对于 Probit 模型，却找不到 u_i 的充分统计量。Logit 模型之所以存在 u_i 的充分统计量，是由逻辑函数的特殊指数形式所决定的。

以最简单的两期模型为例进行说明，即 $T = 2$。此时，对于个体 i，只有以下三种可能情形，即 $n_i = y_{i1} + y_{i2} = 0, 1$ 或 2。下面分别考虑这三种情形。

（1）$n_i = y_{i1} + y_{i2} = 0$。此时，必然有 $y_{i1} = y_{i2} = 0$，即 $P(y_{i1} = 0, y_{i2} = 0 \mid n_i = 0) = 1$，其对数似然函数为 $\ln 1 = 0$，故对整个样本的对数似然函数没有贡献。直观来看，由于此条件似然函数的取值为常数（无论参数 $\boldsymbol{\beta}$ 如何取值），故此观测值不包含任何可用于估计 $\boldsymbol{\beta}$ 的信息。因此，在进行条件似然估计法时，是否包括这些观测值并不影响估计结果；事实上，等于损失了这些样本的观测值。

（2）$n_i = y_{i1} + y_{i2} = 2$。此时，必然有 $y_{i1} = y_{i2} = 1$，即 $P(y_{i1} = 1, y_{i2} = 1 \mid n_i = 2) = 1$。同理，这些观测值也不包含任何有助于估计 $\boldsymbol{\beta}$ 的信息，也可以忽略。

（3）$n_i = y_{i1} + y_{i2} = 1$。此时，或者 $(y_{i1} = 0, y_{i2} = 1)$，或者 $(y_{i1} = 1, y_{i2} = 0)$。分别计算其条件概率如下。

$$P(y_{i1} = 0, y_{i2} = 1 \mid n_i = 1) = \frac{P(y_{i1} = 0, y_{i2} = 1)}{P(y_{i1} = 0, y_{i2} = 1) + P(y_{i1} = 1, y_{i2} = 0)} \tag{17.12}$$

$$P(y_{i1} = 1, y_{i2} = 0 \mid n_i = 1) = \frac{P(y_{i1} = 1, y_{i2} = 0)}{P(y_{i1} = 0, y_{i2} = 1) + P(y_{i1} = 1, y_{i2} = 0)} \tag{17.13}$$

假设在给定 u_i 与 $\boldsymbol{x}_{it}$ 的条件下，y_{i1} 与 y_{i2} 相互独立，则

$$P(y_{i1} = 0, y_{i2} = 1) = \frac{1}{1 + e^{u_i + \boldsymbol{x}_{i1}'\boldsymbol{\beta}}} \cdot \frac{e^{u_i + \boldsymbol{x}_{i2}'\boldsymbol{\beta}}}{1 + e^{u_i + \boldsymbol{x}_{i2}'\boldsymbol{\beta}}} \tag{17.14}$$

$$P(y_{i1} = 1, y_{i2} = 0) = \frac{e^{u_i + \boldsymbol{x}_{i1}'\boldsymbol{\beta}}}{1 + e^{u_i + \boldsymbol{x}_{i1}'\boldsymbol{\beta}}} \cdot \frac{1}{1 + e^{u_i + \boldsymbol{x}_{i2}'\boldsymbol{\beta}}} \tag{17.15}$$

将表达式（17.14）与（17.15）代入方程（17.12）可得

$$\begin{aligned} P(y_{i1} = 0, y_{i2} = 1 \mid n_i = 1) &= \frac{e^{u_i + \boldsymbol{x}_{i2}'\boldsymbol{\beta}}}{e^{u_i + \boldsymbol{x}_{i1}'\boldsymbol{\beta}} + e^{u_i + \boldsymbol{x}_{i2}'\boldsymbol{\beta}}} \\ &= \frac{e^{\boldsymbol{x}_{i2}'\boldsymbol{\beta}}}{e^{\boldsymbol{x}_{i1}'\boldsymbol{\beta}} + e^{\boldsymbol{x}_{i2}'\boldsymbol{\beta}}} = \frac{e^{(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}}}{1 + e^{(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}}} = \Lambda[(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}] \end{aligned} \tag{17.16}$$

注意到在上式中，分子与分母中的 e^{u_i} 被同时消去了，这正是逻辑分布的方便之处。类似地，将表达式（17.14）与（17.15）代入方程（17.13）可得

$$P(y_{i1} = 1, y_{i2} = 0 \mid n_i = 1) = \Lambda[-(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}] = 1 - \Lambda[(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}] \tag{17.17}$$

定义虚拟变量 $d_i = 1$，如果 $(y_{i1} = 0, y_{i2} = 1)$；$d_i = 0$，如果 $(y_{i1} = 1, y_{i2} = 0)$。综合方程（17.16）与（17.17），可将个体 i 的条件对数似然函数写为

$$\ln L_i(\boldsymbol{\beta}) = \{d_i \ln \Lambda[(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}] + (1 - d_i) \ln\{1 - \Lambda[(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})'\boldsymbol{\beta}]\}\} \cdot \mathbf{1}(n_i = 1) \tag{17.18}$$

其中，$\mathbf{1}(\cdot)$为示性函数，$\mathbf{1}(n_i=1)$表示仅考虑 $n_i=1$ 的观测值。将上式对 i 加总，即可得到整个样本的条件对数似然函数。从方程(17.18)可以看出：

(1) 给定 n_i 的条件似然函数不再依赖于 u_i(已在前面的推导中消去)；

(2) 此条件似然函数仍为 Logit 模型，只是解释变量变为$(\boldsymbol{x}_{i2}-\boldsymbol{x}_{i1})$；

(3) 不随时间变化的变量将无法识别其系数，因为$(x_{i2}-x_{i1})=0$。另外，与随机效应模型相比，固定效应的似然函数(17.18)并不包含积分，故没有必要进行数值积分。

更一般地，对于 $T>2$，可以计算给定 $n_i=1, n_i=2, \cdots$，或 $n_i=T-1$ 的条件似然函数。固定效应模型的缺点是，将损失所有 $n_i=0$ 或 $n_i=T$ 的观测值，导致样本容量减少。另外，由于条件似然函数估计法消去了 u_i，故无法估计个体效应 u_i，也因此无法预测 y_i 的发生概率或解释变量对 y_i 的边际效应(它们的表达式都包含 u_i)。解决方法之一是假设 $u_i=0$。

如何在固定效应模型与混合回归之间进行选择？可以使用以下的豪斯曼检验。原假设为不存在个体效应，即“H_0: $u_i=u$”。如果原假设成立，则固定效应模型与混合回归都是一致的，但混合回归更有效率(固定效应模型未利用“H_0: $u_i=u$”的信息，而且可能损失样本容量)。反之，如果原假设不成立，则固定效应模型是一致的，而混合回归不一致。如果二者的系数估计值相差较大(以二次型来衡量)，则倾向于拒绝原假设，接受存在个体效应的替代假设。更进一步，对于固定效应模型与随机效应模型之间的选择，也可以进行豪斯曼检验。

固定效应面板 logit 的 Stata 命令为，

```
xtlogit y x1 x2 x3,fe
```

17.4 面板二值选择模型的 Stata 实例

Chen(2014)使用公元 25—1911 年期间的王朝面板(dynastic panel)来考察中国北方农民起义的决定因素，横截面单位为王朝(dyn)，而时间维度为年(year)。数据集“uprising.dta”包括被解释变量 uprising (=1，发生农民起义；=0，其他)，以及以下解释变量：age(王朝年龄)，pop(人口)，temp(气温)，sfamine(=1，发生严重饥荒；=0，其他)，sfamine1(sfamine 的一阶滞后)，sfamine2(sfamine 的二阶滞后)，sfamine3(sfamine 的三阶滞后)，relief(=1，如果有政府救济；=0，其他)，relief1(relief 的一阶滞后)，relief2(relief 的二阶滞后)，relief3(relief 的三阶滞后)。由于面板 probit 无固定效应模型，故使用面板 logit 模型。

```
. use uprising.dta,clear
```

首先，进行固定效应面板 logit 估计。

```
. xtlogit uprising age pop temp sfamine sfamine1 sfamine2 sfamine3
relief relief1 relief2 relief3,fe nolog
. estimates store FE
```

```
note: multiple positive outcomes within groups encountered.
note: 1 group (43 obs) dropped because of all positive or
      all negative outcomes.
Conditional fixed-effects logistic regression   Number of obs      =      1561
Group variable: dyn                             Number of groups   =        10

                                                Obs per group: min =        35
                                                               avg =     156.1
                                                               max =       287

                                                LR chi2(11)        =    191.55
Log likelihood  = -416.20732                    Prob > chi2        =    0.0000
```

uprising	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0119585	.0019157	6.24	0.000	.0082038	.0157131
pop	.0311968	.0145438	2.15	0.032	.0026914	.0597021
temp	.5265406	.1989067	2.65	0.008	.1366906	.9163906
sfamine	1.130333	.3002773	3.76	0.000	.5418006	1.718866
sfamine1	.5676738	.3271661	1.74	0.083	-.0735599	1.208907
sfamine2	.5911076	.3238025	1.83	0.068	-.0435335	1.225749
sfamine3	.577348	.3177628	1.82	0.069	-.0454556	1.200152
relief	-.5210128	.2390826	-2.18	0.029	-.9896061	-.0524195
relief1	-.6768912	.2469321	-2.74	0.006	-1.160869	-.1929132
relief2	-.8935897	.2521104	-3.54	0.000	-1.387717	-.3994624
relief3	-.7126752	.2436877	-2.92	0.003	-1.190294	-.235056

从上表第二行可知，有一个朝代（共 43 年的观测值）从未发生农民起义，故在估计固定效应模型时被去掉。检查数据可知，该朝代为第二个朝代（朝代变量 dyn = 2），即三国时期的曹魏。变量 age，pop，temp 的系数均至少在 5% 水平上显著为正，这意味着王朝后期更容易发生农民起义；人口压力越大，则农民起义的概率越高；而气温越高，也越易发生农民起义[①]。变量 sfamine（严重饥荒）的当期系数在 1% 水平上显著为正，其前三阶滞后的系数也在 10% 水平上显著为正，这表明严重饥荒显著地增加农民起义的发生概率。另一方面，变量 relief（政府救灾）的各期系数均至少在 5% 水平上显著为负，意味着政府救灾有助于缓解农民起义。

其次，估计混合回归 logit 模型。为保持与固定效应的样本一致，在混合回归中也去掉了“dyn = 2”时的朝代。

```
. logit uprising age pop temp sfamine sfamine1 sfamine2 sfamine3
relief relief1 relief2 relief3 if dyn ~ =2,vce(cluster dyn) nolog

Logistic regression                             Number of obs      =      1561
                                                Wald chi2(8)       =         .
                                                Prob > chi2        =         .
Log pseudolikelihood = -505.18692               Pseudo R2          =    0.1012

                                (Std. Err. adjusted for 10 clusters in dyn)
```

uprising	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.004675	.0021493	2.18	0.030	.0004623	.0088876
pop	.0099646	.0074699	1.33	0.182	-.0046761	.0246053
temp	.4547908	.3623444	1.26	0.209	-.2553911	1.164973
sfamine	1.234068	.4340109	2.84	0.004	.3834224	2.084714
sfamine1	.7412263	.363804	2.04	0.042	.0281836	1.454269
sfamine2	.787409	.2799314	2.81	0.005	.2387534	1.336065
sfamine3	.8675515	.2147783	4.04	0.000	.4465938	1.288509
relief	-.2743782	.1292077	-2.12	0.034	-.5276207	-.0211357
relief1	-.4558084	.2266392	-2.01	0.044	-.900013	-.0116037
relief2	-.5703093	.2113629	-2.70	0.007	-.984573	-.1560457
relief3	-.372896	.221107	-1.69	0.092	-.8062577	.0604657
_cons	-2.466138	.5024458	-4.91	0.000	-3.450914	-1.481362

① 进一步的研究发现，人口与气温的作用并不稳健。

为了在固定效应模型与混合回归之间进行选择，下面进行豪斯曼检验。但豪斯曼检验不允许在混合回归使用选择项“vce(cluster dyn)”，故重新进行混合回归。

```
. quietly logit uprising age pop temp sfamine  sfamine1 sfamine2 sfamine3 relief relief1 relief2 relief3 if dyn ~ =2,nolog
. estimates store POOLED
. hausman FE POOLED
```

	Coefficients			
	(b) FE	(B) POOLED	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
age	.0119585	.004675	.0072835	.0015887
pop	.0311968	.0099646	.0212322	.0140593
temp	.5265406	.4547908	.0717497	.1283301
sfamine	1.130333	1.234068	-.1037347	.1321584
sfamine1	.5676738	.7412263	-.1735526	.149367
sfamine2	.5911076	.787409	-.1963013	.1502582
sfamine3	.577348	.8675515	-.2902035	.1540328
relief	-.5210128	-.2743782	-.2466346	.1026869
relief1	-.6768912	-.4558084	-.2210828	.1042203
relief2	-.8935897	-.5703093	-.3232804	.1141729
relief3	-.7126752	-.372896	-.3397792	.1134979

```
            b = consistent under Ho and Ha; obtained from xtlogit
 B = inconsistent under Ha, efficient under Ho; obtained from logit

Test:  Ho:  difference in coefficients not systematic

              chi2(11) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                       =       71.76
             Prob>chi2 =      0.0000
```

上表结果显示，强烈拒绝混合回归的原假设，认为应使用固定效应模型。下面进行随机效应面板 logit 估计。

```
. xtlogit uprising age pop temp sfamine  sfamine1 sfamine2 sfamine3 relief relief1 relief2 relief3 if dyn ~ =2,re nolog
. estimates store RE
```

```
Random-effects logistic regression              Number of obs      =      1561
Group variable: dyn                             Number of groups   =        10

Random effects u_i ~ Gaussian                   Obs per group: min =        35
                                                               avg =     156.1
                                                               max =       287

                                                Wald chi2(11)      =    127.59
Log likelihood  = -458.18161                    Prob > chi2        =    0.0000
```

uprising	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0112875	.0018141	6.22	0.000	.007732	.014843
pop	.0242926	.0107208	2.27	0.023	.0032802	.0453051
temp	.5079557	.1956755	2.60	0.009	.1244387	.8914726
sfamine	1.147349	.2985703	3.84	0.000	.5621616	1.732536
sfamine1	.5874378	.3250784	1.81	0.071	-.0497041	1.22458
sfamine2	.6138714	.3219423	1.91	0.057	-.0171238	1.244867
sfamine3	.598085	.3153273	1.90	0.058	-.0199452	1.216115
relief	-.5251701	.2378131	-2.21	0.027	-.9912753	-.059065
relief1	-.684916	.2455628	-2.79	0.005	-1.16621	-.2036217
relief2	-.8914661	.2502494	-3.56	0.000	-1.381946	-.4009862
relief3	-.7093608	.242091	-2.93	0.003	-1.18385	-.2348712
_cons	-2.872401	.5282318	-5.44	0.000	-3.907716	-1.837085
/lnsig2u	.5281101	.5038209			-.4593608	1.515581
sigma_u	1.3022	.3280378			.7947876	2.133557
rho	.340125	.1130776			.1610808	.5804774

```
Likelihood-ratio test of rho=0: chibar2(01) =   94.01 Prob >= chibar2 = 0.000
```

从上表最后一行可知，LR 检验的强烈拒绝原假设“H_0: $\rho=0$”，认为应使用面板随机效应模型，不宜进行混合回归。由于随机效应面板 Logit 的计算使用了数值积分，下面使用命令 quadchk 检验此数值积分的稳健性。

```
. quadchk,noout
```

其中，选择项“noout”表示不显示重估模型的结果。命令“xtlogit,re”默认在 12 个点上进行数值积分，而命令“quadchk”则在 8 个点与 16 个点上进行数值积分，重新估计此模型（refitting model），然后与默认的 12 点上数值积分结果相比较。一般来说，只要这几种数值积分的系数估计值的“相对差距”（relative difference）不超过10^{-4}，则认为结果较为稳健。否则，应使用选项“intpoints(#)”来增加数值积分的计算点数，比如“xtlogit,re intpoints(24)”表示在 24 个点上进行数值积分。

```
Refitting model intpoints() =  8
Refitting model intpoints() = 16

                        Quadrature check

             Fitted       Comparison     Comparison
           quadrature     quadrature     quadrature
           12 points      8 points       16 points
-------------------------------------------------------------------
Log        -458.18161     -458.18163     -458.18161
likelihood                -.00001689      9.789e-08   Difference
                           3.687e-08     -2.136e-10   Relative difference
-------------------------------------------------------------------
uprising:    .01128749      .01128749      .01128749
     age                  -4.194e-12     -1.647e-11   Difference
                          -3.716e-10     -1.459e-09   Relative difference
-------------------------------------------------------------------
uprising:    .02429264      .02429264      .02429264
     pop                   2.767e-11      4.978e-10   Difference
                           1.139e-09      2.049e-08   Relative difference
-------------------------------------------------------------------
uprising:    .50795566      .50795566      .50795565
    temp                   1.938e-11     -2.265e-09   Difference
                           3.816e-11     -4.458e-09   Relative difference
-------------------------------------------------------------------
uprising:   1.1473487       1.1473487      1.1473487
 sfamine                   1.586e-10      6.578e-09   Difference
                           1.382e-10      5.733e-09   Relative difference
-------------------------------------------------------------------
uprising:    .58743785      .58743785      .58743785
sfamine1                   1.730e-10      1.531e-09   Difference
                           2.944e-10      2.606e-09   Relative difference
-------------------------------------------------------------------
uprising:    .61387142      .61387142      .61387143
sfamine2                   2.203e-10      1.580e-09   Difference
                           3.588e-10      2.574e-09   Relative difference
-------------------------------------------------------------------
uprising:     .598085        .598085        .598085
sfamine3                   3.440e-10      6.962e-10   Difference
                           5.752e-10      1.164e-09   Relative difference
-------------------------------------------------------------------
uprising:   -.52517014     -.52517014     -.52517015
  relief                    3.558e-10     -5.434e-09   Difference
                           -6.775e-10      1.035e-08   Relative difference
-------------------------------------------------------------------
uprising:   -.68491604     -.68491604     -.68491604
 relief1                   3.761e-10     -4.941e-09   Difference
                           -5.492e-10      7.214e-09   Relative difference
-------------------------------------------------------------------
uprising:   -.89146607     -.89146607     -.89146607
 relief2                   3.875e-10     -5.759e-09   Difference
                           -4.347e-10      6.461e-09   Relative difference
-------------------------------------------------------------------
uprising:   -.70936077     -.70936077     -.70936078
 relief3                   4.232e-10     -5.424e-09   Difference
                           -5.966e-10      7.646e-09   Relative difference
-------------------------------------------------------------------
uprising:   -2.8724008     -2.8724008     -2.8724008
   _cons                  -2.114e-08     -5.972e-08   Difference
                           7.361e-09      2.079e-08   Relative difference
-------------------------------------------------------------------
lnsig2u:     .5281101       .5281101       .52811025
   _cons                  -1.799e-09      1.540e-07   Difference
                          -3.407e-09      2.917e-07   Relative difference
-------------------------------------------------------------------
```

从上表可知，所有系数的相对差距均小于10^{-4}，故数值积分的结果值得信赖。

为了在固定效应与随机效应模型之间进行选择，下面再次进行豪斯曼检验。

```
. hausman FE RE
```

	Coefficients			
	(b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
age	.0119585	.0112875	.000671	.0006156
pop	.0311968	.0242926	.0069041	.0098278
temp	.5265406	.5079557	.0185849	.0357068
sfamine	1.130333	1.147349	-.0170153	.0319724
sfamine1	.5676738	.5874378	-.0197641	.0369005
sfamine2	.5911076	.6138714	-.0227638	.0346586
sfamine3	.577348	.598085	-.020737	.0392664
relief	-.5210128	-.5251701	.0041573	.0246053
relief1	-.6768912	-.684916	.0080249	.0259682
relief2	-.8935897	-.8914661	-.0021236	.030576
relief3	-.7126752	-.7093608	-.0033144	.0278508

b = consistent under Ho and Ha; obtained from xtlogit
B = inconsistent under Ha, efficient under Ho; obtained from xtlogit

Test: Ho: difference in coefficients not systematic

chi2(11) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 21.57
Prob>chi2 = 0.0280
(V_b-V_B is not positive definite)

上表结果表明，可以在5%水平上拒绝随机效应的原假设，认为应使用固定效应模型。

最后，为了演示的目的，估计面板 probit 的随机效应模型。

```
. xtprobit uprising age pop temp sfamine sfamine1 sfamine2 sfamine3
relief relief1 relief2 relief3 if dyn ~ =2,re nolog
```

avg = 156.1
max = 287

Log likelihood = -460.72517
Wald chi2(11) = 135.18
Prob > chi2 = 0.0000

uprising	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0056746	.0009103	6.23	0.000	.0038904	.0074587
pop	.0134784	.0055652	2.42	0.015	.0025708	.024386
temp	.2441783	.1007821	2.42	0.015	.0466491	.4417075
sfamine	.6406771	.1649126	3.88	0.000	.3174543	.9638998
sfamine1	.3719557	.1765811	2.11	0.035	.0258631	.7180483
sfamine2	.349083	.1759537	1.98	0.047	.0042202	.6939458
sfamine3	.3321031	.1728483	1.92	0.055	-.0066732	.6708795
relief	-.2639172	.1241089	-2.13	0.033	-.5071662	-.0206682
relief1	-.3895312	.1271142	-3.06	0.002	-.6386704	-.140392
relief2	-.456395	.1278903	-3.57	0.000	-.7070553	-.2057347
relief3	-.3525801	.1251987	-2.82	0.005	-.5979651	-.1071952
_cons	-1.609203	.2701962	-5.96	0.000	-2.138778	-1.079628
/lnsig2u	-.8042879	.5053684			-1.794792	.1862159
sigma_u	.6688844	.1690165			.4076298	1.09758
rho	.309109	.1079268			.1424862	.5464199

Likelihood-ratio test of rho=0: chibar2(01) = 89.11 Prob >= chibar2 = 0.000

上表最后一行的 LR 检验结果强烈拒绝混合 probit 回归，认为应使用随机效应模型。随机效应面板 probit 的回归结果在性质上与随机效应面板 logit 相似。

17.5 面板泊松回归

考虑被解释变量为计数变量的面板数据。比如，若干企业在一段时间内每年获得专利的个数，全国各省在几年内每年发生矿难的数目，世界各国近几十年每年发生示威游行的次数，一些病人在一段时间内的每期发病次数，等等。

对于个体 i，时期 t，记被解释变量为 Y_{it}，假设 $Y_{it}=y_{it}$ 的概率由参数为 λ_{it} 的泊松分布所决定：

$$P(Y_{it}=y_{it}|\boldsymbol{x}_{it})=\frac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!}\quad(y_{it}=0,1,2,\cdots) \tag{17.19}$$

其中，$\lambda_{it}>0$ 为泊松到达率。为了保证 λ_{it} 为非负，假设

$$\lambda_{it}=\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta}+u_i)=\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})\cdot\exp(u_i)\equiv v_i\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta}) \tag{17.20}$$

其中，$\boldsymbol{x}_{it}$ 不含常数项，而 $v_i\equiv\exp(u_i)$ 为乘积形式的个体效应（multiplicative individual effects）。如果 $v_1=v_2=\cdots=v_n$，则不存在个体效应，为混合泊松回归（pooled Poisson），可将此面板数据作为截面数据处理。此时，可直接用截面数据的泊松回归 Stata 命令进行混合回归，但需使用聚类稳健标准误。

混合泊松回归的 Stata 命令为，

```
poisson y x1 x2 x3,vce(cluster id) irr
```

其中，id 是用来确定面板单位的变量。选择项“irr”表示汇报发生率比（incidence rate ratio）。

更一般地，我们允许个体效应的存在，即不同个体拥有不同的 v_i。如果 v_i 与所有解释变量 $\boldsymbol{x}_{it}$ 均不相关，则为“随机效应模型”。如果 v_i 与某个解释变量相关，则为“固定效应模型”。

首先考虑随机效应模型。由于这是非线性模型，故可使用最大似然估计（MLE）。记 v_i 的密度函数为 $g(v_i)$。假设样本为 iid，则在给定 v_i 的情况下，个体 i 的条件分布为，

$$f(y_{i1},y_{i2},\cdots,y_{iT}|v_i)=\prod_{t=1}^{T}\left(\frac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!}\right)=\prod_{t=1}^{T}\left(\frac{e^{-v_i\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})}[v_i\exp(\boldsymbol{x}_{it}'\boldsymbol{\beta})]^{y_{it}}}{y_{it}!}\right) \tag{17.21}$$

但上式中的 v_i 不可观测。为此，记 $(y_{i1},y_{i2},\cdots,y_{iT},v_i)$ 的联合密度为 $f(y_{i1},y_{i2},\cdots,y_{iT},v_i)$，并进行如下分解，

$$f(y_{i1},y_{i2},\cdots,y_{iT},v_i)=f(y_{i1},y_{i2},\cdots,y_{iT}|v_i)\cdot g(v_i) \tag{17.22}$$

将上式中的 v_i 积分积掉，即可得到 $(y_{i1},y_{i2},\cdots,y_{iT})$ 的边缘密度，

$$\begin{aligned}f(y_{i1},y_{i2},\cdots,y_{iT})&=\int_{-\infty}^{+\infty}f(y_{i1},y_{i2},\cdots,y_{iT},v_i)\,dv_i\\&=\int_{-\infty}^{+\infty}f(y_{i1},y_{i2},\cdots,y_{iT}|v_i)\cdot g(v_i)\,dv_i\end{aligned} \tag{17.23}$$

由于 $v_i=\exp(u_i)>0$，故通常选择 v_i 服从 Gamma 分布（指数分布与卡方分布都是 Gamma 分布的特例）。特别地，假设 $v_i\sim\text{Gamma}(1/\alpha,\alpha)$，其中 $\alpha>0$，则 $E(v_i)=1$，$\text{Var}(v_i)=\alpha$。将 Gamma $(1/\alpha,\alpha)$ 的概率密度代入方程（17.23），可以得到负二项分布的概率密度，然后进行 MLE 估计。由于参数 α 为个体异质性 v_i 的方差，故可对“H_0：$\alpha=0$”（Stata 称为“alpha = 0”）进行 LR 检验来判

断是否存在个体异质性，即究竟应使用随机效应泊松回归还是混合泊松回归。

另一种选择是，假设 $u_i \sim N(0,\sigma_u^2)$，但得不到解析解，一般用“Gauss-Hermite quadrature”（高斯－汉密尔特正交法）方法对方程（17.23）进行数值积分；也需要使用 Stata 命令 quadchk 来检验此数值积分的精确度。类似地，可通过检验“H_0: $\sigma_u = 0$”（Stata 称为“sigma_u = 0”）来判断是否存在个体异质性。

下面考虑固定效应的面板泊松模型。类似于固定效应的面板二值选择模型（参见本章第 3 节），可以使用 $n_i \equiv \sum_{t=1}^{T} y_{it}$ 作为 v_i 的充分统计量，并计算在给定 n_i 情况下的条件似然函数，然后进行条件最大似然估计（conditional MLE）。显然，如果 $n_i = 0$，则意味着 $y_{i1} = \cdots = y_{iT} = 0$，故个体 i 的观测值对于条件似然函数的贡献为 0。因此，所有 y_{it} 观测值均为 0 的个体，在估计固定效应泊松回归时将被去掉，故可能损失样本容量。固定效应模型的另一缺点是，无法估计不随时间而变的变量系数。固定效应泊松回归的最大优点在于，它允许个体异质性 v_i 与解释变量 $\boldsymbol{x}_{it}$ 相关。在固定效应与随机效应泊松回归之间进行选择时，可借助于豪斯曼检验。

面板泊松回归的 Stata 命令为

```
xtpoisson y x1 x2 x3,fe normal irr
```

其中，选择项“fe”表示固定效应，默认为“re”（随机效应）。选择项“normal”表示在估计随机效应模型时，指定个体异质性服从正态分布，默认为服从 Gamma 分布。选择项“irr”表示汇报发生率比（incidence rate ratio）。

17.6 面板负二项回归

泊松回归的缺陷是，它假设均等分散，即方差等于期望。如果存在过度分散，即方差大于期望，则可考虑负二项回归。如果不存在个体异质性，则可直接进行混合负二项回归，其 Stata 命令为

```
nbreg y x1 x2 x3,vce(cluster id)
```

其中，id 为用来确定面板单位的变量。Stata 将输出过度分散参数 α 的 95% 置信区间。根据这个置信区间，可以检验原假设“H_0: $\alpha = 0$”（对应于泊松回归）。

对于随机效应的面板负二项模型，假设个体异质性服从 Beta(r,s) 分布，则可将个体异质性积分掉，然后进行 MLE 估计（此处理方法类似于随机效应的面板泊松模型）。对于固定效应的面板负二项模型，则可考虑在给定 $n_i \equiv \sum_{t=1}^{T} y_{it}$（个体异质性的充分统计量）情况下的条件似然函数，然后进行条件最大似然估计（conditional MLE）。固定效应负二项回归的奇特之处在于，它可以估计不随时间而变的变量系数。在固定效应与随机效应负二项回归之间进行选择时，可使用豪斯曼检验。如何在泊松回归与负二项回归之间进行选择，涉及稳健性与有效性之间的权衡（前者稳健，而后者更有效），有关讨论参见第 13 章第 3 节。

面板负二项回归的 Stata 命令为

```
xtnbreg y x1 x2 x3,fe irr
```

其中，选择项“fe”表示固定效应，默认为“re”（随机效应）。选择项“irr”表示汇报发生率比（incidence rate ratio）。

17.7 面板计数模型的 Stata 实例

下面以数据集 mus18data.dta 为例[①],介绍面板计数模型的 Stata 操作。此数据集来自"Rand Health Insurance Experiment"。该实验随机地给家庭分配不同的健康保险单(health insurance policy),然后跟踪调查几年,故为面板数据。该实验的主要目的是研究不同的共保率(coinsurance rate)如何影响对医疗服务的使用。如果共保率为25%,则意味着投保人支付25%的医疗费用,而保险公司支付75%。

在本例中,被解释变量 mdu 表示个体看医生的次数,为计数变量。解释变量包括 lcoins(共保率加1再取对数,即 log(coinsurance + 1)),ndisease(所患慢性病的数目),female(是否女性),age(年龄),lfam(家庭人数的对数),child(是否有孩子)。

(1) 数据集的结构与统计特征。

```
. use mus18data.dta,clear
. xtset id year
```

```
      panel variable:  id (unbalanced)
       time variable:  year, 1 to 5, but with gaps
               delta:  1 unit
```

由上表可知,面板变量为 id(对应于每位个体),时间跨度为5年,为非平衡面板。

```
. xtdescribe
```

```
     id:  125024, 125025, ..., 632167                    n =        5908
   year:  1, 2, ..., 5                                   T =           5
          Delta(year) = 1 unit
          Span(year)  = 5 periods
          (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%      95%      max
                         1       2        3         3         5        5        5

   Freq.  Percent    Cum.  |  Pattern
  ------------------------+---------
    3710    62.80   62.80  |  111..
    1584    26.81   89.61  |  11111
     156     2.64   92.25  |  1....
     147     2.49   94.74  |  11...
      79     1.34   96.07  |  ..1..
      66     1.12   97.19  |  .11..
      33     0.56   97.75  |  ..111
      33     0.56   98.31  |  .1111
      29     0.49   98.80  |  ...11
      71     1.20  100.00  |  (other patterns)
  ------------------------+---------
    5908   100.00          |  XXXXX
```

① 此数据集取自 Cameron and Trivedi (2010)。

由上表下部可知,89.61% 的个体在前三年或全部五年都在样本中。

```
. sum mdu lcoins ndisease female age lfam child
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mdu	20186	2.860696	4.504765	0	77
lcoins	20186	2.383588	2.041713	0	4.564348
ndisease	20186	11.2445	6.741647	0	58.6
female	20186	.5169424	.4997252	0	1
age	20186	25.71844	16.76759	0	64.27515
lfam	20186	1.248404	.5390681	0	2.639057
child	20186	.4014168	.4901972	0	1

也可使用命令 xtsum 得到更详细的面板数据统计特征。

(2) 进行混合泊松回归,并使用聚类稳健标准误。

```
. poisson mdu lcoins ndisease female age lfam child, vce (cluster id) nolog
```

```
Poisson regression                                Number of obs   =      20186
                                                  Wald chi2(6)    =     476.93
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -62579.401                 Pseudo R2       =     0.0609

                                  (Std. Err. adjusted for 5908 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.0808023   .0080013   -10.10   0.000    -.0964846   -.0651199
    ndisease |   .0339334   .0026024    13.04   0.000     .0288328    .039034
      female |   .1717862   .0342551     5.01   0.000     .1046473    .2389251
         age |   .0040585   .0016891     2.40   0.016      .000748    .0073691
        lfam |  -.1481981   .0323434    -4.58   0.000     -.21159   -.0848062
       child |   .1030453   .0506901     2.03   0.042     .0036944    .2023961
       _cons |    .748789   .0785738     9.53   0.000     .5947872    .9027907
------------------------------------------------------------------------------
```

由上表可知,所有解释变量均至少在 5% 水平上显著。其中,lcoins 与 lfam 的系数显著为负,表明共保率越高、家庭成员越多,则看医生次数越少。ndisease,female,age 与 child 的系数显著为正,表明所患慢性病越多、女性、年龄越大、有子女,则看医生次数越多。

(3) 进行随机效应的面板泊松回归。由于命令"xtpoisson,re"不提供聚类稳健标准误,故使用选择项"vce(boot)"计算聚类自助标准误(有关自助法,参见第 19 章);运行以下命令将花费较长时间。

```
. xtpoisson mdu lcoins ndisease female age lfam child,re vce(boot,reps(50) seed(10101) nodots)
```

其中,"reps(50)"表示重复抽样 50 次,"seed(10101)"表示以 10101 作为随机数的种子(参见第 19 章附录),而"nodots"表示不显示重复抽样过程的点(dots)。

```
Random-effects Poisson regression               Number of obs      =     20186
Group variable: id                              Number of groups   =      5908

Random effects u_i ~ Gamma                      Obs per group: min =         1
                                                               avg =       3.4
                                                               max =         5

                                                Wald chi2(6)       =    476.31
Log likelihood  = -43240.556                    Prob > chi2        =    0.0000

                                (Replications based on 5908 clusters in id)
------------------------------------------------------------------------------
             |               Observed   Bootstrap                         Normal-based
         mdu |                  Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |          -.0878258   .0080482   -10.91   0.000    -.1036   -.0720515
  ndisease   |           .0387629   .0027883    13.90   0.000     .0332979    .044228
      female |           .1667192   .0383524     4.35   0.000     .0915499    .2418885
         age |           .0019159   .0017831     1.07   0.283    -.0015789    .0054107
        lfam |          -.1351786    .034328    -3.94   0.000    -.2024602   -.067897
       child |           .1082678   .0548715     1.97   0.048     .0007218    .2158139
       _cons |           .7574177   .0867116     8.73   0.000      .587466    .9273694
-------------+----------------------------------------------------------------
    /lnalpha |           .0251256   .0278546                     -.0294685    .0797197
-------------+----------------------------------------------------------------
       alpha |           1.025444   .0285634                      .9709615    1.082983
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0: chibar2(01) =  3.9e+04 Prob>=chibar2 = 0.000
```

上表最后一行的 LR 检验结果强烈拒绝“alpha = 0”的原假设(对应于混合泊松回归,其中 α 为个体异质性 v_i 的方差),即拒绝混合泊松模型,认为应使用随机效应的面板泊松模型。命令“xtpoisson,re”默认假设个体异质性 v_i 服从 Gamma 分布。也可使用选择项“normal”指定个体异质性服从正态分布,为了提高速度,下面仅使用普通标准误(不使用自助标准误)。

```
. xtpoisson mdu lcoins ndisease female age lfam child,re normal nolog
```

```
Random-effects Poisson regression               Number of obs      =     20186
Group variable: id                              Number of groups   =      5908

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       3.4
                                                               max =         5

                                                Wald chi2(6)       =    798.96
Log likelihood  = -43226.889                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.1145337   .0072788   -15.74   0.000    -.1287998   -.1002676
  ndisease   |   .0408695   .0022941    17.81   0.000     .0363731    .045366
      female |    .208436   .0304848     6.84   0.000      .148687    .268185
         age |    .002674   .0011961     2.24   0.025     .0003297    .0050182
        lfam |  -.1443327   .0265365    -5.44   0.000    -.1963432   -.0923222
       child |   .0737146   .0344697     2.14   0.032     .0061553    .141274
       _cons |   .2872796   .0641625     4.48   0.000     .1615234    .4130359
-------------+----------------------------------------------------------------
    /lnsig2u |   .0549982   .0254991     2.16   0.031      .005021    .1049755
-------------+----------------------------------------------------------------
     sigma_u |   1.027881    .013105                      1.002514     1.05389
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01) =  3.9e+04 Pr>=chibar2 = 0.000
```

从上表可知，假设个体异质性服从正态分布的估计系数与默认 Gamma 分布的估计系数十分接近。而且，前者的对数似然函数略高于后者，说明前者拟合得更好。由于使用了数值积分，故需要检验数值积分的准确度。

```
. quadchk,nooutput
```

```
Refitting model intpoints() =  8
Refitting model intpoints() = 16

                        Quadrature check

             Fitted       Comparison     Comparison
           quadrature     quadrature     quadrature
           12 points      8 points       16 points
-------------------------------------------------------------------------------
Log        -43226.889     -43226.885     -43226.885
likelihood                 .00461531      .00462363   Difference
                          -1.068e-07     -1.070e-07   Relative difference
-------------------------------------------------------------------------------
mdu:       -.11453372     -.11458364     -.11453628
  lcoins                  -.00004992     -2.567e-06   Difference
                           .00043585      .00002241   Relative difference
-------------------------------------------------------------------------------
mdu:        .04086953      .04088382      .04087027
ndisease                   .00001429      7.382e-07   Difference
                           .00034964      .00001806   Relative difference
-------------------------------------------------------------------------------
mdu:        .20843598      .20851939      .20844042
  female                    .0000834      4.440e-06   Difference
                           .00040013       .0000213   Relative difference
-------------------------------------------------------------------------------
mdu:        .00267396      .00267552      .00267402
     age                  1.561e-06      6.599e-08   Difference
                           .00058392      .00002468   Relative difference
-------------------------------------------------------------------------------
mdu:       -.14433271     -.1443633      -.14433429
    lfam                  -.00003059     -1.580e-06   Difference
                           .00021195      .00001095   Relative difference
-------------------------------------------------------------------------------
mdu:        .07371462      .07371141      .07371365
   child                 -3.210e-06     -9.678e-07   Difference
                          -.00004355     -.00001313   Relative difference
-------------------------------------------------------------------------------
mdu:        .28727964      .28660796      .28726141
   _cons                  -.00067167     -.00001823   Difference
                          -.00233805     -.00006347   Relative difference
-------------------------------------------------------------------------------
lnsig2u:    .05499824      .05529067      .05499982
   _cons                   .00029244      1.580e-06   Difference
                           .00531723      .00002873   Relative difference
-------------------------------------------------------------------------------
```

上表显示,对于某些系数,其相对差距(relative difference)超过了10^{-4}。故下面使用选择项"intpoints(#)"将数值积分的计算点数增加到16(默认为12)。

```
. xtpoisson mdu lcoins ndisease female age lfam child,re normal nolog
intpoints(16)
```

```
Random-effects Poisson regression                Number of obs      =     20186
Group variable: id                               Number of groups   =      5908

Random effects u_i ~ Gaussian                    Obs per group: min =         1
                                                                avg =       3.4
                                                                max =         5

                                                 Wald chi2(6)       =    799.05
Log likelihood  = -43226.885                     Prob > chi2        =    0.0000
```

mdu	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lcoins	-.1145363	.0072785	-15.74	0.000	-.1288019	-.1002706
ndisease	.0408703	.0022941	17.82	0.000	.0363739	.0453666
female	.2084404	.0304838	6.84	0.000	.1486933	.2681875
age	.002674	.001196	2.24	0.025	.0003298	.0050182
lfam	-.1443343	.0265357	-5.44	0.000	-.1963433	-.0923253
child	.0737136	.0344692	2.14	0.032	.0061552	.1412721
_cons	.2872615	.0641609	4.48	0.000	.1615084	.4130146
/lnsig2u	.0549996	.0254986	2.16	0.031	.0050232	.1049759
sigma_u	1.027881	.0131048			1.002515	1.05389

```
Likelihood-ratio test of sigma_u=0: chibar2(01) =  3.9e+04 Pr>=chibar2 = 0.000
```

然后再检验数值积分的精确度。

```
. quadchk,nooutput
```

```
Refitting model intpoints() = 11
Refitting model intpoints() = 21

                        Quadrature check

              Fitted       Comparison     Comparison
            quadrature     quadrature     quadrature
            16 points      11 points      21 points
-------------------------------------------------------------------
Log         -43226.885     -43226.894     -43226.885
likelihood                 -.00925502     -.00029544   Difference
                            2.141e-07      6.835e-09   Relative difference
-------------------------------------------------------------------
mdu:        -.11453627     -.11453627     -.11453611
  lcoins                   -4.418e-11      1.668e-07   Difference
                            3.858e-10     -1.456e-06   Relative difference
-------------------------------------------------------------------
mdu:         .04087026      .04087026      .04087022
ndisease                    1.112e-11     -4.246e-08   Difference
                            2.722e-10     -1.039e-06   Relative difference
-------------------------------------------------------------------
mdu:         .20844041      .20844041      .20844014
  female                    6.809e-11     -2.617e-07   Difference
                            3.266e-10     -1.256e-06   Relative difference
-------------------------------------------------------------------
mdu:         .00267402      .00267402      .00267402
     age                    1.133e-12     -4.253e-09   Difference
                            4.237e-10     -1.591e-06   Relative difference
-------------------------------------------------------------------
mdu:        -.14433431     -.14433431     -.14433418
    lfam                   -2.772e-11      1.299e-07   Difference
                            1.921e-10     -8.998e-07   Relative difference
-------------------------------------------------------------------
mdu:         .07371365      .07371365      .07371368
   child                   -3.389e-13      3.289e-08   Difference
                           -4.598e-12      4.462e-07   Relative difference
-------------------------------------------------------------------
mdu:         .28726149      .28726149      .28726306
   _cons                   -3.947e-10      1.564e-06   Difference
                           -1.374e-09      5.444e-06   Relative difference
-------------------------------------------------------------------
lnsig2u:     .05499955      .05499955       .054999
   _cons                    2.615e-10     -5.539e-07   Difference
                            4.755e-09     -.00001007   Relative difference
-------------------------------------------------------------------
```

上表显示，所有系数的相对差距均小于10^{-4}，故可以认为结果是稳健的。

(4) 进行固定效应的面板泊松回归。但命令“xtpoisson,fe”并不提供聚类稳健标准误。解决方法之一是使用自助标准误的选择项，比如“vce(boot,reps(50) seed(10101) nodots)”。解决方法之二为下载非官方命令“xtpqml”(其中，pqml 表示 Poisson Quasi Maximum Likelihood)，该命令的估计结果与“xtpoisson,fe”相同，但提供聚类稳健标准误。

```
. ssc install xtpqml     (下载安装命令 xtpqml)
. xtpqml mdu lcoins ndisease female age lfam child,i(id)
```

其中，必选项“i(id)”表示以变量 id 来确定面板单位。

```
note: 265 groups (265 obs) dropped because of only one obs per group
note: 666 groups (2130 obs) dropped because of all zero outcomes
note: lcoins dropped because it is constant within group
note: ndisease dropped because it is constant within group
note: female dropped because it is constant within group

Iteration 0:   log likelihood = -24182.852
Iteration 1:   log likelihood = -24173.211
Iteration 2:   log likelihood = -24173.211

Conditional fixed-effects Poisson regression    Number of obs      =     17791
Group variable: id                              Number of groups   =      4977

                                                Obs per group: min =         2
                                                               avg =       3.6
                                                               max =         5

                                                Wald chi2(3)       =     19.24
Log likelihood  = -24173.211                    Prob > chi2        =    0.0002
```

mdu	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0112009	.0039024	-2.87	0.004	-.0188494	-.0035523
lfam	.0877134	.0554606	1.58	0.114	-.0209874	.1964141
child	.1059867	.0437744	2.42	0.015	.0201905	.1917829

Calculating Robust Standard Errors...

mdu	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mdu						
age	-.0112009	.0091493	-1.22	0.221	-.0291331	.0067314
lfam	.0877134	.1160837	0.76	0.450	-.1398064	.3152332
child	.1059867	.0786326	1.35	0.178	-.0481304	.2601037

Wald chi2(3) = 4.58　　Prob > chi2 = 0.2051

上表前五行显示，有 265 组数据由于仅有一期观测值而被去掉，另有 666 组数据由于取值全为 0 而被去掉；而且，固定效应模型无法识别不随时间而变的变量系数（lcoins，ndisease，female），故这些变量也被去掉。回归结果的上半部提供普通标准误（事实上，复制了命令“xtpoisson，fe”的估计结果）；而下半部的标准误则为聚类稳健标准误。可以看出，聚类稳健标准误大约是普通标准误的两倍，导致所有变量系数均不显著。

（5）初步考察是否存在过度分散。

```
. sum mdu,detail
```

number face-to-fact md visits

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	20186
25%	0	0	Sum of Wgt.	20186
50%	1		Mean	2.860696
		Largest	Std. Dev.	4.504765
75%	4	72		
90%	7	74	Variance	20.2929
95%	10	76	Skewness	4.823598
99%	21	77	Kurtosis	44.75799

```
. dis r(Var)/r(mean)
7.093696
```

由此可知，被解释变量 mdu 的方差是平均值的 7 倍多，故可能存在过度分散。存在过度分散的情况下，尽管泊松回归依然是一致的，但负二项回归可能更有效率。

（6）进行混合负二项回归，并使用聚类稳健标准误。

```
. nbreg mdu lcoins ndisease female age lfam child,vce(cluster id) nolog
```

Negative binomial regression — Number of obs = 20186
Dispersion = mean — Wald chi2(6) = 531.13
Log pseudolikelihood = -43403.178 — Prob > chi2 = 0.0000

(Std. Err. adjusted for 5908 clusters in id)

mdu	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lcoins	-.0855356	.0079578	-10.75	0.000	-.1011325	-.0699387
ndisease	.036683	.0025083	14.62	0.000	.0317668	.0415993
female	.1746153	.0344054	5.08	0.000	.1071819	.2420486
age	.0035134	.0015948	2.20	0.028	.0003876	.0066392
lfam	-.1588963	.032843	-4.84	0.000	-.2232673	-.0945253
child	.1067837	.0523271	2.04	0.041	.0042244	.2093429
_cons	.7500703	.0805942	9.31	0.000	.5921086	.9080321
/lnalpha	.2621342	.0248113			.2135049	.3107634
alpha	1.299701	.0322473			1.23801	1.364466

上表最后一行显示，过度分散参数 α 的 95% 置信区间为[1.24,1.36]，故拒绝“H_0: $\alpha=0$”，认为存在过度分散，使用负二项回归可以提高效率。

(7) 进行随机效应的面板负二项回归，并使用自助标准误；运行此命令需较长时间。

```
. xtnbreg mdu lcoins ndisease female age lfam child,nolog vce(boot,
reps(50) seed(10101) nodots)
```

```
Random-effects negative binomial regression     Number of obs      =     20186
Group variable: id                              Number of groups   =      5908

Random effects u_i ~ Beta                       Obs per group: min =         1
                                                               avg =       3.4
                                                               max =         5

                                                Wald chi2(6)       =    617.55
Log likelihood  = -40661.277                    Prob > chi2        =    0.0000

                                (Replications based on 5908 clusters in id)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.1072858   .0071451   -15.02   0.000    -.1212899   -.0932817
   ndisease |   .0333777   .0021562    15.48   0.000     .0291517    .0376038
      female |   .2039408   .0305068     6.69   0.000     .1441485    .2637331
         age |   .0023233   .0014846     1.56   0.118    -.0005865     .005233
        lfam |  -.1434145   .0281385    -5.10   0.000    -.1985651   -.088264
       child |   .1144578   .0410581     2.79   0.005     .0339853    .1949302
       _cons |   .8821153   .0791412    11.15   0.000     .7270014    1.037229
-------------+----------------------------------------------------------------
        /ln_r |   1.127977   .0287538                      1.071621    1.184334
        /ln_s |   .7258626   .0355438                      .6561979    .7955272
-------------+----------------------------------------------------------------
           r |   3.089401   .0888319                      2.920109    3.268508
           s |   2.066513   .0734518                       1.92745    2.215609
------------------------------------------------------------------------------
Likelihood-ratio test vs. pooled: chibar2(01) =  5305.26 Prob>=chibar2 = 0.000
```

上表最后一行的 LR 检验结果强烈拒绝了混合负二项回归的原假设，认为应使用随机效应的面板负二项回归。

(8) 进行固定效应的面板负二项回归，并使用自助标准误；运行此命令需较长时间。

```
. xtnbreg mdu lcoins ndisease female age lfam child,fe nolog vce(boot,
reps(50) seed(10101) nodots)
```

```
Conditional FE negative binomial regression      Number of obs      =     17791
Group variable: id                               Number of groups   =      4977

                                                 Obs per group: min =         2
                                                               avg =       3.6
                                                               max =         5

                                                 Wald chi2(6)       =     67.42
Log likelihood  = -21626.982                     Prob > chi2        =    0.0000

                                 (Replications based on 4977 clusters in id)
------------------------------------------------------------------------------
             |               Observed   Bootstrap                         Normal-based
         mdu |                  Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |          -.088532   .0199968    -4.43   0.000    -.127725   -.0493391
   ndisease |          .0153807   .0061465     2.50   0.012    .0033338    .0274276
      female |          .2459735   .0636349     3.87   0.000    .1212514    .3706956
         age |         -.0020748   .0026177    -0.79   0.428   -.0072054    .0030558
        lfam |         -.0007717   .0570776    -0.01   0.989   -.1126417    .1110983
       child |          .2032379   .0532196     3.82   0.000    .0989293    .3075464
       _cons |          .9243416   .1166414     7.92   0.000    .6957287    1.152955
------------------------------------------------------------------------------
```

由上表可知,固定效应的面板负二项回归依然可以估计不随时间而变的变量系数,这是固定效应负二项回归的优点。

(9) 最后,使用豪斯曼检验在固定效应与随机效应的负二项回归之间进行选择。

```
. quietly xtnbreg mdu lcoins ndisease female age lfam child
. est sto re
. quietly xtnbreg mdu lcoins ndisease female age lfam child,fe
. est sto fe
. hausman fe re
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          S.E.
-------------+----------------------------------------------------------------
      lcoins |    -.088532   -.1072858        .0187537        .0124662
   ndisease |     .0153807    .0333777       -.017997         .00347
      female |     .2459735    .2039408        .0420327        .0523672
         age |    -.0020748    .0023233       -.004398         .0016488
        lfam |    -.0007717   -.1434145        .1426429        .0406003
       child |     .2032379    .1144578        .0887801        .0382981
------------------------------------------------------------------------------
                b = consistent under Ho and Ha; obtained from xtnbreg
     B = inconsistent under Ha, efficient under Ho; obtained from xtnbreg

  Test:  Ho:  difference in coefficients not systematic

              chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       101.59
            Prob>chi2 =       0.0000
```

上表结果显示,豪斯曼检验强烈拒绝随机效应负二项回归,支持固定效应负二项回归。

17.8 面板 Tobit

考虑归并数据(censored data)的面板模型。假设

$$y_{it}^{*} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it} \tag{17.24}$$

其中,y_i^{*} 不可观测,扰动项 $\varepsilon_{it} \sim N(0,\sigma_\varepsilon^2)$,而 u_i 为个体效应。如果 $u_1 = u_2 = \cdots = u_n$,则可直接进行混合 Tobit 回归(参见第 14 章),但须使用聚类稳健标准误。混合 Tobit 回归的 Stata 命令为

```
tobit y x1 x2 x3,ll(#) ul(#) vce(cluster id)
```

其中,选择项"ll(#)"表示下限(lower limit),选择项"ul(#)"表示上限(upper limit),而 id 为指示面板单位的变量。

更一般地,我们允许个体效应的存在。不失一般性,为简便起见,假定在 0 处存在左归并(left-censored)。假设可以观测到 $y_{it} = \begin{cases} y_{it}^{*}, & 若\ y_{it}^{*} > 0 \\ 0, & 若\ y_{it}^{*} \leqslant 0 \end{cases}$。如果 u_i 与解释变量 $\boldsymbol{x}_{it}$ 不相关,则为随机效应模型(RE);反之,则为固定效应模型(FE)。对于固定效应的 Tobit 模型,由于找不到个体异质性 u_i 的充分统计量,故无法像固定效应的 Logit 或计数模型那样进行条件最大似然估计。如果直接在混合 Tobit 回归中加入面板单位的虚拟变量(类似于 LSDV 法),所得的固定效应估计量也是不一致的。因此,下面仅考虑随机效应的 Tobit 模型。

首先,在给定个体效应 u_i 的情况下,个体 i 的条件分布为

$$\begin{aligned} & f(y_{i1},y_{i2},\cdots,y_{iT} \mid u_i) \\ = & \prod_{t=1}^{T} \left[1 - \Phi((\boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i)/\sigma_\varepsilon)\right]^{1(y_{it}=0)} \left[\frac{1}{\sigma_\varepsilon}\phi((y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta} - u_i)/\sigma_\varepsilon)\right]^{1(y_{it}>0)} \end{aligned} \tag{17.25}$$

上式的推导方法与第 14 章第 5 节类似。但上式中的个体异质性 u_i 不可观测。假设 $u_i \sim N(0,\sigma_u^2)$,记其概率密度函数为 $g(u_i)$。记 $(y_{i1},y_{i2},\cdots,y_{iT},u_i)$ 的联合密度为 $f(y_{i1},y_{i2},\cdots,y_{iT},u_i)$,并进行如下分解,

$$f(y_{i1},y_{i2},\cdots,y_{iT},u_i) = f(y_{i1},y_{i2},\cdots,y_{iT} \mid u_i) \cdot g(u_i) \tag{17.26}$$

将上式中的 u_i 积分积掉,即可得到 $(y_{i1},y_{i2},\cdots,y_{iT})$ 的边缘密度(或无条件分布),

$$\begin{aligned} f(y_{i1},y_{i2},\cdots,y_{iT}) &= \int_{-\infty}^{+\infty} f(y_{i1},y_{i2},\cdots,y_{iT},u_i)\,du_i \\ &= \int_{-\infty}^{+\infty} f(y_{i1},y_{i2},\cdots,y_{iT} \mid u_i) \cdot g(u_i)\,du_i \end{aligned} \tag{17.27}$$

上式的积分无解析解,一般使用"Gauss-Hermite quadrature"方法进行数值积分;也需要使用 Stata 命令 quadchk 来检验此数值积分的精确度。类似地,可通过检验"H_0: $\sigma_u = 0$"(Stata 称为"sigma_u = 0")来判断是否存在个体异质性。另外,可以定义同一个体不同期扰动项的自相关系数为(Stata 记为 rho),

$$\rho \equiv \mathrm{Corr}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \quad (t \neq s) \tag{17.28}$$

如果 ρ 越大,则复合扰动项 $(u_i + \varepsilon_{it})$ 中个体效应的部分 (u_i) 越重要。特别地,如果 $\rho = 0$,则说明 $\sigma_u^2 = 0$,即不存在个体随机效应,而应选择混合回归。

随机效应的面板 Tobit 回归的 Stata 命令为

```
xttobit y x1 x2 x3,ll(#) ul(#) tobit
```

其中,选择项“ll(#)”表示下限(lower limit),选择项“ul(#)”表示上限(upper limit),而选择项“tobit”表示对原假设“H_0: $\sigma_u = 0$”进行 LR 检验。

下面继续以数据集 mus18data. dta 为例①,介绍面板 Tobit 模型的 Stata 操作。考虑被解释变量 med(医疗支出),而解释变量与前面的面板计数模型实例相同。显然,医疗开支为非负,故在 0 处存在左归并。

首先,进行混合 Tobit 回归,并使用聚类稳健标准误。

```
. tobit med lcoins ndisease female age lfam child, ll (0) nolog vce (cluster id)
```

```
Tobit regression                                  Number of obs     =      20186
                                                  F(  6, 20180)     =      44.99
                                                  Prob > F          =     0.0000
Log pseudolikelihood = -130419.54                 Pseudo R2         =     0.0033

                                  (Std. Err. adjusted for 5908 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         med |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -29.59847   3.603346    -8.21   0.000    -36.66132   -22.53562
   ndisease |    13.1485   1.509925     8.71   0.000     10.18892    16.10807
      female |    64.7522   14.45226     4.48   0.000      36.4246    93.0798
         age |    4.05935   .9118431     4.45   0.000     2.272063    5.846637
        lfam |  -52.63194   19.81777    -2.66   0.008    -91.47639   -13.78749
       child |  -54.00112    27.0619    -2.00   0.046    -107.0446   -.9575866
       _cons |  -99.85116   39.26419    -2.54   0.011    -176.8122   -22.89014
-------------+----------------------------------------------------------------
       /sigma |    795.977   77.26739                      644.5266    947.4274
------------------------------------------------------------------------------
  Obs. summary:       4453  left-censored observations at med<=0
                     15733     uncensored observations
                         0 right-censored observations
```

其次,进行随机效应的面板 Tobit 回归。

```
. xttobit med lcoins ndisease female age lfam child,ll(0) nolog tobit
```

① 此例来自 Cameron and Trivedi (2010)。

```
Random-effects tobit regression                 Number of obs      =     20186
Group variable: id                              Number of groups   =      5908

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       3.4
                                                               max =         5

                                                Wald chi2(6)       =    573.45
Log likelihood  = -130030.45                    Prob > chi2        =    0.0000
```

med	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lcoins	-31.10247	3.578498	-8.69	0.000	-38.1162	-24.08875
ndisease	13.49452	1.139156	11.85	0.000	11.26182	15.72722
female	60.10112	14.95966	4.02	0.000	30.78072	89.42152
age	4.075582	.7238253	5.63	0.000	2.656911	5.494254
lfam	-57.75023	14.68422	-3.93	0.000	-86.53077	-28.96968
child	-52.02314	24.21619	-2.15	0.032	-99.48599	-4.560284
_cons	-98.27203	36.05977	-2.73	0.006	-168.9479	-27.59618
/sigma_u	371.3134	8.64634	42.94	0.000	354.3668	388.2599
/sigma_e	715.1779	4.704581	152.02	0.000	705.9571	724.3987
rho	.2123246	.0086583			.1957541	.2296872

```
Likelihood-ratio test of sigma_u=0: chibar2(01)=  778.18 Prob>=chibar2 = 0.000

  Observation summary:      4453  left-censored observations
                           15733     uncensored observations
                               0 right-censored observations
```

上表底部的 LR 检验结果强烈拒绝“H_0: $\sigma_u = 0$”，故认为存在个体效应，应使用随机效应的面板 Tobit 回归。

17.9 面板随机前沿模型

第 14 章介绍了横截面数据的随机前沿模型。将该模型用于估计生产函数可得：

$$\ln y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{ki} + \underbrace{v_i - u_i}_{\varepsilon_i}, \quad u_i \geqslant 0 \tag{17.29}$$

其中，$u_i \geqslant 0$ 为“无效率项”(inefficiency term)，而 v_i 为“idiosyncratic error”。为了得到一致估计，通常假设 u_i 与 $\boldsymbol{x}_i$ 不相关。但此假设未必成立，因为企业很可能知道自己的无效率项 u_i，并以此调整其最优投入水平 $\boldsymbol{x}_i$。此难题在横截面数据的情况下很难解决；但如果有面板数据，则可以允许 u_i 与 $\boldsymbol{x}_i$ 相关。事实上，即使不考虑 u_i 与 $\boldsymbol{x}_i$ 相关性，使用横截面数据也得不到对无效率项 u_i 的一致估计，因为随着横截面单位 n 增大，待估计的无效率项也同步增多，故存在“伴生参数问题”(incidental parameters problem)。而如果使用面板数据，则可在 $T \to \infty$ 的条件下得到一致估计(这要求面板数据的时间维度较长，即为长面板)。

取决于无效率项 u_i 是否随时间而变，可将面板随机前沿模型分为以下两类。

1. 技术效率不随时间而变

考虑如下模型：

$$\ln y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{kit} + v_{it} - u_i, \quad u_i \geqslant 0 \tag{17.30}$$

其中，无效率项 u_i 不随时间而变，称为“time-invariant technical efficiency”。此模型与标准的线性面板模型十分类似（参见第 15 章），只是个体效应（$-u_i$）一定是非正的（nonpositive）。如果 u_i 与 $\boldsymbol{x}_i$ 相关，则为固定效应模型；反之，则为随机效应模型。

对于固定效应模型，可将方程（17.30）写为

$$\ln y_{it} = \beta_{0i} + \sum_{k=1}^{K} \beta_k \ln x_{kit} + v_{it} \tag{17.31}$$

其中，$\beta_{0i} \equiv \beta_0 - u_i$ 为企业 i 的特有截距项。显然，方程（17.31）可使用 LSDV 法进行估计，即引入（$n-1$）个企业虚拟变量进行 OLS 估计（参见第 15 章）。记截距项的估计结果为 $\hat{\beta}_{0i}$，并计算其最大值 $\hat{\beta}_0 \equiv \max_i \{\hat{\beta}_{0i}\}$，则可通过下式估计无效率项：

$$\hat{u}_i = \hat{\beta}_0 - \hat{\beta}_{0i} \tag{17.32}$$

显然，$\hat{u}_i \geqslant 0$，且至少有一家企业达到了效率边界；其他企业的效率水平则为离开此效率边界的相对距离。

对于随机效应模型，可对 u_i 与 v_{it} 的分布作出具体假定，然后进行 MLE 估计。这些假定与横截面数据的设定相似，即 u_i，v_{it} 均为 iid 且相互独立，而且独立于解释变量 $\boldsymbol{x}_i$；$v_{it} \sim N(0, \sigma_v^2)$；而 u_i 服从断尾正态分布，即 $u_i \sim N^+(\mu, \sigma_u^2)$，且在原点左边断尾。在这些假定之下，可得出对数似然函数（推导过程类似于横截面数据的情形，参见 Kumbhakar and Lovell, 2000, p. 102 – 105）。类似地，可使用条件期望 $\mathrm{E}(u_i \mid \varepsilon_i)$ 作为无效率项 u_i 的估计，其中 $\varepsilon_{it} \equiv v_{it} - u_i$。

2. 技术效率随时间而变

如果面板数据的时间维度较大，则“技术效率不随时间而变”的假定可能不现实。在上述随机效应模型的框架下，Battese and Coelli（1992）假设无效率项 u_{it} 同时随企业与时间而变：

$$u_{it} = \mathrm{e}^{-\eta(t - T_i)} u_i \tag{17.33}$$

其中，T_i 为企业 i 的时间维度（允许非平衡面板），η 为待估参数，而 $u_i \sim N^+(\mu, \sigma_u^2)$。上式表明，$u_{it}$ 随时间而递减，直至到最后一期 T_i 时，$u_{iT_i} = u_i$；故称为“时变衰减模型”（time-varying decay model）。显然，如果 $\eta = 0$，则退化为技术效率不随时间而变的模型；对此，只要检验“H_0: $\eta = 0$”，即可判断。估计方法仍为 MLE。当无效率项 u_{it}（表示离生产前沿的距离）变化时，生产前沿本身也可能变化；这种技术进步（technical change）的效应可通过加入时间虚拟变量来捕捉。

对于技术效率不随时间而变的固定效应模型，可直接用 Stata 命令 `reg` 进行 LSDV 估计。对于随机效应模型，则有以下专门的 Stata 命令：

```
xtfrontier lny lnx1 lnx2 lnx3,cost ti tvd
```

其中，“`lny`”为对数形式的被解释变量，“`lnx1 lnx2 lnx3`”为对数形式的解释变量。选择项“`cost`”表示估计成本函数，默认估计生产函数。选择项“`ti`”表示估计技术效率不随时间而变的随机效应模型，而选择项“`tvd`”表示估计随机效应的时变衰减模型。

下面以 Stata 提供的数据集“xtfrontier1. dta”为例进行演示。该数据集的主要变量包括：lnwidgets（产量对数）、lnmachines（资本对数）、lnworkers（劳动力对数）、id（个体变量）以及 t（时间变量）。

首先，使用固定效应的 LSDV 法。

```
. use xtfrontier1.dta
```

```
. reg lnwidgets lnmachines lnworkers i.id,vce(cluster id)
```

Linear regression

Number of obs	= 948
F(1, 90)	= .
Prob > F	= .
R-squared	= 0.7243
Root MSE	= 1.006

(Std. Err. adjusted for 91 clusters in id)

lnwidgets	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnmachines	.2918565	.0152287	19.16	0.000	.2616021	.3221109
lnworkers	.2985121	.0174847	17.07	0.000	.2637756	.3332486
id						
2	-1.695635	.041544	-40.82	0.000	-1.778169	-1.613101
3	1.425356	.0319736	44.58	0.000	1.361835	1.488877
4	-.5109693	.0171278	-29.83	0.000	-.5449967	-.4769419
5	.0050908	.0215518	0.24	0.814	-.0377256	.0479072
6	-.6426216	.0216263	-29.71	0.000	-.6855859	-.5996572
7	-1.622634	.0164675	-98.54	0.000	-1.65535	-1.589918
8	-.0976995	.0265487	-3.68	0.000	-.1504431	-.0449559
9	-.9882372	.036626	-26.98	0.000	-1.061001	-.9154732
10	1.236901	.0299493	41.30	0.000	1.177401	1.2964
11	-1.732293	.0110949	-156.13	0.000	-1.754335	-1.710251
12	-2.131603	.017535	-121.56	0.000	-2.166439	-2.096767
13	.9925154	.0365132	27.18	0.000	.9199754	1.065055
14	.5199811	.0242669	21.43	0.000	.4717706	.5681916
15	-1.690762	.0318765	-53.04	0.000	-1.75409	-1.627434
16	-2.948755	.0237505	-124.16	0.000	-2.99594	-2.901571
17	-1.933026	.0235441	-82.10	0.000	-1.9798	-1.886251
18	-3.325573	.039951	-83.24	0.000	-3.404943	-3.246203
19	-.581822	.0434022	-13.41	0.000	-.668048	-.495596
20	-1.204938	.0323211	-37.28	0.000	-1.26915	-1.140727
21	-1.363309	.0422181	-32.29	0.000	-1.447182	-1.279435
22	.0796255	.0159948	4.98	0.000	.0478489	.111402
23	1.216471	.0300302	40.51	0.000	1.156811	1.276131
24	-2.744233	.015575	-176.20	0.000	-2.775176	-2.713291
25	1.040995	.0175595	59.28	0.000	1.00611	1.07588
26	1.131819	.0422106	26.81	0.000	1.04796	1.215678
27	-1.467872	.0298633	-49.15	0.000	-1.527201	-1.408544
28	.8157485	.0242732	33.61	0.000	.7675255	.8639714
29	.7387448	.0197171	37.47	0.000	.6995733	.7779163
30	-1.487877	.0266645	-55.80	0.000	-1.540851	-1.434904
31	1.009767	.0080488	125.46	0.000	.9937771	1.025758
32	-1.808957	.0210879	-85.78	0.000	-1.850852	-1.767062
33	-2.144318	.0310849	-68.98	0.000	-2.206074	-2.082563
34	-2.50431	.0074786	-334.86	0.000	-2.519168	-2.489453
35	-2.921074	.0276509	-105.64	0.000	-2.976008	-2.866141
36	.4856123	.0300699	16.15	0.000	.4258731	.5453514
37	-2.318	.0296372	-78.21	0.000	-2.376879	-2.25912
38	.9231861	.0114178	80.85	0.000	.9005026	.9458697
39	-2.132264	.0287247	-74.23	0.000	-2.18933	-2.075197
40	1.46387	.0142943	102.41	0.000	1.435472	1.492268
41	.2768492	.0148452	18.65	0.000	.2473565	.3063419
42	1.049103	.020893	50.21	0.000	1.007595	1.090611
43	-.7912524	.0346812	-22.82	0.000	-.8601526	-.7223521
44	.9562655	.028856	33.14	0.000	.8989379	1.013593
45	-2.608002	.0232772	-112.04	0.000	-2.654247	-2.561758
46	-1.187144	.0343153	-34.60	0.000	-1.255317	-1.11897
47	-1.694994	.0386311	-43.88	0.000	-1.771742	-1.618247
48	.4930057	.0279105	17.66	0.000	.4375566	.5484548
49	-.8026196	.0343494	-23.37	0.000	-.8708608	-.7343785
50	.3637465	.02856	12.74	0.000	.3070071	.4204858
51	-3.566843	.0314315	-113.48	0.000	-3.629287	-3.504399
52	.6163288	.0374223	16.47	0.000	.5419828	.6906748
53	-2.777265	.0280012	-99.18	0.000	-2.832895	-2.721636
54	.6421726	.0356541	18.01	0.000	.5713394	.7130057
55	1.260894	.0430793	29.27	0.000	1.17531	1.346479
56	-2.00337	.0597843	-33.51	0.000	-2.122142	-1.884598
57	-.1839978	.0266548	-6.90	0.000	-.2369522	-.1310434
58	-1.836363	.0325928	-56.34	0.000	-1.901114	-1.771611
59	.7331885	.0290314	25.26	0.000	.6755125	.7908644
60	1.270419	.0194528	65.31	0.000	1.231772	1.309065
61	-.9638028	.0188156	-51.22	0.000	-1.001183	-.9264224
62	.0719316	.0253706	2.84	0.006	.0215285	.1223347
63	-1.591004	.0331893	-47.94	0.000	-1.65694	-1.525067
64	.9536491	.0334183	28.54	0.000	.8872577	1.02004
65	-2.563908	.0253687	-101.07	0.000	-2.614308	-2.513509
66	-.2381077	.020969	-11.36	0.000	-.2797663	-.1964491
67	-1.145653	.0138836	-82.52	0.000	-1.173235	-1.11807
68	1.218136	.0268255	45.41	0.000	1.164843	1.27143
69	.5956427	.0244782	24.33	0.000	.5470125	.644273
70	.102732	.0190324	5.40	0.000	.0649208	.1405432
71	-2.271472	.0175619	-129.34	0.000	-2.306362	-2.236582
72	.9855415	.0211339	46.63	0.000	.9435552	1.027528
73	-.5846073	.0160393	-36.45	0.000	-.6164721	-.5527425
74	-.1471411	.0128814	-11.42	0.000	-.1727322	-.12155
75	-2.15375	.0180408	-119.38	0.000	-2.189591	-2.117909
76	.7950362	.0132114	60.18	0.000	.7687896	.8212829
77	-.2138768	.0178099	-12.01	0.000	-.2492594	-.1784943
78	-.17695	.0281078	-6.30	0.000	-.2327911	-.1211088
79	.7954589	.0421389	18.88	0.000	.7117427	.8791751
80	.500001	.0251092	19.91	0.000	.4501172	.5498848
81	-1.953061	.0306898	-63.64	0.000	-2.014031	-1.89209
82	-.1321939	.0342241	-3.86	0.000	-.200186	-.0642018
83	-.0354772	.0249803	-1.42	0.159	-.0851048	.0141504
84	-1.027946	.0329124	-31.23	0.000	-1.093332	-.9625596
85	-.6832718	.0342629	-19.94	0.000	-.7513411	-.6152025
86	-.4829249	.0199879	-24.16	0.000	-.5226343	-.4432155
87	.307109	.0385769	7.96	0.000	.2304693	.3837487
88	-1.503912	.0278106	-54.08	0.000	-1.559163	-1.448661
89	.0719344	.0375732	1.91	0.059	-.0027113	.14658
90	-1.147258	.0352705	-32.53	0.000	-1.217329	-1.077187
91	.0047903	.0600026	0.08	0.937	-.1144152	.1239959
_cons	1.692352	.0444898	38.04	0.000	1.603965	1.780738

上表显示,绝大多数的个体虚拟变量均很显著,证明存在个体效应。

其次,估计技术效率不随时间而变的随机效应模型。

```
. xtfrontier lnwidgets lnmachines lnworkers,ti nolog
```

```
Time-invariant inefficiency model                Number of obs      =       948
Group variable: id                               Number of groups   =        91

                                                 Obs per group: min =         6
                                                               avg =      10.4
                                                               max =        14

                                                 Wald chi2(2)       =    661.76
Log likelihood  = -1472.6069                     Prob > chi2        =    0.0000
```

lnwidgets	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnmachines	.2904551	.0164219	17.69	0.000	.2582688	.3226415
lnworkers	.2943333	.0154352	19.07	0.000	.2640808	.3245858
_cons	3.030983	.1441022	21.03	0.000	2.748548	3.313418
/mu	1.125667	.6479217	1.74	0.082	-.144236	2.39557
/lnsigma2	1.421979	.2672745	5.32	0.000	.898131	1.945828
/ilgtgamma	1.138685	.3562642	3.20	0.001	.4404204	1.83695
sigma2	4.145318	1.107938			2.455011	6.999424
gamma	.7574382	.0654548			.6083592	.8625876
sigma_u2	3.139822	1.107235			.9696821	5.309962
sigma_v2	1.005496	.0484143			.9106055	1.100386

下面,预测无效率项 u_{it},将其记为 uhat,并罗列前 15 个观测值。

```
. predict uhat,u
. list id uhat in 1/15
```

	id	uhat
1.	1	1.340585
2.	1	1.340585
3.	1	1.340585
4.	1	1.340585
5.	1	1.340585
6.	1	1.340585
7.	1	1.340585
8.	1	1.340585
9.	1	1.340585
10.	1	1.340585
11.	1	1.340585
12.	2	2.977796
13.	2	2.977796
14.	2	2.977796
15.	2	2.977796

再次,估计随机效应的时变衰减模型。

```
. xtfrontier lnwidgets lnmachines lnworkers,tvd nolog
```

```
Time-varying decay inefficiency model           Number of obs      =       948
Group variable: id                              Number of groups   =        91

Time variable: t                                Obs per group: min =         6
                                                               avg =      10.4
                                                               max =        14

                                                Wald chi2(2)       =    661.93
Log likelihood  = -1472.5289                    Prob > chi2        =    0.0000
```

lnwidgets	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnmachines	.2907555	.0164376	17.69	0.000	.2585384	.3229725
lnworkers	.2942412	.0154373	19.06	0.000	.2639846	.3244978
_cons	3.028939	.1436046	21.09	0.000	2.74748	3.310399
/mu	1.110831	.6452809	1.72	0.085	-.1538967	2.375558
/eta	.0016764	.00425	0.39	0.693	-.0066535	.0100064
/lnsigma2	1.410723	.2679485	5.26	0.000	.885554	1.935893
/ilgtgamma	1.123982	.3584243	3.14	0.002	.4214828	1.82648
sigma2	4.098919	1.098299			2.424327	6.930228
gamma	.7547265	.0663495			.603838	.8613419
sigma_u2	3.093563	1.097606			.9422943	5.244832
sigma_v2	1.005356	.0484079			.9104785	1.100234

上表的 eta 表示 η,估计值仅为 0.001 7(十分接近 0),且 p 值高达 0.693;故可认为技术效率不随时间而变。由于该样本的平均时间跨度为 10.4 年,可能发生技术变迁,下面加入时间虚拟变量重新估计。

```
. xtfrontier lnwidgets lnmachines lnworkers i.t,tvd nolog
```

Time-varying decay inefficiency model — Number of obs = 948
Group variable: id — Number of groups = 91

Time variable: t — Obs per group: min = 6
avg = 10.4
max = 14

Wald chi2(15) = 685.79
Log likelihood = -1465.4647 — Prob > chi2 = 0.0000

lnwidgets	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnmachines	.2932471	.0163939	17.89	0.000	.2611156	.3253786
lnworkers	.2937614	.0153939	19.08	0.000	.2635899	.3239329
t						
2	-.2275212	.1483311	-1.53	0.125	-.5182449	.0632025
3	-.1717603	.1507034	-1.14	0.254	-.4671335	.1236129
4	-.2104515	.1546102	-1.36	0.173	-.5134819	.0925789
5	-.1665324	.1594565	-1.04	0.296	-.4790614	.1459966
6	-.3096068	.165391	-1.87	0.061	-.6337672	.0145536
7	-.2099839	.1758252	-1.19	0.232	-.554595	.1346272
8	-.1819103	.1875663	-0.97	0.332	-.5495335	.185713
9	-.1222577	.2016833	-0.61	0.544	-.5175498	.2730343
10	-.4271937	.2134572	-2.00	0.045	-.8455622	-.0088253
11	-.5210739	.2265704	-2.30	0.021	-.9651437	-.0770041
12	-.4829743	.2504117	-1.93	0.054	-.9737722	.0078236
13	-.2995339	.284717	-1.05	0.293	-.857569	.2585013
14	-.913919	.3452203	-2.65	0.008	-1.590538	-.2372996
_cons	3.264047	.1949113	16.75	0.000	2.882028	3.646066
/mu	1.079811	.5944632	1.82	0.069	-.0853158	2.244937
/eta	.0128377	.0073383	1.75	0.080	-.0015452	.0272206
/lnsigma2	1.314226	.2642142	4.97	0.000	.796376	1.832077
/ilgtgamma	1.015936	.3631487	2.80	0.005	.3041774	1.727694
sigma2	3.72187	.9833712			2.21749	6.246846
gamma	.7341802	.070872			.5754634	.8491172
sigma_u2	2.732523	.9825715			.8067188	4.658328
sigma_v2	.989347	.0476378			.8959787	1.082715

从上表可知，有三个时间虚拟变量在5%水平上显著。而且，时变系数 η(eta)也变得在10%水平上显著(p 值为 0.08)。

习　　题

17.1 使用 Stata 提供的数据集 union. dta 估计面板二值选择模型。被解释变量为 union(是否工会成员)，解释变量包括 age(年龄)、grade(教育年限)、south(是否南方)、black(是否黑人)以及 not_smsa(不在大城市)。

(1) 进行混合 Logit 回归；

(2) 估计随机效应的面板 Logit 模型；

(3) 估计固定效应的面板 Logit 模型；

（4）通过豪斯曼检验，确定应使用以上哪种模型；

（5）估计随机效应的面板 Probit 模型。

17.2 使用数据集 patr7079.dta 估计面板计数模型。此数据集来自 Cameron and Trivedi（2005），最初为 Hall et al（1986）所使用。该数据集包括 346 家企业在 1970—1979 年所获专利数据。被解释变量为 PAT（企业于该年申请并最终获批的专利数），主要解释变量为 LOGR（企业该年的 R&D 支出），其他解释变量包括 SCISECT（企业是否属于科学部门，the scientific sector），以及 LOGK（1972 年的企业账面资本额）。

（1）进行混合泊松回归；

（2）进行随机效应的面板泊松回归；

（3）进行固定效应的面板泊松回归；

（4）进行随机效应的面板负二项回归；

（5）进行固定效应的面板负二项回归；

（6）通过豪斯曼检验，确定应使用以上哪个回归。

第 18 章　随机实验与自然实验

大多数实证分析的目的是为了揭示变量之间的因果关系。然而，一般的经济数据并不足以有力地证明因果关系，除非是实验数据。为此，经济学家越来越热衷于使用“随机实验”或“自然实验”数据。本章内容主要取自 Stock and Watson (2011)。

18.1　实 验 数 据

实验方法是研究因果关系的有力工具。假设我们要研究变量 x_1 是否导致 y，以及此因果关系的强弱。假定 $\{x_1, x_2, \cdots, x_K\}$ 包含了所有可能影响 y 的因素。不同学科可能依条件的不同而采用不同的实验方法，大致可以分为以下几类。

(1) 控制实验(controlled experiment)。比如，在理想的物理实验中，对除 x_1 以外的因素 $\{x_2, \cdots, x_K\}$ 全部控制不变，单独让 x_1 变化，然后观察 y 变化的情况。

(2) 随机(控制)实验(randomized controlled experiment)。比如，医学上对新药 x_1 疗效的实验。由于参加实验者的体质与生活方式不同，不可能完全控制所有其他的因素 $\{x_2, \cdots, x_K\}$ (即使用老鼠做实验，老鼠之间仍然有差异)，故无法进行严格的控制实验。为此，当代统计学之父费舍尔(Ronald Fischer)提出了随机实验的概念。通常将实验人群(或个体)随机地分为两组，其中“实验组”或“处理组”(treatment group)服用真药，而“控制组”(control group，也称“对照组”)服用“安慰药”(placebo)[①]。所有被试者均不知道自己被分在哪一组，以避免心理作用的干扰。有时候，科研人员也不知道被试者究竟在哪一组。这被称为“双盲法”(double blind)。又比如，在农学中将地块随机地分成三组(因为很难找到土壤条件完全一样的地块)，分别给予不同的施肥量，然后考察施肥的效果。在经济学中，“实验经济学”(experimental economics)所做的实验基本上属于随机实验。

(3) 自然实验或准实验(natural experiment or quasi experiment)。由于某些并非为了实验目的而发生的外部突发事件，使得当事人仿佛被随机地分在了实验组或控制组。比如，一个州通过了一个法律，但相邻的另一个州没有通过这一法律。两个州的民众事先并不知道哪个州会通过这个法律，因此无法进行自我选择住在哪个州。这样，从考察这个法律的效果而言，可以近似地认为民众是随机地选择住在哪个州的，或者随机地被分在了哪个组，即实验组(通过法律)与控制组(没通过法律)。这种情形被称为“自然实验”或“准实验”。

(4) 思想实验(thought experiment)。比如，弗里德曼(Milton Friedman)曾设想在一个与世隔绝的小岛上通过直升机投放货币，然后考察该岛的宏观经济会发生怎样的变化。所谓“思想实验”，其实并无实验数据，只是在理想情形下的逻辑推理。

---

① 比如，没有实际作用的糖浆。

18.2 理想的随机实验

在理想的随机实验(ideal randomized experiment)中,实验组与控制组成员的决定完全是随机的,比如,通过抛硬币或计算机随机数来决定。因此,个体究竟被分在哪一组或得到多大的实验"处理水平"(treatment level),与个体的特征或其他可能影响实验结果的因素是完全独立的。因此,解释变量"处理水平"与被遗漏的扰动项是不相关的。这样,就避免了遗漏变量偏差(omitted variable bias)或内生变量偏差(endogeneity bias)。这是随机实验的最大优点。

考虑以下回归模型:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{18.1}$$

其中,x_i 是完全随机地决定的。由于 x_i 与 ε_i 相互独立,故 $\mathrm{Cov}(x_i, \varepsilon_i) = 0$,因此无论遗漏了多少解释变量,OLS 都是一致的。更进一步,由于 x_i 与 ε_i 相互独立,故 $\mathrm{E}(\varepsilon_i \mid x_1, \cdots, x_n) = 0$,也满足小样本理论中的严格外生性假定。因此,OLS 也是无偏的。

在随机实验中,因果效应(causal effect)有着清晰的定义。在理想的随机实验中,X 对 y 的因果效应表现在条件期望的差别,即 $\mathrm{E}(y \mid X = x) - \mathrm{E}(y \mid X = 0)$,也称为"处理效应"(treatment effect)。具体来看两个例子。

例 x_i 为服用药量,比如,$x_i = \{0, 1, 2\}$,而 y_i 为病情康复情况。

例 $x_i = \{0, 1\}$ 表示是否参加过某一就业培训项目(job training program),y_i 为未来的就业状态,而扰动项 ε_i 包含遗漏变量,比如个人能力、工作经验等。由于 x_i 与 ε_i 相互独立,故参加过培训的实验组成员的能力及工作经验等的概率分布与没有参加过的控制组成员是一样的。

如果 $x_i = \{0, 1\}$ 为虚拟变量,则方程(18.1)的 OLS 估计量为

$$\hat{\beta}_{\mathrm{OLS}} = \bar{y}_{\mathrm{treat}} - \bar{y}_{\mathrm{control}} \tag{18.2}$$

其中,$\bar{y}_{\mathrm{treat}}$ 为实验组的样本均值,而 $\bar{y}_{\mathrm{control}}$ 为控制组的样本均值。这是因为,在回归方程中加入虚拟变量的效果就相当于给予实验组与控制组不同的截距项。另一方面,当 $\{y_i\}$ 对常数项进行回归时,其系数估计值便是 $\bar{y}$。因此,$\hat{\alpha}_{\mathrm{OLS}} = \bar{y}_{\mathrm{control}}$,而 $\hat{\alpha}_{\mathrm{OLS}} + \hat{\beta}_{\mathrm{OLS}} = \bar{y}_{\mathrm{treat}}$。由于 $\hat{\beta}_{\mathrm{OLS}}$ 等于实验组均值与控制组均值之差,故也被称为"差分估计量"(differences estimator)。

例 班级规模是否影响学习成绩?班级规模过大(师生比过低)可能影响任课老师给每位学生的注意力,从而影响学习效果。由于遗漏变量的存在,使用观测数据很难回答此问题。比如,规模较小的班级可能位于好学区,师资好,家庭也富有。为此,美国田纳西州进行了为期四年的随机实验(称为 Project STAR,即 Student-Teacher Achievement Ratio),将幼儿园至小学三年级的学生随机分为三组。第一组为普通班,每班 22 ~ 25 名学生;第二组为小班,每班 13 ~ 17 名学生;第三组也为小班,但配备一名教学助理(teacher's aide)。教师也随机分到这三类班级。实验结果发现,尽管班级规模对学习成绩的影响在统计上显著,但在经济上并不显著(即此效应本身比较小,普通班与小班的成绩差距类似于男生与女生的成绩差距)。有关此随机实验的更多信息,参见 Stock and Watson (2011)第 13 章。

18.3 引入更多的解释变量

虽然在理想的随机实验条件下,OLS 估计量(即差分估计量)是一致且无偏的,但由于遗漏

了较多变量，故 ε_i 的方差可能较大，OLS 估计可能效率不高。如果引入某些遗漏变量，则可以改善这个问题。另外，引入其他遗漏变量也提供了一个检验 x_i 是否完全随机的机会。假设引入的其他解释变量为 $\{z_{i1},\cdots,z_{iK}\}$，

$$y_i=\alpha+\beta x_i+\delta_1 z_{i1}+\cdots+\delta_K z_{iK}+\varepsilon_i \tag{18.3}$$

如果 x_i 完全随机，则 $\{z_{i1},\cdots,z_{iK}\}$ 应该对 x_i 没有解释力。因此，我们可以把 x_i 对 $\{z_{i1},\cdots,z_{iK}\}$ 进行回归，并检验该回归方程的整体显著性。如果该方程显著，则认为 x_i 不完全随机。

例 家中有计算机能否改善教育效果(educational outcome)？如果使用观测数据，会遇到遗漏变量偏差问题，比如在教育方面有追求的学生与家庭通常也积极地在家中添置计算机。为此，Fairlie and London (2012)在美国北加州的一个社区学院(community college)[①]的 286 名新生中进行了随机实验。这些学生被随机分成大致相等的两组，其中一组(实验组)得到免费计算机，而另一组为控制组；并进行问卷跟踪调查两年，同时从学院行政部门获取相关信息。该研究的被解释变量包括对教育效果的四种度量(课程通过率、课程中学分课的比例[②]、课程中可转学分课的比例、获得学位的比例)。其他解释变量(即上文的 $\{z_{i1},\cdots,z_{iK}\}$)包括性别、种族、父母教育水平、高中成绩、是否自己有子女、是否与父母同住、家庭收入、是否工作、英语是否为母语等。Fairlie and London (2012)同时汇报了不包含其他解释变量的差分估计量(在文中称为 treatment-control difference)，以及包括其他解释变量的差分估计量(在文中称为 regression-adjusted treatment-control difference)。结果发现，获赠计算机实验组的两种教育效果变量显著高于控制组，而其他两种教育效果变量的效应也为正(但不显著)。

18.4 随机实验执行过程中可能出现的问题

在随机实验的执行过程中，可能并不完全符合“理想的随机实验”的要求。可能出现的问题包括，“内部有效性问题”(internal validity)与“外部有效性问题”(external validity)。

1. 内部有效性问题

(1) 未能完全随机分组(failure to randomize)。比如，以姓氏字母在字母表的前半部或后半部来决定谁进入就业培训项目。然而，在美国，姓氏与种族有关，而种族又与就业机会有关。因此，这种分组其实不完全随机。此时，可把 x_i 对可观测的个体特征 $\{z_{i1},\cdots,z_{iK}\}$ 进行回归，并检验该回归方程的整体显著性。如果该方程显著，则认为 x_i 不完全随机。

(2) 未能完全遵从实验设计(partial compliance)。比如，被指定参加就业培训者却因故没来，而未被指定参加者却自行来参加。即使研究者知道实际上谁参加了，但由于是否参加的决定部分地由个人做出，也违反了 x_i 的随机性假定(x_i 可能与不可观测的个体特性 ε_i 相关)，导致不一致的估计。此时可使用工具变量法来解决 x_i 的内生性问题。具体来说，以“设计处理水平”(assigned treatment level，记为 w_i)作为“实际处理水平”(actual treatment level，记为 x_i)的工具变量。显然，只要有部分人遵循实验设计，则 w_i 就与 x_i 相关，满足相关性；另一方面，w_i 完全由随机数确定，肯定与扰动项 ε_i 不相关，满足外生性；故 w_i 为有效工具变量。

使用工具变量法估计处理效应(instrumental variables estimation of the treatment effect)的前提

① 美国的社区学院通常为两年制，相当于中国的中专，其学生多来自低收入家庭。

② 能获得学分的课程称为“学分课”。有些课程不能获得学分。

是,研究者既知道设计处理水平,也知道实际处理水平。但研究者有时可能不知道实际处理水平,比如参加实验组者是否真吃了药。此时,对实际处理水平 x_i 的测量误差仍会导致不一致估计。

(3) 中途退出实验(attrition)。比如,参加就业培训的优秀者可能在项目进行过程中就找到工作而退出项目。那么,该项目只剩下较不优秀者,造成选择性偏差。但如果中途退出的原因与实验无关(比如,必须退出项目去照顾生病的家人),则不会造成选择性偏差。

(4) 实验效应或霍桑效应(experimental effect 或 Hawthorne effect)。参加实验本身可能改变个体的心理或行为,从而影响实验结果。对于药物疗效实验,可以通过使用双盲法来避免这种效应。但对于就业培训,双盲法并不可行,因为个体一定知道他是否参加了就业培训。这种效应也称为"霍桑效应"。20 世纪二三十年代,美国通用汽车公司(General Motors)在霍桑(地名)的工厂做了一些实验,研究灯光、休息时间、车间布置、工时等因素对劳动生产率的影响。结果发现,无论这些条件如何改变,工人的劳动生产率始终在上升!其原因可能是,工人在实验过程中感到被关注而更加努力地工作。

(5) 样本过小。由于实验成本高,实验的样本容量可能较小。由于个体间差异可能较大,导致估计量的精确度不足。

2. 外部有效性问题

(1) 样本的代表性不足(non-representative sample)。比如,让即将获释的囚犯参与就业培训,其结论可能难以推广到普通人群,因为雇主可能歧视有犯罪记录的人。又比如,项目参加者为自愿报名的义工,而义工的素质通常高于普通人。

(2) 小型实验的条件与大规模推广时的现实条件不同。小型实验的条件与大规模推广时的现实条件可能很不同,而这种差异可能导致项目的结果截然不同。比如,英国《经济学家》杂志(*Economist*)在 2008 年 6 月 12 日报导了经济学家在非洲肯尼亚进行的实验,旨在推广蚊帐以防止疟疾。一种观点认为,免费发放蚊帐能够最快地推广蚊帐。另一种观点则认为,由于不知道蚊帐的价值,人们不会珍惜免费发放的蚊帐,而且免费发放使得蚊帐的长期供给变得困难。经济学家于是随机地对一些村庄提供免费蚊帐,而对另外一些村庄则有偿提供。实验结果证明,免费提供蚊帐更为有效。然而,这个实验结果如果大规模推广仍然有效吗?反驳者指出,在肯尼亚的那个实验地区,蚊帐的价值已广为人知;在实验过程中,蚊帐的供给也是有保证的;而这两点在大规模推广时不见得成立。

(3) 一般均衡效应(general equilibrium effect)。一个小型的就业培训项目可能不会改变社会上雇主的行为。但大面积推广后,雇主可能减少企业自行提供的员工培训,使得该项目给整个社会带来的净福利减少。

(4) 自我选择效应。在随机实验中,可以随机地安排实验组与控制组的成员。但在现实社会中,人们常常可以自由选择是否参与。这样,就会有自我选择问题。比如,在随机实验中,数据可能显示该项目并不有效。但在现实社会中,预期收益最大的人将最有积极性参加此项目。因此,这个项目可能在实际上很有效。

18.5 自然实验

随机实验虽然说服力强,但通常成本较高。另一方面,自然实验为自然发生(非为实验目的

而发生),几乎没有成本,而说服力类似于随机实验,故为实证研究者所青睐;但发现自然实验却需要一双慧眼。自然实验可分为两类。在第一类自然实验中,个体的分组或处理水平完全由自然实验所决定,可直接用 OLS 估计因果效应。在第二类自然实验中,个体的分组或处理水平只是部分地由自然实验所决定,此时应以自然实验带来的随机变动作为工具变量进行估计。不完美的自然实验也可能存在与不完美的随机实验类似的问题(参见上节),只是前者不可能存在实验效应,因为自然实验中的个体不会视自己为实验对象。

例 最低工资对就业的影响。提高法定最低工资(minimum wage)在多大程度上会影响对低技能工人的需求?由于工资与雇佣人数由劳动力市场的供求均衡所决定,故如果使用 OLS 估计劳动力需求函数,将导致联立方程偏差或内生变量偏差。设想这样一个随机实验:将参加实验的雇主分为两组,要求实验组的雇主提高最低工资,而控制组的雇主保持最低工资不变,然后比较这两组的劳动力需求变化。然而,这种实验很难执行,因为经济学家并没有立法的权力或实验成本太高。为此,Card and Krueger (1994)考虑了一个自然实验。在 1992 年,美国新泽西州通过法律将最低工资从每小时 $4.25 提高到 $5.05,但在相邻的宾夕法尼亚州最低工资却保持不变。在这种情况下,这两个州的雇主仿佛被随机地分配到实验组(新泽西州)与控制组(宾夕法尼亚州)。Card and Kruger (1994)收集了两个州的快餐店在实施新法前后雇佣人数的数据,并使用双重差分法(参见下节)进行估计。结果发现,提高最低工资对低技能工人的就业几乎没有影响。也就是说,对低技能工人的需求是没有弹性的,并不会因工资的上升而显著地减少雇佣人数。本例属于第一类自然实验,个体分组完全由自然实验所决定。

例 服兵役是否影响退役后的长期收入?如果使用观测数据比较服过兵役的人群与未服兵役人群的收入,则会产生选择偏差,因为能否参军取决于身体条件,以及个人选择(或许就业前景不佳者才选择参军)。为此,Angrist (1990)考察了越战期间的参军者,当时美国对全国年轻男子以生日抽签的方式进行征兵(draft)。尽管抽签结果完全随机,但是否参军还取决于体检结果,而且有些人得到豁免,另一些人虽未抽中却自愿参军。因此,本例属于第二类自然实验,即个体分组仅部分地由自然实验决定,故应使用抽签结果作为参军与否的工具变量进行估计。估计结果表明,服兵役会减少白人的长期收入,但不影响非白人的长期收入。

例 经济增长对内战的影响。显然,内战多发生于穷国。是否经济增长慢导致内战概率上升?另一方面,内战也会拖累经济增长,故存在内生性。Miguel et al (2004)使用外生的降雨量变化(rainfall variation)作为自然实验,考察 41 个非洲国家在 1981—1999 年间,经济增长对内战概率的因果作用。由于灌溉系统与工业的缺乏,这些非洲国家的经济比较依赖于自然降雨的农业(rain-fed agriculture),故经济增长与降雨量有关,但并不完全取决于降雨量;故属于第二类自然实验。因此,Miguel et al (2004)使用降雨量变化作为经济增长的工具变量,发现经济低增长显著地增加内战概率。这种利用大自然的外生变化而进行的自然实验被称为"大自然的自然实验"(natural natural experiments)[①],在外生变化的随机性方面一般优于由于社会因素所引起的自然实验。大自然的自然实验的随机变化主要来自天气、自然灾害以及生物机制(比如,双胞胎、出生日期、性别)等。需要注意的是,尽管天气变化完全由大自然随机决定,也不一定在计量经济学意义上为外生,即与模型的扰动项不相关。这是因为,虽然经济系统内的变量一般不会影响天气,但天气却会影响经济系统内的变量(人们会根据天气而调整其行为)。因此,仍需要探讨工具变量是否满足排他性约束。为此,Miguel et al (2004)试图排除降雨量影响内战的其他渠道

① 参见 Rosenzweig and Wolpin (2000)。

(除了人均 GDP 这一渠道),比如降雨变化可能影响收入分配、农村贫困率、政府税收、道路情况(水灾可破坏道路)、民众情绪(过分干燥使人易怒好斗),而这些因素在理论上都有可能直接影响内战的概率(不通过人均 GDP 这一渠道)。

例(一个失败的例子) 京杭大运河流经省份的人均 GDP 平均而言高于其他省份。这是否可以归功于京杭大运河对区域经济增长的促进作用?问题在于,当隋炀帝确定京杭大运河的位置时,他是在地图上随机画了一条线吗?

18.6 双重差分法

在做随机实验或自然实验时,实验的效果常常需要一段时间才能显现出来,而我们关心的恰恰是被解释变量实验前后的变化。为此,考虑以下两期面板数据:

$$y_{it}=\alpha+\gamma D_t+\beta x_{it}+u_i+\varepsilon_{it}\quad (i=1,\cdots,n;t=1,2) \tag{18.4}$$

其中,D_t 为实验期虚拟变量($D_t=1$,如果 $t=2$,实验后;$D_t=0$,如果 $t=1$,实验前),u_i 为不可观测的个体特征,而政策虚拟变量(policy dummy)

$$x_{it}=\begin{cases}1, & \text{若 } i\in\text{实验组,且 } t=2\\ 0, & \text{其他}\end{cases} \tag{18.5}$$

因此,当 $t=1$ 时(第一期),实验组与控制组并没有受到任何不同对待,x_{it} 都等于 0。当 $t=2$ 时(第二期),实验组 $x_{it}=1$,而控制组 x_{it} 依然等于 0。如果该实验未能完全地随机化(比如,观测数据),则 x_{it} 可能与被遗漏的个体特征 u_i 相关,从而导致 OLS 估计不一致。由于是面板数据,可以对方程(18.4)进行一阶差分(即第二期减去第一期),以消掉 u_i,

$$\Delta y_i=\gamma+\beta x_{i2}+\Delta\varepsilon_i \tag{18.6}$$

用 OLS 估计上式,即可得到一致估计。根据与差分估计量(differences estimator)同样的推理可知

$$\hat{\beta}_{\text{OLS}}=\Delta\bar{y}_{\text{treat}}-\Delta\bar{y}_{\text{control}}=(\bar{y}_{\text{treat},2}-\bar{y}_{\text{treat},1})-(\bar{y}_{\text{control},2}-\bar{y}_{\text{control},1}) \tag{18.7}$$

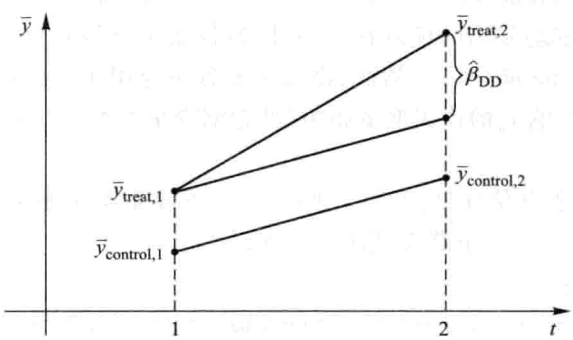


图 18.1 双重差分估计量示意图

因此,这个估计法称为"双重差分估计量"(Difference-in-Differences estimator,简记 DD),记为 $\hat{\beta}_{\text{DD}}$,即实验组的平均变化与控制组的平均变化之差,参见图 18.1。从图中可以直观地看出,双重差分估计量已经剔除了实验组与控制组"实验前差异"(pretreatment differences)的影响。

对于双重差分估计量,也可以引入其他解释变量 $\{z_{i1},\cdots,z_{iK}\}$:

$$\Delta y_i = \gamma + \beta x_{i2} + \delta_1 z_{i1} + \cdots + \delta_K z_{iK} + \Delta \varepsilon_i \tag{18.8}$$

然后用 OLS 估计上式。显然，以 Δy_i 为被解释变量的双重差分法（参见方程(18.6)与(18.8)）不适用于多期的数据。此时，需要回到以 y_{it} 为被解释变量的面板模型。为了简洁，暂时忽略其他解释变量 $\{z_{i1},\cdots,z_{iK}\}$，并仍假设为两期数据。可以证明，方程(18.6)与以下两期面板模型等价：

$$y_{it} = \beta_0 + \beta_1 G_i \cdot D_t + \beta_2 G_i + \gamma D_t + \varepsilon_{it} \quad (i = 1,\cdots,n; t = 1,2) \tag{18.9}$$

其中，G_i 为实验组虚拟变量（$G_i = 1$，如果个体 i 属于实验组；$G_i = 0$，如果个体 i 属于控制组）；D_t 为实验期虚拟变量（$D_t = 1$，如果 $t = 2$；$D_t = 0$，如果 $t = 1$），而互动项 $G_i \cdot D_t = x_{it}$（取值为 1，若 $i \in$ 实验组，且 $t = 2$；反之，取值为 0）。在方程(18.9)中，分组虚拟变量 G_i 刻画的是实验组与控制组本身的差异（即使不进行实验，也存在此差异），时间虚拟变量 D_t 刻画的是实验前后两期本身的差异（即使不进行实验，也存在此时间趋势），而互动项 $G_i \cdot D_t$ 才能真正度量实验组的政策效应。如果有其他解释变量 $\{z_{i1},\cdots,z_{iK}\}$，可直接放入方程(18.9)。

当 $t = 1$ 时，方程(18.9)可以写为

$$y_{i1} = \beta_0 + \beta_2 G_i + \varepsilon_{i1} \tag{18.10}$$

当 $t = 2$ 时，方程(18.9)可以写为

$$y_{i2} = \beta_0 + \beta_1 G_i \cdot D_2 + \beta_2 G_i + \gamma + \varepsilon_{i2} \tag{18.11}$$

将方程(18.11)减去(18.10)可得

$$\Delta y_i = \gamma + \beta_1 G_i \cdot D_2 + (\varepsilon_{i2} - \varepsilon_{i1}) = \gamma + \beta_1 x_{i2} + \Delta \varepsilon_i \tag{18.12}$$

方程(18.12)与方程(18.6)完全相同。因此，对方程(18.9)进行 OLS 估计，得到的 $\hat{\beta}_1$（即互动项 $G_i \times D_t$ 的系数）就是双重差分估计量。使用方程(18.9)面板形式的双重差分估计量的好处在于，它可以很容易推广到多期数据的情形。比如，共有 4 期数据，则可估计如下方程：

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 G_i + \gamma_1 D2_t + \gamma_2 D3_t + \gamma_4 D4_t + \varepsilon_{it} \quad (i = 1,\cdots,n; t = 1,\cdots,4) \tag{18.13}$$

其中，$D2_t,\cdots,D4_t$ 分别为对应于第 2～4 期的时间虚拟变量（以捕捉即使没有政策变化也可能存在的时间效应），而主要感兴趣的政策虚拟变量 x_{it} 定义为

$$x_{it} = \begin{cases} 1, & \text{若 } i \in \text{实验组，且 } t \in \text{实验期} \\ 0, & \text{其他} \end{cases} \tag{18.14}$$

在方程(18.13)中，x_{it} 的系数 OLS 估计值 $\hat{\beta}_1$ 就是双重差分估计量。显然，在两期数据中，x_{it} 就是交叉项 $G_i \times D_t$。双重差分法的优点在于，它同时控制了分组效应 G_i（group-specific effects）与时间效应 D_t（time-specific effects）。然而，将双重差分法应用于观测数据时，如果政策虚拟变量 x_{it} 为内生，比如，x_{it} 与影响 y_{it} 的遗漏变量相关（此遗漏变量为扰动项 ε_{it} 的一部分），则依然得不到一致估计。

在 Stata 中估计双重差分估计量，可手工进行，也可下载非官方命令 `diff`。

```
ssc install diff     (下载安装命令 diff)
```

该命令的基本格式为

```
diff y,treat(varname) period(varname) cov(z1 z2) robust report test
```

其中，“y”为结果变量（outcome variable，即上文的 y_{it}），必选项“treat(varname)”用来指定处理变量（即上文的 G_i），必选项“period(varname)”用来指定实验期虚拟变量（即上文的 D_t）。选择项“robust”表示汇报稳健标准误。选择项“cov(z1 z2)”用来指定其他解释变量（即上文的 $z_{i1},\cdots,z_{iK}$），其中“cov”表示协变量（covariate）；选择项“report”表示汇报对协变量系数的估计结果。选择项“test”表示检验在基期时，各变量在实验组与控制组的均值是否相等。

下面以数据集 cardkrueger1994.dta 为例。该数据集来自 Card and Kruger(1994)的自然实验

(参见上节)。被解释变量为 fte (full time employment 人数),实验组虚拟变量为 treated (treated = 1,如果快餐店在新泽西州;treated = 0,如果快餐店在宾夕法尼亚州),实验期虚拟变量为 t (t = 1,如果时间为 1992 年 11 月;t = 0,如果时间为 1992 年 2 月)。其他解释变量包括以下快餐品牌虚拟变量:bk (Burger King = 1),kfc (Kentucky Fried Chicken = 1),roys (Roy Rogers = 1),wendys (Wendy's = 1)。首先,手工计算双重差分法。

```
. use cardkrueger1994,clear
. gen gd = t* treated    (定义交叉项 gd)
. reg fte gd treated t,r
```

```
Linear regression                               Number of obs =     801
                                                F(  3,   797) =    1.43
                                                Prob > F      =  0.2330
                                                R-squared     =  0.0080
                                                Root MSE      =   9.003

------------------------------------------------------------------------------
             |               Robust
         fte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          gd |    2.913982   1.736818     1.68   0.094    -.4952963    6.323261
     treated |   -2.883534   1.403338    -2.05   0.040    -5.638209   -.1288592
           t |    -2.40651   1.594091    -1.51   0.132    -5.535623    .7226031
       _cons |    19.94872   1.317281    15.14   0.000     17.36297    22.53447
------------------------------------------------------------------------------
```

从上表可知,处理效应(即互动项 gd)的系数为 2.914,但仅在 10% 水平上显著(p 值为 0.094)。其次,使用命令 diff 来估计。

```
. diff fte,t(treated) p(t) robust
```

```
Number of observations in the DIFF-IN-DIFF: 801
           Baseline       Follow-up
   Control: 78             77          155
   Treated: 326            320         646
            404            397

R-square:  0.00805

                        DIFFERENCE IN DIFFERENCES ESTIMATION
------------------------ ------------ BASE LINE ------------ ------------ FOLLOW UP ------------
> -- ----------------
 Outcome Variable      | Control  |  Treated  | Diff(BL)  | Control  |  Treated  | Diff(FU
> ) | DIFF-IN-DIFF
-----------------------+----------+-----------+-----------+----------+-----------+-----------
> -+----------------
fte                    | 19.949   | 17.065    | -2.884    | 17.542   | 17.573    | 0.030
>    | 2.914
Std. Error             | 1.317    | 0.484     | 1.403     | 0.898    | 0.491     | 1.023
>    | 1.737
t                      | 15.14    | 13.99     | -2.05     | 17.27    | 20.59     | -0.04
>    | 1.68
P>|t|                 | 0.000    | 0.000     | 0.040**   | 0.000    | 0.000     | 0.976
>    | 0.094*
-----------------------------------------------------------------------------------------------
> ----------------
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

上表第一列显示，处理效应的系数估计值为 2.914，而 p 值为 0.094，这与手工回归的结果完全相同。上表将“$t=0$”的时期称为“Baseline”（基期），而将“$t=1$”的时期称为“Follow-up”（跟踪期）。再次，引入快餐品牌虚拟变量，再进行手工回归（以 wendys 作为参照虚拟变量）。

```
. reg fte gd treated t bk kfc roys,r
```

```
Linear regression                               Number of obs =     801
                                                F(  6,   794) =   57.30
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1878
                                                Root MSE      =  8.1617
```

fte	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gd	2.93502	1.543422	1.90	0.058	-.0946504	5.96469
treated	-2.323906	1.253701	-1.85	0.064	-4.784867	.1370549
t	-2.402678	1.410265	-1.70	0.089	-5.170966	.3656108
bk	.9168795	.9382545	0.98	0.329	-.9248729	2.758632
kfc	-9.204856	.8991089	-10.24	0.000	-10.96977	-7.439945
roys	-.8970458	1.041071	-0.86	0.389	-2.940623	1.146532
_cons	21.16069	1.307146	16.19	0.000	18.59482	23.72656

从上表可知，引入其他解释变量（即快餐品牌虚拟变量）后，处理效应的系数估计值增大为 2.935，并接近 5% 的显著性水平（p 值为 0.058）。下面再次使用命令 diff。

```
. diff fte,t(treated) p(t) cov(bk kfc roys) robust
```

```
DIFFERENCE-IN-DIFFERENCES WITH COVARIATES


Number of observations in the DIFF-IN-DIFF: 801
            Baseline        Follow-up
   Control: 78              77          155
   Treated: 326             320         646
            404             397

R-square:  0.18784

                              DIFFERENCE IN DIFFERENCES ESTIMATION
                    ---------- BASE LINE ----------  ---------- FOLLOW UP ----------
> --------------
 Outcome Variable   | Control | Treated | Diff(BL) | Control | Treated | Diff(FU
> ) | DIFF-IN-DIFF
--------------------+---------+---------+----------+---------+---------+---------
> ----------------
fte                 | 21.161  | 18.837  | -2.324   | 18.758  | 19.369  | 0.611
>   | 2.935
Std. Error          | 1.307   | 0.886   | 1.254    | 1.042   | 0.908   | 0.900
>   | 1.543
t                   | 16.19   | 18.54   | -1.85    | 18.85   | 19.67   | 0.94
>   | 1.90
P>|t|               | 0.000   | 0.000   | 0.064*   | 0.000   | 0.000   | 0.497
>   | 0.058*
------------------------------------------------------------------------------
> ----------------
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

上表第一列显示的处理效应估计结果与手工回归完全一样。最后，使用命令 diff 检验各变量在实验组与控制组的均值是否相等。

```
. diff fte,t(treated) p(t) cov(bk kfc roys wendys) test
```

```
TWO-SAMPLE T TEST

Number of observations (baseline): 404
            Baseline        Follow-up
   Control: 78              -               78
   Treated: 326             -               326
            404             -

t-test at period = 0:

> ----------
 Variable(s)          |   Mean Control   |  Mean Treated  |    Diff.    |   |t|   |   Pr(|
> T|>|t|)
----------------------+------------------+----------------+-------------+---------+-------
> ----------
fte                    |  19.949          |  17.065        |  -2.884     |  2.44   |  0.015
> 0**
bk                     |  0.443           |  0.411         |  -0.032     |  0.52   |  0.603
> 5
kfc                    |  0.152           |  0.205         |  0.054      |  1.08   |  0.281
> 8
roys                   |  0.215           |  0.248         |  0.033      |  0.61   |  0.544
> 8
wendys                 |  0.190           |  0.136         |  -0.054     |  1.22   |  0.224
> 1
-----------------------------------------------------------------------------------------

> ----------
*** p<0.01; ** p<0.05; * p<0.1
```

从上表可知，在基期，结果变量 fte 的均值在两组之间有显著差异，而各品牌虚拟变量的均值则在两组之间无显著差异。有关命令 diff 的更多功能（比如双重差分倾向得分匹配估计量），将在第 28 章介绍。

18.7 三重差分法

双重差分法的隐含假设是，即使没有政策变化，控制组与实验组的时间趋势也一样；在方程（18.9）中表现为共同的时间趋势项 γD_t。然而，如果控制组与实验组的时间趋势不同，比如前者系数为 γ_0，而后者系数为 $\gamma_0+\gamma_1$，则方程（18.9）应写为：

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 G_i \cdot D_t + \beta_2 G_i + \gamma_0 D_t + \gamma_1 G_i \cdot D_t + \varepsilon_{it} \\ &= \beta_0 + (\beta_1 + \gamma_1) G_i \cdot D_t + \beta_2 G_i + \gamma_0 D_t + \varepsilon_{it} \end{aligned} \tag{18.15}$$

因此，用 OLS 估计上式只能得到对 $(\beta_1+\gamma_1)$ 的一致估计；无法得到对实验效应 β_1 的一致估计。此时，须进一步改进双重差分估计量。

来看一个具体的例子。假设美国 B 州针对 65 岁或以上老年人（记为 E，表示 elderly）引入了一项新的医疗保健政策，而该政策不适用于 65 岁以下人群（记为 N）。欲考察此政策对健康状况 y 的影响。一种做法是以 B 州 65 岁或以上人群作为实验组，而以该州 65 岁以下人群作为控制

组，然后进行双重差分估计。此种 DD 法的潜在缺陷是，年轻人相对于老年人的相对健康状况本身可能随时间发生变化（即使未推出此项新政策），比如由于联邦政府同时期的其他政策变化所引起。因此，另一种做法是以相邻 A 州的 65 岁或以上老年人作为控制组，然后进行双重差分估计。然而，第二种 DD 法同样有缺陷，因为它忽略了 A 州与 B 州老年人的健康状况，即使在没有新政策的情况下也可能有不同的时间趋势，比如两个州的州情不同。更稳健的做法则将以上两种 DD 法结合起来，估计如下方程（为了简洁，省略个体下标 i 与时间下标 t）：

$$y=\beta_0+\beta_1 B+\beta_2 E+\beta_3 B\cdot E+\gamma_0 D+\gamma_1 D\cdot B+\gamma_2 D\cdot E+\delta D\cdot B\cdot E+\varepsilon_{it} \tag{18.16}$$

其中，B 为 B 州虚拟变量（B 州 =1，A 州 =0），E 为老年虚拟变量（老年 =1，其他年龄 =0），D 为时间虚拟变量（第 2 期 =1，第 1 期 =0）。互动项 $D\cdot B\cdot E$（表示第 2 期 B 州老年人）的系数 δ 即为政策效应。OLS 估计量 $\hat{\delta}$ 可以表达为

$$\hat{\delta}=[(\bar{y}_{BE2}-\bar{y}_{BE1})-(\bar{y}_{BN2}-\bar{y}_{BN1})]-[(\bar{y}_{AE2}-\bar{y}_{AE1})-(\bar{y}_{AN2}-\bar{y}_{AN1})] \tag{18.17}$$

其中，$\bar{y}_{BE2}$ 表示第 2 期 B 州老年人的样本均值，$\bar{y}_{BE1}$ 表示第 1 期 B 州老年人的样本均值，以此类推。此估计量称为“三重差分估计量”（Difference-in-Difference-in-Differences estimator，简记 DDD），记为 $\hat{\delta}_{DDD}$。上式右边的第一个方括弧就是以 B 州年轻人为控制组的 DD 估计量，但此 DD 估计量没有控制实验组与控制组（在未引入新政策情况下）的时间趋势差异（trend differential）；上式右边第二个方括弧通过 A 州来捕捉时间趋势差异；如果 A 州与 B 州的时间趋势差异相似，则 $\hat{\delta}_{DDD}$ 就是对政策效应的可信估计。

下面对表达式（18.17）进行说明。根据方程（18.16），对应于 $\bar{y}_{BE2}$ 的总体期望为

$$E(y_{BE2})=\beta_0+\beta_1+\beta_2+\beta_3+\gamma_0+\gamma_1+\gamma_2+\delta \tag{18.18}$$

类似地，对应于 $\bar{y}_{BE1}$ 的总体期望为

$$E(y_{BE1})=\beta_0+\beta_1+\beta_2+\beta_3 \tag{18.19}$$

将方程（18.19）减去（18.18）可得对应于 $(\bar{y}_{BE2}-\bar{y}_{BE1})$ 的总体期望

$$E(y_{BE2})-E(y_{BE1})=\gamma_0+\gamma_1+\gamma_2+\delta \tag{18.20}$$

根据同样推导可得，

$$E(y_{BN2})-E(y_{BN1})=\gamma_0+\gamma_1 \tag{18.21}$$

$$E(y_{AE2})-E(y_{AE1})=\gamma_0+\gamma_2 \tag{18.22}$$

$$E(y_{AN2})-E(y_{AN1})=\gamma_0 \tag{18.23}$$

综合方程（18.20）、（18.21）、（18.22）与（18.23）可知，对应于方程（18.17）的总体期望为

$$\begin{aligned}&\{[E(y_{BE2})-E(y_{BE1})]-[E(y_{BN2})-E(y_{BN1})]\}-\\&\{[E(y_{AE2})-E(y_{AE1})]-[E(y_{AN2})-E(y_{AN1})]\}\\&=[(\gamma_0+\gamma_1+\gamma_2+\delta)-(\gamma_0+\gamma_1)]-[(\gamma_0+\gamma_2)-(\gamma_0)]=[\gamma_2+\delta]-[\gamma_2]=\delta\end{aligned} \tag{18.24}$$

当然，在方程（18.16）中，也可加入其他解释变量 $\{z_{i1},\cdots,z_{iK}\}$，或推广到多期数据。

18.8 观测数据的处理效应

在许多情况下，并没有随机实验或自然实验的数据，而只有观测数据（observational data）。比如，政府提供的就业培训项目，完全由个人决定是否参与。此时，由于实际处理水平 x_i 存在自我选择（self selection），并非随机分组，故可能导致不一致的估计。因此，对于观测数据，需使用特别的计量方法来估计处理效应，详见第 28 章。

习　题

18.1 （选自 Stock and Watson，2011）考虑一项旨在评估在宿舍里安装互联网对大学生成绩的影响的研究。在一个大宿舍里，随机地选择半个宿舍安装宽带互联网（实验组），并收集所有宿舍成员的期末成绩。以下的哪种情况会对该随机实验的“内部有效性”产生威胁？（回答“会”或“不会”）

（1）在学年中间，所有男性运动员都搬出了大学宿舍，住进了兄弟会（fraternity），因此退出了此项研究（观测不到他们的期末成绩）；

（2）被分到控制组的工程专业学生自行安装了一个局部网，分享一个私人无线互联网，并共同付费；

（3）实验组中的艺术专业学生一直没有学会如何使用互联网账户；

（4）实验组中的经济专业学生为控制组成员提供互联网服务，并收费。

第 19 章　蒙特卡罗法与自助法

19.1　蒙特卡罗法的思想与用途

通过计算机模拟从总体抽取大量随机样本的计算方法统称为“蒙特卡罗法”(Monte Carlo Methods,简记 MC)[①]。先来看一个例子。

为了计算圆周率 π,考虑下面的随机实验。在一个边长为 1 的正方形中内接一个 1/4 单位圆,参见图 19.1。显然,正方形的面积为 1,而 1/4 圆的面积为 $\pi/4$。因此,如果知道 1/4 单位圆面积占正方形面积的比例,就可以计算出 π。

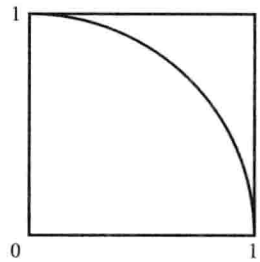


图 19.1　计算圆周率 π 的随机实验

向这个正方形随机地射箭,并假设其落点在该正方形上服从二维均匀分布(通过计算机模拟来实现)。重复实验 n 次(n 很大),其中有 m 次落在 1/4 圆内。根据大数定律,$m/n \xrightarrow{p} \pi/4$,故 $\pi \approx 4m/n$。

这个例子揭示了蒙特卡罗法的精神,但在蒙特卡罗法的具体应用上,由于目的不同可以有很多变化。在计量经济学中,常使用蒙特卡罗法来确定统计量的小样本性质。我们知道,许多统计量的精确分布没有解析解。解决方法之一是,使用大样本理论,用渐近分布来近似真实分布。然而,使用大样本理论的代价是要求样本容量较大[②],而现实中的样本容量可能不够[③]。另外,有时统计量的分布即使在大样本的假设下,也推导不出来。此时,蒙特卡罗法的价值就更加凸显,比如第 21 章中的单位根检验。

例如,对于线性回归模型,$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i\ (i = 1, \cdots, n)$,希望对线性假设“$H_0$: $\boldsymbol{R\beta} = \boldsymbol{r}$”进行显著性水平为 5% 的检验。如果不假定扰动项服从正态分布,则只能用 χ^2 分布对此假设进行大样本检验,即检验统计量 $W \equiv n\,(\boldsymbol{R\hat{\beta}} - \boldsymbol{r})'[\boldsymbol{R}\,\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}']^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{r}) \xrightarrow{d} \chi^2(m)$,其中 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的 OLS 估计量,m 为线性约束的个数。由于是大样本检验,故所使用的 χ^2 分布其实只是统计量真实分布的近似(approximation)。因此,“5%”可能只是“名义显著性水平”(nominal size),而不是“真实显著性水平”(true or actual size),二者之差被称为“显著性水平扭曲”(size distortion)[④]。

此时,可以使用蒙特卡罗法来确定“真实显著性水平”。基本步骤如下。

第一步,给定参数 $\boldsymbol{\beta}$ 的具体取值,以及解释变量 $\boldsymbol{x}$ 与扰动项 ε 的概率分布。

第二步,从 $\boldsymbol{x}$ 与 ε 的分布中随机抽样,得到 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 与 $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$。

---

① 这个名字来源于摩纳哥(在法国附近)的蒙特卡罗赌场(Monte Carlo Casino),这是最早使用这个方法的一位美国物理学家的叔叔常去的赌场。

② 理论上而言,只有当样本容量无穷大时,渐近分布才绝对地等于真实分布。

③ 比如,中国的宏观经济年度数据,常常只有 1978 年改革开放以来才有,故只有三十几个数据。

④ 也就是说,对第 I 类错误(Type I error)发生概率的衡量是不准确的。

第三步,根据方程"$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$"计算$\{y_1, y_2, \cdots, y_n\}$。

第四步,对这个样本进行 OLS 估计,计算检验统计量 W,并与$\chi^2(m)5\%$的临界值比较,确定是否拒绝原假设"H_0: $\boldsymbol{R\beta} = \boldsymbol{r}$"。

第五步,大量重复第二至第四步,得到 M 个随机样本(比如,$M = 1\ 000$),进行 M 次检验,其中拒绝原假设的比例就是真实的显著性水平。

实现蒙特卡罗法的一个前提是,必须能够通过计算机产生随机数,并从任意概率分布进行随机抽样,参见本章附录。本章内容主要来自 Cameron and Trivedi (2005, 2010)。

19.2 蒙特卡罗法实例:模拟中心极限定理

根据中心极限定理(CLT),当样本容量 $n \to \infty$时,样本均值的分布收敛于正态分布。假设 x 服从在(0,1)上的均匀分布,样本容量为 30,我们希望用蒙特卡罗法直观地"看到"样本均值 $\bar{x}_{30}$ 的分布,并与正态分布相比较。为此,从(0,1)上的均匀分布抽取 10 000 个样本容量为 30 的随机样本,得到 10 000 个样本均值 $\bar{x}_{30}$ 的观测值,然后画其直方图。

相应的 Stata 程序如下[①]。基本步骤为,用命令 program 定义一个叫"onesample"的程序,从均匀分布抽取一个随机样本,并计算 $\bar{x}_{30}$;然后用命令 simulate 重复此程序 10 000 次,得到 10 000个 $\bar{x}_{30}$ 的观测值;最后用命令 histogram 画 $\bar{x}_{30}$ 的直方图。

```
program onesample, rclass                    (定义程序 onesample,并以 r()形式储存结果)
drop _all                                    (删去内存中已有数据)
set obs 30                                   (确定随机抽样的样本容量为 30)
gen x = runiform()                           (得到在(0,1)上均匀分布的一个随机样本)
sum x                                        (使用命令"sum"得到样本均值)
return scalar mean_sample = r(mean)          (将样本均值记为 r(mean_sample))
end                                          (程序"onesample"结束)
simulate xbar = r(mean_sample), seed(101) reps(10000): onesample
```

其中,选择项"reps(10000)"表示,命令 simulate 将运行"onesample"程序 10 000 遍,并生成变量 xbar 来记录这 10 000 个样本均值。其中,选择项"seed(101)"用来确定随机数的初始值,以便再次模拟或别人运行此程序时,也能得到完全一样的结果,参见本章附录。

```
hist xbar, normal xtitle("xbar from many sample")
```

其中,选择项"normal"表示画出相应的正态分布,选择项"xtitle"表示横轴标题(如图 19.2)。

从上图的结果来看,$\bar{x}_{30}$(xbar)的直方图与相应的正态分布十分接近,从而验证了中心极限定理适用于均匀分布的随机变量。从这个例子也可以看出蒙特卡罗法的一个局限,即仅仅这个蒙特卡罗模拟本身并不能说明中心极限定理也适用于具有其他分布的随机变量。

---

① 由于 Stata 版本不同,如果发现本章少数命令无法运行,可输入命令"update all",更新 Stata 命令库。

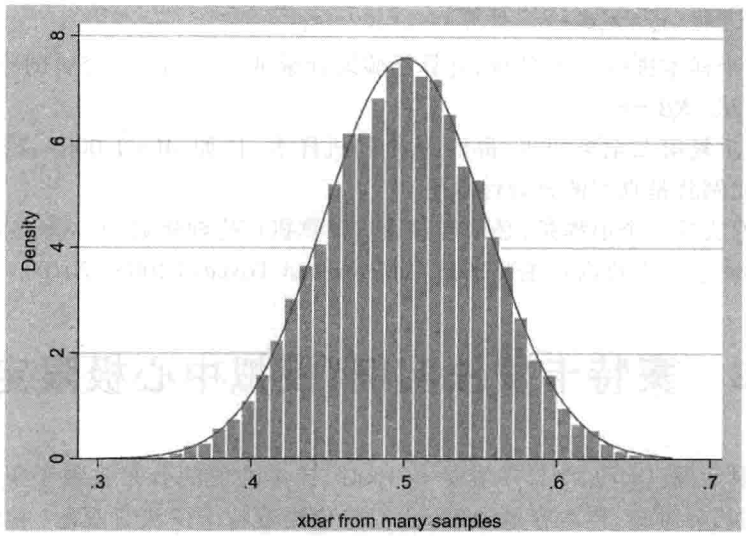


图 19.2 模拟中心极限定理

19.3 蒙特卡罗法实例:服从卡方分布的扰动项

考虑下面的数据生成过程(DGP):

$$y=\alpha+\beta x+\varepsilon,\quad \varepsilon\sim\chi^2(1)-1,\quad x\sim\chi^2(1) \tag{19.1}$$

其中,$\alpha=1$,$\beta=2$,样本容量为 $n=30$。解释变量 x 服从 $\chi^2(1)$ 分布,而扰动项 ε 为经过位移后的 $\chi^2(1)$ 分布(以保证其期望为零),而且 x 与 ε 相互独立。由于小样本理论要求扰动项服从正态分布,这个模型显然不满足小样本理论的假定,但符合大样本理论的要求。

我们将抽取 1 000 个样本容量为 30 的随机样本,计算 1 000 个 $\hat{\beta}_{OLS}$,然后以显著性水平 5% 对原假设"H_0: $\beta=2$"进行 1 000 次 t 检验。由于 t 检验是大样本检验,故 5% 只是"名义显著性水平",而"真实显著性水平"等于在 1 000 次蒙特卡罗模拟中,拒绝"H_0: $\beta=2$"的比例。

为此,先用命令 program 定义一个叫"chi2data"程序进行一次抽样与检验;然后,用命令 simulate 来重复此程序 1 000 次:

```
program chi2data,rclass                    (定义程序 chi2data,并以 r( )形式储存结果)
drop _all                                  (删去内存中已有数据)
set obs 30                                 (确定随机抽样的样本容量为 30)
gen double x = rchi2(1)                    (生成服从χ²(1)分布的解释变量)
gen y = 1 + 2*x + rchi2(1) -1              (生成被解释变量)
reg y x                                    (线性回归)
return scalar t2 = (_b[x] -2)/_se[x]       (计算 t 统计量)
return scalar r2 = abs(return(t2)) > invttail(28,0.025)
                                           (是否拒绝原假设)
end                                        (程序"chi2data"结束)
simulate reject = r(r2),reps(1000) seed(101):chi2data   (模拟 1 000 次)
```

```
mean reject                                    (计算拒绝原假设的比例)
```

在上述程序中,"invttail(28,0.025)"给出了 $t(28)$ 分布所对应的显著性水平为 5% 的双边检验的临界值。该模拟结果显示(参见下表),拒绝原假设的比例为 4.6%,比较接近于 5% 的名义显著性水平。

```
Mean estimation                    Number of obs    =      1000
```

	Mean	Std. Err.	[95% Conf.	Interval]
reject	.046	.0066278	.032994	.059006

19.4 蒙特卡罗积分

蒙特卡罗法的另一常见用途是用来计算复杂或高维的积分,称为"蒙特卡罗积分"(Monte Carlo integration)。考虑计算定积分 $\int_a^b f(x)\mathrm{d}x$,其中 a,b 为有限值。显然,可以通过变量替换,将积分的下限与上限分别变为 0 与 1。因此,不失一般性,仅考虑 $I \equiv \int_0^1 f(x)\mathrm{d}x$。假设随机变量 x 服从在[0,1]上的均匀分布,则随机变量函数 $f(x)$ 的期望值为

$$\mathrm{E}[f(x)] = \int_0^1 f(x)\cdot 1\mathrm{d}x \equiv I \tag{19.2}$$

如果抽取随机变量 x 的样本容量为 S(S 代表 Simulation)的随机样本,记为 $\{x_1,\cdots,x_s,\cdots,x_S\}$,则蒙特卡罗积分估计值即为 $f(x)$ 的样本均值:

$$\hat{I}_{\mathrm{MC}} = \frac{1}{S}\sum_{s=1}^{S} f(x_s) \tag{19.3}$$

根据大数定律,当 $S\to\infty$ 时,样本均值 $\hat{I}_{\mathrm{MC}}$ 依概率收敛于总体均值 $\mathrm{E}[f(x)] = I$。上述方法的缺点是,如果积分上限 a 或下限 b 为无穷,则不适用,因为均匀分布的定义域一定是有限的。为此,考虑从某个适当的概率密度 $g(x)$ 中抽取随机样本 $\{x_1,\cdots,x_s,\cdots,x_S\}$。原积分总可以写为

$$\int_a^b f(x)\mathrm{d}x = \int_a^b \left[\frac{f(x)}{g(x)}\right] g(x)\mathrm{d}x \equiv \int_a^b w(x)g(x)\mathrm{d}x = \mathrm{E}[w(x)] \tag{19.4}$$

其中,$w(x) \equiv \dfrac{f(x)}{g(x)}$。此时,蒙特卡罗积分估计值为

$$\hat{I}_{\mathrm{MC}} = \frac{1}{S}\sum_{s=1}^{S} w(x_s) \tag{19.5}$$

从密度函数 $g(x)$ 中抽样的方法被称为"重要性抽样"(importance sampling),因为函数 $w(x)$ 决定了每个样本点的权重或重要性。

作为实例,考虑计算 $\mathrm{E}\{\exp[-\exp(y)]\}$,其中 $y\sim N(0,1)$。该期望的表达式为

$$\mathrm{E}\{\exp[-\exp(y)]\} = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp[-\exp(y)]\exp(-y^2/2)\mathrm{d}y \tag{19.6}$$

此积分虽然存在,但没有解析解。为此,从标准正态分布中抽取 S 个随机观测值,记其第 s 个抽样值为 y^s,则可使用以下估计值:

$$\hat{\mathrm{E}}\{\exp[-\exp(y)]\} = \frac{1}{S}\sum_{s=1}^{S}\exp[-\exp(y^s)] \tag{19.7}$$

相应的 Stata 命令如下(令 $S=100$):

```
. clear all
. set obs 100
. set seed 10101
. gen double y = rnormal(0,1)
. gen double gy = exp(-exp(y))
. sum gy
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gy	100	.3679463	.2408939	4.50e-11	.8852481

上表显示,当 $S=100$ 时,表达式(19.6)的蒙特卡罗估计值为 0.367 946 3。但我们并不知道 $S=100$ 是否能达到足够的精确度。为此,令 $S=1\ 000$,重复上述命令。

```
. clear all
. set obs 1000
. set seed 10101
. gen double y = rnormal(0,1)
. gen double gy = exp(-exp(y))
. sum gy
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gy	1000	.3821273	.2577481	2.05e-36	.956477

上表显示,当 $S=1\ 000$ 时,表达式(19.6)的蒙特卡罗估计值为 0.382 127 3,与 $S=100$ 时结果有较大差异;故应考虑将 S 增大为 10 000,以此类推。

19.5 最大模拟似然法与模拟矩估计

蒙特卡罗法也可用于最大似然法或矩估计。使用 MLE 的前提是,必须能写出似然函数 $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$。但在有些情况下,该似然函数的表达式可能包含无法求解的积分。比如,在随机效应的非线性面板模型中(参见第 17 章),一般需要将个体效应 u_i 积分掉(因为 u_i 不可观测),才能写出其似然函数。记 u_i 的密度函数为 $g(u_i)$,并假设第 i 个观测值的似然函数为

$$f(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}) = \int h(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta},u_i)g(u_i)\,\mathrm{d}u_i \tag{19.8}$$

如果上式中的积分无解析解,则可使用蒙特卡罗积分进行估计。从分布 $g(u_i)$ 中随机抽取 S 个观测值,记为 $\{u_i^1,\cdots,u_i^S\}$,则上式的估计值可写为

$$\hat{f}(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}h(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta},u_i^s) \tag{19.9}$$

假设样本为 iid,则整个样本的对数似然函数估计值为

$$\ln\hat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\hat{f}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \tag{19.10}$$

其中，n 为样本容量。最大化上式所得到的估计量 $\hat{\boldsymbol{\theta}}_{\text{MSL}}$ 称为“最大模拟似然估计量”（Maximum Simulated Likelihood Estimator，简记 MSL）。在一定的正则条件下，当模拟抽样的次数 $S\to\infty$ 时，$\hat{f}$ 对 f 的近似程度越来越好，即 $(\hat{f}-f)\xrightarrow{p}0$，则 MSL 为一致估计量。进一步，如果 $\sqrt{n}/S\to 0$（即 S 的增长速度快于 $\sqrt{n}$），则 MSL 为渐近有效估计量（渐近等价于 MLE），且服从渐近正态分布。

类似地，在进行矩估计时，如果矩条件中包含无解析解的积分，也可使用蒙特卡罗积分来估计此矩条件，然后再进行矩估计。此方法称为“模拟矩估计”（Method of Simulated Moments），简记 MSM。在一定的正则条件下，MSM 为一致估计量且渐近正态。

19.6 自助法的思想与用途

蒙特卡罗法虽然威力很大，但缺点是必须对总体模型（即数据生成过程）做很具体的假定，比如确定所有参数的取值以及所有解释变量与扰动项的概率分布；而且严格来说，所得到的结论仅适用于那个特定的数据生成过程，并不清楚它在多大意义上能够推广。为此，Efron（1979）提出了一种对原始样本进行“再抽样”（resampling）的方法，即“自助法”或“自举法”（bootstrap）①。

假设从总体抽得样本容量为 n 的随机样本。显然，来自总体的这个样本带有总体的信息。在一定程度上，可以将此样本看作是一个总体，再进行“有放回”（with replacement）地抽样，样本容量仍然为 n。这种样本被称为“自助样本”（bootstrap sample）。由于是有放回地抽样，原来的某些观测值可能不出现，而有些观测值则可能多次出现。自助法的优点是，可以通过计算机模拟毫不费力地获得许多自助样本，然后利用这些自助样本对总体进行统计推断。

假设 $\{x_1, x_2, \cdots, x_n\}$ 是来自总体 F 的一个随机样本，可以定义总体 F 的经验分布函数（empirical distribution function）F_n 如下

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(x_i \leqslant x), \quad -\infty < x < \infty \tag{19.11}$$

其中，$\mathbf{1}(\cdot)$ 为示性函数，而 $\sum_{i=1}^{n}\mathbf{1}(x_i \leqslant x)$ 表示样本中小于或等于 x 的个数。经验分布函数的形状为阶梯函数（step function），在每个 x_i 处向上跳一个台阶。可以证明（参见本章附录），对任意 x，$F_n(x)\xrightarrow{p}F(x)$，即当样本容量 n 越来越大时，经验分布函数 F_n 与真实分布函数 F 越来越像，乃至最终收敛。这正是自助法能够成立的前提，因为自助法可以看成是从经验分布函数中不断地抽样。

自助法在计量经济学中的用途主要有两个方面。首先，对于某些统计量（比如，样本中位数），用常规方法很难得到其标准误的估计量。此时可以使用自助法，计算每个自助样本的样本中位数，从而得到样本中位数的分布，并计算其标准误。其次，对研究者而言，还可以使用自助法得到更加渐近有效的估计量（asymptotic refinement）②。

---

① “Bootstrap”一词的英文原意为“鞋带”。在英文中有句谚语，人应该自己绑鞋带，意即人应该“自助”。由于“bootstrap法”利用现有样本再抽样进行推断（而不依赖于另外的样本或数据），故也是一种“自助”的表现。

② 即在大样本下，估计量能以更快的速度收敛到真实值。

19.7 自助法的分类

根据自助法的操作程序，可以将自助法分为以下几类。

（1）**非参数自助法**（nonparametric bootstrap），也称为“经验分布自助法”（empirical distribution function bootstrap）。这种方法就是前面所介绍的将原始样本进行有放回地随机抽样。在回归模型中，这意味着将被解释变量与所有解释变量，即$(y_i,\boldsymbol{x}_i)$，成对地抽样，故也称为“成对自助法”（paired bootstrap）。这是最简单，也是最常见的自助法。

（2）**参数自助法**（parametric bootstrap）。假设总体分布函数的形式已知，为$F(x,\theta)$，其中θ为未知参数。则可以先得到θ的估计量$\hat{\theta}$（比如，使用最大似然估计法），然后从总体$F(x,\hat{\theta})$中重复抽样。这个方法的前提是必须对总体分布函数的形式比较确信。在此前提下，参数自助法通常比非参数自助法更有效率。在回归模型中，则需要先确定条件分布的具体形式，即$y|\boldsymbol{x}\sim F(\boldsymbol{x},\theta)$。具体来说，一种方法是，在得到估计量$\hat{\theta}$后，给定原先的$\boldsymbol{x}_i$，从$F(\boldsymbol{x}_i,\hat{\theta})$中随机抽样得到对应的$y_i$。这相当于“固定解释变量”（fixed regressors）的情形[①]。另一种方法是，先从$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$中进行再抽样（resample），得到$\boldsymbol{x}_i^*$，然后再从$F(\boldsymbol{x}_i^*,\hat{\theta})$中随机抽样得到对应的$y_i$。这相当于“随机解释变量”（stochastic regressors）的情形。

（3）**残差自助法**（residual bootstrap）。对于回归模型$y_i=g(\boldsymbol{x}_i,\boldsymbol{\beta})+\varepsilon_i$，首先通过估计得到残差$\hat{\varepsilon}_i=y_i-g(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})$，然后对残差$\{\hat{\varepsilon}_1,\hat{\varepsilon}_2,\cdots,\hat{\varepsilon}_n\}$使用自助法，得到残差的自助样本$\{\hat{\varepsilon}_1^*,\hat{\varepsilon}_2^*,\cdots,\hat{\varepsilon}_n^*\}$，然后计算对应的$y_i^*=g(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})+\hat{\varepsilon}_i^*$，进而得到自助样本$\{(y_1^*,\boldsymbol{x}_1),\cdots,(y_n^*,\boldsymbol{x}_n)\}$。

19.8 使用自助法估计标准误

假设原始样本为$\{x_1,x_2,\cdots,x_n\}$。对于未知参数θ的估计量$\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$，需要计算其标准差$\sigma_{\hat{\theta}}\equiv\sqrt{\mathrm{Var}(\hat{\theta})}$，也称为估计量$\hat{\theta}$的“标准误”（standard error）。但有时$\sqrt{\mathrm{Var}(\hat{\theta})}$没有解析表达式，故希望用计算机模拟的方法来估计$\sqrt{\mathrm{Var}(\hat{\theta})}$。

如果可以从真实总体F产生样本容量为n的随机样本（比如，B个样本），对每个样本都可以计算$\hat{\theta}$，从而得到B个估计值$\{\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_B\}$，则可以使用

$$s_{\hat{\theta}}\equiv\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i-\bar{\theta})^2} \tag{19.12}$$

来估计$\sigma_{\hat{\theta}}$，其中$\bar{\theta}\equiv\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i$。但真实总体$F$的分布常常未知，无法从中抽取随机样本，而从实际总体中进行多次实地抽样的成本可能很高。为此，考虑以经验分布函数F_n来近似真实分布函数F，并从F_n中大量抽取随机样本，即在原始样本$\{x_1,x_2,\cdots,x_n\}$中每次有放回地随机抽样，得

① 在计量经济学的本科课程中，有时这样假设。

到样本容量为 n 的自助样本 $\{x_1^*, x_2^*, \cdots, x_n^*\}$，并以此计算 $\hat{\theta}^* = \hat{\theta}(x_1^*, x_2^*, \cdots, x_n^*)$。如此重复，共抽取 B 个自助样本，则得到 θ 的 B 个自助估计值 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\}$。可以定义标准误的自助估计为

$$s_{\hat{\theta}}^* \equiv \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\theta}^*)^2} \tag{19.13}$$

其中，$\bar{\theta}^* \equiv \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$。

19.9 使用自助法进行区间估计

考虑用自助法对 θ 进行置信度为 $(1-\alpha)$ 的区间估计，有以下三种方法。

(1) **百分位法**(percentile method)。根据上一节，我们已经得到自助估计量 $\hat{\theta}^*$ 的经验分布 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\}$。将 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\}$ 按从小到大的顺序排列，并记其 $\alpha/2$ 与 $(1-\alpha/2)$ 上分位数(upper quantile)[①]分别为 $\hat{\theta}_{\alpha/2}^*$ 与 $\hat{\theta}_{1-\alpha/2}^*$，则 θ 的置信区间为

$$[\hat{\theta}_{1-\alpha/2}^*, \hat{\theta}_{\alpha/2}^*] \tag{19.14}$$

(2) **基于正态的置信区间**(normal-based confidence interval)。也可以使用标准正态分布来估计置信区间，即

$$[\hat{\theta} - 1.96 \times s_{\hat{\theta}}^*, \hat{\theta} + 1.96 \times s_{\hat{\theta}}^*] \tag{19.15}$$

其中，$s_{\hat{\theta}}^*$ 是用自助法估计的标准误，并假定置信度为 95%。

(3) **百分位 t 法**(percentile-$t$method)。根据每个自助样本计算对应的自助 t 统计量

$$t_i^* \equiv \frac{\hat{\theta}_i^* - \hat{\theta}}{s_{\hat{\theta}_i^*}}, \quad i = 1, \cdots, B \tag{19.16}$$

其中，$\hat{\theta}$ 为根据原始样本计算的 θ 估计量，而 $s_{\hat{\theta}_i^*}$ 是根据 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*\}$ 计算的标准误。如此，即得到自助 t 统计量的经验分布 $\{t_1^*, t_2^*, \cdots, t_B^*\}$，并记其 $\alpha/2$ 与 $(1-\alpha/2)$ 上分位数(upper quantile)分别为 $t_{\alpha/2}^* > 0$ 与 $t_{1-\alpha/2}^* < 0$，则 θ 的置信区间为 $[\hat{\theta} + t_{1-\alpha/2}^* \times s_{\hat{\theta}}, \hat{\theta} + t_{\alpha/2}^* \times s_{\hat{\theta}}]$，其中 $s_{\hat{\theta}}$ 是根据原始样本计算的标准误。百分位 t 法比百分位法更渐近有效，但在小样本中不一定有优势。

19.10 使用自助法进行假设检验

考虑用自助法进行如下双边检验，H_0: $\theta = \theta_0$ *vs* H_1: $\theta \neq \theta_0$。一种方法是，如果 $\theta_0 \in [\hat{\theta}_{1-\alpha/2}^*, \hat{\theta}_{\alpha/2}^*]$，则接受原假设 H_0；反之，则拒绝。这就是“百分位法”(percentile method)。

另一方法是，在假设 H_0 成立的情况下，计算原始样本的 t 统计量

① “上分位数”指的是从分布的上限往下计算的分位数，而不是通常的从下限往上计算的(下)“分位数”。

$$t \equiv \frac{\hat{\theta} - \theta_0}{s_{\hat{\theta}}} \tag{19.17}$$

如果 $t \in [\hat{\theta} + t^*_{1-\alpha/2} \times s_{\hat{\theta}}, \hat{\theta} + t^*_{\alpha/2} \times s_{\hat{\theta}}]$，则接受原假设 H_0；反之，则拒绝。其中，$t^*_{\alpha/2}$ 与 $t^*_{1-\alpha/2}$ 的定义如上。这就是百分位 t 法，它比百分位法更渐近有效。

19.11 自助法的一致性(选读)

自助法是一致估计量吗？考虑以下一般的自助法。随机样本 $\{x_1, x_2, \cdots, x_n\}$ 来自总体 $F(x, \theta)$。我们希望知道统计量 $T_n = T_n(x_1, x_2, \cdots, x_n)$ 的概率分布 $G_n(t, F) \equiv \mathrm{P}(T_n \leqslant t)$，其中 $G_n(t, F)$ 的下标 n 表示该分布依赖于样本容量 n。另外，$G_n(t, F)$ 也依赖于真实的总体分布 F。

大样本理论(渐近理论)通过让 $n \to \infty$，即用 $G_\infty(t, F)$ 来近似 $G_n(t, F)$；对于 $G_n(t, F)$ 中未知的 F，则使用其一致估计量 $\hat{F} = F(x, \hat{\theta})$ 来替代，其中 $\hat{\theta}$ 是 θ 的一致估计量。自助法则不需要让 $n \to \infty$，而是使用经验分布函数 F_n 来替代 F，即用 $G_n(t, F_n)$ 来近似 $G_n(t, F)$。尽管 $G_n(t, F_n)$ 一般没有解析解，但可以把样本 F_n 近似地看作总体 F，并从 F_n 中进行再抽样，获得自助样本。从一个自助样本 $\{x_1^*, x_2^*, \cdots, x_n^*\}$，就可以得到一个统计量 $T_n^* = T_n(x_1^*, x_2^*, \cdots, x_n^*)$。独立地重复抽样 B 次，得到 B 个统计量 $\{T^*_{n,1}, T^*_{n,2}, \cdots, T^*_{n,B}\}$，从中可以得到 T_n 分布函数的自助估计(bootstrap estimate) $\hat{G}_n(t, F_n)$，即 $\{T^*_{n,1}, T^*_{n,2}, \cdots, T^*_{n,B}\}$ 的经验分布函数：

$$\hat{G}_n(t, F_n) \equiv \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}(T^*_{n,i} \leqslant t) \tag{19.18}$$

其中，$\mathbf{1}(\cdot)$ 为示性函数。显然，当 $B \to \infty$ 时，$\hat{G}_n(t, F_n) \xrightarrow{p} G_n(t, F_n)$，因为前者是后者的经验分布函数(参见本章附录)。然而，当 $n \to \infty$ 时，是否 $G_n(t, F_n) \xrightarrow{p} G_n(t, F)$？显然，这要求 $F_n \xrightarrow{p} F$(即 F_n 是 F 的一致估计)。另外，这还要求总体分布函数 $F(x, \theta)$ 与统计量 T_n 的分布函数 $G_n(t, F)$ 都是光滑的。这些条件在实践中一般都成立，故自助估计量是一致的。

19.12 异方差情况下的自助法

如果回归模型的扰动项存在条件异方差，通常的解决方法是使用稳健标准误，参见第 5 章。尽管稳健标准误在大样本下是一致的，但在小样本中可能不够精确。使用自助法可能得到更精确的估计。

由于条件异方差的存在并不影响观测数据 $(y_i, \boldsymbol{x}_i)$ 仍然为独立同分布[①]，故仍可以使用成对自助法(paired bootstrap)。然而，残差自助法(residual bootstrap)却不成立，因为在条件异方差的情况下，扰动项不是独立同分布的，故经验分布函数 F_n 不是总体分布函数 F 的一致估计，自助估计量也就不是一致估计。

然而，如果对残差 $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n\}$ 先进行某种线性变换后再进行抽样，则可能满足从异方差

① 独立同分布的假设表明，扰动项的条件方差函数 $\mathrm{Var}(\varepsilon_i | \boldsymbol{x}_i)$ 的形式相同，但其具体取值可以依 $\boldsymbol{x}_i$ 的不同而变化。

的扰动项进行抽样的要求。为此，Wu（1986）与 Liu（1988）提出了如下“野自助法”（wild bootstrap）。定义具有两点分布的新残差为

$$\hat{\varepsilon}_i^* = \begin{cases} \dfrac{1-\sqrt{5}}{2}\hat{\varepsilon}_i, & \text{以}\dfrac{\sqrt{5}+1}{2\sqrt{5}}\text{的概率} \\ \dfrac{1+\sqrt{5}}{2}\hat{\varepsilon}_i, & \text{以}\dfrac{\sqrt{5}-1}{2\sqrt{5}}\text{的概率} \end{cases} \tag{19.19}$$

可以证明（参见习题），$\mathrm{E}(\hat{\varepsilon}_i^*)=0$，$\mathrm{Var}(\hat{\varepsilon}_i^*)=\hat{\varepsilon}_i^2$。因此，对$\{\hat{\varepsilon}_1^*,\hat{\varepsilon}_2^*,\cdots,\hat{\varepsilon}_n^*\}$进行再抽样，就是从异方差的扰动项进行抽样。Horowitz（1997，2001）的模拟显示，在存在异方差的情况下，野自助法比成对自助法更有效率；即使在同方差的情况下，也不逊色。

19.13 面板数据与时间序列的自助法

对于 n 大 T 小的短面板（short panel），考虑其离差形式

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}_{it}'\boldsymbol{\beta} + \tilde{\varepsilon}_{it} \tag{19.20}$$

其中，固定效应已被去掉（参见第 15 章）。假设对于不同的个体 i，残差 $\tilde{\varepsilon}_{it}$是相互独立的；但给定个体 i，则 $\tilde{\varepsilon}_{it}$可能存在异方差或自相关。此时，可以使用成对自助法对个体 i 进行再抽样，而不对时间 t 进行再抽样，即如果抽中个体 i，则个体 i 在所有时间的观测值都同时被抽中。这种方法称为“分块自助法”（block bootstrap），对于非线性面板数据或聚类数据（clustered data）也适用。

由于自助法假定样本为独立同分布的，而时间序列数据通常存在自相关（故不是独立同分布的），因此针对时间序列的自助法更为复杂。

19.14 自助法的 Stata 命令

使用自助法的一个具体问题是，如何确定自助样本的个数 B。为了节省计算时间，Stata 的默认值为 $B=50$。然而，对于正式的研究而言，这一般不够，故必须使用选择项来指定自助样本个数 B。Cameron and Trivedi（2010）建议，针对 5% 的显著性水平，如果使用自助法计算标准误差，则取 $B=400$；如果进行区间估计或假设检验，则取 $B=999$。

在 Stata 中有三种方法可以实现自助法。

1. 使用选择项“vce(bootstrap)”

如果只是用自助法估计标准误，对于横截面数据而言，大多数 Stata 命令支持选择项“vce(bootstrap)”进行成对自助法，比如，

```
reg y x1 x2 x3,vce(boot,reps(400) seed(10101) nodots)
```

其中，“boot”为“bootstrap”的简写，选择项“reps(400)”表示 $B=400$，“seed(10101)”表示抽样“种子”为“10101”以保证可重复性[①]，“nodots”表示在执行命令过程中不显示代表抽样进程的“点”（可节省时间）。Stata 的输出结果包括了基于正态的置信区间（normal-based

① 只要每次均使用相同的抽样“种子”（seed），即可保证自助抽样结果的可重复性。

confidence interval)。

对于面板数据而言,某些 Stata 命令也支持选择项"vce(bootstrap)",但进行的是针对个体 i 的分块自助法,而非针对观测值(i,t)的成对自助法。

在执行了含选择项"vce(bootstrap)"的 Stata 命令之后,还可以使用以下命令获得自助法的更多信息:

```
estat boot, all
```

其中,选择项"all"表示显示根据不同方法计算的置信区间。

2. 使用命令"bootstrap"

对于一般的统计量使用自助法,可以使用命令 bootstrap,其格式为

```
bootstrap explist[,options]: command
```

其中,"explist"提供了需要进行自助法估计的统计量。比如,使用如下命令可以得到与上面使用选择项"vce(bootstrap)"相同的结果。

```
bootstrap,reps(400) seed(10101) nodots: reg y x1 x2 x3
```

这个命令只对回归系数用了自助法,如果对标准误也使用自助法,则可以用如下命令。

```
bootstrap _b _se,reps(400) seed(10101) nodots: reg y x1 x2 x3
```

其中,"_b"表示回归系数(得到回归系数估计量的自助分布),"_se"表示标准误(得到标准误估计量的自助分布)。

在使用命令 bootstrap 之后,也可以用"estat boot"得到进一步的自助信息。

3. 使用命令"bsample"

使用命令 bootstrap 的局限是,自助法的整个计算过程必须能够用一个命令(比如,reg y x1 x2 x3)来表达。对于更复杂的情形(比如,参数自助法、残差自助法、野自助法),则可以用命令 bsample 进行一次有放回地再抽样,计算相应的统计量,然后用命令 simulate 进行重复抽样与计算。比如,使用命令"bsample 50"将得到样本容量为 50 的有放回再抽样样本(原始样本的样本容量必须大于或等于 50),这个再抽样的样本数据将替代内存中已有的数据。如果不指定再抽样的样本容量,则默认值为原始样本的样本容量 n。

19.15 使用自助法进行稳健的豪斯曼检验

考虑对解释变量的内生性进行豪斯曼检验(参见第 10 章)[①]。对于模型 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$,记 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 与 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 分别为 OLS 与 IV 估计量。如果原假设"H_0:所有解释变量均为外生变量"成立,则 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 与 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 都是一致的,但 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 更有效率。反之,如果 H_0 不成立,则 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 不一致,而 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 一致。因此,在 H_0 成立的情况下,$(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}}) \xrightarrow{p} \mathbf{0}$,而统计量

$$(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})' \widehat{[\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})]}^{-} (\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}}) \xrightarrow{d} \chi^2(r) \tag{19.21}$$

其中,r 为内生解释变量的个数。为了计算 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})$,需要知道 $\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{IV}}, \hat{\boldsymbol{\beta}}_{\text{OLS}})$。为此,Hausman(1978)假设,在 H_0 成立的情况下,OLS 是最有效率的(fully efficient)。在此条件下,可

① 对于使用豪斯曼检验来决定应该在面板模型中使用固定效应还是随机效应估计量,以下讨论也大致适用。

以证明 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}}-\hat{\boldsymbol{\beta}}_{\text{OLS}})=\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}})-\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$。但如果存在异方差,则 OLS 并不是最有效率的(最有效率的估计量为 GLS)。因此,该检验在异方差情况下不适用。

为了得到异方差稳健的豪斯曼检验,可以使用自助法来直接计算 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}}-\hat{\boldsymbol{\beta}}_{\text{OLS}})$,而不需要假设 OLS 在 H_0 成立的情况下是最有效率的。自助法的操作步骤为,对样本进行 B 次自助抽样,得到 B 个自助估计量 $\{\hat{\boldsymbol{\beta}}^{*}_{\text{IV},b}-\hat{\boldsymbol{\beta}}^{*}_{\text{OLS},b},b=1,\cdots,B\}$,进而得到对 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}}-\hat{\boldsymbol{\beta}}_{\text{OLS}})$ 的自助估计量:

$$\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}}-\hat{\boldsymbol{\beta}}_{\text{OLS}})}=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}^{*}_{\text{IV},b}-\hat{\boldsymbol{\beta}}^{*}_{\text{OLS},b}-\bar{\boldsymbol{\beta}}^{*}_{\text{Diff}})(\hat{\boldsymbol{\beta}}^{*}_{\text{IV},b}-\hat{\boldsymbol{\beta}}^{*}_{\text{OLS},b}-\bar{\boldsymbol{\beta}}^{*}_{\text{Diff}})' \tag{19.22}$$

其中,$\bar{\boldsymbol{\beta}}^{*}_{\text{Diff}}\equiv\frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}^{*}_{\text{IV},b}-\hat{\boldsymbol{\beta}}^{*}_{\text{OLS},b})$。

以数据集 grilic.dta 为例。第 10 章曾使用该数据集估计教育投资的回报率,并进行了传统(非稳健)的豪斯曼检验。为了进行稳健的豪斯曼检验,先定义一个叫“hausmantest”的程序,来计算 $(\hat{\boldsymbol{\beta}}_{\text{OLS}}-\hat{\boldsymbol{\beta}}_{\text{IV}})$:

```
. use grilic.dta,clear
. program hausmantest,eclass
  1. tempname b bols biv
  2. reg lw iq s expr tenure rns smsa,r
  3. matrix `bols'=e(b)
  4. ivregress 2sls lw s expr tenure rns smsa (iq=med kww),r
  5. matrix `biv'=e(b)
  6. matrix `b'=`bols'-`biv'
  7. ereturn post `b'
  8. end
```

然后,用命令“bootstrap”进行 400 次自助抽样,计算 $(\hat{\boldsymbol{\beta}}^{*}_{\text{OLS}}-\hat{\boldsymbol{\beta}}^{*}_{\text{IV}})$ 400 次,并得到其自助协方差矩阵:

```
. bootstrap _b,reps(400) seed(10101) nodots nowarn: hausmantest
```

运行结果为

```
Bootstrap results                                 Number of obs      =       758
                                                  Replications       =       400

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0106493   .0060772    -1.75   0.080    -.0225603    .0012618
           s |    .032007   .0182011     1.76   0.079    -.0036664    .0676804
        expr |  -.0039794   .0036892    -1.08   0.281    -.0112101    .0032513
      tenure |   .0045648    .003909     1.17   0.243    -.0030967    .0122262
         rns |  -.0310054   .0199684    -1.55   0.120    -.0701428     .008132
        smsa |   .0095145   .0128142     0.74   0.458    -.0156008    .0346298
       _cons |   .6771285   .3887073     1.74   0.082    -.0847239    1.438981
------------------------------------------------------------------------------
```

然后检验所有解释变量的显著性:

```
. test s iq expr tenure rns smsa _cons
```

```
(1)  s = 0
(2)  iq = 0
(3)  expr = 0
(4)  tenure = 0
(5)  rns = 0
(6)  smsa = 0
(7)  _cons = 0

          chi2(  7) =    3.19
        Prob > chi2 =    0.8673
```

需要注意的是，上述检验的统计量(3.19)是正确的，但χ^2分布的自由度应该为1(内生解释变量的个数)而非7(所有解释变量的个数)。为此，手工计算$\chi^2(1)$的p值，

```
. dis chi2tail(1,3.19)
. 07409001
```

结果显示，p值为0.07，故可在10%的显著性水平上拒绝“所有解释变量均为外生变量”的原假设。

有关蒙特卡罗模拟与自助法的进一步知识及Stata操作，可参考Cameron and Trivedi (2005, 2010)的相应章节。

习　题

19.1 对于野自助法，证明(1)$E(\hat{\varepsilon}_i^*)=0$；(2)$Var(\hat{\varepsilon}_i^*)=\hat{\varepsilon}_i^2$。

19.2 针对数据集nerlove.dta(参见第4章)，使用自助标准误估计成本函数。

19.3 (选做)针对面板数据集traffic.dta(参见第15章)，使用自助法进行稳健的豪斯曼检验，以确定应该使用固定效应模型，还是随机效应模型。

附　录

A19.1 随机数的产生

服从在(0,1)区间上均匀分布的随机变量的样本值，称为“均匀分布随机数”，简称“随机数”(random number)。有了均匀分布的随机数后，就可以产生几乎任何分布的随机数了。然而，由电脑产生的所谓“随机数”，其实只是“伪随机数”(pseudo random number)，因为它仍然来自确定性的序列(由递推公式产生)。计算机其实并不知道如何掷骰子！尽管如此，这些“伪随机数”都能通过独立性与均匀分布的统计检验，故可以作为随机数来使用。

比如，一个简单的“随机数产生器”(random number generator)可使用以下递推公式：

$$x_j=(kx_{j-1}+c)\bmod m \tag{19.23}$$

其中，k，c为常数，算子“a mod b”表示a除以b后的余数，而m是计算机所能表示的最大数①。这个序列就是介于0与m之间的整数，而$r_j\equiv x_j/m$就是介于0与1之间的随机数。初始值x_0，称为种子(seed)，决定了该随机序列的初始值$r_0\equiv x_0/m$。为了使随机样本具有可重复性，保证以后抽样或别人抽样也能得到完全一样的样本，需要设定种子。如果不给定种子，则按照“计算机内置钟表”(computer clock)来自动选择种子，这样每次抽样的样本就会不同，模拟的结果也就不能完全复制。

在Stata中，使用命令“set seed #”来确定“种子”，例如：

① 比如，如果计算机使用32位数进行计算，则$m=2^{32}-1$。

```
set seed 10101          (确定"种子"为 10101)
set obs 30              (确定样本容量为 30)
gen x = runiform()      (产生均匀分布的随机变量 x)
```

A19.2　离散型随机变量的抽样

假设离散型随机变量的分布律为

$$P(X = x_i) = p_i, \quad i = 1,2,\cdots, \quad \sum_{i=1}^{\infty} p_i = 1 \tag{19.24}$$

其中，x_i 按下标 i 从小到大的顺序排列。显然，随机变量 X 的累积分布函数是一个值域在[0, 1]之间的阶梯函数，参见图 19.3。

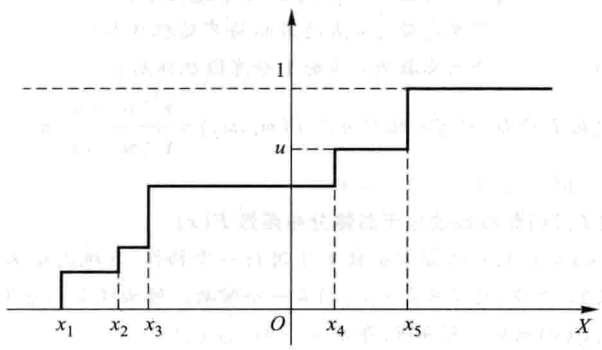


图 19.3　离散型随机变量的抽样

首先产生一个均匀分布的随机数 $u \in (0,1)$，然后在纵轴上沿着累积分布函数找到 u 在横轴上对应的 x_i 取值(在图 19.3 中为 x_4)。在代数上，可以定义 X 为

$$X = \begin{cases} x_1, & u < p_1 \\ x_2, & p_1 \leqslant u < p_1 + p_2 \\ \cdots\cdots & \\ x_i, & \sum_{j=1}^{i-1} p_j \leqslant u < \sum_{j=1}^{i} p_j \\ \cdots\cdots & \end{cases} \tag{19.25}$$

可以证明，这个随机变量 X 的分布律就是 $P(X = x_i) = p_i$，因为

$$P(X = x_i) = P\left\{\sum_{j=1}^{i-1} p_j \leqslant u < \sum_{j=1}^{i} p_j\right\} = \sum_{j=1}^{i} p_j - \sum_{j=1}^{i-1} p_j = p_i \tag{19.26}$$

其中，第二个等号用到了均匀分布"$u \sim U(0,1)$"的性质。

在 Stata 中，对离散型随机变量进行随机抽样的命令有

```
gen x = rpoisson(m)       (产生参数为 m 的泊松分布随机样本)
gen x = rbinomial(n,p)    (产生二项分布 b(n,p)的随机样本)
```

A19.3　连续型随机变量的抽样

上述对离散型随机变量抽样的思想，对于连续型随机变量也是适用的。假设连续型随机变量 X 的累积分布函数为 $F(x)$，在产生一个均匀分布的随机数 $u \in (0,1)$ 后，仍然可以在纵轴上沿着累积分布函数找到 u 在横轴上对应的 x 取值。其理论基础是下面的这个命题。

命题　设随机变量 $U \sim U(0,1)$，$F(x)$ 是随机变量 X 的累积分布函数，且 $F(x)$ 是一个严格单调递增的连续函数，则随机变量 $F^{-1}(U)$ 具有分布函数 $F(x)$，其中 $F^{-1}(\cdot)$ 是 $F(\cdot)$ 的反函数。

证明：由于 $F(x)$ 是一个严格单调递增的连续函数，故其反函数 $F^{-1}(x)$ 存在，而且 $F^{-1}(x)$ 也是一个严格单调递增的连续函数。因此，随机变量 $F^{-1}(U)$ 的分布函数为

$$\begin{aligned} P\{F^{-1}(U) \leqslant x\} &= P\{F[F^{-1}(U)] \leqslant F(x)\} && (\text{因为 } F(x) \text{严格单调递增}) \\ &= P\{U \leqslant F(x)\} && (\text{反函数的性质}) \end{aligned}$$

$$= F(x) \qquad (\text{因为 } U \sim U(0,1))$$

因此，$F^{-1}(U)$的分布函数为$F(x)$。

这个命题表明，要产生以$F(x)$为分布函数的随机变量X，只要先产生随机数$U \sim U(0,1)$，然后再令$X = F^{-1}(U)$即可。这种方法称为“逆概率变换”(inverse-probability transformation)。

在Stata中，对连续型随机变量进行随机抽样的命令有

```
gen x = rnormal()          (产生标准正态分布 N(0,1) 的随机样本)
gen x = rnormal(m,s)       (产生正态分布 N(m,s^2) 的随机样本)
gen x = rt(m)              (产生自由度为 m 的 t 分布随机样本)
gen x = rchi2(m)           (产生自由度为 m 的 χ^2 分布随机样本)
gen x = rbeta(a,b)         (产生参数为 a,b 的 Beta 分布随机样本)
gen x = rgamma(a,b)        (产生参数为 a,b 的 Γ 分布随机样本)
```

对于自由度为(m_1, m_2)的F分布，则可以使用公式$F(m_1, m_2) = \dfrac{\chi^2(m_1)/m_1}{\chi^2(m_2)/m_2}$来产生其随机样本，即

```
gen x = (rchi2(m1)/m1)/(rchi2(m2)/m2)
```

A19.4 经验分布函数$F_n(x)$依概率收敛于总体分布函数$F(x)$

证明：任意给定x，$-\infty < x < \infty$，从总体分布函数F中进行一次抽样，则观测值落入区间$(-\infty, x]$的概率为$F(x)$(根据分布函数的定义)。其中，对于给定x，$F(x)$是一个常数。因为样本容量为n，故在n次抽样中，观测值落在区间$(-\infty, x]$的次数$S(x)$服从二项分布，即$S(x) \sim b(n, F(x))$。

根据二项分布的性质，$\mathrm{E}[S(x)] = nF(x)$，而$\mathrm{Var}[S(x)] = nF(x)[1 - F(x)]$。

根据定义，经验分布函数$F_n(x) = \dfrac{1}{n}S(x)$，故$\mathrm{E}[F_n(x)] = \dfrac{1}{n}\mathrm{E}[S(x)] = F(x)$，$\mathrm{Var}[F_n(x)] = \dfrac{1}{n^2}\mathrm{Var}[S(x)] = \dfrac{1}{n}F(x)[1 - F(x)] \xrightarrow{P} 0$。

因此，$F_n(x)$依均方收敛于$F(x)$。由于依均方收敛是依概率收敛的充分条件(参见第5章)，故随机变量$F_n(x)$依概率收敛于常数$F(x)$。

第 20 章　平稳时间序列

第 20 ~ 22 章介绍“时间序列数据”(time series data)的计量模型。根据时间序列的随机过程特性,可以分为“平稳序列”(stationary)与“非平稳序列”(non-stationary)两大类,需要使用不同的计量方法。本章介绍平稳序列①,下一章介绍非平稳序列。

20.1　时间序列的数字特征

时间序列指的是同一个体在不同时点上的数据。对于离散时间$\{1,2,\cdots,T\}$,记随机变量 y 的相应观测值为$\{y_1,y_2,\cdots,y_T\}$,并假设其为(严格)平稳过程。因此,该时间序列的期望、方差、自协方差、自相关系数等数字特征均不随时间推移而改变。考察时间序列的数字特征,有助于判别应该使用怎样的时间序列模型进行估计。

期望 $\mu \equiv \mathrm{E}(y)$ 反映的是该序列的平均水平,可以用样本均值 $\bar{y} \equiv \frac{1}{T}\sum_{t=1}^{T} y_t$ 来估计。

方差 $\mathrm{Var}(y)$ 反映的是该序列的波动幅度,可以用 $\frac{1}{T-1}\sum_{t=1}^{T}(y_t-\bar{y})^2$ 来估计。

通常,不同期的观测值之间存在相关性②。为此,引入以下两个概念。

定义　时间序列$\{y_t\}$的“**k 阶自协方差**”(autocovariance of order k)为

$$\gamma_k \equiv \mathrm{Cov}(y_t,y_{t+k}) = \mathrm{E}[(y_t-\mu)(y_{t+k}-\mu)] \tag{20.1}$$

其中,$\mu \equiv \mathrm{E}(y)$。它反映的是同一变量(y)相隔 k 期之间的自相关程度。显然,当 $k=0$ 时,$\gamma_0 = \mathrm{Var}(y)$。对 γ_k 的估计值为“样本自协方差”:

$$\hat{\gamma}_k \equiv \frac{1}{T-k}\sum_{t=1}^{T-k}(y_t-\bar{y})(y_{t+k}-\bar{y}) \tag{20.2}$$

定义　时间序列$\{y_t\}$的“**k 阶自相关系数**”(autocorrelation of order k)为

$$\rho_k \equiv \mathrm{Corr}(y_t,y_{t+k}) \equiv \frac{\mathrm{Cov}(y_t,y_{t+k})}{\mathrm{Var}(y_t)} \tag{20.3}$$

它将自协方差标准化为介于$[-1,1]$之间的量。显然,对于严格平稳过程,ρ_k 不依赖于时间,而仅仅是滞后阶数 k 的函数,故被称为“自相关函数”(Autocorrelation Function,简记 ACF)。将(k,ρ_k)画成图,即为“自相关图”(correlogram)。由于自相关函数关于原点对称($\rho_k = \rho_{-k}$),故一般只画自相关图的正的半边。对 ρ_k 的估计值为“样本自相关系数”:

$$\hat{\rho}_k \equiv \hat{\gamma}_k / \hat{\gamma}_0 \tag{20.4}$$

然而,y_t 与 y_{t+k}之间的相关性可能由二者之间的变量$\{y_{t+1},\cdots,y_{t+k-1}\}$所引起,而 $\rho_k \equiv \mathrm{Corr}(y_t,$

---

① 如无特别说明,本章的时间序列均为“严格平稳”(strictly stationary)过程。

② 如果将每个观测值比作一颗“珍珠”,则可以将时间序列想象为将“珍珠”串在一起的项链,即相邻的“珍珠”之间由线连在一起而“相关”。如果时间序列比较短,则每个观测值对于研究者的价值,大概也如“珍珠”般珍贵。

y_{t+k})并未对$\{y_{t+1},\cdots,y_{t+k-1}\}$的作用进行控制。为此,引入以下概念。

定义 时间序列$\{y_t\}$的"**k 阶偏自相关系数**"(partial autocorrelation of order k)为

$$\rho_k^* \equiv \mathrm{Corr}(y_t, y_{t+k} | y_{t+1},\cdots,y_{t+k-1}) \tag{20.5}$$

即给定$\{y_{t+1},\cdots,y_{t+k-1}\}$条件下,$y_t$ 与 y_{t+k}的条件相关系数。显然,ρ_k^* 也只是 k 的函数,故称为"偏自相关函数"(Partial Autocorrelation Function,简记 PACF)。为了得到对 ρ_k^* 的样本估计值 $\hat{\rho}_k^*$,对以下 k 阶自回归方程进行 OLS 估计:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_k y_{t-k} + \varepsilon_t \tag{20.6}$$

则 $\hat{\beta}_k$ 就是"k 阶样本偏自相关系数"$\hat{\rho}_k^*$。以上这些时间序列的数字特征,都是时间序列固有的特征,不依赖于模型的设定。另一方面,我们在设定模型时,希望尽可能地与这些数字特征相一致。

20.2 自回归模型

到目前为止,我们强调的是以回归模型来推断因果关系。如果从一个客户的角度仅关心某变量(比如股价)的未来值,则可以用该变量的过去值来预测其未来值(因为时间序列一般存在自相关)。这种模型被称为"单变量时间序列"(univariate time series)。此时,可以不必理会因果关系,只考虑相关关系即可。比如,看到街上有人带伞,可以预测今天要下雨,但行人带伞并不导致下雨。对于样本数据$\{y_1,y_2,\cdots,y_T\}$,最简单的预测方法为一阶自回归(AR(1)):

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \quad (t=2,\cdots,T) \tag{20.7}$$

其中,扰动项 ε_t 为白噪声,满足零期望 $\mathrm{E}(\varepsilon_t)=0$,同方差 $\mathrm{Var}(\varepsilon_t)=\sigma_\varepsilon^2$,且无自相关 $\mathrm{Cov}(\varepsilon_t,\varepsilon_s)=0, t\neq s$。假设 $|\beta_1|<1$,则 $\{y_t\}$ 为渐近独立的平稳过程(ergodic stationary)。由于 y_{t-1} 依赖于 $\{\varepsilon_{t-1},\cdots,\varepsilon_1\}$,而 ε_t 与 $\{\varepsilon_{t-1},\cdots,\varepsilon_1\}$ 不相关,故 y_{t-1} 与 ε_t 不相关,因此用 OLS 估计方程(20.7)是一致的。然而,使用 OLS 只能用观测值 $t=2,\cdots,T$ 进行回归,将损失一个样本容量。

为了提高估计效率,考虑最大似然估计。为此,进一步假设 $\{\varepsilon_t\}$ 为独立同分布,且服从正态分布 $N(0,\sigma_\varepsilon^2)$。由于 $\{y_t\}$ 为平稳过程,其期望与方差均不随时间而变。对方程(20.7)两边同时取期望可得

$$\mathrm{E}(y) = \beta_0 + \beta_1 \mathrm{E}(y) \tag{20.8}$$

经移项整理可知,$\{y_t\}$ 的无条件期望为 $\dfrac{\beta_0}{1-\beta_1}$。对方程(20.7)两边同时取方差可得

$$\mathrm{Var}(y) = \beta_1^2 \mathrm{Var}(y) + \sigma_\varepsilon^2 \tag{20.9}$$

故 $\{y_t\}$ 的无条件方差为 $\dfrac{\sigma_\varepsilon^2}{1-\beta_1^2}$。因此,$y_1$ 服从正态分布 $N\left(\dfrac{\beta_0}{1-\beta_1},\dfrac{\sigma_\varepsilon^2}{1-\beta_1^2}\right)$,其(无条件)密度函数为

$$f_{y_1}(y_1) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2/(1-\beta_1^2)}} \exp\left\{\frac{-[y_1-(\beta_0/(1-\beta_1))]^2}{2\sigma_\varepsilon^2/(1-\beta_1^2)}\right\} \tag{20.10}$$

在给定 y_1 的条件下,y_2 的条件分布为,$y_2|y_1 \sim N(\beta_0+\beta_1 y_1,\sigma_\varepsilon^2)$,其条件密度为

$$f_{y_2|y_1}(y_2|y_1) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{\frac{-(y_2-\beta_0-\beta_1 y_1)^2}{2\sigma_\varepsilon^2}\right\} \tag{20.11}$$

则 y_1 与 y_2 的联合分布密度为

$$f_{y_1,y_2}(y_1,y_2) = f_{y_1}(y_1) f_{y_2|y_1}(y_2|y_1) \tag{20.12}$$

依此类推得，$y_t|y_{t-1}\sim N(\beta_0+\beta_1 y_{t-1},\sigma_\varepsilon^2)$，$t=2,\cdots,T$。写出整个样本数据$\{y_1,y_2,\cdots,y_T\}$的联合概率密度(即似然函数)为

$$f_{y_1,\cdots,y_T}(y_1,\cdots,y_T)=f_{y_1}(y_1)\prod_{t=2}^{T}f_{y_t|y_{t-1}}(y_t|y_{t-1}) \tag{20.13}$$

取对数可得对数似然函数：

$$\ln L(\beta_0,\beta_1,\sigma_\varepsilon^2;y_1,\cdots,y_T)=\ln f_{y_1}(y_1)+\sum_{t=2}^{T}\ln f_{y_t|y_{t-1}}(y_t|y_{t-1}) \tag{20.14}$$

代入$f_{y_1}(y_1)$与$f_{y_t|y_{t-1}}(y_t|y_{t-1})$的表达式可得

$$\begin{aligned}\ln L=&-\frac{1}{2}\ln 2\pi-\frac{1}{2}\ln[\sigma_\varepsilon^2/(1-\beta_1^2)]-\frac{[y_1-(\beta_0/(1-\beta_1))]^2}{2\sigma_\varepsilon^2/(1-\beta_1^2)}\\&-\frac{T-1}{2}\ln 2\pi-\frac{T-1}{2}\ln\sigma_\varepsilon^2-\sum_{t=2}^{T}\frac{(y_t-\beta_0-\beta_1 y_{t-1})^2}{2\sigma_\varepsilon^2}\end{aligned} \tag{20.15}$$

寻找最优参数$(\beta_0,\beta_1,\sigma_\varepsilon^2)$，使得 $\ln L$ 最大化，这个估计量被称为“精确最大似然估计量”(Exact MLE)。由于 $\ln L$ 为$(\beta_0,\beta_1,\sigma_\varepsilon^2)$的非线性函数，故需要使用迭代法进行数值计算。

如果样本容量 T 较大，则第一个观测值对似然函数的贡献较小，可以忽略。此时，相当于考虑在第一个观测值 y_1 给定的情况下，$\{y_2,\cdots,y_T\}$的条件分布，则对数似然函数简化为

$$\ln L=-\frac{T-1}{2}\ln 2\pi-\frac{T-1}{2}\ln\sigma_\varepsilon^2-\sum_{t=2}^{T}\frac{(y_t-\beta_0-\beta_1 y_{t-1})^2}{2\sigma_\varepsilon^2} \tag{20.16}$$

最大化上式所得到的估计量称为“条件最大似然估计量”(conditional MLE)。显然，(20.16)式右边第一项不含任何参数(可视为常数)，第二项仅与 σ_ε^2 有关，而在第三项中 σ_ε^2 仅出现于分母。因此，为了使上式最大化，必须让上式右边第三项的分子最小化(参见第 6 章)，即

$$\min_{(\beta_0,\beta_1)}\sum_{t=2}^{T}(y_t-\beta_0-\beta_1 y_{t-1})^2 \tag{20.17}$$

其结果与对方程(20.7)进行 OLS 估计是一样的。因此，与“精确 MLE”相比，“条件 MLE”的计算更为简便。更一般地，可以考虑 p 阶自回归模型，记为 AR(p)：

$$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\varepsilon_t \tag{20.18}$$

对 AR(p)的估计与 AR(1)类似。在使用“条件 MLE”时，考虑在给定$\{y_1,\cdots,y_p\}$的情况下，$\{y_{p+1},\cdots,y_T\}$的条件分布。由于条件 MLE 等价于 OLS，而后者并不依赖于正态性假定，故条件 MLE 的一致性也不依赖于正态性假定。

然而，通常我们并不知道滞后期 p。因此，如何估计 $\hat{p}$ 在实践中有着重要的意义。

方法之一：设一个最大滞后期 $p_{\max}$，然后令 $\hat{p}=p_{\max}$进行估计，并对最后一个滞后期系数的显著性进行 t 检验。如果接受该系数为 0，则令 $\hat{p}=p_{\max}-1$，重新进行估计，再对(新的)最后一个滞后期的系数进行 t 检验，如果显著，则停止；否则，令 $\hat{p}=p_{\max}-2$；依此类推。这个准则被称为“由大到小的序贯 t 规则”(general-to-specific sequential t rule)。

方法之二：使用信息准则，选择 $\hat{p}$ 使得 AIC，BIC 或 HQIC 最小化，分别记为 $\hat{p}_{\text{AIC}}$，$\hat{p}_{\text{BIC}}$与 $\hat{p}_{\text{HQIC}}$。比如，

$$\min_{p}\text{AIC}\equiv\ln(\boldsymbol{e}'\boldsymbol{e}/T)+\frac{2}{T}(p+1) \tag{20.19}$$

其中，$\boldsymbol{e}'\boldsymbol{e}$ 为残差平方和。可以证明，$\hat{p}_{\text{BIC}}$与 $\hat{p}_{\text{HQIC}}$是真实滞后阶数 p 的一致估计，但 $\hat{p}_{\text{AIC}}$在大样本中可能高估 p。然而，这几种信息准则在小样本中难分优劣。

在使用信息准则时，通常也需要设定最大滞后期 $p_{\max}$，否则可能得到很大(不合理)的 $\hat{p}$。事

实上，如果令 $\hat{p}=T-1$，则该模型只剩下一个方程：

$$y_T=\beta_0+\beta_1 y_{T-1}+\cdots+\beta_{T-1}y_1+\varepsilon_T \tag{20.20}$$

显然，对这个方程用 OLS，将得到唯一的残差 $e_T=0$。因此，残差平方和 $\boldsymbol{e}'\boldsymbol{e}=0$，$\ln(\boldsymbol{e}'\boldsymbol{e}/T)=-\infty$。也就是说，$\hat{p}=T-1$ 使得 AIC 最小化，但这显然不合理。

在实践中，可以结合以上两种方法来确定 $\hat{p}$。如果二者不一致，为了保守起见（即尽量避免遗漏变量偏差），可取二者滞后阶数的大者。

20.3 移动平均模型

另一类时间序列模型为“移动平均过程”（Moving Average Process，简记 MA）。记一阶移动平均过程为 MA(1)：

$$y_t=\mu+\varepsilon_t+\theta\varepsilon_{t-1} \tag{20.21}$$

其中，$\{\varepsilon_t\}$ 为白噪声，而 ε_t 的系数被标准化为 1[①]。由于 y_t 可以被看成是白噪声的移动平均[②]，故得名。考虑使用条件 MLE 估计。为此，假设 $\{\varepsilon_t\}$ 为独立同分布且服从正态分布 $N(0,\sigma_\varepsilon^2)$。如果已经知道 ε_{t-1}，则

$$y_t|\varepsilon_{t-1}\sim N(\mu+\theta\varepsilon_{t-1},\sigma_\varepsilon^2) \tag{20.22}$$

假设 $\varepsilon_0=0$，则可以知道 $\varepsilon_1=y_1-\mu$。给定 ε_1，则可以知道 $\varepsilon_2=y_2-\mu-\theta\varepsilon_1$。以此类推，则可以使用递推公式“$\varepsilon_t=y_t-\mu-\theta\varepsilon_{t-1}$”计算出全部 $\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T\}$。这样，在给定 $\varepsilon_0=0$ 的条件下，就可以写下样本数据 $\{y_1,y_2,\cdots,y_T\}$ 的似然函数，然后使用数值方法求解其最大化问题。

更一般地，对于 q 阶移动平均过程，记为 MA(q)：

$$y_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q} \tag{20.23}$$

也可以进行条件 MLE 估计，即在给定“$\varepsilon_0=\varepsilon_{-1}=\cdots=\varepsilon_{-q+1}=0$”的条件下，最大化样本数据的似然函数。

20.4 ARMA

为了使模型更好地拟合数据，可以将 AR(p)与 MA(q)结合起来，得到 ARMA(p,q)：

$$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q} \tag{20.24}$$

其中，$\{\varepsilon_t\}$ 为白噪声。在给定 $\{y_1,\cdots,y_p\}$ 与“$\varepsilon_0=\varepsilon_{-1}=\cdots=\varepsilon_{-q+1}=0$”的条件下，可以使用条件 MLE 来估计 ARMA(p,q)。为了估计 ARMA(p,q)，首先必须确定(p,q)，而经济理论通常不能提供这些信息，故只能根据数据来估计($\hat{p}$,$\hat{q}$)。

在实践中，常常先考察数据的自相关函数（ACF）与偏自相关函数（PACF），以判断是否存在 $p=0$ 或 $q=0$ 的情形。如果 $p=0$，则为 MA(q)：

$$y_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q} \tag{20.25}$$

显然，如果 $j>q$，则 $\mathrm{Cov}(y_t,y_{t-j})=0$，因为产生 y_t 的扰动项 $\{\varepsilon_t,\varepsilon_{t-1},\cdots,\varepsilon_{t-q}\}$ 与产生 y_{t-j} 的扰动

① 如果 ε_t 的系数不为 1，可以在方程两边同时除以该系数。

② 其实是一种广义的“移动平均”，因为并不要求权重之和为 1，即($1+\theta$)不一定等于 1。

项$\{\varepsilon_{t-j},\varepsilon_{t-j-1},\cdots,\varepsilon_{t-j-q}\}$没有重叠的部分(无交集)。这意味着,对于MA($q$)模型,ACF函数在$j>q$时都等于零,即出现"截尾"。另一方面,对于AR(1),我们知道其ACF函数呈指数衰减,被称为"拖尾"(tails off to zero),但不存在截尾。

如果$q=0$,则ARMA(p,q)简化为AR(p)模型:

$$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\varepsilon_t \tag{20.26}$$

假设真实模型为AR(p),却用OLS来估计AR($p+1$),即$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\beta_{p+1}y_{t-p-1}+\varepsilon_t$,则$\underset{T\to\infty}{\text{plim}}\,\hat{\beta}_{p+1}=0$,因为$\beta_{p+1}=0$。而$\hat{\beta}_{p+1}$正是对$(p+1)$阶偏自相关函数的估计。由此可知,对于AR($p$)模型,PACF函数在$j>p$时都等于零,即出现截尾。另一方面,对于MA($q$)模型,PACF函数逐渐衰减,即拖尾,但不存在截尾。

综上所述,对于AR(p)模型,其ACF函数拖尾,而PACF函数截尾。如果出现这种情形,则可判断其为AR(p),不包含移动平均的部分。

另一方面,对于MA(q)模型,其ACF函数截尾,而PACF函数拖尾。如果出现这种情形,则可判断其为MA(q),不包含自回归的部分。

如果以上两种情形均不符合,即ACF函数与PACF函数都拖尾[①],则要考虑一般的ARMA(p,q)模型,其中p,q均不为零。时间序列分析的鼻祖,Box,Jenkins and Reinsel(1994)认为,对大多数情况,$p\leqslant 2$与$q\leqslant 2$就足够了。当然,为了保险起见,可以让$p_{\max}$与$q_{\max}$更大些。具体如何确定p,q,可以依据信息准则或由大到小的序贯t规则。

在估计完模型之后,仍然需要进行一些诊断性分析(diagnostic checking),以确定ARMA(p,q)模型的假定是否成立。其中,最重要的假定是,扰动项$\{\varepsilon_t\}$为白噪声。如果模型过小(inadequate),即$\hat{p}<p$或$\hat{q}<q$,则相当于遗漏解释变量。这些被遗漏的解释变量被纳入扰动项中,导致扰动项出现自相关,不再是白噪声。为此,可以使用Q检验来检验模型的残差是否存在自相关。如果残差存在自相关,则应考虑使模型更大些(增加自回归或移动平均的阶数),重新对模型进行估计,然后再检验新模型的残差是否为白噪声,如此反复,直至确认残差为白噪声。

20.5 自回归分布滞后模型

在自回归模型中,也可以引入其他变量来构成"自回归分布滞后模型"(Autoregressive Distributed Lag Model,简记ADL(p,q)):

$$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\gamma_1 x_{t-1}+\cdots+\gamma_q x_{t-q}+\varepsilon_t \tag{20.27}$$

比如,记通货膨胀率为π_t,失业率为un_t,则一个可能的预测模型为

$$\Delta\pi_t=\beta_0+\beta_1\Delta\pi_{t-1}+\beta_2\Delta\pi_{t-2}+\beta_3\Delta\pi_{t-3}+\beta_4\Delta\pi_{t-4}+\gamma_1 un_{t-1}+\varepsilon_t \tag{20.28}$$

其中,$\Delta\pi_t$相当于"y_t"。这个ADL(4,1)反映的是宏观经济学中的经验菲利普斯曲线(empirical Phillip Curve)。如果$\gamma_1<0$,则失业率越低,物价越有上涨的压力。另一方面,通胀的调整也受到过去通胀变化的滞后作用。

也可以引入更多的解释变量。比如,共有K个解释变量$\{x_{1t},\cdots,x_{Kt}\}$,其中第j个解释变量x_{jt}共有q_j个滞后值被包括在模型中,$j=1,\cdots,K$。

---

① 不太可能二者均截尾。

$$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\gamma_{11}x_{1,t-1}+\cdots+\gamma_{1q_1}x_{1,t-q_1}+\cdots+\gamma_{K1}x_{K,t-1}+\cdots+\gamma_{Kq_K}x_{K,t-q_K}+\varepsilon_t \tag{20.29}$$

如果自回归分布滞后模型满足以下假定,则万事大吉,可以用 OLS 来估计它。

(i) $E(\varepsilon_t|y_{t-1},y_{t-2},\cdots,x_{1,t-1},x_{1,t-2},\cdots,x_{K,t-1},x_{K,t-2},\cdots)=0$。这个假定类似于第 3 章的严格外生性假设,它意味着扰动项 ε_t 与所有解释变量的整个历史全部无关。这保证了对滞后期数 $(p,q_1,\cdots,q_K)$ 的设定是正确的。如果滞后期数的设定不正确,比如,真实模型还应该包括 $y_{t-(p+1)}$,但该项 $\beta_{p+1}y_{t-(p+1)}$ 却被纳入扰动项 ε_t 中,则扰动项 ε_t 便与解释变量相关,导致 OLS 不一致。

(ii) $\{y_t,x_{1t},\cdots,x_{Kt}\}$ 为渐近独立的平稳序列。

(iii) $\{y_t,x_{1t},\cdots,x_{Kt}\}$ 有非零的有限四阶矩。

(iv) 解释变量无完全多重共线性。

对滞后期数的选择可以使用信息准则(最小化 AIC 或 BIC),或使用 t,F 检验来检验最后一期系数的显著性。更一般地,可以在 ARMA 模型中引入其他变量,称为“ARMAX”模型。

20.6 ARMA 模型的 Stata 命令及实例

1. 自相关与偏自相关

```
corrgram y,lags(#)  (计算第 1 至第#阶 ACF 与 PACF,以及检验自相关的 Q 统计量)
ac y,lags(#)  (将 ACF 画成自相关图,并给出置信区间)
pac y,lags(#)  (将 PACF 画成偏自相关图,并给出置信区间)
```

2. ARMA

```
arima y,ar(1/#) ma(1/#)
```

其中,选择项“ar(1/#)”表示第 1 至第#阶自回归,而“ma(#)”表示第 1 至第#阶移动平均。ARMA 的另一等价命令格式为

```
arima y,arima(#p,#d,#q)
```

其中,“#p”表示自回归的阶数,“#q”表示移动平均的阶数,而“#d”表示原序列 $\{y_t\}$ 需要经过几次差分才是平稳过程(参见第 21 章)。下面检验其残差项是否存在自相关:

```
predict e1,res  (计算残差,并将其命名为 e1)
corrgram e1,lags(#)  (检验残差是否存在第 1 至第#阶自相关的 Q 检验)
```

3. ADL 与 ARMAX

ARMAX 的 Stata 命令为

```
arima y x1 x2 x3,ar(#) ma(#)
```

下面以数据集 pe. dta 为例[①]。该数据集的主要变量 logpe 为 1871—2002 年美国标准普尔股指(S&P)的市盈率(price earning ratio)的对数。由于 logpe 为非平稳序列(参见第 21 章),故对其差分建立 ARMA 模型。首先,定义其一阶差分为“d_logpe”:

```
. use pe.dta,clear
. tsset year
```

① 此例来自 Verbeek(2004)。

```
        time variable:  year, 1871 to 2002
                delta:  1 unit
```

```
. g d_logpe = d.logpe
```

计算其前 10 阶自相关与偏自相关系数：

```
. corrgram d_logpe,lags(10)
```

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.0651	0.0652	.56741	0.4513		
2	-0.1661	-0.1723	4.2951	0.1168		
3	-0.0611	-0.0393	4.8031	0.1868		
4	-0.1796	-0.2117	9.2305	0.0556		
5	-0.0530	-0.0525	9.6196	0.0868		
6	0.0897	0.0232	10.741	0.0967		
7	0.1288	0.0927	13.07	0.0704		
8	-0.0050	-0.0425	13.074	0.1093		
9	-0.0163	0.0132	13.112	0.1576		
10	0.0583	0.0917	13.601	0.1920		

上表显示，直至第 4 阶的 Q 统计量较为显著（p 值为 0.055 6）。更直观地，可以考察自相关图（结果如图 20.1）与偏自相关图（结果如图 20.2）。

```
. ac d_logpe,lags(10)
```

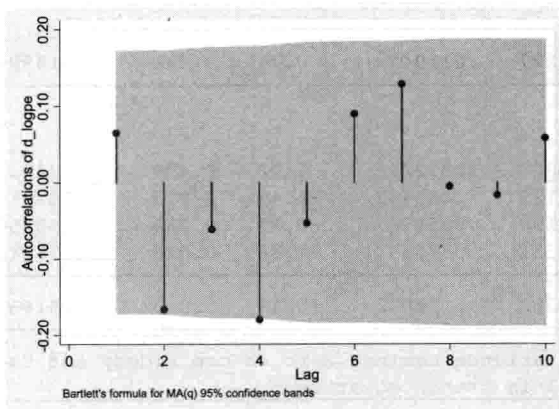


图 20.1 自相关图

```
. pac d_logpe,lags(10)
```

以上两图显示，第 4 阶自相关与偏自相关系数均在 5% 水平显著地不为 0（落在 95% 的置信区间之外），而可以认为 4 阶以上的自相关与偏自相关系数为 0。由于自相关系数与偏自相关系数均存在断尾，故分别考虑 AR(4)与 MA(4)模型。首先，估计 AR(4)模型：

```
. arima d_logpe,ar(1/4) nolog
```

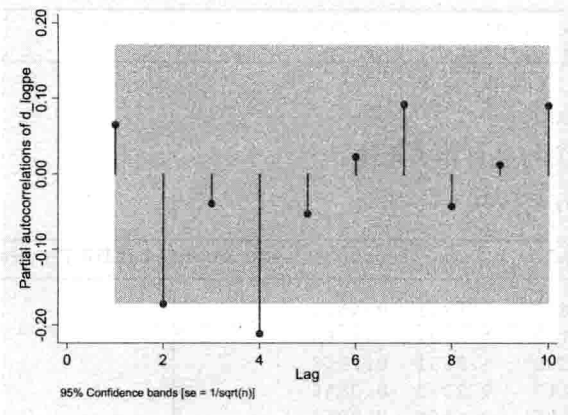


图 20.2 偏自相关图

```
ARIMA regression

Sample:  1872 - 2002                            Number of obs      =       131
                                                Wald chi2(4)       =     10.14
Log likelihood =  43.35448                      Prob > chi2        =    0.0382

------------------------------------------------------------------------------
             |                 OPG
     d_logpe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
d_logpe      |
       _cons |   .0084227   .0119439     0.71   0.481    -.0149868    .0318322
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L1. |   .0601002   .0832216     0.72   0.470    -.1030112    .2232117
         L2. |  -.2029129    .081432    -2.49   0.013    -.3625167   -.0433091
         L3. |  -.0228198   .0854787    -0.27   0.789    -.1903549    .1447153
         L4. |   -.206651   .0875637    -2.36   0.018    -.3782728   -.0350293
-------------+----------------------------------------------------------------
      /sigma |   .1736352   .0112072    15.49   0.000     .1516695     .195601
------------------------------------------------------------------------------
Note: The test of the variance against zero is one sided, and the two-sided
    confidence interval is truncated at zero.
```

计算信息准则：

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	131	.	43.35448	6	-74.70897	-57.45779

Note: N=Obs used in calculating BIC; see **[R] BIC note**

检验其残差项是否存在自相关：

```
. predict e1,res
(1 missing value generated)
. corrgram e1,lags(10)
```

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	-0.0100	-0.0100	.01327	0.9083		
2	-0.0003	-0.0004	.01329	0.9934		
3	0.0103	0.0106	.0276	0.9988		
4	-0.0033	-0.0031	.0291	0.9999		
5	-0.0328	-0.0341	.17765	0.9993		
6	0.0256	0.0260	.26879	0.9996		
7	0.0912	0.0965	1.4376	0.9844		
8	-0.0211	-0.0190	1.5005	0.9927		
9	-0.0429	-0.0481	1.7638	0.9947		
10	0.0783	0.0801	2.6469	0.9886		

结果表明,可以接受残差项“无自相关”的原假设(检验至第 10 阶自相关)。

其次,估计 MA(4)模型:

```
. arima d_logpe,ma(1/4) nolog
```

```
ARIMA regression

Sample:  1872 - 2002                            Number of obs      =       131
                                                Wald chi2(4)       =      8.78
Log likelihood =  42.69439                      Prob > chi2        =    0.0669
```

d_logpe	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
d_logpe						
_cons	.0080427	.0111621	0.72	0.471	-.0138347	.02992
ARMA						
ma						
L1.	.0478866	.0841641	0.57	0.569	-.1170719	.2128451
L2.	-.1875636	.0868965	-2.16	0.031	-.3578777	-.0172496
L3.	-.0400797	.0922776	-0.43	0.664	-.2209405	.1407812
L4.	-.1462202	.0904237	-1.62	0.106	-.3234474	.0310071
/sigma	.1745448	.0108632	16.07	0.000	.1532533	.1958363

```
Note: The test of the variance against zero is one sided, and the two-sided
      confidence interval is truncated at zero.
```

下面计算信息准则:

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	131	.	42.69439	6	-73.38878	-56.1376

Note: N=Obs used in calculating BIC; see **[R] BIC note**

检验其残差项是否存在自相关:

```
. predict e2,res
(1 missing value generated)
. corrgram e2,lags(10)
```

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.0082	0.0082	.00898	0.9245		
2	-0.0137	-0.0139	.03448	0.9829		
3	-0.0033	-0.0031	.03595	0.9982		
4	-0.0276	-0.0282	.14015	0.9977		
5	-0.0502	-0.0522	.48898	0.9925		
6	0.0873	0.0908	1.5509	0.9560		
7	0.0956	0.0986	2.8338	0.8999		
8	0.0024	0.0025	2.8346	0.9443		
9	-0.0390	-0.0426	3.0521	0.9622		
10	0.0687	0.0774	3.7308	0.9587		

上表显示,MA(4)的残差项也不存在自相关。根据信息准则,AR(4)模型似乎略优于MA(4)模型。然而,两个模型中的第 1 阶与第 3 阶系数均不显著。为此,考虑更为简洁的模型,将 AR(4)与 MA(4)模型中的第 1 阶与第 3 阶变量略去:

```
. arima d_logpe,ar(2 4) nolog
```

```
ARIMA regression

Sample:  1872 - 2002                            Number of obs      =       131
                                                Wald chi2(2)       =     11.61
Log likelihood =  43.02852                      Prob > chi2        =    0.0030
```

d_logpe	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
d_logpe						
_cons	.0083801	.0114639	0.73	0.465	-.0140888	.0308489
ARMA						
ar						
L2.	-.201582	.0788054	-2.56	0.011	-.3560378	-.0471263
L4.	-.2119179	.0841825	-2.52	0.012	-.3769126	-.0469231
/sigma	.1740682	.0110923	15.69	0.000	.1523277	.1958086

```
Note: The test of the variance against zero is one sided, and the two-sided
    confidence interval is truncated at zero.
```

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	131	.	43.02852	4	-78.05704	-66.55625

Note: N=Obs used in calculating BIC; see **[R] BIC note**

```
. arima d_logpe,ma(2 4) nolog
```

```
ARIMA regression

Sample:  1872 - 2002                            Number of obs      =       131
                                                Wald chi2(2)       =      9.66
Log likelihood =   42.4702                      Prob > chi2        =    0.0080
```

d_logpe	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
d_logpe						
_cons	.0080273	.0108818	0.74	0.461	-.0133006	.0293552
ARMA						
ma						
L2.	-.1809963	.0866252	-2.09	0.037	-.3507785	-.011214
L4.	-.1563501	.0865606	-1.81	0.071	-.3260057	.0133055
/sigma	.1748415	.0107209	16.31	0.000	.1538288	.1958541

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	131	.	42.4702	4	-76.9404	-65.43961

Note: N=Obs used in calculating BIC; see **[R] BIC note**

上表显示，略去第 1 阶与第 3 阶变量的简洁模型，其信息准则比完整模型更优。根据这个标准，似乎略去第 1 阶与第 3 阶变量的 AR(4)模型最为合适。

20.7 误差修正模型

ADL 是一种动态模型。从经济理论而言，相关的变量之间可能存在长期的均衡关系，而变量的短期变动则是向着这个长期均衡关系的部分调整。“误差修正模型”(Error Correction Model，简记 ECM)正是这一思想在计量经济学的体现。

考虑最简单的 AR(1)模型：

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \tag{20.30}$$

其中，$|\beta_1| < 1$，故 $\{y_t\}$ 为平稳过程。对方程(20.30)两边求期望，并令长期均衡值 $y^* = \mathrm{E}(y_t) = \mathrm{E}(y_{t-1})$，则可得 $y^* = \dfrac{\beta_0}{1-\beta_1}$。将 $\beta_0 = (1-\beta_1)y^*$ 代入方程(20.30)，并在方程两边同时减去 y_{t-1} 可得

$$\Delta y_t = (1-\beta_1)y^* - (1-\beta_1)y_{t-1} + \varepsilon_t \tag{20.31}$$

$$\Delta y_t = \underbrace{(1-\beta_1)(y^* - y_{t-1})}_{\text{error correction}} + \varepsilon_t \tag{20.32}$$

这就是 AR(1)的误差修正模型，它将 Δy_t 表达为对长期均衡的偏离($y^* - y_{t-1}$)的部分调整(即误

差修正)加上扰动项。

考虑略微复杂的 ADL 模型:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + \varepsilon_t \tag{20.33}$$

其中,$|\beta_1| < 1$。假设经济理论认为(y,x)之间存在长期均衡关系 $y = \phi + \theta x$。对方程(20.33)两边求期望,并令 $y^* = \mathrm{E}(y_t) = \mathrm{E}(y_{t-1})$,$x^* = \mathrm{E}(x_t) = \mathrm{E}(x_{t-1})$,可得

$$y^* = \beta_0 + \beta_1 y^* + \gamma_0 x^* + \gamma_1 x^* \tag{20.34}$$

$$(1-\beta_1) y^* = \beta_0 + (\gamma_0 + \gamma_1) x^* \tag{20.35}$$

$$y^* = \frac{\beta_0}{(1-\beta_1)} + \frac{(\gamma_0 + \gamma_1)}{(1-\beta_1)} x^* \tag{20.36}$$

由此可得,$\phi = \dfrac{\beta_0}{1-\beta_1}$,$\theta = \dfrac{\gamma_0 + \gamma_1}{1-\beta_1}$。其中,$\theta = \dfrac{\gamma_0 + \gamma_1}{1-\beta_1}$称为"长期乘数"(long-run multiplier),它衡量的是当 x 永久性地变化一单位时,将会导致 y 的永久性变化幅度。显然,$\beta_0 = (1-\beta_1)\phi$,$\gamma_0 + \gamma_1 = (1-\beta_1)\theta$。对方程(20.33)作如下变形,即可得到误差修正模型。在方程(20.33)两边同时减去 y_{t-1},并在方程右边加上、减去 $\gamma_0 x_{t-1}$,可得

$$\Delta y_t = \beta_0 - (1-\beta_1) y_{t-1} + \gamma_0 (x_t - x_{t-1}) + (\gamma_0 + \gamma_1) x_{t-1} + \varepsilon_t \tag{20.37}$$

代入 $\beta_0 = (1-\beta_1)\phi$,可得

$$\Delta y_t = (1-\beta_1)\phi - (1-\beta_1) y_{t-1} + \gamma_0 \Delta x_t + (\gamma_0 + \gamma_1) x_{t-1} + \varepsilon_t \tag{20.38}$$

代入 $\gamma_0 + \gamma_1 = (1-\beta_1)\theta$,则有

$$\Delta y_t = \gamma_0 \Delta x_t + \underbrace{(\beta_1 - 1)(y_{t-1} - \phi - \theta x_{t-1})}_{\text{error correction}} + \varepsilon_t \tag{20.39}$$

这就是误差修正的形式。其中,$(\beta_1 - 1)(y_{t-1} - \phi - \theta x_{t-1})$称为"误差修正项"(error correction term)。参数$\{\phi, \theta\}$称为"长期参数",而参数$\{\gamma_0, \beta_1 - 1\}$称为"短期参数"。误差修正模型的形式看似非线性,但仍是线性回归,即可以把它还原为 ADL 模型,然后用 OLS 来估计。一般来说,ADL 模型都可以变换成 ECM 模型(参见习题)。误差修正模型的优点是,经济含义十分明确,而且可以分别考察长期效应与短期效应。

20.8 MA(∞)与滞后算子

我们知道,对于 MA(q),其自相关函数(ACF)存在截尾。如果希望其自相关系数永远不为0,可以考虑 MA(∞):

$$y_t = \mu + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \tag{20.40}$$

其中,$\theta_0 = 1$。问题是,这个无穷多个随机变量之和,能收敛到某个随机变量吗?一个常用的充分条件是,序列$\{\theta_j\}_{j=0}^{\infty}$为"绝对值可加总"(Absolutely Summable,简记 AS),即 $\sum_{j=0}^{\infty} |\theta_j| < \infty$(有限)。在 AS 的条件下,不仅 MA(∞)有定义,而且还是一个弱平稳过程(可通过计算二阶矩来验证);更进一步,如果$\{\varepsilon_t\}$为独立同分布的,则 MA(∞)是一个严格平稳过程。虽然样本容量 T 通常有限,而我们也不可能追溯到无穷远的过去,但 MA(∞)在理论上仍然有着重要意义。

为了计算方便,在时间序列分析中常引入"滞后算子"(lag operator 或 backshift operator)。

定义 $Ly_t = y_{t-1}$,$L^2 y_t = L(Ly_t) = y_{t-2}$,$\cdots$,$L^p y_t = y_{t-p}$。特别地,$L^0 y_t = 1 \cdot y_t = y_t$。

显然,滞后算子的运算相当于幂函数。比如,两个滞后算子相乘满足交换律,其乘积等于指

数相加，即 $L^p \cdot L^q = L^{p+q}$。由于 $\Delta y_t = y_t - y_{t-1} = (1-L)y_t$，故差分算子 $\Delta \equiv 1 - L$。

对于 AR(p)，$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t$，可以用滞后算子简洁地表示：

$$y_t = \beta_0 + \beta_1 L y_t + \cdots + \beta_p L^p y_t + \varepsilon_t \tag{20.41}$$

移项后可得

$$y_t - \beta_1 L y_t - \cdots - \beta_p L^p y_t = \beta_0 + \varepsilon_t \tag{20.42}$$

向右提取公因子 y_t

$$(1 - \beta_1 L - \cdots - \beta_p L^p) y_t = \beta_0 + \varepsilon_t \tag{20.43}$$

定义“滞后多项式”(lag polynomial)$\beta(L) \equiv 1 - \beta_1 L - \cdots - \beta_p L^p$，则

$$\beta(L) y_t = \beta_0 + \varepsilon_t \tag{20.44}$$

上式容易让人产生一种愿望，即如果存在 $\beta(L)^{-1}$，则可以在两边同时左乘 $\beta(L)^{-1}$，从而得到 AR(p)的一个“解”。

定义 对于任意一个实数序列$(\alpha_0, \alpha_1, \alpha_2, \cdots)$，定义其对应的“**滤波**”(filter)为 $\alpha(L) = \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots$。因此，滤波的实质就是一个(无穷多项)的滞后多项式。

命题 假设$\{x_t\}$为弱平稳过程，序列$\{\theta_j\}_{j=0}^{\infty}$为绝对值可加总(AS)，则 $y_t = \alpha + \sum_{j=0}^{\infty} \theta_j x_{t-j} = \alpha + \sum_{j=0}^{\infty} \theta_j L^j x_t$ 有定义，且为弱平稳。进一步，如果$\{x_t\}$为独立同分布，则$\{y_t\}$为严格平稳过程。

这个命题告诉我们，弱平稳过程经过 AS 滤波的作用后，仍为弱平稳过程。

定义 对于两个滤波“$\alpha(L) = \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots$”与“$\beta(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots$”，定义其乘积为

$$\begin{aligned} \delta(L) \equiv \alpha(L)\beta(L) &\equiv (\alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots)(\beta_0 + \beta_1 L + \beta_2 L^2 + \cdots) \\ &= \alpha_0\beta_0 + (\alpha_0\beta_1 + \alpha_1\beta_0) L + (\alpha_2\beta_0 + \alpha_1\beta_1 + \alpha_0\beta_2) L^2 + \cdots \end{aligned} \tag{20.45}$$

显然，滤波的乘积满足交换律。我们最感兴趣的情形是 $\delta(L) = 1$，即令上式第二行中的常数项 $\alpha_0\beta_0$ 为 1，而其余各项的系数均为 0。此时，我们称 $\beta(L)$ 为 $\alpha(L)$ 的“逆”(inverse)，并记为 $\alpha(L)^{-1}$。只要 $\alpha_0 \neq 0$，则 $\alpha(L)^{-1}$都有定义且唯一，因为可以得到满足 $\alpha(L)\beta(L) = 1$ 的唯一解 $(\beta_0, \beta_1, \beta_2, \cdots)$，即 $\beta_0 = 1/\alpha_0$，$\beta_1 = -\alpha_1\beta_0/\alpha_0$，等等。

例 $(1-L)^{-1} = 1 + L + L^2 + L^3 + \cdots$

证明：显然，$(1-L)(1+L+L^2+L^3+\cdots) = 1$。需要注意的是，$(1-L)^{-1}$不再是 AS。

例 $(1-\beta L)^{-1} = 1 + \beta L + \beta^2 L^2 + \beta^3 L^3 + \cdots$

证明：只要将上例中的“L”换成“βL”即可。如果 $|\beta| < 1$，则 $(1-\beta L)^{-1}$为 AS。

命题 对于 $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$，假设 $|\beta_1| < 1$，则此 AR(1)是 MA(∞)。

证明：

方法一(迭代法)

$$\begin{aligned} y_t &= \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \\ &= \beta_0 + \beta_1(\beta_0 + \beta_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= (\beta_0 + \beta_0\beta_1) + \beta_1^2 y_{t-2} + \beta_1\varepsilon_{t-1} + \varepsilon_t \\ &= (\beta_0 + \beta_0\beta_1) + \beta_1^2(\beta_0 + \beta_1 y_{t-3} + \varepsilon_{t-2}) + \beta_1\varepsilon_{t-1} + \varepsilon_t \\ &= \beta_0(1 + \beta_1 + \beta_1^2) + \beta_1^3 y_{t-3} + \beta_1^2\varepsilon_{t-2} + \beta_1\varepsilon_{t-1} + \varepsilon_t \\ &= \cdots \\ &= \beta_0(1 + \beta_1 + \beta_1^2 + \cdots) + \varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_1^2\varepsilon_{t-2} + \beta_1^3\varepsilon_{t-3} + \cdots \end{aligned}$$

$$=\frac{\beta_0}{1-\beta_1}+\varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_1^2\varepsilon_{t-2}+\beta_1^3\varepsilon_{t-3}+\cdots$$

方法二(滤波求逆法)

由于$(1-\beta_1 L)y_t=\beta_0+\varepsilon_t$,故

$$\begin{aligned}y_t&=(1-\beta_1 L)^{-1}(\beta_0+\varepsilon_t)\\&=(1+\beta L+\beta^2L^2+\cdots)\beta_0+(1+\beta L+\beta^2L^2+\cdots)\varepsilon_t\\&=\beta_0(1+\beta_1+\beta_1^2+\cdots)+\varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_1^2\varepsilon_{t-2}+\beta_1^3\varepsilon_{t-3}+\cdots\\&=\frac{\beta_0}{1-\beta_1}+\varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_1^2\varepsilon_{t-2}+\beta_1^3\varepsilon_{t-3}+\cdots\end{aligned}$$

因此,可以将平稳的AR(1)看成是过去所有扰动项的总效应之和,而且离现在越远的扰动项其影响力呈几何级数递减。从证明方法二,可以看出滤波求逆方法的威力。从AR(1)的MA(∞)表达式可以看出

$$IRF(j)\equiv\frac{\partial y_{t+j}}{\partial\varepsilon_t}=\beta_1^j \tag{20.46}$$

$(\partial y_{t+j}/\partial\varepsilon_t)$表示的是,当第$t$期的扰动项$\varepsilon_t$变化1单位时(而其他期的扰动项均不变),对相隔$j$期的$y_{t+j}$的影响,被称为"动态乘子"(dynamic multiplier)。显然,动态乘子与绝对时间t无关,而是时间间隔j的函数。将$(\partial y_{t+j}/\partial\varepsilon_t)$视为$j$的函数,则称为"脉冲响应函数"(Impulse Response Function,简记IRF),它刻画的是y_{t+j}对ε_t的1单位脉冲(impulse)的响应(response)。将$(j,\partial y_{t+j}/\partial\varepsilon_t)$画图,就可以得到对IRF的直观认识。有时也考察$\varepsilon_t$从第0期(即当前期)到第$k$期对$y_{t+j}$的累积效应:

$$CIRF(k)\equiv\sum_{j=0}^{k}\frac{\partial y_{t+j}}{\partial\varepsilon_t} \tag{20.47}$$

上式称为"累积脉冲响应函数"(Cumulative Impulse Response Function,简记CIRF)。类似地,AR(p)也是MA(∞)。将AR(p)"$y_t=\beta_0+\beta_1y_{t-1}+\cdots+\beta_py_{t-p}+\varepsilon_t$"写为"$\beta(L)y_t=\beta_0+\varepsilon_t$",如果$\beta(L)^{-1}$为AS,则

$$y_t=\beta(L)^{-1}\beta_0+\beta(L)^{-1}\varepsilon_t \tag{20.48}$$

显然,这是MA(∞)的形式。对于AR(p)的脉冲响应函数,可以通过计算$\beta(L)^{-1}$得到。另一方法是,直接对AR(p)进行如下模拟(simulation):令$y_{-1}=y_{-2}=\cdots=y_{-p}=0$,$\varepsilon_0=1$,而$\varepsilon_t=0(\forall t\neq 0)$,使用AR($p$)方程"$y_t=\beta_0+\beta_1y_{t-1}+\cdots+\beta_py_{t-p}+\varepsilon_t$"(令$t=0$)可得$y_0$。然后,将$\{y_0,y_{-1},\cdots,y_{-p+1}\}$代入AR($p$)方程(令$t=1$)可得$y_1$。依此类推,可以计算出$y_j$,就是相隔$j$期的动态乘子$(\partial y_{t+j}/\partial\varepsilon_t)$。

同样道理,更一般的ARMA(p,q)也是MA(∞):

$$y_t=\beta_0+\beta_1y_{t-1}+\cdots+\beta_py_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q} \tag{20.49}$$

$$y_t-\beta_1Ly_t-\cdots-\beta_pL^py_t=\beta_0+\varepsilon_t+\theta_1L\varepsilon_t+\cdots+\theta_qL^q\varepsilon_t \tag{20.50}$$

$$\beta(L)y_t=\beta_0+\theta(L)\varepsilon_t \tag{20.51}$$

其中,$\theta(L)\equiv1+\theta_1L+\cdots+\theta_qL^q$。假设$\beta(L)^{-1}$为AS,则

$$y_t=\beta(L)^{-1}\beta_0+\beta(L)^{-1}\theta(L)\varepsilon_t \tag{20.52}$$

这还是MA(∞)的形式。总之,ARMA(p,q)在本质上都可以归结为MA(∞)。

20.9 向量自回归过程

我们常常同时关心几个经济变量的预测，比如 GDP 增长率与失业率。一种方法是用单变量时间序列的方法对每个变量分别作预测。另一种方法则是将这些变量放在一起，作为一个系统来预测，以使得预测相互自洽（mutually consistent），这称为“多变量时间序列”（multivariate time series）。由 Sims（1980）所提倡的“向量自回归”（Vector Autoregression，简记 VAR）正是这样一种方法。

假设有两个时间序列变量$\{y_{1t},y_{2t}\}$，分别作为两个回归方程的被解释变量；而解释变量为这两个变量的 p 阶滞后值，构成一个二元（bivariate）的 VAR(p) 系统：

$$\begin{cases} y_{1t}=\beta_{10}+\beta_{11}y_{1,t-1}+\cdots+\beta_{1p}y_{1,t-p}+\gamma_{11}y_{2,t-1}+\cdots+\gamma_{1p}y_{2,t-p}+\varepsilon_{1t} \\ y_{2t}=\beta_{20}+\beta_{21}y_{1,t-1}+\cdots+\beta_{2p}y_{1,t-p}+\gamma_{21}y_{2,t-1}+\cdots+\gamma_{2p}y_{2,t-p}+\varepsilon_{2t} \end{cases} \tag{20.53}$$

其中，$\{\varepsilon_{1t}\}$ 与 $\{\varepsilon_{2t}\}$ 均为白噪声过程（故不存在自相关），但允许两个方程的扰动项之间存在“同期相关性”（contemporaneous correlation）：

$$\mathrm{Cov}(\varepsilon_{1t},\varepsilon_{2s})=\begin{cases}\sigma_{12}, & \text{若 } t=s \\ 0, & \text{其他}\end{cases} \tag{20.54}$$

注意到上面两个方程的解释变量完全一样。将两个方程写在一起：

$$\begin{pmatrix}y_{1t}\\y_{2t}\end{pmatrix}=\begin{pmatrix}\beta_{10}\\\beta_{20}\end{pmatrix}+\begin{pmatrix}\beta_{11}\\\beta_{21}\end{pmatrix}y_{1,t-1}+\cdots+\begin{pmatrix}\beta_{1p}\\\beta_{2p}\end{pmatrix}y_{1,t-p}+\begin{pmatrix}\gamma_{11}\\\gamma_{21}\end{pmatrix}y_{2,t-1}+\cdots+\begin{pmatrix}\gamma_{1p}\\\gamma_{2p}\end{pmatrix}y_{2,t-p}+\begin{pmatrix}\varepsilon_{1t}\\\varepsilon_{2t}\end{pmatrix} \tag{20.55}$$

将同期变量写成列向量，并把相应的系数合并为矩阵，

$$\begin{pmatrix}y_{1t}\\y_{2t}\end{pmatrix}=\begin{pmatrix}\beta_{10}\\\beta_{20}\end{pmatrix}+\begin{pmatrix}\beta_{11}&\gamma_{11}\\\beta_{21}&\gamma_{21}\end{pmatrix}\begin{pmatrix}y_{1,t-1}\\y_{2,t-1}\end{pmatrix}+\cdots+\begin{pmatrix}\beta_{1p}&\gamma_{1p}\\\beta_{2p}&\gamma_{2p}\end{pmatrix}\begin{pmatrix}y_{1,t-p}\\y_{2,t-p}\end{pmatrix}+\begin{pmatrix}\varepsilon_{1t}\\\varepsilon_{2t}\end{pmatrix} \tag{20.56}$$

记 $\boldsymbol{y}_t\equiv\begin{pmatrix}y_{1t}\\y_{2t}\end{pmatrix}$，$\boldsymbol{\varepsilon}_t\equiv\begin{pmatrix}\varepsilon_{1t}\\\varepsilon_{2t}\end{pmatrix}$，则有

$$\boldsymbol{y}_t=\underbrace{\begin{pmatrix}\beta_{10}\\\beta_{20}\end{pmatrix}}_{\boldsymbol{\Gamma}_0}+\underbrace{\begin{pmatrix}\beta_{11}&\gamma_{11}\\\beta_{21}&\gamma_{21}\end{pmatrix}}_{\boldsymbol{\Gamma}_1}\boldsymbol{y}_{t-1}+\cdots+\underbrace{\begin{pmatrix}\beta_{1p}&\gamma_{1p}\\\beta_{2p}&\gamma_{2p}\end{pmatrix}}_{\boldsymbol{\Gamma}_p}\boldsymbol{y}_{t-p}+\boldsymbol{\varepsilon}_t \tag{20.57}$$

定义相应的系数矩阵为 $\boldsymbol{\Gamma}_0,\boldsymbol{\Gamma}_1,\cdots,\boldsymbol{\Gamma}_p$，可得

$$\boldsymbol{y}_t=\boldsymbol{\Gamma}_0+\boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p}+\boldsymbol{\varepsilon}_t \tag{20.58}$$

这个形式与 AR(p) 很相似，故名“VAR(p)”。其中，$\{\boldsymbol{\varepsilon}_t\}$ 是一维白噪声过程的推广，称为“向量白噪声过程”（vector white noise process）。

定义 向量白噪声过程$\{\boldsymbol{\varepsilon}_t\}$是一个弱平稳的随机过程，满足以下条件：

$$\mathrm{E}(\boldsymbol{\varepsilon}_t)=\mathbf{0},\quad \mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')=\boldsymbol{\Sigma},\quad \mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s')=\mathbf{0}\quad(t\neq s) \tag{20.59}$$

其中，$\boldsymbol{\Sigma}$ 为正定矩阵（positive definite）。由于不要求 $\boldsymbol{\Sigma}$ 为对角矩阵，故向量白噪声过程的各个分量之间可以存在同期相关。有时也把白噪声过程称为“新息过程”（innovation process）。

由于 VAR(p) 系统中的解释变量$\{\boldsymbol{y}_{t-1},\cdots,\boldsymbol{y}_{t-p}\}$依赖于$\{\boldsymbol{\varepsilon}_{t-1},\boldsymbol{\varepsilon}_{t-2},\cdots\}$，而 $\boldsymbol{\varepsilon}_t$ 与$\{\boldsymbol{\varepsilon}_{t-1},\boldsymbol{\varepsilon}_{t-2},\cdots\}$不相关，故可视所有解释变量为前定变量（predetermined），与当期扰动项 $\boldsymbol{\varepsilon}_t$ 不相关，故

可以用OLS对每个方程分别进行估计。由此可见,“扰动项$\{\boldsymbol{\varepsilon}_t\}$不存在自相关”是保证OLS一致性的重要条件。如果假设扰动项服从正态分布,则可以推导出条件MLE估计量。条件MLE等价于对每个方程分别用OLS进行估计。由于方程(20.58)右边的解释变量均为滞后变量$\{\boldsymbol{y}_{t-1},\cdots,\boldsymbol{y}_{t-p}\}$,不包含同期变量$\boldsymbol{y}_t$,故也称为“简化式VAR”(reduced-form VAR)。如果解释变量包括同期变量$\boldsymbol{y}_t$,则称为“结构VAR”(structural VAR),需要附加额外的约束条件以解决内生性问题,参见第24章。

在进行VAR建模时,需要确定变量的滞后阶数,以及VAR系统中包含几个变量。

1. 滞后阶数的选择

方法之一是使用信息准则。根据残差$\hat{\boldsymbol{\varepsilon}}_t$可以估计协方差矩阵$\boldsymbol{\Sigma}$,记为$\hat{\boldsymbol{\Sigma}}$。矩阵$\hat{\boldsymbol{\Sigma}}$的$(i,j)$元素为$\hat{\boldsymbol{\Sigma}}_{ij}\equiv\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}$,其中$T$为样本容量。则VAR模型的AIC与BIC分别为

$$\mathrm{AIC}(p)\equiv\ln|\hat{\boldsymbol{\Sigma}}|+n(np+1)\frac{2}{T} \tag{20.60}$$

$$\mathrm{BIC}(p)\equiv\ln|\hat{\boldsymbol{\Sigma}}|+n(np+1)\frac{\ln T}{T} \tag{20.61}$$

其中,n为VAR系统中变量的个数,p为滞后阶数,$|\hat{\boldsymbol{\Sigma}}|$为$\hat{\boldsymbol{\Sigma}}$的行列式,而$n(np+1)$为VAR模型中待估系数之总数(每个方程共有$(np+1)$个系数,共有$n$个方程)。将$|\hat{\boldsymbol{\Sigma}}|$视为多维情形下的残差平方和,则以上表达式为单一方程的信息准则向多方程情形的推广。比如,当$n=1$时,$\hat{\boldsymbol{\Sigma}}$为1阶矩阵,故$|\hat{\boldsymbol{\Sigma}}|=\hat{\boldsymbol{\Sigma}}=\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2=\mathrm{SSR}/T$,故$\mathrm{AIC}(p)=\ln(\mathrm{SSR}/T)+(p+1)\frac{2}{T}$,这正是单一方程的AIC表达式(其中,$p+1$为解释变量个数,含常数项)。

方法之二是检验最后一阶系数的显著性(类似于由大到小的序贯t规则)。在上面的例子中,假设要确定使用VAR(p)还是VAR$(p-1)$,则可以检验原假设“$H_0:\beta_{1p}=\beta_{2p}=\gamma_{1p}=\gamma_{2p}=0$”。

方法之三是检验VAR模型的残差是否为白噪声,即是否存在自相关。如果真实模型为VAR(p),但被错误地设置为Var$(p-1)$,则解释变量的最后一阶滞后$\boldsymbol{y}_{t-p}$被纳入扰动项$\boldsymbol{\varepsilon}_t$,导致扰动项出现自相关。更糟糕的是,由于$\{\boldsymbol{y}_t\}$的相关性,包含$\boldsymbol{y}_{t-p}$的扰动项$\boldsymbol{\varepsilon}_t$将与解释变量$\{\boldsymbol{y}_{t-1},\cdots,\boldsymbol{y}_{t-(p-1)}\}$相关,导致OLS估计不一致。为此,需要检验VAR模型的残差是否存在自相关。如果存在自相关,则可能意味着应该在解释变量中加入更高阶的滞后变量。

2. VAR变量个数的选择

VAR系统中包含的变量个数越多,则需要估计的系数越多。假设有5个变量,滞后4期,则每个方程中共有21个待估系数(含截距项),而整个VAR系统共有105个待估系数!待估系数过多将使得样本容量过小,增大估计误差,降低预测精度。因此,VAR模型通常仅包含为数不多的几个变量。

另一方面,如果VAR模型太小,则可能存在遗漏变量偏差。比如,早期的VAR模型常发现一个奇怪结果,即紧缩货币政策之后通货膨胀反而上升,称为“物价之谜”(price puzzle)。Sims(1992)认为,这是由于美联储在调整基准利率时,会考虑许多因素,而有些因素未包括在VAR模型中。如果某遗漏变量的变化预示着通货膨胀加剧,而美联储据此提高利率,则所谓的货币冲击(monetary shocks)不过是预期通胀的反映。此后,在VAR模型常使用“大宗商品价格”(commodity price)来控制未来的预期通胀。

在设定VAR模型是,主要应根据经济理论来确定哪些变量应在VAR模型中。比如,经济理论

告诉我们，通货膨胀率、失业率、短期利息率互相关联，可以构成一个三变量的 VAR 模型。如果 VAR 模型包含不相关的变量，则会增大估计量方差，降低预测能力。另外，也可以在 VAR 系统中引入其他外生变量，比如 $\{z_{1t}, z_{2t}, \cdots, z_{Kt}\}$，与扰动项不相关。

20.10 VAR 的脉冲响应函数

由于 VAR 模型包含许多参数，而这些参数的经济意义很难解释，故将注意力集中于脉冲响应函数。为了研究 VAR 的脉冲响应函数，首先定义向量移动平均过程。将 MA(∞)向多维推广，可以定义 n 维“无穷阶向量移动平均过程”(Vector Moving Average，简记 VMA(∞))：

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\psi}_0\boldsymbol{\varepsilon}_t + \boldsymbol{\psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots = \boldsymbol{\alpha} + \sum_{j=0}^{\infty} \boldsymbol{\psi}_j\boldsymbol{\varepsilon}_{t-j} \tag{20.62}$$

其中 $\boldsymbol{\psi}_0 = \boldsymbol{I}_n$，$\boldsymbol{\psi}_j$ 皆为 n 维方阵。定义“多维滤波”(multivariate filter)为(无穷项)“滞后矩阵多项式”(lag matrix polynomial)：

$$\boldsymbol{\psi}(L) \equiv \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots \tag{20.63}$$

因此，可以将 VMA(∞)简洁地写为 $\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\psi}(L)\boldsymbol{\varepsilon}_t$。对于两个多维滤波，可以类似于一维滤波那样地定义其乘积[①]，以及多维滤波 $\boldsymbol{\psi}(L)$ 的“逆”(inverse) $\boldsymbol{\psi}(L)^{-1}$。

使用滞后算子，可以把上述 VAR(p)系统“$\boldsymbol{y}_t = \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t$”写成 VMA(∞)的形式：

$$(\boldsymbol{I} - \boldsymbol{\Gamma}_1 L - \cdots - \boldsymbol{\Gamma}_p L^p)\boldsymbol{y}_t = \boldsymbol{\Gamma}_0 + \boldsymbol{\varepsilon}_t \tag{20.64}$$

$$\boldsymbol{\Gamma}(L)\boldsymbol{y}_t = \boldsymbol{\Gamma}_0 + \boldsymbol{\varepsilon}_t \tag{20.65}$$

其中，$\boldsymbol{\Gamma}(L) \equiv \boldsymbol{I} - \boldsymbol{\Gamma}_1 L - \cdots - \boldsymbol{\Gamma}_p L^p$。在方程(20.65)两边同时左乘 $\boldsymbol{\Gamma}(L)^{-1}$可得

$$\boldsymbol{y}_t = \boldsymbol{\Gamma}(L)^{-1}\boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}(L)^{-1}\boldsymbol{\varepsilon}_t \tag{20.66}$$

记 $\boldsymbol{\Gamma}(L)^{-1} \equiv \boldsymbol{\psi}(L) = \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots$，$\boldsymbol{\Gamma}(L)^{-1}\boldsymbol{\Gamma}_0 \equiv \boldsymbol{\alpha}$，则得到 VAR 模型的 VMA 表示法(Vector Moving Average Representation)：

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\psi}_0\boldsymbol{\varepsilon}_t + \boldsymbol{\psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots = \boldsymbol{\alpha} + \sum_{i=0}^{\infty} \boldsymbol{\psi}_i\boldsymbol{\varepsilon}_{t-i} \tag{20.67}$$

其中，$\boldsymbol{\psi}_0 \equiv \boldsymbol{I}_n$，而其余的 $\boldsymbol{\psi}_i$ 可通过递推公式 $\boldsymbol{\psi}_i = \sum_{j=1}^{i} \boldsymbol{\psi}_{i-j}\boldsymbol{\Gamma}_j$ 来确定(参见习题)。根据向量微分法则可知

$$\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} = \boldsymbol{\psi}_s \tag{20.68}$$

其中，$(\partial \boldsymbol{y}_{t+s}/\partial \boldsymbol{\varepsilon}_t')$ 为 n 维列向量 $\boldsymbol{y}_{t+s}$ 对 n 维行向量 $\boldsymbol{\varepsilon}_t'$ 求偏导数，故得到 $n \times n$ 矩阵 $\boldsymbol{\psi}_s$。矩阵 $\boldsymbol{\psi}_s$ 是一维情形下相隔 s 期的动态乘子向多维的推广，其第 i 行、第 j 列元素等于 $(\partial y_{i,t+s}/\partial \varepsilon_{jt})$。它表示的是，当第 j 个变量在第 t 期的扰动项 ε_{jt} 增加 1 单位时(而其他变量与其他期的扰动项均不变)，对第 i 个变量在第 $(t+s)$ 期的取值 $y_{i,t+s}$ 的影响。将 $(\partial y_{i,t+s}/\partial \varepsilon_{jt})$ 视为时间间隔 s 的函数，就是“脉冲响应函数”(IRF)。

有时也考察累积脉冲响应函数(CIRF)，即经过 k 期后的累积效应：

$$\sum_{s=0}^{k} \frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} = \sum_{s=0}^{k} \boldsymbol{\psi}_s \tag{20.69}$$

---

① 需要注意的是，多维滤波的乘积不满足交换律，正如矩阵乘法不满足交换律一样。

在方程(20.69)中,让 $k\to\infty$,则得到长期效应(long-run effect):

$$\sum_{s=0}^{\infty}\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'}=\sum_{s=0}^{\infty}\boldsymbol{\psi}_s=\boldsymbol{\psi}_0+\boldsymbol{\psi}_1+\boldsymbol{\psi}_2+\cdots=\boldsymbol{\psi}(1) \tag{20.70}$$

根据定义,$\boldsymbol{\Gamma}(L)\boldsymbol{\psi}(L)=\boldsymbol{I}$。令 $L=1$ 可得,$\boldsymbol{\Gamma}(1)\boldsymbol{\psi}(1)=\boldsymbol{I}$。因此,长期效应可以写为:

$$\boldsymbol{\psi}(1)=\boldsymbol{\Gamma}(1)^{-1}=(\boldsymbol{I}-\boldsymbol{\Gamma}_1-\cdots-\boldsymbol{\Gamma}_p)^{-1} \tag{20.71}$$

脉冲响应函数的缺点是,它假定在计算($\partial y_{i,t+s}/\partial \varepsilon_{jt}$)时,只让 ε_{jt} 变动,而所有其他同期扰动项均不变。此假定只有当扰动项的协方差矩阵 $\boldsymbol{\Sigma}\equiv \mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')$ 为对角矩阵时才成立(即同期扰动项之间正交);否则,当 ε_{jt} 变动时,必然伴随着其他方程的同期扰动项发生相应的变动。为此,需要考虑"正交化的脉冲响应函数"(Orthogonalized Impulse Response Function,简记 OIRF),即从扰动项 $\boldsymbol{\varepsilon}_t$ 中分离出相互正交的部分,记为 $\boldsymbol{u}_t$,然后计算当 $\boldsymbol{u}_t$ 中的某个分量变动时,对各变量在不同时期的影响。可以证明[①],对任何对称正定矩阵 $\boldsymbol{\Sigma}_{n\times n}$ 均可进行以下"三角分解"(triangular factorization):

$$\boldsymbol{\Sigma}=\boldsymbol{ADA}' \tag{20.72}$$

其中,$\boldsymbol{D}$ 为 n 级对角矩阵,其所有元素均为正数;$\boldsymbol{A}$ 为 n 阶下三角矩阵(lower triangular matrix),其主对角线元素均为 1;而且,矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{D}$ 唯一确定。对扰动项的协方差矩阵 $\boldsymbol{\Sigma}$ 进行三角分解,并定义"正交化的扰动项"为

$$\boldsymbol{u}_t\equiv\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}_t \tag{20.73}$$

可以证明,新扰动项 $\boldsymbol{u}_t$ 不再存在同期相关,因为其协方差矩阵为对角矩阵:

$$\begin{aligned}\mathrm{E}(\boldsymbol{u}_t\boldsymbol{u}_t')&=\mathrm{E}(\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\boldsymbol{A}^{-1'})=\boldsymbol{A}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')\boldsymbol{A}^{-1'}\\&=\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}'^{-1}=\underbrace{\boldsymbol{A}^{-1}\boldsymbol{A}}_{=\boldsymbol{I}}\boldsymbol{D}\underbrace{\boldsymbol{A}'\boldsymbol{A}'^{-1}}_{=\boldsymbol{I}}=\boldsymbol{D}\end{aligned} \tag{20.74}$$

其中,$\boldsymbol{D}$ 主对角线上的每一元素即为新扰动项 $\boldsymbol{u}_t$ 相应分量的方差。在实践中,常使用以下与三角分解相关联的"乔利斯基分解"(Cholesky decomposition)。由于 $\boldsymbol{\Sigma}=\boldsymbol{ADA}'$,而 $\boldsymbol{D}$ 为对角矩阵且所有元素均为正数,故可进一步分解为

$$\boldsymbol{\Sigma}=\boldsymbol{A}\,\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{A}'\equiv\boldsymbol{PP}' \tag{20.75}$$

其中,$\boldsymbol{P}\equiv\boldsymbol{A}\,\boldsymbol{D}^{1/2}$。由于 $\boldsymbol{A}$ 为下三角矩阵,$\boldsymbol{D}^{1/2}$ 为对角矩阵,故 $\boldsymbol{P}$ 依然为下三角矩阵;而 $\boldsymbol{P}'$ 依然为上三角矩阵。定义经乔利斯基分解所得的正交化扰动项为

$$\boldsymbol{v}_t\equiv\boldsymbol{P}^{-1}\boldsymbol{\varepsilon}_t \tag{20.76}$$

容易证明,新扰动项 $\boldsymbol{v}_t$ 的协方差矩阵为单位矩阵:

$$\begin{aligned}\mathrm{E}(\boldsymbol{v}_t\boldsymbol{v}_t')&=\mathrm{E}(\boldsymbol{P}^{-1}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\boldsymbol{P}^{-1'})=\boldsymbol{P}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')\boldsymbol{P}^{-1'}\\&=\boldsymbol{P}^{-1}\boldsymbol{\Sigma}\boldsymbol{P}'^{-1}=\boldsymbol{P}^{-1}\boldsymbol{PP}'\boldsymbol{P}'^{-1}=\boldsymbol{I}\end{aligned} \tag{20.77}$$

显然,新扰动项 $\boldsymbol{v}_t$ 各分量正交,且方差均被标准化为 1;因此,新扰动项 $\boldsymbol{v}_t$ 的一个分量变化一单位,也就是变化一个标准差。利用(20.75)的乔利斯基分解,可将 VAR 模型的 VMA 表示法(20.67)写为:

$$\boldsymbol{y}_t=\boldsymbol{\alpha}+\sum_{i=0}^{\infty}\boldsymbol{\psi}_i\boldsymbol{\varepsilon}_{t-i}=\boldsymbol{\alpha}+\sum_{i=0}^{\infty}\underbrace{\boldsymbol{\psi}_i\boldsymbol{P}}\underbrace{\boldsymbol{P}^{-1}\boldsymbol{\varepsilon}_{t-i}}=\boldsymbol{\alpha}+\sum_{i=0}^{\infty}\boldsymbol{\Phi}_i\boldsymbol{v}_{t-i} \tag{20.78}$$

其中,$\boldsymbol{\Phi}_i\equiv\boldsymbol{\psi}_i\boldsymbol{P}$。由向量微分法则可知:

$$\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{v}_t'}=\boldsymbol{\Phi}_s \tag{20.79}$$

---

① 参见 Hamilton(1994),p. 87。

其中，矩阵 $\boldsymbol{\Phi}_s$ 的第 i 行、第 j 列元素等于 $(\partial y_{i,t+s}/\partial v_{jt})$，即正交化的脉冲响应函数（OIRF）。在实际估计过程中，可根据残差 $\hat{\boldsymbol{\varepsilon}}_t$，估计 $\hat{\boldsymbol{\Sigma}}$，对 $\hat{\boldsymbol{\Sigma}}$ 进行乔利斯基分解得到 $\hat{\boldsymbol{P}}$ 以及 $\hat{\boldsymbol{v}}_t = \hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{\varepsilon}}_t$。

然而，正交化的脉冲响应函数依然有缺点。首先，正交化冲击（orthogonalized shocks）$\boldsymbol{v}_t$ 的经济含义不易解释。其次，由于乔利斯基分解所用矩阵 $\boldsymbol{P}$ 为下三角矩阵，故正交化的脉冲响应函数依赖于变量的次序（order of variables）；如果改变变量次序，则可能得到很不相同的结果。将方程（20.76）两边同时左乘 $\boldsymbol{P}$ 可得 $\boldsymbol{P}\boldsymbol{v}_t = \boldsymbol{\varepsilon}_t$，即

$$\begin{pmatrix} p_{11} & 0 & \cdots & 0 \\ p_{21} & p_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \\ \vdots \\ v_{nt} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{nt} \end{pmatrix} \tag{20.80}$$

作为一个思想实验，假设 y_{2t} 受到一个单位的正交化冲击，即 $v_{2t} = 1$，而向量 $\boldsymbol{v}_t$ 的其他分量均为 0。由于矩阵 $\boldsymbol{P}$ 的第一行只有一个非零元素 p_{11}，故对 v_{2t} 的冲击对于 ε_{1t} 没有影响，也不影响 y_{1t}。表达式（20.80）的第二个方程可写为

$$\varepsilon_{2t} = p_{21}v_{1t} + p_{22}v_{2t} = p_{21} \times 0 + p_{22} \times 1 = p_{22} \tag{20.81}$$

上式表明，v_{2t} 对 y_{2t} 的冲击为 p_{22}。类似地，可以证明 $\varepsilon_{3t} = p_{32}$，$\varepsilon_{4t} = p_{42}$，等等。总之，对 v_{2t} 的单位冲击会影响 $\varepsilon_{2t} = \varepsilon_{3t} = \cdots$，进而影响 $y_{2t} = y_{3t} = \cdots$，但不会影响 ε_{1t} 与 y_{1t}。更一般地，对 v_{jt} 的一单位冲击，只对第 t 期的第 j 至第 n 个变量有影响，而对第 t 期的第 1 至第 $(j-1)$ 个变量无影响。因此，按照一定变量顺序进行乔利斯基分解，意味着在 VAR 模型之上附加了一种“递归结构”（recursive structure），故也称为“递归 VAR”（recursive VAR）。

在递归 VAR 中，对第一个方程的正交化冲击会在同期影响所有方程；对第二个方程的正交化冲击会在同期影响除第一个方程之外的所有方程；依此类推，对最后一个方程的正交化冲击则只会在同期影响最后一个方程本身。这个结果完全是由于在对协方差矩阵 $\hat{\boldsymbol{\Sigma}}$ 进行乔利斯基分解时选择了某种变量次序；如果选择另一次序，就会得到不同的结果。因此，OIRF 虽然使得因果关系更清楚，但代价是需要对变量起作用的次序作较强的先验假设，而经济理论通常无法对变量次序给出明确的指南。

在实践中，可借助以下三种方法来确定 OIRF 的变量次序。

（1）通过考察两个变量之间的“交叉相关图”（cross correlogram），来确定两个变量的次序。对于变量 y_t 与 x_t，定义其“交叉相关函数”（cross-correlation function）为

$$\rho_{yx}(k) \equiv \text{corr}(y_t, x_{t+k}) \tag{20.82}$$

即 y_t 与 x_{t+k} 的相关系数。将 $\rho_{yx}(k)$ 视为 k 的函数，并画图，即得到交叉相关图。需要注意的是，$\rho_{yx}(k)$ 并不关于原点对称。一般来说，$\rho_{yx}(k) \neq \rho_{yx}(-k)$；但 $\rho_{yx}(k) \equiv \text{corr}(y_t, x_{t+k}) = \text{corr}(x_t, y_{t-k}) \equiv \rho_{xy}(-k)$。通过交叉相关图，可看出使得 $|\rho_{yx}(k)|$（绝对值）最大的 k 值，记为 k^*。如果 $k^* > 0$，则意味着 y_t 与未来的 x_{t+k^*} 最相关（变量排序应为 y_t, x_t）；反之，如果 $k^* < 0$，则意味着 y_t 与过去的 x_{t+k^*} 最相关（变量排序应为 x_t, y_t）。

（2）借助格兰杰因果检验来确定两个变量之间的排序，参见本章第 12 节。

（3）在难以确定变量次序的情况下，为了稳健起见，对于不同的变量排序，分别画其正交化脉冲响应图，然后进行比较。如果不同变量排序的 OIRF 差别不大，则结果较可信；反之，则值得怀疑。

20.11 预测误差的方差分解

VAR 模型的用途之一是预测。对于 VAR(p)模型 $\boldsymbol{y}_t = \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t$,在得到参数估计后,很容易进行向前一期的预测(one-step-ahead forecast):

$$\hat{\boldsymbol{y}}_{t+1} = \hat{\boldsymbol{\Gamma}}_0 + \hat{\boldsymbol{\Gamma}}_1 \boldsymbol{y}_t + \cdots + \hat{\boldsymbol{\Gamma}}_p \boldsymbol{y}_{t-p+1} \tag{20.83}$$

其中,假设 $\boldsymbol{y}_t$ 为已知,即当前时间为第 t 期。为了预测 $\boldsymbol{y}_{t+2}$,则可将 $\hat{\boldsymbol{y}}_{t+1}$ 视为 $\boldsymbol{y}_{t+1}$,即

$$\hat{\boldsymbol{y}}_{t+2} = \hat{\boldsymbol{\Gamma}}_0 + \hat{\boldsymbol{\Gamma}}_1 \hat{\boldsymbol{y}}_{t+1} + \hat{\boldsymbol{\Gamma}}_2 \boldsymbol{y}_t + \cdots + \hat{\boldsymbol{\Gamma}}_p \boldsymbol{y}_{t-p+2} \tag{20.84}$$

类似地,可计算向前 h 期的预测 $\hat{\boldsymbol{y}}_{t+h}$。利用 VAR 模型的 VMA 表示法,可将向前 h 期的预测误差(h-step forecast error)写为(Lutkepohl,2005,sec. 2.2.2):

$$\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h} = \boldsymbol{\psi}_0 \boldsymbol{\varepsilon}_{t+h} + \boldsymbol{\psi}_1 \boldsymbol{\varepsilon}_{t+h-1} + \cdots + \boldsymbol{\psi}_{h-1} \boldsymbol{\varepsilon}_{t+1} = \sum_{i=0}^{h-1} \boldsymbol{\psi}_i \boldsymbol{\varepsilon}_{t+h-i} \tag{20.85}$$

我们希望度量各方程的扰动项对预测误差的单独贡献。由于 $\boldsymbol{\varepsilon}_{t+h-i}$ 的各分量存在同期相关(contemporaneously correlated),故利用乔利斯基分解将(20.85)式写为

$$\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h} = \sum_{i=0}^{h-1} \boldsymbol{\psi}_i \boldsymbol{\varepsilon}_{t+h-i} = \sum_{i=0}^{h-1} \underbrace{\boldsymbol{\psi}_i \boldsymbol{P} \boldsymbol{P}^{-1}}_{\boldsymbol{\Phi}_i} \boldsymbol{\varepsilon}_{t+h-i} = \sum_{i=0}^{h-1} \boldsymbol{\Phi}_i \boldsymbol{v}_{t+h-i} \tag{20.86}$$

其中,$\boldsymbol{v}_{t+h-i}$ 为正交化冲击。记矩阵 $\boldsymbol{\Phi}_i$ 的 (m,k) 元素为 $\phi_{i,mk}$,向量 $\boldsymbol{v}_{t+h-i}$ 的第 k 个元素为 $v_{k,t+h-i}$,则根据表达式(20.86),可将 $\boldsymbol{y}_{t+h}$ 的第 j 个分量 $y_{j,t+h}$ 的预测误差写为:

$$\begin{aligned} y_{j,t+h} - \hat{y}_{j,t+h} &= \sum_{i=0}^{h-1} (\phi_{i,j1} v_{1,t+h-i} + \cdots + \phi_{i,jn} v_{n,t+h-i}) \\ &= \sum_{i=0}^{h-1} \sum_{k=1}^{n} \phi_{i,jk} v_{k,t+h-i} \\ &= \sum_{k=1}^{n} \sum_{i=0}^{h-1} \phi_{i,jk} v_{k,t+h-i} \\ &= \sum_{k=1}^{n} (\phi_{0,jk} v_{k,t+h} + \cdots + \phi_{h-1,jk} v_{k,t+1}) \end{aligned} \tag{20.87}$$

在上式的推导中交换了求和的次序,即从“先对 k 求和,再对 i 求和”变为“先对 i 求和,再对 k 求和”。由于 $\boldsymbol{v}_{t+h-i}$ 的每个正交化分量均不相关,且方差为 1,故很容易计算对 $y_{j,t+h}$ 预测的均方误差(MSE):

$$\mathrm{MSE}(\hat{y}_{j,t+h}) \equiv \mathrm{E}[(y_{j,t+h} - \hat{y}_{j,t+h})^2] = \sum_{k=1}^{n} (\phi_{0,jk}^2 + \cdots + \phi_{h-1,jk}^2) \tag{20.88}$$

根据方程(20.88),可以计算第 l 个变量的正交化冲击对 $y_{j,t+h}$ 预测均方误差的贡献比例:

$$\frac{\phi_{0,jl}^2 + \cdots + \phi_{h-1,jl}^2}{\sum_{k=1}^{n} (\phi_{0,jk}^2 + \cdots + \phi_{h-1,jk}^2)} \tag{20.89}$$

这是度量第 l 个变量对第 j 个变量的影响的又一方法(另一方法为脉冲响应函数)。显然,由表达式(20.89)计算的贡献比例是预测时期 h 的函数,故可将其列表或画图。而且,所有变量对预测误差均方误差的贡献比例之和为 1。此性质称为“预测误差的方差分解”(Forecast-error Variance Decomposition,简记 FEVD)。由于它将预测误差的来源分别归因于各变量的正交化新息(orthogonalized innovation),故也称为“新息核算”(innovation accounting)。

由于在推导过程中使用了乔利斯基分解,故 FEVD 的结果也依赖于变量排序。这是 FEVD 与 IRF 的共同缺陷。为了稳健起见,一般建议比较不同变量排序的 FEVD 结果;如果结果对于变量排序不敏感,则较为可信。

20.12 格兰杰因果检验

经济学中常常需要确定因果关系究竟是从 x 到 y，还是从 y 到 x，抑或双向因果关系。格兰杰[Granger(1969)]提出的检验方法基于以下思想。如果 x 是 y 的因，但 y 不是 x 的因，则 x 的过去值可以帮助预测 y 的未来值，但 y 的过去值却不能帮助预测 x 的未来值。考虑以下时间序列模型①：

$$y_t = \gamma + \sum_{m=1}^{p} \alpha_m y_{t-m} + \sum_{m=1}^{p} \beta_m x_{t-m} + \varepsilon_t \tag{20.90}$$

其中，滞后阶数 p 可根据“信息准则”或“由大到小的序贯 t 规则”来确定。检验原假设“$H_0: \beta_1 = \cdots = \beta_p = 0$”，即 x 的过去值对预测 y 的未来值没有帮助。如果拒绝 H_0，则称 x 是 y 的“格兰杰因”(Granger cause)。将以上回归模型中 x 与 y 的位置互换，则可以检验 y 是否为 x 的格兰杰因。在实际操作中，常将 (x, y) 构成一个二元 VAR 系统，然后在 VAR 的框架使用 Stata 命令 `vargranger` 进行格兰杰因果检验。也可使用非官方命令 `gcause`，下载方法为“`ssc install gcause`”。

需要指出的是，格兰杰因果关系并非真正意义上的因果关系。它充其量只是一种动态相关关系，表明的是一个变量是否对另一变量有“预测能力”(predictability)。从某种意义上来说，它顶多是因果关系的必要条件(如果不考虑非线性的因果关系)。另外，格兰杰因果关系也可能由第三个变量所引起。在经济学的实证研究中，由于通常不可能进行“控制实验”，能够最有说服力地说明因果关系的当属随机实验与自然实验(参见第 18 章)。

另外，格兰杰因果检验仅适用于平稳序列，或者有协整关系的单位根过程(参见第 21 章)。对于不存在协整关系的单位根变量，则只能先差分，得到平稳序列后再进行格兰杰因果检验。

20.13 面板格兰杰因果检验

对于面板数据，可进行面板格兰杰因果检验。考虑以下面板模型：

$$y_{it} = \gamma + \sum_{m=1}^{p} \alpha_m y_{i,t-m} + \sum_{m=1}^{p} \beta_m x_{i,t-m} + u_i + \varepsilon_{it} \quad (i = 1, \cdots, n; t = 1, \cdots, T) \tag{20.91}$$

其中，u_i 为个体 i 的异质性(个体效应)，可能与解释变量相关。显然，这是一个动态面板模型。如果时间维度较短，应使用差分 GMM 或系统 GMM 进行估计；反之，如果时间维度较长，可考虑偏差校正 LSDV 法(参见第 16 章)。Hartwig(2009, 2010)提供了进行面板格兰杰因果检验的两个实例。

20.14 VAR 的 Stata 命令及实例

与 VAR 相关的 Stata 命令包括(假设变量为 x, y, z)

---

① 更一般地，可以考虑 $y_t = \gamma + \sum_{l=1}^{p} \alpha_l y_{t-l} + \sum_{m=1}^{q} \beta_m x_{t-m} + \varepsilon_t$，其中 p 与 q 不一定相等。

varsoc x y z,maxlag(#)

此命令用来计算不同滞后期的信息准则,其中“soc”表示“Selection-Order Criteria”,选择项“maxlag(#)”表示最大滞后期,默认值为4。在未使用选择项“maxlag(#)”的情况下,如果恰好选择最优滞后4期,则应通过选择项“maxlag(#)”增加最大滞后期数。

varbasic x y z,lags(numlist) irf fevd

这是估计VAR模型的便捷命令。选择项“lags(numlist)”表示滞后阶数,默认为“lags(1 2)”,即滞后二阶。选择项“irf”表示画(未正交化)脉冲响应图,选择项“fevd”表示画预测方差分解图,默认为“oirf”(画正交化脉冲响应图)。

更正式而深入的VAR命令为

var x y z,lags(numlist) dfk small exog(w1 w2)

其中,选择项“lags(numlist)”表示滞后阶数,默认为“lags(1 2)”,即滞后二阶。如果样本容量较小,可使用选择项“dfk”进行自由度调整。选择项“small”表示显示小样本的 t 或 F 统计量,而非大样本的标准正态或 χ^2 统计量。选择项“exog(w1 w2)”表示在VAR模型中引入外生变量w1,w2。

在使用命令var完成VAR模型的估计之后,Stata提供了一系列“估计后命令”(postestimation commands)来检验VAR的性质与假设。

varlmar

估计VAR后,对残差是否存在自相关进行LM检验。

varnorm

估计VAR后,检验残差是否服从正态分布。

varstable,graph

估计VAR后,通过特征值检验该VAR系统是否为平稳过程。如果所有特征值都在单位圆内部,则为平稳过程,参见第21章。选择项“graph”表示画出特征值的几何分布图。

varwle

估计VAR后,对每个方程以及所有方程的各阶系数的联合显著性进行Wald检验,其中“wle”表示Wald lag-exclusion statistics。

vargranger

估计VAR后,进行格兰杰因果检验。

irf create irfname,set(filename) step(#) replace

估计VAR后,将有关脉冲响应的结果存为“irfname”(可自行命名)。选择项“set(filename)”表示建立脉冲文件“filename”,使之成为当前的脉冲文件(make filename active),并将脉冲结果“irfname”存入此脉冲文件“filename”(若未使用选择项“set(filename)”指定脉冲文件,则将脉冲响应结果存入当前的脉冲文件);选择项“step(#)”表示考察#期的脉冲响应函数,默认值为“step(8)”;选择项“replace”表示替代已有的同名脉冲响应结果irfname(如果有)。一个脉冲文件“filename”可存储多个脉冲响应结果“irfname”。

irf graph irf,impulse(varname) response(varname) noci

画脉冲响应图(未正交化)。其中,选择项“impulse(varname)”用于指定脉冲变量,而选择项“response(varname)”用来指定反应变量,默认画出所有变量的脉冲响应图。选择项“noci”表示不画置信区间,默认画置信区间。

irf graph cirf,impulse(varname) response(varname)

画累积脉冲响应图(未正交化)

```
irf graph oirf,impulse(varname) response(varname)
```

画正交化的脉冲响应图

```
irf graph coirf,impulse(varname) response(varname)
```

画正交化的累积脉冲响应图

```
irf graph fevd,impulse(varname) response(varname)
```

画预测方差分解图

如果将以上命令中的“irf graph”改为“irf table”,则将相应信息列表而非画图。

```
fcast compute prefix,step(#)
```

估计 VAR 后,计算被解释变量未来#期的预测值,并把预测值赋予被解释变量加上前缀“prefix”(自行确定)的变量名。

```
fcast graph varlist,observed
```

运行命令“fcast compute”后,将变量“varlist”的预测值画图,其中选择项“observed”表示与实际观测值相比较。

下面以数据集 varexample. dta 为例演示①。该数据集包含自 1960 年第 1 季以来的美国宏观季度数据,最初来自 Stock and Watson(2001),后经 Becketti(2013)拓展至 2012 年第 1 季。主要变量包括:date(季度),fedfunds(联邦基金利率),inflation(通货膨胀率)②,以及 unrate(失业率)③。我们希望估计一个关于利率、通胀率与失业率的三元 VAR 系统。

```
. use varexample.dta,clear
```

首先,看一下这三个变量的时间趋势图,结果参见图 20.3。

```
. tsline inflation unrate fedfunds,lpattern("-" "_") xline(169)
```

其中,选择项“xline(169)”表示在 x 轴上 $t=169$(即 2002 年第 1 季)的位置画一条垂直线。

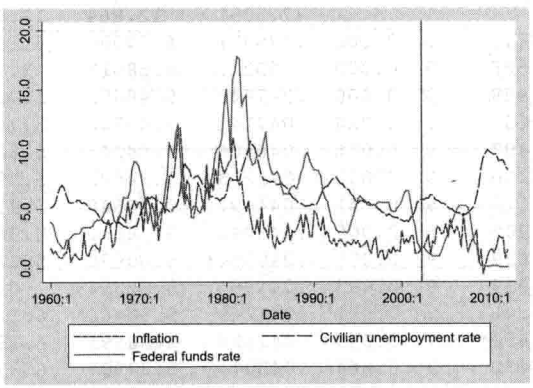


图 20.3 利率、通胀率与失业率的时间趋势图

在本例中,预留最近十年的数据(即 2002 年第 2 季—2012 年第 1 季)作为“样本外预测”(out-of-sample forecasts)。图 20.3 中的垂直线为样本内(in-sample)与样本外(out-of-sample)的分

① 本例主要借鉴了 Becketti(2013)。

② 以年化的季度 GDP 平减指数的百分比变化(annualized quarterly percentage change in GDP price deflator)来衡量,计算公式为 $400\times\ln(P_t/P_{t-1})$,其中 P_t 为 GDP 平减指数。

③ 其中,联邦基金利率与失业率的季度值为其相应月度值的平均。

界线。从图 20.3 可以看出,在样本内的 20 世纪 70 年代中期与 80 年代早期,联邦基金利率与通胀率都曾达到很高的水平。另一方面,这两个变量在样本外都曾达到很低的水平,反映出 2001 年 9·11 事件后宽松的货币政策以及 2008 年全球金融风暴的影响。与此相反,无论在样本内还是样本外,失业率的波动范围大体相同。以上特点也可通过对比样本内外的统计指标来看出。

```
. sum inflation unrate fedfunds if date < =tq(2002q1)
```

其中,"tq(2002q1)"表示季度数据格式。

Variable	Obs	Mean	Std. Dev.	Min	Max
inflation	169	3.792319	2.494935	.5674297	11.79167
unrate	169	5.927811	1.502554	3.4	10.66667
fedfunds	169	6.497337	3.190249	1.683333	17.78

```
. sum inflation unrate fedfunds if date >tq(2002q1)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
inflation	40	2.23731	1.063853	-.448682	4.558042
unrate	40	6.591667	1.930454	4.433333	9.933333
fedfunds	40	1.929667	1.851749	.0733333	5.256667

为了估计 VAR,首先需要根据信息准则确定 VAR 模型的阶数。

```
. varsoc inflation unrate fedfunds if date < =tq(2002q1),maxlag(13)
```

其中,选择项"maxlag(13)"表示最多滞后 13 阶。

Selection-order criteria
Sample: 1963:2 - 2002:1 Number of obs = 156

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-1000.39				77.5091	12.864	12.8878	12.9227
1	-461.535	1077.7	9	0.000	.086931	6.07096	6.16625	6.30557
2	-414.252	94.566	9	0.000	.053221	5.58015	5.7469	5.99071*
3	-396.258	35.988	9	0.000	.047442	5.46485	5.70306*	6.05136
4	-386.713	19.09	9	0.024	.047141	5.45786	5.76754	6.22033
5	-376.51	20.407	9	0.016	.046466	5.44243	5.82358	6.38085
6	-359.906	33.206	9	0.000	.042213	5.34495	5.79756	6.45933
7	-354.848	10.117	9	0.341	.044491	5.39549	5.91956	6.68581
8	-343.743	22.209	9	0.008	.043423	5.36851	5.96404	6.83478
9	-329.929	27.629	9	0.001	.040964*	5.30678	5.97379	6.94901
10	-322.47	14.919	9	0.093	.041961	5.32654	6.065	7.14472
11	-311.765	21.41	9	0.011	.041269	5.30468*	6.11461	7.29882
12	-302.939	17.652*	9	0.039	.041623	5.30691	6.18831	7.477
13	-299.568	6.7431	9	0.664	.045074	5.37907	6.33193	7.72511

Endogenous: inflation unrate fedfunds
Exogenous: _cons

其中,"LL"表示对数似然函数;"LR"表示似然比检验,即对最后一阶系数的联合显著性进行似然比检验,随后的 df 与 p 分别表示此似然比统计量的自由度与 p 值;"FPE"表示 Akaike's Final Prediction Error,度量向前一期预测的均方误差(MSE of one-step ahead forecast)。上表显示,不同信息准则所选择的滞后阶数并不一致(上表中打星号者)。根据最简洁的 SBIC 准则,只要滞后 2 阶就够了。根据 HQIC 准则,需滞后 3 阶。根据 FPE,需滞后 9 阶。根据 AIC 准则,需滞后 11 阶。而根据 LR 检验,则需滞后 12 阶之多。如果根据 SBIC 准则,选择滞后 2 阶,可能过于简洁;反

之，如果根据 AIC 准则，选择滞后 11 阶，共需估计 3×34=102 个参数，将损失较多样本容量。根据 Lutkepohl(2005)，SBIC 与 HQIC 提供了对真实滞后阶数的一致估计，而 FPE 与 AIC 可能高估滞后阶数。作为折中，Stock and Watson(2001)与 Becketti(2013)均选择滞后 4 阶。然而，滞后 4 阶的 VAR 模型依然不能保证扰动项为白噪声，故本例选择滞后 5 阶，刚好能保证扰动项为白噪声(参见下文的检验)。

下面估计 5 阶向量自回归模型：

```
. var inflation unrate fedfunds if date<=tq(2002q1),lags(1/5)
```

```
Vector autoregression

Sample:  1961:2 - 2002:1                        No. of obs      =       164
Log likelihood = -394.4564                      AIC             =   5.39581
FPE            =  .0443362                      HQIC            =  5.764131
Det(Sigma_ml)  =  .0246449                      SBIC            =  6.303088

Equation           Parms      RMSE     R-sq      chi2     P>chi2
inflation             16   1.03136   0.8446  891.4284   0.0000
unrate                16   .230874   0.9791  7676.319   0.0000
fedfunds              16   .886433   0.9292  2153.037   0.0000

                    Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
inflation
  inflation
         L1.     .5168961   .0764965     6.76   0.000     .3669658    .6668265
         L2.     .1584442   .0840164     1.89   0.059    -.0062249    .3231133
         L3.     .1001178    .084663     1.18   0.237    -.0658185    .2660542
         L4.      .28014   .0849989     3.30   0.001     .1135452    .4467348
         L5.    -.0425256   .0806997    -0.53   0.598    -.2006942    .1156429

      unrate
         L1.    -.8904664    .386067    -2.31   0.021    -1.647144   -.133789
         L2.     1.350818   .6414896     2.11   0.035     .0935211    2.608114
         L3.     -.237796   .6696613    -0.36   0.723    -1.550308    1.074716
         L4.    -1.403182   .6447665    -2.18   0.030    -2.666901   -.1394632
         L5.     1.068276   .3606713     2.96   0.003     .3613729    1.775178

    fedfunds
         L1.     .1095813   .1023397     1.07   0.284    -.0910007    .3101634
         L2.    -.0274367   .1390409    -0.20   0.844    -.2999518    .2450784
         L3.    -.0821163   .1371351    -0.60   0.549    -.3508962    .1866636
         L4.    -.0217318   .1356754    -0.16   0.873    -.2876508    .2441872
         L5.    -.0059257   .1012347    -0.06   0.953    -.2043421    .1924907

       _cons     .7982082   .3714537     2.15   0.032     .0701724    1.526244
unrate
  inflation
         L1.     .0088075    .017124     0.51   0.607    -.0247549     .04237
         L2.    -.0080949   .0188074    -0.43   0.667    -.0449567    .0287668
         L3.     .0154295   .0189521     0.81   0.416     -.021716    .0525749
         L4.    -.0134909   .0190273    -0.71   0.478    -.0507837     .023802
         L5.     .0095475   .0180649     0.53   0.597    -.0258591    .0449541

      unrate
         L1.      1.43892   .0864225    16.65   0.000     1.269535    1.608305
         L2.    -.5031447   .1435998    -3.50   0.000     -.784595   -.2216944
         L3.     .0987478   .1499061     0.66   0.510    -.1950627    .3925583
         L4.    -.2194761   .1443333    -1.52   0.128    -.5023641     .063412
         L5.     .1231384   .0807376     1.53   0.127    -.0351044    .2813811

    fedfunds
         L1.    -.0071153   .0229091    -0.31   0.756    -.0520163    .0377857
         L2.     .0609474   .0311248     1.96   0.050    -.0000561    .1219508
         L3.    -.0378636   .0306982    -1.23   0.217     -.098031    .0223037
         L4.     .0238004   .0303714     0.78   0.433    -.0357266    .0833273
         L5.    -.0085719   .0226618    -0.38   0.705    -.0529881    .0358443

       _cons     .1091412   .0831512     1.31   0.189    -.0538322    .2721146
fedfunds
  inflation
         L1.     .1414521    .065747     2.15   0.031     .0125904    .2703138
         L2.     .1786844   .0722101     2.47   0.013     .0371551    .3202137
         L3.    -.0471542   .0727659    -0.65   0.517    -.1897727    .0954643
         L4.    -.0023731   .0730546    -0.03   0.974    -.1455574    .1408113
         L5.    -.1204686   .0693595    -1.74   0.082    -.2564108    .0154736

      unrate
         L1.    -1.586394   .3318157    -4.78   0.000    -2.236741   -.936047
         L2.     1.783011   .5513455     3.23   0.001     .7023939    2.863629
         L3.    -1.057614   .5755584    -1.84   0.066    -2.185687    .0704601
         L4.     1.101875   .5541619     1.99   0.047     .0157373    2.188012
         L5.    -.3179867   .3099886    -1.03   0.305    -.9255533    .2895798

    fedfunds
         L1.     .9448705   .0879586    10.74   0.000     .7724748    1.117266
         L2.    -.3834779   .1195024    -3.21   0.001    -.6176984   -.1492575
         L3.     .3654699   .1178645     3.10   0.002     .1344598     .59648
         L4.     .0063154   .1166099     0.05   0.957    -.2222358    .2348666
         L5.      .005741   .0870089     0.07   0.947    -.1647933    .1762753

       _cons     .2710914   .3192558     0.85   0.396    -.3546385    .8968214
```

即使仅滞后 5 阶，此 VAR 模型依然包含了 48 个参数；这些系数如此之多，以至于无法解释其经济含义。因此，在实证论文中，甚至不汇报 VAR 的回归系数，而主要汇报脉冲响应函数、预测方差分解与格兰杰因果检验。在上述命令 var 中，由于样本容量为 164，故没有使用选择项“dfk”或“small”进行小样本自由度调整。

下面检验各阶系数的联合显著性：

```
. varwle
```

Equation: inflation

lag	chi2	df	Prob > chi2
1	63.62012	3	0.000
2	8.46243	3	0.037
3	1.819772	3	0.611
4	17.74984	3	0.000
5	12.07998	3	0.007

Equation: unrate

lag	chi2	df	Prob > chi2
1	356.3783	3	0.000
2	22.35949	3	0.000
3	2.902869	3	0.407
4	4.008033	3	0.261
5	3.539652	3	0.316

Equation: fedfunds

lag	chi2	df	Prob > chi2
1	255.0479	3	0.000
2	35.38302	3	0.000
3	16.31482	3	0.001
4	4.265025	3	0.234
5	3.71692	3	0.294

Equation: All

lag	chi2	df	Prob > chi2
1	572.2714	9	0.000
2	46.94062	9	0.000
3	20.02311	9	0.018
4	27.19206	9	0.001
5	20.22245	9	0.017

虽然单一方程的某些阶系数不显著，但作为三个方程的整体，各阶系数均高度显著。

下面检验残差是否为白噪声（residual whiteness），即残差是否存在自相关：

```
. varlmar
```

Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	12.0759	9	0.20906
2	10.3352	9	0.32404

H0: no autocorrelation at lag order

结果显示,可以接受残差"无自相关"的原假设,即认为扰动项为白噪声。

进一步检验此 VAR 系统是否稳定(为平稳过程),结果见下表与图 20.4:

```
. varstable,graph
```

Eigenvalue stability condition

Eigenvalue	Modulus
.9694669 + .04441834*i*	.970484
.9694669 - .04441834*i*	.970484
.01442202 + .7336039*i*	.733746
.01442202 - .7336039*i*	.733746
.6764154 + .2053908*i*	.706911
.6764154 - .2053908*i*	.706911
-.6844224	.684422
-.3249446 + .5288642*i*	.620714
-.3249446 - .5288642*i*	.620714
-.0073735 + .5922748*i*	.592321
-.0073735 - .5922748*i*	.592321
.4677774 + .154557*i*	.49265
.4677774 - .154557*i*	.49265
-.4349533	.434953
.4285352	.428535

All the eigenvalues lie inside the unit circle.
VAR satisfies stability condition.

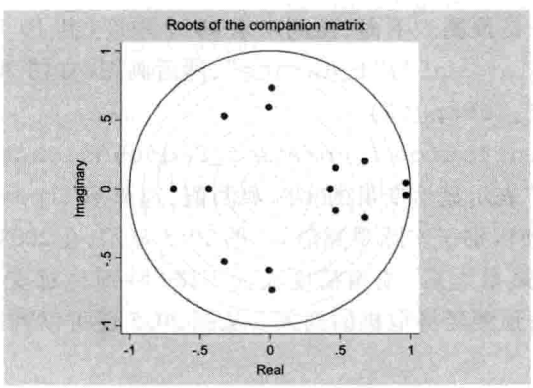


图 20.4 VAR 系统稳定性的判别图

上表与图 20.4 显示,所有特征值均在单位圆之内,故此 VAR 系统是稳定的;但有两个根十分接近单位圆,这意味着有些冲击有较强的持续性(persistence),详见第 21 章。

进一步检验 VAR 模型的残差是否服从正态分布：

```
. varnorm
```

Jarque-Bera test

Equation	chi2	df	Prob > chi2
inflation	7.020	2	0.02989
unrate	51.470	2	0.00000
fedfunds	217.041	2	0.00000
ALL	275.531	6	0.00000

Skewness test

Equation	Skewness	chi2	df	Prob > chi2
inflation	.33178	3.009	1	0.08281
unrate	.79646	17.339	1	0.00003
fedfunds	1.1369	35.332	1	0.00000
ALL		55.680	3	0.00000

Kurtosis test

Equation	Kurtosis	chi2	df	Prob > chi2
inflation	3.7662	4.011	1	0.04520
unrate	5.2349	34.131	1	0.00000
fedfunds	8.1567	181.709	1	0.00000
ALL		219.851	3	0.00000

上表显示，绝大多数的检验结果均可在 5% 的显著性水平上拒绝这三个变量的扰动项服从正态分布的原假设。尽管扰动项不服从正态分布对 VAR 模型的影响不大，但残差项的非正态性暗示模型可能偏离了真实的数据生成过程（DGP），并且使得对变量未来值的预测区间（forecast intervals）变得不可信（该预测区间依赖于正态假设）。

VAR 模型的用途之一是预测。下面，预测未来 40 个季度（共 10 年）的变量取值，分别记为“f_fedfunds”，“f_inflation”与“f_unrate”，然后画图（如图 20.5）：

```
. fcast compute f_,step(40)
. fcast graph f_inflation f_unrate f_fedfunds,observed lpattern("_")
```

其中，选择项“observed”表示显示变量的实际观测值，选择项“lpattern("_")”表示以虚线来表示变量的预测值（以便区别于实际观测值）。图 20.5 显示，在 2008 年金融风暴之前，预测效果尚可；但在 2008 年金融风暴之后，预测精度大大下降（特别是对变量 unrate 与 fedfunds 的预测）。显然，VAR 模型无法预测经济危机的到来。从图 20.5 还可以看出，预测的时期越长，则预测的精确度越低。

除了预测之外，我们还想知道对某变量的冲击会对该变量及其他变量产生怎样的动态影响，比如提高联邦基金利率一个百分点会对一年后的通胀率有多大影响。这就需要用到正交化的脉冲响应函数（未正交化的脉冲响应函数无法厘清各变量冲击的单独影响，意义不大），但正交化的脉冲响应函数依赖于变量的排序（上文的预测不依赖于变量排序）。为此，分别考察变量之间

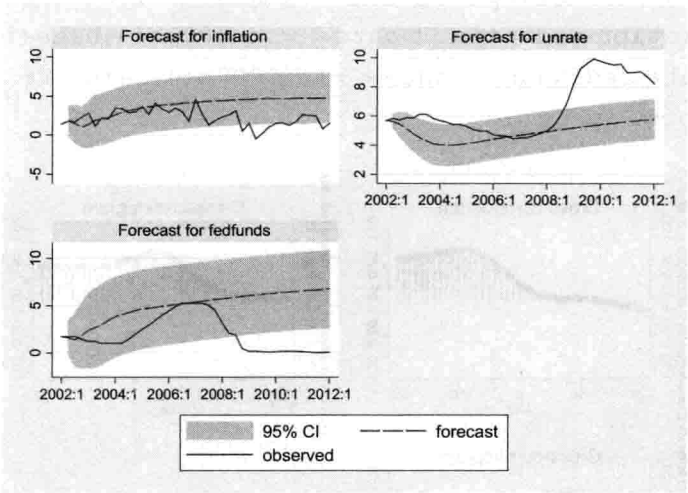


图 20.5 对未来 40 个季度的预测

的格兰杰因果关系与交叉相关图。首先,考察这三个变量之间的格兰杰因果关系。

```
. vargranger
```

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
inflation	unrate	14.702	5	0.012
inflation	fedfunds	2.5547	5	0.768
inflation	ALL	31.159	10	0.001
unrate	inflation	2.0977	5	0.835
unrate	fedfunds	17.344	5	0.004
unrate	ALL	40.496	10	0.000
fedfunds	inflation	26.041	5	0.000
fedfunds	unrate	35.561	5	0.000
fedfunds	ALL	50.517	10	0.000

上表上部显示,在以 inflation 为被解释变量的方程中,如果检验变量 unrate 系数的联合显著性(即在方程中排除变量 unrate),其卡方统计量为 14.702,相应的 p 值为 0.012,故可认为 unrate 是 inflation 的格兰杰原因。类似地,如果检验变量 fedfunds 系数的联合显著性(即在方程中排除变量 fedfunds),其卡方统计量为 2.554 7,相应的 p 值为 0.768,故可认为 fedfunds 不是 inflation 的格兰杰原因。如果同时检验变量 unrate 与 fedfunds 系数的联合显著性(即在方程中同时排除变量 inflation 与 unrate),其卡方统计量为 31.159,相应的 p 值为 0.001,强烈拒绝"unrate 与 fedfunds 都不是 inflation 的格兰杰原因"的原假设。

上表中部汇报了以 unrate 为被解释变量的检验结果。结果表明,inflation 不是 unrate 的格兰杰原因,而 fedfunds 是 unrate 的格兰杰原因。上表下部汇报了以 fedfunds 为被解释变量的检验结果。结果表明,inflation 与 unrate 都是 fedfunds 的格兰杰原因。显然,格兰杰因果关系并未给出唯一的变量作用次序。为此,进一步考察交叉相关图,结果参见图 20.6。

```
. xcorr inflation unrate if date < =tq(2002q1),name(iu)
. xcorr inflation fedfunds if date < =tq(2002q1),name(ir)
. xcorr unrate fedfunds if date < =tq(2002q1),name(ur)
. graph combine iu ir ur
```

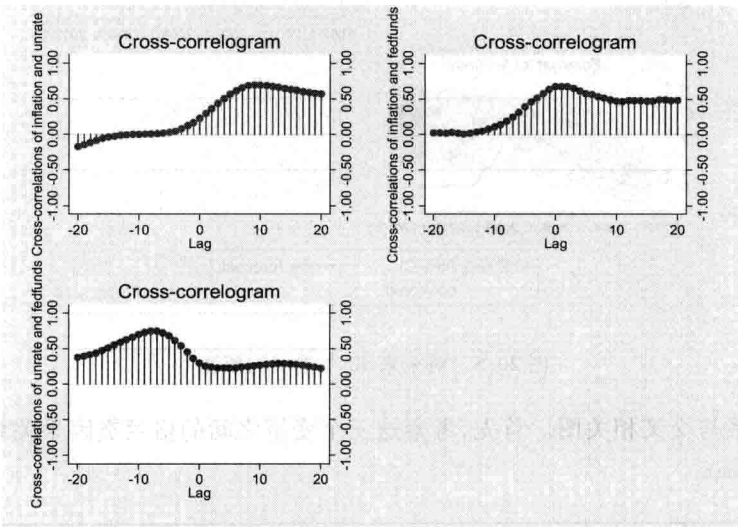
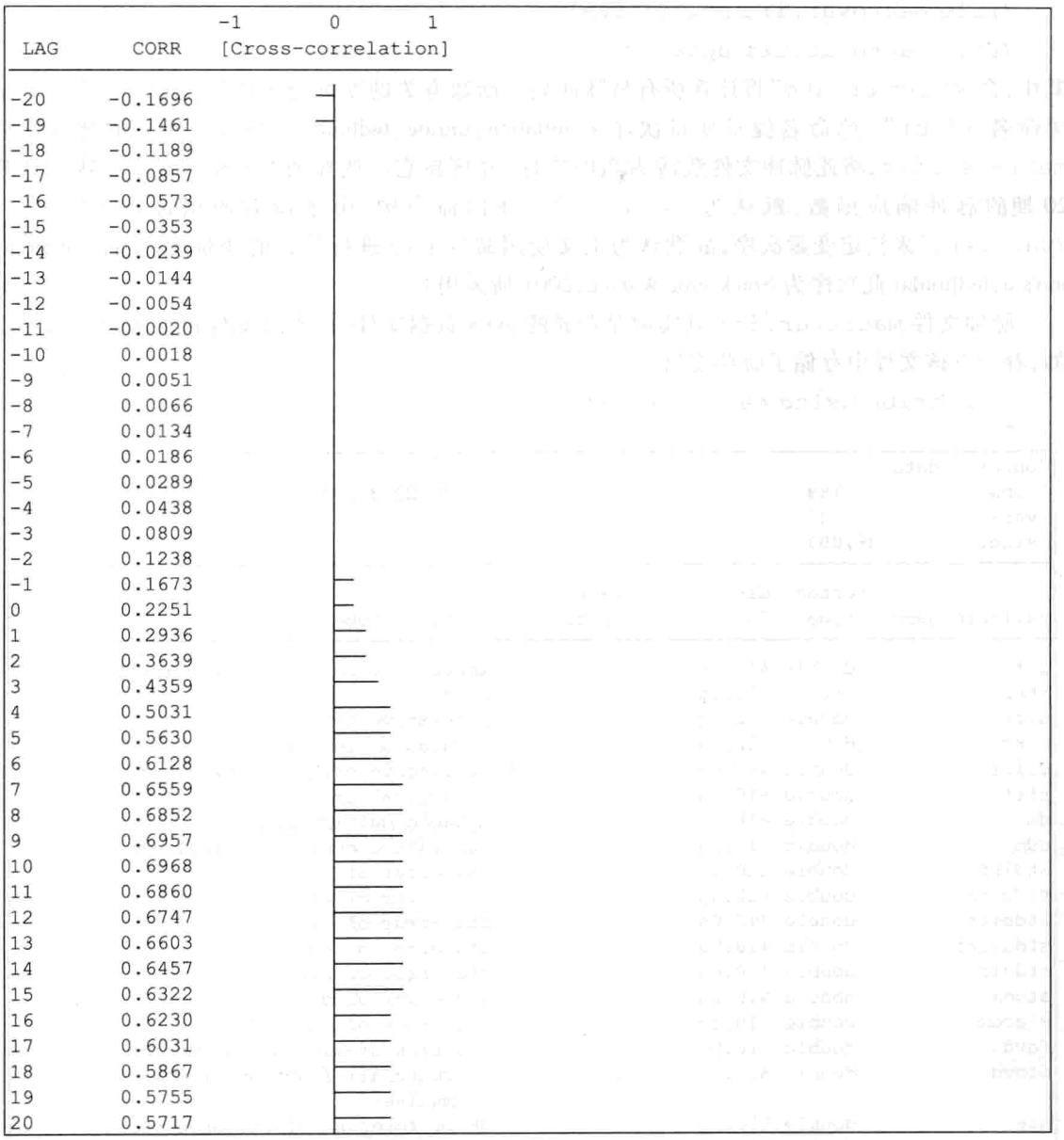


图 20.6 三个变量之间的交叉相关图

从图 20.6 左上部可看出，inflation 与滞后 10 个季度的 unrate 最相关；右上部显示，inflation 与滞后 1 个季度的 fedfunds 最相关；左下部显示，unrate 与提前 8 个季度的 fedfunds 最相关。如果无法从图 20.6 准确地看出交叉相关系数的最大值，可通过命令“xcorr,table”将交叉相关系数列表，例如：

```
. xcorr inflation unrate if date < =tq(2002q1),table
```



LAG	CORR	-1 0 1 [Cross-correlation]
-20	-0.1696	
-19	-0.1461	
-18	-0.1189	
-17	-0.0857	
-16	-0.0573	
-15	-0.0353	
-14	-0.0239	
-13	-0.0144	
-12	-0.0054	
-11	-0.0020	
-10	0.0018	
-9	0.0051	
-8	0.0066	
-7	0.0134	
-6	0.0186	
-5	0.0289	
-4	0.0438	
-3	0.0809	
-2	0.1238	
-1	0.1673	
0	0.2251	
1	0.2936	
2	0.3639	
3	0.4359	
4	0.5031	
5	0.5630	
6	0.6128	
7	0.6559	
8	0.6852	
9	0.6957	
10	0.6968	
11	0.6860	
12	0.6747	
13	0.6603	
14	0.6457	
15	0.6322	
16	0.6230	
17	0.6031	
18	0.5867	
19	0.5755	
20	0.5717	

总之，交叉相关图提示的变量次序为：inflation→fedfunds→unrate。从经济理论上，这可以解释为对通货膨胀的正冲击引起美联储提高基准利率，而后者导致失业率上升。当然，这并非唯一的理论解释。比如，根据宏观经济学中的菲利普斯曲线，失业率下降会引起通货膨胀上升，即 unrate→inflation；而根据泰勒规则（the Taylor rule），美联储在制定利率时，也会考虑失业率或产出缺口（output gap），即 unrate→fedfunds。并且，上文的格兰杰因果检验表明，inflation 不是 unrate 的格兰杰原因，与交叉相关图提示的变量次序相悖。因此，在计算脉冲响应函数与预测方差分解时，仍需变换变量次序以考察结果的稳健性。

下面考察脉冲响应函数，结果如图 20.7。

```
. irf create iuf,set(macrovar) step(20)
(file macrovar.irf created)
```

```
(file macrovar.irf now active)
(file macrovar.irf updated)
```

其中,命令“irf create”将计算所有与脉冲响应函数有关的变量与统计量。此命令将计算结果命名为“iuf”(此命名提示变量次序为 inflation, unrate, fedfunds),存入新建立的脉冲文件 macrovar.irf,将此脉冲文件激活为当前文件,并更新它。选择项“step(20)”意味着计算 20 期的脉冲响应函数,默认为“step(8)”。在此命令中,由于没有使用选择项“order(varlist)”来指定变量次序,故默认为上文使用命令 var 进行估计的变量次序,即 inflation, unrate, fedfunds(此次序为 Stock and Watson, 2001 所采用)。

脉冲文件 macrovar.irf 其实就是普通的 Stata 数据文件,只不过文件扩展名为 irf。比如,看一下该文件中存储了哪些变量:

```
. describe using macrovar.irf
```

```
Contains data
  obs:           189                          1 Sep 2013 10:32
 vars:            23
 size:        36,099
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
irf             double %10.0g                 impulse response function (irf)
step            int    %10.0g                 step
cirf            double %10.0g                 cumulative irf
oirf            double %10.0g                 orthogonalized irf
coirf           double %10.0g                 cumulative orthogonalized irf
sirf            double %10.0g                 structural irf
dm              double %10.0g                 dynamic multipliers
cdm             double %10.0g                 cumulative dynamic multipliers
stdirf          double %10.0g                 std error of irf
stdcirf         double %10.0g                 std error of cirf
stdoirf         double %10.0g                 std error of oirf
stdcoirf        double %10.0g                 std error of coirf
stdsirf         double %10.0g                 std error of sirf
stddm           double %10.0g                 std error of dm
stdcdm          double %10.0g                 std error of cdm
fevd            double %10.0g                 fraction of mse due to impulse
sfevd           double %10.0g                 (structural) fraction of mse due to
                                                impulse
mse             double %10.0g                 SE of forecast of response variable
stdfevd         double %10.0g                 std error of fevd
stdsfevd        double %10.0g                 std error of sfevd
irfname         str15  %15s                  name of results
response        str9   %9s                   response variable
impulse         str9   %9s                   impulse variable
-------------------------------------------------------------------------------
Sorted by:
```

上表中的“dm”为“动态乘数”(dynamic multipliers),指的是当 VAR 模型中的外生变量变动一单位时对内生变量的动态影响(本例无外生变量,故变量 dm 为空,无观测值)。也可用命令“use macro.irf”打开此脉冲文件(正如打开 Stata 数据文件),但这将覆盖内存中的原有数据。

下面根据此脉冲文件来画正交脉冲响应图(未正交的脉冲响应没有太大意义)。

```
. irf graph oirf,yline(0)
```

其中，选择项“yline(0)”表示在纵轴 $y=0$ 的位置画一条水平线，作为正交脉冲响应函数的参照线。

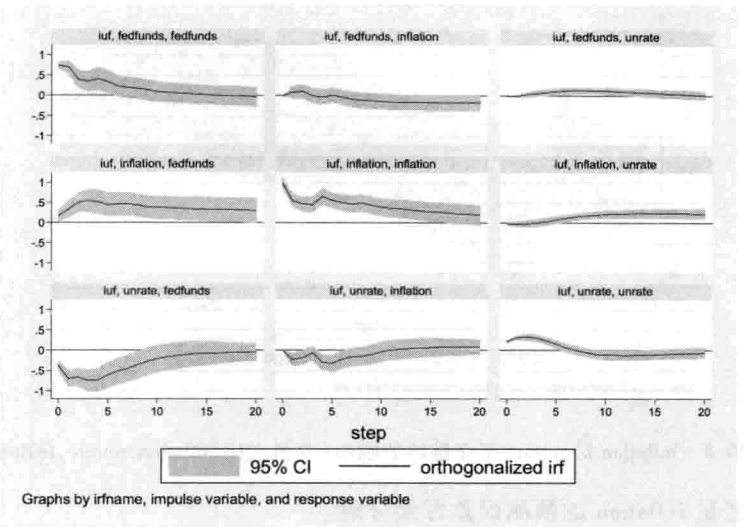


图 20.7　正交化脉冲响应图(变量次序 inflation, unrate, fedfunds)

图 20.7 包含 9 个小图，每个图的标题依次为“irfname”(此处为 iuf)，“impulse variable”(脉冲变量)，以及“response variable”(响应变量)。比如，第一行的三个小图均以 fedfunds 为脉冲变量，分别描绘 fedfunds 对 fedfunds, inflation 与 unrate 的动态效应。从第一行可以看出，联邦基金利率对于通货膨胀与失业率几乎没有作用，与预期的货币政策效应不符；这可能是因为乔利斯基分解将 fedfunds 排在变量次序的最后(参见下文变换次序的稳健性检验)。

类似地，第二行的三个小图分别描绘 inflation 对 fedfunds, inflation 与 unrate 的动态效应。从第二行可以看出，通货膨胀上升会引起美联储长久地提高联邦基金利率，并导致失业率上升。而且，最初的通胀冲击会引起通胀长期存在，即通胀的惯性(可能由于通胀预期)；换言之，通胀一旦形成则不容易根除。

第三行的三个小图分别描绘 unrate 对 fedfunds, inflation 与 unrate 的动态效应。失业率上升将引起美联储在未来三年里(约 12 个月)保持较低的联邦基金利率以刺激经济，而失业率上升还将通过菲利普斯曲线的作用使得通货膨胀降低。

也可以显示单个脉冲响应图。比如，想单独看 inflation 对 unrate 正交脉冲的响应，可输入如下命令(结果如图 20.8)：

```
. irf graph oirf,yline(0) i(unrate) r(inflation)
```

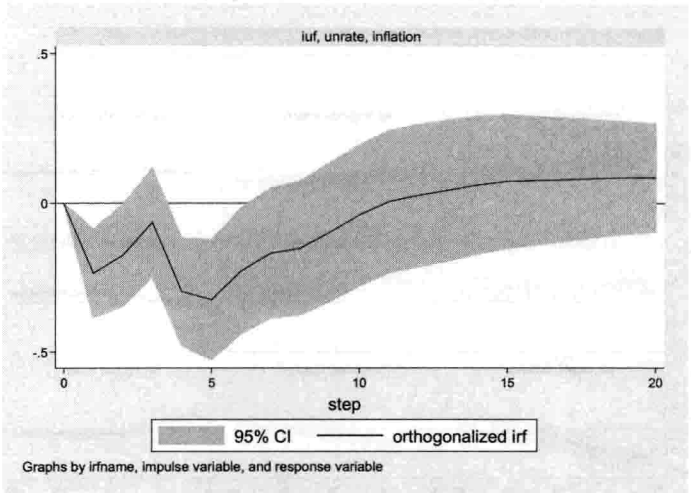


图 20.8 inflation 对 unrate 正交脉冲的响应(变量次序 inflation,unrate,fedfunds)

下面看一下变量 inflation 的预测误差方差分解。

```
. irf table fevd,r(inflation) noci
```

其中,选择项"r(inflation)"表示以 inflation 为响应变量,而选择项"noci"表示不显示置信区间(为了使结果更简洁)。

Results from iuf

step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	1	0	0
2	.953093	.04199	.004917
3	.934426	.054805	.010769
4	.939476	.050951	.009573
5	.913917	.078127	.007957
6	.887961	.105279	.00676
7	.881257	.112039	.006703
8	.878773	.112222	.009005
9	.87767	.11012	.01221
10	.877838	.106494	.015668
11	.877551	.102269	.020179
12	.87583	.098461	.025709
13	.873281	.095157	.031563
14	.870045	.092718	.037237
15	.865953	.091104	.042943
16	.861174	.09009	.048736
17	.856246	.089351	.054403
18	.851252	.088909	.059838
19	.846121	.088754	.065125
20	.840902	.088808	.07029

(1) irfname = iuf, impulse = inflation, and response = inflation
(2) irfname = iuf, impulse = unrate, and response = inflation
(3) irfname = iuf, impulse = fedfunds, and response = inflation

Results from iuf

step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	1	0	0
2	.953093	.04199	.004917
3	.934426	.054805	.010769
4	.939476	.050951	.009573
5	.913917	.078127	.007957
6	.887961	.105279	.00676
7	.881257	.112039	.006703
8	.878773	.112222	.009005
9	.87767	.11012	.01221
10	.877838	.106494	.015668
11	.877551	.102269	.020179
12	.87583	.098461	.025709
13	.873281	.095157	.031563
14	.870045	.092718	.037237
15	.865953	.091104	.042943
16	.861174	.09009	.048736
17	.856246	.089351	.054403
18	.851252	.088909	.059838
19	.846121	.088754	.065125
20	.840902	.088808	.07029

(1) irfname = iuf, impulse = inflation, and response = inflation
(2) irfname = iuf, impulse = unrate, and response = inflation
(3) irfname = iuf, impulse = fedfunds, and response = inflation

上表显示，对 inflation 进行向前 1 个季度的预测，其预测方差完全来自于 inflation 本身；即使向前作 20 季度的预测，也依然有 84.1% 的预测方差来自 inflation 本身，其余的 8.9% 与 7% 分别来自 unrate 与 fedfunds。这意味着，inflation 主要受自身的影响，变量 unrate 与 fedfunds 的作用很小（至少在目前的变量排序下）。也可以直观地将上述结果通过图示来表达，参见图 20.9。

```
. irf graph fevd,r(inflation)
```

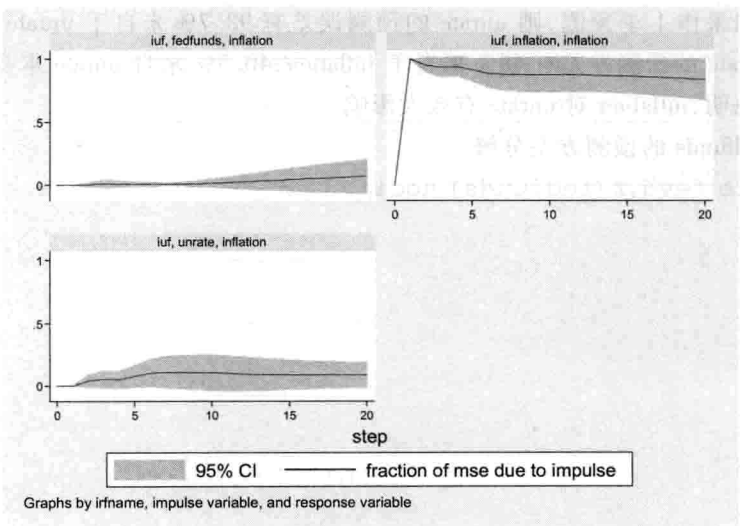


图 20.9 inflation 的预测方差分解图（变量次序 inflation，unrate，fedfunds）

下面考察 unrate 的预测方差分解。

```
. irf table fevd,r(unrate) noci
```

Results from iuf

step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	.01264	.98736	0
2	.009345	.990471	.000184
3	.007585	.988049	.004366
4	.00551	.980447	.014042
5	.007125	.963736	.029139
6	.018249	.9292	.052551
7	.043133	.876104	.080764
8	.080901	.811523	.107576
9	.126238	.745167	.128595
10	.174162	.684145	.141693
11	.219919	.632196	.147886
12	.261628	.589172	.149199
13	.299129	.553461	.147409
14	.332917	.523301	.143782
15	.363443	.497355	.139202
16	.391168	.474643	.134189
17	.416454	.454492	.129054
18	.439605	.436412	.123984
19	.460828	.42007	.119102
20	.480289	.405224	.114487

(1) irfname = iuf, impulse = inflation, and response = unrate

(2) irfname = iuf, impulse = unrate, and response = unrate

(3) irfname = iuf, impulse = fedfunds, and response = unrate

上表显示，如果作 1 季预测，则 unrate 的预测误差有 98.7% 来自于 unrate 自身。但如果作 20 季预测，则 unrate 的预测方差有 48% 来自于 inflation，40.5% 来自 unrate 本身，其余 11.4% 来自 fedfunds。这表明，inflation 对 unrate 有较大影响。

下面考察 fedfunds 的预测方差分解。

```
. irf table fevd,r(fedfunds) noci
```

Results from iuf

step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	.036696	.200998	.762306
2	.075248	.354469	.570283
3	.146628	.407824	.445548
4	.187531	.453521	.358948
5	.204253	.476606	.319141
6	.216086	.485133	.298782
7	.232608	.487134	.280258
8	.24991	.484575	.265515
9	.264633	.479535	.255832
10	.276911	.473724	.249365
11	.289319	.466917	.243764
12	.301628	.459708	.238664
13	.312878	.452849	.234273
14	.32306	.446469	.230471
15	.332846	.440272	.226882
16	.342299	.434262	.223438
17	.351197	.428579	.220224
18	.359497	.423241	.217263
19	.367347	.418152	.2145
20	.374771	.4133	.211929

(1) irfname = iuf, impulse = inflation, and response = fedfunds

(2) irfname = iuf, impulse = unrate, and response = fedfunds

(3) irfname = iuf, impulse = fedfunds, and response = fedfunds

上表显示，如果作 20 季预测，则 fedfunds 的预测方差大部分来自 unrate 与 inflation（分别占 41.3%与 37.5%），只有少部分源于 fedfunds 自身（占 21.2%）。这表明，fedfunds 受 unrate 与 inflation 的影响较大。

下面将所有的预测方差分解画图，结果参见图 20.10。

```
. irf graph fevd
```

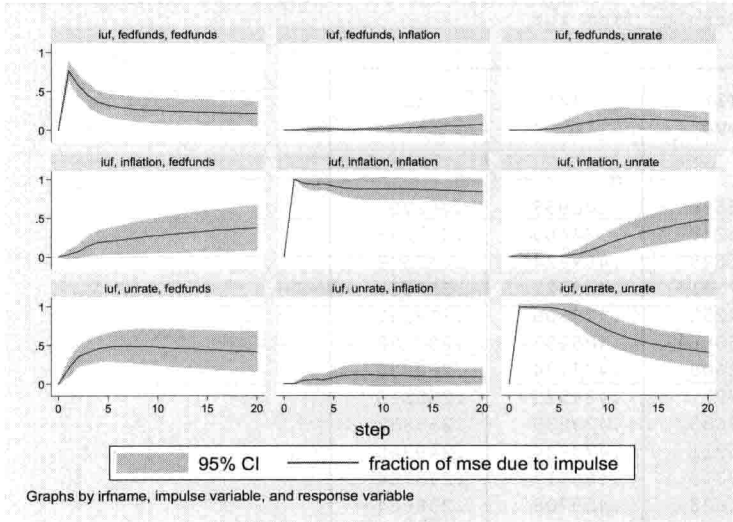


图 20.10 所有变量的预测方差分解图(变量次序 inflation,unrate,fedfunds)

由于以上的脉冲响应与方差分解结果依赖于变量次序,故下面将变量次序变为交叉相关图所推荐的次序 inflation,fedfunds,unrate,以考察结果的稳健性(对于三个变量的 VAR 系统,共有 6 种可能排序,其他情形从略)。为此,建立一个新的脉冲响应结果,并命名为 ifu。

```
. irf create ifu,order(inflation fedfunds unrate) step(20)
(file macrovar.irf updated)
```

下面比较在以上两种变量排序下,inflation 对 fedfunds 脉冲的响应,结果参见图 20.11。

```
.irf graph oirf,i(fedfunds) r(inflation) yline(0) noci
```

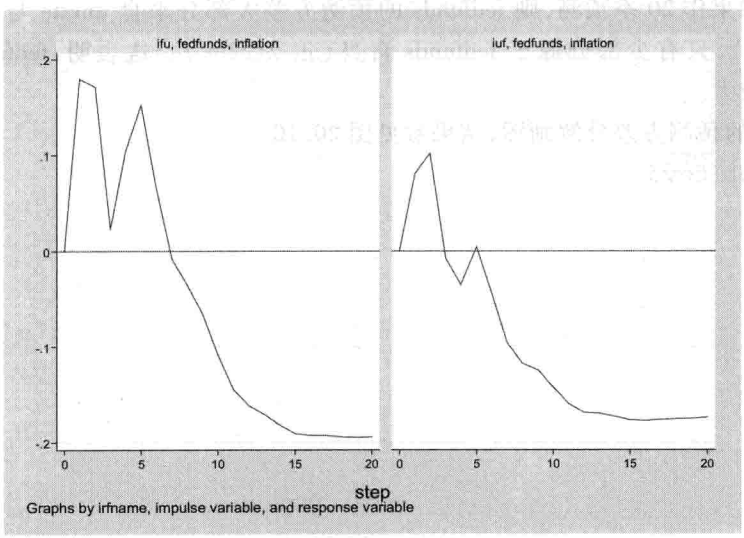


图 20.11 对比两种变量排序下,inflation 对 fedfunds 脉冲的响应

在图 20.11 中,左图为在变量排序 ifu 的情况下,fedfunds 对 inflation 的动态效应;而右图为在变量排序 iuf 的情况下,fedfunds 对 inflation 的动态效应。显然,在左图中,fedfunds 对 inflation

的作用幅度明显更大。

下面，比较在两种变量排序下，inflation 的预测方差来源于 fedfunds 的比重，结果参见图 20.12。

```
. irf graph fevd,i(fedfunds) r(inflation) noci
```

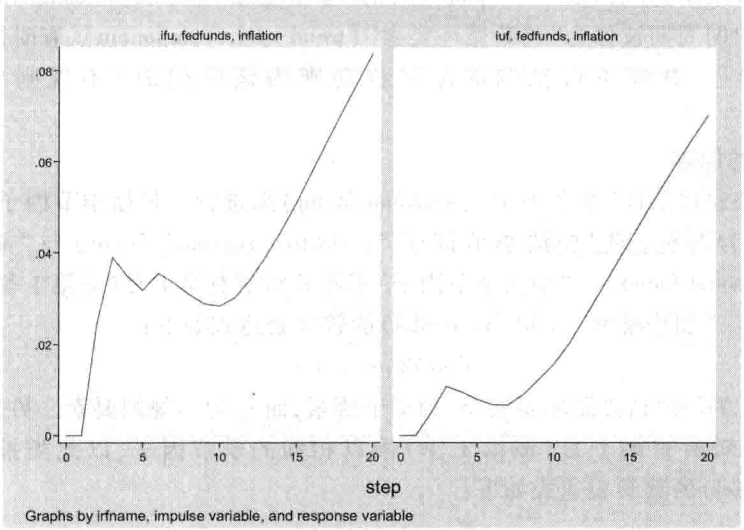


图 20.12 对比两种变量排序下，unrate 对 fedfunds 脉冲的响应

在图 20.12 中，左图为在变量排序 ifu 的情况下，inflation 预测方差中来源于 fedfunds 的比重，而右图则为在变量排序 iuf 的情况下，inflation 预测方差中来源于 fedfunds 的比重。显然，在左图中，fedfunds 对于 inflation 的作用明显大于右图。

总之，上述结果显示，无论脉冲响应函数还是预测方差分解，都在一定程度上依赖于变量排序。因此，在使用或解释其结果时应十分谨慎。对于 n 个变量的 VAR 系统，共有 $n!$种可能的变量排序。如果无法确定唯一的变量排序，则应对可能的变量排序进行稳健性检验。

20.15 季节调整

1. 季节效应

对于月度或季度时间序列，常常需要对其进行“季节调整”(seasonal adjustment)，去掉“季节效应”后才能使用。比如，考察中国的季度 GDP 数据。由于第一季度包含春节，故通常第一季度的 GDP 偏低。如果直接以第二季度 GDP 除以第一季度 GDP 来计算环比增长率，则会高估第二季度的 GDP 增长率；同样道理，将第一季度 GDP 除以上年第四季度 GDP 则会低估第一季度的 GDP 增长率。总之，包含季节效应的时间序列不能直接计算环比增长率。如果不进行季节调整，则只能计算同比增长率，即与去年同一季(月)相比。年度数据不需要进行季节调整。

可能导致季节效应的因素包括：

(1) 天气因素。比如，在冬季由于取暖而增加能源消耗。

(2) 行政因素。比如，学校开学与放假的日期对交通量的影响。

(3) 固定假日。比如,十一国庆节对旅游与交通的影响。

(4) 移动假日(moving holiday)。比如,春节期间,铁路运输量增加而 GDP 下降。

(5) 日历因素。比如,闰年与闰月的影响。

(6) 交易日效应。比如,五金店销售额在有五个周末的月份高于只有四个周末的月份。

所有这些季节因素共同构成一个时间序列的"季节要素"(seasonal component)。该时间序列的长期走势与中期周期被称为"趋势循环要素"(trend cycle component),有时简称"趋势要素"(trend component)[①]。其他不可预测的随机扰动则为该序列的"不规则要素"(irregular component)。

2. 季节调整的原理

季节调整通常通过估计"季节因子"(seasonal factor)来进行。根据季节因子起作用的方式,季节因子主要分为两种,即"加法季节因子"(additive seasonal factor)与"乘法季节因子"(multiplicative seasonal factor)。"加法季节因子"意味着对所有第 1 月(或第 1 季)加上相同的季节因子,以此类推。"加法模型"(additive model)的数学表达式如下:

$$Y_t = TC_t + S_t + I_t \tag{20.92}$$

其中,Y_t 为原序列,TC_t 为趋势循环要素,S_t 为季节要素,而 I_t 为不规则要素。另一方面,"乘法季节因子"则意味着对所有第 1 月(或第 1 季)乘以相同的季节因子,以此类推。"乘法模型"(multiplicative model)的数学表达式如下:

$$Y_t = TC_t \times S_t \times I_t \tag{20.93}$$

使用乘法模型要求 Y_t 序列中不包含零或负数。对方程(20.93)两边同时取对数可得:

$$\ln Y_t = \ln TC_t + \ln S_t + \ln I_t \tag{20.94}$$

方程(20.94)在形式上与加法模型相同,故称为"对数加法模型"。季节调整的目标就是将原序列 Y_t 分解为趋势循环要素、季节要素与不规则要素,然后去掉季节要素 S_t,得到季节调整序列(seasonally adjusted series)。季节调整的具体方法有多种,使用不同方法,会得到不同的季节调整序列,带有一定的主观性;这是季节调整的局限性。因此,国外的统计部门一般同时公布原序列与季节调整序列。下面以加法模型为例,介绍常用的回归法、移动平均比率法与 X12 方法。

(1) 回归法。[②] 回归法的基本步骤为,首先生成月度(或季度)虚拟变量,然后把时间序列对这些虚拟变量进行 OLS 回归,所得到的残差就是经季节调整后的序列。

例 Kennan(1985)在使用美国制造业的月度产量数据时,首先取对数,然后将其对时间、时间平方以及月度虚拟变量进行回归,所得残差即为经过季节调整并去掉时间趋势的产量对数。

下面以 turksales.dta 为例,该数据集包括 1990 年第 1 季至 1999 年第 4 季的火鸡销售额(sales)与时间变量(t)。首先看一下火鸡销售额的时间趋势,结果参见图 20.13。

```
. use turksales.dta,clear
. tsline sales
```

从图 20.13 可知,火鸡销售额有时明显的季节波动。其中,第四季度的火鸡销售最好,因为感恩节在第四季度,而火鸡是感恩节的传统食物;第一与第二季度的销售较为平淡。

为了生成季度虚拟变量,首先从时间变量 t 中提取季度信息,记为变量 quarter。

```
. gen quarter = quarter((dofq(t)))
```

---

① 事实上,很难区分"趋势"与"周期",人为的区分总是带有主观性,故将此二者并称。

② 有关季节调整的回归法,主要借鉴了 Baum(2006)。

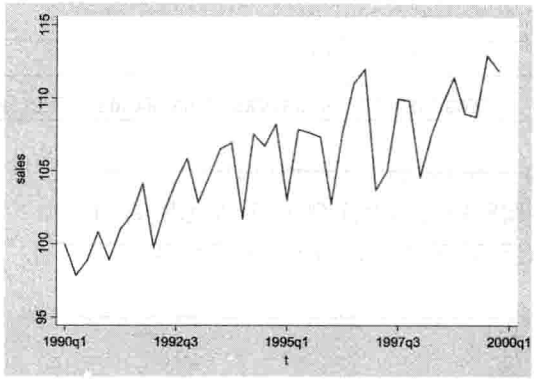


图 20.13 火鸡销售额的时间趋势

其次,使用命令 tab 来生成季度虚拟变量。

```
. tab quarter,gen(q)
```

quarter	Freq.	Percent	Cum.
1	10	25.00	25.00
2	10	25.00	50.00
3	10	25.00	75.00
4	10	25.00	100.00
Total	40	100.00	

然后,以第 1 季度为参照值,把变量 sales 对常数项以及第 2 ~ 4 季度虚拟变量进行回归。

```
. reg sales q2 - q4
```

Source	SS	df	MS	Number of obs =	40
				F(3, 36) =	4.03
Model	161.370376	3	53.7901254	Prob > F =	0.0143
Residual	480.52796	36	13.3479989	R-squared =	0.2514
				Adj R-squared =	0.1890
Total	641.898336	39	16.4589317	Root MSE =	3.6535

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
q2	2.389294	1.633891	1.46	0.152	-.9243908	5.702978
q3	4.33511	1.633891	2.65	0.012	1.021425	7.648794
q4	5.232047	1.633891	3.20	0.003	1.918362	8.545731
_cons	102.6287	1.155335	88.83	0.000	100.2856	104.9719

上表中的 F 统计量的 p 值为 0.014 3,这个回归方程高度显著。而且 $R^2 = 0.25$,表明季度虚拟变量对于变量 sales 有较强的解释力。为了获得经季度调整的序列,下面使用命令 predict 来获得上述回归的残差项(记为 sales_sa)。

```
.predict sales_sa,r
```

然而,OLS 残差项的平均值一定为 0,故需要把原序列的均值加回去,并记季节调整序列为 sales_sa_reg。

```
. sum sales
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sales	40	105.6178	4.056961	97.84603	112.9617

```
.gen sales_sa_reg = sales_sa + r(mean)
```

下面,将回归法的季节调整序列与原序列画图,参见图 20.14。

```
. tsline sales_sa_reg sales,lpattern("_")
```

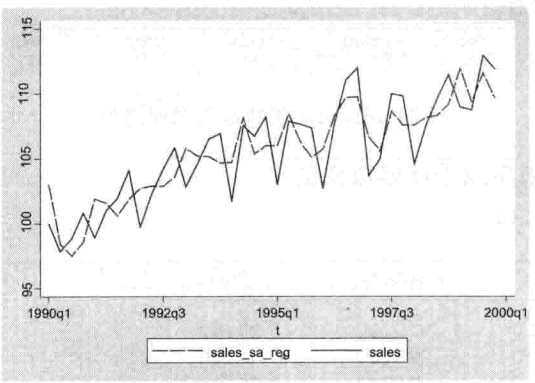


图 20.14 回归法的季节调整效果

从图 20.14 可以看出,经回归法进行季节调整后,序列变得更为光滑,季节性波动的特征已不太明显。回归法的缺点是,它假设每年的季节调整因子均不变(可以理解为季节固定效应),但对于移动假日(比如,春节有时落在 1 月,有时落在 2 月),每年的季节调整因子可能不尽相同。为此,下面转而介绍移动平均比率法。

(2) 移动平均比率法。移动平均比率法的基础是移动平均(Moving Average)。奇数项移动平均很容易进行。考虑数据$\{y_1,\cdots,y_T\}$,则时点 t 的 3 项移动平均为

$$MA_t=\frac{y_{t-1}+y_t+y_{t+1}}{3}\quad(t=2,\cdots,T-1) \tag{20.95}$$

显然,进行 3 项移动平均后的序列 MA_t 缺失第一项与最后一项,可通过插值来补齐,比如,$MA_1=(2y_1+y_2)/3$,而 $MA_T=(2y_T+y_{T-1})/3$。消除季节影响的简单方法为,对季度数据进行 4 项移动平均;而对月度数据进行 12 项移动平均。然而,由于 4 与 12 均为偶数,故进行以下“中心化移动平均”。下面以季度数据为例,求 $t=3$ 时点上的 4 项平均。

$$MA_{2.5}=\frac{y_1+y_2+y_3+y_4}{4} \tag{20.96}$$

$$MA_{3.5}=\frac{y_2+y_3+y_4+y_5}{4} \tag{20.97}$$

定义相应的中心化移动平均值为

$$MA_3=\frac{1}{2}(MA_{2.5}+MA_{3.5})=\frac{1}{2}\left(\frac{y_1+y_2+y_3+y_4}{4}+\frac{y_2+y_3+y_4+y_5}{4}\right) \tag{20.98}$$

以此类推,中心化移动平均的一般公式为

$$MA_t = \frac{1}{2}\left(\frac{y_{t-2}+y_{t-1}+y_t+y_{t+1}}{4}+\frac{y_{t-1}+y_t+y_{t+1}+y_{t+2}}{4}\right) = \frac{1}{4}\left(\frac{y_{t-2}}{2}+y_{t-1}+y_t+y_{t+1}+\frac{y_{t+2}}{2}\right) \tag{20.99}$$

从公式(20.99)可知,中心化移动平均也可视为加权平均,即各期权重不完全相同。类似地,可以计算适用于月度数据的12项中心化移动平均。下面主要以加法模型为例,说明季节调整的"移动平均比率法"(ratio to moving average)。该法分为以下五个步骤:

① 对季度(月度)数据 Y_t 进行4项(12项)中心化移动平均,得到趋势循环序列 TC_t。

② 计算季节要素 S_t 与不规则要素 I_t 之和:$SI_t \equiv S_t + I_t = Y_t - TC_t$。注:如果使用乘法模型,则 $SI_t \equiv S_t \times I_t = Y_t / TC_t$。

③ 对于季(月)度数据的 SI_t,分别计算其第 j 季(月)的季(月)度平均值,得到季(月)节因子 s_t。例如,如果 $j=2$,则计算 SI_t 序列在整个数据期间所有第2季(月)数据的平均值。

④ 调整季节因子,使得它们的和为0,得到标准化的季节因子 $S_t \equiv s_t - \frac{1}{k}\sum_{j=1}^{k} s_j$,其中 $k=4$(季度数据),或 $k=12$(月度数据)。注:如果使用乘法模型,则 $S_t \equiv \frac{s_t}{\sqrt[k]{s_1 \times \cdots \times s_k}}$。

⑤ 计算季节调整的最终结果:$TCI_t \equiv Y_t - S_t$。注:如果使用乘法模型,则 $TCI_t \equiv Y_t / S_t$。

其中,第③—④步的目的是将季节序列 S_t 从序列 $SI_t \equiv S_t + I_t$ 中分离出来。

下面继续以数据集 turksales. dta 为例,使用加法模型进行移动平均比率法的季节调整。首先进行中心化移动平均,得到趋势循环序列 TC_t。

```
. tssmooth ma sales1 = sales,window(2 1 1)
```

此命令表示对序列 sales 进行移动平均,并将结果记为 sales1;选择项"window(2 1 1)"表示在移动平均时,包括过去2项,当期1项,以及未来1项。

```
The smoother applied was
     (1/4)*[x(t-2) + x(t-1) + 1*x(t) + x(t+1)]; x(t)= sales
```

```
. tssmooth ma sales2 = sales,window(1 1 2)
```

其中,选择项"window(1 1 2)"表示在移动平均时,包括过去1项,当期1项,以及未来2项。然后进行中心化。

```
The smoother applied was
     (1/4)*[x(t-1) + 1*x(t) + x(t+1) + x(t+2)]; x(t)= sales
```

```
.gen sales_tc = (sales1 + sales2) / 2
```

再将原序列减去趋势循环序列 TC_t。

```
. gen sales_si = sales - sales_tc
```

下面,通过循环语句 forvalues 来计算每一季度的平均值,即季节因子 s_i[①]。

```
. forvalues i = 1 / 4 {
  2. quietly sum sales_si if q`i'= =1
  3. scalar s`i'= r(mean)
```

① 也可以不使用循环语句,而直接对每一季度计算一次。但如果是月度数据,需要重复12次,比较麻烦。

```
  4. }
```

然后将这四个季度的平均值再进行平均。

```
. scalar s_mean = (s1 + s2 + s3 + s4) / 4
```

接下来,再以循环语句将每个季节因子减去季节因子的平均值,得到标准化的季节因子,然后将原序列减去标准化的季节因子,即得到季节调整序列。

```
. gen sales_sa_ma = 0  (初始化季节调整序列为 0,以便后面更新)
. forvalues i = 1 / 4 {
  2. scalar s`i' = s`i' - s_mean
  3. replace sales_sa_ma = sales - s`i' if q`i' = = 1
  4. }
```

最后,看一下移动平均比率法的季节调整序列与原序列的时间序列图,结果见图 20.15。

```
.tsline sales_sa_reg sales_sa_ma sales,lpattern("_"" - ")
```

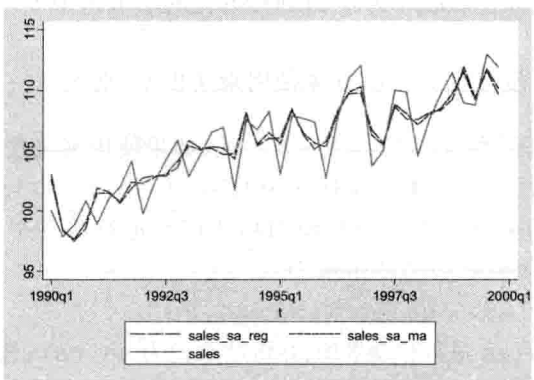


图 20.15 回归法与移动平均比率法的季节调整结果比较

从图 20.15 可以看出,经过回归法与移动平均比率法的季节调整序列十分接近。为了与加法模型对比,下面使用乘法模型(相关变量名加后缀 *m* 以示区别,*m* 表示 multiplicative)。

```
. gen sales_sim = sales /sales_tc
. forvalues i = 1 / 4 {
  2.quietly sum sales_sim if q`i' = = 1
  3. scalar sm`i' = r(mean)
  4. }
. scalar s_meanm = (sm1 * sm2 * sm3 * sm4)^0.25
. gen sales_sa_mam = 0  (初始化季节调整序列为 0,以便后面更新)
. forvalues i = 1 / 4 {
  2. scalar sm`i' = sm`i' / s_meanm
  3. replace sales_sa_mam = sales / sm`i' if q`i' = = 1
  4. }
```

下面,对比加法模型与乘法模型的季节调整结果,参见图 20.16。

```
.tsline sales_sa_ma sales_sa_mam,lpattern("_")
```

从图 20.16 可知,乘法模型与加法模型的季节调整结果十分接近,没有实质区别。但二者的

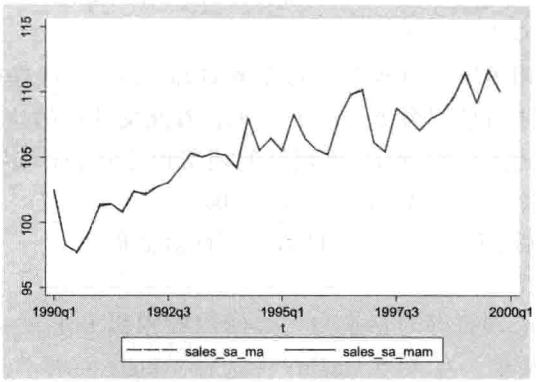


图 20.16 加法模型与乘法模型的季节调整结果对比

取值并不完全相同。下面看一下前 5 个观测值。

```
. list sales_sa_ma sales_sa_mam in 1/5
```

	sales~ma	sales~am
1.	102.4991	102.3909
2.	98.29412	98.24789
3.	97.64602	97.74304
4.	99.07458	99.17122
5.	101.4089	101.2747

(3) X12 方法。1954 年美国商务部人口普查局首次开发了可在计算机上运行的季节调整程序,称为 X1[①]。以后该程序每次改进都以 X 加上序号表示。1965 年推出基于移动平均法的 X11,很快成为全世界统计机构进行季节调整的标准方法。X11 的基本方法为通过多次移动平均来分解原序列,并对极端值(outlier)进行自动调整[②],以得到每一步都更为准确的结果。对于时间序列的两端观测值,X11 进行单边移动平均(one-sided moving average),但这种非对称滤波法(asymmetric filter)的质量不高。为此,加拿大统计局于 1980 年提出 X11 - ARIMA,使用 ARIMA 建模方法对原序列进行前推(forecast)与后推(backcast),然后对加长的序列进行 X11 季节调整。

美国人口普查局于 1998 年推出 X12 - ARIMA,基本方法与 X11 - ARIMA 一致,主要增加了对交易日、节假日影响的调节功能,以及对各种极端值的处理。这些极端值包括"单点异常值"(Additive Outlier,简记 AO),即发生于某时刻的异常值;"水平移动"(Level Shift,简记 LS),表示在瞬间变化到一个新水平并维持在此新水平上;以及"暂时变化"(Temporary Change,简记 TC),表示在瞬间变化到一个新水平,但逐渐恢复到原水平。X12 - ARIMA 分为两个模块,即 regARIMA 与 X11。模块 regARIMA 使用线性回归的方法(带 ARIMA 扰动项)来处理交易日、移动假日与极端值的影响,并对原序列进行前推与后推拓展。

在 Stata 12 中[③],可通过下载南开大学王群勇老师撰写的 `sax12` 命令来实现 X12 - ARIMA。

---

① X 表示 eXperimental。

② 极端值会影响季节调整的准确性。在季节调整后,再把极端值加回去。

③ 在 Stata 11 中也可以运行命令 `sax12`,但无法运行后续命令 `sax12im`。

下载方法为(或输入命令"findit sax12",找到此命令后下载):

```
.net install st0255.pkg
```

命令 sax12 需要调用美国人口普查局的程序 x12a.exe。下载程序 x12a.exe 后[①],将其放入 Stata 的 PLUS 文件夹即可(可使用命令 sysdir 来查看此文件夹的位置)。由于命令 sax12 包含很多选择项,且有些选项无法同时使用,故建议使用菜单方式(menu-driven)来执行此命令。有关此命令的详细说明与示例,参见 Wang and Wu(2012)。

下载命令 sax12 之后,输入以下命令即可打开此对话框。

```
.db sax12
```

其中,db 表示 dialog begin,即打开对话窗口。继续以数据集 turksales.dta 中的变量 sales 为例。对话窗口提供了一系列菜单。在菜单"main"中选择"single series",然后在 Variable 栏中选择 sales,作为季节调整的对象。菜单"if/in"用于限制限制样本区间,在此忽略。菜单"prior"用于指定对原序列先进行某种变换(比如,取对数),再进行季节调整。为了与上文的做法一致,此处不进行任何变换,即在"Data transformation"栏中选择"none"[②]。在菜单"regression"中选择"constant"(常数项)与"trading day with leap year"以剔除贸易日与闰年的影响。在菜单"regression"之下,还允许设定"User defined variables",比如用于捕捉中国数据中的春节效应;在本例中不作设定。在菜单"outlier"中选择所有的极端值种类,即 AO,LS 与 TC;并选择"Automatic selection given maximum lag and difference"。在菜单"adjustment"中选择"X11 mode:additive"[③]。菜单"other"提供了对季节调整结果的诊断方法。最后点击 OK。

此时,会发现 Stata 命令窗口出现如下命令:

```
. sax12 sales, satype(single) inpref(sales.spc) outpref(sales) transfunc(none) regpre(const td) outauto(ao ls tc) outlsrun(0) ammaxlag(21) ammaxdiff(2 1) ammaxlead(12) x11mode(add) x11seas(x11default)
```

显然,此命令的选择项过多,故适合进行菜单操作。为了调用(import)季节调整序列,打开另一菜单对话框:

```
. db sax12im
```

在此对话框中,点击 Browse,找到文件 sales.out 之后将其打开。在"Select the series or insert the suffix"选项中选择"seasonally adjusted series(d11)",然后点击 OK。此时,会发现变量 sales_d11 出现在变量窗口,这就是经过 X12-ARIMA 季节调整后的序列。下面,对比 X12-ARIMA 与移动平均比率法的季节调整结果,参见图 20.17。

```
. tsline sales_sa_ma sales_d11,lpattern("-")
```

从图 20.17 可知,经 X12-ARIMA 季节调整的序列(图中实线)与移动平均比率法(图中虚线)在某些区间十分类似,但在有些区间则前者明显比后者更为平滑,这可能是因为 X12-ARIMA 进行了多次移动平均,并且处理了极端值。

X12 季节调整也可以在 Eviews 中实现。假设数据存于 Excel 表中,基本步骤如下:File→Open→Eviews Workfile→文件类型选择"Excel file(*.xls)"→打开→双击 Workfile 中需要进行季节调整的变量→点击 proc→Seasonal Adjustment→Census X12→根据对话框进行选项,包括指

① 下载地址为 http://www.census.gov/srd/www/x12a/x12downv03_pc.html。

② 如果选择"Auto(log or none)"则将根据信息准则自动选择是否先对原序列取对数,再进行季节调整。

③ 这也是为了与上文的加法模型相一致。如果使用乘法模型,则选择"multiplicative"。

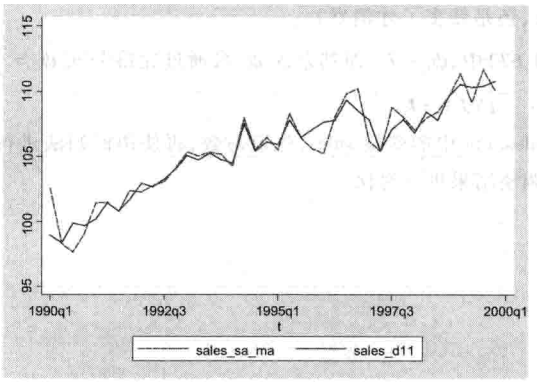


图 20.17　对比移动平均比率法与 X12 - ARIMA

定需要保存的序列(component series to save)。详见高铁梅(2009)。

需要注意的是,Eviews 中的 X12 程序仅针对美国的节假日而设,并不完全适用于中国。为此,2009 年国家统计局与南开大学数量经济研究所组成联合研究小组,开发了针对中国节假日的 NBS - SA 季节调整软件。以此软件为基础,国家统计局于 2011 年 4 月开始发布 GDP 等指标的环比数据,结束了中国基本不生产环比统计数据的历史。Wang and Wu(2012)介绍了一个实例,利用 Stata 命令 sax12 来处理中国数据的春节效应。

总之,X12 依然是目前最权威的季节调整方法,为美国、英国、加拿大等统计局所采用。有关季节调整的更详细说明,参见栾惠德、张晓峒(2007),高铁梅(2009),以及《时间序列 X - 12 - ARIMA 季节调整——原理与方法》(中国人民银行调查统计司,2006)。

习　题

20.1　(部分调整模型,partial adjustment model)记 y_t^* 为 y_t 的最优值或理想值(比如,y_t 为资本存量,而 y_t^* 为最优资本存量),满足以下关系,$y_t^* = \alpha + \beta x_t + u_t$,其中 u_t 为独立于 $\{x_t\}$ 的扰动项。假设 y_t 不能立即调整到最优值 y_t^*,而只能进行部分调整,即 $y_t - y_{t-1} = \theta(y_t^* - y_{t-1})$,其中 $0 < \theta < 1$。证明 $\{y_t\}$ 为 ADL 模型。

20.2　(使用 Yule-Walker equations 求解自协方差)对于 AR(2)模型,$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$,其中 ε_t 为白噪声,对此方程两边分别求与 y_{t-k} 的协方差,其中 $k = 0,1,2$,得到关于 $\gamma_0, \gamma_1, \gamma_2$(0,1,2 阶自协方差)的三元一次方程组。对于 $k > 2$,给出 γ_k 的递推公式(将 γ_k 表示为 γ_{k-1} 与 γ_{k-2} 的函数)。

20.3　假设模型为 $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_0 x_t + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \varepsilon_t$,推导其对应的误差修正模型。提示:假设 (x,y) 之间的长期均衡关系为 $y = \phi + \theta x$。

20.4　使用美国 1960—2002 年季度数据集 macro_swatson. dta,估计如下的经验菲利普斯曲线 ADL(2,1)模型:

$$\Delta \text{ inf}_t = \beta_0 + \beta_1 \Delta \text{ inf}_{t-1} + \beta_2 \Delta \text{ inf}_{t-2} + \beta_3 \text{ unem}_{t-1} + \varepsilon_t \tag{20.100}$$

其中,inf 为通货膨胀率,unem 为失业率。使用 BG 检验,检验是否存在一阶自相关;如果有,则使用滞后一阶的 HAC 标准误。$\hat{\beta}_3$ 的符号是否与宏观理论相一致?

20.5　数据集 lutkepohl2. dta 包含了联邦德国 1960—1982 年的以下季度宏观变量:inv(投资),inc(收入),consump(消费),ln_inv(投资之对数),ln_inc(收入的对数),ln_consump(消费的对数),dln_inv(ln_inv 的一阶差分),dln_inc(ln_inc 的一阶差分),dln_consump(ln_consump 的一阶差分)。估计一个关于(dln_inv,dln_inc,dln_consump)的 VAR 模型,并进行相关的检验。

20.6　使用数据集 consumption_china. dta 估计一个有关(c,y)的 VAR 模型,画脉冲响应图,并进行格兰杰因

果检验(究竟是消费拉动增长,还是钱多了才消费)。

20.7 证明在表达式(20.67)中,$\boldsymbol{\psi}_0 \equiv \boldsymbol{I}_n$,而其余的 $\boldsymbol{\psi}_i$ 可通过递推公式 $\boldsymbol{\psi}_i = \sum_{j=1}^{i} \boldsymbol{\psi}_{i-j}\boldsymbol{\Gamma}_j$ 来确定。提示:使用关系式$(\boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots)\boldsymbol{\Gamma}(L) = \boldsymbol{I}_n$。

20.8 对于数据集 turksales. dta 中的变量 sales,先取对数,再使用回归法进行季节调整,并将此结果通过画图与正文中不取对数的季节调整结果进行对比。

第 21 章　单位根与协整

21.1　非平稳序列

如果一个时间序列不是平稳过程，则称为“非平稳序列”（non-stationary time series）。在以下几种情况下，都有可能出现非平稳序列。

（1）确定性趋势：如果一个时间序列有一个“确定性趋势”（deterministic trend），则为非平稳序列。比如，$y_t=\beta_0+\beta_1 t+\varepsilon_t$。显然，$\mathrm{E}(y_t)=\beta_0+\beta_1 t$ 随时间而改变，故不是平稳序列。对于这种非平稳序列，只要把时间趋势去掉，就变成平稳序列，故称为“趋势平稳”（trend stationary）序列。

（2）结构变动（structural break）：如果一个时间序列存在结构变动，则为非平稳序列。对此，可用邹检验（Chow test）进行检验，参见第 9 章。

（3）随机趋势：另一种导致非平稳的趋势为“随机趋势”（stochastic trend）。比如，随机游走模型（random walk）：

$$y_t=y_{t-1}+\varepsilon_t \tag{21.1}$$

其中，$\{\varepsilon_t\}$ 为白噪声。由于 $\Delta y_t=\varepsilon_t$，故来自 $\{\varepsilon_t\}$ 的任何扰动对 $\{y_t\}$ 都具有永久性的冲击，其影响力不随时间而衰减，故称 $\{\varepsilon_t\}$ 为这个模型的“随机趋势”。在上式中，如果包含常数项，则为“带漂移的随机游走”（random walk with drift）：

$$y_t=\beta_0+y_{t-1}+\varepsilon_t,\quad \beta_0\neq 0 \tag{21.2}$$

其中，β_0 为每个时期的平均“漂移”（drift），因为 $\mathrm{E}(y_t)=\beta_0+\mathrm{E}(y_{t-1})$。显然，随机游走是 AR(1) 的特例。对于 AR(1)，$y_t=\beta_0+\beta_1 y_{t-1}+\varepsilon_t$，如果 $\beta_1=1$，则为随机游走。对于随机游走，只要对其进行一阶差分，就可以得到平稳序列，故也称为“差分平稳”（difference stationary）序列。

定义　称平稳的时间序列为“零阶单整”（Integrated of order zero）[①]，记为 I(0)。如果时间序列的一阶差分为平稳过程，则称为“一阶单整”（Integrated of order one），记为 I(1)，也称为“单位根过程”（unit root process）。更一般地，如果时间序列的 d 阶差分为平稳过程，则称为“d 阶单整”（Integrated of order d），记为 I(d)。

对于 I(0) 序列，由于它是平稳的，故长期而言有回到其期望值的趋势。这种性质被称为“均值回复”（mean-reverting）。非平稳的 I(1) 序列则会“到处乱跑”（wander widely），没有上述性质。另外，I(0) 序列对于其过去的行为只有有限的记忆，即发生在过去的扰动项对未来的影响随时间而衰减；而 I(1) 序列则对过去的行为具有无限长的记忆，即任何过去的冲击都将永久性地改变未来的整个序列。比如，假设 $\{y_t\}$ 是 GDP 的时间序列，且为 I(1)，则任何对货币政策或财政政

① 由于只考虑单一变量，故称为“单整”，与后面的多变量“协整”概念相呼应。

策的调整都将对未来的 GDP 产生永久(permanent)的影响①。

定义 如果时间序列$\{y_t\}$的 d 阶差分为平稳的 ARMA(p,q)过程,则称$\{y_t\}$为 ARIMA(p,d,q)过程。最常见的为 ARIMA$(p,1,q)$,即经过一次差分就得到平稳的 ARMA(p,q)。

21.2 ARMA 的平稳性

在什么情况下,ARMA(p,q)才平稳呢?显然,MA(q)是平稳的,因为它是有限个白噪声的线性组合。因此,ARMA(p,q)的平稳性取决于其 AR(p)的部分。从第 5 章已经知道,对于AR(1),$y_t=\beta_0+\beta_1 y_{t-1}+\varepsilon_t$,如果 ε_t 平稳且$|\beta_1|<1$,则为平稳过程。更一般地,考虑 AR(p)的平稳性,即$y_t=\beta_0+\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}+\varepsilon_t$。

回顾 AR(1)的情形,"$y_t=\beta_0+\beta_1 y_{t-1}+\varepsilon_t$"其实是一阶随机差分方程,其稳定性与对应的确定性差分方程"$y_t=\beta_0+\beta_1 y_{t-1}$"是一样的。因此,只要考虑一阶差分方程"$y_t=\beta_0+\beta_1 y_{t-1}$"是否有稳定解即可,而这个非齐次(含常数项$\beta_0$)差分方程的解取决于对应的齐次(不含常数项)差分方程"$y_t=\beta_1 y_{t-1}$"的通解 $y_t=y_0\beta_1^t$(解的形式为指数函数)。因此,其稳定条件为$|\beta_1|<1$。

对于 AR(p),考虑其对应的确定性齐次差分方程:$y_t=\beta_1 y_{t-1}+\cdots+\beta_p y_{t-p}$。假设其解的形式仍为指数函数,即 $y_t=z^{-t}=(1/z)^t$,其中 z 待定。将此解代入差分方程可得

$$z^{-t}-\beta_1 z^{-(t-1)}-\cdots-\beta_p z^{-(t-p)}=0 \tag{21.3}$$

将上式两边同乘以 z^t 可得特征方程:

$$\phi(z)\equiv 1-\beta_1 z-\cdots-\beta_p z^p=0 \tag{21.4}$$

这个多项式方程在复数域中一定有 p 个根(包括重根)。与此对应,齐次差分方程也有 p 个形如$(1/z)^t$的解②,而其通解则是这 p 个解的线性组合。给定初始条件$\{y_0,y_1,\cdots,y_{p-1}\}$,则可求出此齐次差分方程的唯一特解。显然,如果要求$\{y_t\}$收敛于一个稳定值,则特征方程所有解的范数$\|z\|$(即在复平面上 z 离原点的距离)都必须大于 1,故所有解必须都落在复平面上的单位圆之外,参见图 21.1③。如果某个根正好落在单位圆之上,则称为"单位根"(unit root),比如随机

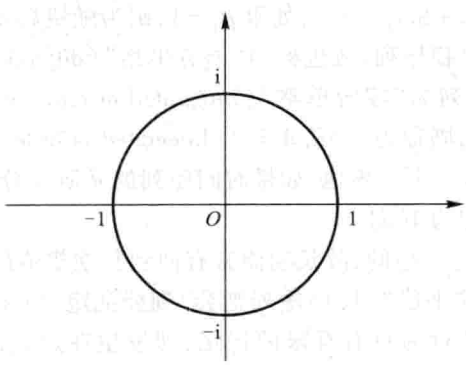


图 21.1 复平面上的单位圆

① 这是一个很强的结论,因为通常我们认为货币政策或财政政策会对宏观经济产生持久(persistent)的影响(自回归系数β_1接近于 1),即其作用将持续一段时间,但不会是永久的(permanent)。

② 如果有重根,则其解的形式为 $t^m(1/z)^t$,其中 m 表示共有 m 重根。这并不影响有关 AR(p)稳定性的结论。

③ 如果将特征方程定义为"$z^p-\beta_1 z^{p-1}-\cdots-\beta_p=0$",则结论与此相反。

游走的情形。如果特征方程的某个根落在单位圆之内，则为爆炸式(explosive)增长的非平稳过程。

例 对于 AR(1)，其特征方程为 $1-\beta_1 z=0$，故 $z=1/\beta_1$。因此 $\|z\|=|z|>1\Leftrightarrow|\beta_1|<1$。显然，有关 AR($p$)稳定性的结论是对 AR(1)情形的推广。

21.3 VAR 的平稳性

有关一维 AR(p)的平稳性条件可以推广到多维 VAR(p)的情形。考虑以下 VAR(p)模型：

$$\boldsymbol{y}_t=\boldsymbol{\Gamma}_0+\boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p}+\boldsymbol{\varepsilon}_t \tag{21.5}$$

其中，$\{\boldsymbol{\varepsilon}_t\}$为向量白噪声过程。可以证明(参见 Lutkepohl,2005)，如果对于复数 z，特征方程

$$|\boldsymbol{I}_n-\boldsymbol{\Gamma}_1 z-\cdots-\boldsymbol{\Gamma}_p z^p|=0 \tag{21.6}$$

的所有根都落在复平面的单位圆之外(即 $\|z\|>1$)，则此 VAR(p)为平稳过程。在方程(21.6)中，$|\cdot|$表示行列式。特别地，当 $p=1$ 时，VAR(1)的平稳性要求 $|\boldsymbol{I}_n-\boldsymbol{\Gamma}_1 z|=0$ 的所有根都满足 $\|z\|>1$，即 $\|1/z\|<1$。由于 $|\boldsymbol{I}_n-\boldsymbol{\Gamma}_1 z|=|z||(1/z)\boldsymbol{I}_n-\boldsymbol{\Gamma}_1|$，故 $|(1/z)\boldsymbol{I}_n-\boldsymbol{\Gamma}_1|=0$，因此 $1/z$ 为矩阵 $\boldsymbol{\Gamma}_1$ 的特征值(根据特征值的定义)。这意味着，VAR(1)的平稳性要求 $\boldsymbol{\Gamma}_1$ 的所有特征值都落在单位圆之内(即 $\|1/z\|<1$)。由于矩阵的特征值求解方便，故此判断条件容易应用。对于一般的 VAR(p)，可以先将其写为 VAR(1)的形式，然后再通过特征值来判断其平稳性。

首先，定义如下三个 $np\times1$ 的列向量：

$$\tilde{\boldsymbol{y}}_t=\begin{pmatrix}\boldsymbol{y}_t\\ \boldsymbol{y}_{t-1}\\ \vdots\\ \boldsymbol{y}_{t-p+1}\end{pmatrix}_{np\times1},\quad \tilde{\boldsymbol{\Gamma}}_0=\begin{pmatrix}\boldsymbol{\Gamma}_0\\ \boldsymbol{0}\\ \vdots\\ \boldsymbol{0}\end{pmatrix}_{np\times1},\quad \tilde{\boldsymbol{\varepsilon}}_t=\begin{pmatrix}\boldsymbol{\varepsilon}_t\\ \boldsymbol{0}\\ \vdots\\ \boldsymbol{0}\end{pmatrix}_{np\times1} \tag{21.7}$$

其次，定义 $np\times np$“伴随矩阵”(companion matrix)：

$$\tilde{\boldsymbol{\Gamma}}=\begin{pmatrix}\boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \cdots & \boldsymbol{\Gamma}_p\\ \boldsymbol{I}_n & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \vdots & \vdots & & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_n & \boldsymbol{0}\end{pmatrix}_{np\times np} \tag{21.8}$$

由此，可将 VAR(p)模型(21.5)写为以下 VAR(1)的形式：

$$\tilde{\boldsymbol{y}}_t=\tilde{\boldsymbol{\Gamma}}_0+\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{y}}_{t-1}+\tilde{\boldsymbol{\varepsilon}}_t \tag{21.9}$$

因此，VAR(p)平稳性要求其伴随矩阵 $\tilde{\boldsymbol{\Gamma}}$ 的所有特征值都落在单位圆之内。

21.4 单位根所带来的问题

对于 AR(1)，一般从理论上认为，不太可能出现 $|\beta_1|>1$ 的情形，否则任何对经济的扰动都将被无限放大。因此，经济学家通常只担心存在单位根的情形，即 $\beta_1=1$。如果时间序列存在单位根，则为非平稳序列，可能带来以下问题。

（1）**自回归系数的估计值向左偏向于0**。假设对于AR(1)，$y_t=\beta_0+\beta_1 y_{t-1}+\varepsilon_t$，其真实值为$\beta_1=1$。然而，$\beta_1$的OLS估计量$\hat{\beta}_1$却不是渐近正态分布，甚至不是对称分布（即使是在大样本中），而是向左偏向于0。这是因为，由于$\{y_t\}$不是平稳序列，中心极限定理不再适用。虽然$\operatorname{plim}_{n\to\infty}\hat{\beta}_1=\beta_1$（仍为一致估计），但在有限样本下可能存在较大偏差。使用蒙特卡罗法可以得到$\hat{\beta}_1$的大样本分布，参见图21.2（产生该图的Stata程序参见附录）。

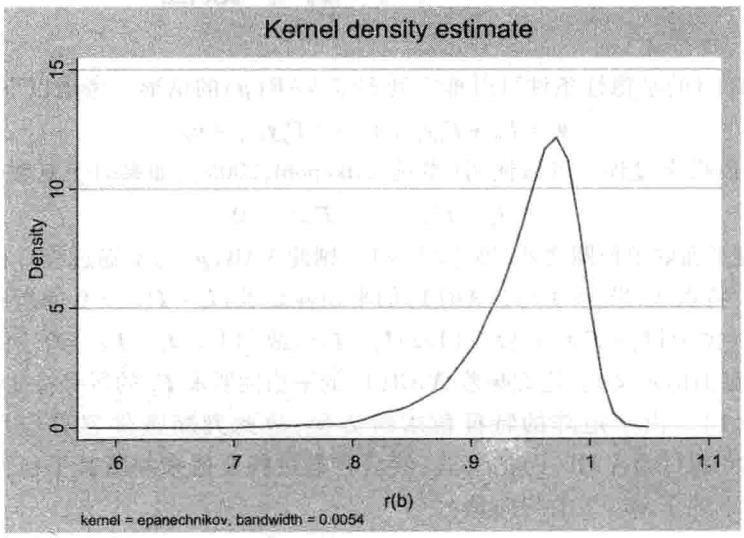


图21.2 在单位根情况下$\hat{\beta}_1$的大样本分布（$n=10\,000$）

（2）**传统的t检验失效**。由于$\hat{\beta}_1$不是渐近正态分布，t统计量也不服从渐近标准正态分布，传统的区间估计与假设检验是无效的。更一般地，建立于平稳性假设基础之上的大样本理论不再适用。

（3）**两个相互独立的单位根变量可能出现伪回归**（spurious regression）**或伪相关**。比如，假设$y_t=y_{t-1}+u_t$，$x_t=x_{t-1}+v_t$，其中，u_t，v_t均为独立同分布且相互独立。因此，y_t与x_t也是相互独立的。考虑OLS回归，$y_t=\alpha+\beta x_t+\varepsilon_t$。由于$y_t$与$x_t$相互独立，故真实参数$\beta=0$。如果样本容量足够大，我们期待OLS估计值$\hat{\beta}\approx 0$，$R^2\approx 0$，然而实际结果并非如此，因为扰动项$\varepsilon_t=y_t-\alpha-\beta x_t$也是非平稳的。这一结论最初由Granger and Newbold（1974）通过蒙特卡罗模拟而发现。

下面，在Stata中模拟此过程。假设$y_0=0$，则$y_1=0+u_1=u_1$，$y_2=y_1+u_2=u_1+u_2$，$y_3=y_2+u_3=u_1+u_2+u_3$；以此类推，$y_t=\sum_{s=1}^{t}u_s=u_1+\cdots+u_t$。根据此性质，很容易在Stata中定义随机游走或单位根变量。

```
. drop _all                (删去内存中已有数据)
. set obs 10000            (确定样本容量为 T = 10 000)
. set seed 1234
. gen u = rnormal()        (产生服从标准正态分布的扰动项 u_t)
. gen y = sum(u)           (定义随机游走 y_t = Σ_{s=1}^{t} u_s)
. set seed 12345
```

```
. gen v = rnormal()   (产生服从标准正态分布的扰动项 v_t)
. gen x = sum(v)      (定义随机游走 x_t = Σ_{s=1}^{t} v_s)
. reg y x
```

Source	SS	df	MS	Number of obs = 10000
				F(1, 9998) = 8835.74
Model	3877265.25	1	3877265.25	Prob > F = 0.0000
Residual	4387283.57	9998	438.81612	R-squared = 0.4691
				Adj R-squared = 0.4691
Total	8264548.82	9999	826.537536	Root MSE = 20.948

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.5621503	.0059804	94.00	0.000	.5504275	.5738731
_cons	-8.53695	.252395	-33.82	0.000	-9.031695	-8.042205

从上表可知,尽管 y 与 x 相互独立(因为 u 与 v 相互独立),但 y 对 x 的回归系数在 1% 水平上显著,而且 R^2 高达 0.47,存在“伪回归”。进一步,把 u 对 v 回归。

```
. reg u v
```

Source	SS	df	MS	Number of obs = 10000
				F(1, 9998) = 0.05
Model	.0535855	1	.0535855	Prob > F = 0.8157
Residual	9865.58665	9998	.986756016	R-squared = 0.0000
				Adj R-squared = -0.0001
Total	9865.64024	9999	.98666269	Root MSE = .99336

u	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
v	.0023061	.0098959	0.23	0.816	-.0170918	.021704
_cons	-.0039404	.0099338	-0.40	0.692	-.0234125	.0155318

由上表可知,u 与 v 的样本值高度不相关(p 值高达 0.816,R^2 = 0.000 0);但由 u 与 v 所产生的随机游走过程 y 与 x 却显著相关。更直观地,看一下 y 与 x 的时间趋势图。

```
. gen t = _n   (定义时间变量 t)
. line y x t, lpattern(dash)
```

从图 21.3 可知,虽然 y 与 x 为相互独立的单位根变量,但二者的样本值却比较相关,存在“伪相关”。如何避免伪相关或伪回归?方法之一,先对变量作一阶差分,然后再回归。方法之二为“协整”(Cointegration),参见下文。但首先必须对是否存在单位根进行检验。

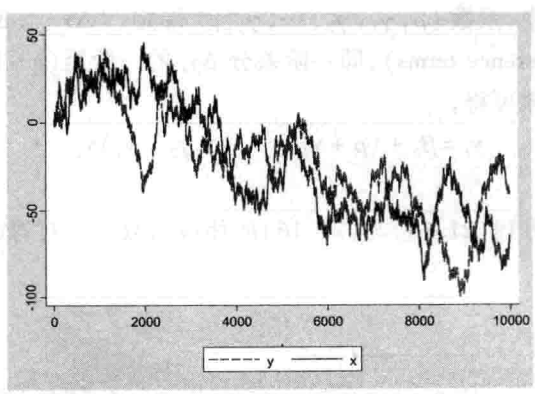


图 21.3 伪相关示意图

21.5 单位根检验与平稳性检验

1. Dickey-Fuller 单位根检验(DF 检验)

定义 如果$\{\varepsilon_t\}$独立同分布,期望为0,方差有限(finite variance),则称$\{\varepsilon_t\}$为"**独立白噪声**"(independent white noise)。换言之,独立白噪声就是期望为0,方差存在的iid序列。

考虑以下AR(1)模型:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma t + \varepsilon_t \tag{21.10}$$

其中,γt为时间趋势(如果不含时间趋势,可令$\gamma = 0$);ε_t为独立白噪声。考虑以下单边检验:

$$H_0: \beta_1 = 1 \quad vs \quad H_1: \beta_1 < 1 \tag{21.11}$$

其中,替代假设为"$H_1: \beta_1 < 1$",因为我们从理论上认为$\beta_1 > 1$不可能[①]。如果H_0成立,则y_t为带漂移项β_0的随机游走;如果不带漂移项,可令$\beta_0 = 0$。在方程(21.10)两边同时减去y_{t-1}可得

$$\Delta y_t = \beta_0 + \delta y_{t-1} + \gamma t + \varepsilon_t \tag{21.12}$$

其中,$\delta \equiv \beta_1 - 1$。对应的原假设与替代假设变为

$$H_0: \delta = 0 \quad vs \quad H_1: \delta < 0 \tag{21.13}$$

对方程(21.13)作OLS回归,可得估计量$\hat{\delta}$及相应的t统计量。此t统计量被称为"Dickey-Fuller统计量"(简记DF)[②],在Stata中记为$Z(t)$。但$Z(t)$并不服从渐近正态分布,其临界值须通过蒙特卡罗模拟来获得[③]。显然,$Z(t)$越小(绝对值很大的负数),则越倾向于拒绝原假设。因此,DF检验是左边单侧检验,即其拒绝域只在分布的最左边。

2. Augmented Dickey-Fuller 单位根检验(ADF 检验)

DF检验使用一阶自回归来检验单位根,要求扰动项$\{\varepsilon_t\}$为独立白噪声,故扰动项无自相关。如果$\{\varepsilon_t\}$存在自相关,可以引入更高阶的滞后项来控制。假设选择了适当的滞后期p,使得以下AR(p)模型的扰动项$\{\varepsilon_t\}$为独立白噪声:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \gamma t + \varepsilon_t \tag{21.14}$$

为了方便检验,将上式转换为以下形式:

$$y_t = \beta_0 + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \cdots + \gamma_{p-1} \Delta y_{t-p+1} + \gamma t + \varepsilon_t \tag{21.15}$$

其中,系数$(\rho, \gamma_1, \gamma_2, \cdots, \gamma_{p-1})$待定,$\{\Delta y_{t-1}, \Delta y_{t-2}, \cdots, \Delta y_{t-p+1}\}$称为"滞后差分项"(lagged difference terms),即一阶差分Δy_t的一阶至$(p-1)$阶滞后项。将上式中的差分算子去掉,合并同类项可得,

$$y_t = \beta_0 + (\rho + \gamma_1) y_{t-1} + (\gamma_2 - \gamma_1) y_{t-2} + \cdots + (\gamma_{p-1} - \gamma_{p-2}) y_{t-p+1} - \gamma_{p-1} y_{t-p} + \gamma t + \varepsilon_t \tag{21.16}$$

将方程(21.14)与(21.16)的相应系数一一对等起来可得

---

① 如果$\beta_1 > 1$,将导致y_t呈指数增长。对于有指数增长趋势的经济变量,如GDP,通常先取对数去掉其指数趋势。

② 参见Dickey and Fuller(1979)。

③ Phillips(1987)证明,DF统计量的渐近分布为布朗运动(Brownian motion)的函数,但仍然没有解析解,必须进行数值计算。

$$\begin{cases}\beta_1=\rho+\gamma_1\\ \beta_2=\gamma_2-\gamma_1\\ \vdots\\ \beta_{p-1}=\gamma_{p-1}-\gamma_{p-2}\\ \beta_p=-\gamma_{p-1}\end{cases}\tag{21.17}$$

在方程组(21.17)中,把$(\beta_1,\beta_2,\cdots,\beta_p)$作为已知数,可以解出$(\rho,\gamma_1\gamma_2,\cdots,\gamma_{p-1})$的表达式。由方程组(21.17)中的最后一个方程倒推上去可得

$$\begin{cases}\rho=\beta_1+\cdots+\beta_p\\ \gamma_1=-(\beta_2+\cdots+\beta_p)\\ \vdots\\ \gamma_{p-2}=-(\beta_{p-1}+\beta_p)\\ \gamma_{p-1}=-\beta_p\end{cases}\tag{21.18}$$

命题 如果$\rho=1$,则 AR(p)有一个单位根。

证明:由于$\rho=1$,故$\phi(1)=1-\beta_1-\cdots-\beta_p=1-\rho=0$。因此,1 是特征方程$\phi(z)=1-\beta_1 z-\cdots-\beta_p z^p=0$的一个根,正好落在单位圆上,故 AR($p$)有一个单位根。

命题 如果$\rho>1$,则特征方程至少有一个根在单位圆之内,故 AR(p)为非平稳。

证明:首先,$\phi(0)=1-\beta_1\cdot 0-\cdots-\beta_p\cdot 0=1$。

其次,由于$\rho>1$,故$\phi(1)=1-\beta_1-\cdots-\beta_p=1-\rho<0$。

然而$\phi(z)$为连续函数,根据中值定理,存在复数z^*,满足$0<\|z^*\|<1$,使得$\phi(z^*)=0$,参见图 21.4 ①。由于$\|z^*\|<1$,z^*落在单位圆之内,故 AR(p)为非平稳。

综上所述,为了检验 AR(p)是否有单位根,可以考虑对方程(21.15)进行回归,并检验

$$H_0:\rho=1 \quad vs \quad H_1:\rho<1 \tag{21.19}$$

在方程(21.15)两边同时减去y_{t-1}可得

$$\Delta y_t=\beta_0+\delta y_{t-1}+\gamma_1\Delta y_{t-1}+\gamma_2\Delta y_{t-2}+\cdots+\gamma_{p-1}\Delta y_{t-p+1}+\gamma t+\varepsilon_t \tag{21.20}$$

其中,$\delta\equiv\rho-1$。则原假设与替代假设变为

$$H_0:\delta=0 \quad vs \quad H_1:\delta<0 \tag{21.21}$$

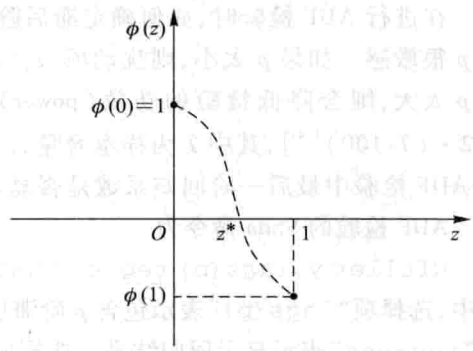


图 21.4 $\rho>1$ 时 AR(p)的非平稳性

对方程(21.20)使用 OLS 可得估计量$\hat{\delta}$及相应的 t 统计量。此 t 统计量被称为“Augmented Dickey-Fuller 统计量”(简记 ADF),Stata 仍记其为 $Z(t)$。ADF 检验是最常用的单位根检验。Nelson and Plosser (1982)使用 ADF 检验考察美国 14 个年度宏观时间序列,结果发现只有其中一个变量可拒绝单位根的原假设。此文一出,引起经济学界对单位根的广泛关注。

与 DF 检验一样,ADF 检验也是左边单侧检验,其拒绝域只在分布的最左边。ADF 统计量的临界值也要通过蒙特卡罗模拟得到。具体而言,ADF 统计量的临界值取决于真实模型(H_0)是否

① 由于 z 为复数(在整个复平面上取值),故实际上应该为三维图,图 21.4 只是示意图。

带漂移项,以及回归方程(21.20)是否包含常数项或时间趋势。Stata 手册(Stata Manual)总结了四种情形,参见表 21.1。

表 21.1 ADF 检验的四种情形

情形	真实模型(H_0)	对回归方程(21.20)的约束	Stata 选择项
1	不带漂移项	$\beta_0=0,\gamma=0$	noconstant
2	不带漂移项	$\gamma=0$	默认值
3	带漂移项	$\gamma=0$	drift
4	不带或带漂移项	无约束	trend

表 21.1 中的情形 2,虽然真实模型不包含漂移项(即无常数项),但在 ADF 检验的回归方程(21.20)中依然包括了常数项。情形 2 与情形 3 对回归方程的约束相同,故检验统计量也相同,但临界值不同(因为真实模型不同)。

关于常数项与时间趋势项

是否应该带常数项或时间趋势项,主要应从理论上考虑。比如,考察 GDP 之对数是否有单位根,一般应包含时间趋势项;而利率、汇率等则不应有时间趋势项。也可以通过画变量的时间序列图来大致判断有无长期增长趋势。另外,在作 ADF 检验时,使用选择项"regress",可以看到常数项或时间趋势项是否显著。如果无从判断,为了稳健起见,可以把各种情况都进行检验,将结果以(c,t,p)格式列表,其中"$c=1$"表示带常数项,"$c=0$"表示不带常数项;"$t=1$"表示带趋势项,"$t=0$"表示不带趋势项;而 p 表示滞后期数。

关于滞后阶数 p 的确定

在进行 ADF 检验时,如何确定滞后阶数 p 是个实际问题。ADF 检验的结果常常对滞后阶数 p 很敏感。如果 p 太小,则扰动项$\{\varepsilon_t\}$可能存在自相关,使得检验出现偏差。另一方面,如果 p 太大,则会降低检验的功效(power)。Schwert (1989)建议,取最大滞后阶数为 $p_{\max}=[12\cdot(T/100)^{1/4}]$,其中 T 为样本容量,[·]表示整数部分,然后使用由大到小的序贯 t 规则,看 ADF 检验中最后一阶回归系数是否显著。也可使用信息准则,比如 AIC 或 BIC。

ADF 检验的 Stata 命令为

```
dfuller y,lags(p) regress noconstant drift trend
```

其中,选择项"lags(p)"表示包含 p 阶滞后差分项,默认为"lags(0)",对应于 DF 检验。选择项"regress"表示显示回归结果。选择项"noconstant drift trend"的含义参见表 21.1。

3. Phillips-Perron 单位根检验(PP 检验)

ADF 检验通过引入高阶滞后项来保证扰动项$\{\varepsilon_t\}$没有自相关。"Phillips-Perron 检验"(Phillips and Perron,1988,简记 PP)仍然使用一阶自回归,但使用异方差自相关稳健的标准误(Newey-West standard errors)对 DF 统计量进行修正:

$$y_t=\beta_0+\rho y_{t-1}+\gamma t+\varepsilon_t \tag{21.22}$$

其中,$\{\varepsilon_t\}$可以存在异方差或自相关。经过修正的 Z(t)统计量,其渐近分布与 DF 检验统计量相同,故临界值也相同,也是左边单侧检验。Phillips and Perron (1988)还提供了另一检验统计量 $Z(\rho)$。使用 PP 检验须指定用于计算 Newey-West 标准误的滞后阶数(Newey-West lags)。Stata 默认的 Newey-West 滞后阶数为$[4(T/100)^{2/9}]$,其中 T 为样本容量,而[·]表示取整数。由于金融变量常存在异方差与自相关,故 PP 检验在金融数据中应用较广。在某种意义上,PP 检验相当于

异方差稳健的 ADF 检验。PP 检验的另一优点是，不必指定差分滞后项的滞后阶数。

PP 检验的 Stata 命令为

```
pperron y,noconstant trend regress lags(#)
```

其中，选择项“lags(#)”用来指定 Newey-West 滞后阶数。其他选择项的含义与 ADF 检验(dfuller)相同。命令 pperron 不提供选择项“drift”(对应于表 21.1 的情形 3)，但表 21.1 的情形 3 只是情形 4(对应于选择项“trend”)的特例，故没有太大妨碍。

4. DF-GLS 单位根检验

ADF 检验与 PP 检验的共同缺点是，检验的功效较低(犯第Ⅱ类错误的概率很大)，尤其当样本容量不大，或真实模型接近于单位根的情形。具体来说，当一个时间序列为很接近于 I(1)的 I(0)序列时，即持续性很强的序列(highly persistent)(比如，一阶自回归系数为 0.98)，则单位根检验将很难根据数据区分其究竟为 I(1)或 I(0)。如果在上述检验中加入常数项或时间趋势项，则将进一步降低检验的功效。

为此，Elliot，Rothenberg and Stock (1996)提出以下的两步检验。第一步，用 GLS 估计原序列 $\{y_t\}$ 的常数项与时间趋势项 $\hat{\delta}_0+\hat{\delta}_1 t$，计算去势后(detrended)的序列 $\{y_t^d \equiv y_t-\hat{\delta}_0-\hat{\delta}_1 t\}$[①]。第二步，对 $\{y_t^d\}$ 使用 ADF 检验。这个检验被称为“DF-GLS 检验”，是目前最有功效的单位根检验。如果真实模型接近于单位根(自回归系数接近于 1)，则应考虑使用 DF-GLS 检验。

DF-GLS 检验的 Stata 命令为

```
dfgls y,notrend maxlag(#) ers
```

其中，选择项“notrend”表示不带时间趋势项，即令 $\hat{\delta}_1=0$；默认带时间趋势项。选择项“maxlag(#)”用来确定 ADF 检验的最大滞后阶数，默认值为 $p_{\max}=[12\cdot(T/100)^{1/4}]$(Schwert，1989)。该命令将汇报上至 $p_{\max}$ 的各阶检验结果，并提供三种方法选择最优滞后阶数，即 Ng-Perron 序贯 t 准则(Ng and Perron，1995)，SIC 信息准则，以及 MAIC 信息准则(Modified AIC)(Ng and Perron，2000)。选择项“ers”表示使用 Elliot，Rothenberg and Stock (1996)提供的临界值，默认临界值来自 Cheung and Lai (1995)。

5. KPSS 平稳性检验

以上单位根检验的原假设皆为“H_0：有单位根”。由于单位根检验的功效不高，对于宏观经济变量，经常无法拒绝原假设，而被迫接受单位根的存在。此时，犯第Ⅱ类错误(即在“H_1：无单位根”为真的情况下却接受“H_0：有单位根”)的概率较高。在某种意义上，单位根检验就相当于先假设“某人有罪”，如果他确实做了一辈子好事，才推翻此“有罪”之原假设。这种“假定有罪”的前提似乎有失公平，至少在法庭判案时，一般假设当事人无罪，除非找到有力的相反证据。

为此，Kwiatkowski，Phillips，Schmidt and Shin (1992)提出的平稳性检验(KPSS)将原假设改为“H_0：时间序列为平稳”，而替代假设变为“H_1：有单位根”。假设时间序列 y_t 可分解为时间趋势、随机游走与平稳过程之和：

$$\begin{aligned} y_t &= \beta t + u_t + \varepsilon_t \\ u_t &= u_{t-1} + v_t, v_t \sim WN(0,\sigma_v^2) \end{aligned} \tag{21.23}$$

---

① 根据 Stock and Watson (2004，p. 549)，在 DF-GLS 检验的第一步，使用广义最小二乘法估计常数项与趋势项。具体方法如下。记被检验变量为 Y_t。首先计算 3 个新变量：V_t, X_{1t}, X_{2t}，即 $V_1=Y_1$，$V_t=Y_t-\alpha^* Y_{t-1}$，$t=2\cdots,T$，$X_{11}=1$，$X_{1t}=1-\alpha^*$，$t=2,\cdots,T$，以及 $X_{21}=1$，$X_{2t}=t-\alpha^*(t-1)$，其中，$\alpha^*=1-13.5/T$。然后将 V_t 对 X_{1t} 与 X_{2t} 回归，即用 OLS 来估计回归方程 $V_t=\delta_0 X_{1t}+\delta_1 X_{2t}+e_t(t=1,\cdots,T)$ 的系数，其中 e_t 为扰动项。

其中，βt 为时间趋势，u_t 是随机游走，而 ε_t 为平稳过程（允许存在异方差与自相关）。假设 u_t 的初始值 u_0 为固定，则可视 u_0 为 y_t 的常数项。假设 u_t 的扰动项 v_t 为白噪声（White Noise，简记WN），记其方差为 σ_v^2。在这个模型中，“y_t 为趋势平稳”（trend stationary）的原假设等价于“H_0：$\sigma_v^2=0$”（此时，$v_t\equiv 0$，故 u_t 为常数），而替代假设为“H_1：$\sigma_v^2>0$”。如果不含时间趋势（$\beta=0$），则原假设为“y_t 为平稳过程”（level stationary）。对此原假设进行拉格朗日乘子检验（LM），即得到KPSS统计量。KPSS检验是单边右侧检验（正如 χ^2 检验一样），其临界值须通过蒙特卡罗模拟得到。Kwiatkowski，Phillips，Schmidt and Shin（1992）使用KPSS检验考察Nelson and Plosser（1982）所使用的美国年度宏观时间序列，结果发现许多变量均无法拒绝“趋势平稳”的原假设。因此，究竟宏观经济变量是否为单位根过程仍然是存疑的。

KPSS检验的Stata命令为

```
ssc install kpss  （下载安装命令 kpss）
kpss y,notrend maxlag(#)
```

其中，选择项“notrend”表示不包括时间趋势，即在原模型中 $\beta=0$；默认包括时间趋势。选择项“maxlag(#)”用来确定最大滞后阶数，默认值为 $p_{\max}=[12\cdot(T/100)^{1/4}]$（Schwert，1989），并汇报上至 $p_{\max}$ 的各阶检验结果。

6. 单整阶数（order of integration）的确定

对时间序列 $\{y_t\}$ 进行单位根检验后，如果认为 $\{y_t\}$ 为非平稳，则要进一步判断其为I(1)或I(2)。可以对一阶差分 $\{\Delta y_t\}$ 进行单位根检验，如果 $\{\Delta y_t\}$ 为平稳，则 $\{y_t\}$ 是I(1)。否则，要继续对二阶差分 $\{\Delta^2 y_t\}$ 进行单位根检验。如果 $\{\Delta^2 y_t\}$ 为平稳，则 $\{y_t\}$ 为I(2)，以此类推。

21.6 单位根检验的Stata实例

下面以数据集 macro_swatson.dta 为例[①]，检验美国的季度通货膨胀率 inf 是否含有单位根。首先，看一下 inf 的时间趋势图（如图21.5）：

```
. use macro_swatson.dta,clear
. line inf quarter
```

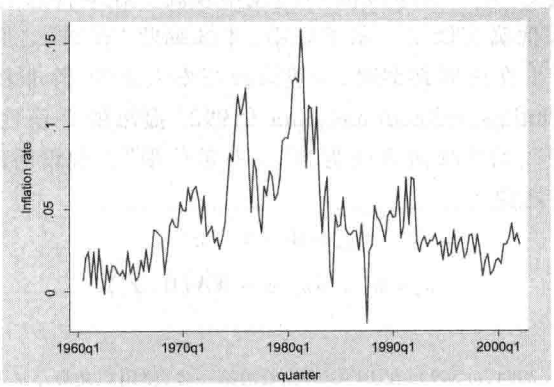


图21.5 美国通货膨胀率的时间趋势

① 此数据集截取自Stock and Watson（2004）。

从上图可以大致看出，通胀率 inf 应该有常数项，但没有明显的时间趋势。为此，首先考虑带常数项，但不带趋势项的 DF 检验：

```
.dfuller inf
```

```
Dickey-Fuller test for unit root                    Number of obs   =       166

                                 ------------ Interpolated Dickey-Fuller ------------
                 Test            1% Critical       5% Critical      10% Critical
                 Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)            -3.718             -3.488            -2.886           -2.576
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0039
```

由于 DF 统计量为 $-3.718 < -3.488$（左边单侧检验），故可以在 1% 的水平上拒绝“存在单位根”的原假设。由于 DF 检验中的扰动项可能存在自相关，故要考虑更高阶的 ADF 检验。首先，计算 Schwert (1989) 建议的最大滞后阶数 $p_{max} = [12 \cdot (T/100)^{1/4}]$：

```
.di 12 * (167/100)^(1/4)
13.641445
```

这表明，$p_{max} = 13$。下面，令 $\hat{p} = 13$，进行 ADF 检验：

```
.dfuller inf,lags(12) reg
```

```
Augmented Dickey-Fuller test for unit root          Number of obs   =       154

                                 ------------ Interpolated Dickey-Fuller ------------
                 Test            1% Critical       5% Critical      10% Critical
                 Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)            -1.782             -3.492            -2.886           -2.576
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3893

------------------------------------------------------------------------------
       D.inf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inf |
         L1. |  -.0874144   .0490495    -1.78   0.077    -.1843878    .009559
         LD. |  -.1754754   .0893062    -1.96   0.051    -.3520386   .0010877
        L2D. |  -.2646741   .0894366    -2.96   0.004     -.441495   -.0878533
        L3D. |   .1934371   .0912662     2.12   0.036      .012999    .3738752
        L4D. |  -.0410425    .091365    -0.45   0.654    -.2216761    .1395911
        L5D. |   .0994029   .0884824     1.12   0.263    -.0755316    .2743374
        L6D. |    .079468   .0887313     0.90   0.372    -.0959585    .2548946
        L7D. |    .083617   .0885392     0.94   0.347    -.0914297    .2586638
        L8D. |  -.196626    .0885273    -2.22   0.028    -.3716493   -.0216028
        L9D. |  -.1043182   .0890171    -1.17   0.243    -.2803097    .0716733
       L10D. |  -.0788603   .0888823    -0.89   0.376    -.2545854    .0968648
       L11D. |  -.0765305   .0845429    -0.91   0.367    -.2436763    .0906153
       L12D. |  -.1506183   .0809574    -1.86   0.065    -.3106754    .0094388

       _cons |   .0041168   .0025237     1.63   0.105    -.0008727    .0091064
------------------------------------------------------------------------------
```

上表显示，最后一阶滞后项（L12D.）在5%的水平上并不显著。依次令 $\hat{p}=12,11,10$，进行ADF检验，最后一阶滞后项仍然不显著（过程略）。下面，令 $\hat{p}=9$，再进行ADF检验。

```
.dfuller inf,lags(8) reg
```

Augmented Dickey-Fuller test for unit root　　Number of obs = 158

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	-2.607	-3.491	-2.886	-2.576

MacKinnon approximate p-value for Z(t) = 0.0915

D.inf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inf						
L1.	-.1168873	.0448301	-2.61	0.010	-.2054771	-.0282974
LD.	-.1219152	.0824072	-1.48	0.141	-.2847619	.0409315
L2D.	-.2372779	.0826736	-2.87	0.005	-.4006511	-.0739047
L3D.	.2436394	.0841557	2.90	0.004	.0773374	.4099414
L4D.	-.0054838	.0863573	-0.06	0.949	-.1761365	.1651688
L5D.	.1182747	.0853734	1.39	0.168	-.0504336	.2869829
L6D.	.0637782	.0852927	0.75	0.456	-.1047706	.2323271
L7D.	.1232936	.0811298	1.52	0.131	-.0370289	.2836161
L8D.	-.1584475	.0785274	-2.02	0.045	-.3136273	-.0032678
_cons	.0053964	.0023139	2.33	0.021	.0008239	.009969

上表显示，最后一阶滞后项在5%的水平上显著地不等于0。ADF统计量 $Z(t)$ 显示，无法在5%的水平上拒绝存在单位根的原假设（$-2.607>-2.886$），即可认为通胀率inf含有单位根。“麦金农的近似 p 值”（MacKinnon approximate p-value）为0.0915，与此结论一致。

解决DF检验中扰动项自相关问题的另一方法为PP检验：

```
.pperron inf
```

Phillips-Perron test for unit root　　Number of obs = 166
　　Newey-West lags = 4

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(rho)	-21.002	-20.020	-13.832	-11.088
Z(t)	-3.427	-3.488	-2.886	-2.576

MacKinnon approximate p-value for Z(t) = 0.0101

由于PP检验也是左边单侧检验，而检验统计量均小于5%的临界值，故可在5%的水平上拒绝“存在单位根”的原假设（此结论与ADF检验相反）。

下面进行目前最有功效的单位根检验，DF-GLS检验：

```
.dfgls inf
```

```
DF-GLS for inf                                          Number of obs =    153
Maxlag = 13 chosen by Schwert criterion

                DF-GLS tau      1% Critical       5% Critical       10% Critical
  [lags]     Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
    13           -1.259          -3.500            -2.797            -2.521
    12           -1.250          -3.500            -2.813            -2.536
    11           -1.537          -3.500            -2.829            -2.551
    10           -1.638          -3.500            -2.845            -2.565
     9           -1.724          -3.500            -2.860            -2.579
     8           -2.004          -3.500            -2.874            -2.592
     7           -2.474          -3.500            -2.888            -2.605
     6           -2.158          -3.500            -2.901            -2.617
     5           -2.025          -3.500            -2.914            -2.628
     4           -2.057          -3.500            -2.926            -2.639
     3           -2.109          -3.500            -2.937            -2.649
     2           -1.671          -3.500            -2.947            -2.659
     1           -2.483          -3.500            -2.957            -2.667

Opt Lag (Ng-Perron seq t) = 12 with RMSE   .0142593
Min SC   = -8.269403 at lag  3 with RMSE   .0149887
Min MAIC = -8.310303 at lag 12 with RMSE   .0142593
```

上表显示，从 1 阶至 13 阶滞后[①]，均无法在 5% 的水平上拒绝“存在单位根”的原假设（甚至无法在 10% 水平上拒绝原假设）。下面将原假设变为平稳序列，进行 KPSS 检验：

```
.kpss inf,notrend
```

```
KPSS test for inf

Maxlag = 4 chosen by Schwert criterion
Autocovariances weighted by Bartlett kernel

Critical values for H0: inf is level stationary

10%: 0.347  5% : 0.463  2.5%: 0.574  1% : 0.739

Lag order    Test statistic
    0            2.14
    1            1.16
    2             .81
    3            .625
    4            .514
```

上表显示，5% 的临界值为 0.463，而从 0 阶至 4 阶滞后，其检验统计量均大于 0.463。由于 KPSS 检验为右边单侧检验，故可在 5% 的水平上拒绝“平稳序列”的原假设，即认为存在单位根。

总之，以上各种检验的证据似乎支持通胀率 inf 为单位根过程（尽管 PP 检验的结果支持平稳过程）[②]。为此，进一步检验是否通胀率之差分 dinf 为平稳过程。以 DF - GLS 检验与 KPSS 检验为例：

```
.dfgls dinf
```

① 根据序贯 t 规则（Ng-Perron seq t），应选择滞后 12 阶。但在此例中，无论使用哪阶滞后，均不改变 DF - GLS 检验的结果。

② 究竟应将 inf 视为平稳过程还是单位根过程，文献中并无共识。

```
DF-GLS for dinf                                              Number of obs =    152
Maxlag = 13 chosen by Schwert criterion

                     DF-GLS tau       1% Critical       5% Critical      10% Critical
  [lags]          Test Statistic        Value             Value            Value
------------------------------------------------------------------------------------
    13               -2.477           -3.501            -2.796           -2.520
    12               -2.692           -3.501            -2.813           -2.535
    11               -3.060           -3.501            -2.829           -2.550
    10               -2.980           -3.501            -2.844           -2.565
     9               -3.177           -3.501            -2.860           -2.579
     8               -3.428           -3.501            -2.874           -2.592
     7               -3.366           -3.501            -2.888           -2.605
     6               -3.050           -3.501            -2.901           -2.617
     5               -3.752           -3.501            -2.914           -2.629
     4               -4.514           -3.501            -2.926           -2.640
     3               -5.217           -3.501            -2.937           -2.650
     2               -6.150           -3.501            -2.948           -2.659
     1              -12.079           -3.501            -2.958           -2.668

Opt Lag (Ng-Perron seq t) =  7 with RMSE  .0153809
Min SC   = -8.179493 at lag  2 with RMSE  .0159336
Min MAIC = -7.206234 at lag  6 with RMSE  .0155328
```

上表显示,信息准则或序贯 t 规则的最优滞后阶数介于2与7之间。在此区间,检验统计量均小于5%的临界值,故可在5%的水平上拒绝"存在单位根"的原假设,即认为 dinf 为平稳过程。下面进行 KPSS 检验:

```
. kpss dinf,notrend
```

```
KPSS test for dinf

Maxlag = 4 chosen by Schwert criterion
Autocovariances weighted by Bartlett kernel

Critical values for H0: dinf is level stationary

10%: 0.347  5% : 0.463  2.5%: 0.574  1% : 0.739

Lag order    Test statistic
    0           .0349
    1           .0461
    2           .0702
    3           .0662
    4           .0672
```

从上表可知,检验统计量均远小于5%的临界值(0.463),故可接受"平稳过程"的原假设。综上所述,可以接受 inf 为一阶单整 I(1)过程。

21.7 面板单位根检验

如果样本容量较小,则对单个变量进行单位根检验的功效可能很弱,即如果接受"存在单位根"的原假设,则犯第Ⅱ类错误的概率很大。此时,如果有面板数据,可以找到更有效的检验方法。比如,可同时检验多个国家的实际汇率是否含有单位根。

文献中已有一系列的面板单位根检验，可通过 Stata 命令 xtunitroot 来实现。为了检验 $\{y_{it}\}$ 是否包含单位根，考虑如下面板自回归模型：

$$y_{it}=\rho_i y_{i,t-1}+\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i+\varepsilon_{it} \tag{21.24}$$

其中，$i=1,\cdots,n$ 表示横截面单位，$t=1,\cdots,T_i$ 表示时间，而 ε_{it} 为平稳的扰动项。根据命令 xtunitroot 的默认设置，$\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i$ 表示个体固定效应（即 $\boldsymbol{z}_{it}=1$），也称为“panel-specific means”；如果加上选择项“trend”，则 $\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i$ 表示个体固定效应与线性时间趋势（linear time trend），即 $\boldsymbol{z}_{it}'=(1,t)$；如果加上选择项“noconstant”，则忽略 $\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i$ 这一项。下文的 IPS 检验、费雪式检验与 Hadri LM 检验允许非平衡面板；其他检验则要求平衡面板，即 $T_i=T,\forall i$。

面板单位根检验的原假设为“$H_0:\rho_i=1,\forall i$”，而替代假设为“$H_1:\rho_i<1$”。方程（21.24）可写为等价形式：

$$\Delta y_{it}=\delta_i y_{i,t-1}+\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i+\varepsilon_{it} \tag{21.25}$$

其中，$\delta_i\equiv\rho_i-1$。相应的原假设与替代假设变为

$$H_0:\delta_i=0,\forall i \quad vs \quad H_1:\delta_i<0 \tag{21.26}$$

有些面板单位根检验（LLC 检验、HT 检验与 Breitung 检验），假设各面板单位的自回归系数均相同，也称为“共同根”（common root），即在方程（21.24）中，$\rho_i=\rho,\forall i$。其他检验则允许各面板单位的自回归系数不同。另外，为了导出检验统计量的大样本分布，这些检验对于横截面维度 n 或时间维度 T 是否固定，或趋于无穷的速度所作的渐近假定也不尽相同。因此，对于具体数据，究竟适用何种面板单位根检验，主要取决于样本容量。比如，基于 $n/T\to0$ 的检验，要求时间维度 T 增长速度快于横截面维度 n，故适用于长面板。又比如，基于 T 固定而 $n\to\infty$ 的检验显然适用于短面板。Stata 手册将这些检验分类总结如表 21.2。

表 21.2 面板单位根检验的特征

检验	Stata 选择项	适用的渐近理论	允许不同的自回归系数	允许非平衡面板
LLC	noconstant	$\sqrt{n}/T\to0$	否	否
LLC		$n/T\to0$	否	否
LLC	trend	$n/T\to0$	否	否
HT	noconstant	$n\to\infty$，T 固定	否	否
HT		$n\to\infty$，T 固定	否	否
HT	trend	$n\to\infty$，T 固定	否	否
Breitung	noconstant	$(T,n)\to_{seq}\infty$	否	否
Breitung		$(T,n)\to_{seq}\infty$	否	否
Breitung	trend	$(T,n)\to_{seq}\infty$	否	否
IPS		$n\to\infty$，T 固定；或 n，T 都固定	是	是
IPS	trend	$n\to\infty$，T 固定；或 n，T 都固定	是	是
IPS	lags()	$(T,n)\to_{seq}\infty$	是	是
IPS	trend lags()	$(T,n)\to_{seq}\infty$	是	是
费雪式		$T\to\infty$，n 有限或趋无穷	是	是
Hadri LM		$(T,n)\to_{seq}\infty$	—	否
Hadri LM	trend	$(T,n)\to_{seq}\infty$	—	否

在表21.1中，$(T,n)\to_{seq}\infty$表示“序贯极限”(sequential limit)，即首先给定n，让$T\to\infty$，然后再让$n\to\infty$。在实践中，这要求T较大(large)，而且n也不能太小(at least moderate)。另外，由于Hadri LM检验为面板平稳性检验(原假设为平稳过程)，故不存在是否“允许不同的自回归系数”的问题。下面具体介绍这些面板单位根检验。

1. LLC检验

由于方程(21.25)的扰动项可能存在自相关，Levin，Lin and Chu (2002)(简记LLC)在方程(21.25)的基础上引入高阶差分滞后项(类似于ADF检验的形式)：

$$\Delta y_{it}=\delta y_{i,t-1}+\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i+\sum_{j=1}^{p_i}\theta_{ij}\Delta y_{i,t-j}+\varepsilon_{it} \tag{21.27}$$

其中，δ为共同的自回归系数(共同根)；不同个体的滞后阶数p_i可以不同；$\{\varepsilon_{it}\}$为平稳的ARMA过程；不同个体的ε_{it}相互独立(不存在截面相关)，但允许异方差。通过引入足够高阶的差分滞后项，可以保证ε_{it}为白噪声。

由于方程(21.27)为动态模型且包含个体固定效应，故存在动态面板偏差。因此，如果直接进行OLS回归，估计量$\hat{\delta}$及相应的t统计量将存在偏差，且不服从渐近正态分布；故需要对此t统计量进行校正。由于方程(21.27)中的p_i未知，LLC检验具体分为以下三个步骤。第一步，对于每个面板单位i，把Δy_{it}对$\left(\boldsymbol{z}_{it}',\ \sum_{j=1}^{p_i}\Delta y_{i,t-j}\right)$回归(通过信息准则选择滞后阶数$p_i$)，得到残差$\hat{e}_{it}$；把$y_{i,t-1}$对$\left(\boldsymbol{z}_{it}',\sum_{j=1}^{p_i}\Delta y_{i,t-j}\right)$回归，得到残差$\hat{\nu}_{i,t-1}$。第二步，考虑到不同面板单位可能存在异方差，故将第一步中的残差做如下标准化：

$$\tilde{e}_{it}\equiv\hat{e}_{it}/\hat{\sigma}_{\varepsilon i},\quad \tilde{\nu}_{i,t-1}\equiv\hat{\nu}_{i,t-1}/\hat{\sigma}_{\varepsilon i} \tag{21.28}$$

其中，$\hat{\sigma}_{\varepsilon i}$为$\varepsilon_{it}$的标准误(可通过$\hat{e}_{it}$对$\hat{\nu}_{i,t-1}$的回归残差来计算)。第三步，使用全部数据，进行以下混合回归：

$$\tilde{e}_{it}=\delta\tilde{\nu}_{i,t-1}+\tilde{\varepsilon}_{it} \tag{21.29}$$

即可得$\hat{\delta}$及相应的t统计量t_δ[①]，但如果存在个体固定效应，则t_δ发散至负无穷(diverges to negative infinity)。为此，Levin，Lin and Chu (2002)提出“偏差校正t统计量”(bias-adjusted t statistic)，记为t_δ^*，在大样本下服从标准正态分布。与ADF检验类似，LLC检验也是左边单侧检验，即拒绝域仅在分布的最左边。

LLC检验假设不存在截面相关。如果此假设不成立，则LLC检验将存在“显著性水平扭曲”(size distortion)(O'Connell，1998)。为了缓解可能存在的截面相关，Levin，Lin and Chu (2002)建议先将面板数据减去各截面单位的均值(cross-sectional means)，再进行LLC检验。

LLC检验的Stata命令格式为

```
xtunitroot llc y,trend noconstant demean lags(#) lags(aic #) lags(bic #) lags(hqic #)
```

其中，“y”为进行检验的变量；选择项“trend”表示加入个体固定效应与线性时间趋势，选择项“noconstant”表示这两项都不加，默认仅加入个体固定效应；选择项“demean”表示先将面板数据减去各截面单位的均值，再进行检验；选择项“lags(#)”用于指定差分滞后项$\Delta y_{i,t-j}$的滞后阶数p(要求所有个体的滞后阶数都相同)；选择项“lags(aic #)”、“lags(bic #)”与“lags

① 此种偏回归的理论基础是“Frisch-Waugh-Lovell Theorem”(参见第3章)。

(hqic #)"分别表示使用 AIC、BIC 或 HQIC 信息准则来选择 p_i 并指定其最大值#,且不同个体的滞后阶数 p_i 可以不同。

下面以 Stata 提供的数据集 pennxrate. dta 为例。该平衡面板来自 Penn World Table 6.2,包含 151 个国家,1970—2003 年的实际汇率数据。目标是检验"购买力平价"(Purchasing Power Parity,简记 PPP)是否成立。购买力平价假说认为,两国之间的名义汇率反映两国之间的物价水平,经物价调整后的实际汇率在长期内趋于均衡值,故应为平稳过程。因此,检验 lnrxrate (实际汇率的对数)是否为单位根过程;如果是,则拒绝 PPP 假说。该数据集还包括两个虚拟变量 g7 与 oecd,分别表示 G7 与 OECD 国家。另外,由于选择美国作为参照国来考察世界各国的汇率,故美国不在此数据集中。

在理论上,没有理由认为 lnrxrate 有时间趋势,故不使用选择项"trend",而使用默认设置,即仅加入个体固定效应。在这种情况下,LLC 检验的前提是 $n/T\to 0$,即要求横截面维度小于时间维度。因此,为了演示的目的,仅使用 G7 中的六个国家(不含美国)进行检验。

```
. use pennxrate.dta,clear
. xtunitroot llc lnrxrate if g7,lags(aic 10)
```

其中,选择项"lags(aic 10)"表示将差分滞后项的最大滞后阶数设为 10,并根据 AIC 信息准则选择最优滞后阶数。

```
Levin-Lin-Chu unit-root test for lnrxrate
-----------------------------------------

Ho: Panels contain unit roots                Number of panels  =       6
Ha: Panels are stationary                    Number of periods =      34

AR parameter: Common                         Asymptotics: N/T -> 0
Panel means:  Included
Time trend:   Not included

ADF regressions: 1.00 lags average (chosen by AIC)
LR variance:     Bartlett kernel, 10.00 lags average (chosen by LLC)
-------------------------------------------------------------------------------
                     Statistic      p-value
-------------------------------------------------------------------------------
 Unadjusted t         -6.7538
 Adjusted t*          -4.0277       0.0000
-------------------------------------------------------------------------------
```

上表显示,根据 AIC 信息准则选择的平均滞后阶数为 1。偏差校正 t_δ^* 统计量(Adjusted t*)为 -4.03,显著为负(p 值为 0.000 0),故强烈拒绝面板包含单位根的原假设,认为面板为平稳过程。显然,此结论支持 PPP。上表中的"未校正 t 统计量"(Unadjusted t)为传统的 t 统计量,不能用于检验。

由于 G7 国家经济发展水平相近且联系密切,故每个国家的扰动项可能存在截面相关。为此,下面使用选择项"demean"来缓解此截面相关。

```
. xtunitroot llc lnrxrate if g7,lags(aic 10) demean
```

```
Levin-Lin-Chu unit-root test for lnrxrate
---------------------------------------------------------------------------
Ho: Panels contain unit roots                   Number of panels  =       6
Ha: Panels are stationary                       Number of periods =      34

AR parameter: Common                            Asymptotics: N/T -> 0
Panel means:  Included
Time trend:   Not included                      Cross-sectional means removed

ADF regressions: 1.50 lags average (chosen by AIC)
LR variance:     Bartlett kernel, 10.00 lags average (chosen by LLC)
---------------------------------------------------------------------------
                     Statistic      p-value
---------------------------------------------------------------------------
 Unadjusted t         -5.5473
 Adjusted t*          -2.0813        0.0187
---------------------------------------------------------------------------
```

上表显示，使用选择项“demean”将面板数据减去截面均值后，偏差校正 t_δ^* 统计量的 p 值上升为0.018 7，但依然在5%水平上显著为负。

2. HT检验

LLC检验仅适用于长面板，而许多微观面板数据的时间维度 T 较小。为此，Harris and Tzavalis（1999）（简记HT）提出了基于 T 固定而 $n\to\infty$ 的检验统计量。令方程(21.24)中的自回归系数均相等可得：

$$y_{it}=\rho y_{i,t-1}+\boldsymbol{z}_{it}'\boldsymbol{\gamma}_i+\varepsilon_{it} \tag{21.30}$$

其中，ρ 为共同根；ε_{it} 服从 iid 正态分布，故为同方差。在 $H_0:\rho=1$ 成立的情况下，Harris and Tzavalis（1999）推导出OLS估计量 $\hat{\rho}$ 的期望 μ 与方差 σ^2 的表达式（为 T 的函数）①，并证明当 T 固定而 $n\to\infty$ 时，$z\equiv\dfrac{\hat{\rho}-\mu}{\sqrt{\sigma/n}}\xrightarrow{d}N(0,1)$。基于此大样本分布，然后进行左边单侧检验。

HT检验的Stata句式为

```
xtunitroot ht y,trend noconstant demean
```

其中，各选择项的含义如上。

继续以数据集 pennxrate.dta 为例演示。由于HT检验的前提为 T 固定而 $n\to\infty$，故使用全部151个国家进行检验。为了缓解截面相关，依然使用选择项“demean”。

```
.use pennxrate.dta,clear
.xtunitroot ht lnrxrate,demean
```

```
Harris-Tzavalis unit-root test for lnrxrate
---------------------------------------------------------------------------
Ho: Panels contain unit roots                   Number of panels  =     151
Ha: Panels are stationary                       Number of periods =      34

AR parameter: Common                            Asymptotics: N -> Infinity
Panel means:  Included                                       T Fixed
Time trend:   Not included                      Cross-sectional means removed
---------------------------------------------------------------------------
                     Statistic          z          p-value
---------------------------------------------------------------------------
 rho                    0.8184     -13.1239          0.0000
---------------------------------------------------------------------------
```

① 由于存在动态面板偏差，OLS估计有偏差，故在原假设成立的情况下，$\hat{\rho}$ 的期望值也不为1。

从上表可知，$\hat{\rho}=0.82$，而 $z=-13.12$，相应的 p 值为 0.000 0，故强烈拒绝面板单位根的原假设，依然支持 PPP。

3. Breitung 检验

LLC 检验与 HT 检验的共同特点是直接用 OLS 估计回归方程，然后再对自回归系数或 t 统计量进行校正，以消除动态面板偏差。Breitung 检验（Breitung 2000）的基本思路与 LLC 检验类似；主要区别在于，首先对数据进行“向前正交变换”（forward orthogonalization），即减去未来各期的平均值，然后再进行回归，使得回归后不再需要偏差校正。具体步骤参见 Stata 手册。所得检验统计量记为 λ，服从渐近标准正态分布，然后进行左边单侧检验。

Breitung 检验假设数据生成过程为 AR(1)。如果存在更高阶的自回归项，则应先进行“预白噪声化”（prewhitening），以消除原序列的自相关，即分别把 Δy_{it} 与 $y_{i,t-1}$ 对（$\Delta y_{i,t-1},\cdots,\Delta y_{i,t-p}$）进行回归，然后以这两个回归的残差来替代 Δy_{it} 与 $y_{i,t-1}$ 进行 Breitung 检验。

Breitung（2000）的蒙特卡罗模拟结果显示，LLC 检验的偏差校正 t_δ^* 统计量的功效（power）较低，特别在存在个体固定效应且自回归系数接近 1 时。而在这些情况下，Breitung 检验的功效则高很多（much higher）。Breitung（2000）假设不同个体的扰动项不存在截面相关，而 Breitung and Das（2005）则提出在截面相关情况下也成立的检验统计量。

Breitung 检验的 Stata 命令句式为

```
xtunitroot breitung y,trend noconstant demean robust lags(#)
```

其中，选择项“robust”表示使用截面相关稳健的统计量（Breitung and Das，2005）；选择项“lags(#)”用于指定进行预白噪声化的滞后阶数，默认不进行预白噪声化；其他选择项的含义如上。

继续以数据集 pennxrate. dta 为例。由于 Breitung 检验的渐近理论假设 $(T,n)\to_{seq}\infty$，故选择 OECD 国家作为样本数据。我们将使用选择项“robust”来控制截面相关，故不再使用选择项“demean”。由于上文 LLC 检验选择的平均滞后期为 1，故假设数据生成过程为 AR(1)，不进行预白噪声化处理。

```
. xtunitroot breitung lnrxrate if oecd,robust
```

```
Breitung unit-root test for lnrxrate
-------------------------------------------------------------------------------
Ho: Panels contain unit roots                  Number of panels =      27
Ha: Panels are stationary                      Number of periods =     34

AR parameter: Common                           Asymptotics:  T,N -> Infinity
Panel means:  Included                                       sequentially
Time trend:   Not included                     Prewhitening: Not performed
-------------------------------------------------------------------------------
                     Statistic      p-value
-------------------------------------------------------------------------------
 lambda*              -1.6794        0.0465
-------------------------------------------------------------------------------
 * Lambda robust to cross-sectional correlation
```

上表显示，检验统计量 $\lambda=-1.68$，相应的 p 值为 0.046 5，故可在 5% 水平上拒绝面板单位根的原假设。

4. IPS 检验

LLC 检验、HT 检验与 Breitung 检验的共同局限在于，它要求每位个体的自回归系数 δ

都相等，此共同根假设在实践中可能过强。比如，不同国家由于制度与文化的原因，经济规律可能不同。为了克服此缺点，Im，Pesaran and Shin（2003）（简记 IPS）提出了如下面板单位根检验。假设面板数据中共有 n 个相互独立的个体，对每位个体分别进行如下 DF 式回归：

$$\Delta y_{it} = \delta_i y_{i,t-1} + \boldsymbol{z}_{it}'\boldsymbol{\gamma}_i + \varepsilon_{it} \tag{21.31}$$

其中，δ_i 为个体 i 的自回归系数；ε_{it} 服从相互独立的正态分布（扰动项无自相关），但允许异方差。假设 T 固定，而 $n\to\infty$ 或固定。面板单位根的原假设为“$H_0: \delta_i = 0, \forall i$”，而替代假设为“服从平稳过程的个体比例大于零”，即当 $n\to\infty$ 时，n_1/n 收敛至某非零正数，其中 n_1 为服从平稳过程的个体数。

记个体 i 的 t 统计量（即 ADF 统计量）为 t_i，计算所有个体 t 统计量的样本均值 $\bar{t} \equiv \frac{1}{n}\sum_{i=1}^{n} t_i$（Stata 称为 t-bar）。Im，Pesaran and Shin（2003）给出了 $\bar{t}$ 分布的临界值。Stata 还汇报另一统计量 $\widetilde{\bar{t}}$（Stata 称为 t-tilde-bar）[①]，它与 $\bar{t}$ 的区别仅在对方程（21.31）扰动项的方差估计量不同。进一步，可将 $\widetilde{\bar{t}}$ 标准化，构造如下统计量 $Z_{\widetilde{\bar{t}}}$（Stata 称为 Z-t-tilde-bar）：

$$Z_{\widetilde{\bar{t}}} \equiv \frac{\widetilde{\bar{t}} - \mathrm{E}(\widetilde{\bar{t}})}{\sqrt{\mathrm{Var}(\widetilde{\bar{t}})/n}} \xrightarrow{d} N(0,1) \tag{21.32}$$

其中，$\mathrm{E}(\widetilde{\bar{t}})$ 与 $\mathrm{Var}(\widetilde{\bar{t}})$ 为 $\bar{t}$ 的理论均值与方差，可通过查表获得。由于假设这 n 个时间序列相互独立，故适用中心极限定理，因此 $Z_{\widetilde{\bar{t}}}$ 的渐近分布为标准正态。IPS 检验也是左边单侧检验。

如果方程（21.31）的扰动项 ε_{it} 存在自相关，可通过引入差分滞后项来消除，即对每位个体分别进行如下 ADF 式回归：

$$\Delta y_{it} = \delta_i y_{i,t-1} + \boldsymbol{z}_{it}'\boldsymbol{\gamma}_i + \sum_{j=1}^{p_i} \theta_{ij}\Delta y_{i,t-j} + \varepsilon_{it} \tag{21.33}$$

其中，不同个体的滞后阶数 p_i 可以不同（可通过信息准则来确定），且假设 $(T,n)\to_{\mathrm{seq}}\infty$。其余检验步骤与扰动项无自相关的情形类似，记其统计量为 $W_{\bar{t}}$（Stata 称为 W-t-bar），对应于上文的 $Z_{\widetilde{\bar{t}}}$ 统计量。

IPS 检验的 Stata 命令句式为

```
xtunitroot ips y,trend demean lags(#) lags(aic #) lags(bic #) lags(hqic #)
```

其中，各选择项的含义类似于 LLC 检验。下面继续以数据集 pennxrate. dta 为例，检验 OECD 国家是否符合 PPP 假说。首先假设扰动项无自相关，但使用选择项“demean”来缓解可能存在的自相关。

```
. xtunitroot ips lnrxrate if oecd,demean
```

① 统计量 $\widetilde{\bar{t}}$ 中的波浪式符号读为“tilde”。

```
Im-Pesaran-Shin unit-root test for lnrxrate
------------------------------------------

Ho: All panels contain unit roots          Number of panels  =     27
Ha: Some panels are stationary             Number of periods =     34

AR parameter: Panel-specific               Asymptotics: T,N -> Infinity
Panel means:  Included                                  sequentially
Time trend:   Not included                 Cross-sectional means removed

ADF regressions: No lags included
-----------------------------------------------------------------------
                                                Fixed-N exact critical values
                     Statistic      p-value        1%       5%       10%
-----------------------------------------------------------------------
 t-bar                -3.1327                    -1.810   -1.730   -1.680
 t-tilde-bar          -2.5771
 Z-t-tilde-bar        -7.3911         0.0000
-----------------------------------------------------------------------
```

上表显示，$\bar{t}$ 统计量为 -3.13，小于1%水平的临界值 -1.81，故拒绝面板单位根的原假设。而统计量 $Z_{\bar{t}}$ 的 p 值为 0.000 0，同样拒绝原假设。下面，考虑扰动项存在自相关的情形，并引入差分滞后项。

```
. xtunitroot ips lnrxrate if oecd, lags(aic 8) demean
```

```
Im-Pesaran-Shin unit-root test for lnrxrate
------------------------------------------

Ho: All panels contain unit roots          Number of panels  =     27
Ha: Some panels are stationary             Number of periods =     34

AR parameter: Panel-specific               Asymptotics: T,N -> Infinity
Panel means:  Included                                  sequentially
Time trend:   Not included                 Cross-sectional means removed

ADF regressions: 1.48 lags average (chosen by AIC)
-----------------------------------------------------------------------
                     Statistic      p-value
-----------------------------------------------------------------------
 W-t-bar              -7.3075         0.0000
-----------------------------------------------------------------------
```

上表显示，根据 AIC 准则选择的平均滞后期为 1.48，而统计量 $W_{\bar{t}}$ 的 p 值为 0.000 0，依然强烈拒绝原假设。

5. 费雪式检验

费雪式检验的基本思路类似于 IPS 检验，即对每位个体分别进行检验，然后再将这些信息综合起来。具体来说，对面板数据中的每位个体分别进行单位根检验（ADF 检验或 PP 检验），得到 n 个检验统计量及相应的 p 值 $\{p_1, \cdots, p_n\}$。Choi (2001) 提出以下四种方法将这些 p 值综合成“费雪式”(Fisher type) 统计量[①]。方法一为“逆卡方变换”(inverse chi-squared transformation)：

$$P \equiv -2\sum_{i=1}^{n} \ln p_i \xrightarrow{d} \chi^2(2n) \quad (T_i \to \infty) \tag{21.34}$$

其中，T_i 为个体 i 的时间维度（因个体而异，允许非平衡面板数据）。由于取了负号，故这是一个单边右侧检验，即统计量 P 越大，则越倾向于拒绝“面板单位根”的原假设。方法二为“逆正态变

① Ronald Fisher 最早提出这类检验，把几个统计检验的证据综合为一个检验统计量。

换”(inverse normal transformation):

$$Z \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi^{-1}(p_i) \xrightarrow{d} N(0,1) \quad (T_i \to \infty) \tag{21.35}$$

其中,$\Phi^{-1}(\cdot)$为标准正态累积分布函数的逆函数,故名。如果使用方法二,则为单边左侧检验。方法三为“逆逻辑变换”(inverse logit transformation):

$$L^{*} \equiv \sqrt{k} \sum_{i=1}^{n} \ln\left(\frac{p_i}{1-p_i}\right) \xrightarrow{d} t(5n+4) \quad (T_i \to \infty) \tag{21.36}$$

其中,$k \equiv \frac{3(5n+4)}{\pi^2 n(5n+2)}$。方法三也是单边左侧检验。如果面板中的个体数 n 很大,可使用“修正逆卡方变换”(modified inverse chi-squared transformation):

$$P_m \equiv -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\ln p_i + 1) \xrightarrow{d} N(0,1) \quad (T_i, n \to \infty) \tag{21.37}$$

费雪式检验的 Stata 命令句式为

```
xtunitroot fisher y,dfuller pperron demean lags(#)
```

其中,选择项“dfuller”表示根据 ADF 检验获得 p 值,选择项“pperron”表示根据 PP 检验获得 p 值。选择项“lags(#)”如果与选择项“dfuller”同时使用,表示 ADF 检验的滞后阶数;如果与选择项“pperron”同时使用,表示用于计算标准误的滞后阶数。进一步,如果使用选择项“dfuller”,则命令 dfuller 的所有选择项也都适用于此命令;类似地,如果使用选择项“pperron”,则命令 pperron 的所有选择项也都适用于此命令。

继续以数据集 pennxrate.dta 为例,使用滞后两期的 ADF 回归检验所有 151 个国家的实际汇率。由于许多国家的实际汇率对数的平均值都不为零,故假设真实模型存在漂移项,并加入命令 dfuller 的选择项“drift”。另外,使用选择项“demean”来缓解可能存在的截面相关。

```
.xtunitroot fisher lnrxrate,dfuller drift lags(2) demean
```

```
Fisher-type unit-root test for lnrxrate
Based on augmented Dickey-Fuller tests
------------------------------------------

Ho: All panels contain unit roots          Number of panels  =     151
Ha: At least one panel is stationary       Number of periods =      34

AR parameter: Panel-specific               Asymptotics: T -> Infinity
Panel means:  Included
Time trend:   Not included                 Cross-sectional means removed
Drift term:   Included                     ADF regressions: 2 lags
------------------------------------------------------------------------
                                  Statistic      p-value
------------------------------------------------------------------------
 Inverse chi-squared(302)  P       975.9130       0.0000
 Inverse normal            Z       -19.6183       0.0000
 Inverse logit t(759)      L*      -20.9768       0.0000
 Modified inv. chi-squared Pm       27.4211       0.0000
------------------------------------------------------------------------
 P statistic requires number of panels to be finite.
 Other statistics are suitable for finite or infinite number of panels.
------------------------------------------------------------------------
```

上表显示,所有四个统计量均强烈拒绝面板单位根的原假设,相应的 p 值为 0.000 0。

6. Hadri LM 检验

Hadri (2000)把 KPSS 平稳性检验推广到面板数据,提出了检验面板平稳性的 LM 检验(原假设为平稳过程)。考虑以下面板形式的 KPSS 检验模型:

$$\begin{aligned} y_{it} &= \beta_i t + u_{it} + \varepsilon_{it} \\ u_{it} &= u_{i,t-1} + \nu_{it} \end{aligned} \tag{21.38}$$

其中,$\beta_i t$ 为个体 i 的时间趋势(panel-specific time trend),而扰动项 ε_{it} 与 ν_{it} 均服从 iid 正态分布,其方差分别为 σ_ε^2 与 σ_ν^2。面板平稳性的原假设等价于"$H_0: \lambda = \dfrac{\sigma_\nu^2}{\sigma_\varepsilon^2} = 0$",而替代假设为"$H_1: \lambda > 0$"。

Hadri LM 检验的 Stata 命令句式为

```
xtunitroot hadri y, trend demean robust kernel (bartlett #) kernel (parzen #) kernel (quadraticspectral #)
```

其中,选择项"robust"表示允许不同个体的 ε_{it} 存在异方差(不再是 iid,但仍为正态);选择项"kernel (bartlett #) kernel (parzen #) kernel (quadraticspectral #)"分别指定使用 bartlett, parzen 或 quadraticspectral 核函数[①]以及滞后阶数#,来估计扰动项的长期方差(long-run variance),即渐近方差。使用选择项"kernel ()"可使得检验结果在存在异方差与自相关的情况下也成立。

仍以数据集 pennxrate. dta 为例,对 OECD 国家进行检验。为了控制自相关,使用滞后 5 阶的 bartlett 核函数。

```
. xtunitroot hadri lnrxrate if oecd, kernel (bartlett 5) demean
```

```
Hadri LM test for lnrxrate
------------------------------------------------------------------------------
Ho: All panels are stationary                 Number of panels   =        27
Ha: Some panels contain unit roots            Number of periods  =        34

Time trend:           Not included            Asymptotics: T, N -> Infinity
Heteroskedasticity:   Robust                                   sequentially
LR variance:          Bartlett kernel, 5 lags Cross-sectional means removed
------------------------------------------------------------------------------
                      Statistic      p-value
------------------------------------------------------------------------------
 z                      9.6473       0.0000
------------------------------------------------------------------------------
```

上表显示,可以拒绝"所有面板单位均为平稳过程"的原假设,与上文的面板单位根检验结果有所不同。Banerjee et al (2005)探讨了面板单位根检验对于 PPP 假说的适用性,并指出由于不同国家之间的汇率可能存在协整或长期关系,面板单位根检验经常在原假设正确的情况下也拒绝原假设。

以上介绍的面板单位根检验大多要求不同个体的扰动项相互独立,在实践中可能很难满足。比如,由于国际经济的互动,跨国面板数据中的变量常存在"共同运动趋势"(comovement),故不同个体的扰动项可能相关。近些年出现的新一代面板单位根检验开始允许不同个体的扰动项之间存在相关性(参见 Silva, 2009;陈海燕, 2012),但 Stata 12 尚未包含这些前沿结果。

---

① 有关核函数的介绍,参见第 27 章。

21.8 协整的思想与初步检验

对于有单位根的变量，传统的处理方法是对其进行一阶差分而得到平稳序列。但是，一阶差分后变量的经济含义与原序列并不相同，而有时我们仍然希望使用原序列进行回归。如果多个单位根变量之间由于某种经济力量而存在“长期均衡关系”(long-run equilibrium)，则有可能进行这种回归。其基本思想是，如果多个单位根序列拥有“共同的随机趋势”(common stochastic trend)，则可以对这些变量作线性组合而消去此随机趋势。

例 短期利率与长期利率可能都是单位根过程，而二者的走势也很相似。从经济理论上来看[①]，长期利率是未来预期短期利率的平均值与“风险溢价”(risk premium)之和，故存在长期均衡关系。

例(非正式) 当你遛狗时，假设你与狗的每一步位置为随机游走过程(带漂移项)，故均为单位根过程。由于你与狗之间有一根皮带相连(“长期均衡关系”)，故你与狗的位置之间不会相离太远(尽管二者都是单位根过程)。

假设两个I(1)过程$\{y_t\}$，$\{x_t\}$可以分别表示为

$$\begin{cases} y_t = \alpha + \beta w_t + \varepsilon_t \\ x_t = \gamma + \delta w_t + u_t \end{cases} \tag{21.39}$$

其中，w_t为随机游走，$w_t = w_{t-1} + \nu_t$；而$\varepsilon_t, u_t, \nu_t$均为白噪声。由于$\{y_t\}$与$\{x_t\}$拥有共同的随机趋势$w_t$，故二者的如下线性组合为平稳过程：

$$\delta y_t - \beta x_t = (\alpha\delta - \beta\gamma) + (\delta\varepsilon_t - \beta u_t) \tag{21.40}$$

在这种情况下，称$\{y_t\}$与$\{x_t\}$是“协整的”(cointegrated)，而称向量$(\delta, -\beta)$为“协整向量”(cointegrating vector)或“协整系数”。显然，可以把$(\delta, -\beta)$标准化为$(1, -\beta/\delta)$。从这个例子可以看出，对于两个I(1)变量，只可能存在一个协整关系；而对于n个I(1)变量，则最多可能存在$(n-1)$个协整关系。一组I(1)变量之间协整关系的个数被称为“协整秩”(cointegration rank)，即线性无关的协整向量的个数。

如何判断一组I(1)变量间是否存在协整关系呢？首先，这些变量必须在理论上可能存在长期均衡关系；否则，进行协整分析将没有意义。其次，如果只有两个变量，则可以直接画图，看二者的时间趋势是否有相似性。但这个方法不严格，也不适用于两个以上的变量。作为正式的协整检验，Engle and Granger (1987)提出了如下的“EG - ADF检验”。

原假设为$\{y_t, x_t\}$存在协整关系，且协整系数为$\{1, -\theta\}$，则$\{z_t \equiv y_t - \theta x_t\}$为平稳过程。如果$\theta$已知(比如，通过经济理论而知)，则可以用ADF检验来确定$\{z_t\}$是否平稳。如果接受“$\{z_t\}$为平稳”，则认为$\{y_t, x_t\}$存在协整关系。然而，通常我们并不知道θ，故“EG - ADF方法”分两步进行。

第一步 用OLS估计协整系数θ，即$y_t = \phi + \theta x_t + z_t$。在“$\{y_t, x_t\}$存在协整关系”的原假设下，虽然$\{y_t, x_t\}$为非平稳的I(1)过程，但$\{z_t\}$为平稳过程。在这种情况下，OLS的估计量$\hat{\phi}$与$\hat{\theta}$都是一致估计量。

---

① 根据“利率期限结构的预期理论”(The expectation theory of the term structure of interest rates)。

第二步 对残差序列$\{\hat{z}_t \equiv y_t - \hat{\phi} - \hat{\theta}x_t\}$进行 ADF 检验，确定其是否平稳。由于协整系数$\hat{\theta}$是估计出来的，不一定是真实的协整系数，故 EG - ADF 统计量的临界值与普通的 ADF 检验不同，参见 Hayashi (2000, p. 646)或 Stock and Watson (2004, p. 557)。

如果检验结果确认$\{\hat{z}_t\}$为平稳，则接受"$\{y_t, x_t\}$存在协整关系"的原假设。此时，估计出的协整关系"$y_t = \hat{\phi} + \hat{\theta}x_t$"即为$\{y_t, x_t\}$之间的长期均衡关系($\hat{\phi}, \hat{\theta}$为长期参数)。如果要估计$\{y_t, x_t\}$之间的短期关系(短期参数)，则需要使用误差修正模型(ECM)。

假设$\{y_t, x_t\}$之间的关系可由一个 ADL 模型来表示，比如，$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + \varepsilon_t$，则其对应的 ECM 模型为(参见第 20 章)

$$\Delta y_t = \gamma_0 \Delta x_t + (\beta_1 - 1)(y_{t-1} - \phi - \theta x_{t-1}) + \varepsilon_t \tag{21.41}$$

显然，方程(21.41)左边的Δy_t为平稳过程。如果$\{y_t, x_t\}$存在协整关系，则方程右边的误差修正项$(y_{t-1} - \phi - \theta x_{t-1})$为平稳，而$\Delta x_t$也为平稳，故方程右边整体上也是平稳的。反之，如果$\{y_t, x_t\}$不存在协整关系，则方程右边的误差修正项$(y_{t-1} - \phi - \theta x_{t-1})$为非平稳，而方程左边依然平稳，故 ECM 模型不能成立[①]。在$\{y_t, x_t\}$存在协整关系的前提下，将残差$\{\hat{z}_t \equiv y_t - \hat{\alpha} - \hat{\theta}x_t\}$代入 ECM 模型可得

$$\Delta y_t = \gamma_0 \Delta x_t + (\beta_1 - 1)\hat{z}_{t-1} + error_t \tag{21.42}$$

使用 OLS 估计上式，即可得到对短期参数的估计。

EG - ADF 方法的一个缺点是，它不能处理同时存在多个协整关系(即协整秩大于 1)的情形。另外，由于 EG - ADF 方法分两步进行，第一步估计的误差会被带到第二步中，故不是最有效率的方法。比 EG - ADF 方法更有效率的方法是，用 MLE 同时估计长期与短期参数(也是目前最流行的方法)[②]。为此，必须更深入地研究 I(1)过程，并给出协整的严格定义。由于最简单的 I(1)过程就是随机游走，故下面首先把一般的 I(1)过程分解为随机游走、时间趋势与平稳过程之和。

21.9 Beveridge-Nelson 分解公式

定义 称时间序列$\{y_t\}$为"**线性 I(0)过程**"，如果$y_t = \delta + u_t$，其中δ为常数，$u_t = \psi(L)\varepsilon_t$，$\{\varepsilon_t\}$为独立白噪声，滤波$\psi(L) \equiv \psi_0 + \psi_1 L + \psi_2 L^2 + \cdots$，满足$\sum_{j=0}^{\infty} j|\psi_j| < \infty$(称为"一可加总"，one-summable，简记为 OS)，$\psi(1) = \psi_0 + \psi_1 + \psi_2 + \cdots \neq 0$。

显然，OS 比 AS(绝对值可加总)的假定更强，前者是后者的充分条件。"$\psi(1) \neq 0$"是一个技术性条件，防止出现退化的情形(详见下文)。本章下面讨论的 I(0)皆为线性 I(0)过程。假设序列$\{y_t\}$为 I(1)，则其差分为 I(0)，故可表示为$\Delta y_t = \delta + u_t$，其中$u_t$为线性 I(0)过程。假设时间从$t = 0$开始，则$y_1 = y_0 + \delta + u_1$，$y_2 = \delta + \underbrace{y_0 + \delta + u_1}_{=y_1} + u_2 = y_0 + 2\delta + u_1 + u_2$，$\cdots$，$y_t = y_0 + \delta t + \sum_{s=1}^{t} u_s$。

① 如果$\{y_t, x_t\}$为平稳过程，则 ECM 模型总能成立。

② 由于 EG - ADF 方法不如 MLE 有效率，故在 Stata 中没有现成的命令，但可以较方便地手工进行(需要查表以获得临界值)。

由于 u_t 为线性 $I(0)$ 过程，故可写为 $u_t=\psi(L)\varepsilon_t$，其中

$$\psi(L)=\psi_0+\psi_1L+\psi_2L^2+\cdots$$

可以将滤波 $\psi(L)$ 分解为

$$\begin{aligned}\psi(L)&=\psi_0+\psi_1L+\psi_2L^2+\cdots\\&=(\psi_0+\psi_1+\psi_2+\cdots)+(\psi_1L-\psi_1)+(\psi_2L^2-\psi_2)+(\psi_3L^3-\psi_3)+\cdots\\&=\psi(1)+[-\psi_1(1-L)-\psi_2(1-L^2)-\psi_3(1-L^3)-\cdots]\\&=\psi(1)+(1-L)[-\psi_1-\psi_2(1+L)-\psi_3(1+L+L^2)-\cdots]\\&=\psi(1)+(1-L)[-(\psi_1+\psi_2+\cdots)-(\psi_2+\psi_3+\cdots)L-(\psi_3+\psi_4+\cdots)L^2-\cdots]\\&=\psi(1)+(1-L)[\alpha_0+\alpha_1L+\alpha_2L^2+\cdots]，\text{其中 }\alpha_j\equiv-(\psi_{j+1}+\psi_{j+2}+\cdots)\\&=\psi(1)+(1-L)\alpha(L)\end{aligned}$$

其中，$\alpha(L)\equiv\alpha_0+\alpha_1L+\alpha_2L^2+\cdots$。由于 $\psi(L)$ 为 OS，故可以证明（参见附录），$\alpha(L)$ 为 AS。

定义 $\eta_t\equiv\alpha(L)\varepsilon_t$。由于 $\alpha(L)$ 为 AS，故 η_t 为平稳过程。因此，

$$u_t=\psi(L)\varepsilon_t=[\psi(1)+(1-L)\alpha(L)]\varepsilon_t=\psi(1)\varepsilon_t+(1-L)\eta_t=\psi(1)\varepsilon_t+\eta_t-\eta_{t-1}\tag{21.43}$$

$$y_t=y_0+\delta t+\sum_{s=1}^{t}u_s=\underbrace{\delta t}_{\text{time trend}}+\underbrace{\psi(1)\sum_{s=1}^{t}\varepsilon_s}_{\text{random walk}}+\underbrace{\eta_t}_{\text{stationary}}+\underbrace{(y_0-\eta_0)}_{\text{initial condition}}\tag{21.44}$$

上式就是著名的“Beveridge-Nelson 分解公式”。它将 I(1) 过程分解为时间趋势 δt、随机游走 $\psi(1)\sum_{s=1}^{t}\varepsilon_s$[①]、平稳序列 η_t 及初始条件 $(y_0-\eta_0)$ 之和。由于协整关系涉及多个变量之间的关系，下面把 Beveridge－Nelson 分解公式向多维推广。

定义 称 $\{\boldsymbol{y}_t\}$ 为 n 维“**线性向量 I(0) 过程**”，如果 $\boldsymbol{y}_t=\boldsymbol{\delta}+\boldsymbol{u}_t$，其中 $\boldsymbol{\delta}$ 为常数向量，$\boldsymbol{u}_t=\boldsymbol{\psi}(\boldsymbol{L})\boldsymbol{\varepsilon}_t$，多维滤波 $\boldsymbol{\psi}(\boldsymbol{L})\equiv\boldsymbol{I}_n+\boldsymbol{\psi}_1\boldsymbol{L}+\boldsymbol{\psi}_2\boldsymbol{L}^2+\cdots$，$\{\boldsymbol{\varepsilon}_t\}$ 为独立同分布的，$\mathrm{E}(\boldsymbol{\varepsilon}_t)=\boldsymbol{0}$，$\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')$ 为正定矩阵，$\{\boldsymbol{\psi}_j\}$ 为 OS（即矩阵的每个元素均为 OS），而且

$$\boldsymbol{\psi}(1)=\boldsymbol{I}_n+\boldsymbol{\psi}_1+\boldsymbol{\psi}_2+\cdots\neq\boldsymbol{0}_{n\times n}$$

因此，可将 n 维向量 I(1) 过程 $\boldsymbol{y}_t$ 表示为

$$\Delta\boldsymbol{y}_t=\boldsymbol{\delta}+\boldsymbol{u}_t=\boldsymbol{\delta}+\boldsymbol{\psi}(\boldsymbol{L})\boldsymbol{\varepsilon}_t\tag{21.45}$$

这称为 I(1) 系统的“向量移动平均表示法”（VMA Representation）。将其写成绝对水平的形式：

$$\boldsymbol{y}_t=\boldsymbol{y}_0+\boldsymbol{\delta t}+\sum_{s=1}^{t}\boldsymbol{u}_s\tag{21.46}$$

依据类似的推导，有如下向量形式的 Beveridge－Nelson 分解公式：

$$\boldsymbol{y}_t=\underbrace{\boldsymbol{\delta t}}_{\text{time trend}}+\underbrace{\boldsymbol{\psi}(1)\sum_{s=1}^{t}\boldsymbol{\varepsilon}_s}_{\text{random walk}}+\underbrace{\boldsymbol{\eta}_t}_{\text{stationary}}+\underbrace{(\boldsymbol{y}_0-\boldsymbol{\eta}_0)}_{\text{initial condition}}\tag{21.47}$$

21.10 协整的定义与最大似然估计

一般来说，由于存在随机趋势项 $\boldsymbol{\psi}(1)\sum_{s=1}^{t}\boldsymbol{\varepsilon}_s$，上述 I(1) 系统 $\{\boldsymbol{y}_t\}$ 为非平稳。然而，如果在方程(21.47)两边同时乘以某 $1\times n$ 的非零行向量 $\boldsymbol{a}'$，则可能将此随机趋势项消去：

① 如果 $\psi(1)=0$，则这个随机游走部分退化为 0，故要求 $\psi(1)\neq0$。

$$\boldsymbol{a}'\boldsymbol{y}_t = \boldsymbol{a}'\boldsymbol{\delta}t + \boldsymbol{a}'\boldsymbol{\psi}(1)\sum_{s=1}^{t}\boldsymbol{\varepsilon}_s + \boldsymbol{a}'\boldsymbol{\eta}_t + \boldsymbol{a}'(\boldsymbol{y}_0 - \boldsymbol{\eta}_0) \tag{21.48}$$

如果 $\underbrace{\boldsymbol{a}'}_{1\times n}\underbrace{\boldsymbol{\psi}(1)}_{n\times n} = \underbrace{\boldsymbol{0}'}_{1\times n}$，则上式可简化为

$$\boldsymbol{a}'\boldsymbol{y}_t = \boldsymbol{a}'\boldsymbol{\delta}t + \boldsymbol{a}'\boldsymbol{\eta}_t + \boldsymbol{a}'(\boldsymbol{y}_0 - \boldsymbol{\eta}_0) \tag{21.49}$$

$\{\boldsymbol{a}'\boldsymbol{y}_t\}$ 是一个趋势平稳(trend stationary)的过程，即只要将时间趋势项 $\boldsymbol{a}'\boldsymbol{\delta}t$ 去掉，就是平稳过程。此时，称 $\boldsymbol{y}_t$ 是“协整的”(cointegrated)，而称 $\boldsymbol{a}$ 为“协整向量”(cointegrating vector)。

问题的关键在于，是否存在 n 维非零列向量 $\boldsymbol{a}$，使得 $\boldsymbol{a}'\boldsymbol{\psi}(1) = \boldsymbol{0}'$，即 $\boldsymbol{\psi}(1)'\boldsymbol{a} = \boldsymbol{0}$。

考虑关于 $\boldsymbol{a}$ 的 n 元方程组 $\boldsymbol{\psi}(1)'\boldsymbol{a} = \boldsymbol{0}$，称其线性无关解的个数为 $\{\boldsymbol{y}_t\}$ 的“协整秩”(cointegration rank)，即线性无关的协整向量的个数，记为 h。根据线性代数知识，$h = n - \text{rank}[\boldsymbol{\psi}(1)]$。由于 $1 \leqslant \text{rank}[\boldsymbol{\psi}(1)] \leqslant n$，故 $0 \leqslant h \leqslant n-1$。如果 $h = 0$，则不存在协整关系。如果 $h = 1$，则存在一个协整关系，可以解释为长期均衡关系(long-run equilibrium)。如果 $h > 1$，则存在多个协整关系，通常需要用经济理论来剔除不合理的协整向量，而将最合理的协整向量解释为长期均衡关系。

不失一般性，假设协整向量 $\boldsymbol{a}$ 的第一个分量不为 0，并将其标准化为 1，即 $\boldsymbol{a} = \begin{pmatrix} 1 \\ \underbrace{-\boldsymbol{\gamma}}_{(n-1)\times 1} \end{pmatrix}$。将 $\boldsymbol{y}_t$ 也同样地分块，即 $\boldsymbol{y}_t = \begin{pmatrix} y_{1t} \\ \underbrace{\boldsymbol{y}_{2t}}_{(n-1)\times 1} \end{pmatrix}$。假设 $\boldsymbol{a}'\boldsymbol{\delta} = 0$(不存在时间趋势项)，则可以将方程(21.49)写成

$$y_{1t} = \alpha + \boldsymbol{\gamma}'\boldsymbol{y}_{2t} + z_t \tag{21.50}$$

其中，$\alpha \equiv (1 - \boldsymbol{\gamma}')(\boldsymbol{y}_0 - \boldsymbol{\eta}_0)$ 为初始条件，可以视为截距项；而 $z_t \equiv (1 - \boldsymbol{\gamma}')\boldsymbol{\eta}_t$ 为平稳过程，可以视为扰动项。这个回归被称为“协整回归”(cointegrating regression)。在 $\boldsymbol{y}_t$ 存在协整关系的前提下，OLS 估计是一致的。

以上讨论的是 I(1) 系统的 VMA 表示法。但由于扰动项不可观测，为了进行 MLE 估计，考虑以下的“向量自回归表示法”(VAR Representation)：

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\delta}t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{21.51}$$

如何知道此 VAR(p) 何时为协整系统呢？为此，先导出其对应的 VMA 表示法。定义 $\boldsymbol{\Phi}(L) \equiv \boldsymbol{I}_n - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p$，则

$$\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\delta}t + \boldsymbol{\varepsilon}_t \tag{21.52}$$

在方程(21.52)两边同时左乘 $\boldsymbol{\Phi}(L)^{-1}$ 可得

$$\boldsymbol{y}_t = \boldsymbol{\Phi}(L)^{-1}\boldsymbol{\alpha} + \boldsymbol{\Phi}(L)^{-1}\boldsymbol{\delta}t + \boldsymbol{\Phi}(L)^{-1}\boldsymbol{\varepsilon}_t \tag{21.53}$$

对方程(21.53)两边同时进行差分可得

$$\Delta\boldsymbol{y}_t = \boldsymbol{\Phi}(L)^{-1}\delta(1-L)t + \boldsymbol{\Phi}(L)^{-1}(1-L)\boldsymbol{\varepsilon}_t = \boldsymbol{\delta}^* + \boldsymbol{\psi}(L)\boldsymbol{\varepsilon}_t \tag{21.54}$$

其中，$(1-L)t = t - (t-1) = 1$，$\boldsymbol{\delta}^* \equiv \boldsymbol{\Phi}(1)^{-1}\boldsymbol{\delta}$，$\boldsymbol{\psi}(L) \equiv \boldsymbol{\Phi}(L)^{-1}(1-L)$。显然，如果此 VAR 系统为协整秩为 h 的协整系统，则 $\psi(L)$ 必须为 OS，而且 $\text{rank}[\psi(1)] = n - h$。

由于 $\boldsymbol{\psi}(L) \equiv \boldsymbol{\Phi}(L)^{-1}(1-L)$，故 $\boldsymbol{\Phi}(L)\boldsymbol{\psi}(L) = (1-L)\boldsymbol{I}_n$。令 $L = 1$，则有

$$\boldsymbol{\Phi}(1)_{n\times n}\boldsymbol{\psi}(1)_{n\times n} = \boldsymbol{0}_{n\times n} \tag{21.55}$$

可以证明，如果 $\{\boldsymbol{y}_t\}$ 的协整秩为 h，则 $\text{rank}[\boldsymbol{\Phi}(1)] = h$。根据线性代数知识，可以将 $\boldsymbol{\Phi}(1)$ 分

解为

$$\boldsymbol{\Phi}(1)_{n\times n}=\boldsymbol{B}_{n\times h}(\boldsymbol{A}')_{h\times n} \tag{21.56}$$

其中，$\boldsymbol{B},\boldsymbol{A}$ 为两个 $n\times h$ 的满列秩矩阵（$\boldsymbol{B},\boldsymbol{A}$ 不唯一）。此条件称为“降秩条件”（reduced rank condition）。因此

$$\boldsymbol{B}_{n\times h}(\boldsymbol{A}')_{h\times n}\boldsymbol{\Psi}(1)_{n\times n}=\boldsymbol{0}_{n\times n} \tag{21.57}$$

因为 $\boldsymbol{B}_{n\times h}$ 满列秩，故

$$\boldsymbol{A}'_{h\times n}\boldsymbol{\psi}(1)_{n\times n}=\boldsymbol{0}_{h\times n} \tag{21.58}$$

由此可见，矩阵 $\boldsymbol{A}_{n\times h}$ 中的列向量都是协整向量（满足协整向量的定义）。从 VAR 方程(21.51)出发，可以导出其对应的“向量误差修正表示法”（VECM Representation）。根据与推导 ADF 检验类似的方法可得

$$\Delta\boldsymbol{y}_t=\boldsymbol{\alpha}+\boldsymbol{\delta}t+(\boldsymbol{\rho}-\boldsymbol{I}_n)\boldsymbol{y}_{t-1}+\boldsymbol{\Gamma}_1\Delta\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_{p-1}\Delta\boldsymbol{y}_{t-p+1}+\boldsymbol{\varepsilon}_t \tag{21.59}$$

其中，$\boldsymbol{\rho}\equiv\boldsymbol{\Phi}_1+\boldsymbol{\Phi}_2+\cdots+\boldsymbol{\Phi}_p$，$\boldsymbol{\Gamma}_s\equiv-(\boldsymbol{\Phi}_{s+1}+\cdots+\boldsymbol{\Phi}_p)$，$s=1,2,\cdots,p-1$。

由于 $\boldsymbol{I}_n-\boldsymbol{\rho}=\boldsymbol{I}_n-\boldsymbol{\Phi}_1-\boldsymbol{\Phi}_2-\cdots-\boldsymbol{\Phi}_p=\boldsymbol{\Phi}(1)=\boldsymbol{B}\boldsymbol{A}'$，故

$$\Delta\boldsymbol{y}_t=\boldsymbol{\alpha}+\boldsymbol{\delta}t-\boldsymbol{B}\underbrace{\boldsymbol{A}'\boldsymbol{y}_{t-1}}_{z_{t-1}}+\boldsymbol{\Gamma}_1\Delta\boldsymbol{y}_{t-1}+\boldsymbol{\Gamma}_2\Delta\boldsymbol{y}_{t-2}+\cdots+\boldsymbol{\Gamma}_{p-1}\Delta\boldsymbol{y}_{t-p+1}+\boldsymbol{\varepsilon}_t \tag{21.60}$$

其中，$\boldsymbol{z}_{t-1}\equiv\boldsymbol{A}'\boldsymbol{y}_{t-1}$ 为误差修正项（因为 $\boldsymbol{A}$ 中的列向量皆为协整向量）。因此，一个协整的 I(1) 系统同时有 VMA，VAR 与 VECM 表示法，此结论称为“格兰杰表示法定理”（Granger Representation Theorem）。

定义 $\boldsymbol{\Gamma}_0\equiv-\boldsymbol{\Phi}(1)=-\boldsymbol{B}\boldsymbol{A}'$。Johansen(1988)使用 MLE 来估计如下 VECM 模型：

$$\Delta\boldsymbol{y}_t=\boldsymbol{\alpha}+\boldsymbol{\Gamma}_0\boldsymbol{y}_{t-1}+\boldsymbol{\Gamma}_1\Delta\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_{p-1}\Delta\boldsymbol{y}_{t-p+1}+\boldsymbol{\varepsilon}_t \tag{21.61}$$

其中，为了简化推导，假定没有时间趋势项。假设样本容量为 $T+p$，即观测数据为 $\{\boldsymbol{y}_{-p+1},\boldsymbol{y}_{-p+2},\cdots,\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T\}$。显然，只有当 $\boldsymbol{y}_t$ 存在协整关系时，这个 VECM 模型才能成立；否则，方程(21.61)左边为平稳过程，而右边为非平稳过程（因为 $\boldsymbol{\Gamma}_0\boldsymbol{y}_{t-1}$ 不平稳）。假设协整秩为 h，则系数矩阵 $\boldsymbol{\Gamma}_0$ 必须满足约束条件“$\text{rank}(\boldsymbol{\Gamma}_0)=h$”。Johansen 的方法是，在满足“$\text{rank}(\boldsymbol{\Gamma}_0)=h$”以及给定 $\{\boldsymbol{y}_{-p+1},\boldsymbol{y}_{-p+2},\cdots,\boldsymbol{y}_0\}$ 的条件下，最大化 $\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T\}$ 的对数似然函数（即条件 MLE）。

假设扰动项为 n 维正态分布，即 $\boldsymbol{\varepsilon}_t\sim N(\boldsymbol{0},\boldsymbol{\Omega})$，而且 $\{\boldsymbol{\varepsilon}_t\}$ 为独立同分布的。n 维随机向量 $\boldsymbol{\varepsilon}_t$ 的联合密度为（参见第 2 章）

$$\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Omega}|^{1/2}}\exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}_t\right\} \tag{21.62}$$

将(21.62)式取对数可得

$$-\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln|\boldsymbol{\Omega}|-\frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}_t \tag{21.63}$$

记 $\boldsymbol{\Pi}\equiv(\boldsymbol{\alpha}\ \boldsymbol{\Gamma}_0\ \boldsymbol{\Gamma}_1\cdots\boldsymbol{\Gamma}_{p-1})'$，$\boldsymbol{x}_t\equiv(1\ \boldsymbol{y}_{t-1}\ \Delta\boldsymbol{y}_{t-1}\cdots\Delta\boldsymbol{y}_{t-p+1})'$，则 $\boldsymbol{\varepsilon}_t=\Delta\boldsymbol{y}_t-\boldsymbol{\Pi}'\boldsymbol{x}_t$，故 $\{\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,\cdots,\boldsymbol{\varepsilon}_T\}$ 所对应的对数条件似然函数为

$$\begin{aligned}\max\ &-\frac{nT}{2}\ln 2\pi-\frac{T}{2}\ln|\boldsymbol{\Omega}|-\frac{1}{2}\sum_{t=1}^{T}[(\Delta\boldsymbol{y}_t-\boldsymbol{\Pi}'\boldsymbol{x}_t)'\boldsymbol{\Omega}^{-1}(\Delta\boldsymbol{y}_t-\boldsymbol{\Pi}'\boldsymbol{x}_t)]\\ &s.t.\ \text{rank}(\boldsymbol{\Gamma}_0)=h\end{aligned} \tag{21.64}$$

其中，n 为该系统中变量的个数，T 为样本容量。为了求解以上约束极值问题，必须确定协整秩 h。当协整秩为 h 时，系数矩阵 $\boldsymbol{\Gamma}_0$ 有 h 个自由（线性无关）的列向量。因此，协整秩 h 越大，则对矩阵 $\boldsymbol{\Gamma}_0$ 的约束越少，其对应的似然函数最大值应该越大。因此，可以进行以下似然比检验：

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = 0 \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) > 0 \tag{21.65}$$

由于矩阵 $\boldsymbol{\Gamma}_0$ 的秩取决于其非零特征值的个数，故检验统计量被称为“迹统计量”，记为 λ_{trace}。由于“迹检验”(trace test)是似然比检验，故为单边右侧检验，即 λ_{trace} 越大(加上 H_0 约束后，似然函数的最大值下降越多)，则越倾向于拒绝原假设。如果接受“H_0: rank($\boldsymbol{\Gamma}_0$) = 0”，则认为不存在协整关系。反之，则继续检验是否存在多个协整关系：

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = 1 \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) > 1 \tag{21.66}$$

依此顺序不断进行检验，直到接受 H_0，确认协整秩 h 为止。确认协整秩 h 后，就可以用条件 MLE 来估计 VECM 模型中的所有参数，包括长期参数(协整系数)与短期参数。

Johansen 还考虑了另一类检验：

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = p \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) = p + 1 \tag{21.67}$$

其检验统计量为“最大特征值统计量”(maximum eigenvalue statistics)，记为 $\lambda_{\max}$，称为“最大特征值检验”。一般认为，迹检验的效果比特征值检验更好，故前者为 Stata 检验协整秩的默认方法。

21.11 协整分析的 Stata 实例

检验协整秩的 Stata 命令为

```
vecrank y1 y2... yn,lags(#)   (默认包括常数项,但不包括时间趋势)
vecrank y1 y2... yn,lags(#) trend(none)   (不包括常数项或时间趋势)
vecrank y1 y2... yn,lags(#) trend(trend)   (包括常数项与时间趋势)
vecrank y1 y2... yn,max   (显示最大特征值统计量及其临界值)
```

其中，选择项“lags(#)”表示对应的 VAR 模型中滞后的阶数，默认为“lags(2)”。命令 vecrank 的输出结果将列出“$h = 0, 1, \cdots, n-1$”的一系列检验，并以星号(*)标出所接受的 h 值。

在作完协整秩检验，并确定 $h \geqslant 1$ 后，就可以对 VECM 模型进行最大似然估计。

```
vec y1 y2 ... yn,lags(#) rank(#)   (默认设置包括常数项,但不包括时间趋势)
vec y1 y2 ... yn,lags(#) rank(#) trend(none)   (不包括常数项,也不包括时间趋势)
vec y1 y2 ... yn,lags(#) rank(#) trend(trend)   (包括常数项,也包括时间趋势)
```

其中，选择项“lags(#)”表示对应的 VAR 模型中滞后的阶数，默认值为“lags(2)”；“rank(#)”表示协整秩的阶数，默认值为“rank(1)”。以上命令将同时估计短期的 VECM 模型，以及长期的协整回归。

与 VAR 模型类似，估计完 VECM 模型后，应对模型的假设进行检验(diagnostic checking)，然后考察 VECM 模型的脉冲响应函数，并进行预测。有关的 Stata 命令包括

```
veclmar              (估计 VECM 后,对残差是否存在自相关进行 LM 检验)
vecnorm              (估计 VECM 后,检验残差是否服从正态分布)
vecstable,graph      (估计 VECM 后,通过特征值检验该 VECM 系统是否为平稳过程,如
                      果所有特征值都在单位圆内部,则为平稳过程。选择项“graph”将
                      画出特征值的几何分布图)
irf create irfname,set(filename) step(#) replace
```

估计 VECM 后，将有关脉冲响应的结果存为“irfname”(可自行命名)。选择项“set

(filename)"表示建立脉冲文件"filename",使之成为当前的脉冲文件(make filename active),并将脉冲结果"irfname"存入此脉冲文件"filename"(若未使用选择项"set (filename)"指定脉冲文件,则将脉冲响应结果存入当前的脉冲文件);选择项"step(#)"表示考察#期的脉冲响应函数,默认值为"step(8)";选择项"replace"表示替代已有的同名脉冲响应结果 irfname(如果有)。一个脉冲文件"filename"可存储多个脉冲响应结果"irfname"。

irf graph irf,impulse(varname) response(varname) noci

画脉冲响应图(未正交化)。其中,选择项"impulse(varname)"用于指定脉冲变量,而选择项"response(varname)"用来指定反应变量,默认画出所有变量的脉冲响应图。选择项"noci"表示不画置信区间,默认画置信区间。

irf graph cirf,impulse(varname) response(varname)

画累积脉冲响应图(未正交化)

irf graph oirf,impulse(varname) response(varname)

画正交化的脉冲响应图

irf graph coirf,impulse(varname) response(varname)

画正交化的累积脉冲响应图

irf graph fevd,impulse(varname) response(varname)

画预测方差分解图

如果将以上命令中的"irf graph"改为"irf table",则将相应信息列表而非画图。

fcast compute prefix,step(#)

估计 VECM 后,计算被解释变量未来#期的预测值,并把预测值赋予被解释变量加上前缀"prefix"(自行确定)的变量名。

fcast graph varlist,observed

运行命令"fcast compute"后,将变量"varlist"的预测值画图,其中选择项"observed"表示与实际观测值相比较。

下面以数据集"mpyr.dta"为例[①],对美国的货币需求函数进行协整分析。该数据集包含了美国1900—1989年的以下年度宏观变量:logp(价格水平的对数),logy(名义净国民生产总值[②]的对数),logm1(M_1的对数),logmr(实际货币[③]的对数,即 logm1-logp),r(名义利率[④]),logv(货币流通速度的对数,即 logmr-logy)。

从经济理论出发,通常将货币需求函数写为

$$\log mr_t = \log m1_t - \log p_t = \beta_0 + \beta_1 \log y_t + \beta_2 r_t + \varepsilon_t \tag{21.68}$$

其中,β_1 为货币需求的收入弹性,一般认为接近于1;而 β_2 为货币需求的利率"半弹性"(semielasticity),一般为负。假定以上变量均为单整I(1)过程(参见习题),故应进行协整分析。首先,从图形上大致考察($\log mr_t$ $\log y_t$ r_t)是否存在协整关系(如图21.6):

.use mpyr.dta,clear

① 此例来自 Hayashi(2000),原始数据来自 Stock and Watson(1993)。

② 即 Net National Product(NNP)。

③ 实际货币的定义为 M1/p。

④ 6月商业票据的年利率,6-month commercial paper rate in percentage at an annual rate.

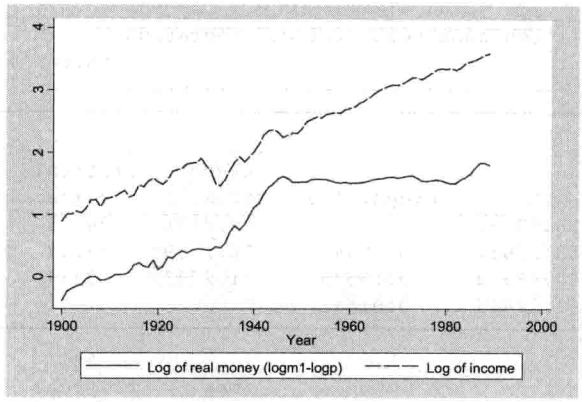


图 21.6 实际货币与收入的时间趋势

```
.line logmr logy year,lpattern("l" "_")
```

从图 21.6 可以看出,实际货币对数与收入对数的时间走势比较接近。在货币需求函数中,如果令 $\beta_1 = 1$ 并把 $\log y_t$ 移项可得

$$\log mr_t - \log y_t = \beta_0 + \beta_2 r_t + \varepsilon_t \tag{21.69}$$

上式左边可以写为,$\log mr - \log y = \log(M_1/py) \equiv \log v$,其中 $v \equiv M_1/py$ 为 M_1 的货币流通速度(M_1 velocity)。这意味着,货币流通速度之对数与名义利率存在线性关系。由于货币流通速度小于 1,故其对数小于 0,为此考虑其负数:

```
.gen _logv = -logv
```

另外,为了在数量级上更匹配,将名义利率除以 10:

```
.gen r10 =r/10
```

然后,考察_logv 与 r10 的走势图(如图 21.7):

```
.tsline _logv r10,lpattern("_")
```

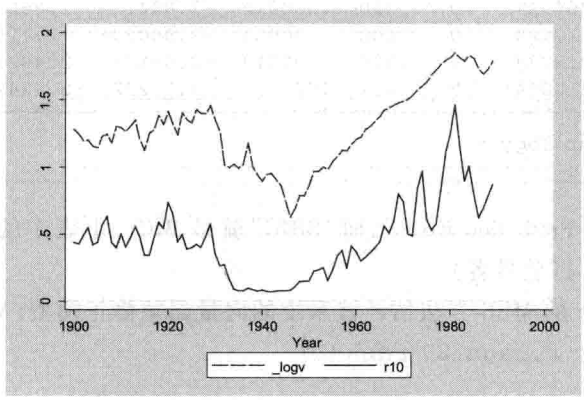


图 21.7 货币流通速度与名义利率的时间趋势

从图 21.7 可以看出,名义利率与货币流通速度对数的升降性具有一定的联动性。以上两图表明,($\log mr_t$ $\log y_t$ r_t)很可能存在长期均衡关系,即为协整系统。为此,首先要确定协整秩,即共有多少个线性无关的协整向量:

```
.vecrank logmr logy r,trend(trend) max
```

Johansen tests for cointegration

Trend: trend　　Number of obs = 88
Sample: 1902 - 1989　　Lags = 2

maximum rank	parms	LL	eigenvalue	trace statistic	5% critical value
0	15	138.03791	.	46.3731	34.55
1	20	153.13651	0.29047	16.1759*	18.17
2	23	160.58579	0.15575	1.2773	3.74
3	24	161.22445	0.01441		

maximum rank	parms	LL	eigenvalue	max statistic	5% critical value
0	15	138.03791	.	30.1972	23.78
1	20	153.13651	0.29047	14.8985	16.87
2	23	160.58579	0.15575	1.2773	3.74
3	24	161.22445	0.01441		

包含常数项与时间趋势项的协整秩迹检验(trace statistic)结果表明,只有一个线性无关的协整向量(上表中打星号者)。而最大特征值检验(max statistic)也表明,可以在5%的水平上拒绝"协整秩为0"的原假设,但无法拒绝"协整秩为1"的原假设。

其次,检验该系统所对应的VAR表示法(VAR representation)的滞后阶数:

```
.varsoc logmr logy r
```

Selection-order criteria
Sample: 1904 - 1989　　Number of obs = 86

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-251.056				.073876	5.90827	5.94272	5.99388
1	132.578	767.27	9	0.000	.000012	-2.80415	-2.66632	-2.46168*
2	148.293	31.429	9	0.000	.00001*	-2.96029*	-2.7191*	-2.36098
3	151.979	7.3723	9	0.598	.000012	-2.83672	-2.49215	-1.98055
4	162.506	21.054*	9	0.012	.000011	-2.87222	-2.42429	-1.75921

Endogenous: logmr logy r
Exogenous: cons

其中,"FPE"表示"Final Prediction Error",而"SBIC"就是BIC。上表中的大多数准则表明(包括AIC),应选择滞后二阶(打星号者)。

下面,使用Johansen的MLE方法估计该系统的向量误差修正模型(VECM):

```
.vec logmr logy r,lags(2) rank(1)
```

```
Vector error-correction model

Sample:  1902 - 1989                              No. of obs      =        88
                                                  AIC             = -3.037506
Log likelihood =  150.6503                        HQIC            =   -2.8447
Det(Sigma_ml)  =  6.54e-06                        SBIC            =  -2.55893

Equation           Parms      RMSE     R-sq      chi2     P>chi2
------------------------------------------------------------------
D_logmr                5    .050841   0.2758  31.61646   0.0000
D_logy                 5    .056773   0.3483  44.35204   0.0000
D_r                    5     1.1287   0.2133  22.50845   0.0004
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D_logmr						
_cel						
L1.	-.0533601	.0410678	-1.30	0.194	-.1338515	.0271314
logmr						
LD.	.2079032	.1107256	1.88	0.060	-.0091149	.4249214
logy						
LD.	.0086587	.101984	0.08	0.932	-.1912263	.2085438
r						
LD.	-.0063968	.0052854	-1.21	0.226	-.0167559	.0039624
_cons	.0186333	.0064766	2.88	0.004	.0059393	.0313273
D_logy						
_cel						
L1.	.0298268	.0458591	0.65	0.515	-.0600554	.119709
logmr						
LD.	.2666361	.1236437	2.16	0.031	.0242989	.5089733
logy						
LD.	.2330244	.1138823	2.05	0.041	.0098191	.4562296
r						
LD.	-.0145323	.005902	-2.46	0.014	-.0261001	-.0029646
_cons	.0157173	.0072323	2.17	0.030	.0015424	.0298923
D_r						
_cel						
L1.	-3.482578	.9117297	-3.82	0.000	-5.269536	-1.695621
logmr						
LD.	2.663613	2.458173	1.08	0.279	-2.154318	7.481544
logy						
LD.	.6533844	2.264106	0.29	0.773	-3.784182	5.09095
r						
LD.	.294868	.1173386	2.51	0.012	.0648885	.5248475
_cons	-.0001509	.1437852	-0.00	0.999	-.2819648	.281663

```
Cointegrating equations

Equation           Parms      chi2     P>chi2
------------------------------------------------------------------
_cel                   2   794.1155     0.0000

Identification:  beta is exactly identified

                 Johansen normalization restriction imposed
```

beta	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cel						
logmr	1	.	.	.	.	.
logy	-.9754246	.0346169	-28.18	0.000	-1.043273	-.9075767
r	.1124051	.0097191	11.57	0.000	.093356	.1314542
_cons	.7299535	.	.	.	.	.

上表上部为误差修正模型，下部则为协整方程（cointegrating equation），以“_ce1”来表示。我们主要对货币需求函数感兴趣，即上表中协整方程所代表的长期均衡关系。根据表中信息，可将估计的货币需求函数写为

$$\widehat{\log mr_t} = -0.73 + 0.98\log y_t - 0.11 r_t \tag{21.70}$$

其中，货币需求的收入弹性为0.98，而货币需求的利率半弹性为-0.11，符合经济理论的预期。作为对比，下面直接用OLS估计此长期均衡关系（即EG-ADF两步法）：

```
. reg logmr logy r
```

Source	SS	df	MS	Number of obs =	90
				F(2, 87) =	1169.93
Model	41.216226	2	20.608113	Prob > F =	0.0000
Residual	1.53248421	87	.017614761	R-squared =	0.9642
				Adj R-squared =	0.9633
Total	42.7487102	89	.480322587	Root MSE =	.13272

logmr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logy	.9418376	.0196178	48.01	0.000	.9028451	.9808302
r	-.0832229	.0053829	-15.46	0.000	-.0939219	-.0725239
_cons	-.7737089	.0426886	-18.12	0.000	-.8585572	-.6888606

从上表可以看出，OLS系数估计值与Johansen的MLE估计结果比较接近。当然，从理论来说，MLE估计更有效率。

下面检验VECM模型的残差是否存在自相关。如果存在自相关，则预示着需要增加滞后阶数：

```
. qui vec logmr logy r,lags(2) rank(1)
. veclmar
```

Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	6.6260	9	0.67599
2	12.5541	9	0.18384

H0: no autocorrelation at lag order

结果显示，可以接受“无自相关”的原假设。进一步检验残差的正态性：

```
. vecnorm
```

Jarque-Bera test

Equation	chi2	df	Prob > chi2
D_logmr	15.620	2	0.00041
D_logy	23.598	2	0.00001
D_r	6.147	2	0.04626
ALL	45.365	6	0.00000

Skewness test

Equation	Skewness	chi2	df	Prob > chi2
D_logmr	.04087	0.024	1	0.87564
D_logy	-.90936	12.128	1	0.00050
D_r	.04973	0.036	1	0.84897
ALL		12.189	3	0.00676

Kurtosis test

Equation	Kurtosis	chi2	df	Prob > chi2
D_logmr	5.0624	15.595	1	0.00008
D_logy	4.7687	11.470	1	0.00071
D_r	4.291	6.111	1	0.01344
ALL		33.176	3	0.00000

上表显示，可在5%的显著性水平上拒绝 D. logmr，D. logy 与 D. r 的残差项服从正态分布的原假设，尽管对 D. logmr 与 D. r 残差的偏度检验并不拒绝正态性（说明这两个变量的残差在对称性方面与正态分布较接近）。尽管 Johansen 的 MLE 估计是基于扰动项服从正态分布而推导出来的，但作为准最大似然估计量（QMLE），在更弱的非正态条件下也成立（参见 Lutkepohl，2005，p. 297）。因此，残差的非正态性对 VECM 模型的估计影响不大（但可能对区间预测产生影响）。

下面检验此 VECM 系统是否稳定，结果如图 21.8。

```
.vecstable,graph
```

Eigenvalue stability condition

Eigenvalue	Modulus
1	1
1	1
.4092107 + .4061819*i*	.576574
.4092107 - .4061819*i*	.576574
.2217304 + .07266624*i*	.233334
.2217304 - .07266624*i*	.233334

The VECM specification imposes 2 unit moduli.

结果显示，除了 VECM 模型本身所假设的单位根之外，伴随矩阵的所有特征值均落在单位圆之内，故是稳定的。

下面考察此 VECM 模型的正交化脉冲响应函数（将脉冲文件命名为 money）（结果如图 21.9）。

```
.irf create money, set (money)
step(10) replace
(file money.irf created)
(file money.irf now active)
(file money.irf updated)
.irf graph oirf
```

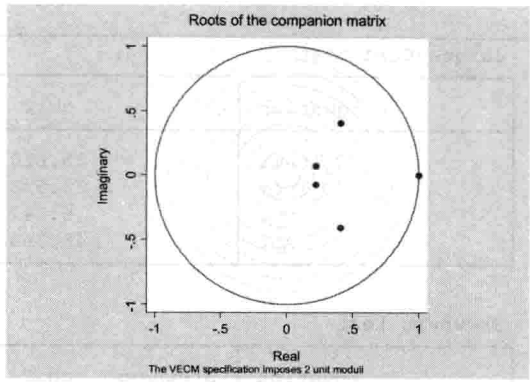


图 21.8 VECM 系统稳定性的判别图

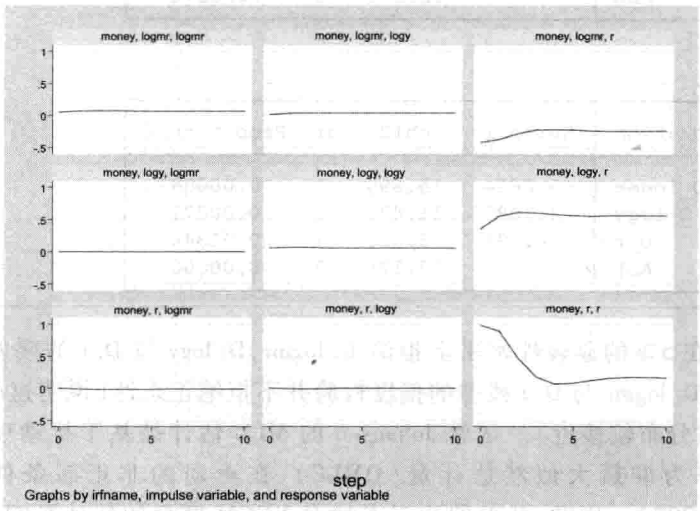


图 21.9 正交化的脉冲响应图

从图 21.9 可知，对于一个含单位根的协整系统，对一个变量的冲击可能对其自身与其他变量具有持久的影响（平稳 VAR 过程的脉冲响应不具有此性质）。估计 VECM 模型后，可以用它进行预测。假设我们仅用 1980 年以前的数据来估计 VECM 模型，然后预测 1980—1989 年的十年数据，并与实际观测值比较，结果如图 21.10。

```
.quietly vec logmr logy r if year<1980,lags(2) rank(1)
.fcast compute f_,step(10)
.fcast graph f_logmr f_logy f_r,observed lpattern("_")
```

图 21.10 似乎表明，对国民收入（logy）的预测最为准确，对货币供给（logmr）的预测次之，而利率（r）最难预测（但要注意它们的纵轴坐标单位与绝对位置都不同）。其中，利率（r）的实际观测值曾一度落在预测值 95% 的置信区间之外。

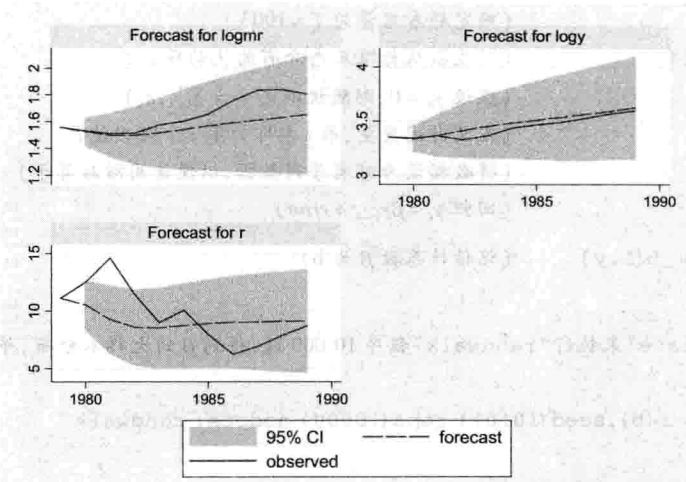


图 21.10　对未来 10 年的预测

习　题

21.1　对于 AR(2)，$y_t = 2 + \frac{5}{2}y_{t-1} - y_{t-2} + \varepsilon_t$，写出其特征方程 $\varphi(z)$，并确定其稳定性。

21.2　考虑 AR(1)模型，$y_t = \rho y_{t-1} + \varepsilon_t, t = 2, \cdots, T$，其中 $|\rho| < 1$，ε_t 为白噪声且独立于 $\{y_t\}$。记 ρ 的 OLS 估计量为 $\hat{\rho}$。

(1) 计算 $E(y_t^2)$。

(2) 根据定理 5.1，$\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{d} N(0, \text{Avar}(\hat{\rho}))$。证明 $\text{Avar}(\hat{\rho}) = 1 - \rho^2$（提示：使用同方差情形下 Avar $(\hat{\rho})$ 的简化公式）。

(3) 假设在 $\rho = 1$ 时，(2)的结论依然成立。证明 $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{p} 0$。

(4) (此题不必做)可以证明，当 $\rho = 1$ 时，$T(\hat{\rho} - \rho) \xrightarrow{d}$ 非退化分布。因此，在单位根情形下，$\hat{\rho}$ 的收敛速度为 T，快于平稳情况下 OLS 估计量的收敛速度 $\sqrt{T}$，称为“超一致”(superconsistency)，详见 Hamilton (1994, p. 486 -488)。

21.3　使用 ADF, PP, DF-GLS 及 KPSS 检验，检验数据集 mpyr. dta 中各主要变量是否含有单位根。

21.4　使用 ADF, PP, DF-GLS 及 KPSS 检验，检验数据集 lutkepohl2. dta 中以下变量是否含有单位根：ln_inc (收入的对数)，ln_consump(消费的对数)，以及 ln_inv(投资的对数)。

21.5　假设数据集 lutkepohl2. dta 中变量 ln_inv, ln_inc 及 ln_consump 均为一阶单整。使用 Johansen 方法对 (ln_inc, ln_consump, ln_inv)进行协整分析。长期均衡关系的系数符号合理吗？

附　录

A21.1　产生图 21.2 的 Stata 程序

首先定义一个叫“randwalk”的程序来产生不带漂移项的随机游走，并进行一阶自回归的 OLS 估计：

```
program randwalk,rclass
  drop _all                    (删去内存中已有数据)
```

```
set obs 100                    (确定样本容量为 T=100)
gen eps=rnormal()              (产生服从标准正态分布的扰动项 ε_t)
gen y=sum(eps)                 (假设 y_0=0,则随机游走 y_t=Σ_{s=1}^{t} ε_s)
gen t=_n                       (定义时间变量,第 t 期即为第 i 个观测值)
tsset t                        (将数据设为时间序列数据,以便使用滞后算子)
reg y L.y                      (回归 y_t=βy_{t-1}+error)
return scalar b=_b[L.y]        (记估计系数 β̂ 为 b)
end
```

然后,使用命令"simulate"来执行"randwalk"程序10 000遍,得到$\hat{\beta}$的大样本分布,并画其(核密度)经验分布图:

```
simulate beta=r(b),seed(10101) reps(10000) nodots: randwalk
kdensity beta
```

A21.2 $\alpha(L)$为绝对值可加总(AS)

证明:由于$\alpha_j \equiv -(\psi_{j+1}+\psi_{j+2}+\cdots) = -\sum_{i=j+1}^{\infty}\psi_i$,故

$$
\begin{aligned}
\sum_{j=0}^{\infty}|\alpha_j| &= \sum_{j=0}^{\infty}\left|-\sum_{i=j+1}^{\infty}\psi_i\right| \\
&= \sum_{j=0}^{\infty}\left|\sum_{i=j+1}^{\infty}\psi_i\right| \\
&\leqslant \sum_{j=0}^{\infty}\sum_{i=j+1}^{\infty}|\psi_i| \quad (\text{根据三角不等式}\ |a+b|\leqslant|a|+|b|) \\
&= \sum_{j=0}^{\infty}(|\psi_{j+1}|+|\psi_{j+2}|+\cdots) \\
&= |\psi_1|+|\psi_2|+|\psi_3|+\cdots \quad (j=0) \\
&\qquad\quad |\psi_2|+|\psi_3|+\cdots \quad (j=1) \\
&\qquad\qquad |\psi_3|+\cdots \quad (j=2) \\
&\qquad\qquad\quad +\cdots \\
&= \sum_{i=1}^{\infty} i|\psi_i| \\
&= \sum_{i=0}^{\infty} i|\psi_i| < \infty \quad (\text{因为}\ \psi(L)\ \text{为 OS})
\end{aligned}
$$

因此,$\sum_{j=0}^{\infty}|\alpha_j|<\infty$,故$\alpha(L)$为AS。

第 22 章　自回归条件异方差模型

22.1　条件异方差模型的例子

通常认为，横截面数据容易存在异方差，而时间序列数据常存在自相关。然而，Engle（1982）指出，时间序列数据也常存在一种特殊的异方差，即“自回归条件异方差”（Autoregressive Conditional Heteroskedasticity，简记 ARCH）[①]。Bollerslev（1986）对 ARCH 进行了推广，称为“Generalized ARCH”，简记 GARCH。

考察美国道琼斯股指在 1953—1990 年期间日收益率的波动，参见图 22.1。

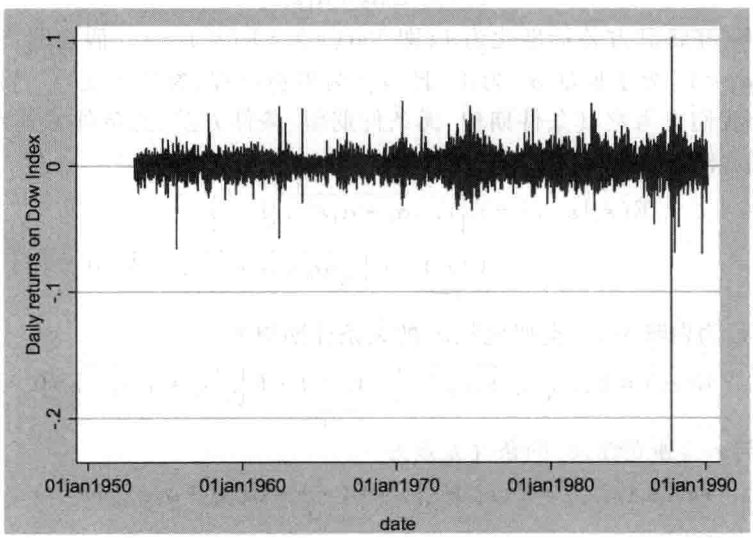


图 22.1　美国道琼斯股指 1953—1990 的日收益率[②]

从图 22.1 可以看出，股指日收益率在某一段时间内剧烈波动，而在另一段时间内风平浪静。从理论上，这可以抽象为，当本期或过去若干期的波动（方差）较大时，未来几期的波动（方差）很可能也较大；反之亦然。换言之，方差大的观测值似乎集聚在一起，而方差小的观测值似乎也集聚在一起。这被称为“波动性集聚”（volatility clustering）或“扎堆”。

在 Engle（1982）的论文发表之前，由于缺乏更好的度量，经济学家一直假设时间序列的方差是恒定的。由于 ARCH 模型考虑了方差的波动性，故可以更好地预测方差（variance forecast），在

① Robert Engle 因此于 2003 年获诺贝尔经济学奖。

② 根据数据集 dow1. dta 计算，参见习题。

金融领域有着重要的应用价值。比如,金融学中使用“VaR 方法”(Value-at-Risk)来度量金融资产所面临的风险,就依赖于对未来收益率方差的预测。因此,Engle 的贡献是一个重要突破。

22.2 ARCH 模型的性质

考虑一般的线性回归模型:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_t \tag{22.1}$$

记扰动项 ε_t 的条件方差为 $\sigma_t^2 \equiv \mathrm{Var}(\varepsilon_t \mid \varepsilon_{t-1}, \cdots)$,其中 σ_t^2 的下标 t 表示条件方差可以随时间而变。受到波动性集聚现象的启发,假设 σ_t^2 取决于上一期扰动项之平方:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \tag{22.2}$$

这就是“ARCH(1)扰动项”。更一般地,假设 σ_t^2 依赖于前 p 期扰动项之平方:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 \tag{22.3}$$

这就是“ARCH(p)扰动项”。不失一般性,以 ARCH(1)为例来考察 ARCH 扰动项的性质。假设扰动项 ε_t 的生成过程为

$$\varepsilon_t = v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \tag{22.4}$$

其中,v_t 为白噪声,并将其方差标准化为 1,即 $\mathrm{Var}(v_t) = \mathrm{E}(v_t^2) = 1$。假定 v_t 与 ε_{t-1} 相互独立,而且 $\alpha_0 > 0, 0 < \alpha_1 < 1$(为了保证 σ_t^2 为正,且 $\{\varepsilon_t\}$ 为平稳过程,参见下文)。序列 $\{\varepsilon_t\}$ 具有怎样的性质呢?下面我们来考察其条件期望、无条件期望、条件方差、无条件方差及序列相关。

由于 v_t 与 ε_{t-1} 相互独立,ε_t 的条件期望为

$$\begin{aligned} \mathrm{E}(\varepsilon_t \mid \varepsilon_{t-1}) &= \mathrm{E}\left\{ v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \,\middle|\, \varepsilon_{t-1} \right\} \\ &= \underbrace{\mathrm{E}(v_t)}_{=0} \cdot \mathrm{E}\left\{ \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \,\middle|\, \varepsilon_{t-1} \right\} = 0 \end{aligned} \tag{22.5}$$

其中,$\mathrm{E}(v_t) = 0$(v_t 为白噪声)。类似地[①],ε_t 的无条件期望为

$$\mathrm{E}(\varepsilon_t) = \mathrm{E}\left\{ v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \right\} = \underbrace{\mathrm{E}(v_t)}_{=0} \cdot \mathrm{E}\left\{ \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \right\} = 0 \tag{22.6}$$

同样地,根据 v_t 与 ε_{t-1} 独立性,ε_t 的条件方差为

$$\begin{aligned} \mathrm{Var}(\varepsilon_t \mid \varepsilon_{t-1}) &= \mathrm{E}(\varepsilon_t^2 \mid \varepsilon_{t-1}) = \underbrace{\mathrm{E}(v_t^2)}_{=1} \cdot \mathrm{E}(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \mid \varepsilon_{t-1}) \\ &= \alpha_0 + \alpha_1 \mathrm{E}(\varepsilon_{t-1}^2 \mid \varepsilon_{t-1}) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned} \tag{22.7}$$

其中,$\mathrm{E}(v_t^2) = 1$。上式就是 ARCH(1)的定义式“$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$”。$\alpha_1$ 越大,则说明上一期扰动项之平方对条件方差 σ_t^2 的冲击越大。进一步考察 ε_t 的无条件方差:

$$\begin{aligned} \mathrm{Var}(\varepsilon_t) = \mathrm{E}(\varepsilon_t^2) &= \mathrm{E}[v_t^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)] \\ &= \underbrace{\mathrm{E}(v_t^2)}_{=1} \cdot \mathrm{E}(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) = \alpha_0 + \alpha_1 \mathrm{E}(\varepsilon_{t-1}^2) \end{aligned} \tag{22.8}$$

对于差分方程“$\mathrm{E}(\varepsilon_t^2) = \alpha_0 + \alpha_1 \mathrm{E}(\varepsilon_{t-1}^2)$”,由于 $0 < \alpha_1 < 1$,故该差分方程有稳定解。令 $\mathrm{E}(\varepsilon_t^2) = \mathrm{E}(\varepsilon_{t-1}^2)$,可得 $\mathrm{E}(\varepsilon_t^2) = \dfrac{\alpha_0}{1 - \alpha_1}$。因此,ARCH 扰动项的无条件方差为常数,不随时间而

① 也可以使用迭代期望定律来证明。

变化。再来看 ε_t 与 $\varepsilon_{t-i}(i\neq 0)$ 的序列相关：

$$\begin{aligned}\mathrm{E}(\varepsilon_t\varepsilon_{t-i}) &= \mathrm{E}\{v_t v_{t-i}\sqrt{(\alpha_0+\alpha_1\varepsilon_{t-1}^2)(\alpha_0+\alpha_1\varepsilon_{t-i}^2)}\}\\ &= \underbrace{\mathrm{E}(v_t v_{t-i})}_{=0}\cdot \mathrm{E}\{\sqrt{(\alpha_0+\alpha_1\varepsilon_{t-1}^2)(\alpha_0+\alpha_1\varepsilon_{t-i}^2)}\}=0\end{aligned} \tag{22.9}$$

其中，由于 v_t 为白噪声，故 $\mathrm{E}(v_t v_{t-i})=0$。从上面的推导可以看出，扰动项 $\{\varepsilon_t\}$ 完全满足古典模型关于“同方差”[①]与“无自相关”的假定。事实上，虽然 $\{\varepsilon_t\}$ 存在条件异方差，却是白噪声！因此，高斯－马尔可夫定理成立，OLS 是最佳线性无偏估计(BLUE)。然而，OLS 显然忽略了条件异方差这一重要信息。如果我们跳出线性估计的范围，则可以找到更优的非线性估计，即最大似然估计。

22.3 ARCH 模型的 MLE 估计

对于 ARCH(1)模型，为了保证条件方差 $\sigma_t^2=\alpha_0+\alpha_1\varepsilon_{t-1}^2$ 始终为非负，必须限制参数 α_0,α_1 均为正数。如果 $\alpha_0<0$ 或 $\alpha_1<0$，则可能出现“$\sigma_t^2<0$”的情形，参见图 22.2。另外，$\alpha_1<1$ 是为了保证 $\{\varepsilon_t\}$ 为平稳过程。如果 $\alpha_1>1$，则 $\mathrm{Var}(\varepsilon_t)$ 将随时间而增大，不是平稳过程。

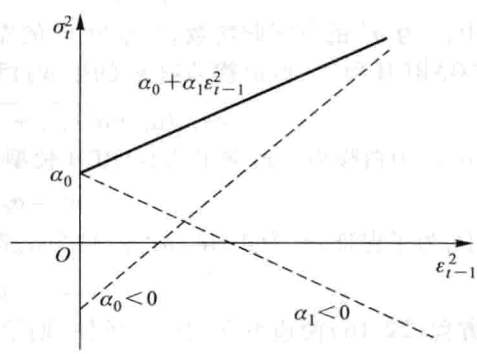


图 22.2 对 ARCH(1)参数的正约束

假设样本容量为 T。显然，在 ARCH(1)模型中，$\{\varepsilon_t\}$ 并非独立同分布的，因为相邻的扰动项通过公式“$\varepsilon_t=v_t\sqrt{\alpha_0+\alpha_1\varepsilon_{t-1}^2}$”而联系在一起。此时，如何计算样本的似然函数呢？由于 ε_t 仅依赖于 ε_{t-1}，而不依赖于 $\{\varepsilon_{t-2},\varepsilon_{t-3},\cdots\}$，故可以将 $\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T\}$ 的联合密度函数分解如下：

$$\begin{aligned}f(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T) &= f(\varepsilon_1)f(\varepsilon_2|\varepsilon_1)f(\varepsilon_3|\varepsilon_2,\varepsilon_1)\cdots f(\varepsilon_T|\varepsilon_{T-1},\cdots)\\ &= f(\varepsilon_1)f(\varepsilon_2|\varepsilon_1)f(\varepsilon_3|\varepsilon_2)\cdots f(\varepsilon_T|\varepsilon_{T-1})\end{aligned} \tag{22.10}$$

由于无条件密度函数 $f(\varepsilon_1)$ 不易计算(要用到 ε_1 的无条件方差的表达式，导致在似然函数中出现非线性项)，常将 $f(\varepsilon_1)$ 忽略不计，即考虑在 ε_1 给定情况下的条件最大似然估计法(conditional MLE)。假设 $\varepsilon_t\sim N(0,\sigma_t^2)$，而 $\sigma_t^2=\alpha_0+\alpha_1\varepsilon_{t-1}^2$，可得似然函数：

$$L=\prod_{t=2}^{T}\frac{1}{\sqrt{2\pi(\alpha_0+\alpha_1\varepsilon_{t-1}^2)}}\exp\left\{-\frac{\varepsilon_t^2}{2(\alpha_0+\alpha_1\varepsilon_{t-1}^2)}\right\} \tag{22.11}$$

$$\ln L=-\frac{T-1}{2}\ln 2\pi-\frac{1}{2}\sum_{t=2}^{T}\ln(\alpha_0+\alpha_1\varepsilon_{t-1}^2)-\frac{1}{2}\sum_{t=2}^{T}\frac{\varepsilon_t^2}{\alpha_0+\alpha_1\varepsilon_{t-1}^2} \tag{22.12}$$

根据方程(22.1)，将“$\varepsilon_t=y_t-\boldsymbol{x}_t'\boldsymbol{\beta}$”代入上式，则对数似然函数 $\ln L$ 成为参数 $(\alpha_0,\alpha_1,\beta_0,\beta_1)$

① 指的是“无条件方差”相同。需要注意的是，本章的条件方差指的是，给定扰动项的滞后值与解释变量 $\boldsymbol{X}$，即 $\mathrm{Var}(\varepsilon_t|\varepsilon_{t-1},\cdots;\boldsymbol{X})$；而第 3 章的条件方差指的是，给定解释变量 $\boldsymbol{X}$，即 $\mathrm{Var}(\varepsilon_t|\boldsymbol{X})$。因此，本章的“无条件方差”就是第 3 章的“条件方差”，故适用高斯－马尔可夫定理。

的函数。可以对$(\alpha_0,\alpha_1,\beta_0,\beta_1)$求偏导得到 ln L 的最大值，通常由计算机数值计算来进行。因此，对 ARCH 模型进行 MLE 估计的特点是，对原方程$(y_t=\boldsymbol{x}_t'\boldsymbol{\beta}+\varepsilon_t)$与条件方差方程$(\sigma_t^2=\alpha_0+\alpha_1\varepsilon_{t-1}^2)$同时进行估计。

类似地，如果要估计 ARCH(p)，则将$\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p\}$(即前 p 个观测值)视为给定，然后使用条件 MLE。即使扰动项不服从正态分布，作为准最大似然估计量(QMLE)，仍可能是一致的。

22.4 GARCH 模型

在 ARCH(p)模型中，如果 p 很大，则要估计很多参数，会损失样本容量。Bollerslev (1986)提出 GARCH，使得待估计参数减少，而对未来条件方差的预测更加准确。其基本思想是，在 ARCH 模型的基础上，再加上 σ_t^2 的自回归部分，即 σ_t^2 还是$\{\sigma_{t-1}^2,\cdots,\sigma_{t-p}^2\}$的函数。GARCH($p$, q)的模型设定为

$$\sigma_t^2=\alpha_0+\alpha_1\varepsilon_{t-1}^2+\cdots+\alpha_q\varepsilon_{t-q}^2+\gamma_1\sigma_{t-1}^2+\cdots+\gamma_p\sigma_{t-p}^2 \tag{22.13}$$

其中，p 为 σ_t^2 的自回归阶数，而 q 为 ε_t^2 的滞后阶数。在 Stata 中，称 ε_{t-i}^2 为“ARCH 项”，而称 σ_{t-i}^2 为“GARCH 项”。假定扰动项 ε_t 的生成过程为

$$\varepsilon_t=v_t\sqrt{\alpha_0+\alpha_1\varepsilon_{t-1}^2+\cdots+\alpha_q\varepsilon_{t-q}^2+\gamma_1\sigma_{t-1}^2+\cdots+\gamma_p\sigma_{t-p}^2} \tag{22.14}$$

其中，v_t 为白噪声。最常用的 GARCH 模型为 GARCH(1,1)：

$$\sigma_t^2=\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1\sigma_{t-1}^2 \tag{22.15}$$

其中，为了保证 σ_t^2 为正，$\alpha_0,\alpha_1,\gamma_1$ 均为正数。GARCH(1,1)扰动项 ε_t 的生成过程为

$$\varepsilon_t=v_t\sqrt{\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1\sigma_{t-1}^2} \tag{22.16}$$

将方程(22.16)两边平方，取(无条件)期望可得

$$\begin{aligned}
\mathrm{E}(\varepsilon_t^2)&=\underbrace{\mathrm{E}(v_t^2)}_{=1}\cdot\mathrm{E}(\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1\sigma_{t-1}^2)\\
&=\alpha_0+\alpha_1\mathrm{E}(\varepsilon_{t-1}^2)+\gamma_1\mathrm{E}(\sigma_{t-1}^2) && (\sigma_{t-1}^2\equiv\mathrm{E}(\varepsilon_{t-1}^2|\varepsilon_{t-2}))\\
&=\alpha_0+\alpha_1\mathrm{E}(\varepsilon_{t-1}^2)+\gamma_1\mathrm{E}[\mathrm{E}(\varepsilon_{t-1}^2|\varepsilon_{t-2})] && (\text{迭代期望定律})\\
&=\alpha_0+\alpha_1\mathrm{E}(\varepsilon_{t-1}^2)+\gamma_1\mathrm{E}(\varepsilon_{t-1}^2)\\
&=\alpha_0+(\alpha_1+\gamma_1)\mathrm{E}(\varepsilon_{t-1}^2)
\end{aligned} \tag{22.17}$$

在上式的推导中使用了迭代期望定律。由此可知，为了保证$\{\varepsilon_t\}$为平稳过程(无条件方差不发散)，必须要求 $\alpha_1+\gamma_1<1$。

为何使用 GARCH 模型能节省待估参数？直观来说，因为 σ_{t-1}^2 中已经包含了$\{\varepsilon_{t-2}^2,\cdots,\varepsilon_{t-p-1}^2\}$的信息。比如，对 GARCH(1,1)使用迭代法可得

$$\begin{aligned}
\sigma_t^2&=\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1\sigma_{t-1}^2\\
&=\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1(\alpha_0+\alpha_1\varepsilon_{t-2}^2+\gamma_1\sigma_{t-2}^2)\\
&=\alpha_0+\alpha_0\gamma_1+\alpha_1\varepsilon_{t-1}^2+\alpha_1\gamma_1\varepsilon_{t-2}^2+\gamma_1^2\sigma_{t-2}^2\\
&=\cdots\\
&=\alpha_0(1+\gamma_1+\gamma_1^2+\cdots)+\alpha_1(\varepsilon_{t-1}^2+\gamma_1\varepsilon_{t-2}^2+\gamma_1^2\varepsilon_{t-3}^2+\cdots)\\
&=\frac{\alpha_0}{1-\gamma_1}+\alpha_1(\varepsilon_{t-1}^2+\gamma_1\varepsilon_{t-2}^2+\gamma_1^2\varepsilon_{t-3}^2+\cdots)
\end{aligned} \tag{22.18}$$

由此可见，在某种意义上[①]，GARCH(1,1)等价于无穷阶ARCH模型。因此，如果将σ_{t-1}^2作为解释变量引入，常可把高阶ARCH(p)模型简化为GARCH(1,1)。对GARCH模型可同样使用MLE估计。

22.5 何时使用ARCH或GARCH模型

只有在扰动项存在条件异方差时，才需要使用ARCH或GARCH模型。那么，如何判断扰动项是否存在条件异方差呢？初步的方法可以画时间序列图，看看是否存在“波动性集聚”。

严格的统计检验包括以下三种方法。

方法一 首先，用OLS估计原方程“$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_t$”，得到残差序列$\{e_t\}$。其次，用OLS估计辅助回归，$e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2 + error_t$，并检验原假设“$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$”（不存在条件异方差）。Engle（1982）提出进行LM检验，其检验统计量为$TR^2 \xrightarrow{d} \chi^2(p)$，其中$T$为样本容量，$R^2$为上述辅助回归的可决系数。如果拒绝$H_0$，则认为应使用ARCH或GARCH模型。

在Stata中，此LM检验可通过命令`reg`的“后估计命令”（postestimation command）`estat archlm`来实现。

方法二 可以对残差平方序列$\{e_t^2\}$进行Q检验，检验其序列相关性。如果$\{e_t^2\}$存在自相关，则认为ε_t存在条件异方差。

方法三 最为直接的方法是，在估计ARCH或GARCH模型之后，看条件方差方程中的系数（即所有α与γ）是否显著。

22.6 ARCH与GARCH模型的扩展

1. ARCH-M

金融理论认为，金融资产的风险越高，其期望收益率也应该越高，这样才会有人愿意持有它。超出正常期望收益率的部分，称为“风险溢价”（risk premium）。但在标准的ARCH模型中，变量的均值与条件方差却没有关系。Engle, Lilien and Robins（1987）提出了如下“ARCH-in-Mean模型”（简记ARCH-M）。

假设金融资产的超额收益率满足以下方程：

$$y_t = \underbrace{\beta + \delta\sigma_t^2}_{\text{risk premium}} + \varepsilon_t \tag{22.19}$$

其中，y_t为“超额收益率”（excess return），即超出无风险的国库券收益率的部分；ε_t为对超额收益率不可预见的冲击；而$(\beta + \delta\sigma_t^2)$为风险溢价，是条件方差$\sigma_t^2$的增函数，即$\delta > 0$。假设$\varepsilon_t$服从ARCH($p$)过程，$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2$。可以使用MLE对方程(22.19)与条件方差方程同时进行估计。由于超额收益率y_t的期望（mean）中包含一个条件方差项$(\delta\sigma_t^2)$，故名“ARCH-in-Mean”。

---

① 要求此ARCH(∞)的系数呈几何级数递减。

2. TARCH

"坏消息"对资产价格波动性的影响可能大于好消息的影响。Glosten, Jagannathan and Runkle (1993) 提出了非对称 (asymmetric) 的"门限 GARCH"模型 (Threshold GARCH, 简记 TARCH)。

假设条件方差方程为

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \lambda_1 \underbrace{\varepsilon_{t-1}^2 \cdot \mathbf{1}(\varepsilon_{t-1} > 0)}_{\text{TARCH}} + \beta_1 \sigma_{t-1}^2 \tag{22.20}$$

其中,$\mathbf{1}(\cdot)$为示性函数,即当 $\varepsilon_{t-1} > 0$ 时,取值为1;反之,则为0。Stata 称"$\varepsilon_{t-1}^2 \cdot \mathbf{1}(\varepsilon_{t-1} > 0)$"为"TARCH"项。

3. EGARCH

在标准的 GARCH 模型中,对参数的取值有所限制,给 MLE 估计带来不便。为此,考虑以下对数形式的条件方差方程:

$$\ln \sigma_t^2 = \alpha_0 + \alpha_1 \underbrace{(\varepsilon_{t-1}/\sigma_{t-1})}_{\text{EARCH}} + \lambda_1 \underbrace{|\varepsilon_{t-1}/\sigma_{t-1}|}_{\text{EARCH_a}} + \beta_1 \underbrace{\ln \sigma_{t-1}^2}_{\text{EGARCH}} \tag{22.21}$$

其中,$(\varepsilon_{t-1}/\sigma_{t-1})$为 ε_{t-1} 的标准化(除以自身的标准差),Stata 称为"EARCH"项。只要 $\alpha_1 \neq 0$,则这个模型也包括了非对称效应(类似于 TARCH)。$|\varepsilon_{t-1}/\sigma_{t-1}|$表示对称效应,Stata 称为"EARCH_a"项(a 表示"absolute value",即绝对值)。由于 σ_t^2 为指数形式,故称为"指数 GARCH"(Exponential GARCH,简记 EGARCH)。Stata 称 $\ln \sigma_{t-1}^2$ 为"EGARCH"项。EGARCH 的优点在于,无论 $\ln \sigma_t^2$ 取何值,都有 $\sigma_t^2 = \exp(\ln \sigma_t^2) > 0$,故对方程(22.21)中的所有参数都没有任何限制。

4. 带 ARMA 的 GARCH

考虑如下线性回归模型:

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + u_t \tag{22.22}$$

其中,扰动项 u_t 为 ARMA(p,q)过程:

$$u_t = \sum_{i=1}^{p} \rho_i u_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{22.23}$$

其中,ε_t 为 GARCH(或 ARCH)扰动项。将方程(22.23)代入(22.22)可得

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \sum_{i=1}^{p} \rho_i (y_{t-i} - \boldsymbol{x}_{t-i}' \boldsymbol{\beta}) + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{22.24}$$

方程(22.24)被称为"带 ARMA 的 GARCH"(GARCH with ARMA terms)。

5. 在条件方差方程中引入解释变量

例如,为了考虑在"9·11"恐怖袭击后是否波动性增大了,可以引入虚拟变量

$$D_t = \begin{cases} 1, t \geqslant 2001/09/11 \\ 0, t < 2001/09/11 \end{cases} \tag{22.25}$$

然后考虑以下 GARCH(1,1)模型:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma D_t \tag{22.26}$$

6. 使用非正态扰动项

如果被解释变量(比如,某些金融变量)的分布函数存在厚尾,则小概率事件比在正态分布情况下更容易发生。此时,可以选择让扰动项服从 $t(k)$ 分布而非正态分布来估计 ARCH 或 GARCH 模型。在进行 MLE 估计时,将 t 分布的自由度 k 也作为待估参数。

22.7 ARCH 与 GARCH 的 Stata 命令及实例

有关 ARCH 与 GARCH 的 Stata 命令包括

```
arch y x1 x2,arch(1/3)                (ARCH(3))
arch y x1 x2,arch(1) garch(1)         (GARCH(1,1))
arch y x1 x2,ar(1) ma(1) arch(1) garch(1)  (带 ARMA(1,1)的 GARCH(1,1))
arch y x1 x2,arch(1) dist(t)          (ARCH(1),扰动项服从 t 分布)
arch y x1 x2,arch(1) het(z1 z2)       (ARCH(1),将 z1,z2 加入条件方差方程)
arch y x1 x2,arch(1) garch(1) tarch(1)  (GARCH(1,1)加上 TARCH(1))
arch y x1 x2,earch(1) egarch(1)       (EGARCH(1,1))
arch y x1 x2,arch(1/3) archm          (ARCH(3)加上 ARCH-M)
```

Stata 允许对 ARCH 或 GARCH 的模型设定进行更多的变化,详见"help arch"。Stata 手册提示,对 GARCH 模型进行 MLE 估计,通常须花费较长时间进行迭代计算,甚至会出现不收敛的情形。

下面以数据集 sp500.dta 为例,对 1981 年 1 月至 1991 年 4 月美国标准普尔股指(S&P 500)的日收益率进行 ARCH/GARCH 分析。该数据集包含以下变量:r(股指日收益率),t(日期,$1 \leqslant t \leqslant 2\,783$)。

首先,看日收益率的时间趋势图,结果如图 22.3:

```
.use sp500.dta,clear
.line r t
```

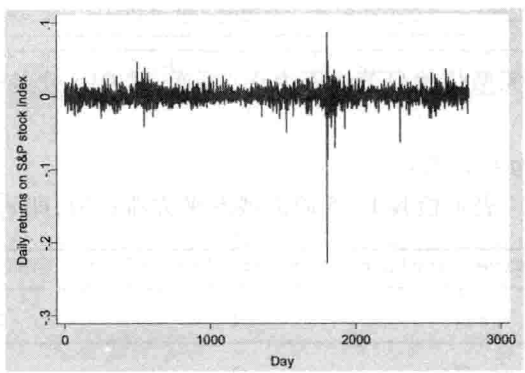


图 22.3 股指日收益率的时间趋势

从图 22.3 可以比较明显地看出,存在波动性集聚。作为对照,先考虑一个自回归模型。为此,用信息准则来确定自回归模型的阶数。将 AR(p)视为 1 维 VAR(p),则可使用 VAR 系列的命令:

```
.varsoc r,maxlag(8)
```

Selection-order criteria
Sample: 9 - 2783 Number of obs = 2775

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	8612.3				.000118	-6.20634	-6.20557	-6.2042
1	8616.31	8.0317	1	0.005	.000118	-6.20851	-6.20697	-6.20424*
2	8619.54	6.443	1	0.011	.000118	-6.21012	-6.2078	-6.20371
3	8620.26	1.4422	1	0.230	.000118	-6.20991	-6.20683	-6.20137
4	8622.07	3.6252	1	0.057	.000118	-6.2105	-6.20664	-6.19982
5	8625.86	7.5913*	1	0.006	.000117*	-6.21251*	-6.20789*	-6.1997
6	8625.9	.0624	1	0.803	.000117	-6.21182	-6.20642	-6.19686
7	8626.13	.46369	1	0.496	.000117	-6.21126	-6.20509	-6.19417
8	8626.62	.9867	1	0.321	.000118	-6.2109	-6.20395	-6.19167

Endogenous: r
Exogenous: cons

上表显示,大多数准则均选择 AR(5)模型。因此,用 OLS 估计 AR(5)模型[①]:

```
.reg r L(1/5).r
```

其中,"L(1/5)"表示 1 - 5 阶滞后。

Source	SS	df	MS	Number of obs = 2778
				F(5, 2772) = 5.49
Model	.00321088	5	.000642176	Prob > F = 0.0000
Residual	.324428204	2772	.000117038	R-squared = 0.0098
				Adj R-squared = 0.0080
Total	.327639084	2777	.000117983	Root MSE = .01082

r	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r						
L1.	.0562165	.0189661	2.96	0.003	.0190273	.0934056
L2.	-.047366	.0189723	-2.50	0.013	-.0845672	-.0101647
L3.	-.0191588	.0189899	-1.01	0.313	-.0563945	.0180769
L4.	-.0389725	.0189717	-2.05	0.040	-.0761726	-.0017723
L5.	.0524083	.018952	2.77	0.006	.0152468	.0895697
_cons	.0004126	.000206	2.00	0.045	8.66e-06	.0008166

上表显示,5 阶滞后的系数依然显著地不为 0。下面,对 OLS 残差是否存在 ARCH 效应进行 LM 检验。

```
.estat archlm,lags(1/5)
```

其中,选择项"lags(1/5)"表示检验 1 - 5 阶的残差平方滞后项,即 $e_{t-1}^2,\cdots,e_{t-5}^2$。

LM test for autoregressive conditional heteroskedasticity (ARCH)

lags(*p*)	chi2	df	Prob > chi2
1	45.415	1	0.0000
2	72.001	2	0.0000
3	80.514	3	0.0000
4	80.693	4	0.0000
5	103.418	5	0.0000

H0: no ARCH effects *vs.* H1: ARCH(*p*) disturbance

① 也可使用命令"var r,lags(1/5)"来得到同样的结果。此处为了使用后估计命令"estat archlm",故使用命令"reg r L(1/5).r"。

上表显示，对 ARCH(1)—ARCH(5)的检验结果均表明，存在显著的 ARCH 效应。

下面，通过画图更直观地考察 OLS 的残差平方是否存在自相关：

```
.predict e1,res
(5 missing values generated)
.g e2 = e1^2
(5 missing values generated)
.ac e2（画残差平方的自相关图，结果如图 22.4）
.pac e2 （画残差平方的偏自相关图，结果如图 22.5）
```

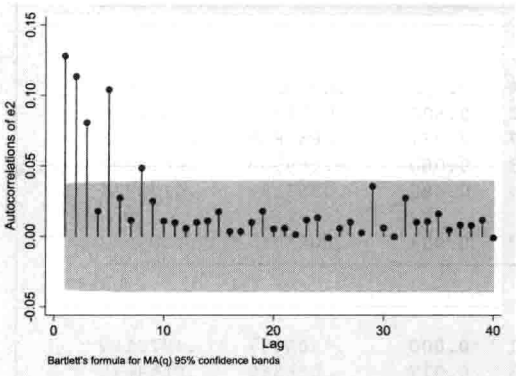


图 22.4 残差平方的自相关图

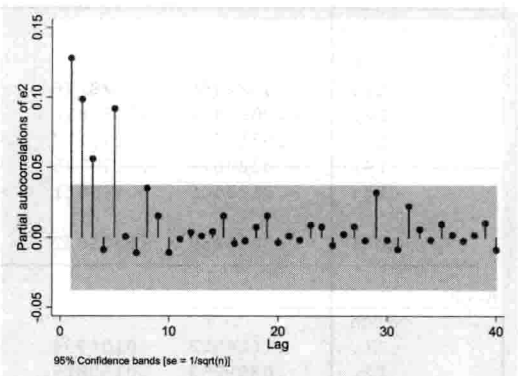


图 22.5 残差平方的偏自相关图

```
.corrgram e2,lags(10)
```

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.1279	0.1279	45.48	0.0000		
2	0.1134	0.0987	81.28	0.0000		
3	0.0804	0.0562	99.291	0.0000		
4	0.0179	-0.0088	100.18	0.0000		
5	0.1041	0.0919	130.35	0.0000		
6	0.0273	0.0007	132.42	0.0000		
7	0.0113	-0.0111	132.77	0.0000		
8	0.0486	0.0353	139.36	0.0000		
9	0.0250	0.0155	141.11	0.0000		
10	0.0107	-0.0107	141.43	0.0000		

从以上结果可以看出，无论是自相关图、偏自相关图，还是 Q 检验，均显示 OLS 残差之平方序列$\{e_t^2\}$存在自相关，故扰动项存在条件异方差，即波动性集聚。此结论与 LM 检验的结果相一致。为此，考察 ARCH(p)模型。为了确定 p，估计序列$\{e_t^2\}$的自回归阶数：

```
.varsoc e2
```

Selection-order criteria
Sample: 10 - 2783 Number of obs = 2774

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	15210.3				1.0e-06	-10.9656	-10.9649	-10.9635
1	15233.2	45.742	1	0.000	1.0e-06	-10.9814	-10.9798	-10.9771
2	15246.8	27.155	1	0.000	9.9e-07	-10.9905	-10.9881	-10.984
3	15251.2	8.7769*	1	0.003	9.8e-07*	-10.9929*	-10.9898*	-10.9844*
4	15251.3	.21651	1	0.642	9.9e-07	-10.9923	-10.9884	-10.9816

Endogenous: e2
Exogenous: _cons

所有的准则均显示(上表中打星号者),应考虑 ARCH(3)模型:

```
.arch r L(1/5).r,arch(1/3) nolog
```

```
ARCH family regression

Sample: 6 - 2783                                  Number of obs   =      2778
Distribution: Gaussian                            Wald chi2(5)    =     12.72
Log likelihood =  8947.899                        Prob > chi2     =    0.0261

------------------------------------------------------------------------------
             |                 OPG
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r            |
           r |
         L1. |   .0522458   .0228176     2.29   0.022     .0075241    .0969676
         L2. |  -.0094867   .0181812    -0.52   0.602    -.0451213    .0261478
         L3. |  -.0311578   .0176172    -1.77   0.077    -.0656869    .0033713
         L4. |  -.0320676   .0170775    -1.88   0.060     -.065539    .0014037
         L5. |  -.0103408   .0149861    -0.69   0.490    -.0397131    .0190314
             |
       _cons |   .0005433   .0001832     2.97   0.003     .0001843    .0009023
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .1606503   .0104934    15.31   0.000     .1400837    .1812169
         L2. |   .0359253   .0150815     2.38   0.017      .006366    .0654845
         L3. |    .118723   .0192852     6.16   0.000     .0809247    .1565214
             |
       _cons |   .0000687   1.67e-06    41.10   0.000     .0000654     .000072
------------------------------------------------------------------------------
```

上表显示,所有 ARCH 项均很显著。下面估计更为简洁的 GARCH(1,1)模型:

```
.arch r L(1/5).r,arch(1) garch(1) nolog
```

```
ARCH family regression

Sample: 6 - 2783                                  Number of obs   =      2778
Distribution: Gaussian                            Wald chi2(5)    =     13.94
Log likelihood =  8999.049                        Prob > chi2     =    0.0160

------------------------------------------------------------------------------
             |                 OPG
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r            |
           r |
         L1. |   .0666247   .0231129     2.88   0.004     .0213243    .1119252
         L2. |  -.0153185   .0226133    -0.68   0.498    -.0596398    .0290027
         L3. |  -.0119301   .0212921    -0.56   0.575    -.0536618    .0298016
         L4. |  -.0434335   .0208996    -2.08   0.038    -.0843959   -.0024711
         L5. |  -.0011428    .019924    -0.06   0.954    -.0401931    .0379075
             |
       _cons |   .0005794   .0001715     3.38   0.001     .0002433    .0009155
-------------+----------------------------------------------------------------
ARCH         |
        arch |
         L1. |   .0893397   .0033301    26.83   0.000     .0828128    .0958666
             |
       garch |
         L1. |   .8642674   .0084824   101.89   0.000     .8476422    .8808925
             |
       _cons |   4.94e-06   5.10e-07     9.69   0.000     3.94e-06    5.94e-06
------------------------------------------------------------------------------
```

上表显示，ARCH(1)与 GARCH(1)项均很显著。

考虑到股市中坏消息与好消息的效应可能不对称，下面在 GARCH(1,1)模型中加入一个 TARCH 项：

```
.arch r L(1/5).r,arch(1) garch(1) tarch(1) nolog
```

ARCH family regression

Sample: 6 - 2783 — Number of obs = 2778
Distribution: Gaussian — Wald chi2(5) = 12.30
Log likelihood = 9016.976 — Prob > chi2 = 0.0309

r	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
r						
r						
L1.	.06502	.0226963	2.86	0.004	.0205361	.109504
L2.	-.006217	.0223478	-0.28	0.781	-.0500179	.0375839
L3.	-.0101253	.0206436	-0.49	0.624	-.0505861	.0303354
L4.	-.0373281	.0200826	-1.86	0.063	-.0766892	.0020329
L5.	.0001177	.0194354	0.01	0.995	-.0379749	.0382103
_cons	.0003849	.0001735	2.22	0.027	.0000448	.000725
ARCH						
arch						
L1.	.126726	.0049348	25.68	0.000	.117054	.136398
tarch						
L1.	-.0915795	.008728	-10.49	0.000	-.108686	-.074473
garch						
L1.	.8681879	.0094734	91.65	0.000	.8496205	.8867554
_cons	5.20e-06	4.86e-07	10.70	0.000	4.25e-06	6.15e-06

上表显示，TARCH 项很显著，即存在不对称效应。而且，不对称效应的规模(-0.09)几乎接近对称效应(0.13)。TARCH 项的负号表明，“好消息”对资产价格波动性的影响小于“坏消息”。

考虑到收益率中可能包含风险溢价，下面估计“ARCH-in-Mean”模型。

```
.arch r L(1/5).r,arch(1) archm nolog
```

ARCH family regression

Sample: 6 - 2783 — Number of obs = 2778
Distribution: Gaussian — Wald chi2(6) = 452.49
Log likelihood = 8897.475 — Prob > chi2 = 0.0000

r	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
r						
r						
L1.	.0387874	.0241306	1.61	0.108	-.0085077	.0860825
L2.	.0078756	.0158632	0.50	0.620	-.0232156	.0389668
L3.	-.0193001	.0122494	-1.58	0.115	-.0433085	.0047082
L4.	-.1074494	.0089398	-12.02	0.000	-.1249712	-.0899277
L5.	.0043577	.0148722	0.29	0.770	-.0247913	.0335068
_cons	.0006864	.0002508	2.74	0.006	.0001949	.001178
ARCHM						
sigma2	-1.18805	1.739309	-0.68	0.495	-4.597033	2.220933
ARCH						
arch						
L1.	.2360421	.0095681	24.67	0.000	.217289	.2547952
_cons	.0000794	1.56e-06	50.75	0.000	.0000763	.0000825

上表显示,ARCHM 项并不显著,且符号为负(风险越高,则收益率越低)。之所以出现这种反常现象,可能因为这里考察的是股指(S&P 500),而非个股。

下面考虑 EGARCH(1,1)模型。但输入命令"arch r L(1/5).r,earch(1) egarch(1) nolog"却无法得到收敛的结果。为了演示目的,转而使用以下命令:

```
.arch r L(1/3).r,earch(1) egarch(1) nolog
```

```
ARCH family regression

Sample: 4 - 2783                                  Number of obs   =      2780
Distribution: Gaussian                            Wald chi2(3)    =      8.40
Log likelihood =  9011.402                        Prob > chi2     =    0.0385

                                 OPG
           r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r            |
           r |
         L1. |    .059919   .0217624     2.75   0.006     .0172654    .1025726
         L2. |   .0068176   .0209431     0.33   0.745    -.0342302    .0478653
         L3. |  -.0165195   .0201316    -0.82   0.412    -.0559768    .0229378
             |
       _cons |    .000252   .0001718     1.47   0.142    -.0000848    .0005889
-------------+----------------------------------------------------------------
ARCH         |
      earch  |
         L1. |  -.0781144   .0070599   -11.06   0.000    -.0919515   -.0642773
             |
    earch_a  |
         L1. |   .1558427   .0117046    13.31   0.000     .1329021    .1787834
             |
     egarch  |
         L1. |   .9635746   .0042339   227.58   0.000     .9552763     .971873
             |
       _cons |  -.3295685   .0393753    -8.37   0.000    -.4067427   -.2523942
------------------------------------------------------------------------------
```

上表显示,非对称效应(earch)与对称效应(earch_a)均十分显著。而且,前者的规模约为后者的一半。非对称效应(earch)的符号为负,表明"坏消息"的作用更大。

在以上的 ARCH, GARCH, TARCH, ARCHM, EGARCH 估计中,均假设扰动项服从正态分布。但股指收益率可能存在厚尾。为此,将日收益率的核密度图(参见第 27 章)与正态分布对比(如图 22.6):

```
.kdensity r,normopt(lpattern("-"))
```

其中,选择项"normopt(lpattern("-"))"表示以虚线作为比较的正态密度。

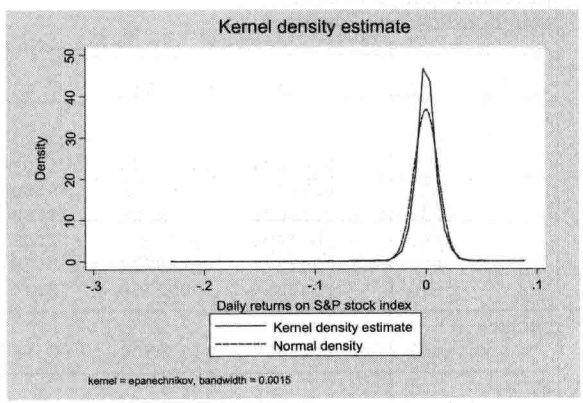


图 22.6 日收益率的核密度与正态密度

上图显示，日收益率的核密度图很可能存在厚尾，尤其在分布的左端。

下面对扰动项的正态性进行严格的统计检验：

```
.quietly var r,lags(1/5)
.varnorm
```

Jarque-Bera test

Equation	chi2	df	Prob > chi2
r	6.1e+05	2	0.00000
ALL	6.1e+05	2	0.00000

Skewness test

Equation	Skewness	chi2	df	Prob > chi2
r	-3.4412	5482.932	1	0.00000
ALL		5482.932	1	0.00000

Kurtosis test

Equation	Kurtosis	chi2	df	Prob > chi2
r	75.047	6.0e+05	1	0.00000
ALL		6.0e+05	1	0.00000

以上各检验均强烈拒绝“扰动项服从正态分布”的原假设。为此，假设扰动项服从 t 分布，重新用 GARCH(1,1)进行估计：

```
.arch r L(1/5).r,arch(1) garch(1) dist(t) nolog
```

ARCH family regression

Sample: 6 - 2783　　Number of obs = 2778
Distribution: t　　Wald chi2(5) = 12.82
Log likelihood = 9135.102　　Prob > chi2 = 0.0251

r	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
r						
r						
L1.	.0487162	.0188628	2.58	0.010	.0117457	.0856866
L2.	-.0194403	.0188372	-1.03	0.302	-.0563606	.01748
L3.	-.027825	.0184064	-1.51	0.131	-.0639009	.008251
L4.	-.0263707	.0183287	-1.44	0.150	-.0622943	.0095528
L5.	-.0080478	.0179622	-0.45	0.654	-.0432531	.0271576
_cons	.0005129	.0001579	3.25	0.001	.0002035	.0008224
ARCH						
arch						
L1.	.0354574	.0070274	5.05	0.000	.021684	.0492308
garch						
L1.	.9391238	.0107548	87.32	0.000	.9180448	.9602028
_cons	2.21e-06	5.67e-07	3.90	0.000	1.10e-06	3.32e-06
/lndfm2	1.35163	.1309665	10.32	0.000	1.094941	1.60832
df	5.86372	.506018			4.989006	6.994414

最后,对 GARCH(1,1)模型的条件方差进行预测:

```
.quietly arch r L(1/5).r,arch(1) garch(1) nolog
.predicth,variance
.line h t
```

(画条件方差的时间趋势,结果如图 22.7 所示)

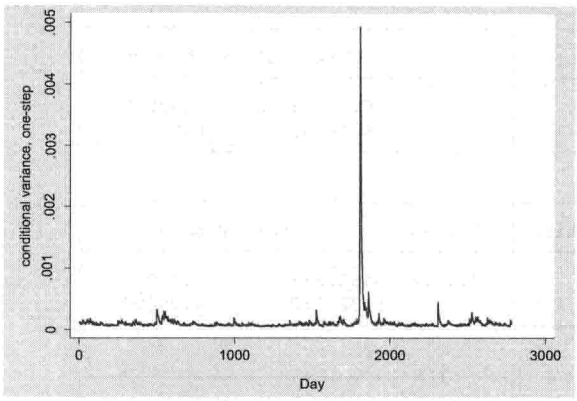


图 22.7 条件方差的时间趋势

图 22.7 显示,日收益率的条件方差时有波动,有时甚至急剧上升。如果使用 OLS 估计,则无法得到这些信息(OLS 将方差假定为常数,即一条水平线)。

22.8 多维 GARCH 模型(选读)

从概念上看,单变量的 GARCH 模型不难推广到多维 GARCH,即让当期扰动项 $\boldsymbol{\varepsilon}_t$ 的条件协方差矩阵 $\boldsymbol{H}_t$ 依赖于上一期扰动项的"平方项" $\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}'$ 与上一期的条件协方差矩阵 $\boldsymbol{H}_{t-1}$。考虑最简单的双变量情形:

$$\boldsymbol{y}_t \equiv \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix} \begin{pmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \equiv \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \tag{22.27}$$

其中,$\boldsymbol{C} \equiv \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix}$ 为待估系数矩阵,$\boldsymbol{x}_t \equiv (x_{1t} \quad x_{2t} \quad x_{3t})'$ 为解释变量(可以包含 $\boldsymbol{y}_t$ 的滞后项)。记扰动项 $\boldsymbol{\varepsilon}_t$ 的条件协方差矩阵为 $\boldsymbol{H}_t$:

$$\boldsymbol{H}_t \equiv \mathrm{Var}\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} = \begin{pmatrix} \sigma_{11,t}^2 & \sigma_{12,t}^2 \\ \sigma_{21,t}^2 & \sigma_{22,t}^2 \end{pmatrix} \tag{22.28}$$

由于 $\boldsymbol{H}_t$ 为对称矩阵,故所有信息均包含于主对角线及以下的元素中。为此,将此对称矩阵的下三角部分取出,按顺序叠放成列向量:

$$\mathrm{vech}\begin{pmatrix} \sigma_{11,t}^2 & \sigma_{12,t}^2 \\ \sigma_{21,t}^2 & \sigma_{22,t}^2 \end{pmatrix} \equiv \begin{pmatrix} \sigma_{11,t}^2 \\ \sigma_{21,t}^2 \\ \sigma_{22,t}^2 \end{pmatrix} \tag{22.29}$$

其中,vech 称为“半向量化算子”(half vectorization)。类似地,可将“平方项”矩阵 $\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}'$ 半向量化:

$$\operatorname{vech}(\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}') = \operatorname{vech}\left[\begin{pmatrix}\varepsilon_{1,t-1}\\ \varepsilon_{2,t-1}\end{pmatrix}\begin{pmatrix}\varepsilon_{1,t-1} & \varepsilon_{2,t-1}\end{pmatrix}\right] = \operatorname{vech}\begin{pmatrix}\varepsilon_{1,t-1}^2 & \varepsilon_{1,t-1}\varepsilon_{2,t-1}\\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} & \varepsilon_{2,t-1}^2\end{pmatrix} = \begin{pmatrix}\varepsilon_{1,t-1}^2\\ \varepsilon_{1,t-1}\varepsilon_{2,t-1}\\ \varepsilon_{2,t-1}^2\end{pmatrix} \tag{22.30}$$

记 $\operatorname{vech}(\boldsymbol{H}_t)\equiv\boldsymbol{h}_t$,借助 vech 算子,Bollerslev,Engle and Wooldridge (1988)提出以下一般的多维 GARCH(p,q)模型,简记 MGARCH(p,q):

$$\boldsymbol{h}_t = \boldsymbol{\gamma} + \sum_{i=1}^{p}\boldsymbol{A}_i\operatorname{vech}(\boldsymbol{\varepsilon}_{t-I}\boldsymbol{\varepsilon}_{t-I}') + \sum_{j=1}^{q}\boldsymbol{B}_j\boldsymbol{h}_{t-j} \tag{22.31}$$

其中,$\boldsymbol{\gamma}$ 为待估常数向量,$\boldsymbol{A}_i$ 与 $\boldsymbol{B}_j$ 分别为待估系数矩阵。方程(22.31)也称为“General Vech GARCH”模型。回到双变量的例子,并令 $p=1$,$q=1$,则方程(22.31)可以写为:

$$\begin{pmatrix}\sigma_{11,t}^2\\ \sigma_{21,t}^2\\ \sigma_{22,t}^2\end{pmatrix} = \begin{pmatrix}\gamma_1\\ \gamma_2\\ \gamma_3\end{pmatrix} + \begin{pmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix}\begin{pmatrix}\varepsilon_{1,t-1}^2\\ \varepsilon_{1,t-1}\varepsilon_{2,t-1}\\ \varepsilon_{2,t-1}^2\end{pmatrix} + \begin{pmatrix}b_{11} & b_{12} & b_{13}\\ b_{21} & b_{22} & b_{23}\\ b_{31} & b_{32} & b_{33}\end{pmatrix}\begin{pmatrix}\sigma_{11,t-1}^2\\ \sigma_{21,t-1}^2\\ \sigma_{22,t-1}^2\end{pmatrix} \tag{22.32}$$

根据方程(22.32)可知:

$$\sigma_{11,t}^2 = \gamma_1 + a_{11}\varepsilon_{1,t-1}^2 + a_{12}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + a_{13}\varepsilon_{2,t-1}^2 + b_{11}\sigma_{11,t-1}^2 + b_{12}\sigma_{21,t-1}^2 + b_{13}\sigma_{22,t-1}^2 \tag{22.33}$$

对于 $\sigma_{21,t}^2$ 与 $\sigma_{22,t}^2$ 也有类似的表达式。显然,即使对于最简单的双变量 MGARCH(1,1)模型(22.32),也有很多待估参数。因此,MGARCH(p,q)模型可能过于灵活,不便拟合数据。为此,Bollerslev,Engle and Wooldridge(1988)提出限制所有 $\boldsymbol{A}_i$ 与 $\boldsymbol{B}_j$ 矩阵为对角矩阵,称为“对角半向量化 GARCH”模型(Diagonal Vech GARCH),简记 DVECH。此时,方程(22.32)简化为

$$\begin{pmatrix}\sigma_{11,t}^2\\ \sigma_{21,t}^2\\ \sigma_{22,t}^2\end{pmatrix} = \begin{pmatrix}\gamma_1\\ \gamma_2\\ \gamma_3\end{pmatrix} + \begin{pmatrix}a_{11} & 0 & 0\\ 0 & a_{22} & 0\\ 0 & 0 & a_{33}\end{pmatrix}\begin{pmatrix}\varepsilon_{1,t-1}^2\\ \varepsilon_{1,t-1}\varepsilon_{2,t-1}\\ \varepsilon_{2,t-1}^2\end{pmatrix} + \begin{pmatrix}b_{11} & 0 & 0\\ 0 & b_{22} & 0\\ 0 & 0 & b_{33}\end{pmatrix}\begin{pmatrix}\sigma_{11,t-1}^2\\ \sigma_{21,t-1}^2\\ \sigma_{22,t-1}^2\end{pmatrix} \tag{22.34}$$

方程(22.33)也相应地简化为

$$\sigma_{11,t}^2 = \gamma_1 + a_{11}\varepsilon_{1,t-1}^2 + b_{11}\sigma_{11,t-1}^2 \tag{22.35}$$

估计 DVECH 模型的困难之处在于,在任何时期 t,都需要保证条件协方差矩阵 $\boldsymbol{H}_t$ 为正定矩阵;且待估参数依然较多。为此,学者提出了一类“条件相关模型”(conditional correlation model)①。由于其构建方法,条件相关模型可自然地保证条件协方差矩阵 $\boldsymbol{H}_t$ 为正定矩阵,而且可以减少待估参数。$\boldsymbol{H}_t$ 的(i,j)元素,即变量 i 与变量 j 在时期 t 的条件协方差 $\sigma_{ij,t}$ 可写为:

$$\sigma_{ij,t} = \rho_{ij,t}\sqrt{\sigma_{ii,t}\sigma_{jj,t}} \tag{22.36}$$

其中,$\rho_{ij,t}$ 为变量 i 与变量 j 在时期 t 的条件相关系数,而 $\sigma_{ii,t}$,$\sigma_{jj,t}$ 分别为变量 i 与变量 j 在时期 t 的条件方差。条件相关模型让条件方差 $\sigma_{ii,t}$($\forall\, i=1,\cdots,m$)(假设共有 m 个变量)服从一维的 GARCH 过程②,而集中刻画条件相关系数 $\rho_{ij,t}$ 的行为。如果假设条件相关系数为常数,即 $\rho_{ij,t}\equiv$

① 其他多维 GARCH 模型包括 BEKK 模型(Baba,Engle,Kraft and Kroner,1990;Engle and Kroner,1995),但尚无法在 Stata 12 中实现。

② 可在此一维的条件方差方程中加入额外的解释变量。在 Stata 命令 mgarch 中,可通过选择项 het(varlist)来指定这些解释变量。

ρ_{ij},不随时间而变,则称为“常条件相关模型”(Constant Conditional Correlation,简记 CCC)(Bollerslev,1990)。但常条件相关的假设较强,在实践中未必满足。于是,Engle(2002)与 Tse and Tsui(2002)分别提出了“动态条件相关模型”(Dynamic Conditional Correlation,简记 DCC)与“可变条件相关模型”(Varying Conditional Correlation,简记 VCC)。

将方程(22.36)写为矩阵形式可得:

$$\boldsymbol{H}_t = \boldsymbol{D}_t^{1/2}\boldsymbol{R}_t\boldsymbol{D}_t^{1/2} \tag{22.37}$$

其中,$\boldsymbol{D}_t$ 为条件方差对角矩阵,而 $\boldsymbol{R}_t$ 为条件相关系数矩阵:

$$\boldsymbol{D}_t = \begin{pmatrix} \sigma_{11,t} & & \\ & \ddots & \\ & & \sigma_{mm,t} \end{pmatrix}, \quad \boldsymbol{R}_t = \begin{pmatrix} 1 & \rho_{12,t} & \cdots & \rho_{1m,t} \\ \rho_{12,t} & 1 & \cdots & \rho_{2m,t} \\ \vdots & \vdots & & \vdots \\ \rho_{1m,t} & \rho_{2m,t} & \cdots & 1 \end{pmatrix} \tag{22.38}$$

显然,如果 $\boldsymbol{R}_t$ 为常数矩阵,则为 CCC 模型。Engle(2002)的 DCC 模型假定,$\rho_{ij,t}$ 由标准化扰动项的几何加权平均(geometrically weighted average of standardized residuals)来决定:

$$\rho_{ij,t} = \frac{\sum_{s=1}^{t-1}\lambda^s \tilde{\varepsilon}_{i,t-s}\tilde{\varepsilon}_{j,t-s}}{\sqrt{\left(\sum_{s=1}^{t-1}\lambda^s \tilde{\varepsilon}_{i,t-s}^2\right)\left(\sum_{s=1}^{t-1}\lambda^s \tilde{\varepsilon}_{i,t-s}^2\right)}} \tag{22.39}$$

其中,$\tilde{\varepsilon}_{i,t-s}$ 为标准化扰动项(方差标准化为 1),而 λ^s 为几何权重(离时期 t 越远,则权重依几何级数减少)。写为矩阵形式,$\boldsymbol{R}_t$ 的动态过程由以下两个方程来决定:

$$\begin{aligned} \boldsymbol{R}_t &= \mathrm{diag}(\boldsymbol{Q}_t)^{-1/2}\boldsymbol{Q}_t\mathrm{diag}(\boldsymbol{Q}_t)^{-1/2}, \\ \boldsymbol{Q}_t &= (1-\lambda_1-\lambda_2)\boldsymbol{R} + \lambda_1\tilde{\boldsymbol{\varepsilon}}_{t-1}\tilde{\boldsymbol{\varepsilon}}_{t-1}' + \lambda_2\boldsymbol{Q}_{t-1} \end{aligned} \tag{22.40}$$

其中,$\tilde{\boldsymbol{\varepsilon}}_t$ 为标准化的扰动项,即 $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{D}_t^{-1/2}\boldsymbol{\varepsilon}_t$;参数 λ_1 与 λ_2 皆为非负,而且 $0 \leqslant \lambda_1 + \lambda_2 < 1$。显然,$\lambda_1$ 与 λ_2 决定了的动态过程。特别地,如果 $\lambda_1 = \lambda_2 = 0$,则 DCC 模型简化为 CCC 模型。因此,通过检验联合假设“$\lambda_1 = \lambda_2 = 0$”,可在 DCC 与 CCC 模型之间进行选择。由于 $\boldsymbol{R}_t$ 中的元素不一定介于 -1 与 1 之间,故也称为“条件准相关系数矩阵”(matrix of conditional quasicorrelations)。

Tse and Tsui (2002)的 VCC 模型则假设 $\boldsymbol{R}_t$ 服从类似于 ARMA 的过程:

$$\boldsymbol{R}_t = (1-\lambda_1-\lambda_2)\boldsymbol{R} + \lambda_1\boldsymbol{\Psi}_{t-1} + \lambda_2\boldsymbol{R}_{t-1} \tag{22.41}$$

其中,$\boldsymbol{R}$ 为 $\boldsymbol{R}_t$ 的均值;$\boldsymbol{\Psi}_{t-1}$ 为对标准化扰动项 $\tilde{\boldsymbol{\varepsilon}}_t$ 的相关系数矩阵的滚动估计量(rolling estimator),即在方程(22.39)中,令 $\lambda = 1$(简单平均,而非加权平均);参数 λ_1 与 λ_2 皆非负,且 $0 \leqslant \lambda_1 + \lambda_2 < 1$。

对于多维 GARCH 模型,通常假设扰动项 $\boldsymbol{\varepsilon}_t$ 服从多维正态分布,然后进行 MLE 估计。如果扰动项不服从正态分布,则为 QMLE 估计,在大样本下依然是一致且渐近正态的。QMLE 与 MLE 的估计系数相同,只是对协方差矩阵(标准误)的估计不同。也可以假设扰动项 $\boldsymbol{\varepsilon}_t$ 来自多维 t 分布,然后进行 MLE 估计。如果扰动项 $\boldsymbol{\varepsilon}_t$ 确实服从多维 t 分布,则基于多维 t 分布假设的 MLE 估计是一致且有效的;而基于多维正态分布假设的 MLE 依然是一致的,但并非最有效。反之,如果扰动项 $\boldsymbol{\varepsilon}_t$ 并非来自多维 t 分布或多维正态分布,则基于多维 t 分布假设的 MLE 估计是不一致的,而基于多维正态分布假设的 QMLE 估计依然是一致的(尽管并非最有效)。

在 Stata 12 中,估计多维 GARCH 模型的命令句式为

```
mgarch model eq … eq,arch(#) garch(#) constraints(#) het(varlist) dist
(t) robust noconstant
```

其中,"model"可以是 dvech,ccc,dcc 或 vcc,分别表示 DVECH,CCC,DCC 与 VCC 模型。选择项"arch(#)"表示 ARCH 项,"garch(#)"表示 GARCH 项;选择项"constraints(#)"用于定义对参数的约束;选择项"het(varlist)"用来指定条件方差方程中的额外解释变量;选择项"dist(t)"假设扰动项服从多维 t 分布,而默认假设为多维正态分布;选择项"robust"表示使用稳健标准误(进行 QMLE 估计);选择项"noconstant"表示忽略常数项;而每个"eq"表示一个方程,其一般格式为(depvars = indepvars,noconstant)。

下面以 Stata 提供的数据集 irates4. dta 为例。该数据集包含美国 6 月期国库券二级市场利率(tbill)与 AAA 级公司债券收益率(bond)。这两个变量的一阶差分,进行 VAR(1)建模,并带 ARCH(1)扰动项[①]。首先估计 DVECH 模型。

```
.use irates4.dta,clear
.mgarch dvech(D.bond D.tbill = LD.bond LD.tbill),arch(1) nolog
```

Diagonal vech MGARCH model

Sample: 3 - 2456 Number of obs = 2454
Distribution: Gaussian Wald chi2(4) = 1183.52
Log likelihood = 4221.658 Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D.bond						
bond						
LD.	.2967674	.0247149	12.01	0.000	.2483271	.3452077
tbill						
LD.	.0947949	.0098683	9.61	0.000	.0754533	.1141364
_cons	.0003991	.00143	0.28	0.780	-.0024036	.0032019
D.tbill						
bond						
LD.	.0108373	.0301501	0.36	0.719	-.0482558	.0699304
tbill						
LD.	.4344747	.0176497	24.62	0.000	.3998819	.4690675
_cons	.0011611	.0021033	0.55	0.581	-.0029612	.0052835
Sigma0						
1_1	.004894	.0002006	24.40	0.000	.0045008	.0052871
2_1	.0040986	.0002396	17.10	0.000	.0036289	.0045683
2_2	.0115149	.0005227	22.03	0.000	.0104904	.0125395
L.ARCH						
1_1	.4514942	.0456835	9.88	0.000	.3619562	.5410323
2_1	.2518879	.036736	6.86	0.000	.1798866	.3238893
2_2	.843368	.0608055	13.87	0.000	.7241914	.9625446

上表下部的 Sigma0 表示方程(22.31)中的常数向量 γ,而 L.ARCH 表示 ARCH(1)项的估计结果,1_1,2_1 与 2_2 分别表示 $\varepsilon_{1,t-1}^2$,$\varepsilon_{1,t-1}\varepsilon_{2,t-1}$与 $\varepsilon_{2,t-1}^2$的相应系数,均显著为正。上表显示,在以 D. tbill 为被解释变量的方程中,解释变量 LD. bond(即 bond 的差分滞后项)并不显著。而且,两个方程的常数项均不显著。下面,为了去掉这些不显著项,在命令中将两个方程分开写。

① 在此例中,如果加上 GARCH(1)项则很难收敛,这是估计多维 GARCH 模型经常碰到的问题。

```
. mgarch dvech (D.bond = LD.bond LD.tbill, noconstant) (D.tbill =
LD.tbill,noconstant),arch(1) nolog
```

Diagonal vech MGARCH model

Sample: 3 - 2456 Number of obs = 2454
Distribution: Gaussian Wald chi2(3) = 1197.76
Log likelihood = 4221.433 Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D.bond						
bond						
LD.	.2941649	.0234734	12.53	0.000	.2481579	.3401718
tbill						
LD.	.0953158	.0098077	9.72	0.000	.076093	.1145386
D.tbill						
tbill						
LD.	.4385945	.0136672	32.09	0.000	.4118072	.4653817
Sigma0						
1_1	.0048922	.0002005	24.40	0.000	.0044993	.0052851
2_1	.0040949	.0002394	17.10	0.000	.0036256	.0045641
2_2	.0115043	.0005184	22.19	0.000	.0104883	.0125203
L.ARCH						
1_1	.4519233	.045671	9.90	0.000	.3624099	.5414368
2_1	.2515474	.0366701	6.86	0.000	.1796752	.3234195
2_2	.8437212	.0600839	14.04	0.000	.7259589	.9614836

上表显示,所有解释变量的系数都很显著了。下面,使用 CCC 模型来估计带 GARCH(1,1)扰动项的原模型。

```
. mgarch ccc (D.bond D.tbill = LD.bond LD.tbill),arch(1) nolog
```

Constant conditional correlation MGARCH model

Sample: 1 - 2456 Number of obs = 2454
Distribution: Gaussian Wald chi2(4) = 1516.13
Log likelihood = 4257.064 Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D.bond						
bond						
LD.	.2683797	.0227155	11.81	0.000	.2238581	.3129012
tbill						
LD.	.0822348	.0090725	9.06	0.000	.064453	.1000167
_cons	.0005556	.0014135	0.39	0.694	-.0022147	.003326
ARCH_D.bond						
arch						
L1.	.5233444	.049989	10.47	0.000	.4253678	.621321
_cons	.0046334	.000191	24.26	0.000	.004259	.0050077
D.tbill						
bond						
LD.	-.0118455	.0272737	-0.43	0.664	-.065301	.04161
tbill						
LD.	.4393633	.0155774	28.21	0.000	.408832	.4698945
_cons	.0009441	.0020204	0.47	0.640	-.0030159	.0049041
ARCH_D.tbill						
arch						
L1.	.951192	.0663333	14.34	0.000	.8211811	1.081203
_cons	.0106312	.0004776	22.26	0.000	.0096952	.0115673
Correlation						
D.bond						
D.tbill	.4575932	.0161694	28.30	0.000	.4259017	.4892847

上表显示,ARCH 项显著为正。上表最后一行显示,国库券收益率差分与债券收益率差分的相关系数为 0.46,且在 1%水平上显著。这意味着,国库券收益率上涨(下跌)时,企业债券收益率也倾向于上涨(下跌)。但常相关模型假设二者的相关系数为常数,未必现实,故下面转向 VCC 模型。

```
.mgarch dcc(D.bond D.tbill = LD.bond LD.tbill),arch(1) nolog
```

Dynamic conditional correlation MGARCH model

Sample: 1 - 2456 Number of obs = 2454
Distribution: Gaussian Wald chi2(4) = 1563.32
Log likelihood = 4267.592 Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D.bond						
bond						
LD.	.2663008	.0226708	11.75	0.000	.2218668	.3107348
tbill						
LD.	.082495	.0090404	9.13	0.000	.0647762	.1002139
_cons	.0003347	.0014168	0.24	0.813	-.0024422	.0031116
ARCH_D.bond						
arch						
L1.	.5115416	.0489595	10.45	0.000	.4155827	.6075005
_cons	.0045801	.0001892	24.20	0.000	.0042092	.0049511
D.tbill						
bond						
LD.	-.0107307	.0269353	-0.40	0.690	-.0635229	.0420614
tbill						
LD.	.4385706	.0153425	28.59	0.000	.4084999	.4686413
_cons	.0009644	.0020006	0.48	0.630	-.0029568	.0048856
ARCH_D.tbill						
arch						
L1.	.9250511	.0640242	14.45	0.000	.7995659	1.050536
_cons	.0103025	.0004579	22.50	0.000	.009405	.0112
Correlation						
D.bond						
D.tbill	.4057807	.0237458	17.09	0.000	.3592397	.4523217
Adjustment						
lambda1	.0040023	.0014576	2.75	0.006	.0011456	.0068591
lambda2	.9795958	.0052797	185.54	0.000	.9692478	.9899438

上表显示，国库券收益率差分与企业债券收益率差分的条件准相关系数为 0.41，且在 1% 水平上显著。上表最后两行显示，λ_1 与 λ_2 均显著为正，故应使用 DCC 而非 CCC 模型。下面，估计 VCC 模型。

```
.mgarch vcc(D.bond D.tbill = LD.bond LD.tbill),arch(1) nolog
```

Varying conditional correlation MGARCH model

Sample: 1 - 2456　　Number of obs = 2454
Distribution: Gaussian　　Wald chi2(4) = 1629.49
Log likelihood = 4273.118　　Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D.bond						
bond						
LD.	.2633396	.0227876	11.56	0.000	.2186767	.3080024
tbill						
LD.	.0841621	.0095937	8.77	0.000	.0653589	.1029654
_cons	.00055	.001428	0.39	0.700	-.0022488	.0033488
ARCH_D.bond						
arch						
L1.	.5212864	.0501374	10.40	0.000	.4230189	.619554
_cons	.0047296	.0001995	23.71	0.000	.0043386	.0051206
D.tbill						
bond						
LD.	-.0306537	.0233759	-1.31	0.190	-.0764696	.0151622
tbill						
LD.	.447364	.0145085	30.83	0.000	.4189278	.4758002
_cons	.0012041	.0019613	0.61	0.539	-.0026401	.0050482
ARCH_D.tbill						
arch						
L1.	.9315652	.0637373	14.62	0.000	.8066423	1.056488
_cons	.010033	.0004448	22.55	0.000	.0091611	.0109049
Correlation						
D.bond						
D.tbill	.5708136	.1365947	4.18	0.000	.303093	.8385342
Adjustment						
lambda1	.0030634	.0007849	3.90	0.000	.001525	.0046019
lambda2	.9959393	.001035	962.31	0.000	.9939108	.9979678

上表显示，国库券收益率差分与企业债券收益率差分的条件准相关系数为 0.57，且在 1% 水平上显著。上表最后两行显示，λ_1 与 λ_2 均显著为正，故应使用 VCC 而非 CCC 模型。

习　题

22.1　考虑 ARCH(p)模型，$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2$。为了保证 σ_t^2 为正，以及 $\{\varepsilon_t\}$ 为平稳过程，需要对参数 $\{\alpha_0, \alpha_1, \cdots, \alpha_p\}$ 进行哪些限制？

22.2　数据集 dow1. dta 包含了 1953—1990 年美国道琼斯股指的收盘价。计算道琼斯股指的日收益率，画其时间趋势图，并参照本章实例进行 ARCH/GARCH 估计。

第 23 章　似不相关回归

23.1　单一方程估计与系统估计

到目前为止，我们仅考虑对单一方程的估计，但有时候会出现多个方程的情形。如果多个方程之间有某种联系，那么将这些方程同时进行联合估计有可能提高估计的效率，这被称为“系统估计”（system estimation）。

有时多个方程是从同一个最大化问题推导而来（比如，从企业的利润最大化问题导出对资本与劳动力的需求），故在理论上存在“跨方程的参数约束”（cross-equation restrictions）。多方程联合估计为检验这些跨方程约束提供了可能。也可以在加上这些约束条件后再进行系统估计。

对多方程系统进行联合估计的缺点是，如果某个方程的误差较大，则系统估计会将这一方程的误差带入其他方程中，进而污染（contaminate）整个方程系统。在某种意义上，选择单一方程估计或系统估计，也是在“有效性”与“稳健性”之间的权衡。

多方程系统主要分为两类。一类为“联立方程组”（simultaneous equations），即不同方程之间的变量存在内在的联系，一个方程的解释变量是另一方程的被解释变量。另一类为“似不相关回归”（Seemingly Unrelated Regression Estimation，简记 SUR 或 SURE），即各方程的变量之间没有内在联系，但各方程的扰动项之间存在相关性。本章介绍似不相关回归，下一章介绍联立方程模型。

例（似不相关回归）　以研一学生的计量成绩与英语成绩作为两个被解释变量。这两个方程所包含的解释变量可以不同，比如，第一个方程可以包括虚拟变量“是否学过本科计量学”，而第二个方程可以包括“考研英语成绩”。这两个方程的变量之间貌似没有联系，但由于同一学生的不可观测因素同时对计量成绩与英语成绩造成影响，故两个方程的扰动项应该是相关的。如果将这两个方程同时进行联合估计，则可以提高估计效率。

23.2　似不相关回归的假定

似不相关回归模型的设定如下。假设共有 n 个方程（n 个被解释变量），每个方程共有 T 个观测值[①]，$T>n$。在第 i 个方程中，共有 K_i 个解释变量。第 i 个方程可以写为

$$\underbrace{\boldsymbol{y}_i}_{T\times 1}=\underbrace{\boldsymbol{X}_i}_{T\times K_i}\underbrace{\boldsymbol{\beta}_i}_{K_i\times 1}+\underbrace{\boldsymbol{\varepsilon}_i}_{T\times 1}\quad(i=1,2,\cdots,n)\tag{23.1}$$

将所有的方程叠放在一起可得

① 此处的 T 不一定指的是时间。但为了叙述方便，我们认为它是时间。

$$\boldsymbol{y} \equiv \underbrace{\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n \end{pmatrix}}_{nT\times \sum_{i=1}^{n} K_i} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_n \end{pmatrix}}_{\sum_{i=1}^{n} K_i \times 1} + \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}}_{nT\times 1} \equiv \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{23.2}$$

可以想象,在计算机中,也把数据排成这个样子(需要加上许多零)。考察"大"扰动项 $\boldsymbol{\varepsilon}$ 之协方差矩阵

$$\boldsymbol{\Omega} \equiv \mathrm{Var}\begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} = \mathrm{E}\begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}(\boldsymbol{\varepsilon}_1' \ \boldsymbol{\varepsilon}_2' \cdots \boldsymbol{\varepsilon}_n') = \mathrm{E}\begin{pmatrix} \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_n' \\ \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_n' \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n' \end{pmatrix}_{nT\times nT} \tag{23.3}$$

假设同一方程不同期的扰动项不存在自相关,且方差也相同,记第 i 个方程的方差为 σ_{ii}。则协方差阵 $\boldsymbol{\Omega}$ 中主对角线上的第(i,i)个矩阵为

$$\mathrm{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i') = \sigma_{ii}\boldsymbol{I}_T \tag{23.4}$$

假设不同方程的扰动项之间存在同期相关①,即

$$\mathrm{E}(\varepsilon_{it}\varepsilon_{js}) = \begin{cases} \sigma_{ij}, t = s \\ 0, t \neq s \end{cases} \tag{23.5}$$

则协方差阵 $\boldsymbol{\Omega}$ 中的第(i,j)个矩阵$(i\neq j)$为

$$\mathrm{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j') = \sigma_{ij}\boldsymbol{I}_T \tag{23.6}$$

综合以上结果可知

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_{11}\boldsymbol{I}_T & \sigma_{12}\boldsymbol{I}_T & \cdots & \sigma_{1n}\boldsymbol{I}_T \\ \sigma_{21}\boldsymbol{I}_T & \sigma_{22}\boldsymbol{I}_T & \cdots & \sigma_{2n}\boldsymbol{I}_T \\ \vdots & \vdots & & \vdots \\ \sigma_{n1}\boldsymbol{I}_T & \sigma_{n2}\boldsymbol{I}_T & \cdots & \sigma_{nn}\boldsymbol{I}_T \end{pmatrix} \tag{23.7}$$

由于 $\boldsymbol{\Omega}$ 中的每个小块矩阵都有共同的因子 $\boldsymbol{I}_T$,我们自然想把 $\boldsymbol{I}_T$ 从右边提取出来。这可以通过矩阵的"克罗内克尔乘积"(Kronecker product)来实现。

定义 对于任意两个矩阵$\boldsymbol{A}_{m\times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$与$\boldsymbol{B}_{p\times q}$(矩阵 $\boldsymbol{A},\boldsymbol{B}$ 的维度可以完全不同),

克罗内克尔乘积为 $\boldsymbol{A}\otimes\boldsymbol{B} \equiv \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix}_{mp\times nq}$。

容易看出,对于任意矩阵 $\boldsymbol{A},\boldsymbol{B}$,其克罗内克尔乘积 $\boldsymbol{A}\otimes\boldsymbol{B}$ 总是有定义的。可以证明(参见习题),克罗内克尔乘积具有以下性质:

(1) $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C})\otimes(\boldsymbol{B}\boldsymbol{D})$;

(2) $(\boldsymbol{A}\otimes\boldsymbol{B})' = \boldsymbol{A}'\otimes\boldsymbol{B}'$;

---

① 如果是横截面数据,则指的是同一个体在不同回归方程中的对应扰动项之间存在相关性。

(3) $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1}=\boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$

使用克罗内克尔乘积,可以将扰动项 $\boldsymbol{\varepsilon}$ 的协方差矩阵简化为

$$\boldsymbol{\Omega}=\begin{pmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1n}\\ \sigma_{21}&\sigma_{22}&\cdots&\sigma_{2n}\\ \vdots&\vdots&&\vdots\\ \sigma_{n1}&\sigma_{n2}&\cdots&\sigma_{nn}\end{pmatrix}\otimes\boldsymbol{I}_T\equiv\boldsymbol{\Sigma}\otimes\boldsymbol{I}_T \tag{23.8}$$

其中,$\boldsymbol{\Sigma}\equiv\begin{pmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1n}\\ \sigma_{21}&\sigma_{22}&\cdots&\sigma_{2n}\\ \vdots&\vdots&&\vdots\\ \sigma_{n1}&\sigma_{n2}&\cdots&\sigma_{nn}\end{pmatrix}$ 为同期协方差矩阵。根据克罗内克尔乘积的性质,$\boldsymbol{\Omega}$ 的逆矩阵可以写为

$$\boldsymbol{\Omega}^{-1}=\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T \tag{23.9}$$

23.3 SUR 的 FGLS 估计

由于 $\boldsymbol{\Omega}$ 不是单位矩阵,故用 OLS 估计这个多方程系统 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 不是最有效率的。假设 $\boldsymbol{\Omega}$ 已知,则 GLS 是最有效率的估计方法:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}}=(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}=[\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{y} \tag{23.10}$$

一般来说,此 GLS 估计量与单一方程 OLS 估计量不同。但如果出现以下两种情形之一,则 GLS 与单一方程 OLS 的结果完全相同,使用 GLS 并不会提高效率。

(1) 各方程的扰动项互不相关。在似不相关回归模型中,各方程间唯一的联系就是扰动项之间的相关性。如果扰动项互不相关,则 $\boldsymbol{\Omega}$ 是单位矩阵,那么系统估计与单一方程估计并无区别(参见习题)。

(2) 每个方程包含的解释变量完全相同(证明参见附录)。比如,向量自回归模型(VAR)中的每个方程包含完全相同的解释变量,故使用单一方程 OLS 估计 VAR 就够了。

除了以上两种特殊情形外,通常来说,各方程的扰动项之间的相关性越大,则 GLS 所能带来的效率改进就越大;各方程的数据矩阵 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$ 之间的相关性越小,则 GLS 所能带来的效率改进也越大(如果数据矩阵完全相同,则 GLS 还原为单一方程 OLS)。

然而,在现实中 $\boldsymbol{\Omega}$ 一般是未知的,故首先需要估计 $\hat{\boldsymbol{\Omega}}$,然后进行 FGLS 估计。由于对每个方程分别进行 OLS 回归也是一致的,故可以使用单一方程 OLS 的残差来一致地估计 σ_{ij}。假设第 i 个方程的 OLS 残差向量为 $\boldsymbol{e}_i$,则 σ_{ij} 的一致估计量为

$$\hat{\sigma}_{ij}=\frac{1}{T}\boldsymbol{e}_i'\boldsymbol{e}_j=\frac{1}{T}\sum_{t=1}^{T}e_{it}e_{jt} \tag{23.11}$$

因此,$\hat{\boldsymbol{\Omega}}=\begin{pmatrix}\hat{\sigma}_{11}&\hat{\sigma}_{12}&\cdots&\hat{\sigma}_{1n}\\ \hat{\sigma}_{21}&\hat{\sigma}_{22}&\cdots&\hat{\sigma}_{2n}\\ \vdots&\vdots&&\vdots\\ \hat{\sigma}_{n1}&\hat{\sigma}_{n2}&\cdots&\hat{\sigma}_{nn}\end{pmatrix}\otimes\boldsymbol{I}_T$。将 $\hat{\boldsymbol{\Omega}}$ 代入方程(23.10)可得

$$\hat{\boldsymbol{\beta}}_{\mathrm{SUR}}=(\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{y} \tag{23.12}$$

这就是“似不相关估计量”(Zellner,1962),记为 $\hat{\boldsymbol{\beta}}_{\mathrm{SUR}}$。使用 FGLS 后得到新的残差,可以再一次计算$\hat{\hat{\boldsymbol{\Omega}}}$,不断迭代直至系数估计值 $\hat{\boldsymbol{\beta}}_{\mathrm{SUR}}$ 收敛为止。

23.4 SUR 的假设检验

在对多方程系统进行 SUR 估计后,对线性假设“$H_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$”的检验可以照常进行。由于 $\boldsymbol{\beta}$ 包含了所有方程的参数,故可以检验跨方程的参数约束。如果接受“$H_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$”,则可把“$\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$”作为约束条件,进行有约束的 FGLS 估计。

需要指出的是,即使各方程的解释变量完全相同①,有时也使用 SUR 而不使用单一方程 OLS,以便检验跨方程的参数约束。如果存在跨方程的参数约束,则即使各方程的解释变量完全相同,SUR 估计与单一方程 OLS 也不再等价。

SUR 模型的基本假设是,各方程的扰动项之间存在同期相关。为此,需要检验原假设“H_0:各方程的扰动项无同期相关”,即“$H_0:\boldsymbol{\Sigma}$ 为对角矩阵”。Breusch and Pagan(1980)建议使用以下 LM 统计量:

$$\lambda_{\mathrm{LM}}=T\sum_{i=2}^{n}\sum_{j=1}^{i-1}r_{ij}^2\xrightarrow{d}\chi^2(n(n-1)/2) \tag{23.13}$$

其中,$r_{ij}=\dfrac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$为根据残差计算的扰动项 $\boldsymbol{\varepsilon}_i$ 与 $\boldsymbol{\varepsilon}_j$ 之间的同期相关系数,而 $\sum_{i=2}^{n}\sum_{j=1}^{i-1}r_{ij}^2$ 为同期相关系数矩阵$\begin{pmatrix}r_{11}&r_{12}&\cdots&r_{1n}\\r_{21}&r_{22}&\cdots&r_{2n}\\\vdots&\vdots&&\vdots\\r_{n1}&r_{n2}&\cdots&r_{nn}\end{pmatrix}$主对角线以下各项之平方和(该矩阵为对称矩阵)。

23.5 似不相关回归的 Stata 命令及实例

似不相关回归的 Stata 命令基本格式为

```
sureg (depvar1 varlist1)(depvar2 varlist2) ...(depvarn
varlistn),isure corr
```

其中,选择项“isure”表示迭代至收敛(默认值为不迭代),选择项“corr”表示汇报对无同期相关的检验。

如果要检验联合假设“第 1 个方程中变量 x 的系数与第 2 个方程中变量 z,以及第 3 个方程中变量 w 的系数相等”,可使用如下命令:

---

① 这种情况在经济学中常常出现。比如,对一组商品的需求取决于这组商品的价格及收入;对一组资产的需求份额(portfolio share of assets)取决于这组资产的回报率及总财富。当需求份额必须加总为 1 时,可以选择将其中一个方程去掉,否则扰动项的协方差矩阵将为不可逆的退化矩阵。

```
test ([depvar1]x = [depvar2]z)([depvar1]x = [depvar3]w)
```

如果要在满足以上两个约束条件的情况下进行 SUR 估计，则可使用以下命令：

```
constraint 1 [depvar1]x = [depvar2]z  (定义第 1 个约束条件)
constraint 2 [depvar1]x = [depvar2]w  (定义第 2 个约束条件)
sureg (depvar1 varlist1) (depvar2 varlist2) ... (depvarn varlistn),c(1 2)
```

下面以来自美国 High School & Beyond Study 的数据集 hsb2.dta 为例。该数据集包含 200 名高中生的以下变量，read(阅读成绩)，math(数学成绩)，science(自然科学成绩)，socst(社会科学成绩)，female(是否女性)，schtyp(是否私立学校)，以及 ses(社会经济地位，social economic status，low = 1，middle = 2，high = 3)。考虑以下两个回归方程：

$$\text{read}_i = \alpha_0 + \alpha_1\ \text{female}_i + \alpha_2\ \text{ses}_i + \alpha_3\ \text{schtyp}_i + \alpha_4\ \text{socst}_i + u_i \tag{23.14}$$

$$\text{math}_i = \beta_0 + \beta_1\ \text{female}_i + \beta_2\ \text{ses}_i + \beta_3\ \text{schtyp}_i + \beta_4\ \text{science}_i + v_i \tag{23.15}$$

作为参照系，首先对以上两个方程进行单一方程 OLS 估计：

```
. use hsb2.dta,clear
. reg3 (read female ses schtyp socst) (math female ses schtyp science),ols
. estimates store OLS
```

```
Multivariate regression
------------------------------------------------------------------------------
Equation              Obs   Parms        RMSE    "R-sq"       F-Stat       P
------------------------------------------------------------------------------
read                  200       4    8.020526    0.4004        32.55  0.0000
math                  200       4    7.250452    0.4131        34.31  0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
read         |
      female |  -1.521471   1.154157    -1.32   0.188    -3.790619     .747678
         ses |   1.186993   .8473698     1.40   0.162    -.4789918    2.852977
      schtyp |   .5071815    1.56462     0.32   0.746    -2.568964    3.583327
       socst |   .5689789   .0565024    10.07   0.000     .4578916    .6800662
       _cons |   20.21426   3.316836     6.09   0.000     13.69314    26.73538
-------------+----------------------------------------------------------------
math         |
      female |   1.131112   1.043281     1.08   0.279    -.9200471    3.182272
         ses |   1.324103   .7491005     1.77   0.078    -.1486777    2.796883
      schtyp |   1.126373   1.413143     0.80   0.426    -1.651959    3.904706
     science |    .573999   .0544043    10.55   0.000     .4670367    .6809614
       _cons |   18.23907    3.28543     5.55   0.000      11.7797    24.69844
------------------------------------------------------------------------------
```

其中，命令 reg3 的默认估计法为三阶段最小二乘法(3SLS)[①]，但加上选择项“ols”后则进行单一方程 OLS 估计。由于被解释变量 read 与 math 为同一学生的阅读与数学成绩，故这两个方程的扰动项在理论上很可能存在相关性。为此，进行 SUR 估计：

```
. sureg (read female ses schtyp socst) (math female ses schtyp
```

① 参见第 24 章。

science),corr①

其中,选择项"corr"表示,汇报对各方程扰动项之间"无同期相关"的检验结果。

```
. estimates store SUR
```

```
Seemingly unrelated regression
------------------------------------------------------------------------------
Equation              Obs   Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
read                  200       4    7.938329    0.3975     117.51    0.0000
math                  200       4    7.189084    0.4082     117.78    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
read         |
      female |  -1.412142   1.139458    -1.24   0.215    -3.64544     .8211552
         ses |   1.457827   .8351991     1.75   0.081    -.1791329    3.094788
      schtyp |   .5854049   1.544871     0.38   0.705    -2.442486    3.613295
       socst |    .514735   .0548756     9.38   0.000     .4071808    .6222891
       _cons |   22.35003   3.251032     6.87   0.000     15.97812    28.72193
-------------+----------------------------------------------------------------
math         |
      female |   1.000534   1.029996     0.97   0.331    -1.018222    3.01929
         ses |   1.577474   .7388309     2.14   0.033     .1293925    3.025556
      schtyp |   1.181735   1.395346     0.85   0.397    -1.553093    3.916562
     science |   .5045673   .0528379     9.55   0.000     .4010069    .6081277
       _cons |   21.32537   3.215353     6.63   0.000     15.02339    27.62735
------------------------------------------------------------------------------

Correlation matrix of residuals:

             read    math
read    1.0000
math    0.1985  1.0000

Breusch-Pagan test of independence: chi2(1) =     7.880, Pr = 0.0050
```

上表最后一行显示,各方程扰动项之间"无同期相关"的检验 p 值为 0.005,故可以在 1% 的显著性水平上拒绝各方程的扰动项相互独立的原假设。因此,使用 SUR 进行系统估计可以提高估计的效率。下面进行迭代式 SUR 估计:

```
. sureg(read female ses schtyp socst)(math female ses schtyp science),
i nolog
. estimates store SUR_i
```

```
Seemingly unrelated regression, iterated
------------------------------------------------------------------------------
Equation              Obs   Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
read                  200       4    7.965347    0.3934     109.40    0.0000
math                  200       4    7.229906    0.4014     105.94    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
read         |
      female |  -1.350362   1.145814    -1.18   0.239    -3.596118    .8953927
         ses |   1.610872   .8381693     1.92   0.055     -.03191     3.253653
      schtyp |   .6296076   1.553704     0.41   0.685    -2.415596    3.674812
       socst |   .4840826   .0540519     8.96   0.000     .3781427    .5900224
       _cons |   23.55691    3.24013     7.27   0.000     17.20638    29.90745
-------------+----------------------------------------------------------------
math         |
      female |   .9298813   1.039964     0.89   0.371    -1.108411    2.968173
         ses |   1.714568   .7450829     2.30   0.021     .2542327    3.174904
      schtyp |    1.21169   1.409091     0.86   0.390    -1.550077    3.973457
     science |   .4669993   .0522569     8.94   0.000     .3645777    .5694209
       _cons |    22.9953   3.211463     7.16   0.000     16.70095    29.28965
------------------------------------------------------------------------------
```

① 此命令也可用"reg3(read female ses schtyp socst)(math female ses schtyp science),sur"来实现,二者的输出结果相同。

为了便于比较,将以上各种方法的系数估计值及标准误列表:

```
. esttab OLS SUR SUR_i,se mtitle r2 star(* 0.1 ** 0.05 *** 0.01)
```

	(1) OLS	(2) SUR	(3) SUR_i
read			
female	-1.521	-1.412	-1.350
	(1.154)	(1.139)	(1.146)
ses	1.187	1.458*	1.611*
	(0.847)	(0.835)	(0.838)
schtyp	0.507	0.585	0.630
	(1.565)	(1.545)	(1.554)
socst	0.569***	0.515***	0.484***
	(0.0565)	(0.0549)	(0.0541)
_cons	20.21***	22.35***	23.56***
	(3.317)	(3.251)	(3.240)
math			
female	1.131	1.001	0.930
	(1.043)	(1.030)	(1.040)
ses	1.324*	1.577**	1.715**
	(0.749)	(0.739)	(0.745)
schtyp	1.126	1.182	1.212
	(1.413)	(1.395)	(1.409)
science	0.574***	0.505***	0.467***
	(0.0544)	(0.0528)	(0.0523)
_cons	18.24***	21.33***	23.00***
	(3.285)	(3.215)	(3.211)
N	200	200	200
R-sq	0.400	0.398	0.393

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

从上表可知,SUR 与迭代 SUR 的估计结果比较接近,但与 OLS 有明显差别。这是因为,SUR 比单一方程 OLS 更有效率。

由于在两个方程中,变量 ses 的系数估计值比较接近,下面检验这两个系数是否相等:

```
. quietly sureg (read female ses schtyp socst) (math female ses schtyp
science)
. test ([read]ses = [math]ses)
```

```
 ( 1)  [read]ses - [math]ses = 0

           chi2(  1) =    0.01
         Prob > chi2 =    0.9053
```

由于 p 值为 0.91,故可以接受“这两个系数相等”的原假设。下面,我们加上这个跨方程约束条件,再进行 SUR 估计:

```
. constraint 1 [read]ses = [math]ses
. sureg (read female ses schtyp socst) (math female ses schtyp science),
c(1)
```

```
Seemingly unrelated regression
------------------------------------------------------------------------------
Equation             Obs   Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
read                 200       4    7.939578    0.3973     119.44    0.0000
math                 200       4    7.188205    0.4083     119.06    0.0000
------------------------------------------------------------------------------

 ( 1)  [read]ses - [math]ses = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
read         |
      female |  -1.397722   1.132922    -1.23   0.217    -3.618209    .8227652
         ses |   1.526381    .602663     2.53   0.011     .3451827    2.707578
      schtyp |    .571229   1.540347     0.37   0.711    -2.447795    3.590253
       socst |   .5131128   .0530443     9.67   0.000     .4091479    .6170776
       _cons |   22.30275    3.22883     6.91   0.000     15.97436    28.63114
-------------+----------------------------------------------------------------
math         |
      female |   .9939827   1.028492     0.97   0.334    -1.021824    3.009789
         ses |   1.526381    .602663     2.53   0.011     .3451827    2.707578
      schtyp |   1.193675   1.391732     0.86   0.391     -1.53407    3.921419
     science |   .5056923   .0520817     9.71   0.000     .4036141    .6077706
       _cons |   21.36176   3.198553     6.68   0.000     15.09271    27.63081
------------------------------------------------------------------------------
```

23.6 变系数面板数据的 SUR 估计

在第 16 章曾讨论过变系数面板数据模型：

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it} \tag{23.16}$$

其中，$\boldsymbol{\beta}_i$ 为个体 i 的系数(常数)。虽然不同个体拥有自己的系数(含截距项与斜率)，但不同个体的扰动项却可能相关，故应使用 SUR 进行系统估计。

以数据集 mus08cigar. dta 为例(参见第 16 章)。由于该面板数据集为长形(long form)，数据排列形式如第 15 章表 15. 1 所示，故首先要把它转换为宽形(wide form)，参见表 23. 1。

表 23. 1 面板数据的宽形排列

	y^1	x_1^1	x_2^1	x_3^1	…	y^n	x_1^n	x_2^n	x_3^n
$t=1$									
$t=2$									
$t=3$									
…									
$t=T$									

其中，y^i 表示个体 i 的被解释变量($i=1,\cdots,n$)，x_k^i 表示个体 i 的第 k 个解释变量($i=1,\cdots,n;k=1,\cdots,K$)。使用命令 reshape 来完成面板数据在长形与宽形之间的转换[①]：

① 还可以使用命令 reshape 从面板数据中提取横截面数据。

```
. use mus08cigar.dta,clear
. reshape wide lnc lnp lnpmin lny,i(year) j(state)
```

```
(note: j = 1 2 3 4 5 6 7 8 9 10)

Data                                long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       300   ->     30
Number of variables                    6   ->     41
j variable (10 values)             state   ->   (dropped)
xij variables:
                                     lnc   ->   lnc1 lnc2 ... lnc10
                                     lnp   ->   lnp1 lnp2 ... lnp10
                                  lnpmin   ->   lnpmin1 lnpmin2 ... lnpmin10
                                     lny   ->   lny1 lny2 ... lny10
-----------------------------------------------------------------------------
```

从上表可知,转换成宽形后,样本容量变为 30,即每位个体都有 30 年的观测数据。另一方面,变量的数量则大大增加。比如,变量 lnc 变为 lnc1,lnc2,…,lnc10,分别表示第 1 个州的 lnc,第 2 个州的 lnc,…,第 10 个州的 lnc;依此类推。

列出此宽形数据的前两个观测值:

```
. list in 1/2
```

```
1.
| year | lnp1 | lnpmin1 | lnc1 | lny1 | lnp2 | lnpmin2 |
|   63 | 4.537577 | 4.446105 | 4.542231 | 7.351354 | 4.358048 | 4.399038 |

| lnc2 | lny2 | lnp3 | lnpmin3 | lnc3 | lny3 | lnp4 |
| 4.828314 | 7.572886 | 4.480007 | 4.434545 | 4.638605 | 7.300023 | 4.414975 |

| lnpmin4 | lnc4 | lny4 | lnp5 | lnpmin5 | lnc5 | lny5 |
| 4.358048 | 4.955827 | 7.928801 | 4.472572 | 4.44993 | 5.051137 | 7.977172 |

| lnp6 | lnpmin6 | lnc6 | lny6 | lnp7 | lnpmin7 | lnc7 |
| 4.472572 | 4.472572 | 5.110782 | 7.723767 | 4.336906 | 4.390974 | 5.506956 |

| lny7 | lnp8 | lnpmin8 | lnc8 | lny8 | lnp9 | lnpmin9 |
| 7.913238 | 4.453739 | 4.476296 | 4.923624 | 7.591199 | 4.476296 | 4.446105 |

| lnc9 | lny9 | lnp10 | lnpmin10 | lnc10 | lny10 |
| 4.684905 | 7.463902 | 4.516375 | 4.399038 | 4.574711 | 7.569832 |

2.
| year | lnp1 | lnpmin1 | lnc1 | lny1 | lnp2 | lnpmin2 |
|   64 | 4.565691 | 4.485369 | 4.558079 | 7.428971 | 4.349237 | 4.409862 |

| lnc2 | lny2 | lnp3 | lnpmin3 | lnc3 | lny3 | lnp4 |
| 4.795791 | 7.632167 | 4.47807 | 4.417674 | 4.630838 | 7.384039 | 4.409862 |

| lnpmin4 | lnc4 | lny4 | lnp5 | lnpmin5 | lnc5 | lny5 |
| 4.349237 | 4.929425 | 7.997976 | 4.496219 | 4.463309 | 4.966335 | 8.040493 |

| lnp6 | lnpmin6 | lnc6 | lny6 | lnp7 | lnpmin7 | lnc7 |
| 4.470717 | 4.521087 | 5.141078 | 7.751622 | 4.345061 | 4.398027 | 5.460861 |

| lny7 | lnp8 | lnpmin8 | lnc8 | lny8 | lnp9 | lnpmin9 |
| 7.970397 | 4.56233 | 4.485369 | 4.832306 | 7.660546 | 4.485369 | 4.448328 |

| lnc9 | lny9 | lnp10 | lnpmin10 | lnc10 | lny10 |
| 4.678421 | 7.531511 | 4.552177 | 4.436943 | 4.50092 | 7.611648 |
```

下面进行 SUR 回归：

```
. sureg (lnc1 lnp1 lnpmin1 lny1)(lnc2 lnp2 lnpmin2 lny2)(lnc3 lnp3
lnpmin3 lny3)(lnc4 lnp4 lnpmin4 lny4)(lnc5 lnp5 lnpmin5 lny5)(lnc6 lnp6
lnpmin6 lny6)(lnc7 lnp7 lnpmin7 lny7)(lnc8 lnp8 lnpmin8 lny8)(lnc9 lnp9
lnpmin9 lny9)(lnc10 lnp10 lnpmin10 lny10),corr
```

Seemingly unrelated regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
lnc1	30	3	.0361064	0.8696	250.40	0.0000
lnc2	30	3	.0576629	0.7873	143.72	0.0000
lnc3	30	3	.031615	0.8852	249.10	0.0000
lnc4	30	3	.0409479	0.9578	727.41	0.0000
lnc5	30	3	.0460177	0.9072	359.52	0.0000
lnc6	30	3	.0310659	0.8903	241.90	0.0000
lnc7	30	3	.0849558	0.9447	525.99	0.0000
lnc8	30	3	.0398698	0.7859	179.43	0.0000
lnc9	30	3	.033639	0.8369	211.75	0.0000
lnc10	30	3	.0646011	0.7739	118.50	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnc1						
lnp1	-.7439516	.0926778	-8.03	0.000	-.9255968	-.5623064
lnpmin1	.2611873	.1049528	2.49	0.013	.0554835	.466891
lny1	.0954658	.0097772	9.76	0.000	.0763029	.1146287
_cons	6.066224	.2589029	23.43	0.000	5.558784	6.573664
lnc2						
lnp2	-.3412215	.0842515	-4.05	0.000	-.5063514	-.1760916
lnpmin2	-.2761468	.0863833	-3.20	0.001	-.445455	-.1068386
lny2	-.0758077	.0159596	-4.75	0.000	-.1070879	-.0445274
_cons	8.134837	.2955115	27.53	0.000	7.555645	8.714029
lnc3						
lnp3	-.1949179	.0954677	-2.04	0.041	-.3820312	-.0078046
lnpmin3	-.1936758	.1022655	-1.89	0.058	-.3941125	.0067609
lny3	.1272329	.0084641	15.03	0.000	.1106436	.1438222
_cons	5.401923	.2095664	25.78	0.000	4.99118	5.812665
lnc4						
lnp4	-.5295479	.0650439	-8.14	0.000	-.6570316	-.4020642
lnpmin4	-.1832105	.0870083	-2.11	0.035	-.3537436	-.0126774
lny4	-.1915741	.0127586	-15.02	0.000	-.2165806	-.1665677
_cons	9.645321	.2376458	40.59	0.000	9.179544	10.1111
lnc5						
lnp5	-.8264925	.0857025	-9.64	0.000	-.9944662	-.6585187
lnpmin5	.2775828	.0893441	3.11	0.002	.1024715	.4526941
lny5	-.1236889	.0137497	-9.00	0.000	-.1506378	-.0967401
_cons	8.441914	.2686479	31.42	0.000	7.915373	8.968454
lnc6						
lnp6	-.1827293	.0654166	-2.79	0.005	-.3109435	-.0545152
lnpmin6	-.0800041	.0637778	-1.25	0.210	-.2050062	.0449981
lny6	-.1152345	.0088816	-12.97	0.000	-.132642	-.097827
_cons	7.199805	.217034	33.17	0.000	6.774426	7.625184

lnc7						
lnp7	-.3231466	.2431009	-1.33	0.184	-.7996156	.1533224
lnpmin7	-.0444938	.148016	-0.30	0.764	-.3345998	.2456122
lny7	-.4557646	.0434102	-10.50	0.000	-.540847	-.3706821
_cons	10.76327	.4644809	23.17	0.000	9.852906	11.67364
lnc8						
lnp8	-.4363926	.0599955	-7.27	0.000	-.5539817	-.3188035
lnpmin8	-.130435	.0690307	-1.89	0.059	-.2657326	.0048626
lny8	-.0394183	.0110457	-3.57	0.000	-.0610675	-.0177691
_cons	7.771211	.2481084	31.32	0.000	7.284928	8.257495
lnc9						
lnp9	-.2814291	.0596311	-4.72	0.000	-.3983038	-.1645544
lnpmin9	-.2136093	.0535691	-3.99	0.000	-.3186029	-.1086157
lny9	.0723731	.0085299	8.48	0.000	.0556549	.0890913
_cons	6.349793	.1943468	32.67	0.000	5.968881	6.730706
lnc10						
lnp10	-1.062045	.1564136	-6.79	0.000	-1.36861	-.7554796
lnpmin10	.4134775	.1739583	2.38	0.017	.0725255	.7544294
lny10	-.0121777	.0182374	-0.67	0.504	-.0479223	.0235668
_cons	7.662863	.3629212	21.11	0.000	6.951551	8.374176

```
Correlation matrix of residuals:

             lnc1     lnc2     lnc3     lnc4     lnc5     lnc6     lnc7     lnc8
 lnc1   1.0000
 lnc2   0.1650   1.0000
 lnc3   0.5892   0.2456   1.0000
 lnc4   0.3870   0.6052   0.3259   1.0000
 lnc5  -0.2228   0.1637  -0.0553   0.0186   1.0000
 lnc6  -0.0216   0.3286   0.1553   0.3569   0.4797   1.0000
 lnc7   0.0642   0.2279  -0.0223   0.0084   0.5667   0.2576   1.0000
 lnc8   0.2957   0.8255   0.1734   0.5548   0.1800   0.1973   0.3246   1.0000
 lnc9   0.5874   0.7296   0.5889   0.6558   0.1733   0.3400   0.1681   0.6903
lnc10   0.6368   0.3761   0.2247   0.4361  -0.1687  -0.2482  -0.0390   0.4449

             lnc9    lnc10
 lnc9   1.0000
lnc10   0.4798   1.0000

Breusch-Pagan test of independence: chi2(45) =   208.986, Pr = 0.0000
```

上表最后一行显示，Breusch-Pagan LM 检验结果强烈拒绝无同期相关的原假设，故 SUR 比单一方程 OLS 更有效率，应该使用 SUR 来估计此变系数面板模型。

习 题

23.1 证明：如果 $\boldsymbol{\Omega}$ 是单位矩阵，则系统 GLS 估计等价于单一方程 OLS 估计。

23.2 证明克罗内克尔乘积的三个性质。

23.3 面板数据集 grunfeld.dta 包含了 10 个公司 1935—1954 年的以下变量：invest（投资额），mvalue（公司市场市值），kstock（公司资本存量）。考虑以下投资函数：

$$\text{invest}_{it} = \beta_0 + \beta_1\ \text{mvalue}_{it} + \beta_2\ \text{kstock}_{it} + u_i + \varepsilon_{it} \tag{23.17}$$

使用 SUR 估计一个变系数面板模型，即允许 $\beta_0, \beta_1, \beta_2$ 随公司的不同而不同。

附　录

A23.1　如果每个方程包含完全相同的解释变量，则系统 GLS 还原为单一方程 OLS。

证明：考虑将 $\hat{\boldsymbol{\beta}}_{\text{GLS}} = [\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{y}$ 按矩阵分块展开。

记 $\boldsymbol{\Sigma}^{-1}$ 的第 (i,j) 个元素为 σ^{ij}。由于 $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n \end{pmatrix}$，故

$$\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2' & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n' \end{pmatrix} \begin{pmatrix} \sigma^{11}\boldsymbol{I}_T & \sigma^{12}\boldsymbol{I}_T & \cdots & \sigma^{1n}\boldsymbol{I}_T \\ \sigma^{21}\boldsymbol{I}_T & \sigma^{22}\boldsymbol{I}_T & \cdots & \sigma^{2n}\boldsymbol{I}_T \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}\boldsymbol{I}_T & \sigma^{n2}\boldsymbol{I}_T & \cdots & \sigma^{nn}\boldsymbol{I}_T \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2' & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n' \end{pmatrix} \begin{pmatrix} \sigma^{11}\boldsymbol{X}_1 & \sigma^{12}\boldsymbol{X}_2 & \cdots & \sigma^{1n}\boldsymbol{X}_n \\ \sigma^{21}\boldsymbol{X}_1 & \sigma^{22}\boldsymbol{X}_2 & \cdots & \sigma^{2n}\boldsymbol{X}_n \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}\boldsymbol{X}_1 & \sigma^{n2}\boldsymbol{X}_2 & \cdots & \sigma^{nn}\boldsymbol{X}_n \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^{11}\boldsymbol{X}_1'\boldsymbol{X}_1 & \sigma^{12}\boldsymbol{X}_1'\boldsymbol{X}_2 & \cdots & \sigma^{1n}\boldsymbol{X}_1'\boldsymbol{X}_n \\ \sigma^{21}\boldsymbol{X}_2'\boldsymbol{X}_1 & \sigma^{22}\boldsymbol{X}_2'\boldsymbol{X}_2 & \cdots & \sigma^{2n}\boldsymbol{X}_2'\boldsymbol{X}_n \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}\boldsymbol{X}_n'\boldsymbol{X}_1 & \sigma^{n2}\boldsymbol{X}_n'\boldsymbol{X}_2 & \cdots & \sigma^{nn}\boldsymbol{X}_n'\boldsymbol{X}_n \end{pmatrix}$$

$$\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_T)\boldsymbol{y} = \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2' & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n' \end{pmatrix} \begin{pmatrix} \sigma^{11}\boldsymbol{I}_T & \sigma^{12}\boldsymbol{I}_T & \cdots & \sigma^{1n}\boldsymbol{I}_T \\ \sigma^{21}\boldsymbol{I}_T & \sigma^{22}\boldsymbol{I}_T & \cdots & \sigma^{2n}\boldsymbol{I}_T \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}\boldsymbol{I}_T & \sigma^{n2}\boldsymbol{I}_T & \cdots & \sigma^{nn}\boldsymbol{I}_T \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2' & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_n' \end{pmatrix} \begin{pmatrix} \sigma^{11}\boldsymbol{y}_1 + \sigma^{12}\boldsymbol{y}_2 + \cdots + \sigma^{1n}\boldsymbol{y}_n \\ \sigma^{21}\boldsymbol{y}_1 + \sigma^{22}\boldsymbol{y}_2 + \cdots + \sigma^{2n}\boldsymbol{y}_n \\ \vdots \\ \sigma^{n1}\boldsymbol{y}_1 + \sigma^{n2}\boldsymbol{y}_2 + \cdots + \sigma^{nn}\boldsymbol{y}_n \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^{11}\boldsymbol{X}_1'\boldsymbol{y}_1 + \sigma^{12}\boldsymbol{X}_1'\boldsymbol{y}_2 + \cdots + \sigma^{1n}\boldsymbol{X}_1'\boldsymbol{y}_n \\ \sigma^{21}\boldsymbol{X}_2'\boldsymbol{y}_1 + \sigma^{22}\boldsymbol{X}_2'\boldsymbol{y}_2 + \cdots + \sigma^{2n}\boldsymbol{X}_2'\boldsymbol{y}_n \\ \vdots \\ \sigma^{n1}\boldsymbol{X}_n'\boldsymbol{y}_1 + \sigma^{n2}\boldsymbol{X}_n'\boldsymbol{y}_2 + \cdots + \sigma^{nn}\boldsymbol{X}_n'\boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n}\sigma^{1j}\boldsymbol{X}_1'\boldsymbol{y}_j \\ \sum_{j=1}^{n}\sigma^{2j}\boldsymbol{X}_2'\boldsymbol{y}_j \\ \vdots \\ \sum_{j=1}^{n}\sigma^{nj}\boldsymbol{X}_n'\boldsymbol{y}_j \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}=\begin{pmatrix}\sigma^{11}\boldsymbol{X}_1'\boldsymbol{X}_1 & \sigma^{12}\boldsymbol{X}_1'\boldsymbol{X}_2 & \cdots & \sigma^{1n}\boldsymbol{X}_1'\boldsymbol{X}_n\\ \sigma^{21}\boldsymbol{X}_2'\boldsymbol{X}_1 & \sigma^{22}\boldsymbol{X}_2'\boldsymbol{X}_2 & \cdots & \sigma^{2n}\boldsymbol{X}_2'\boldsymbol{X}_n\\ \vdots & \vdots & & \vdots\\ \sigma^{n1}\boldsymbol{X}_n'\boldsymbol{X}_1 & \sigma^{n2}\boldsymbol{X}_n'\boldsymbol{X}_2 & \cdots & \sigma^{nn}\boldsymbol{X}_n'\boldsymbol{X}_n\end{pmatrix}^{-1}\begin{pmatrix}\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{X}_1'\boldsymbol{y}_j\\ \sum_{j=1}^{n}\sigma^{2j}\boldsymbol{X}_2'\boldsymbol{y}_j\\ \vdots\\ \sum_{j=1}^{n}\sigma^{nj}\boldsymbol{X}_n'\boldsymbol{y}_j\end{pmatrix} \tag{23.18}$$

当每个方程包含完全相同的解释变量时，则 $\boldsymbol{X}_1=\boldsymbol{X}_2=\cdots=\boldsymbol{X}_n\equiv\boldsymbol{X}$①。因此，

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}=\begin{pmatrix}\sigma^{11}\boldsymbol{X}'\boldsymbol{X} & \sigma^{12}\boldsymbol{X}'\boldsymbol{X} & \cdots & \sigma^{1n}\boldsymbol{X}'\boldsymbol{X}\\ \sigma^{21}\boldsymbol{X}'\boldsymbol{X} & \sigma^{22}\boldsymbol{X}'\boldsymbol{X} & \cdots & \sigma^{2n}\boldsymbol{X}'\boldsymbol{X}\\ \vdots & \vdots & & \vdots\\ \sigma^{n1}\boldsymbol{X}'\boldsymbol{X} & \sigma^{n2}\boldsymbol{X}'\boldsymbol{X} & \cdots & \sigma^{nn}\boldsymbol{X}'\boldsymbol{X}\end{pmatrix}^{-1}\begin{pmatrix}\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{X}'\boldsymbol{y}_j\\ \sum_{j=1}^{n}\sigma^{2j}\boldsymbol{X}'\boldsymbol{y}_j\\ \vdots\\ \sum_{j=1}^{n}\sigma^{nj}\boldsymbol{X}'\boldsymbol{y}_j\end{pmatrix} \tag{23.19}$$

记第 j 个方程的单一方程 OLS 估计值向量、拟合值向量与残差向量分别为 $\boldsymbol{b}_j,\hat{\boldsymbol{y}}_j,\boldsymbol{e}_j$，则 $\boldsymbol{y}_j=\hat{\boldsymbol{y}}_j+\boldsymbol{e}_j$。因此，

$$\boldsymbol{X}'\boldsymbol{y}_j=\boldsymbol{X}'(\hat{\boldsymbol{y}}_j+\boldsymbol{e}_j)=\boldsymbol{X}'\hat{\boldsymbol{y}}_j+\boldsymbol{X}'\boldsymbol{e}_j=\boldsymbol{X}'\hat{\boldsymbol{y}}_j=\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_j$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}=\left[\begin{pmatrix}\sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n}\\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n}\\ \vdots & \vdots & & \vdots\\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn}\end{pmatrix}\otimes(\boldsymbol{X}'\boldsymbol{X})\right]^{-1}\begin{pmatrix}\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_j\\ \sum_{j=1}^{n}\sigma^{2j}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_j\\ \vdots\\ \sum_{j=1}^{n}\sigma^{nj}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_j\end{pmatrix}$$

$$=[\boldsymbol{\Sigma}^{-1}\otimes(\boldsymbol{X}'\boldsymbol{X})]^{-1}\begin{pmatrix}(\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{b}_j\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{2j}\boldsymbol{b}_j\\ \vdots\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{nj}\boldsymbol{b}_j\end{pmatrix}$$

$$=[\boldsymbol{\Sigma}\otimes(\boldsymbol{X}'\boldsymbol{X})^{-1}]\begin{pmatrix}(\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{b}_j\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{2j}\boldsymbol{b}_j\\ \vdots\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{nj}\boldsymbol{b}_j\end{pmatrix}$$

$$=\begin{pmatrix}\sigma_{11}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \cdots & \sigma_{1n}(\boldsymbol{X}'\boldsymbol{X})^{-1}\\ \sigma_{21}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \cdots & \sigma_{2n}(\boldsymbol{X}'\boldsymbol{X})^{-1}\\ \vdots & \vdots & & \vdots\\ \sigma_{n1}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \sigma_{n2}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \cdots & \sigma_{nn}(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{pmatrix}\begin{pmatrix}(\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{b}_j\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{2j}\boldsymbol{b}_j\\ \vdots\\ (\boldsymbol{X}'\boldsymbol{X})\sum_{j=1}^{n}\sigma^{nj}\boldsymbol{b}_j\end{pmatrix}$$

上式中左边矩阵的第一行与右边矩阵之乘积为

① 为了书写方便，此证明的 $\boldsymbol{X}$ 与前面章节中的 $\boldsymbol{X}$ 定义不同。

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{1,\mathrm{GLS}} &= \sigma_{11}\sum_{j=1}^{n}\sigma^{1j}\boldsymbol{b}_j + \sigma_{12}\sum_{j=1}^{n}\sigma^{2j}\boldsymbol{b}_j + \cdots + \sigma_{1n}\sum_{j=1}^{n}\sigma^{nj}\boldsymbol{b}_j \\ &= (\sigma_{11}\sigma^{11} + \sigma_{12}\sigma^{21} + \cdots + \sigma_{1n}\sigma^{n1})\boldsymbol{b}_1 + (\sigma_{11}\sigma^{12} + \sigma_{12}\sigma^{22} + \cdots + \sigma_{1n}\sigma^{n2})\boldsymbol{b}_2 \\ &\quad + \cdots + (\sigma_{11}\sigma^{1n} + \sigma_{12}\sigma^{2n} + \cdots + \sigma_{1n}\sigma^{nn})\boldsymbol{b}_n \quad (\text{合并同类项}) \\ &= 1\cdot\boldsymbol{b}_1 + 0\cdot\boldsymbol{b}_2 + \cdots + 0\cdot\boldsymbol{b}_n = \boldsymbol{b}_1\end{aligned}$$

最后一步能成立是因为

$$\boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1} = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n}\\ \vdots & \vdots & & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}\end{pmatrix}\begin{pmatrix}\sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n}\\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n}\\ \vdots & \vdots & & \vdots\\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn}\end{pmatrix} = \boldsymbol{I}_n$$

同理可证，$\hat{\boldsymbol{\beta}}_{i,\mathrm{GLS}} = \boldsymbol{b}_i \quad (i = 1, \cdots, n)$。

第 24 章　联立方程模型

24.1　联立方程模型的结构式与简化式

迄今为止，我们主要关注单一方程。即使在上一章使用 SUR 对多个方程同时进行估计，各方程的变量之间也并没有内在关系（除了各方程扰动项的相关性以外）。然而，经济理论常常推导出一组相互联系的方程，其中一个方程的解释变量是另一方程的被解释变量，这就是联立方程组。

例　农产品市场均衡模型，由需求函数、供给函数及市场均衡条件组成，参见第 10 章。

例　简单的宏观经济模型[①]，参见第 10 章。

即使我们只关心单个方程，但如果该方程包含内生解释变量，则完整的模型仍然是联立方程组（另一方程以此内生解释变量为被解释变量）。为此，有必要研究如何估计联立方程模型。

由 M 个方程构成的联立方程模型的“结构式”（structural form）可以表示为

$$\begin{cases} \gamma_{11}y_{t1}+\gamma_{21}y_{t2}+\cdots+\gamma_{M1}y_{tM}+\beta_{11}x_{t1}+\cdots+\beta_{K1}x_{tK}=\varepsilon_{t1} \\ \gamma_{12}y_{t1}+\gamma_{22}y_{t2}+\cdots+\gamma_{M2}y_{tM}+\beta_{12}x_{t1}+\cdots+\beta_{K2}x_{tK}=\varepsilon_{t2} \\ \cdots\cdots\cdots\cdots \\ \gamma_{1M}y_{t1}+\gamma_{2M}y_{t2}+\cdots+\gamma_{MM}y_{tM}+\beta_{1M}x_{t1}+\cdots+\beta_{KM}x_{tK}=\varepsilon_{tM} \end{cases} \tag{24.1}$$

其中，$\{y_{ti}\}$ 为内生变量，$\{x_{tj}\}$ 为外生变量[②]，其第一个下标表示第 t 个观测值（$t=1,\cdots,T$）[③]，第二个下标表示第 i 个内生变量（$i=1,\cdots,M$），或第 j 个外生变量（$j=1,\cdots,K$）。内生变量的系数为 $\{\gamma_{ik}\}$，其第一个下标表示它是第 i 个内生变量的系数，而第二个下标表示它在第 k 个方程中（$k=1,\cdots,M$）。类似地，外生变量的系数为 $\{\beta_{jk}\}$，其第一个下标表示它是第 j 个外生变量的系数，而第二个下标表示它在第 k 个方程中。结构方程的扰动项为 $\{\varepsilon_{tk}\}$，其第一个下标表示第 t 个观测值（$t=1,\cdots,T$），而第二个下标表示它在第 k 个方程中。“完整的方程系统”（complete system of equations）要求，内生变量个数等于方程个数 M。

将上述方程组写成更简洁的“横排”矩阵形式

$$\begin{aligned} &(y_{t1}\quad y_{t2}\quad\cdots\quad y_{tM})\begin{pmatrix}\gamma_{11}&\gamma_{12}&\cdots&\gamma_{1M}\\ \gamma_{21}&\gamma_{22}&\cdots&\gamma_{2M}\\ \vdots&\vdots&&\vdots\\ \gamma_{M1}&\gamma_{M2}&\cdots&\gamma_{MM}\end{pmatrix}+(x_{t1}\quad x_{t2}\quad\cdots\quad x_{tK})\begin{pmatrix}\beta_{11}&\beta_{12}&\cdots&\beta_{1M}\\ \beta_{21}&\beta_{22}&\cdots&\beta_{2M}\\ \vdots&\vdots&&\vdots\\ \beta_{K1}&\beta_{K2}&\cdots&\beta_{KM}\end{pmatrix} \\ &=(\varepsilon_{t1}\quad\varepsilon_{t2}\quad\cdots\quad\varepsilon_{tM}) \end{aligned} \tag{24.2}$$

---

① 联立方程形式的宏观经济模型在 20 世纪 70 年代较流行，但由于方程的稳定性及预测能力不理想，现已基本为时间序列模型所替代，比如 VAR 模型。

② 外生变量指的是，由系统外部决定的变量（包括常数项）；而内生变量则在系统内部决定。

③ 数据形式既可以是时间序列（则 t 表示时间），也可以是横截面数据（则 t 表示个体）。

用矩阵来表示即

$$\boldsymbol{y}_t'\boldsymbol{\Gamma}+\boldsymbol{x}_t'\boldsymbol{B}=\boldsymbol{\varepsilon}_t' \tag{24.3}$$

其中，系数矩阵 $\boldsymbol{\Gamma}_{M\times M}$ 与 $\boldsymbol{B}_{K\times M}$ 的每一列对应于一个方程。比如，第一个方程为

$$(y_{t1}\quad y_{t2}\quad\cdots\quad y_{tM})\begin{pmatrix}\gamma_{11}\\ \gamma_{21}\\ \vdots\\ \gamma_{M1}\end{pmatrix}+(x_{t1}\quad x_{t2}\quad\cdots\quad x_{tK})\begin{pmatrix}\beta_{11}\\ \beta_{21}\\ \vdots\\ \beta_{K1}\end{pmatrix}=\varepsilon_{t1} \tag{24.4}$$

扰动项 $\boldsymbol{\varepsilon}_t$ 由第 t 期各方程的扰动项所构成。假设扰动项 $\boldsymbol{\varepsilon}_t$ 满足 $\mathrm{E}(\boldsymbol{\varepsilon}_t|\boldsymbol{x}_t)=\mathbf{0}$（因为 $\boldsymbol{x}_t$ 是外生变量），记其协方差矩阵为，

$$\boldsymbol{\Sigma}\equiv\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\boldsymbol{x}_t) \tag{24.5}$$

由于存在内生变量，如果直接用 OLS 估计联立方程组中的每一方程，将导致内生变量偏差（endogeneity bias）或联立方程偏差（simultaneity bias），得不到一致估计。为此，求解联立方程组（24.3），将 $\boldsymbol{y}_t$ 表示为 $\boldsymbol{x}_t$ 与 $\boldsymbol{\varepsilon}_t$ 的函数：

$$\boldsymbol{y}_t'\boldsymbol{\Gamma}=-\boldsymbol{x}_t'\boldsymbol{B}+\boldsymbol{\varepsilon}_t' \tag{24.6}$$

假设 $\boldsymbol{\Gamma}$ 非退化，在上式两边同时右乘 $\boldsymbol{\Gamma}^{-1}$，

$$\boldsymbol{y}_t'=-\boldsymbol{x}_t'\boldsymbol{B}\boldsymbol{\Gamma}^{-1}+\boldsymbol{\varepsilon}_t'\boldsymbol{\Gamma}^{-1} \tag{24.7}$$

$$\boldsymbol{y}_t'=\boldsymbol{x}_t'\boldsymbol{\Pi}+\boldsymbol{v}_t' \tag{24.8}$$

方程（24.8）被称为“简化式”（reduced form），其系数矩阵为 $\underbrace{\boldsymbol{\Pi}}_{K\times M}\equiv-\underbrace{\boldsymbol{B}}_{K\times M}\underbrace{\boldsymbol{\Gamma}^{-1}}_{M\times M}$，其扰动项为 $\boldsymbol{v}_t'\equiv\boldsymbol{\varepsilon}_t'\boldsymbol{\Gamma}^{-1}$，故 $\boldsymbol{v}_t\equiv\boldsymbol{\Gamma}^{-1\prime}\boldsymbol{\varepsilon}_t$。简化式扰动项 $\boldsymbol{v}_t$ 仍然与外生变量 $\boldsymbol{x}_t$ 不相关，因为

$$\mathrm{E}(\boldsymbol{v}_t|\boldsymbol{x}_t)=\mathrm{E}(\boldsymbol{\Gamma}^{-1\prime}\boldsymbol{\varepsilon}_t|\boldsymbol{x}_t)=\boldsymbol{\Gamma}^{-1\prime}\mathrm{E}(\boldsymbol{\varepsilon}_t|\boldsymbol{x}_t)=\mathbf{0} \tag{24.9}$$

简化式扰动项 $\boldsymbol{v}_t$ 的协方差矩阵为

$$\boldsymbol{\Omega}\equiv\mathrm{E}(\boldsymbol{v}_t\boldsymbol{v}_t'|\boldsymbol{x}_t)=\mathrm{E}(\boldsymbol{\Gamma}^{-1\prime}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\boldsymbol{\Gamma}^{-1}|\boldsymbol{x}_t)=\boldsymbol{\Gamma}^{-1\prime}\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\boldsymbol{x}_t)\boldsymbol{\Gamma}^{-1}=\boldsymbol{\Gamma}^{-1\prime}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} \tag{24.10}$$

由于每个简化式方程仅包含一个内生变量（即方程左边的被解释变量），而方程右边的解释变量全部为外生变量 $\boldsymbol{x}_t$，故可以使用 OLS 得到简化式参数 $\boldsymbol{\Pi}$ 与 $\boldsymbol{\Omega}$ 的一致估计。然而，通常我们最终关心的是结构式参数。那么，在什么情况下，才能从简化式参数 $(\boldsymbol{\Pi},\boldsymbol{\Omega})$ 反推出结构式参数 $(\boldsymbol{\Gamma},\boldsymbol{B},\boldsymbol{\Sigma})$ 呢？这就涉及联立方程模型的“识别问题”（problem of identification）。

24.2 联立方程模型的识别

在对模型的总体参数进行估计之前，其参数必须“可识别”（identified）。可识别是进行参数估计的前提条件。如果一个总体参数可识别，则该参数的任意两个不同取值，都会在随机样本中显示出系统差异，即如果样本容量足够大，则应该能够在统计意义上区分这两个不同的参数值。反之，如果无论多大的样本都区分不开，即由不同参数值的总体产生的观测数据在统计意义上是一样的，则存在“观测等价”（observational equivalence），该参数“不可识别”（unidentified）。

下面以一个简单例子来说明可识别的概念。考虑以下回归模型：

$$y_i=\alpha_1+\alpha_2+\beta x_i+\varepsilon_i \tag{24.11}$$

显然，仅通过样本数据 $\{y_i,x_i\}_{i=1}^n$ 是无法对 α_1 与 α_2 分别进行识别的，但可以识别二者之和

$(\alpha_1+\alpha_2)$。回到联立方程模型的情形,“可识别”意味着,可以从简化式参数$(\boldsymbol{\Pi},\boldsymbol{\Omega})$求出结构式参数$(\boldsymbol{\Gamma},\boldsymbol{B},\boldsymbol{\Sigma})$的唯一解(unique solution)。从上面的推导可知,这两组参数之间的关系如下:

$$\boldsymbol{\Pi}\equiv -\boldsymbol{B}\boldsymbol{\Gamma}^{-1} \tag{24.12}$$

$$\boldsymbol{\Omega}\equiv \boldsymbol{\Gamma}^{-1\prime}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} \tag{24.13}$$

显然,如果$\boldsymbol{\Gamma}$已知,则可以通过$\boldsymbol{\Pi}$与$\boldsymbol{\Omega}$求得$\boldsymbol{B}$与$\boldsymbol{\Sigma}$。然而,一般来说,$\boldsymbol{\Gamma}$是由未知参数组成的矩阵。事实上,结构式的参数个数比简化式的参数个数多出M^2个。这是因为,简化式参数$(\boldsymbol{\Pi},\boldsymbol{\Omega})$的总个数为$[K\times M+M(M+1)/2]$(其中,$\boldsymbol{\Pi}_{K\times M}$含$K\times M$个参数,而对称矩阵$\boldsymbol{\Omega}_{M\times M}$含$M(M+1)/2$个参数);而结构式参数$(\boldsymbol{\Gamma},\boldsymbol{B},\boldsymbol{\Sigma})$的总个数为$[M^2+K\times M+M(M+1)/2]$(其中,$\boldsymbol{\Gamma}_{M\times M}$含$M^2$个参数,$\boldsymbol{B}_{K\times M}$含$K\times M$个参数,对称矩阵$\boldsymbol{\Sigma}$含$M(M+1)/2$个参数)。这表明,在一般情况下,不可能从$(\boldsymbol{\Pi},\boldsymbol{\Omega})$求出$(\boldsymbol{\Gamma},\boldsymbol{B},\boldsymbol{\Sigma})$的唯一解。因此,如果不对结构式参数进行约束,将不可能从简化式参数得到结构式参数的唯一解,结构方程也就不可识别。为了识别结构方程,常对结构参数施加如下约束。

(1) 标准化(normalization)。在每个结构方程中,可以将一个内生变量视为被解释变量,并将其系数标准化为1。即使这样,结构式参数仍然比简化式参数多(M^2-M)个。

(2) 恒等式(identity)。比如,供需相等的均衡条件、会计恒等式、定义式。恒等式中每个变量的系数均为已知,不需要识别或估计。

(3) 排斥约束(exclusion restrictions)。在结构方程中排斥某些内生或外生变量,这相当于对结构矩阵$(\boldsymbol{\Gamma},\boldsymbol{B})$施以“零约束”(zero restrictions),即让$(\boldsymbol{\Gamma},\boldsymbol{B})$中的某些元素为0。

(4) 线性约束(linear restriction)。比如,在理论上可以假设生产函数为规模报酬不变(constant returns to scale),则资本的产出弹性与劳动力的产出弹性之和为1。

(5) 对扰动项协方差矩阵的约束(restrictions on the disturbance covariance matrix)。比如,在某些情况下,可以假设不同方程的扰动项之间不相关。

在实践中,最重要的约束方法是“排斥变量”(即零约束)。对于线性约束,可以通过重新定义变量而将其转化为“排斥变量”约束。那么,究竟需要多少零约束才可以保证结构方程可识别呢?

由于系数矩阵$\boldsymbol{\Gamma}$与$\boldsymbol{B}$的每一列对应于一个方程,故矩阵$\boldsymbol{\Gamma}$的第k列$\boldsymbol{\Gamma}_k$与矩阵$\boldsymbol{B}$的第k列$\boldsymbol{B}_k$包含了第k个方程的系数。由于$\boldsymbol{\Pi}=-\boldsymbol{B}\boldsymbol{\Gamma}^{-1}$,故$\boldsymbol{\Pi}\boldsymbol{\Gamma}=-\boldsymbol{B}$。将$\boldsymbol{\Gamma}$与$\boldsymbol{B}$进行矩阵分块可得

$$\underbrace{\boldsymbol{\Pi}}_{K\times M}\underbrace{(\boldsymbol{\Gamma}_1\cdots\boldsymbol{\Gamma}_M)}_{M\times M}=\underbrace{(\boldsymbol{\Pi}\boldsymbol{\Gamma}_1\cdots\boldsymbol{\Pi}\boldsymbol{\Gamma}_M)}_{K\times M}=-\underbrace{(\boldsymbol{B}_1\cdots\boldsymbol{B}_M)}_{K\times M} \tag{24.14}$$

由此可知,$\boldsymbol{\Gamma}_k$与$\boldsymbol{B}_k$满足以下方程组:

$$\boldsymbol{\Pi}\boldsymbol{\Gamma}_k=-\boldsymbol{B}_k \tag{24.15}$$

不失一般性,我们仅考虑第一个结构方程,即$k=1$[①]。假设在第一个方程中,内生变量y_1的系数已被标准化为1,另有M_1个内生变量(以$\boldsymbol{y}_1$来表示)也包括在此方程中,而其余M_1^*个内生变量(以$\boldsymbol{y}_1^*$来表示)则被排斥在此方程之外,故$1+M_1+M_1^*=M$。不失一般性,将$\boldsymbol{y}_1$排在y_1之后、$\boldsymbol{y}_1^*$之前,即$(y_1\ \boldsymbol{y}_1\ \boldsymbol{y}_1^*)'$。同时假设,第一个方程包含$K_1$个外生变量(以$\boldsymbol{x}_1$来表示),而其余$K_1^*$个外生变量(以$\boldsymbol{x}_1^*$来表示)则被排斥在此方程之外,故$K_1+K_1^*=K$。不失一般性,将$\boldsymbol{x}_1$排在$\boldsymbol{x}_1^*$之前,即$(\boldsymbol{x}_1\ \boldsymbol{x}_1^*)'$。以右上角的$*$表示被排斥的变量及其对应的系数,则第一个方程可以写为

$$y_1=\boldsymbol{y}_1'\boldsymbol{\gamma}_1+\boldsymbol{y}_1^{*\prime}\boldsymbol{\gamma}_1^*+\boldsymbol{x}_1'\boldsymbol{\beta}_1+\boldsymbol{x}_1^{*\prime}\boldsymbol{\beta}_1^*+\varepsilon_1 \tag{24.16}$$

---

① 如果$k>1$,可以通过调整方程的排序而使得$k=1$。

显然，根据“排斥变量”约束，$\boldsymbol{\gamma}_1^* = \mathbf{0}, \boldsymbol{\beta}_1^* = \mathbf{0}$。因此，矩阵 $\boldsymbol{\Gamma}$ 的第 1 列为 $\boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 \\ -\boldsymbol{\gamma}_1 \\ \mathbf{0} \end{pmatrix}$，矩阵 $\boldsymbol{B}$ 的第 1 列为 $\boldsymbol{B}_1 = \begin{pmatrix} -\boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}$，故 $\boldsymbol{\Pi}_{K\times M} \cdot \begin{pmatrix} 1 \\ -\boldsymbol{\gamma}_1 \\ \mathbf{0} \end{pmatrix}_{M\times 1} = \begin{pmatrix} -\boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}_{K\times 1}$。

将简化式矩阵 $\boldsymbol{\Pi}$ 进行相应的分块可得（当将 $\boldsymbol{y}_1$ 排在 $\boldsymbol{y}_1^*$ 之前，$\boldsymbol{x}_1$ 排在 $\boldsymbol{x}_1^*$ 之前时，对矩阵 $\boldsymbol{\Pi}$ 的相应行与列也进行相应的调整）

$$\begin{array}{c} \\ K_1\text{行} \\ K_1^*\text{行} \end{array} \overset{\begin{array}{ccc} 1\text{列} & M_1\text{列} & M_1^*\text{列} \end{array}}{\begin{pmatrix} \boldsymbol{\pi}_1 & \underline{\boldsymbol{\Pi}}_1 & \overline{\boldsymbol{\Pi}}_1 \\ \boldsymbol{\pi}_1^* & \underline{\boldsymbol{\Pi}}_1^* & \overline{\boldsymbol{\Pi}}_1^* \end{pmatrix}_{K\times M}} \begin{pmatrix} 1 \\ -\boldsymbol{\gamma}_1 \\ \mathbf{0} \end{pmatrix}_{M\times 1} = \begin{pmatrix} -\boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}_{K\times 1} \begin{array}{l} K_1\text{行} \\ K_1^*\text{行} \end{array} \tag{24.17}$$

将上式左边的两个矩阵相乘后，可以得到以下两个方程组

$$\underbrace{\boldsymbol{\pi}_1}_{K_1\times 1} - \underbrace{\underline{\boldsymbol{\Pi}}_1}_{K_1\times M_1} \underbrace{\boldsymbol{\gamma}_1}_{M_1\times 1} = -\underbrace{\boldsymbol{\beta}_1}_{K_1\times 1} \tag{24.18}$$

$$\underbrace{\underline{\boldsymbol{\Pi}}_1^*}_{K_1^*\times M_1} \underbrace{\boldsymbol{\gamma}_1}_{M_1\times 1} = \underbrace{\boldsymbol{\pi}_1^*}_{K_1^*\times 1} \tag{24.19}$$

如果可以根据 $\boldsymbol{\Pi}$ 求出 $(\boldsymbol{\gamma}_1, \boldsymbol{\beta}_1)$ 的唯一解，则第一个结构方程可识别。假如知道 $\boldsymbol{\gamma}_1$，就可以通过方程组(24.18)求解 $\boldsymbol{\beta}_1$。因此，问题的关键在于求解 $\boldsymbol{\gamma}_1$。第一个结构方程可识别的充分必要条件是，在方程组 $\underline{\boldsymbol{\Pi}}_1^* \boldsymbol{\gamma}_1 = \boldsymbol{\pi}_1^*$ 中，$\boldsymbol{\gamma}_1$ 有唯一解。根据线性代数的知识，这要求满足以下秩条件，

$$\operatorname{rank}(\underline{\boldsymbol{\Pi}}_1^*) = \operatorname{rank}(\underline{\boldsymbol{\Pi}}_1^* \quad \boldsymbol{\pi}_1^*) = M_1 \tag{24.20}$$

即系数矩阵 $\underline{\boldsymbol{\Pi}}_1^*$ 的秩等于增广矩阵 $(\underline{\boldsymbol{\Pi}}_1^*, \boldsymbol{\pi}_1^*)$ 的秩，且等于未知数的个数 M_1。秩条件是可识别的充分必要条件。如果秩条件不满足，则不可识别。由于矩阵的秩一定小于或等于矩阵的行数，而 $\underline{\boldsymbol{\Pi}}_1^*$ 为 $K_1^* \times M_1$ 级矩阵，故 $K_1^* \geqslant \operatorname{rank}(\underline{\boldsymbol{\Pi}}_1^*) = M_1$。因此，秩条件成立的必要条件为

$$K_1^* \geqslant M_1 \tag{24.21}$$

这被称为“阶条件”(order condition)。它意味着，结构方程所排斥的外生变量的个数(K_1^*)应大于或等于该方程所包含的内生解释变量的个数(M_1)。显然，阶条件只是可识别的必要条件，而非充分条件。

从工具变量法的角度(这是估计联立方程组的主要方法，参见下文)，如果有 M_1 个或更多的有效工具变量，则可以对含有 M_1 个内生解释变量的结构方程进行一致估计。容易看出，被第一个结构方程排斥的所有外生变量都是有效工具变量，因为根据外生变量的定义，它们与扰动项不相关(外生性)；另一方面，根据简化式，内生变量可以表示为外生变量的函数，故它们与内生解释变量相关(相关性)。因此，阶条件“$K_1^* \geqslant M_1$”的含义是，需要有足够多的工具变量，才能对该方程进行一致估计。由于阶条件很容易检验，而秩条件不容易检验(矩阵 $\underline{\boldsymbol{\Pi}}_1^*$ 包含未知参数，如何计算它的秩呢?)，并且阶条件满足而秩条件不满足的情况很少见，故在实践中，常常只检验阶条件(只要数变量的个数即可)，而忽略秩条件。

在可识别(即秩条件满足)的情况下，如果恰好 $K_1^* = M_1$，则称该结构方程“恰好识别”(just identified)，即工具变量个数正好相等内生解释变量的个数；如果 $K_1^* > M_1$，则称该结构方程“过

度识别"(overidentified),即工具变量个数大于内生解释变量的个数。在过度识别的情况下,方程组 $\underline{\boldsymbol{\Pi}}_1^* \boldsymbol{\gamma}_1 = \boldsymbol{\pi}_1^*$ 存在多余的方程。在理论上,无论使用 K_1^* 方程中的任意 M_1 个方程,都能得到 $\boldsymbol{\gamma}_1$ 相同的唯一解。

24.3 单一方程估计法

估计联立方程组的方法可以分为两类,即"单一方程估计法"(single equation estimation),也称"有限信息估计法"(limited information estimation);以及系统估计法,也称"全信息估计法"(full information estimation)。前者对联立方程组中的每一个方程分别进行估计,而后者则将其作为一个系统进行联合估计。单一方程估计法主要包括,普通最小二乘法、间接最小二乘法、二阶段最小二乘法以及广义矩估计法。下一节考虑系统估计法。

1. 普通最小二乘法

由于存在内生解释变量,一般来说,OLS 是不一致的。但由于其计算简单,OLS 仍然可以作为一种参照系。另外,对于一种特殊的递归模型(recursive model)[①],即 $\boldsymbol{\Gamma}$ 为下三角矩阵(lower triangular matrix)而协方差矩阵 $\boldsymbol{\Sigma}$ 为对角矩阵(不同方程之间的扰动项不相关)的情形,OLS 依然是一致的。以一个三方程的系统为例:

$$\begin{cases} y_1 = \boldsymbol{x}'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2 = \boldsymbol{x}'\boldsymbol{\beta}_2 + \gamma_{12} y_1 + \varepsilon_2 \\ y_3 = \boldsymbol{x}'\boldsymbol{\beta}_3 + \gamma_{13} y_1 + \gamma_{23} y_2 + \varepsilon_3 \end{cases} \tag{24.22}$$

显然,第一个方程不含内生解释变量,可以用 OLS 得到一致估计。在第二个方程中,唯一的内生解释变量为 y_1,而且与扰动项不相关,因为

$$\mathrm{Cov}(y_1, \varepsilon_2) = \mathrm{Cov}(\boldsymbol{x}'\boldsymbol{\beta}_1 + \varepsilon_1, \varepsilon_2) = \underbrace{\mathrm{Cov}(\boldsymbol{x}'\boldsymbol{\beta}_1, \varepsilon_2)}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_1, \varepsilon_2)}_{=0} = 0 \tag{24.23}$$

因此,可以用 OLS 来估计第二个方程。同理,在第三个方程中,内生解释变量为 (y_1, y_2),而且 $\mathrm{Cov}(y_1, \varepsilon_3) = \mathrm{Cov}(y_2, \varepsilon_3) = 0$,故也可以用 OLS 来估计。

2. 间接最小二乘法

在恰好识别的情况下,可以先用 OLS 来一致地估计简化式参数,然后通过结构式参数与简化式参数的关系来求解结构式参数,这被称为"间接最小二乘法"(Indirect Least Square,简记 ILS)。然而,虽然 ILS 在恰好识别的情况下是一致的,但却不是最有效率的,故不常用。

ILS 的另一缺点是,在过度识别的情况下无法使用。这是因为,在过度识别时,根据简化式参数求解结构式参数的联立方程组 $\underline{\boldsymbol{\Pi}}_1^* \boldsymbol{\gamma}_1 = \boldsymbol{\pi}_1^*$ 有多余的方程。理论上,无论使用哪些方程,都能得到相同的唯一解。但在实践中,却会得到不同的估计量,而我们不知道如何取舍。以简单的商品供需模型为例,

$$\begin{cases} q = \beta p + \gamma y & (\text{需求}) \\ q = \delta p & (\text{供给}) \end{cases} \tag{24.24}$$

其中,内生变量 q 与 p 分别表示商品的数量与价格,外生变量 y 表示收入,为了简便略去了扰动项。由于供给方程包含一个内生解释变量 p,而排斥了一个外生变量 y,故为恰好识别;但需求方

① 也称为"三角系统"(triangular system)。

程不可识别。这个联立方程模型的简化式为

$$\begin{cases} p = \dfrac{\gamma}{\delta - \beta} y \equiv \pi_1 y \\ q = \dfrac{\gamma\delta}{\delta - \beta} y \equiv \pi_2 y \end{cases} \tag{24.25}$$

其中，$\pi_1 \equiv \dfrac{\gamma}{\delta - \beta}$，$\pi_2 \equiv \dfrac{\gamma\delta}{\delta - \beta}$。显然，$\pi_2 / \pi_1 = \delta$。记 $\hat{\pi}_1$，$\hat{\pi}_2$ 为简化式参数 π_1，π_2 的 OLS 估计量，则结构参数 δ 的 ILS 估计量为 $\hat{\delta}_{\text{ILS}} = \hat{\pi}_2 / \hat{\pi}_1$。ILS 在恰好识别的情况下成立。

现在假设需求方程中还有另一外生变量 x，而供给方程不变：

$$\begin{cases} q = \beta p + \gamma y + \theta x & \text{（需求）} \\ q = \delta p & \text{（供给）} \end{cases} \tag{24.26}$$

此时，供给方程包含一个内生解释变量 p，排斥两个外生变量 (y, x)，故为过度识别。需求方程仍然不可识别。这个新模型的简化式为

$$\begin{cases} p = \dfrac{\gamma}{\delta - \beta} y + \dfrac{\theta}{\delta - \beta} x \equiv \pi_1 y + \pi_3 x \\ q = \dfrac{\gamma\delta}{\delta - \beta} y + \dfrac{\theta\delta}{\delta - \beta} x \equiv \pi_2 y + \pi_4 x \end{cases} \tag{24.27}$$

其中，$\pi_1 \equiv \dfrac{\gamma}{\delta - \beta}$，$\pi_2 \equiv \dfrac{\gamma\delta}{\delta - \beta}$，$\pi_3 \equiv \dfrac{\theta}{\delta - \beta}$，$\pi_4 \equiv \dfrac{\theta\delta}{\delta - \beta}$。显然，$\pi_2 / \pi_1 = \delta$，$\pi_4 / \pi_3 = \delta$。记 $\hat{\pi}_1$，$\hat{\pi}_2$，$\hat{\pi}_3$，$\hat{\pi}_4$ 为简化式参数 π_1，π_2，π_3，π_4 的 OLS 估计量，则存在对结构参数 δ 的两个 ILS 估计，即 $\hat{\pi}_2 / \hat{\pi}_1$ 与 $\hat{\pi}_4 / \hat{\pi}_3$。一般来说，$\hat{\pi}_2 / \hat{\pi}_1 \neq \hat{\pi}_4 / \hat{\pi}_3$，而我们不知道究竟该用何者作为 $\hat{\delta}_{\text{ILS}}$，故 ILS 不适用①。

3. 二阶段最小二乘法

在结构方程可识别的情况下，其排斥的外生变量个数大于或等于包含的内生解释变量个数，而所有排斥的外生变量都是有效工具变量，故可以用工具变量法来估计。如果结构方程的扰动项满足同方差、无自相关的古典假定，则二阶段最小二乘法（2SLS）是最有效率的工具变量法，故也是最常见的单一方程估计法，参见第 10 章。

4. 广义矩估计法

在过度识别的情况下，如果结构方程的扰动项存在异方差或自相关，则 GMM 比 2SLS 更有效率。在恰好识别的情况下，GMM 等价于 2SLS，参见第 10 章。

5. 有限信息最大似然估计法

如果假定结构方程的扰动项服从正态分布，则可以使用最大似然估计法对单一方程进行估计，这被称为“有限信息最大似然估计法”（Limited Information Maximum Likelihood Estimation，简记 LIML）。LIML 与 2SLS 在大样本下是渐近等价的，但小样本性质似乎不如 2SLS。然而，如果存在弱工具变量（weak instruments），则 LIML 比 2SLS 更稳健。

24.4 三阶段最小二乘法

在使用单一方程估计法时，由于忽略了各方程之间的联系（包括各方程扰动项之间的联

① 除非在用 OLS 估计简化式参数 π_1，π_2，π_3，π_4 时加上约束条件“$\pi_2 / \pi_1 = \pi_4 / \pi_3$”。从此例还可看出，结构参数的过度识别相当于给简化式参数附加了约束条件。

系),故不如将所有方程作为一个整体进行估计(即系统估计法)更有效率。系统估计法的缺点是,如果其中的某个方程估计得不准确,则可能影响系统中其他方程的估计。

最常见的系统估计法为"三阶段最小二乘法"(Three Stage Least Square,简记 3SLS)①。在某种意义上,3SLS 是将 2SLS 与 SUR 相结合的一种估计方法。对于一个多方程的系统,如果各方程中都不包含内生解释变量,则对每个方程进行 OLS 估计是一致的;但却不是最有效率的,因为单一方程 OLS 忽略了不同方程的扰动项之间可能存在相关性。此时,用 SUR 对整个方程系统同时进行估计是有效率的。

对于一个多方程系统,如果方程中包含内生解释变量,则对每个方程进行 2SLS 估计是一致的;但却不是最有效率的,因为单一方程 2SLS 忽略了不同方程的扰动项之间可能存在相关性。此时,用 3SLS 对整个联立方程系统同时进行估计是有效率的。

3SLS 的基本步骤如下。

前两步:对每个方程进行 2SLS 估计。

第三步:根据前两步的估计,得到对整个系统的扰动项之协方差矩阵的估计。然后,据此对整个系统进行 GLS 估计(类似于 SUR 的做法)。具体操作如下。

对于联立方程模型的第 j 个方程,忽略不在方程中的内生变量 $\boldsymbol{y}_j^*$ 与外生变量 $\boldsymbol{x}_j^*$,并同时考虑所有 T 个观测值,则可以将第 j 个方程写为

$$\underbrace{\boldsymbol{y}_j}_{T\times 1} = \underbrace{\boldsymbol{Y}_j}_{T\times M_j}\underbrace{\boldsymbol{\gamma}_j}_{M_j\times 1} + \underbrace{\boldsymbol{X}_j}_{T\times K_j}\underbrace{\boldsymbol{\beta}_j}_{K_j\times 1} + \underbrace{\boldsymbol{\varepsilon}_j}_{T\times 1} \equiv \underbrace{\boldsymbol{Z}_j}_{T\times(M_j+K_j)}\underbrace{\boldsymbol{\delta}_j}_{(M_j+K_j)\times 1} + \underbrace{\boldsymbol{\varepsilon}_j}_{T\times 1} \quad (j=1,\cdots,M) \tag{24.28}$$

其中,第 j 个方程中的解释变量为 $\boldsymbol{Z}_j \equiv (\boldsymbol{Y}_j \boldsymbol{X}_j)$(包含内生解释变量 $\boldsymbol{Y}_j$ 与外生变量 $\boldsymbol{X}_j$),$\boldsymbol{\delta}_j \equiv \begin{pmatrix}\boldsymbol{\gamma}_j \\ \boldsymbol{\beta}_j\end{pmatrix}$(同时包含内生变量与外生变量的系数)。将所有 M 个方程叠放在一起可得

$$\boldsymbol{y} \equiv \underbrace{\begin{pmatrix}\boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_M\end{pmatrix}}_{MT\times 1} = \begin{pmatrix}\boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_M\end{pmatrix}\begin{pmatrix}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_M\end{pmatrix} + \underbrace{\begin{pmatrix}\boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_M\end{pmatrix}}_{MT\times 1} \equiv \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{24.29}$$

假设 $\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0}$,$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \boldsymbol{\Sigma}\otimes\boldsymbol{I}$,其中 $\boldsymbol{X}$ 包含整个方程系统中所有的外生变量(都可以作为工具变量),$\otimes$ 为克罗内克尔乘积(参见第 23 章)。记 $\hat{\boldsymbol{Z}}_j \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}_j$ 为第 j 个方程解释变量 $\boldsymbol{Z}_j$ 对所有外生变量(工具变量)$\boldsymbol{X}$ 进行回归的拟合值(第一阶段回归),则第 j 个方程的 2SLS 估计量为

$$\hat{\boldsymbol{\delta}}_{j,2\mathrm{SLS}} \equiv (\hat{\boldsymbol{Z}}_j'\hat{\boldsymbol{Z}}_j)^{-1}\hat{\boldsymbol{Z}}_j'\boldsymbol{y}_j \tag{24.30}$$

定义 $\hat{\boldsymbol{Z}} \equiv \begin{pmatrix}\hat{\boldsymbol{Z}}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{Z}}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \hat{\boldsymbol{Z}}_M\end{pmatrix}$,则可以将所有方程的单一方程 2SLS 估计量简洁地写在一起(参见习题):

① 参见 Zellner and Theil(1962)。

$$\hat{\boldsymbol{\delta}}_{2SLS} \equiv \begin{pmatrix} \hat{\boldsymbol{\delta}}_{1,2SLS} \\ \hat{\boldsymbol{\delta}}_{2,2SLS} \\ \vdots \\ \hat{\boldsymbol{\delta}}_{M,2SLS} \end{pmatrix} = (\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y} \tag{24.31}$$

为了进行 3SLS 估计，必须先得到对协方差矩阵 $\boldsymbol{\Sigma}$ 的估计值 $\hat{\boldsymbol{\Sigma}}$。记矩阵 $\hat{\boldsymbol{\Sigma}}$ 的 (i,j) 元素为 $\hat{\sigma}_{ij}$，利用单一方程 2SLS 估计的残差可得

$$\hat{\sigma}_{ij} = \frac{1}{T}(\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{\delta}}_{i,2SLS})'(\boldsymbol{y}_j - \boldsymbol{Z}_j\hat{\boldsymbol{\delta}}_{j,2SLS}) \tag{24.32}$$

一般来说，使用 GLS 会比单一方程 2SLS 更有效率。类比 SUR，可定义 3SLS 估计量为

$$\hat{\boldsymbol{\delta}}_{3SLS} = [\hat{\boldsymbol{Z}}'(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\boldsymbol{I})\hat{\boldsymbol{Z}}]^{-1}\hat{\boldsymbol{Z}}'(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\boldsymbol{I})\boldsymbol{y} \tag{24.33}$$

对于 3SLS，也可以进行迭代，即用 3SLS 的残差重新估计协方差矩阵 $\boldsymbol{\Sigma}$，然后再使用 GLS，如此反复，直至收敛。但迭代 3SLS 并不能提高其渐近效率。

除了 3SLS 外，系统估计法还包括，“系统广义矩估计”（System GMM），以及“全信息最大似然估计法”（Full Information Maximum Likelihood Estimation，简记 FIML），参见 Greene（2012）或 Hayashi（2000）。

24.5 三阶段最小二乘法的 Stata 实例

3SLS 的 Stata 命令格式为

```
reg3 (depvar1 varlist1) (depvar2 varlist2) ... (depvarN varlistN),ols 2sls sure ireg3 exog(varlist) endog(varlist) inst(varlist)
```

其中，选择项“ols”表示进行 OLS 估计，“2sls”表示进行 2SLS 估计，“sure”表示进行 SUR 估计，而默认值为进行 3SLS 估计。选择项“ireg3”表示进行迭代式 3SLS 估计。而选择项“exog(varlist)”、“endog(varlist)”以及“inst(varlist)”则用于指定额外的外生变量，内生变量（除被解释变量以外），以及工具变量（除方程组系统自带的工具变量以外）。详见“help reg3”。

下面以数据集 klein. dta 为例。该数据集包含 1920—1941 年的以下宏观经济变量，consump（消费）、wagepriv（私企工资）、wagegovt（政府工资）、govt（政府开支）、capitall（资本存量的滞后值）。考虑以下联立方程模型：

$$\begin{cases} \text{consump}_t = \alpha_0 + \alpha_1\ \text{wagepriv}_t + \alpha_2\ \text{wagegovt}_t + u_t \\ \text{wagepriv}_t = \beta_0 + \beta_1\ \text{consump}_t + \beta_2\ \text{govt}_t + \beta_3\text{capitall}_t + v_t \end{cases} \tag{24.34}$$

其中，第一个方程以私企工资与政府工资（相当于收入）来解释消费，即消费函数；而第二个方程以消费、政府开支、资本存量的滞后值（相当于总需求）来解释私企工资。这个联立方程系统共有两个内生变量，即 consump 与 wagepriv；三个外生变量，即 wagegovt，govt 与 capitall。每个方程均包含一个内生解释变量。第一个方程排斥了两个外生变量 govt 与 capitall，即有两个工具变量可用，故为过度识别；而第二个方程排斥了一个外生变量 wagegovt，即有一个工具变量可用，故为

恰好识别。

作为参照系，首先用 OLS 对每个方程进行单一方程估计：

```
. use klein.dta,clear
. reg3 (consump wagepriv wagegovt)(wagepriv consump govt capital1),ols
. estimates store OLS
```

```
Multivariate regression
------------------------------------------------------------------------------
Equation              Obs   Parms        RMSE    "R-sq"       F-Stat         P
------------------------------------------------------------------------------
consump                22       2    1.60651    0.9567       210.15    0.0000
wagepriv               22       3   1.553524    0.9489       111.33    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
consump      |
    wagepriv |   .9918122   .0678073    14.63   0.000     .8544216    1.129203
    wagegovt |   .6780964   .2147332     3.16   0.003     .2430055    1.113187
       _cons |   14.24549   2.045098     6.97   0.000     10.10173    18.38925
-------------+----------------------------------------------------------------
wagepriv     |
     consump |   .7742524   .0654305    11.83   0.000     .6416777    .9068272
       govt  |   .4048119   .1969143     2.06   0.047     .0058257    .8037981
    capital1 |  -.0443646   .0356482    -1.24   0.221    -.1165947    .0278656
       _cons |   1.668479   6.744839     0.25   0.806    -11.99786    15.33482
------------------------------------------------------------------------------
```

其中，命令 reg3 的默认估计法为 3SLS，但加上选择项“ols”后则进行单一方程 OLS 估计。

下面进行单一方程 2SLS 估计：

```
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1),2sls
. estimates store Two_SLS
```

```
Two-stage least-squares regression
------------------------------------------------------------------------------
Equation              Obs   Parms        RMSE    "R-sq"       F-Stat         P
------------------------------------------------------------------------------
consump                22       2    1.911394    0.9388        89.83    0.0000
wagepriv               22       3    2.720166    0.8432        21.67    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
consump      |
    wagepriv |   .8012754   .1376629     5.82   0.000     .5223438    1.080207
    wagegovt |   1.029531   .3280273     3.14   0.003     .3648848    1.694178
       _cons |   19.3559    3.856336     5.02   0.000     11.54222    27.16958
-------------+----------------------------------------------------------------
wagepriv     |
     consump |   .3752562   .2848669     1.32   0.196    -.2019389    .9524514
       govt  |   1.155399   .5996727     1.93   0.062    -.0596528    2.370452
    capital1 |   .0107234    .072061     0.15   0.883    -.1352862    .1567329
       _cons |   8.443596   12.61305     0.67   0.507    -17.11287    34.00007
------------------------------------------------------------------------------
Endogenous variables:  consump wagepriv
Exogenous variables:   wagegovt govt capital1
------------------------------------------------------------------------------
```

其中，选择项“2sls”表示进行单一方程 2SLS 估计。下面进行 3SLS 估计。

```
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt
capital1),first
. estimates store Three_SLS
```

其中,选择项“first”表示显示第一阶段回归的结果(即内生解释变量对工具变量的回归)。

First-stage regressions

Source	SS	df	MS		
Model	661.220563	3	220.406854	Number of obs	22
Residual	472.554316	18	26.2530175	F(3, 18)	8.40
				Prob > F	0.0011
Total	1133.77488	21	53.9892799	R-squared	0.5832
				Adj R-squared	0.5137
				Root MSE	5.1238

consump	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wagegovt	1.397195	.7497098	1.86	0.079	-.1778867	2.972277
govt	1.214701	.6203894	1.96	0.066	-.0886885	2.518091
capital1	.084001	.1173772	0.72	0.483	-.1625994	.3306014
_cons	23.92661	22.25599	1.08	0.297	-22.83148	70.68471

Source	SS	df	MS		
Model	480.949809	3	160.316603	Number of obs	22
Residual	368.542899	18	20.4746055	F(3, 18)	7.83
				Prob > F	0.0015
Total	849.492709	21	40.4520338	R-squared	0.5662
				Adj R-squared	0.4939
				Root MSE	4.5249

wagepriv	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wagegovt	.5243061	.6620813	0.79	0.439	-.8666751	1.915287
govt	1.611224	.5478763	2.94	0.009	.4601782	2.762269
capital1	.0422453	.1036578	0.41	0.688	-.1755317	.2600222
_cons	17.42221	19.65464	0.89	0.387	-23.87065	58.71507

Three-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
consump	22	2	1.776297	0.9388	208.02	0.0000
wagepriv	22	3	2.372443	0.8542	80.04	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
consump						
wagepriv	.8012754	.1279329	6.26	0.000	.5505314	1.052019
wagegovt	1.029531	.3048424	3.38	0.001	.432051	1.627011
_cons	19.3559	3.583772	5.40	0.000	12.33184	26.37996
wagepriv						
consump	.4026076	.2567312	1.57	0.117	-.1005764	.9057916
govt	1.177792	.5421253	2.17	0.030	.1152461	2.240338
capital1	-.0281145	.0572111	-0.49	0.623	-.1402462	.0840173
_cons	14.63026	10.26693	1.42	0.154	-5.492552	34.75306

Endogenous variables: consump wagepriv
Exogenous variables: wagegovt govt capital1

上表显示,第一阶段的两个回归方程均很显著(F 统计量的 p 值很小)。

下面进行迭代式 3SLS 估计:

```
. reg3 (consump wagepriv wagegovt)(wagepriv consump govt capital1),
ireg3
. estimates store Three_SLS_iter
```

```
Iteration 1:   tolerance =    .655117
Iteration 2:   tolerance =  .00433981
Iteration 3:   tolerance =  .00004779
Iteration 4:   tolerance =  5.240e-07

Three-stage least-squares regression, iterated
------------------------------------------------------------------------------
Equation             Obs  Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
consump               22      2    1.776297    0.9388     208.02    0.0000
wagepriv              22      3    2.373113    0.8542      86.04    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
consump      |
    wagepriv |   .8012754   .1279329     6.26   0.000     .5505314    1.052019
    wagegovt |   1.029531   .3048424     3.38   0.001      .432051    1.627011
       _cons |   19.3559    3.583772     5.40   0.000     12.33184    26.37996
-------------+----------------------------------------------------------------
wagepriv     |
     consump |    .402311   .2475678     1.63   0.104    -.0829131     .887535
        govt |   1.177549   .5228599     2.25   0.024     .1527627    2.202336
    capital1 |  -.0276932    .054752    -0.51   0.613    -.1350052    .0796188
       _cons |   14.56316   9.841946     1.48   0.139    -4.726701    33.85302
------------------------------------------------------------------------------

Endogenous variables:  consump wagepriv
Exogenous variables:   wagegovt govt capital1
------------------------------------------------------------------------------
```

为了便于比较,将以上方法的系数估计值及标准误列表:

```
. esttab OLS Two_SLS Three_SLS Three_SLS_iter,r2 mtitles star(*0.1
**0.05 ***0.01)
```

	(1) OLS	(2) Two_SLS	(3) Three_SLS	(4) Three_SLS_~r
consump				
wagepriv	0.992***	0.801***	0.801***	0.801***
	(14.63)	(5.82)	(6.26)	(6.26)
wagegovt	0.678***	1.030***	1.030***	1.030***
	(3.16)	(3.14)	(3.38)	(3.38)
_cons	14.25***	19.36***	19.36***	19.36***
	(6.97)	(5.02)	(5.40)	(5.40)
wagepriv				
consump	0.774***	0.375	0.403	0.402
	(11.83)	(1.32)	(1.57)	(1.63)
govt	0.405**	1.155*	1.178**	1.178**
	(2.06)	(1.93)	(2.17)	(2.25)
capital1	-0.0444	0.0107	-0.0281	-0.0277
	(-1.24)	(0.15)	(-0.49)	(-0.51)
_cons	1.668	8.444	14.63	14.56
	(0.25)	(0.67)	(1.42)	(1.48)
N	22	22	22	22
R-sq	0.957	0.939	0.939	0.939

t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01

从上表可以看出,单一方程 2SLS,3SLS 与迭代 3SLS 的估计结果很接近,但与单一方程 OLS 的估计结果差别较大。

24.6 结构 VAR

在 20 世纪 80 年代之前,传统的联立方程模型流行一时,尤其在宏观经济领域。这些结构模型越建越大,似乎能很好地拟合样本数据;但对于样本外数据的预测能力却较弱。而且,为了识别结构方程组,常需要附加很多难以置信(incredible)的约束条件。作为解决方法,Sims(1980)提出了 VAR 模型(参见第 20 章)。然而,简化式 VAR 的脉冲响应函数并不唯一(依赖于变量次序),而且简化式 VAR 无法揭示经济结构(变量之间没有当期影响)。

为此,经济学家又试图将结构重新纳入 VAR 模型中,允许变量之间存在当期影响,形成"结构 VAR"的方法(Sims, 1981, 1986; Bernanke, 1986; Shapiro and Watson, 1988; Blanchard and Quah, 1989)。作为例子,考虑如下二元动态联立方程组(为叙述方便,忽略常数项):

$$\begin{cases} y_{1t} = -a_{12}y_{2t} + \gamma_{11}y_{1,t-1} + \gamma_{12}y_{2,t-1} + \varepsilon_{1t} \\ y_{2t} = -a_{21}y_{1t} + \gamma_{21}y_{1,t-1} + \gamma_{22}y_{2,t-1} + \varepsilon_{2t} \end{cases} \tag{24.35}$$

其中,扰动项的分布满足

$$\begin{pmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\end{pmatrix}\sim \text{iid}\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right] \tag{24.36}$$

方程组(24.35)的显著特征是在方程右边的解释变量中包含了当期变量,即 y_{1t} 的解释变量包括 y_{2t},而 y_{2t} 的解释变量也包括 y_{1t}。一般认为,方程组(24.35)来自于经济理论对于经济结构的建模,故称为"结构 VAR"(Structural VAR,简记 SVAR)。结构方程的扰动项 ε_{1t} 与 ε_{2t} 相互独立,称为"结构新息"(structural innovation)。比如,y_{1t} 为去势(detrended)的实际 GDP 对数,y_{2t} 为去势的名义货币供给对数;则结构新息的假设意味着,对产出的意外冲击(unexpected shocks to output)与对货币供给的意外冲击不相关。又比如,y_{1t} 为实际 GDP 增长率,y_{2t} 为失业率;则 ε_{1t} 与 ε_{2t} 可分别解释为需求冲击(demand shock)与供给冲击(supply shock),而需求冲击(例如消费者偏好变化)与供给冲击(例如石油价格波动)不相关(Blanchard and Quah, 1989)。

将方程组(24.35)写为矩阵形式可得

$$\underbrace{\begin{pmatrix}1 & a_{12}\\ a_{21} & 1\end{pmatrix}}_{\boldsymbol{A}}\underbrace{\begin{pmatrix}y_{1t}\\ y_{2t}\end{pmatrix}}_{\boldsymbol{y}_t}=\underbrace{\begin{pmatrix}\gamma_{11} & \gamma_{12}\\ \gamma_{21} & \gamma_{22}\end{pmatrix}}_{\boldsymbol{\Gamma}_1}\underbrace{\begin{pmatrix}y_{1,t-1}\\ y_{2,t-1}\end{pmatrix}}_{\boldsymbol{y}_{t-1}}+\underbrace{\begin{pmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\end{pmatrix}}_{\boldsymbol{\varepsilon}_t} \tag{24.37}$$

更简洁地,上式可写为

$$\boldsymbol{A}\,\boldsymbol{y}_t=\boldsymbol{\Gamma}_1\,\boldsymbol{y}_{t-1}+\boldsymbol{\varepsilon}_t \tag{24.38}$$

其中,矩阵 $\boldsymbol{A}$ 反映了 y_{1t} 与 y_{2t} 的当期互动,即内生性。假设矩阵 $\boldsymbol{A}$ 非退化,在方程(24.38)两边同时左乘 $\boldsymbol{A}^{-1}$,即可得到其相应的简化式 VAR(reduced-form VAR):

$$\boldsymbol{y}_t=\boldsymbol{A}^{-1}\boldsymbol{\Gamma}_1\,\boldsymbol{y}_{t-1}+\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}_t \tag{24.39}$$

其中,简化式 VAR 的扰动项 $\boldsymbol{u}_t\equiv\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}_t$ 为结构式 VAR 扰动项 $\boldsymbol{\varepsilon}_t$ 的线性组合。简化式扰动项的协方差矩阵为

$$\text{Var}(\boldsymbol{u}_t)=\text{Var}(\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}_t)=\boldsymbol{A}^{-1}\text{Var}(\boldsymbol{\varepsilon}_t)\boldsymbol{A}^{-1\prime} \tag{24.40}$$

其中,$\text{Var}(\boldsymbol{\varepsilon}_t)$ 为对角矩阵;但 $\text{Var}(\boldsymbol{u}_t)$ 不再是对角矩阵,故包含 3 个参数。方程(24.38)可识别的必要条件(阶条件)是,结构 VAR(24.38)的待估参数个数小于或等于简化 VAR(24.39)的待估参数个数。在本例中,SVAR 的待估参数为 8 个(6 个系数,2 个方差),而 VAR 的待估参数为 7 个(4 个系数,3 个协方差)。因此,为了识别此 SVAR,至少需要对方程(24.38)施加一个约束,比如 $a_{12}=0$(这意味着 y_{2t} 对 y_{1t} 无直接影响)。

下面,考虑一般形式的 SVAR。从 p 阶简化 VAR 出发:

$$\boldsymbol{y}_t=\boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p}+\boldsymbol{u}_t \tag{24.41}$$

其中,$\boldsymbol{y}_t$ 为 $M\times 1$ 向量;$\boldsymbol{u}_t$ 为简化式扰动项,允许存在同期相关(contemporaneous correlation)。在方程(24.41)两边同时左乘某非退化矩阵 $\boldsymbol{A}$:

$$\boldsymbol{A}\boldsymbol{y}_t=\boldsymbol{A}\boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{A}\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p}+\boldsymbol{A}\boldsymbol{u}_t \tag{24.42}$$

经移项整理可得:

$$\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{\Gamma}_1L-\cdots-\boldsymbol{\Gamma}_pL^p)\boldsymbol{y}_t=\boldsymbol{A}\boldsymbol{u}_t \tag{24.43}$$

我们希望 SVAR 的扰动项正交,一种简单的作法为令 $\boldsymbol{A}\boldsymbol{u}_t=\boldsymbol{\varepsilon}_t$,其中 $\boldsymbol{\varepsilon}_t$ 为 SVAR 的结构扰动项,不存在同期相关;但此假定可能过强(矩阵 A 来自经济理论对经济结构的建模,未必能同时使得 $\boldsymbol{A}\boldsymbol{u}_t$ 同期不相关)。更一般地,假设 $\boldsymbol{A}\boldsymbol{u}_t=\boldsymbol{B}\boldsymbol{\varepsilon}_t$,其中 $\boldsymbol{B}$ 为 $M\times M$ 矩阵;则方程(24.43)可写为

$$\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{\Gamma}_1L-\cdots-\boldsymbol{\Gamma}_pL^p)\boldsymbol{y}_t=\boldsymbol{A}\boldsymbol{u}_t=\boldsymbol{B}\boldsymbol{\varepsilon}_t \tag{24.44}$$

其中，结构扰动项 $\boldsymbol{\varepsilon}_t$ 的协方差矩阵被标准化为单位矩阵 $\boldsymbol{I}_M$。方程(24.44)称为 SVAR 的“AB 模型”(AB - Model)(Amisano and Giannini, 1997)。对于传统的联立方程模型，分析的重点在于解释变量的边际效应，故一般不要求结构扰动项正交；而对于 AB 模型，分析的重点在于正交化冲击的效应，故一般假设结构扰动项 $\boldsymbol{\varepsilon}_t$ 正交。如果令 $\boldsymbol{A}=\boldsymbol{I}_M$，则为 B 模型；如果令 $\boldsymbol{B}=\boldsymbol{I}_M$，则为 A 模型。A 模型与 B 模型都是 AB 模型的特例。在方程(24.44)两边同时左乘 $\boldsymbol{A}^{-1}$，即可得到相应的简化 VAR：

$$\boldsymbol{y}_t=\boldsymbol{\Gamma}_1\boldsymbol{y}_{t-1}+\cdots+\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p}+\underbrace{\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\varepsilon}_t}_{\boldsymbol{u}_t} \tag{24.45}$$

由于 $\boldsymbol{u}_t=\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\varepsilon}_t$，故简化式扰动项 $\boldsymbol{u}_t$ 的协方差矩阵为

$$\mathrm{Var}(\boldsymbol{u}_t)=\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{A}^{-1'} \tag{24.46}$$

对于结构 VAR 模型(24.44)，其待估参数总数为“M^2(矩阵 $\boldsymbol{A}$ 的参数个数) + M^2(矩阵 $\boldsymbol{B}$ 的参数个数) + pM^2(矩阵 $\boldsymbol{\Gamma}_1,\cdots,\boldsymbol{\Gamma}_p$ 的参数个数)”，即 $2M^2+pM^2$。另一方面，对于简化 VAR 模型(24.45)，其待估计参数总数为“$M(M+1)/2$(对称协方差矩阵 $\mathrm{Var}(\boldsymbol{u}_t)$ 的参数个数) + pM^2(矩阵 $\boldsymbol{\Gamma}_1,\cdots,\boldsymbol{\Gamma}_p$ 的参数个数)”，即 $[M(M+1)/2]+pM^2$。因此，在一般情况下，SVAR 的参数比 VAR 的参数多出 $[2M^2-M(M+1)/2]$ 个。

因此，为了识别 AB 模型(24.44)，至少需要对矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 中元素施加 $[2M^2-M(M+1)/2]$ 个约束。即使将矩阵 $\boldsymbol{A}$ 的主对角线元素都标准化为 1，也还需要附加 $[2M^2-M-M(M+1)/2]$ 个约束条件。如果正好施加如此多约束，则为恰好识别；如果施加更多约束，则为过度识别。此阶条件(order condition)为识别 AB 模型的必要条件①。

为了估计 SVAR 模型，一般假设结构扰动项 $\boldsymbol{\varepsilon}_t$ 服从多维正态分布，即 $\boldsymbol{\varepsilon}_t\sim N(\boldsymbol{0},\boldsymbol{I}_M)$，然后进行带约束条件的最大似然估计(通常无解析解，须数值计算)。虽然此 MLE 估计量在多维正态的假设下导出，但在更弱的条件下，QMLE 估计量依然是一致的。

在具体操作上，如何施加约束条件，是一门艺术。一般来说，应从经济理论或对简化式 VAR 的估计结果出发，来设置约束条件。比较常用的方法沿用了乔利斯基分解的思路，将矩阵 $\boldsymbol{A}$ 设为下三角矩阵且主对角线元素全部为 1，并将矩阵 $\boldsymbol{B}$ 设为对角矩阵，称为“乔利斯基约束”(Cholesky restrictions)。以 $M=3$ 为例，约束条件可写为

$$\boldsymbol{A}=\begin{pmatrix}1&0&0\\ \cdot&1&0\\ \cdot&\cdot&1\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}\cdot&0&0\\ 0&\cdot&0\\ 0&0&\cdot\end{pmatrix} \tag{24.47}$$

其中，缺失值“.”表示自由参数(即没有约束)。在上式中，从矩阵 $\boldsymbol{A}$ 的第一行可以看出，y_{2t} 与 y_{3t} 对 y_{1t} 无直接影响。类似地，从矩阵 $\boldsymbol{A}$ 的第二行可看出，y_{1t} 对 y_{2t} 有直接影响，但 y_{3t} 对 y_{2t} 无直接影响。最后，从矩阵 $\boldsymbol{A}$ 的第三行可看出，y_{1t} 与 y_{2t} 对 y_{3t} 都有直接影响。显然，使用乔利斯基约束来识别 SVAR，其估计结果依赖于变量的次序。因此，对于所选择的特定变量次序，需要从理论上进行说明，或进行敏感度分析，即变换变量次序，并对比其结果。

针对矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 所施加的约束也称为“短期约束”(short-run restrictions)，其 SVAR 模型称为“短期 SVAR”(short-run SVAR)。另一类约束为“长期约束”(long-run restrictions)，即对结构冲击 $\boldsymbol{\varepsilon}_t$ 对于 $\boldsymbol{y}_t$ 的长期效应进行约束，由 Blanchard and Quah(1989)所首倡；其 SVAR 模型称为

① 充分必要条件为秩条件(rank condition)，在此从略；详见 Lutkepohl(2005, p. 365)。

“长期 SVAR”(long-run SVAR)。比如,根据货币中性假说(money neutrality hypothesis),货币在长期内是中性的,即货币供给在长期内对于实际产出的累积影响为零。从简化式 VAR(24.45)出发,可推导出 SVAR 模型的脉冲响应函数。进一步,根据与第 20 章类似的推导可知,结构冲击 $\boldsymbol{\varepsilon}_t$ 对 $\boldsymbol{y}_t$ 的长期效应为

$$\boldsymbol{C} \equiv (\boldsymbol{I} - \boldsymbol{\Gamma}_1 - \cdots - \boldsymbol{\Gamma}_p)^{-1} \boldsymbol{A}^{-1} \boldsymbol{B} \tag{24.48}$$

因此,在长期内可将 SVAR 模型简洁地写为

$$\boldsymbol{y}_t = \boldsymbol{C} \boldsymbol{\varepsilon}_t \tag{24.49}$$

例如,假设 $M=2$,第一个变量为实际 GDP,第二个变量为货币供给;则可约束长期效应矩阵 $\boldsymbol{C}$ 的(1,2)元素为 0,即对第二个变量(货币供给)的结构冲击在长期内对第一个变量(实际 GDP)无作用。

24.7 SVAR 的 Stata 实例

短期 SVAR 的 Stata 命令格式为

```
svar y1 y2 y3, aconstraints(constraints_a) aeq(matrix_aeq) acns(matrix_acns) bconstraints(constraints_a) beq(matrix_aeq) bcns(matrix_acns) varconstraints(constraints_v) lags(numlist)
```

此命令的前三个选择项提供了对矩阵 $\boldsymbol{A}$ 施加约束的三种方法,其中“aconstraints(constraints_a)”表示由命令 constraint 所定义的约束;“aeq(matrix_aeq)”表示由命令 matrix 所定义的约束,该矩阵的元素取值或为实数(可以为 0),或为缺失值“.”(表示自由参数);“acns(matrix_acns)”表示由命令 matrix 所定义的跨参数约束,其矩阵元素取值或为缺失值“.”(表示自由参数),或为 0,或为正整数,而且每个正整数至少出现两次,比如,所有取值为“1”的元素都相等,而所有取值为“2”的元素也都相等,但取值为 1 的元素可以不同于取值为 2 的元素。类似地,选择项“bconstraints(constraints_a)”,“beq(matrix_aeq)”与“bcns(matrix_acns)”表示对矩阵 $\boldsymbol{B}$ 施加这三种约束。选择项“varconstraints(constraints_v)”表示对简化 VAR 的参数进行约束。最后,选择项“lags(numlist)”表示简化 VAR 的阶数,比如“lags(1/4)”表示 VAR(4);默认为“lags(1/2)”,即 VAR(2)。

下面使用 Stata 提供的数据集 lutkepohl2.dta 进行演示。该数据集包含了联邦德国的季度宏观变量(已经过季节调整):inv(投资),inc(收入),consump(消费),ln_inv(投资的对数),ln_inc(收入的对数),ln_consump(消费的对数),dln_inv(ln_inv 的一阶差分),dln_inc(ln_inc 的一阶差分),dln_consump(ln_consump 的一阶差分)。下面估计一个关于(dln_inv, dln_inc, dln_consump)的 SVAR 模型(对数一阶差分可解释为百分比变化),并对矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 施以乔利斯基约束,参见表达式(24.47)。这意味着,我们假定投资不受收入与消费的当期影响;收入受到投资的当期影响,但不受消费的当期影响;而消费受到投资与收入的当期影响[①]。

---

① 此假定并无太多理论支持,故应变换变量次序,进行敏感度分析(比如,比较不同变量排序的脉冲响应图),在此从略。

首先,定义矩阵 ***A*** 与 ***B***。

```
. use lutkepohl2.dta,clear
. matrix A = (1,0,0 \.,1,0 \.,.,1)
. matrix B = (.,0,0 \0,.,0 \0,0,.)
```

看一下定义矩阵 ***A*** 与 ***B*** 的结果。

```
. matrix list A
```

```
A[3,3]
    c1  c2  c3
r1   1   0   0
r2   .   1   0
r3   .   .   1
```

```
. matrix list B
```

```
symmetric B[3,3]
    c1  c2  c3
r1   .
r2   0   .
r3   0   0   .
```

其次,确定 VAR 模型的滞后阶数。

```
. varsoc dln_inv dln_inc dln_consump
```

```
Selection-order criteria
Sample:  1961q2 - 1982q4                        Number of obs      =        87
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	696.398				2.4e-11	-15.9402	-15.9059	-15.8552*
1	711.682	30.568	9	0.000	2.1e-11	-16.0846	-15.9477*	-15.7445
2	724.696	26.028	9	0.002	1.9e-11*	-16.1769*	-15.9372	-15.5817
3	729.124	8.8557	9	0.451	2.1e-11	-16.0718	-15.7294	-15.2215
4	738.353	18.458*	9	0.030	2.1e-11	-16.0771	-15.632	-14.9717

```
Endogenous:  dln_inv dln_inc dln_consump
 Exogenous:  _cons
```

根据 AIC 准则,可使用默认的二阶 VAR[①]。下面,使用对矩阵 ***A*** 与 ***B*** 的上述约束,估计 SVAR 模型。

```
. svar dln_inv dln_inc dln_consump, aeq(A) beq(B) nolog
```

① 根据 BIC 准则,可使用一阶 VAR。为了保证扰动项为白噪声,此处使用二阶 VAR。

```
Estimating short-run parameters

Structural vector autoregression

 ( 1)  [a_1_1]_cons = 1
 ( 2)  [a_1_2]_cons = 0
 ( 3)  [a_1_3]_cons = 0
 ( 4)  [a_2_2]_cons = 1
 ( 5)  [a_2_3]_cons = 0
 ( 6)  [a_3_3]_cons = 1
 ( 7)  [b_1_2]_cons = 0
 ( 8)  [b_1_3]_cons = 0
 ( 9)  [b_2_1]_cons = 0
 (10)  [b_2_3]_cons = 0
 (11)  [b_3_1]_cons = 0
 (12)  [b_3_2]_cons = 0

Sample:  1960q4 - 1982q4                        No. of obs      =         89
Exactly identified model                        Log likelihood  =   742.2131
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a_1_1	1	(constrained)				
/a_2_1	-.031361	.0266532	-1.18	0.239	-.0836003	.0208782
/a_3_1	-.0566791	.0187076	-3.03	0.002	-.0933453	-.0200128
/a_1_2	0	(omitted)				
/a_2_2	1	(constrained)				
/a_3_2	-.4792407	.0738282	-6.49	0.000	-.6239414	-.3345399
/a_1_3	0	(omitted)				
/a_2_3	0	(omitted)				
/a_3_3	1	(constrained)				
/b_1_1	.0425173	.0031868	13.34	0.000	.0362712	.0487633
/b_2_1	0	(omitted)				
/b_3_1	0	(omitted)				
/b_1_2	0	(omitted)				
/b_2_2	.0106908	.0008013	13.34	0.000	.0091202	.0122613
/b_3_2	0	(omitted)				
/b_1_3	0	(omitted)				
/b_2_3	0	(omitted)				
/b_3_3	.0074461	.0005581	13.34	0.000	.0063522	.0085399

上表上部列出了所有的约束条件，共有 12 个，为恰好识别(exactly identified)。上表下部列出了对矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 自由元素的估计值及标准误。其中，对 a_{21}，a_{31}，a_{32}的估计值均为负，经移项后可知，这些当期效应均为正(与理论预期相符)。估计 SVAR 的主要兴趣通常在于考察其脉冲响应函数。为此，输入如下命令，结果参见图 24.1。

```
. irf create germany,set(germany)
(file germany.irf created)
(file germany.irf now active)
(file germany.irf updated)
. irf graph sirf
```

其中，“sirf”表示“结构脉冲响应函数”(structural irf)，等价于“正交化的脉冲响应函数”(如果

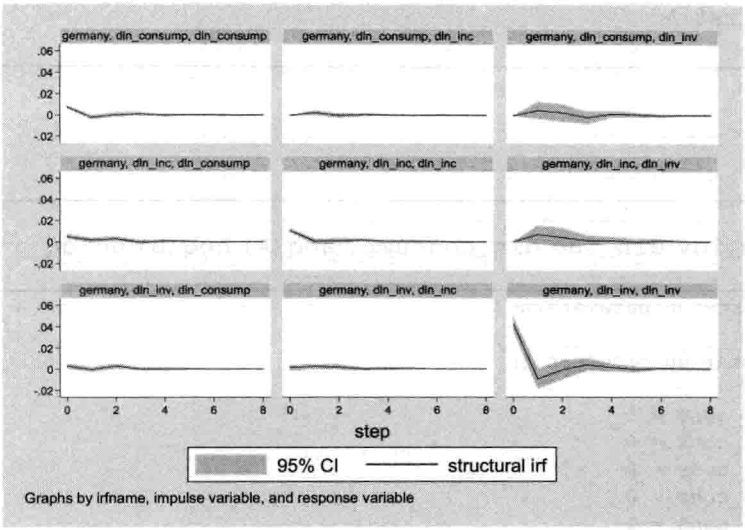


图 24.1 SVAR 的结构脉冲响应函数

输入命令“irf graph oirf”,可得到完全相同的结果)。

下面,考察 SVAR 模型的预测误差方差分解,结果参见图 24.2。

```
. irf graph sfevd
```

其中,“sfevd”表示“结构预测误差方差分解”(structural FEVD)。此命令等价于“irf graph fevd”。

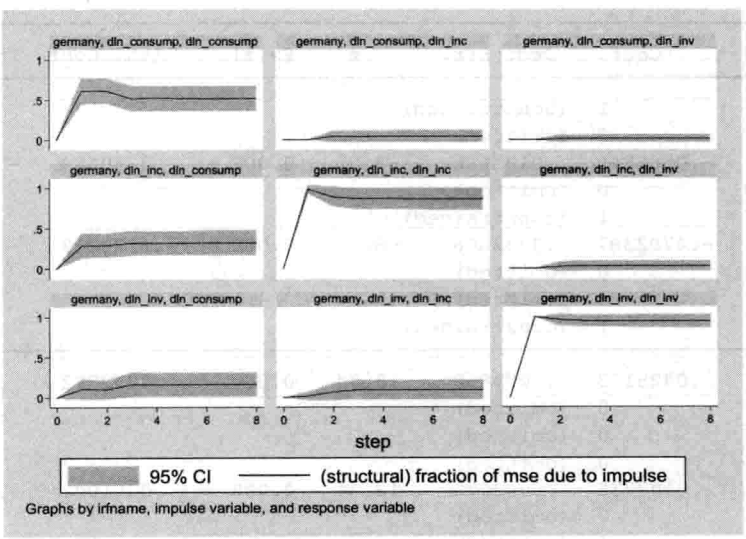


图 24.2 SVAR 的结构预测误差方差分解

以上为恰好识别的 SVAR 模型。注意到,矩阵 $\boldsymbol{A}$ 的 a_{21} 元素并不显著(p 值为 0.239)。为了演示过度识别的 SVAR 模型,下面约束 $a_{21}=0$,重新定义矩阵 $\boldsymbol{A}$ 如下(矩阵 $\boldsymbol{B}$ 的定义不变)。

```
. matrix A = (1,0,0\0,1,0\.,.,1)
```

```
. matrix B = (.,0,0\0,.,0\0,0,.)
. matrix list A
```

```
A[3,3]
    c1  c2  c3
r1   1   0   0
r2   0   1   0
r3   .   .   1
```

```
. svar dln_inv dln_inc dln_consump, aeq(A) beq(B) nolog
```

```
Estimating short-run parameters

Structural vector autoregression

 ( 1)  [a_1_1]_cons = 1
 ( 2)  [a_1_2]_cons = 0
 ( 3)  [a_1_3]_cons = 0
 ( 4)  [a_2_1]_cons = 0
 ( 5)  [a_2_2]_cons = 1
 ( 6)  [a_2_3]_cons = 0
 ( 7)  [a_3_3]_cons = 1
 ( 8)  [b_1_2]_cons = 0
 ( 9)  [b_1_3]_cons = 0
 (10)  [b_2_1]_cons = 0
 (11)  [b_2_3]_cons = 0
 (12)  [b_3_1]_cons = 0
 (13)  [b_3_2]_cons = 0

Sample:  1960q4 - 1982q4                        No. of obs      =         89
Overidentified model                            Log likelihood  =   741.5262

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /a_1_1 |           1  (constrained)
      /a_2_1 |           0  (omitted)
      /a_3_1 |  -.0566846   .0185638    -3.05   0.002    -.0930689   -.0203002
      /a_1_2 |           0  (omitted)
      /a_2_2 |           1  (constrained)
      /a_3_2 |  -.4792397   .0732606    -6.54   0.000    -.6228279   -.3356515
      /a_1_3 |           0  (omitted)
      /a_2_3 |           0  (omitted)
      /a_3_3 |           1  (constrained)
-------------+----------------------------------------------------------------
      /b_1_1 |   .0425173   .0031868    13.34   0.000     .0362712    .0487633
      /b_2_1 |           0  (omitted)
      /b_3_1 |           0  (omitted)
      /b_1_2 |           0  (omitted)
      /b_2_2 |   .0107736   .0008075    13.34   0.000     .0091909    .0123563
      /b_3_2 |           0  (omitted)
      /b_1_3 |           0  (omitted)
      /b_2_3 |           0  (omitted)
      /b_3_3 |   .0074461   .0005581    13.34   0.000     .0063522    .0085399
------------------------------------------------------------------------------
LR test of identifying restrictions:  chi2(  1)=    1.374  Prob > chi2 = 0.241
```

上表显示，共有 13 个约束条件（比恰好识别多出 1 个约束条件）。参数估计值及显著性变化不大。上表底部提供了对过度识别的似然比检验结果，该检验的原假设为“所有约束条件均成立”，结果无法拒绝此原假设。

长期 SVAR 的 Stata 命令基本格式为

```
svar y1 y2 y3,lrconstraints(constraints_lr) lreq(matrix_lreq) lrcns
(matrix_lrcns) varconstraints(constraints_v) lags(numlist)
```

此命令的前三个选择项提供了对长期效应矩阵 $\boldsymbol{C}$ 施加约束的三种方法，其含义与短期 SVAR 命令相同。其余选择项的含义也类似。下面以 Stata 提供的季度数据集 m1gdp. dta 为例。该数据集包括的主要变量为 ln_gdp（经季节调整的实际 GDP 对数）与 ln_m1（经季节调整的货币供给量 M1 对数）。根据货币中性假说，货币供给对实际产出的长期效应为 0。因此，对长期效应矩阵 $\boldsymbol{C}$ 作如下约束：

$$\boldsymbol{C}=\begin{pmatrix} \cdot & 0 \\ \cdot & \cdot \end{pmatrix} \tag{24.50}$$

首先，定义矩阵 $\boldsymbol{C}$。

```
. use m1gdp.dta,clear
. matrix C =(.,0 \.,.)
. matrix list C
```

```
C[2,2]
    c1  c2
r1   .   0
r2   .   .
```

其次，根据信息准则，确定简化 VAR 的滞后阶数。

```
. varsoc ln_gdp ln_m1
```

```
Selection-order criteria
Sample:  1960q1 - 2002q2                        Number of obs      =       170
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	13.412				.002997	-.134258	-.119288	-.097367
1	1084.15	2141.5	4	0.000	1.1e-08	-12.6841	-12.6392	-12.5734
2	1142.65	117.01	4	0.000	5.6e-09	-13.3253	-13.2504	-13.1408*
3	1149.76	14.217*	4	0.007	5.4e-09*	-13.3619*	-13.2571*	-13.1036
4	1153.27	7.0269	4	0.134	5.4e-09	-13.3561	-13.2214	-13.0241

```
Endogenous:  ln_gdp ln_m1
 Exogenous:  _cons
```

上表显示，大多数信息准则支持 VAR(3)。下面，估计长期 SVAR 模型。

```
. svar ln_gdp ln_m1,lreq(C) lags(1/3) nolog
```

```
Estimating long-run parameters

Structural vector autoregression

 ( 1)  [c_1_2]_cons = 0

Sample:  1959q4 - 2002q2                        No. of obs      =       171
Exactly identified model                        Log likelihood  =  1154.855
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/c_1_1	3.853233	.208359	18.49	0.000	3.444857	4.261609
/c_2_1	7.419362	.4066705	18.24	0.000	6.622303	8.216422
/c_1_2	0	(omitted)				
/c_2_2	.8698669	.047037	18.49	0.000	.7776761	.9620578

上表显示，此长期 SVAR 模型为恰好识别，且 $\boldsymbol{C}$ 矩阵中的所有元素都很显著。

习　题

24.1 证明方程(24.31)。

24.2 对于数据集 supDem.dta，考虑以下商品市场的供需均衡模型：

$$\begin{cases} \text{quantity}_t = \alpha_0 + \alpha_1\ \text{price}_t + \alpha_2\ \text{pcompete}_t + u_t \\ \text{quantity}_t = \beta_0 + \beta_1\ \text{price}_t + \beta_2\ \text{praw}_t + v_t \end{cases} \tag{24.51}$$

其中，第一个方程为需求函数，第二个方程为供给函数，quantity 为商品数量，price 为商品价格，pcompete 为替代商品的价格，praw 为原材料价格。

(1) 需求方程与供给方程为不可识别、恰好识别或过度识别？

(2) 使用 OLS 分别估计需求与供给方程；

(3) 使用 2SLS 分别估计需求与供给方程；

(4) 使用 3SLS 估计整个方程系统(提示：由于两个方程的被解释变量均为 quantity，故在使用 Stata 命令 reg3 时，应加上选择项“endog(price)”，指定 price 也为内生变量)。

第 25 章　非线性回归与门限回归

25.1　非线性最小二乘法

对于非线性回归模型，除了使用最大似然估计法（MLE），还可以使用“非线性最小二乘法”（Nonlinear Least Square，简记 NLS）。考虑以下非线性回归模型：

$$y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i \quad (i = 1, \cdots, n) \tag{25.1}$$

其中，$\boldsymbol{\beta}$ 为 K 维未知（真实）参数向量，而 $g(\cdot)$ 是 $\boldsymbol{\beta}$ 的非线性函数，且无法通过变量转换变为 $\boldsymbol{\beta}$ 的线性函数。如果 $g(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$，则回到古典线性回归模型。记 $\tilde{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的一个假想值（hypothetical value），其对应的残差为 $e_i = y_i - g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})$。非线性最小二乘法通过选择 $\tilde{\boldsymbol{\beta}}$，使得残差平方和最小：

$$\min_{\tilde{\boldsymbol{\beta}}} \mathrm{SSR}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})]^2 \tag{25.2}$$

最小化的一阶条件为

$$\frac{\partial \mathrm{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = -2 \sum_{i=1}^{n} [y_i - g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})] \frac{\partial g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \boldsymbol{0} \tag{25.3}$$

可以简化为

$$\sum_{i=1}^{n} [y_i - g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})] \frac{\partial g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \boldsymbol{0} \tag{25.4}$$

$$\sum_{i=1}^{n} e_i \frac{\partial g(\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \boldsymbol{0} \tag{25.5}$$

这是一个 K 个方程、K 个未知数的非线性方程组。满足这个非线性方程组的估计量被称为“非线性最小二乘估计量”，记为 $\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$。方程（25.5）表明，残差向量 $\boldsymbol{e}$ 与 $\frac{\partial g(\boldsymbol{x}, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 正交，而不是与 $\boldsymbol{x}$ 正交（线性回归的情形）。这个非线性方程组通常没有解析解，要使用数值迭代方法来求解，比如牛顿 - 拉弗森法。

例[①]　考虑如下非线性回归模型：

$$y_i = \beta_1 + \beta_2 \exp(\beta_3 x_i) + \varepsilon_i \tag{25.6}$$

---

① 此例来自 Greene（2003，p. 165）。

显然,这个模型含有三个未知参数(β_1,β_2,β_3),即 $K=3$。使用 NLS 进行估计,残差平方和为

$$\min_{\tilde{\boldsymbol{\beta}}} \mathrm{SSR}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)]^2 \tag{25.7}$$

NLS 估计量的一阶条件为

$$\frac{\partial \mathrm{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_1} = -2\sum_{i=1}^{n} [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] = 0 \tag{25.8}$$

$$\frac{\partial \mathrm{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_2} = -2\sum_{i=1}^{n} [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] \exp(\tilde{\beta}_3 x_i) = 0 \tag{25.9}$$

$$\frac{\partial \mathrm{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_3} = -2\sum_{i=1}^{n} [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] \tilde{\beta}_2 x_i \exp(\tilde{\beta}_3 x_i) = 0 \tag{25.10}$$

这是一个含有三个方程、三个未知数的非线性方程组,没有解析解,只能用数值方法求解。

NLS 的大样本性质

可以证明,如果 $\mathrm{E}(\varepsilon_i | \boldsymbol{x}_i) = 0$(即扰动项与当期解释变量不相关),再加上一些技术性条件,则 $\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$ 为真实参数 $\boldsymbol{\beta}$ 的一致估计量,且 $\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$ 服从渐近正态分布①。进一步地,如果扰动项为球形扰动项(即满足同方差与无自相关的假设),则 $\hat{\boldsymbol{\beta}}_{\mathrm{NLS}}$ 是渐近有效的(asymptotically efficient)。

25.2 非线性回归的 Stata 命令及实例

NLS 的 Stata 命令格式为

```
nl(depvar = <sexp>)[,options]
```

其中,"sexp"表示"substitutable expression",用来定义非线性回归模型。比如,对于方程(25.6),可以使用以下命令:

```
nl(y = {beta1} + {beta2} * exp({beta3 = 1} * x)),robust
```

其中,括弧"{}"中表示的是待估参数,"{beta3 = 1}"表示 β_3 的初始值为 1。如果担心存在异方差,可以使用选择项"robust"以得到稳健标准误。

下面以数据集 consumption_china.dta 为例,来估计以下非线性消费函数:

$$c_i = \beta_1 + \beta_2 y_i^{\beta_3} + \varepsilon_i \tag{25.11}$$

由于参数 β_3 未知,故这是非线性回归。如果 $\beta_3 = 1$,则为线性回归。因此,选择 $\beta_3 = 1$ 作为迭代计算的初始值:

```
. use consumption_china.dta,clear
. nl(c = {beta1} + {beta2} * y^{beta3 = 1}),nolog
```

① 参见 Cameron and Trivedi(2005), Proposition 5.6, p. 153。

Source	SS	df	MS		
Model	93414757.2	2	46707378.6	Number of obs =	29
Residual	184339.472	26	7089.97971	R-squared =	0.9980
				Adj R-squared =	0.9979
Total	93599096.7	28	3342824.88	Root MSE =	84.20202
				Res. dev. =	336.2584

c	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
/beta1	-235.1029	55.98706	-4.20	0.000	-350.186	-120.0199
/beta2	3.942834	.8282788	4.76	0.000	2.240283	5.645386
/beta3	.7634422	.0215016	35.51	0.000	.7192451	.8076393

Parameter beta1 taken as constant term in model & ANOVA table

从上表可知，非线性参数 β_3 的 p 值为 0.000，这说明非线性消费函数的模型设定有其合理性[①]，而线性消费函数过于简化了。在线性消费函数的假设下，边际消费倾向为常数。事实上，由于 $\hat{\beta}_3 = 0.76$，故边际消费倾向随着收入的增加而递减。

下面用稳健标准误重新进行估计：

```
. nl (c = {beta1} + {beta2} * y^{beta3 = 1}),r nolog
```

Nonlinear regression					Number of obs =	29
					R-squared =	0.9980
					Adj R-squared =	0.9979
					Root MSE =	84.20202
					Res. dev. =	336.2584

c	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
/beta1	-235.1029	47.5145	-4.95	0.000	-332.7704	-137.4355
/beta2	3.942834	.88375	4.46	0.000	2.12626	5.759408
/beta3	.7634422	.0229917	33.21	0.000	.7161821	.8107023

Parameter beta1 taken as constant term in model

总的来说，稳健标准误与普通标准误相差不大。这说明，大概不存在异方差问题。

25.3 门限回归

在回归分析中，我们常常关心系数估计值是否稳定，即如果将整个样本分成若干个子样本（subsample）分别进行回归，是否还能得到大致相同的估计系数。对于时间序列数据，这意味着经济结构是否随着时间的推移而改变（参见第 9 章）。对于横截面数据，比如，样本中有男性与女性，则可以根据性别将样本一分为二，分别估计男性样本与女性样本。如果用来划分样本的变量不是离散型变量而是连续型变量，比如，企业规模、人均国民收入，则需要给出一个划分的标

① 也可以引入收入的平方项，而仍然使用线性回归。

准,即“门限(门槛)值”(threshold level)。

在应用研究中,人们常常怀疑大企业与小企业的投资行为不同,那么如何区分大企业与小企业呢?另外,受到流动性约束(liquidity constraint)的企业与没有流动性约束企业的投资行为也可能不同,如何通过债务股本比(debt to equity ratio)或其他指标来区分这两类企业?再比如,发达国家与发展中国家的经济增长规律可能不同,如何通过人均国民收入这一指标来区分一个国家发达与否?总之,经济规律可能是非线性的,其函数形式可能依赖于某个变量(称为“门限变量”)而改变。

传统的做法是,由研究者主观(随意)地确定一个门限值,然后根据此门限值把样本一分为二(或分成更多子样本),既不对门限值进行参数估计,也不对其显著性进行统计检验。显然,这样得到的结果并不可靠。为此,Hansen(2000)提出“门限(门槛)回归”(threshold regression),以严格的统计推断方法对门限值进行参数估计与假设检验。

假设样本数据为$\{y_i,\boldsymbol{x}_i,q_i\}_{i=1}^n$,其中$q_i$为用来划分样本的“门限变量”(threshold variable),q_i可以是解释变量$\boldsymbol{x}_i$的一部分。考虑以下门限回归模型:

$$\begin{cases} y_i=\boldsymbol{\beta}_1'\boldsymbol{x}_i+\varepsilon_i, & \text{若 } q_i\leqslant\gamma \\ y_i=\boldsymbol{\beta}_2'\boldsymbol{x}_i+\varepsilon_i, & \text{若 } q_i>\gamma \end{cases} \tag{25.12}$$

其中,γ为待估计的门限值,$\boldsymbol{x}_i$为外生解释变量,与扰动项ε_i不相关。可以将上面这个分段函数合并写为

$$y_i=\boldsymbol{\beta}_1'\underbrace{\boldsymbol{x}_i\cdot\mathbf{1}(q_i\leqslant\gamma)}_{=\boldsymbol{z}_{i1}}+\boldsymbol{\beta}_2'\underbrace{\boldsymbol{x}_i\cdot\mathbf{1}(q_i>\gamma)}_{=\boldsymbol{z}_{i2}}+\varepsilon_i \tag{25.13}$$

其中,$\mathbf{1}(\cdot)$为示性函数,即如果括号中的表达式为真,则取值为 1;反之,取值为 0。显然,这是一个非线性回归,因为它无法写成参数$(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\gamma)$的线性函数。可以用非线性最小二乘法(NLS)来估计,即最小化残差平方和。事实上,如果γ的取值已知,则可以通过定义$\boldsymbol{z}_{i1}\equiv\boldsymbol{x}_i\cdot\mathbf{1}(q_i\leqslant\gamma)$与$\boldsymbol{z}_{i2}\equiv\boldsymbol{x}_i\cdot\mathbf{1}(q_i>\gamma)$,将方程(25.13)转化为参数为$(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$的线性回归模型:

$$y_i=\boldsymbol{\beta}_1'\boldsymbol{z}_{i1}+\boldsymbol{\beta}_2'\boldsymbol{z}_{i2}+\varepsilon_i \tag{25.14}$$

因此,在实际计算上,常分两步来最小化残差平方和。首先,给定γ的取值,对方程(25.14)使用 OLS 来估计$\hat{\boldsymbol{\beta}}_1(\gamma)$与$\hat{\boldsymbol{\beta}}_2(\gamma)$(显然$\hat{\boldsymbol{\beta}}_1$与$\hat{\boldsymbol{\beta}}_2$依赖于$\gamma$),并计算残差平方和SSR($\gamma$)(称为Concentrated Sum of Squared Residuals),也是γ的函数。其次,选择γ使得SSR(γ)最小化。注意到,给定q_i,由于示性函数$\mathbf{1}(q_i\leqslant\gamma)$与$\mathbf{1}(q_i>\gamma)$只能取值 0 或 1,故是$\gamma$的阶梯函数,而“阶梯的升降点”正好是$q_i$(只有一级“台阶”)。由此可知,SSR($\gamma$)也是$\gamma$的阶梯函数,而阶梯的升降点恰好在$\{q_i\}_{i=1}^n$不重叠的观测值上,因为如果$\gamma$取$\{q_i\}_{i=1}^n$以外的其他值,不会对子样本的划分产生影响,故不改变 SSR(γ)。因此,最多只需要考虑γ取n个值即可,即$\gamma\in\{q_1,q_2,\cdots,q_n\}$。这使得 SSR($\gamma$)的最小化计算得以简化。记最后的参数估计量为$(\hat{\boldsymbol{\beta}}_1(\hat{\gamma}),\hat{\boldsymbol{\beta}}_2(\hat{\gamma}),\hat{\gamma})$。

在一定的条件下,Hansen(2000)导出了$\hat{\gamma}$的大样本渐近分布,在此基础上构造$\hat{\gamma}$的置信区间,并对原假设“$H_0:\gamma=\gamma_0$”进行似然比检验(参见下文)。

类似地,可以考虑包含“多个门限值”(multiple thresholds)的门限回归。比如,对于门限变量q_i,假设两个门限值为$\gamma_1<\gamma_2$,则门限回归模型为

$$y_i=\boldsymbol{\beta}_1'\boldsymbol{x}_i\cdot\mathbf{1}(q_i\leqslant\gamma_1)+\boldsymbol{\beta}_2'\boldsymbol{x}_i\cdot\mathbf{1}(\gamma_1<q_i\leqslant\gamma_2)+\boldsymbol{\beta}_3'\boldsymbol{x}_i\cdot\mathbf{1}(q_i>\gamma_2)+\varepsilon_i \tag{25.15}$$

25.4 面板数据的门限回归

对于面板数据$\{y_{it},\boldsymbol{x}_{it},q_{it}:1\leqslant i\leqslant n,1\leqslant t\leqslant T\}$，其中$i$表示个体，$t$表示时间，Hansen(1999)考虑了如下的固定效应(fixed effects)门限回归模型：

$$\begin{cases}y_{it}=\mu_i+\boldsymbol{\beta}_1'\boldsymbol{x}_{it}+\varepsilon_{it},\text{若 } q_{it}\leqslant\gamma\\ y_{it}=\mu_i+\boldsymbol{\beta}_2'\boldsymbol{x}_{it}+\varepsilon_{it},\text{若 } q_{it}>\gamma\end{cases}\tag{25.16}$$

其中，q_{it}为门限变量(可以是解释变量$\boldsymbol{x}_{it}$的一部分)，γ为待估计的门限值，扰动项ε_{it}为独立同分布的。假设解释变量$\boldsymbol{x}_{it}$为外生变量，与扰动项ε_{it}不相关。因此，$\boldsymbol{x}_{it}$不包含被解释变量y_{it}的滞后值，不是动态面板(dynamic panel)。个体截距项μ_i的存在表明，这是固定效应模型。使用示性函数$\mathbf{1}(\cdot)$，可以将模型更简洁地表示为

$$y_{it}=\mu_i+\boldsymbol{\beta}_1'\boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}\leqslant\gamma)+\boldsymbol{\beta}_2'\boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}>\gamma)+\varepsilon_{it}\tag{25.17}$$

假设n较大，T较小(短面板)，故大样本的渐近理论基于“$n\to\infty$”而展开。定义$\boldsymbol{\beta}\equiv\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}$，$\boldsymbol{x}_{it}(\gamma)\equiv\begin{pmatrix}\boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}\leqslant\gamma)\\ \boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}>\gamma)\end{pmatrix}$，则方程(25.17)可进一步简化为

$$y_{it}=\mu_i+\boldsymbol{\beta}'\boldsymbol{x}_{it}(\gamma)+\varepsilon_{it}\tag{25.18}$$

对于第i位个体，将方程(25.18)两边对时间求平均可得

$$\bar{y}_i=\mu_i+\boldsymbol{\beta}'\bar{\boldsymbol{x}}_i(\gamma)+\bar{\varepsilon}_i\tag{25.19}$$

其中，$\bar{y}_i\equiv\frac{1}{T}\sum_{t=1}^{T}y_{it}$，$\bar{\boldsymbol{x}}_i(\gamma)\equiv\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_{it}(\gamma)$，$\bar{\varepsilon}_i\equiv\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}$。将方程(25.18)减去方程(25.19)，可得模型的离差形式：

$$y_{it}-\bar{y}_i=\boldsymbol{\beta}'[\boldsymbol{x}_{it}(\gamma)-\bar{\boldsymbol{x}}_i(\gamma)]+(\varepsilon_{it}-\bar{\varepsilon}_i)\tag{25.20}$$

记$y_{it}^*\equiv y_{it}-\bar{y}_i$，$\boldsymbol{x}_{it}^*(\gamma)\equiv\boldsymbol{x}_{it}(\gamma)-\bar{\boldsymbol{x}}_i(\gamma)$，$\varepsilon_{it}^*\equiv\varepsilon_{it}-\bar{\varepsilon}_i$，则可得

$$y_{it}^*=\boldsymbol{\beta}'\boldsymbol{x}_{it}^*(\gamma)+\varepsilon_{it}^*\tag{25.21}$$

仍然使用两步法进行估计。首先，给定γ的取值，用OLS对方程(25.21)进行一致估计(组内估计量)，得到估计系数$\hat{\boldsymbol{\beta}}(\gamma)$以及残差平方和SSR($\gamma$)。其次，对于$\gamma\in\{q_{it}:1\leqslant i\leqslant n,1\leqslant t\leqslant T\}$($\gamma$最多有$nT$个可能取值)，选择$\hat{\gamma}$，使得SSR($\hat{\gamma}$)最小。最后得到估计系数$\hat{\boldsymbol{\beta}}(\hat{\gamma})$。如果不希望某个子样本中的观测值过少，则可以限制$\gamma$的取值，比如不考虑$\{q_{it}\}$中最大5%或最小5%的取值。

对于是否存在“门限效应”(threshold effect)，可以检验以下原假设：

$$H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2\tag{25.22}$$

如果此原假设成立，则不存在门限效应。此时，模型简化为

$$y_{it}=\mu_i+\boldsymbol{\beta}_1'\boldsymbol{x}_{it}+\varepsilon_{it}\tag{25.23}$$

对于这个标准的固定效应面板模型，可以将其转化为离差形式，然后用OLS来估计(组内估计量)。记在“$H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$”约束下所得到的残差平方和为SSR*，以区别于无约束的残差平方和SSR($\hat{\gamma}$)。显然，SSR$^*\geqslant$SSR($\hat{\gamma}$)。如果[SSR* − SSR($\hat{\gamma}$)]越大，加上约束条件后使得SSR增大

越多，则越应该倾向于拒绝“$H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$”。

Hansen(1999)提出使用以下似然比检验(LR)统计量：

$$\mathrm{LR}\equiv[\mathrm{SSR}^{*}-\mathrm{SSR}(\hat{\gamma})]/\hat{\sigma}^2 \tag{25.24}$$

其中，$\hat{\sigma}^2\equiv\dfrac{\mathrm{SSR}(\hat{\gamma})}{n(T-1)}$为对扰动项方差的一致估计。然而，如果原假设“$H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$”成立，则不存在门限效应，也就无所谓门限值 γ 等于多少。因此，在 H_0 成立的情况下，无论 γ 取什么值，对模型都没有影响，故参数 γ 不可识别。因此，检验统计量 LR 的渐近分布并非标准的 χ^2 分布，而依赖于样本矩(sample moments)，无法将其临界值列表，但可以用自助法(bootstrap)来得到其临界值。

如果拒绝“$H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$”，则认为存在门限效应，可以进一步对门限值进行检验，即检验“$H_0:\gamma=\gamma_0$”。定义似然比检验统计量为

$$\mathrm{LR}(\gamma)\equiv[\mathrm{SSR}(\gamma)-\mathrm{SSR}(\hat{\gamma})]/\hat{\sigma}^2 \tag{25.25}$$

可以证明，在“$H_0:\gamma=\gamma_0$”成立的情况下，$\mathrm{LR}(\gamma)$的渐近分布虽然仍然是非标准的，但其累积分布函数为$(1-\mathrm{e}^{-x/2})^2$，可以直接算出其临界值。由此，可以利用统计量 $\mathrm{LR}(\gamma)$来计算 γ 的置信区间。

类似地，可以考虑多门限值的面板回归模型。以两个门限值为例：

$$y_{it}=\mu_i+\boldsymbol{\beta}_1'\boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}\leqslant\gamma_1)+\boldsymbol{\beta}_2'\boldsymbol{x}_{it}\cdot\mathbf{1}(\gamma_1<q_{it}\leqslant\gamma_2)+\boldsymbol{\beta}_3'\boldsymbol{x}_{it}\cdot\mathbf{1}(q_{it}>\gamma_2)+\varepsilon_{it} \tag{25.26}$$

其中，门限值 $\gamma_1<\gamma_2$。同样地，可以将这个模型转换为离差形式，并仍用两步法进行估计。首先，给定(γ_1,γ_2)，使用 OLS 估计离差模型，得到残差平方和 $\mathrm{SSR}(\gamma_1,\gamma_2)$。其次，选择$(\gamma_1,\gamma_2)$使得 $\mathrm{SSR}(\gamma_1,\gamma_2)$最小化。

25.5 门限回归的计算机操作

目前，尚无门限回归的官方 Stata 程序。门限回归的发明者 Bruce Hansen 在其个人网站[①]提供了进行门限回归的 Matlab 与 Gauss 程序。

习 题

25.1 使用数据集 mpyr.dta，估计货币流通速度的对数($\log v$)与名义利率(r)的以下非线性回归模型：

$$\log v_t=\beta_1+\beta_2 r_t^{\beta_3}+\varepsilon_t \tag{25.27}$$

将其结果与线性回归比较。其中，非线性参数 β_3 显著吗？

① 网址为 http://www.ssc.wisc.edu/~bhansen/，或搜索“Bruce Hansen economics”。

第 26 章　分位数回归

26.1　为什么需要分位数回归

在迄今为止的回归模型中，我们着重考察解释变量 $\boldsymbol{x}$ 对被解释变量 y 的条件期望 $\mathrm{E}(y|\boldsymbol{x})$ 的影响，实际上是均值回归。但我们真正关心的是 $\boldsymbol{x}$ 对整个条件分布 $y|\boldsymbol{x}$ 的影响，而条件期望 $\mathrm{E}(y|\boldsymbol{x})$ 只是刻画条件分布 $y|\boldsymbol{x}$ 集中趋势的一个指标而已。如果条件分布 $y|\boldsymbol{x}$ 不是对称分布（symmetric distribution），则条件期望 $\mathrm{E}(y|\boldsymbol{x})$ 很难反映整个条件分布的全貌。如果能够估计出条件分布 $y|\boldsymbol{x}$ 的若干重要的条件分位数（conditional quantiles），比如中位数（median）、1/4 分位数（lower quartile）、3/4 分位数（upper quartile），就能对条件分布 $y|\boldsymbol{x}$ 有更全面的认识。另一方面，使用 OLS 的古典“均值回归”，由于最小化的目标函数为残差平方和 $\left(\sum_{i=1}^{n} e_i^2\right)$，故容易受极端值（outliers）的影响。

为此，Koenker and Bassett（1978）提出“分位数回归”（Quantile Regression，简记 QR），使用残差绝对值的加权平均（比如，$\sum_{i=1}^{n}|e_i|$）作为最小化的目标函数，故不易受极端值影响，较为稳健。更重要的是，分位数回归还能提供关于条件分布 $y|\boldsymbol{x}$ 的全面信息。下面首先回顾有关总体分位数与样本分位数的概念。

26.2　总体分位数

假设 Y 为连续型随机变量，其累积分布函数为 $F_y(\cdot)$，则 Y 的“总体 q 分位数”（population q^{th} quantile，$0<q<1$），记为 y_q，满足以下定义式：

$$q=\mathrm{P}(Y\leqslant y_q)=F_y(y_q) \tag{26.1}$$

即总体 q 分位数 y_q 正好将总体分布分为两部分，其中小于或等于 y_q 的概率为 q，而大于 y_q 的概率为 $(1-q)$。如果 $q=1/2$，则为中位数，正好将总体分为两个相等的部分，一半在中位数之上，而另一半在中位数之下。如果 $F_y(\cdot)$ 严格单调递增，则有

$$y_q=F_y^{-1}(q) \tag{26.2}$$

其中，$F_y^{-1}(\cdot)$ 为 $F_y(\cdot)$ 的逆函数，参见图 26.1。

以上讨论的是单一变量的分位数。对于回归模型而言，记条件分布 $y|\boldsymbol{x}$ 的累积分布函数为 $F_{y|x}(\cdot)$。条件分布 $y|\boldsymbol{x}$ 的总体 q 分位数，记为 y_q，满足以下定义式：

$$q=F_{y|x}(y_q) \tag{26.3}$$

假设 $F_{y|x}(\cdot)$ 严格单调递增，则有

$$y_q=F_{y|x}^{-1}(q) \tag{26.4}$$

由于条件累积分布函数 $F_{y|x}(\cdot)$ 依赖于 $\boldsymbol{x}$，故条件分布 $y|\boldsymbol{x}$ 的总体 q 分位数 y_q 也依赖于 $\boldsymbol{x}$，可以明确地写为 $y_q(\boldsymbol{x})$，称为“条件分位数函数”(conditional quantile function)。也就是说，条件分位数 $y_q(\boldsymbol{x})$ 是解释变量 $\boldsymbol{x}$ 的函数。更进一步，对于线性回归模型而言，如果扰动项满足同方差的假定，或扰动项异方差的形式为乘积形式，则 $y_q(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的线性函数。证明如下。考虑以下模型：

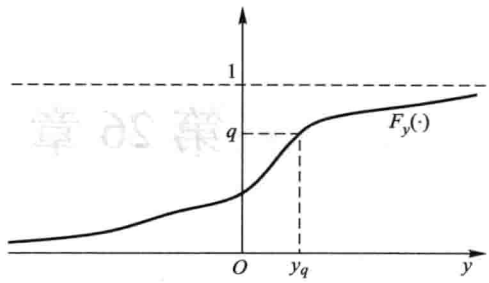


图 26.1 总体 q 分位数与累积分布函数

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + u \\ u &= \boldsymbol{x}'\boldsymbol{\alpha} \cdot \varepsilon \\ \varepsilon &\sim \text{iid}(0, \sigma^2) \end{aligned} \tag{26.5}$$

其中，不失一般性，假设 $\boldsymbol{x}'\boldsymbol{\alpha} > 0$[①]。如果 $\boldsymbol{x}'\boldsymbol{\alpha}$ 为常数，则扰动项 u 为同方差；反之，则为乘积形式的异方差。根据定义，条件分位数函数 $y_q(\boldsymbol{x})$ 满足

$$\begin{aligned} q &= \mathrm{P}\{y \leqslant y_q(\boldsymbol{x})\} && \text{(条件分位数的定义)} \\ &= \mathrm{P}\{\boldsymbol{x}'\boldsymbol{\beta} + u \leqslant y_q(\boldsymbol{x})\} && \text{(代入 } y = \boldsymbol{x}'\boldsymbol{\beta} + u\text{)} \\ &= \mathrm{P}\{u \leqslant y_q(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\} && \text{(移项)} \\ &= \mathrm{P}\{\boldsymbol{x}'\boldsymbol{\alpha} \cdot \varepsilon \leqslant y_q(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\} && \text{(代入 } u = \boldsymbol{x}'\boldsymbol{\alpha} \cdot \varepsilon\text{)} \\ &= \mathrm{P}\left\{\varepsilon \leqslant \frac{y_q(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}}{\boldsymbol{x}'\boldsymbol{\alpha}}\right\} && \text{(两边同除以 } \boldsymbol{x}'\boldsymbol{\alpha} > 0\text{)} \\ &= F_\varepsilon\left(\frac{y_q(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}}{\boldsymbol{x}'\boldsymbol{\alpha}}\right) && \text{(累积分布函数的定义)} \end{aligned}$$

其中，$F_\varepsilon(\cdot)$ 为 ε 的累积分布函数。因此，

$$\frac{y_q(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}}{\boldsymbol{x}'\boldsymbol{\alpha}} = F_\varepsilon^{-1}(q) \tag{26.6}$$

$$y_q(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{x}'\boldsymbol{\alpha}F_\varepsilon^{-1}(q) = \boldsymbol{x}'[\boldsymbol{\beta} + \boldsymbol{\alpha}F_\varepsilon^{-1}(q)] \tag{26.7}$$

从方程(26.7)可以看出，$y_q(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的线性函数。在同方差的情况下，$\boldsymbol{x}'\boldsymbol{\alpha}$ 为常数，则所有条件分位数函数 $\{y_q(\boldsymbol{x}), 0 < q < 1\}$ 的“斜率”都等于 $\boldsymbol{\beta}$，只有截距项 $\boldsymbol{x}'\boldsymbol{\alpha}F_\varepsilon^{-1}(q)$ 依赖于 q。在一般情况下，条件分位数函数的“斜率”也依赖于 q，通常记为 $\boldsymbol{\beta}_q$。

由于上面这个结果，在下文中，我们假设条件分位数函数是解释变量 $\boldsymbol{x}$ 的线性函数。

26.3 样本分位数

对于随机变量 Y，如果总体的 q 分位数 y_q 未知，则可以使用样本 q 分位数 $\hat{y}_q$ 来估计 y_q。通常的做法是，首先将样本数据 $\{y_1, y_2, \cdots, y_n\}$ 按照从小到大的顺序排列为 $\{y_{(1)}, y_{(2)}, \cdots, y_{(n)}\}$，则 $\hat{y}_q$ 等于第 $[nq]$ 个最小观测值，其中 n 为样本容量，$[nq]$ 表示大于或等于 nq 并离 nq 最近的正整

① 如果 $\boldsymbol{x}'\boldsymbol{\alpha} < 0$，则可以把 $\boldsymbol{x}'\boldsymbol{\alpha} \cdot \varepsilon$ 写为 $(-\boldsymbol{x}'\boldsymbol{\alpha}) \cdot (-\varepsilon)$。

数。比如，$n=97$，$q=0.25$，则$[nq]=[97\times0.25]=[24.25]=25$。

但是，上述样本分位数的计算方法不容易推广到回归模型的情形。一种更方便的等价方法是，将样本分位数看成某个最小化问题的解。事实上，样本均值也可以看成是最小化残差平方和问题的最优解，即

$$\min_{\mu}\sum_{i=1}^{n}(y_i-\mu)^2\Rightarrow\mu=\bar{y}\equiv\frac{1}{n}\sum_{i=1}^{n}y_i \tag{26.8}$$

类似地，样本中位数可以视为是“最小化残差绝对值之和”问题的最优解，即

$$\min_{\mu}\sum_{i=1}^{n}|y_i-\mu|\Rightarrow\mu=\text{median}\{y_1,y_2,\cdots,y_n\} \tag{26.9}$$

为什么求解这个最小化问题会得到样本中位数呢？因为只要上式中μ的取值偏离中位数，就会使得残差绝对值之和上升。直观来看，考虑一个样本容量为99的样本，假设其样本中位数（即第50个最小观测值）为10，参见图26.2。假设第51个最小观测值为12。如果在上述最小化问题中，让$\mu=12$而不是10，则对于前50个观测值而言，其对应的残差绝对值$|y_i-\mu|$都将增加2；而对于后49个观测值而言，其对应的残差绝对值$|y_i-\mu|$都将减少2。故总变动为$(50\times2)-(49\times2)=2$，故第51个最小观测值不如第50个最小观测值（中位数）更能使目标函数最小化。根据同样的道理，第49个最小观测值也不如第50个最小观测值。由此可知（严格证明见下文），第50个最小观测值（中位数）是最优解。

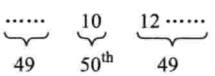


图26.2

命题 可以将样本q分位数视为以下最小化残差绝对值的加权平均问题的最优解：

$$\min_{\mu}\sum_{i:y_i\geq\mu}^{n}q|y_i-\mu|+\sum_{i:y_i<\mu}^{n}(1-q)|y_i-\mu|\Rightarrow\mu=\hat{y}_q \tag{26.10}$$

例 如果$q=1/4$，则满足“$y_i\geq\mu$”条件的观测值只得到1/4的权重，而满足“$y_i<\mu$”条件的其余观测值则得到3/4的权重。直观来看，因为估计的是1/4分位数（位于总体的底部），故较大的观测值得到的权重较小，而较小的观测值得到的权重较大。

证明：将目标函数(26.10)中的绝对值去掉可得

$$\min_{\mu}\quad\sum_{i:y_i\geq\mu}^{n}q(y_i-\mu)+\sum_{i:y_i<\mu}^{n}(1-q)(\mu-y_i) \tag{26.11}$$

对μ求一阶导数可得

$$\sum_{i:y_i\geq\mu}^{n}q(-1)+\sum_{i:y_i<\mu}^{n}(1-q)=0 \tag{26.12}$$

假设$y_{(k)}<\mu\leqslant y_{(k+1)}$，其中$y_{(k)}$为第$k$个最小观测值，则共有$k$个观测值满足“$y_i<\mu$”，$(n-k)$个观测值满足“$y_i\geq\mu$”，故

$$-(n-k)q+k(1-q)=0 \tag{26.13}$$

经整理可得

$$k=nq \tag{26.14}$$

当然，k还必须是整数。由此可知，最优解$\mu=y_{([nq])}=\hat{y}_q$，即样本分位数。为了证明最小化问题的二阶条件也满足，只要说明目标函数为凸函数即可。为此，定义函数$\rho_q(\cdot)$为

$$\rho_q(y_i-\mu)\equiv\begin{cases}q|y_i-\mu|,\text{若 }y_i\geq\mu\\(1-q)|y_i-\mu|,\text{若 }y_i<\mu\end{cases} \tag{26.15}$$

函数$\rho_q(y_i-\mu)$的形状如图26.3，故也称为“倾斜的绝对值函数”（tilted absolute value function）或“打钩函数”（check function）。从图形易知，$\rho_q(\cdot)$为凸函数。而目标函数(26.11)可以写为$\sum_{i=1}^{n}\rho_q(y_i-\mu)$，即$n$个凸函数之和，故仍然是凸函数。

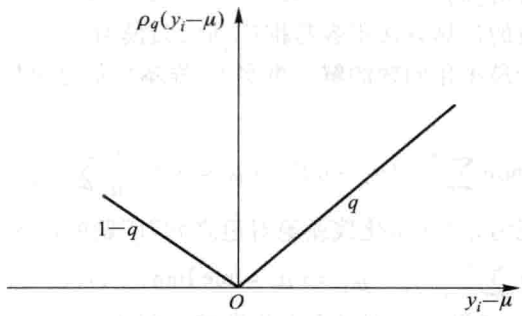


图 26.3 打钩函数 $\rho_q(\cdot)$及其斜率

26.4 分位数回归的估计方法

下面将单变量情形下对样本分位数的估计方法推广到线性回归模型。假设条件分布 $y|\boldsymbol{x}$ 的总体 q 分位数 $y_q(\boldsymbol{x})$是 $\boldsymbol{x}$ 的线性函数,即

$$y_q(\boldsymbol{x}_i)=\boldsymbol{x}_i'\boldsymbol{\beta}_q \tag{26.16}$$

其中,$\boldsymbol{\beta}_q$ 被称为"q 分位数回归系数",其估计量 $\hat{\boldsymbol{\beta}}_q$ 可以由以下最小化问题来定义:

$$\min_{\boldsymbol{\beta}_q}\sum\nolimits_{i:y_i\geqslant \boldsymbol{x}_i'\boldsymbol{\beta}_q}^{n} q\,|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_q|+\sum\nolimits_{i:y_i<\boldsymbol{x}_i'\boldsymbol{\beta}_q}^{n}(1-q)\,|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_q| \tag{26.17}$$

如果 $q=1/2$,则为"中位数回归"(median regression)。此时,目标函数简化为

$$\min_{\boldsymbol{\beta}_q}\sum\nolimits_{i=1}^{n}|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_q| \tag{26.18}$$

故中位数回归也被称为"最小绝对离差估计量"(Least Absolute Deviation Estimator,简记 LAD)。显然,它比均值回归(OLS)更不易受到极端值的影响,故更加稳健。

由于分位数回归的目标函数带有绝对值,不可微分,故通常使用线性规划(linear programming)[①]的方法来计算 $\hat{\boldsymbol{\beta}}_q$。可以证明,样本分位数回归系数 $\hat{\boldsymbol{\beta}}_q$ 是总体分位数回归系数 $\boldsymbol{\beta}_q$ 的一致估计量,而且 $\hat{\boldsymbol{\beta}}_q$ 服从渐近正态分布,即

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_q-\boldsymbol{\beta}_q)\xrightarrow{d}N(\boldsymbol{0},\mathrm{Avar}(\hat{\boldsymbol{\beta}}_q)) \tag{26.19}$$

其中,渐近方差 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_q)=\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$(形式上仍为夹心估计量),$\boldsymbol{A}\equiv\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum\nolimits_{i=1}^{n}f_{u_q}(0|\boldsymbol{x}_i)\,\boldsymbol{x}_i\boldsymbol{x}_i'$,$\boldsymbol{B}\equiv\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum\nolimits_{i=1}^{n}q(1-q)\,\boldsymbol{x}_i\boldsymbol{x}_i'$,而 $f_{u_q}(0|\boldsymbol{x}_i)$ 是扰动项 $u_q\equiv y-\boldsymbol{x}'\boldsymbol{\beta}_q$ 的条件密度函数在 $u_q=0$ 处的取值。因此,要计算 $\hat{\boldsymbol{\beta}}_q$ 的协方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_q)$,首先要估计 $f_{u_q}(0|\boldsymbol{x}_i)$。这是 Stata 的默认方法。Stata 也提供自助法作为计算 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_q)$ 的另一方法。

对于 q 分位数回归,可使用准 R^2 度量其拟合优度,其定义为:

① 即目标函数与约束条件均为线性函数的最优化问题。

$$1-\frac{\sum_{i:y_i\geqslant x_i'\hat{\boldsymbol{\beta}}_q}^{n} q\,|y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_q|+\sum_{i:y_i<x_i'\hat{\boldsymbol{\beta}}_q}^{n}(1-q)\,|y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_q|}{\sum_{i:y_i\geqslant \hat{y}_q}^{n} q\,|y_i-\hat{y}_q|+\sum_{i:y_i<\hat{y}_q}^{n}(1-q)\,|y_i-\hat{y}_q|} \tag{26.20}$$

其中，$\hat{y}_q$ 为样本 q 分位数，上式第二项的分子为 q 分位数回归目标函数的最小值(sum of weighted deviations about estimated quantiles)，而分母为“sum of weighted deviations about raw quantiles”。

例 Buchinsky(1994)使用分位数回归研究 1963—1987 年间美国工资结构的变化。

例 张车伟、薛欣欣(2008)使用分位数回归分析国有与非国有部门工资差异的决定因素。

26.5 分位数回归的 Stata 命令及实例

Stata 提供了以下有关分位数回归的命令。

(1) 只作一个分位数回归，使用默认的协方差矩阵计算方法[①]

```
qreg y x1 x2 x3  (默认为中位数回归)
qreg y x1 x2 x3,q(#)  (#分位数回归)
```

(2) 只作一个分位数回归，使用自助法计算协方差矩阵

```
set seed #  (指定产生随机数的种子，以便每次均能得到同样的结果)
bsqreg y x1 x2 x3,reps(#) q(#)  (自助法重复#次，#分位数回归)
```

(3) 同时作多个分位数回归(simultaneous quantile regressions)，使用自助法计算协方差矩阵

```
sqreg y x1 x2 x3,q(.25 .5 .75) reps(#)  (同时计算 0.25,0.5 与 0.75 分
                                         位数回归，自助法重复#次)
test [q25=q50=q75]: x1  (检验这三个分位数回归 x1 的系数是否相等)
```

(4) 将不同分位数回归的系数及其置信区间进行画图比较

首先，下载非官方命令 grqreg (表示 graph quantile regression)：

```
net install grqreg.pkg  (下载安装命令 grqreg)
```

然后进行如下操作：

```
set seed #
bsqreg y x1 x2 x3,reps(#) q(.5)  (为了得到自助标准误而先作中位数回归)
grqreg,cons ci ols olsci
```

其中，选择项“cons”表示对常数项也进行比较，选择项“ci”表示包括估计系数的 95% 置信区间，选择项“ols”表示提供 OLS 估计系数作为参照系，选择项“olsci”表示提供 OLS 估计系数的 95% 置信区间。

下面以数据集 grilic. dta 为例(参见第 10 章)。作为参照系，首先进行 OLS 回归[②]：

```
. use grilic.dta,clear
. reg lw s iq expr tenure rns smsa,r
```

---

① 也可使用非官方命令 qreg2，下载方法为“ssc install qreg2”。该命令提供异方差稳健的标准误。

② 此处复制了第 10 章的结果，且不考虑内生解释变量的问题。

```
Linear regression                                   Number of obs =     758
                                                    F(  6,   751) =   71.89
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.3600
                                                    Root MSE      =  .34454
```

lw	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	.0927874	.0069763	13.30	0.000	.0790921	.1064826
iq	.0032792	.0011321	2.90	0.004	.0010567	.0055016
expr	.0393443	.0066603	5.91	0.000	.0262692	.0524193
tenure	.034209	.0078957	4.33	0.000	.0187088	.0497092
rns	-.0745325	.0299772	-2.49	0.013	-.1333815	-.0156834
smsa	.1367369	.0277712	4.92	0.000	.0822186	.1912553
_cons	3.895172	.1159286	33.60	0.000	3.667589	4.122754

下面进行中位数回归：

```
. qreg lw s iq expr tenure rns smsa,nolog
```

```
Median regression                                   Number of obs =     758
  Raw sum of deviations   261.36 (about 5.684)
  Min sum of deviations 203.7577                    Pseudo R2     =  0.2204
```

lw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	.1014672	.0085368	11.89	0.000	.0847084	.1182261
iq	.0045716	.0013901	3.29	0.001	.0018426	.0073005
expr	.0359681	.0080899	4.45	0.000	.0200865	.0518497
tenure	.0425991	.0098489	4.33	0.000	.0232645	.0619338
rns	-.0362275	.0369023	-0.98	0.327	-.1086715	.0362164
smsa	.1318406	.0358176	3.68	0.000	.0615262	.202155
_cons	3.629197	.1400148	25.92	0.000	3.354331	3.904064

从上表可知，增加一年教育(s)能够使工资的中位数增加 10.1%，略大于对工资平均数的影响(OLS 系数估计值为 9.3%)。

下面使用自助法来计算分位数回归的标准误。为便于复制结果，指定随机数的种子：

```
. set seed 10101
. bsqreg lw s iq expr tenure rns smsa,reps(400) q(.5)
```

```
(fitting base model)
(bootstrapping ..................................................................
> ..................................................................
> ..................................................................
> ..................................................................)
Median regression, bootstrap(400) SEs               Number of obs =     758
  Raw sum of deviations   261.36 (about 5.684)
  Min sum of deviations 203.7577                    Pseudo R2     =  0.2204
```

lw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	.1014672	.0094649	10.72	0.000	.0828863	.1200481
iq	.0045716	.001477	3.10	0.002	.0016719	.0074712
expr	.0359681	.0093081	3.86	0.000	.0176951	.0542411
tenure	.0425991	.0096584	4.41	0.000	.0236384	.0615599
rns	-.0362275	.0472525	-0.77	0.444	-.1289901	.0565351
smsa	.1318406	.0368775	3.58	0.000	.0594454	.2042359
_cons	3.629197	.1736001	20.91	0.000	3.288398	3.969997

对比自助标准误与 Stata 的默认标准误可知，二者相差不大。

也可以同时估计多个分位数回归，比如，1/10，5/10 与 9/10 分位数：

```
. set seed 10101
. sqreg lw s iq expr tenure rns smsa,reps(400) q(.1 .5 .9) nodots
```

其中，选择项“nodots”表示不显示自助抽样过程中的点。

```
Simultaneous quantile regression                    Number of obs =       758
  bootstrap(400) SEs                                .10 Pseudo R2 =    0.1580
                                                    .50 Pseudo R2 =    0.2204
                                                    .90 Pseudo R2 =    0.2230
```

lw	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
s	.0761644	.0139931	5.44	0.000	.0486941	.1036347
iq	.0052221	.0019361	2.70	0.007	.0014213	.009023
expr	.0336056	.0108269	3.10	0.002	.0123511	.0548601
tenure	.0335323	.0182892	1.83	0.067	-.0023717	.0694363
rns	-.0730452	.0527144	-1.39	0.166	-.1765304	.03044
smsa	.1328733	.0482216	2.76	0.006	.0382082	.2275385
_cons	3.493085	.2372351	14.72	0.000	3.027363	3.958808
q50						
s	.1014672	.0094601	10.73	0.000	.0828958	.1200387
iq	.0045716	.0013566	3.37	0.001	.0019084	.0072348
expr	.0359681	.0103368	3.48	0.001	.0156757	.0562605
tenure	.0425991	.0100634	4.23	0.000	.0228434	.0623549
rns	-.0362275	.0456606	-0.79	0.428	-.1258652	.0534101
smsa	.1318406	.0365317	3.61	0.000	.0601242	.2035571
_cons	3.629197	.1687068	21.51	0.000	3.298005	3.96039
q90						
s	.0825558	.0115389	7.15	0.000	.0599035	.1052082
iq	.0042117	.001989	2.12	0.035	.0003071	.0081163
expr	.0484154	.01051	4.61	0.000	.027783	.0690478
tenure	.0256796	.008211	3.13	0.002	.0095603	.0417989
rns	-.072669	.0525121	-1.38	0.167	-.175757	.0304191
smsa	.1281147	.0520661	2.46	0.014	.0259023	.2303271
_cons	4.36716	.165961	26.31	0.000	4.041358	4.692963

进一步，可以检验在以上三个分位数回归中，教育年限(s)的系数是否相等：

```
. test [q10 = q50 = q90]: s
```

```
 ( 1)  [q10]s - [q50]s = 0
 ( 2)  [q10]s - [q90]s = 0

       F(  2,   751) =    2.32
            Prob > F =    0.0989
```

结果表明，可以在 10% 的显著性水平上认为，以上分位数回归系数不完全相等。

为了便于比较，下面把 OLS 与“1/10，5/10，9/10 分位数”的系数估计值及标准误列表：

```
. qui reg lw s iq expr tenure rns smsa
. est sto OLS
. qui qreg lw s iq expr tenure rns smsa,q(.1)
. est sto QR_10
. qui qreg lw s iq expr tenure rns smsa,q(.5)
```

```
. est sto QR_50
. qui qreg lw s iq expr tenure rns smsa,q(.9)
. est sto QR_90
. esttab OLS QR_10 QR_50 QR_90,se mtitles star(* 0.1 ** 0.05 *** 0.01)
```

	(1) OLS	(2) QR_10	(3) QR_50	(4) QR_90
s	0.0928***	0.0762***	0.101***	0.0826***
	(0.00667)	(0.0154)	(0.00854)	(0.0117)
iq	0.00328***	0.00522**	0.00457***	0.00421**
	(0.00108)	(0.00237)	(0.00139)	(0.00206)
expr	0.0393***	0.0336***	0.0360***	0.0484***
	(0.00631)	(0.0113)	(0.00809)	(0.0135)
tenure	0.0342***	0.0335**	0.0426***	0.0257
	(0.00771)	(0.0144)	(0.00985)	(0.0171)
rns	-0.0745***	-0.0730	-0.0362	-0.0727
	(0.0288)	(0.0602)	(0.0369)	(0.0520)
smsa	0.137***	0.133**	0.132***	0.128**
	(0.0279)	(0.0576)	(0.0358)	(0.0505)
_cons	3.895***	3.493***	3.629***	4.367***
	(0.109)	(0.241)	(0.140)	(0.226)
N	758	758	758	758

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

以教育投资的回报率(s 的系数)为例。上表显示,随着分位数的增加(1/10→5/10→9/10),教育年限(s)的分位数回归系数呈现先升后降的趋势(7.6%→10.1%→8.3%)。这表明,教育年限对工资的条件分布的两端之影响小于对其中间部分的影响。也就是说,增加教育年限对于低工资者与高工资者的影响都比较小,而最大受益者为中间阶层。

另一方面,估计系数的标准误则呈现先降后升的趋势(0.015→0.009→0.012)。这说明,对于条件分布两端的分位数回归系数的估计较不准确。

进一步,把分位数回归系数随着分位数的变化情形更直观地图示,结果如图 26.4。

```
. qui bsqreg lw s iq expr tenure rns smsa,q(.50) reps(400)
. grqreg,cons ci ols olsci
```

以图 26.4 第一行第二列的小图为例。该图显示,随着分位数的变化,教育年限(s)的分位数回归系数(即教育回报率)的变化。此图的基本形状印证了在前面表格中,教育年限的分位数回归系数先升后降的格局。另外,上图还显示,在条件分布的两端,95% 的置信区间通常变得更宽了(因为系数估计值的标准误变大了)。

目前,分位数回归仍是一个活跃的前沿领域,特别是分位数回归与各种类型的数据相结合,

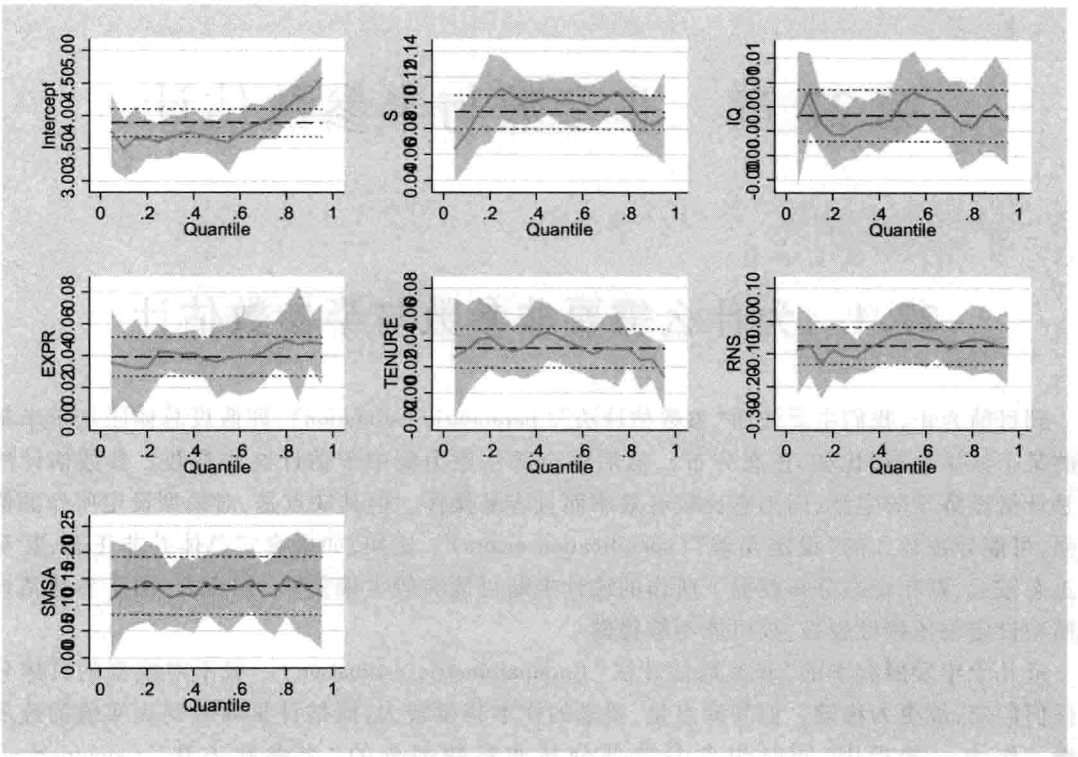


图 26.4　分位数回归系数的变化

比如面板分位数回归(Koenker, 2004)、单位根分位数回归(Koenker and Xiao, 2004)、分位数自回归(Koenker and Xiao, 2006)等。

习　题

26.1　参照本章实例,使用数据集 nerlove. dta 对成本函数进行分位数回归(参见第 4 章)。

26.2　参照本章实例,使用数据集 hprice2a. dta 对房价的决定因素进行分位数回归(参见第 7 章)。

第27章　非参数与半参数估计

27.1　为什么需要非参数与半参数估计

到目前为止，我们主要使用“参数估计法”（parametric estimation），即假设总体服从带未知参数的某个具体分布（比如，正态分布），然后将全部注意力集中于估计这些参数。参数估计法依然是计量经济学的主流，因为它比较有效率而且容易操作。但其缺点是，对模型设定所作的假定较强，可能导致较大的“设定误差”（specification errors）。比如，如果真实总体并非正态，甚至偏离正态较远，则在正态分布前提下所作的统计推断可能有较大偏差[①]。换言之，由于参数估计法对模型设定的依赖性较强，故可能不够稳健。

近几十年发展起来的“非参数估计法”（nonparametric estimation）一般不对模型的具体分布作任何假定，故更为稳健。但其缺点是，要求的样本容量较大，而估计量收敛到真实值的速度也较慢。作为一种折中，同时包含参数部分与非参数部分的“半参数方法”（semiparametric estimation）应运而生，它降低了对样本容量的要求，又具有一定的稳健性。总之，非参数及半参数方法与传统的参数估计法是互补关系，当后者不太适用时，则可以考虑前者。

27.2　对密度函数的非参数估计

在统计学中，经常需要根据样本数据来推断总体的分布，即密度函数。如果采用参数估计法，则要先对总体分布的具体形式进行假定。比如，假设总体服从正态分布 $N(\mu,\sigma^2)$，然后估计参数 (μ,σ^2)，就可以得到对密度函数的估计。如上所述，如果真实总体与正态分布相去甚远，则根据参数估计法作出的统计推断可能有较大偏差。

在估计密度函数时，如果不假设总体分布的具体形式，则为非参数方法。最原始的非参数方法是画直方图（histogram），即将数据的取值范围等分为若干组（bin），然后计算数据落入每一组的频率，以此画图，作为对密度函数的估计。用直方图来估计密度函数的缺点是，即使随机变量是连续的，直方图也始终是不连续的阶梯函数。

为了得到对密度函数的光滑估计，Rosenblatt（1956）提出“核密度估计法”（kernel density estimation）。为此，首先考察直方图的数学本质，然后再推广到核密度估计。

假设要估计连续型随机变量 x 在 x_0 处的概率密度 $f(x_0)$。由于概率密度 $f(x_0)$ 是累积分布函数 $F(x)$ 在 x_0 处的导数，根据微积分中导数的定义：

---

① 参见 DiNardo and Tobias(2001)的例子。

$$f(x_0)=\lim_{h\to 0}\frac{F(x_0+h)-F(x_0-h)}{2h}$$
$$=\lim_{h\to 0}\frac{\mathrm{P}(x_0-h<x<x_0+h)}{2h} \tag{27.1}$$

其中，区间(x_0-h,x_0+h)为在x_0附近的小邻域。对于样本$\{x_1,x_2,\cdots,x_n\}$，可以用数据落入区间(x_0-h,x_0+h)的频率来估计概率$\mathrm{P}(x_0-h<x<x_0+h)$，得到以下直方图估计量：

$$\hat{f}_{\mathrm{HIST}}(x_0)=\frac{\sum_{i=1}^{n}\mathbf{1}(x_0-h<x_i<x_0+h)/n}{2h}$$
$$=\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\cdot\mathbf{1}\left\{\left|\frac{x_i-x_0}{h}\right|<1\right\} \tag{27.2}$$

其中，$\mathbf{1}(\cdot)$为示性函数，即当括弧中的表达式为真时，取值为1；反之，则取值为0。从方程(27.2)的第二个表达式可以清楚地看出，直方图估计量$\hat{f}_{\mathrm{HIST}}(x_0)$对于所有在区间$(x_0-h,x_0+h)$内的观测值都给予相同权重，而对于在此区间以外的观测值则权重为0。这个区间的半径h定义了“在x_0附近邻域的大小”(size of neighborhood around x_0)，被称为“带宽”(bandwidth)。这个区间的直径$2h$则被称为“窗宽”(window width)。

直方图之所以得到不光滑的密度估计，根本原因是由于它使用了示性函数作为“权重函数”(weighting function)，以及它的各组之间不允许交叠。为了得到光滑的密度估计，核密度估计法使用更一般的权重函数，并允许各组之间交叠。核密度估计量为

$$\hat{f}(x_0)=\frac{1}{nh}\sum_{i=1}^{n}K[(x_i-x_0)/h] \tag{27.3}$$

其中，函数$K(\cdot)$称为“核函数”(kernel function)，本质上就是权重函数。带宽h越大，在x_0附近邻域越大，则估计的密度函数$\hat{f}(x)$越光滑①，故称带宽h为“光滑参数”(smoothing parameter)。一般假设核函数$K(z)$满足以下性质：

(i) $K(z)$连续且关于原点对称(偶函数)；

(ii) $\int_{-\infty}^{+\infty}K(z)\mathrm{d}z=1$，$\int_{-\infty}^{+\infty}zK(z)\mathrm{d}z=0$，$\int_{-\infty}^{+\infty}|K(z)|\mathrm{d}z<+\infty$；

(iii) 或者①存在$z_0>0$，使得当$|z|>z_0$时，$K(z)=0$；或者②当$|z|\to+\infty$时，$|z|K(z)\to 0$；

(iv) $\int_{-\infty}^{+\infty}z^2K(z)\mathrm{d}z=\gamma$，其中$\gamma$为常数。

条件(ii)要求核函数的曲线下面积(积分)为1(将核函数标准化)，并满足一些有界条件(boundedness conditions)。条件(iii)①比条件(iii)②更强，在实践中，常常采用条件(iii)①，即如果超出某个邻域范围$[-z_0,z_0]$，则权重变为0。不失一般性，常将邻域$[-z_0,z_0]$标准化为$[-1,1]$。条件(iv)也是一个有界条件(在某些证明中使用)。

比较常见的核函数参见表27.1。其中，均匀核也用于直方图，只是在用均匀核进行核密度估计时并不固定分组，而在每个点上进行估计。最为流行的核函数为二次核(也称Epanechnikov核，参见图27.1)与高斯核。除高斯核满足条件(iii)②外，表27.1所列其他核函数均满足条件(iii)①。这些核函数的共同特点是，离原点越近，则核函数取值越大，并在原点处达到最大值；

① 比如，在直方图的情形下，如果只分一组，则估计的密度函数将是无穷光滑的均匀分布。然而，如果带宽h很大，即只分很少的几个组，则直方图将很难反映真实分布的概貌(即过度光滑)。

这意味着,越近的点给予的权重越大。

表 27.1 常用的核函数

核函数名称	核函数的数学形式	δ
均匀核(uniform or rectangular)	$\frac{1}{2}\cdot\mathbf{1}(\lvert z\rvert<1)$	1.351 0
三角核(triangular or Bartlett)	$(1-\lvert z\rvert)\cdot\mathbf{1}(\lvert z\rvert<1)$	—
伊番科尼可夫核(Epanechnikov)① 或二次核(quadratic)	$\frac{3}{4}(1-z^2)\cdot\mathbf{1}(\lvert z\rvert<1)$	1.718 8
四次核(quartic) 或双权核(biweight)	$\frac{15}{16}(1-z^2)^2\cdot\mathbf{1}(\lvert z\rvert<1)$	2.036 2
三权核(Triweight)	$\frac{35}{32}(1-z^2)^3\cdot\mathbf{1}(\lvert z\rvert<1)$	2.312 2
三三核(Tricubic)	$\frac{70}{81}(1-\lvert z\rvert^3)^3\cdot\mathbf{1}(\lvert z\rvert<1)$	—
高斯核(Gaussian or Normal)	$\frac{1}{\sqrt{2\pi}}\exp\{-z^2/2\}$	0.776 4

注:其中 δ 为用来计算"Silverman 嵌入估计"的常数,参见下文。

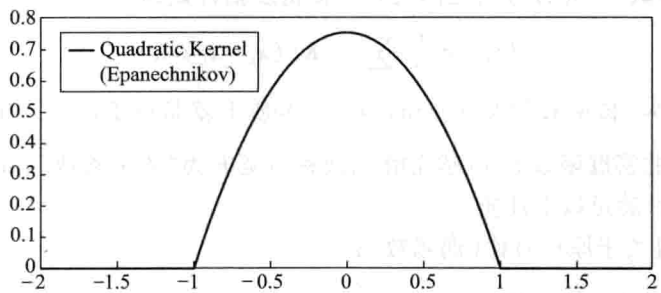


图 27.1 二次核(Epanechnikov 核)

给定核函数 $K(\cdot)$ 与带宽 h,就可以估计核密度 $\hat{f}(x_0)$。在 Stata 中,默认设置为在等距离的 $\min(n,50)$ 个点来计算 $\hat{f}(x_0)$,然后连成光滑的密度函数。当然,也可以人为指定估计的点数。

27.3 核密度估计的性质

由于核密度估计使用了在 x_0 附近的点 x 来估计 $\hat{f}(x_0)$,而一般地,如果 $x\neq x_0$,则 $f(x)\neq f(x_0)$,故核密度估计通常是有偏的。通过二阶泰勒近似,可以证明(见附录),其偏差为

$$\mathrm{Bias}(x_0)\equiv \mathrm{E}[\hat{f}(x_0)]-f(x_0)\approx\frac{1}{2}h^2f''(x_0)\int_{-\infty}^{+\infty}z^2K(z)\,\mathrm{d}z \tag{27.4}$$

① Stata 记此函数为"epan2",而默认的核函数"epan"与此略有不同。

即偏差与 h^2 成正比，为 h^2 的同阶无穷小，记为 $O(h^2)$①。直观来看，带宽 h 越大，则将使用离 x_0 更远的点在估计 $f(x_0)$，导致偏差增大。更进一步，由于偏差大致与 h^2 成正比，故随着 h 增大，偏差将以平方的速度迅速上升。如果当样本容量 $n\to\infty$ 时，让带宽 $h\to 0$（正如在画直方图时，样本容量越大，则可以把每组分得越细），则偏差将在大样本中消失。另一方面，密度函数的二阶导数 $f''(x_0)$ 越大，即在 x_0 处的曲率越大，则 x_0 附近的函数值波动越大，导致 x_0 附近观测点所包含的信息量下降，也会引起偏差增大。最后，偏差还取决于核函数 $K(z)$。

类似地，可以证明核密度估计的方差为（参见附录）

$$\mathrm{Var}[\hat{f}(x_0)] = \frac{1}{nh} f(x_0) \int_{-\infty}^{+\infty} K(z)^2 \mathrm{d}z + o(1/nh) \tag{27.5}$$

其中，$o(1/nh)$ 表示是 $(1/nh)$ 的高阶无穷小②。因此，$\mathrm{Var}[\hat{f}(x_0)] = O(1/nh)$，即是 $(1/nh)$ 的同阶无穷小。直观来看，样本容量 n 越大，则方差越小；另一方面，带宽 h 越大，由于使用了更多观测点来估计 $f(x_0)$，故方差越小。如果当样本容量 $n\to\infty$ 时，让 $nh\to\infty$（虽然 $h\to 0$，但 h 趋于 0 的速度比样本容量 $n\to\infty$ 的速度更慢，故二者的乘积仍然趋于无穷），则此方差将在大样本中消失。

核密度估计的一致性

如果当 $n\to\infty$ 时，让带宽 $h\to 0$ 且 $nh\to\infty$，则偏差 $\mathrm{Bias}(x_0)$ 与方差 $\mathrm{Var}[\hat{f}(x_0)]$ 在大样本下都趋于 0，故根据均方收敛可知，$\hat{f}(x_0)$ 是 $f(x_0)$ 的一致估计量。

核密度估计的渐近正态性

如果核函数 $K(z)$ 的条件(iv)得到满足，则通过中心极限定理可以证明，$\hat{f}(x_0)$ 服从渐近正态分布：

$$\sqrt{nh}[\hat{f}(x_0) - f(x_0) - \mathrm{Bias}(x_0)] \xrightarrow{d} N\left(0, f(x_0)\int_{-\infty}^{+\infty} K(z)^2 \mathrm{d}z\right) \tag{27.6}$$

利用估计量 $\hat{f}(x_0)$ 的渐近正态性，可以进行区间估计。由此可见，核密度估计量的收敛速度为 $\sqrt{nh}$。由于最优带宽 h^* 与 $n^{-0.2}$ 成正比（参见下节），故 $\sqrt{nh} = \sqrt{n\cdot n^{-0.2}} = \sqrt{n^{0.8}} = n^{0.4} < n^{0.5} = \sqrt{n}$，这意味着非参估计量的收敛速度 $n^{0.4}$ 慢于参数估计量的通常收敛速度 $n^{0.5}$。

27.4 最优带宽

如果带宽 h 越大，则 x_0 附近的邻域越大，故偏差也越大（偏差与 h^2 成正比）；另一方面，带宽 h 越大，则 $\hat{f}(x_0)$ 越光滑，即方差 $\mathrm{Var}[\hat{f}(x_0)]$ 越小，参见方差的表达式(27.5)。在选择“最优带宽”(optimal bandwidth) h^* 时，通常希望最小化均方误差(MSE)，即估计量方差与偏差平方之和：

$$\min_h \mathrm{MSE}[\hat{f}(x_0)] = [\mathrm{Bias}(x_0)]^2 + \mathrm{Var}[\hat{f}(x_0)] \tag{27.7}$$

由于 $\mathrm{Bias}(x_0) = O(h^2)$，故 $[\mathrm{Bias}(x_0)]^2 = O(h^4)$，而 $\mathrm{Var}[\hat{f}(x_0)] = O(1/nh)$，故此最小化问题可以大致写为

---

① 记函数 $a(h)$ 为 $O(h^k)$，如果当 $h\to 0$ 时，$a(h)/h^k$ 的极限是一个有限的数。

② 记函数 $a(h)$ 为 $o(h^k)$，如果当 $h\to 0$ 时，$a(h)/h^k$ 的极限为 0。

$$\min_h \mathrm{MSE}[\hat{f}(x_0)] = k_1 h^4 + (k_2/nh) \tag{27.8}$$

其中，k_1，k_2 为常数。对 h 求导，可得一阶条件为

$$4k_1 h^3 + k_2 \frac{1}{n}(-1/h^2) = 0 \tag{27.9}$$

$$h = (4k_1/k_2)^{-0.2} n^{-0.2} \tag{27.10}$$

由此可知，最优带宽为 $h^* = O(n^{-0.2})$。显然，随着 n 增大，$n^{-0.2} = 1/\sqrt[5]{n}$ 的下降速度远慢于 $n^{-1} = 1/n$，参见图 27.2。故当 $n \to \infty$ 时，最优带宽 $h^* \to 0$，而且 $nh^* = n \cdot O(n^{-0.2}) = O(n^{0.8}) \to \infty$。因此，选择最优带宽 h^*，就能保证核密度估计的一致性。

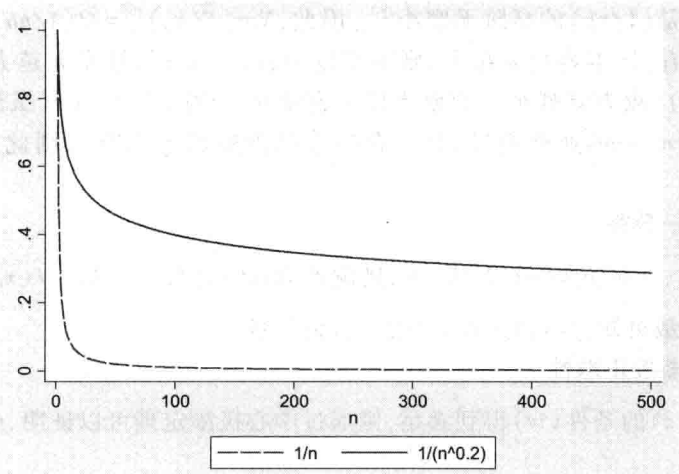


图 27.2 对比 $n^{-0.2}$ 与 n^{-1} 的下降速度

然而，均方误差 $\mathrm{MSE}[\hat{f}(x_0)]$ 依然取决于 x_0。如果希望得到对于 x_0 所有可能取值的均方误差的整体度量，则可以最小化如下的“积分均方误差”（Integrated Mean Squared Error，简记 IMSE）：

$$\min_h \mathrm{IMSE} \equiv \int_{-\infty}^{+\infty} \mathrm{MSE}[\hat{f}(x_0)]\,\mathrm{d}x_0 \tag{27.11}$$

求解这个最小化问题，Silverman（1986）证明最优带宽为（参见附录）

$$h^* = \delta\left[\int_{-\infty}^{+\infty} f''(x_0)^2\,\mathrm{d}x_0\right]^{-0.2} n^{-0.2} \tag{27.12}$$

其中，常数 $\delta \equiv \left[\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z \Big/ \left(\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right)^2\right]^{0.2}$ 仅依赖于核函数。从方程（27.12）可以看出，最优带宽 h^* 还取决于密度函数的曲率（$f''(x_0)$）。当密度函数波动较大（highly variable）时，如果选择较大的邻域，将带来更大的偏差，故最优带宽 h^* 较小。

由于 δ 依赖于核函数，故最优带宽 h^* 也依赖于核函数。可以证明，如果对于不同的核函数分别使用相应的最优带宽，则积分均方误差 IMSE（h^*）差别不大。能使 IMSE（h^*）最小化的核函数为“伊番科尼可夫核”（Epanechnikov or quadratic），也是 Stata 默认的核函数，尽管它只有微弱的优势。总之，对于最优带宽的选择远比核函数的选择更重要。使用不同核函数得到的密度估计一般非常接近。

但是，最优带宽 h^* 仍然依赖于待估密度函数的二阶导数 $f''(x_0)$。如果样本来自正态总体，

则可以计算出，$\int_{-\infty}^{+\infty} f''(x_0)^2 \mathrm{d}x_0 = 3/(8\sqrt{\pi}\sigma^5) = 0.211\ 6/\sigma^5$，故

$$h^* = 1.364\ 3\delta n^{-0.2} s \tag{27.13}$$

其中，s 为样本标准差。为了防止样本标准差受极端值的影响，常使用“Silverman 嵌入估计”(Silverman's plug-in estimate)：

$$h^* = 1.364\ 3\delta n^{-0.2} \min(s, iqr/1.349) \tag{27.14}$$

其中，“iqr”为样本四分位距(sample interquartile range)，即样本 3/4 分位数与 1/4 分位数之间的距离。使用 $\min(s, iqr/1.349)$ 是为了防止存在极端值而导致 s 被高估。使用嵌入估计通常能得到很好的结果，但为了保险起见，可以比较两倍嵌入估计与一半嵌入估计的效果。在实践中，也常使用“眼球法”(eyeball method)，即用肉眼对带宽进行判断，是否使用当前带宽导致密度函数的图形“过度光滑”(oversmoothed)或“不够光滑”(undersmoothed)，然后微调到合适的带宽。

27.5 多元密度函数的核估计

对于 k 维随机变量 $\boldsymbol{x}$，可以类似地进行“多元密度函数的核估计”(multivariate kernel density estimator)：

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K[(\boldsymbol{x}_i - \boldsymbol{x}_0)/h] \tag{27.15}$$

其中，$K(\cdot)$ 是 k 维核函数，即权重函数。$K(\cdot)$ 通常为一维核函数的乘积，也可以使用多维正态的密度函数。多元密度函数核估计的性质与一元的情形相似。但最优带宽为 $h^* = O(n^{-1/(k+4)})$(大于一元情形下的最优带宽)，而 $\hat{f}(x_0)$ 收敛到真实值的速度也更慢。

在多维情况下，容易出现“数据稀疏”(sparseness of data)的问题，即在 x_0 附近的观测点很少。另外，即便是二元密度函数，也要画成三维的立体图；更高维则无法画图。

估计多维密度函数的用途之一是为了估计条件密度函数(conditional density function)。由于条件密度 $f(y|x) = f(x,y)/f(x)$，故可以用 $\hat{f}(y|x) = \hat{f}(x,y)/\hat{f}(x)$ 作为条件密度的估计量，其中 $\hat{f}(x,y)$ 与 $\hat{f}(x)$ 分别为二维与一维的密度函数核估计。

27.6 非参数核回归

计量经济学的首要方法为回归。考虑以下非参数一元回归模型：

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i \\ \varepsilon_i &\sim \mathrm{iid}(0, \sigma_\varepsilon^2) \end{aligned} \tag{27.16}$$

该模型的困难(与优点)在于，$m(\cdot)$ 是未知函数(连函数形式也未知)。也正因为如此，该模型可以更好地反映真实的回归关系。非参数回归的思想是，对于每一个 $i(i = 1, \cdots, n)$，分别估计 $m(x_i)$，从而得到对回归函数 $m(x)$ 的估计。在某种意义上，我们不寻求 $m(x)$ 的解析解，而是寻找其数值解。

考虑最简单的情形。假设对于 x 的某个特定取值，比如 x_0，都有若干个 y 的观测值，比如 n_0

个。显然,可以把这 n_0 个 y 观测值的平均值作为 $m(x_0)$ 的估计量。但在现实数据中,n_0 可能很小(对于连续变量,很可能 n_0 仅为1),导致估计量的方差过大。解决数据稀疏问题的一种方法是,对 x_0 附近邻域中的观测值也进行加权平均,即"局部加权平均估计量"(local weighted average estimator):

$$\hat{m}(x_0) = \sum_{i=1}^{n} w_{i0,h} y_i \tag{27.17}$$

其中,权重 $w_{i0,h}$ 是 (x_i, x_0, h) 的函数,即 $w_{i0,h} = w(x_i, x_0, h)$,且满足 $\sum_{i=1}^{n} w_{i0,h} = 1$。$x_i$ 是 x_0 附近的点,而 h 仍然是带宽或光滑参数。Nadaraya(1964)与 Watson(1964)使用核函数来定义以下权重,得到"核回归估计量"(kernel regression estimator):

$$w_{i0,h} = \frac{K[(x_i - x_0)/h]}{\sum_{i=1}^{n} K[(x_i - x_0)/h]} \tag{27.18}$$

因此,核回归估计量可以写为

$$\hat{m}(x_0) = \frac{\sum_{i=1}^{n} K[(x_i - x_0)/h] y_i}{\sum_{i=1}^{n} K[(x_i - x_0)/h]} \tag{27.19}$$

由于使用了 x_0 附近邻域中观测值的信息,核回归估计量 $\hat{m}(x_0)$ 也是有偏的。可以证明(参见附录),在大样本下其偏差为

$$\text{Bias}(x_0) \equiv \text{E}[\hat{m}(x_0)] - m(x_0) = h^2 \left[m'(x_0) \frac{f'(x_0)}{f(x_0)} + \frac{1}{2} m''(x_0) \right] \int_{-\infty}^{+\infty} z^2 K(z) \, \text{d}z \tag{27.20}$$

因此,$\text{Bias}(x_0) = O(h^2)$。而核回归估计的方差为

$$\text{Var}[\hat{m}(x_0)] = \frac{1}{nh} \frac{\sigma_\varepsilon^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 \, \text{d}z + o(1/nh) \tag{27.21}$$

故 $\text{Var}[\hat{m}(x_0)] = O(1/nh)$。因此,与核密度估计的情形类似,如果当 $n \to \infty$ 时,让带宽 $h \to 0$,而且 $nh \to \infty$,则根据均方收敛,核回归估计是一致的。可以证明(参见附录),如果 $\{x_1, x_2, \cdots, x_n\}$ 为独立同分布的,则核回归估计量 $\hat{m}(x_0)$ 服从渐近正态分布:

$$\sqrt{nh}[\hat{m}(x_0) - m(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 \, \text{d}z\right) \tag{27.22}$$

由于 $\text{Bias}(x_0) = O(h^2)$ 且 $\text{Var}[\hat{m}(x_0)] = O(1/nh)$,最小化 IMSE 的结果显示,最优带宽为 $h^* = O(n^{-0.2})$。然而,由于最优带宽 h^* 取决于待估计回归函数的导数(因为 $m'(x_0)$,$m''(x_0)$ 出现在偏差的表达式(27.20)中),而要估计 $m'(x_0)$,$m''(x_0)$ 又需要指定最优带宽 h^*,故在实践上常使用以下"交叉核实"(Cross Validation,简记 CV)的方法来确定最优带宽 h^*。其思想是,在估计 $\hat{m}(x_i)$ 时,不使用 y_i 的信息,看其余观测值预测 y_i 的能力有多强;而这个能力又取决于带宽 h。故选择带宽 h,使得此预测能力最强,即最小化以下目标函数:

$$\min_h \text{CV}(h) \equiv \sum_{i=1}^{n} [y_i - \hat{m}_{-1}(x_i)]^2 \pi(x_i) \tag{27.23}$$

其中,$\hat{m}_{-1}(x_i) \equiv \dfrac{\sum_{j \neq i} w_{ji,h} y_j}{\sum_{j \neq i} w_{ji,h}}$ 是对 $m(x_i)$ 的"去掉一个观测值"估计量(leave-one-out estimate),即 $j = 1, \cdots, n$,但 $j \neq i$。$\pi(x_i)$ 是权重函数(weighting function),主要是为了给边界附近的端点(end points)更小的权重,以避免端点可能对估计量带来较大的扭曲。比如,可以不考虑 x_i 的5%分位数以下与95%分位数以上的观测值,即对这些观测值令 $\pi(x_i) = 0$,而对其余观测值令 $\pi(x_i) = 1$。

之所以要去掉自身第 i 个观测值是因为，如果将它保留，则总可以选择足够小的带宽 h 使得对于任何 i，都有 $\hat{m}(x_i)=y_i$[①]，故 $\mathrm{CV}(h)=0$ 得到最小化。但这个过小的带宽 h 并不最优。$\mathrm{CV}(h)$ 也被称为“预测误差的估计值”(estimated prediction error)，因为它衡量的是用 $\hat{m}_{-1}(x_i)$ 来预测 y_i 会带来多大的预测误差。可以证明，最小化 $\mathrm{CV}(h)$ 与最小化 IMSE 是渐近等价的。在实践上，交叉核实并非决定最优带宽的完美方法，故仍经常辅之以不严格的眼球法。另外，可以证明，能使 $\mathrm{IMSE}(h^*)$ 最小化的核函数为“伊番科尼可夫核”(Epanechnikov 或 quadratic kernel)，尽管它只有微弱的优势。

27.7 多元核回归

更一般地，对于 k 维解释向量 $\boldsymbol{x}$，考虑如下非参数多元回归模型：

$$y_i=m(\boldsymbol{x}_i)+\varepsilon_i=m(x_{1i},x_{2i},\cdots,x_{ki})+\varepsilon_i \tag{27.24}$$

其中，$m(\cdot)$ 是未知的多元函数。在 $\boldsymbol{x}_0$ 处的核回归估计量可以写为

$$\hat{m}(\boldsymbol{x}_0)=\frac{\sum_{i=1}^{n}K[(\boldsymbol{x}_i-\boldsymbol{x}_0)/h]y_i}{\sum_{i=1}^{n}K[(\boldsymbol{x}_i-\boldsymbol{x}_0)/h]} \tag{27.25}$$

其中，$K(\cdot)$ 为 k 维核函数。多元核回归估计量的性质与一元核回归相似。但最优带宽为 $h^*=O(n^{-1/(k+4)})$(大于一元情形下的最优带宽)，而 $\hat{m}(\boldsymbol{x}_0)$ 收敛到真实值的速度也更慢。多元回归的解释变量越多，则收敛的速度越慢，对样本容量的要求也就越大。这种“维度的诅咒”(curse of dimensionality)大大限制了多元非参数回归的应用。一种解决方法是，使用半参数估计(semiparametric estimation)来降低模型中非参数部分的维度，参见下文。

27.8 k 近邻回归

事实上，核回归估计是局部加权平均估计量(local weighted average estimator)的一个特例，因为它使用的是一个特别的权重。选择权重的另一方式是，对于最靠近 x_0 的 k 个 x_i 的观测值都给予相同的权重，而对其余观测值则给予权重 0，即只对最靠近 x_0 的 k 个观测值进行简单算术平均(不加权)。记 $N_k(x_0)$ 为最靠近 x_0 的 k 个 x_i 观测值的集合(包括 x_0 自身)，则 k 近邻估计量(k-nearest neighbor estimator)可以定义为

$$\hat{m}_{\mathrm{KNN}}(x_0)\equiv\frac{1}{k}\sum_{i=1}^{n}\mathbf{1}\{x_i\in N_k(x_0)\}\cdot y_i \tag{27.26}$$

这个估计量可以看成是使用“均匀核”(uniform kernel)的核估计，但其带宽却是可变的，而且可以不对称(左边的带宽不一定等于右边的带宽)。“对称化”(symmetrized)的 k 近邻估计量则对小于 x_0 的 $(k-1)/2$ 观测值与大于 x_0 的 $(k-1)/2$ 观测值进行简单算术平均。

在某种意义上，k 近邻估计量相当于移动平均(moving average)。由于 k 近邻估计量使用的是简单算术平均，而不是核回归估计所使用的加权平均，故前者可能不如后者光滑(正如直方图

① 可以选择足够小的带宽，使得 x_i 邻域中的观测值只剩下 y_i，故 $\hat{m}(x_i)=y_i$。

不如核密度估计光滑)。而且,越靠近端点,则可用于进行移动平均的样本点就越少,估计量也就越不准确。这种边界问题(boundary problem)可通过下面的"局部线性回归"来缓解。

27.9 局部线性回归

前面介绍的核回归估计量,实际上是一个"局部常数估计"(local constant estimator),因为它假定在 x_0 附近的某个邻域里,$m(x)$ 均等于一个常数。局部线性回归则假定 $m(x)$ 在 x_0 附近的某个邻域里为线性函数,即在该邻域里,$m(x) = a_0 + b_0(x - x_0)$,然后使用加权最小二乘法(WLS)来估计这个线性函数(因为不同观测点到 x_0 的距离不同,所包含的信息量也因此不同),即最小化以下目标函数:

$$\min_{\{a_0,b_0\}} \sum_{i=1}^{n} K[(x_i - x_0)/h][y_i - a_0 - b_0(x_i - x_0)]^2 \tag{27.27}$$

其中,$K(\cdot)$ 为核函数,即权重函数。显然,离 x_0 越近,则权重越大(除非使用均匀核,则权重都一样,等价于标准 OLS 回归)。因此,在 x_0 附近的小邻域里,$\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$。当 x 正好等于 x_0 时,$\hat{m}(x_0) = \hat{a}_0$,而一阶导数 $\hat{m}'(x_0) = \hat{b}_0$。这种方法称为"局部线性回归"(local linear regression),由 Fan(1992)首倡,故也称为"范回归"(Fan regression)。Fan(1992)表明,局部线性回归不仅能较好地解决"边界问题",而且比核回归更有效率且适用于更多数据类型。直观来看,如果带宽足够小,则在此小邻域内,一般的函数都可以很好地用线性函数来近似,故局部线性回归具有较好的性质。

更一般地,可以假定 $m(x)$ 在 x_0 附近的某个邻域为 p 级多项式。"局部 p 级多项式估计量"(local polynomial estimator of degree p)最小化以下目标函数:

$$\min_{\{a_0,b_0\}} \sum_{i=1}^{n} K[(x_i - x_0)/h]\left[y_i - a_{0,0} - a_{0,1}(x_i - x_0) - \cdots - \frac{a_{0,p}}{p!}(x_i - x_0)^p\right]^2 \tag{27.28}$$

局部线性或局部多项式估计量有助于缓解前面提及的在两端估计不准确的边界问题。一个常用的局部回归估计量为 Cleveland(1979)所提出的"局部加权散点光滑估计量"(Locally weighted scatterplot smoothing,简记 Lowess),这是局部多项式估计量的一个变种(variant)或升级版。该估计量使用"三三核"(tricubic kernel),同时使用可变带宽 $h_{0,k}$(由 x_0 到其最近的 k 个观测值的距离所决定[①]),以及对较大的残差 $[y_i - \hat{m}(x_i)]$ 给予较小的权重。Lowess 估计的优点是,它使用了可变带宽(依数据的稠密程度而定),对于极端值更稳健,而且缓解了在两端估计不准的边界问题。

27.10 非参数估计的 Stata 命令及实例

Stata 提供了如下有关非参数估计的命令。

1. 直方图与核密度估计

```
hist y,width(#) start(#)        (画 y 的直方图,指定组宽与起始点)
```

① 在 Stata 中,默认 $k = 0.8n$,其中 n 为样本容量。

```
kdensity y,bwidth(#) n(#)      (对 y 进行核密度估计,指定带宽与估计点数)
kdensity y,norm                (同时画一个正态分布,作为对比)
```

有关直方图与核密度估计的 Stata 实例,参见第 4 章。

2. 核密度回归

```
ssc install kernreg1           (下载安装命令 kernreg1)
kernreg1 y x,bwidth(#) kercode(#) npoint(#) gen(mhvar gridvar)
```

其中,选择项"bwidth(#)"用于指定带宽,默认值为最优带宽;"kercode(#)"为必选项,用于指定核函数,其中 1 = Uniform,2 = Triangle,3 = Epanechnikov,4 = Quartic(Biweight),5 = Triweight,6 = Gaussian,7 = Cosinus;"npoint(#)"为必选项,用于指定将解释变量"x"的取值范围分为多少个等距离的点(称为"网格点",grid point),并在这些点上进行核密度回归;选择项"gen(mhvar gridvar)"产生两个变量,其中"gridvar"为解释变量"x"的网格点,而"mhvar"为对应的核密度回归值。另外,该程序将最优带宽记为"全局宏"(global macro)"S_1"(字母"S"为大写),故可以用"$S_1"来调用它。

下面以数据集 mpyr.dta 为例(参见第 21 章),将 logv(货币流通速度的对数)对 r(名义利率)进行核密度回归,结果如图 27.3:

```
. use mpyr.dta,clear
. kernreg1 logv r,k(3) np(100) gen(logv_kern r_grid)
```

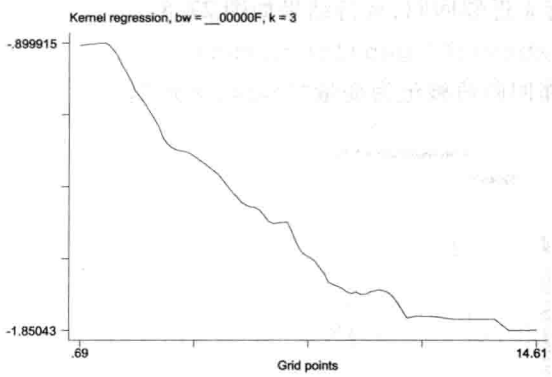


图 27.3 核密度回归

此命令使用了伊番科尼可夫核(k = 3),在 100 个等距离的网格点进行估计,并将解释变量 r(利率)的网格点记为变量"r_grid",对应的核密度回归值记为"logv_kern"。

更直观地,可将散点图、线性回归线及核密度回归线画在一张图上(如图 27.4):

```
. graph twoway(scatter logv r)(lfit logv r,lpattern("-"))(line logv_
kern r_grid)
```

由于该命令的默认带宽为最优带宽,故可以利用该程序计算最优带宽。

```
. dis $S_1
. 90306211
```

这表明,最优带宽为 0.90。

3. k 近邻回归

k 近邻回归可通过非官方命令 knnreg 来实现:

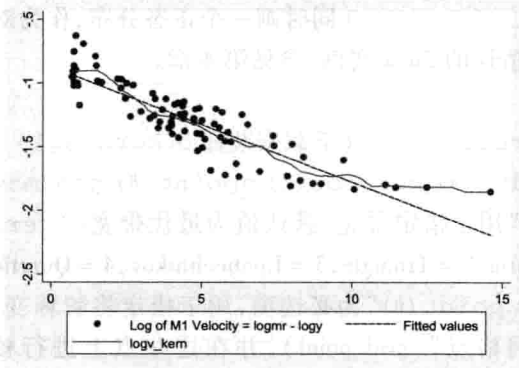


图 27.4 散点图、线性回归线与核密度回归

```
net install snp10.pkg （下载安装命令 knnreg）
```

该命令的格式为

```
knnreg y x,knum(#) gen(mkvar)
```

其中，选择项“knum(#)”为必选项，用于指定 k 近邻回归中 k 的取值；选择项“gen(mkvar)”将 k 近邻回归值赋予变量“mkvar”。

继续上面的例子进行 k 近邻回归，运行结果如图 27.5：

```
. knnreg logv r,knum(10) gen(logv_knn)
```

在上述命令中，k 近邻回归值被记为变量“logv_knn”。

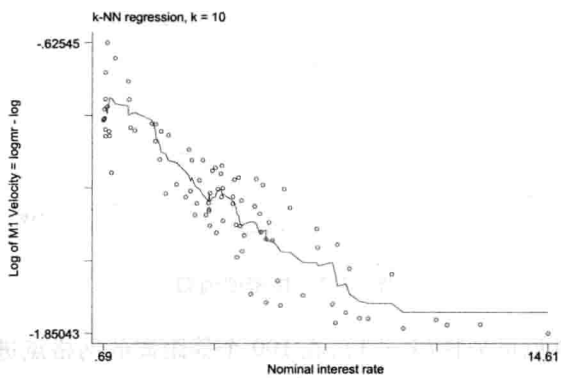


图 27.5 k 近邻回归

4. 局部多项式回归

```
lpoly y x,bwidth(#) kernel(kernel) degree(#) gen(gridvar newvar)
```

其中，选择项“bwidth(#)”用于指定带宽，默认值为“rule-of-thumb(ROT) bandwidth estimator”；“kernel(kernel)”用于指定核函数，默认值为“kernel(epanechnikov)”；选择项“degree(#)”用于指定多项式的级数，默认值为“degree(0)”，即只有常数项（局部均值平滑，local-mean smoothing）；选择项“gen(gridvar newvar)”将产生两个变量，其中“gridvar”为解释变量“x”的网格点，而“newvar”为对应的局部多项式回归值。

继续上面的例子进行局部多项式回归，结果如图 27.6：

```
. lpoly logv r,b($S_1) degree(1) gen(r_grid_1 logv_lpoly)
```

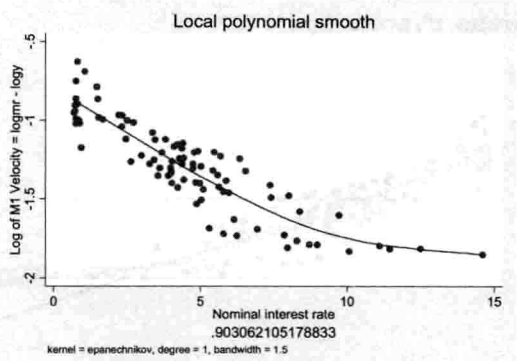


图 27.6 局部多项式回归

其中,选择项“b($S_1)”表示使用前面核密度回归的最优带宽(0.90);选择项“degree(1)”表明,这是一个局部线性回归;选择项“gen(r_grid_1 logv_lpoly)”将解释变量 r(利率)的网格点记为“r_grid_1”,而把对应的局部回归值记为“logv_lpoly”。

5. Lowess 回归

```
lowess y x,bwidth(#) gen(newvar)
```

其中,选择项“bwidth(#)”用于指定带宽,默认值为 0.8;选择项“gen(newvar)”将被解释变量的光滑值(smoothed values)赋予变量“newvar”。

继续以数据集 mpyr.dta 为例,结果如图 27.7:

```
. lowess logv r,b($S_1) gen(logv_lowess)
```

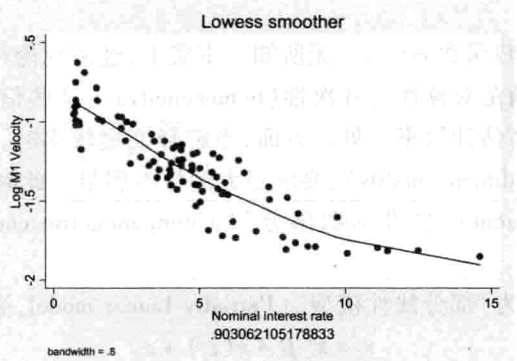


图 27.7 Lowess 回归

其中,选择项“b($S_1)”表示使用前面核密度回归的最优带宽(0.90);选择项“gen(logv_lowess)”把被解释变量的光滑值记为“logv_lowess”。

最后,将以上所有各图合并在一起(如图 27.8):

```
. graph twoway(scatter logv r)(lfit logv r,lp("-"))(line logv_knn r,lp
("_"))(line logv_kern r_grid)(line logv_lpoly r_grid_1,lp("_."))
(line logv_lowess r,lp("."))
```

```
graph twoway(scatter logv r)(lfit logv r,lp("-"))(line logv_knn r,sort
lp("_"))(line logv_kern r_grid,sort)(line logv_lpoly r_grid_1,sort lp
("_."))(line logv_lowess r,sort lp("."))
```

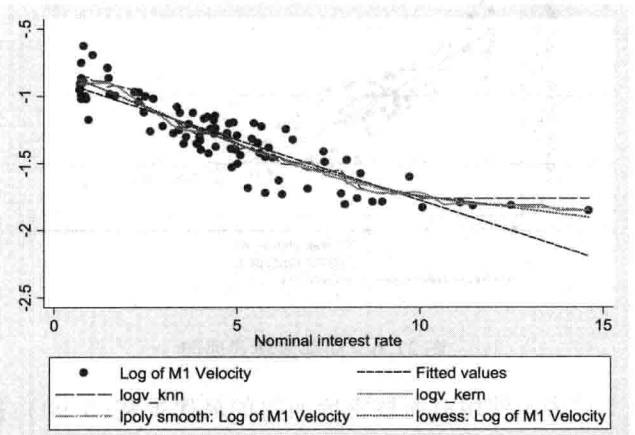


图 27.8 各种非参数回归的比较

从图 27.8 可知,以上各种非参数回归的结果很接近,但与线性回归有所不同。非参数估计对数据的拟合程度明显优于线性回归(尤其在右端)。总之,非参数回归以牺牲部分光滑度为代价(线性回归的光滑度为无穷),可以更好地拟合数据。

27.11 半参数估计

非参数回归假设对回归函数 $m(x)$ 一无所知。事实上,经济理论可能对 $m(x)$ 的具体形式有所限制,比如需求函数要满足对称性与齐次性(homogeneity)。这些信息应该而且可以被纳入非参数回归的框架中,以提高估计效率。另一方面,当解释变量较多时,完全的非参数方法可能面临"维度的诅咒"(curse of dimensionality),要求很大的样本容量。此时,可以使用同时包含"参数部分"(a parametric component)与"非参数部分"(a nonparametric component)的"半参数模型"(semiparametric model)。

最常见的半参数模型为"部分线性模型"(Partially Linear model,简记 PL),即

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + g(z_i) + \varepsilon_i \tag{27.29}$$

其中,参数部分 $\boldsymbol{x}_i'\boldsymbol{\beta}$ 为线性函数,而非参数部分 $g(z_i)$ 为未知函数(连函数形式也不知道),并假设扰动项 ε_i 均值独立于 $\boldsymbol{x}_i, z_i$,即 $\mathrm{E}(\varepsilon_i|\boldsymbol{x}_i, z_i) = 0$。Robinson(1988)提出了以下"罗宾逊差分估计量"(Robinson difference estimator)。给定 z_i,对方程(27.29)两边取条件期望可得

$$\mathrm{E}(y_i|z_i) = \mathrm{E}(\boldsymbol{x}_i|z_i)'\boldsymbol{\beta} + g(z_i) + \underbrace{\mathrm{E}(\varepsilon_i|z_i)}_{=0} \tag{27.30}$$

其中,根据迭代期望定律,

$$\mathrm{E}(\varepsilon_i|z_i) = \mathrm{E}_{x_i}\mathrm{E}[(\varepsilon_i|z_i)|\boldsymbol{x}_i] = \mathrm{E}_{x_i}\underbrace{\mathrm{E}(\varepsilon_i|\boldsymbol{x}_i, z_i)}_{=0} = 0 \tag{27.31}$$

将方程(27.29)减去方程(27.30)可得

$$y_i - \mathrm{E}(y_i|\boldsymbol{z}_i) = [\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{z}_i)]'\boldsymbol{\beta} + \varepsilon_i \tag{27.32}$$

在这个“差分方程”中,未知函数 $g(\boldsymbol{z}_i)$ 被消去了,而条件期望 $\mathrm{E}(y_i|\boldsymbol{z}_i)$ 与 $\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ 则可以用非参数方法来估计(比如,核回归)。假设 $\hat{\mathrm{E}}(y_i|\boldsymbol{z}_i)$ 与 $\hat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ 分别为对 $\mathrm{E}(y_i|\boldsymbol{z}_i)$ 与 $\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ 的非参数估计,则可对以下线性方程进行最小二乘估计:

$$y_i - \hat{\mathrm{E}}(y_i|\boldsymbol{z}_i) = [\boldsymbol{x}_i - \hat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{z}_i)]'\boldsymbol{\beta} + u_i \tag{27.33}$$

记此估计量为 $\hat{\boldsymbol{\beta}}_{\mathrm{PL}}$。由于使用了条件期望的估计量来替代条件期望本身,故扰动项不再是 ε_i,记新扰动项为 u_i。假设 ε_i 为 $\mathrm{iid}(0,\sigma^2)$,根据大样本 OLS 理论,可以证明,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{PL}} - \boldsymbol{\beta}) \xrightarrow{d} N\left[0, \sigma^2\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum\nolimits_{i=1}^{n}\boldsymbol{w}_i\boldsymbol{w}_i'\right)^{-1}\right] \tag{27.34}$$

其中,$\boldsymbol{w}_i \equiv \boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{z}_i)$。在上式中,只要用 $\hat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ 来替代 $\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ 就可以估计渐近协方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{\mathrm{PL}})$。如果存在异方差,则可以使用标准的夹心估计量来得到稳健标准误。

最后,可以得到对 $g(\boldsymbol{z}_i)$ 的非参数估计,即

$$\hat{g}(\boldsymbol{z}_i) = \hat{\mathrm{E}}(y_i|\boldsymbol{z}_i) - \hat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{z}_i)'\hat{\boldsymbol{\beta}} \tag{27.35}$$

对于部分线性模型使用罗宾逊差分估计量,可通过非官方命令 semipar 来实现:

```
.ssc install semipar  (下载安装命令 semipar)
```

该命令的格式为

```
semipar y x1 x2 x3,nonpar(z)robust cluster(varname) kernel(kernel)gen
(varname) ci partial(varname)
```

其中,必选项“nonpar(z)”用来指定半参模型中非参部分,选择项“robust”表示使用异方差稳健标准误,选择项“cluster(varname)”表示使用聚类稳健标准误。选择项“gen(varname)”用来记录 $\hat{\mathrm{E}}(y_i|\boldsymbol{z}_i)$(nonparametric fit of the dependent variable);选择项“ci”表示在图上给出此 nonparametric fit 的置信区间;选择项“partial(varname)”用来记录 $[y_i - \hat{\mathrm{E}}(y_i|\boldsymbol{z}_i)]$(dependent variable partialled out from the nonparametric fit)。选择项“kernel(kernel)”用来指定用于核回归的核函数,默认为高斯核,即“kernel(gaussian)”,更多备选核函数参见“help semipar”。

下面以第 4 章的数据集 nerlove.dta 为例。假设我们不确定原材料价格($\ln PF$)对总成本($\ln TC$)的作用函数形式,故考虑以下部分线性模型:

$$\ln TC_i = \beta_1 \ln Q_i + \beta_2 \ln PL_i + \beta_3 \ln PK_i + g(\ln PF_i) + \varepsilon_i \tag{27.36}$$

其中,$g(\cdot)$ 为未知函数。

作为参照,首先进行 OLS 回归。

```
.use nerlove.dta,clear
.reg lntc lnq lnpl lnpk lnpf,r
.est sto reg
```

```
Linear regression                               Number of obs =       145
                                                F(  4,   140) =    177.19
                                                Prob > F      =    0.0000
                                                R-squared     =    0.9260
                                                Root MSE      =    .39227
```

lntc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0325376	22.16	0.000	.656585	.785242
lnpl	.4559645	.260326	1.75	0.082	-.0587139	.9706429
lnpk	-.2151476	.3233711	-0.67	0.507	-.8544698	.4241745
lnpf	.4258137	.0740741	5.75	0.000	.2793653	.5722622
_cons	-3.566513	1.718304	-2.08	0.040	-6.963693	-.1693331

其次,使用罗宾逊差分估计量进行半参估计。

```
.semipar lntc lnq lnpl lnpk,nonpar(lnpf) robust xtitle(lnPF) ytitle
(lnTC) ci
.est sto semipar
```

```
                                                Number of obs =       145
                                                R-squared     =    0.9273
                                                Adj R-squared =    0.9258
                                                Root MSE      =    0.3749
```

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7237013	.030765	23.52	0.000	.6628846	.784518
lnpl	.3398899	.2931278	1.16	0.248	-.2395684	.9193481
lnpk	-.3549753	.3252955	-1.09	0.277	-.9980231	.2880725

该命令还会输出被解释变量 ln*TC* 对非参变量 ln*PF* 的核回归图(参见图 27.9)。

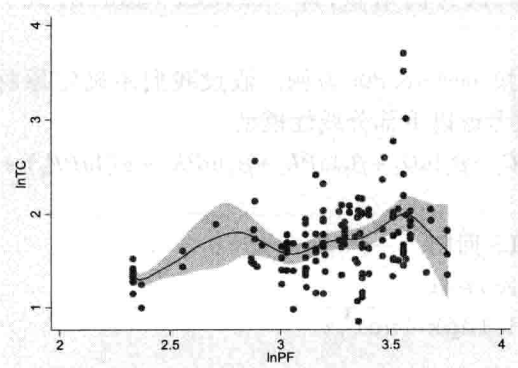


图 27.9 ln*TC* 对 ln*PF* 的核回归图

下面,对比这两种回归方法的结果。

```
. esttab reg semipar,mtitles r2 se star(* 0.1 * * 0.05 * * * 0.01)
```

	(1) reg	(2) semipar
lnq	0.721***	0.724***
	(0.0325)	(0.0308)
lnpl	0.456*	0.340
	(0.260)	(0.293)
lnpk	-0.215	-0.355
	(0.323)	(0.325)
lnpf	0.426***	
	(0.0741)	
_cons	-3.567**	
	(1.718)	
N	145	145
R-sq	0.926	0.927

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

从上表可知,半参估计结果与线性回归差别不大。当然,本例纯粹为演示而设,在其他情况下,二者可能有较大差别。

习　题

27.1　使用数据集 mpyr. dta,进行以下估计。

(1) 画变量 logv 的直方图;

(2) 画变量 logv 的核密度图,使用眼球法选择适当的带宽;

(3) 作为对照,在 logv 的核密度图上加上正态分布密度图。

27.2　参照本章的实例,使用数据集 consumption_china. dta 对中国的消费函数进行以下非参数估计。

(1) 核密度回归;

(2) k 近邻回归;

(3) 局部多项式回归(一阶);

(4) Lowess 回归。

附　录

A27.1　证明核密度估计的偏差

证明:假设样本 $\{x_1, x_2, \cdots, x_n\}$ 为独立同分布的。核密度估计量的公式为 $\hat{f}(x_0) = \frac{1}{nh}\sum_{i=1}^{n} K[(x_i - x_0)/h]$,其期望值为

$$\mathrm{E}[\hat{f}(x_0)] = \mathrm{E}\left[\frac{1}{h}K((x - x_0)/h)\right] \qquad (\text{对 } x \text{ 求期望})$$

$$= \int_{-\infty}^{+\infty} \frac{1}{h}K[(x - x_0)/h]f(x)\,\mathrm{d}x \qquad (\text{期望的定义})$$

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} K(z) f(x_0 + hz)\,\mathrm{d}z \qquad \text{（积分变换 } z \equiv (x - x_0)/h, x = x_0 + hz\text{）} \\
&\approx \int_{-\infty}^{+\infty} K(z)\left[f(x_0) + f'(x_0)hz + \frac{1}{2}f''(x_0)(hz)^2\right]\mathrm{d}z \quad \text{（二阶泰勒近似）} \\
&= f(x_0)\underbrace{\int_{-\infty}^{+\infty} K(z)\,\mathrm{d}z}_{=1} + hf'(x_0)\underbrace{\int_{-\infty}^{+\infty} zK(z)\,\mathrm{d}z}_{=0} + \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z \quad \text{（核函数的性质）} \\
&= f(x_0) + \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z
\end{aligned}
$$

因此，$\mathrm{Bias}(x_0) \equiv \mathrm{E}[\hat{f}(x_0)] - f(x_0) \approx \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z$ （移项）。

A27.2 证明核密度估计的方差公式

证明：假设$\{y_i\}$为独立同分布的，则 $\mathrm{Var}(\bar{y}) = \frac{1}{n}\mathrm{Var}(y) = \frac{1}{n}\{\mathrm{E}(y^2) - [\mathrm{E}(y)]^2\}$，故

$$
\mathrm{Var}[\hat{f}(x_0)] = \frac{1}{n}\mathrm{E}\left\{\frac{1}{h}K[(x - x_0)/h]\right\}^2 - \frac{1}{n}\left\{\mathrm{E}\left[\frac{1}{h}K((x - x_0)/h)\right]\right\}^2 \tag{27.37}
$$

其中，上式右边的第一项等于

$$
\begin{aligned}
\mathrm{E}\left\{\frac{1}{h}K[(x - x_0)/h]\right\}^2 &= \int_{-\infty}^{+\infty} \frac{1}{h^2}[K((x - x_0)/h)]^2 f(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \frac{1}{h}K(z)^2 f(x_0 + hz)\,\mathrm{d}z \qquad \text{（积分变换 } z \equiv (x - x_0)/h\text{）} \\
&\approx \int_{-\infty}^{+\infty} \frac{1}{h}K(z)^2[f(x_0) + f'(x_0)hz]\,\mathrm{d}z \qquad \text{（一阶泰勒近似）} \\
&= \frac{1}{h}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z + f'(x_0)\int_{-\infty}^{+\infty} zK(z)^2\,\mathrm{d}z \quad \text{（分成两项）}
\end{aligned}
$$

根据附录 A27.1，$\mathrm{E}\left[\frac{1}{h}K((x - x_0)/h)\right] \approx f(x_0) + \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z$，故方程(27.37)第二项中的 $\mathrm{E}\left[\frac{1}{h}K((x - x_0)/h)\right]^2 = \left[f(x_0) + \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2$。综合以上结果，

$$
\begin{aligned}
\mathrm{Var}[\hat{f}(x_0)] \approx{}& \frac{1}{nh}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z + \frac{1}{n}f'(x_0)\int_{-\infty}^{+\infty} zK(z)^2\,\mathrm{d}z - \\
&\frac{1}{n}\left[f(x_0) + \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2
\end{aligned} \tag{27.38}
$$

当 $n \to \infty$ 且 $h \to 0$ 时，上式的性质由第一项 $\frac{1}{nh}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z$ 所决定（其余两项为高阶无穷小），故 $\mathrm{Var}[\hat{f}(x_0)] = \frac{1}{nh}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z + o(1/nh)$。

A27.3 证明最优带宽

证明：由于 $\mathrm{IMSE} = \int_{-\infty}^{+\infty} \mathrm{MSE}[\hat{f}(x_0)]\,\mathrm{d}x_0 = \int_{-\infty}^{+\infty} \{[\mathrm{Bias}(x_0)]^2 + \mathrm{Var}[\hat{f}(x_0)]\}\,\mathrm{d}x_0$，

而 $\mathrm{Bias}(x_0) = \frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z$，$\mathrm{Var}[\hat{f}(x_0)] \approx \frac{1}{nh}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z$，故

$$
\begin{aligned}
\mathrm{IMSE} &= \int_{-\infty}^{+\infty}\left(\left[\frac{1}{2}h^2 f''(x_0)\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2\right)\mathrm{d}x_0 + \int_{-\infty}^{+\infty}\left[\frac{1}{nh}f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z\right]\mathrm{d}x_0 \\
&= \frac{1}{4}h^4\int_{-\infty}^{+\infty}\left([f''(x_0)]^2\left[\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2\right)\mathrm{d}x_0 + \frac{1}{nh}\int_{-\infty}^{+\infty}\left[f(x_0)\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z\right]\mathrm{d}x_0 \\
&= \frac{1}{4}h^4\left[\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2\int_{-\infty}^{+\infty}[f''(x_0)]^2\,\mathrm{d}x_0 + \frac{1}{nh}\left[\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z\right]\underbrace{\left[\int_{-\infty}^{+\infty} f(x_0)\,\mathrm{d}x_0\right]}_{=1} \\
&= \frac{1}{4}h^4\left[\int_{-\infty}^{+\infty} z^2 K(z)\,\mathrm{d}z\right]^2\int_{-\infty}^{+\infty}[f''(x_0)]^2\,\mathrm{d}x_0 + \frac{1}{nh}\int_{-\infty}^{+\infty} K(z)^2\,\mathrm{d}z
\end{aligned}
$$

将上式对 h 求导，并令其导数为 0，

$$\frac{\partial\ \mathrm{IMSE}}{\partial h} = h^3\left[\int_{-\infty}^{+\infty} z^2K(z)\,\mathrm{d}z\right]^2\int_{-\infty}^{+\infty}[f''(x_0)]^2\,\mathrm{d}x_0 + \frac{1}{n}\int_{-\infty}^{+\infty}K(z)^2\,\mathrm{d}z\left(-\frac{1}{h^2}\right) = 0 \tag{27.39}$$

整理可得最优带宽为 $h^* = \delta\left[\int_{-\infty}^{+\infty} f''(x_0)^2\,\mathrm{d}x_0\right]^{-0.2} n^{-0.2}$，其中常数

$$\delta \equiv \left[\int_{-\infty}^{+\infty}K(z)^2\,\mathrm{d}z\Big/\left(\int_{-\infty}^{+\infty}z^2K(z)\,\mathrm{d}z\right)^2\right]^{0.2} \tag{27.40}$$

A27.4 证明核回归估计量 $\hat{m}(x_0)$ 服从渐近正态分布

$$\sqrt{nh}[\hat{m}(x_0) - m(x_0) - \mathrm{Bias}(x_0)] \xrightarrow{d} N\left(0,\ \frac{\sigma_\varepsilon^2}{f(x_0)}\int_{-\infty}^{+\infty}K(z)^2\,\mathrm{d}z\right) \tag{27.41}$$

证明：由于核回归估计量是一个加权平均，即 $\hat{m}(x_0) = \sum_{i=1}^{n} w_{i0,h}y_i$，故

$$\begin{aligned}
\hat{m}(x_0) - m(x_0) &= \sum_{i=1}^{n} w_{i0,h}y_i - \sum_{i=1}^{n} w_{i0,h}m(x_0) && \text{（权重之和为 1）}\\
&= \sum_{i=1}^{n} w_{i0,h}[y_i - m(x_0)] && \text{（两项合并）}\\
&= \sum_{i=1}^{n} w_{i0,h}[m(x_i) - m(x_0) + \varepsilon_i] && \text{（代入模型的假定 } y_i = m(x_i) + \varepsilon_i\text{）}
\end{aligned}$$

由于权重 $w_{i0,h} = \dfrac{K[(x_i - x_0)/h]}{\sum_{i=1}^{n}K[(x_i - x_0)/h]}$，而核密度估计 $\hat{f}(x_0) = \dfrac{1}{nh}\sum_{i=1}^{n}K[(x_i - x_0)/h]$，故权重可以写为 $w_{i0,h} = \dfrac{\frac{1}{nh}K[(x_i - x_0)/h]}{\hat{f}(x_0)}$。因此，用 $\sqrt{nh}$ 来标准化变量可得

$$\begin{aligned}
\sqrt{nh}[\hat{m}(x_0) - m(x_0)] &= \sqrt{nh}\sum_{i=1}^{n}\frac{\frac{1}{nh}K[(x_i - x_0)/h]}{\hat{f}(x_0)}[m(x_i) - m(x_0) + \varepsilon_i] && \text{（代入 } w_{i0,h}\text{）}\\
&= \frac{1}{\sqrt{nh}}\frac{\sum_{i=1}^{n}K[(x_i - x_0)/h][m(x_i) - m(x_0) + \varepsilon_i]}{\hat{f}(x_0)} && \text{（合并整理）}
\end{aligned}$$

我们已经知道 $\hat{f}(x_0)$ 是 $f(x_0)$ 的一致估计量，即 $\hat{f}(x_0) \xrightarrow{p} f(x_0)$。故只需要考虑上式中分子的渐近分布。将上式中的分子分成两个部分：

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K[(x_i - x_0)/h][m(x_i) - m(x_0)] + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K[(x_i - x_0)/h]\varepsilon_i \tag{27.42}$$

对于第一部分，在大数定律成立的情况下，它将收敛到其期望值，因为

$$\begin{aligned}
&\mathrm{E}\left\{\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K[(x_i - x_0)/h][m(x_i) - m(x_0)]\right\}\\
&= \frac{\sqrt{n}}{\sqrt{h}}\mathrm{E}\{K[(x - x_0)/h][m(x) - m(x_0)]\} \quad (\{x_1, x_2, \cdots, x_n\}\text{为独立同分布的，省略下标 } i)\\
&= \frac{\sqrt{n}}{\sqrt{h}}\int_{-\infty}^{+\infty}K((x - x_0)/h)[m(x) - m(x_0)]f(x)\,\mathrm{d}x \quad \text{（期望的定义）}\\
&= \sqrt{nh}\int_{-\infty}^{+\infty}K(z)[m(x_0 + hz) - m(x_0)]f(x_0 + hz)\,\mathrm{d}z \quad \left(\text{积分变换 } z \equiv \frac{x - x_0}{h}\right)\\
&\approx \sqrt{nh}\int_{-\infty}^{+\infty}K(z)\left[hzm'(x_0) + \frac{1}{2}h^2z^2m''(x_0)\right][f(x_0) + f'(x_0)hz]\,\mathrm{d}z \quad \text{（泰勒展开）}
\end{aligned}$$

将上式的被积函数展开成四项。根据核函数的性质，其中第一项为

$$\sqrt{nh}\int_{-\infty}^{+\infty}K(z)hzm'(x_0)f(x_0)\,\mathrm{d}z = h\sqrt{nh}m'(x_0)f(x_0)\int_{-\infty}^{+\infty}zK(z)\,\mathrm{d}z = 0 \tag{27.43}$$

而第四项为

$$\frac{1}{2}\sqrt{nh}\int_{-\infty}^{+\infty}K(z)h^3z^3m''(x_0)f'(x_0)\,\mathrm{d}z=\frac{1}{2}h^3\sqrt{nh}m''(x_0)f'(x_0)\int_{-\infty}^{+\infty}z^3K(z)\,\mathrm{d}z \tag{27.44}$$

显然，第四项是第二项与第三项的高阶无穷小，故可以忽略。因此，只要考虑第二项与第三项即可：

$$\begin{aligned}
&\mathrm{E}\left\{\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K[(x_i-x_0)/h][m(x_i)-m(x_0)]\right\}\\
&\approx\sqrt{nh}\int_{-\infty}^{+\infty}K(z)\left[h^2z^2m'(x_0)f'(x_0)+\frac{1}{2}h^2z^2m''(x_0)f(x_0)\right]\mathrm{d}z \quad (\text{泰勒展开})\\
&=\sqrt{nh}h^2\left[m'(x_0)f'(x_0)+\frac{1}{2}m''(x_0)f(x_0)\right]\int_{-\infty}^{+\infty}z^2K(z)\,\mathrm{d}z \quad (\text{合并整理})\\
&=\sqrt{nh}f(x_0)\mathrm{Bias}(x_0)
\end{aligned}$$

其中，$\mathrm{Bias}(x_0)=h^2\left[m'(x_0)\dfrac{f'(x_0)}{f(x_0)}+\dfrac{1}{2}m''(x_0)\right]\displaystyle\int_{-\infty}^{+\infty}z^2K(z)\,\mathrm{d}z$。

对于第二部分$\dfrac{1}{\sqrt{nh}}\displaystyle\sum_{i=1}^{n}K[(x_i-x_0)/h]\varepsilon_i$，显然其中每一项的期望值均为 0，而每一项的方差为

$$\begin{aligned}
\mathrm{Var}\{K[(x_i-x_0)/h]\varepsilon_i\}&=\mathrm{E}\{K^2[(x_i-x_0)/h]\varepsilon_i^2\} && (\text{公式 }\mathrm{Var}(y)=\mathrm{E}(y^2)-[\mathrm{E}(y)]^2)\\
&=\mathrm{E}_x\mathrm{E}_\varepsilon\{K^2[(x_i-x_0)/h]\varepsilon_i^2\mid x\} && (\text{迭代期望公式})\\
&=\mathrm{E}_x\{K^2[(x_i-x_0)/h]\mathrm{Var}(\varepsilon_i^2\mid x)\} && (\text{对 }\varepsilon\text{ 求期望})\\
&=\int_{-\infty}^{+\infty}K^2[(x_i-x_0)/h]\mathrm{Var}(\varepsilon_i^2\mid x)f(x)\,\mathrm{d}x && (\text{对 }x\text{ 求期望})\\
&=h\int_{-\infty}^{+\infty}K(z)^2\mathrm{Var}(\varepsilon_i^2\mid x_0+hz)f(x_0+hz)\,\mathrm{d}z && \left(\text{积分变换 }z\equiv\frac{x-x_0}{h}\right)\\
&=h\ \mathrm{Var}(\varepsilon_i^2\mid x_0)f(x_0)\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z && (\text{大样本下 }h\to0)\\
&=h\sigma_\varepsilon^2f(x_0)\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z && (\text{根据假定，}\varepsilon_i\sim\mathrm{iid}(0,\sigma_\varepsilon^2))
\end{aligned}$$

因此，$\mathrm{Var}\left\{\dfrac{1}{\sqrt{nh}}\displaystyle\sum_{i=1}^{n}K[(x_i-x_0)/h]\varepsilon_i\right\}=\dfrac{1}{nh}\displaystyle\sum_{i=1}^{n}\mathrm{Var}\{K[(x_i-x_0)/h]\varepsilon_i\}$

$$\begin{aligned}
&=\frac{1}{nh}\sum_{i=1}^{n}\mathrm{Var}\{K[(x_i-x_0)/h]\varepsilon_i\}\\
&=\frac{1}{nh}n\cdot h\sigma_\varepsilon^2f(x_0)\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z\\
&=\sigma_\varepsilon^2f(x_0)\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z
\end{aligned}$$

综合第一部分与第二部分可知，

$$\sqrt{nh}[\hat{m}(x_0)-m(x_0)]\xrightarrow{d}\frac{1}{f(x_0)}N\left(\sqrt{nh}f(x_0)\mathrm{Bias}(x_0),\sigma_\varepsilon^2f(x_0)\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z\right) \tag{27.45}$$

因此，

$$\sqrt{nh}[\hat{m}(x_0)-m(x_0)]\xrightarrow{d}N\left(\sqrt{nh}\,\mathrm{Bias}(x_0),\frac{\sigma_\varepsilon^2}{f(x_0)}\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z\right) \tag{27.46}$$

故

$$\sqrt{nh}[\hat{m}(x_0)-m(x_0)-\mathrm{Bias}(x_0)]\xrightarrow{d}N\left(0,\frac{\sigma_\varepsilon^2}{f(x_0)}\int_{-\infty}^{+\infty}K(z)^2\mathrm{d}z\right) \tag{27.47}$$

第28章 处理效应

28.1 处理效应与选择难题

在经济学中，我们常希望评估某项目或政策实施后的效应，比如政府推出的就业培训项目（job training program）。此类研究被称为“项目效应评估”（program evaluation）[①]，而项目效应也常被称为“处理效应”（treatment effect）[②]。项目参与者的全体构成“实验组”或“处理组”（treatment group，或 the treated），而未参与项目者则构成“控制组”（control group）或“对照组”（comparison group）。有关处理效应的专著包括 Morgan and Winship（2007），Guo and Fraser（2009），Rosenbaum（2010）以及 Imbens and Wooldridge（2009）的文献综述。

考虑就业培训的处理效应评估。一个自然（天真）的做法是直接对比实验组与控制组的未来收入或就业状况。如果这样做，常会发现参加就业培训者的未来收入比未参加者更低。难道就业培训反而有害？值得注意的是，是否参加培训是参加者自我选择（self selection）的结果，岗位好收入高的人群并不需要参加就业培训，而就业培训的参加者多为失业、低收入者，甚至刑满释放犯。由于实验组与对照组成员的初始条件不完全相同，故存在“选择偏差”（selection bias）。另外，即使实验组的未来收入低于对照组，我们真正感兴趣的问题是，实验组的未来收入是否会比这些人如果未参加培训项目的（假想）未来收入更高[③]。

为此，Rubin（1974）提出了以下“反事实框架”（a counterfactual framework），称为“鲁宾因果模型”（Rubin Causal Model，简记 RCM）。以虚拟变量 $D_i = \{0,1\}$ 表示个体 i 是否参与此项目，即 1 为参与，而 0 为未参与。通常称 D_i 为“处理变量”（treatment variable），反映个体 i 是否得到了“处理”（treatment）。记其未来收入或其他感兴趣的结果（outcome of interest）为 y_i。我们想知道 D_i 是否对 y_i 有因果作用。对于个体 i，其未来收入 y_i 可能有两种状态，取决于是否参加此项目，即

$$y_i = \begin{cases} y_{1i} & \text{若 } D_i = 1 \\ y_{0i} & \text{若 } D_i = 0 \end{cases} \tag{28.1}$$

其中，y_{0i} 表示个体 i 未参加项目的未来收入，而 y_{1i} 表示个体 i 参加项目的未来收入。我们想知道 $(y_{1i} - y_{0i})$，即个体 i 参加该项目的因果效应。如果个体 i 参加了项目，则可观测到 y_{1i}，但看不到 y_{0i}；除非让此人“穿越”到过去，改写历史而选择不参加此项目，才能观测到 y_{0i}。反之，如果个体 i 未参加项目，则可观测到 y_{0i}，但看不到 y_{1i}。总之，由于个体只能处于一种状态（要么参加项目，要

---

① “项目评估”在汉语中主要指投资项目的可行性分析，而“项目效应评估”（program evaluation）的含义完全不同。

② 直译为“治疗效应”，因为它最早盛行于医学领域对于药物疗效的评估。也译为“处置效应”。

③ 类似地，1993 年诺贝尔经济学奖得主罗伯特·福格尔（Robert Fogel）在使用“历史计量学”（Cliometrics）评估铁路对美国经济发展的贡献时，将 1890 年的美国经济与“假想”（hypothetical）的没有铁路的 1890 年美国经济进行了对比。

么不参加),故只能观测到 y_{0i} 或 y_{1i},而无法同时观测到 y_{0i} 与 y_{1i}。这实际上是一种"数据缺失"(missing data)问题。

表达式(28.1)将 y_i 写为分段函数。更简洁地,可将 y_i 写为

$$y_i = (1 - D_i)y_{0i} + D_i y_{1i} = y_{0i} + \underbrace{(y_{1i} - y_{0i})}_{\text{处理效应}} D_i \tag{28.2}$$

其中,$(y_{1i} - y_{0i})$ 为个体 i 参加项目的因果效应或处理效应(treatment effect)。显然,不同个体的处理效应可能不同,故应将 (y_{0i}, y_{1i}, D_i) 视为来自三维随机向量 (y_0, y_1, D) 总体的一个随机抽样(random draw)。假设样本为 iid,即对于任何 $i \neq j$,(y_{0i}, y_{1i}, D_i) 的概率分布与 (y_{0j}, y_{1j}, D_j) 相同,而且二者相互独立。这意味着不存在溢出效应,即个体 i 是否参加项目,不影响任何其他个体。此假定被称为"个体处理效应稳定假设"(Stable Unit Treatment Value Assumption,简记 SUTVA)。SUTVA 假定排除了个体间的社会互动(social interactions)或一般均衡效应(general equilibrium effects)。由于处理效应 $(y_{1i} - y_{0i})$ 为随机变量,故我们关心其期望值,即"平均处理效应"(Average Treatment Effect,简记 ATE),也称为"平均因果效应"(Average Causal Effect,简记 ACE):

$$\text{ATE} \equiv \text{E}(y_{1i} - y_{0i}) \tag{28.3}$$

ATE 表示从总体中随机抽取某个体的期望处理效应,无论该个体是否参与项目。有些学者批评此定义过于宽泛,因为总体中的某些个体可能根本无资格参加项目。比如,在估计培训项目的平均处理效应时,我们并不希望将百万富翁也包括在内。但 ATE 的定义依然十分有用,因为总可以通过重新定义总体而将无资格参加项目者排除在外,比如,限定培训前收入(pretraining income)在某临界水平之下。

另一常用概念为仅考虑项目实际参加者的平均处理效应,称为"参与者平均处理效应"(Average Treatment Effect on the Treated,简记 ATT 或 ATET)或"参与者处理效应"(Treatment Effect on the Treated,简记 TOT),即

$$\text{ATT} \equiv \text{E}(y_{1i} - y_{0i} \mid D_i = 1) \tag{28.4}$$

对于政策制定者而言,ATT 可能更为重要,因为它衡量的是项目参与者的毛收益(比如,可对比此毛收益与项目成本,进行成本收益分析)。在特殊情况下,ATE 与 ATT 可能相等,但一般不等。由于不能同时观测 y_{0i} 与 y_{1i},应如何估计 ATE 或 ATT?如果简单地比较项目参与者与未参与者的未来收入,则会导致选择偏差,因为

$$\begin{aligned}\underbrace{\text{E}(y_{1i} \mid D_i = 1) - \text{E}(y_{0i} \mid D_i = 0)}_{\text{参与者与未参与者的平均差异}} &= \underbrace{\text{E}(y_{1i} \mid D_i = 1) - \text{E}(y_{0i} \mid D_i = 1)}_{\text{ATT}} \\ &+ \underbrace{\text{E}(y_{0i} \mid D_i = 1) - \text{E}(y_{0i} \mid D_i = 0)}_{\text{选择偏差}}\end{aligned} \tag{28.5}$$

在上式中,先减去 $\text{E}(y_{0i} \mid D_i = 1)$,然后再加上它。上式将项目参与者与未参与者的平均收入之差分解为两部分,其中第一项为 ATT,而第二项为参与者的平均 y_{0i} 与未参与者的平均 y_{0i} 之差(即这两类人如果都未参与项目的收入差距),即选择偏差。由于低收入者通常更倾向于选择参加培训项目,故选择偏差一般为负,导致实验组与控制组的收入之差[即 $\text{E}(y_{1i} \mid D_i = 1) - \text{E}(y_{0i} \mid D_i = 0)$]低估参与者平均处理效应(ATT)。如果选择偏差的绝对值足够大,则可能导致 $\text{E}(y_{1i} \mid D_i = 1) - \text{E}(y_{0i} \mid D_i = 0) < 0$,出现参加培训者的收入反而低于未参加者的情形。

类似地,可以定义"非参与者平均处理效应"(Average Treatment Effect on the Untreated,简记 ATU)为

$$\text{ATU} \equiv \text{E}(y_{1i} - y_{0i} \mid D_i = 0) \tag{28.6}$$

由于个体通常会根据其参加项目的预期收益 $E(y_{1i}-y_{0i})$ 而自我选择是否参加项目，导致对平均处理效应的估计带来困难，这被称为“选择难题”(the selection problem)。

需要注意的是，在估计处理效应时面临的选择难题，与第 14 章的样本选择问题(Heckman, 1979)有所不同，后者考虑的是所获样本是否为总体的代表性样本。具体来说，样本选择问题通常不考虑某项目或政策的效应，故个体间的差异并不在于是否得到处理，而在于是否能进入样本(即被解释变量 y_i 是否可观测)，通常 $D_i=1$ 意味着 y_i 可观测，而 $D_i=0$ 则意味着 y_i 不可观测。而在处理效应模型中，无论 $D_i=1$ 或 0，结果变量 y_i 均可观测。

28.2 通过随机分组解决选择难题

选择难题并非不可克服。解决方法之一是通过随机分组(random assignment)[①]，使得个体 i 的 D_i(即是否参加项目)通过抛硬币或电脑随机数而决定，则 D_i 独立于 (y_{0i},y_{1i})。此时，ATE = ATT，因为 $E(y_{1i}-y_{0i}\mid D_i=1)=E(y_{1i}-y_{0i})$[由于 $(y_{1i}-y_{0i})$ 独立于 D_i，故条件期望等于无条件期望]。对 ATE 的估计也非常简单，只要比较实验组与控制组的平均收入即可，因为

$$E(y_{1i}\mid D_i=1)-E(y_{0i}\mid D_i=0)=E(y_{1i})-E(y_{0i})=\text{ATE}=\text{ATT} \tag{28.7}$$

上式能够成立，也是因为 D_i 独立于 (y_{0i},y_{1i})。因此，在随机分组的情况下，只需要计算样本中实验组与控制组的平均收入之差，即可一致地估计平均处理效应。此估计量即第 18 章介绍的“差额估计量”(differences estimator)，也称为“difference-in-means”，服从渐近正态分布。

事实上，上述结果在更弱的均值独立(mean independence)条件下也成立，即只要 y_{0i}，y_{1i} 都均值独立于 D_i，则表达式(28.7)依然成立。进一步，如果我们只关心 ATT，则只需要 y_{0i} 均值独立于 D_i 即可，不需要对 y_{1i} 与 D_i 的关系进行任何限制。这是因为，如果 y_{0i} 均值独立于 D_i，即 $E(y_{0i}\mid D_i)=E(y_{0i})$，则公式(28.5)中的选择偏差为 0，因为 $E(y_{0i}\mid D_i=1)-E(y_{0i}\mid D_i=0)=E(y_{0i})-E(y_{0i})=0$。由于随机分组可以很好地解决选择偏差，故随机实验或准实验的研究方法在经济学各领域日益盛行。

然而，随机分组并非在所有情况下都可行(可能成本太高)。如果只有观测数据(observational data)，则很可能不满足“y_{0i} 均值独立于 D_i”的假设，因为该假设意味着个体 i 是否参加项目(D_i)与其不参加项目的收入(y_{0i})不相关。因此，在大多数情况下，需要使用以下两类方法。第一类方法假设个体依可测变量选择是否参加项目(参见本章第 3 ~ 7 节)，而第二类方法假设个体依不可测变量选择(参见本章第 8 节)。

28.3 依可测变量选择

通常，除了 (y_i,D_i) 之外，还可以观测到个体 i 的一些特征，比如年龄、性别、培训前收入，记为向量 $\boldsymbol{x}_i$，也称为“协变量”(covariates)。这样，总体可由 $(y_0,y_1,D,\boldsymbol{x})$ 来表示。如果个体 i 对 D_i

① 有关随机实验，详见第 18 章。

的选择完全取决于可观测的$\boldsymbol{x}_i$，称为“依可测变量选择”(selection on observables)[①]，则可以找到估计处理效应的合适方法(即使没有合适的工具变量)。如果个体对D_i的选择完全取决于$\boldsymbol{x}_i$，则在给定$\boldsymbol{x}_i$的情况下，潜在结果(y_{0i}, y_{1i})将独立于D_i，这就是 Rosenbaum and Rubin(1983)所引入的“可忽略性”假设：

假定 28.1 可忽略性(Ignorability)。给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})独立于D_i，记为$(y_{0i}, y_{1i}) \perp D_i \mid \boldsymbol{x}_i$，其中“$\perp$”表示相互独立。

“可忽略性”的含义是，给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})对于D_i的影响可以忽略，故名。可忽略性也称为“无混淆性”(unconfoundedness)[②]，“条件独立假定”(Conditional Independence Assumption，简记CIA)，或“依可测变量选择”(selection on observables)。假定 28.1 意味着，给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})在处理组与控制组的分布完全一样，即

$$F(y_{0i}, y_{1i} \mid \boldsymbol{x}_i, D_i = 1) = F(y_{0i}, y_{1i} \mid \boldsymbol{x}_i, D_i = 0) \tag{28.8}$$

其中，$F(\cdot)$表示分布函数。在很多情况下，只需要更弱的均值独立即可：

假定 28.2 均值可忽略性(Ignorability in Mean)。$\mathrm{E}(y_{0i} \mid \boldsymbol{x}_i, D_i) = \mathrm{E}(y_{0i} \mid \boldsymbol{x}_i)$，而且$\mathrm{E}(y_{1i} \mid \boldsymbol{x}_i, D_i) = \mathrm{E}(y_{1i} \mid \boldsymbol{x}_i)$。这意味着，在给定$\boldsymbol{x}_i$的情况下，$y_{0i}$与$y_{1i}$都均值独立于$D_i$。

如果假定 28.1 或 28.2 成立，则原则上可将$\boldsymbol{x}_i$直接作为控制变量引入以下回归方程，以解决遗漏变量问题：

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \gamma D_i + \varepsilon_i \tag{28.9}$$

然而，我们并不清楚$\boldsymbol{x}_i$是否应以线性形式进入上述方程。如果遗漏了非线性项，则依然可能存在遗漏变量偏差。解决方法之一为基于鲁宾反事实框架的匹配估计量，参见下一节。从回归方程(28.9)也可以看出，可忽略性假定是一个很强的假定；它意味着回归方程已包括了所有相关变量，故不存在任何与解释变量相关的遗漏变量。另一方面，如果$\boldsymbol{x}_i$中已包含较丰富的协变量(a rich set of covariates)，则可认为可忽略性假定基本得到满足，遗漏变量偏差较小。

28.4 匹配估计量的思想

假设个体i属于处理组，匹配估计量的基本思路是，找到属于控制组的某个体j，使得个体j与个体i的可测变量取值尽可能相似(匹配)，即$\boldsymbol{x}_i \approx \boldsymbol{x}_j$。基于可忽略性假设，则个体$i$与个体$j$进入处理组的概率相近，具有可比性；故可将$y_j$作为$y_{0i}$的估计量，即$\hat{y}_{0i} = y_j$。因此，可将$(y_i - \hat{y}_{0i}) = y_i - y_j$作为对个体$i$处理效应的度量。对处理组中的每位个体都如此进行匹配；类似地，对控制组每位个体也进行匹配，然后对每位个体的处理效应进行平均，即可得到“匹配估计量”(matching estimators)。

由于匹配的具体方法不同，故存在不同的匹配估计量。这里有两个技术细节。首先，是否放回。如果不放回(no replacement)，则每次都将匹配成功的个体(i, j)从样本中去掉，不再参与其余匹配；如果有放回(with replacement)，则依然将匹配成功个体留在样本中，参与其余匹配(这导

---

① 这并不要求D_i是$\boldsymbol{x}_i$的确定性函数。仍然可允许除$\boldsymbol{x}_i$以外的随机因素对D_i产生影响，只要这些随机因素独立于$(\boldsymbol{x}_i, y_{0i}, y_{1i})$。

② “unconfoundedness”的含义是，只要把$\boldsymbol{x}_i$包括在回归方程中，就能完全解决遗漏变量偏差，从而避免各变量间作用关系的混淆(confounding)。

致一位个体可能与多位不同组个体匹配)。其次,是否允许并列(ties),比如控制组个体 j 与 k 的可测变量都与处理组个体 i 一样接近。如果允许并列,则将 y_j 与 y_k 的平均值作为 y_{0i} 的估计量,即 $\hat{y}_{0i}=(y_j+y_k)/2$。如果不允许并列,则计算机程序将根据数据排序选择个体 j 或 k;此时,匹配结果可能与数据排序有关,故一般建议先将样本随机排序①,再进行匹配。

以上为一对一(one-to-one)匹配,也可以进行一对多匹配,比如一对四匹配,即针对每位个体寻找四位不同组的最近个体进行匹配。一般来说,匹配估计量存在偏差(bias)②,除非在"精确匹配"(exact matching)的情况下,即对于所有匹配都有 $\boldsymbol{x}_i=\boldsymbol{x}_j$。更常见的为"非精确匹配"(inexact matching),即只能保证 $\boldsymbol{x}_i\approx\boldsymbol{x}_j$。在非精确匹配的情况下,如果进行一对一匹配,则偏差较小,但方差较大;而进行一对多匹配可降低方差(因为使用了更多信息),但代价是偏差增大(因为使用了更远的信息)。Abadie et al(2004)建议进行一对四匹配,在一般情况下可最小化均方误差(MSE)。

下面以一个具体例子来说明匹配的过程③。假设样本容量为 7,其中包括 3 位控制组个体与 4 位处理组个体。同时假设 $\boldsymbol{x}_i$ 仅包含一个变量 x_i。具体数据参见表 28.1。下面,进行有放回的一对一匹配,且允许并列。

表 28.1 匹配估计量的简单例子

i	D_i	x_i	y_i	匹配结果	$\hat{y}_{0i}$	$\hat{y}_{1i}$
1	0	2	7	{5}	7	8
2	0	4	8	{4,6}	8	7.5
3	0	5	6	{4,6}	6	7.5
4	1	3	9	{1,2}	7.5	9
5	1	2	8	{1}	7	8
6	1	3	6	{1,2}	7.5	6
7	1	1	5	{1}	7	5

首先,考虑个体 1 的匹配。由于个体 1 属于控制组($D_1=0$),故在处理组(即个体 4—7)中寻找最佳匹配。由于 $x_1=2$,而 $x_5=2$,故个体 1 的匹配为个体 5,记为{5}。因此,$\hat{y}_{01}=y_1=7$,而 $\hat{y}_{11}=y_5=8$。

其次,考虑个体 2 的匹配。由于 $x_2=4$,在处理组中没有完全相同的匹配,而最近匹配为 $x_4=x_6=3$,故个体 2 的匹配结果为{4,6}。因此,$\hat{y}_{02}=y_2=8$,而 $\hat{y}_{12}=(y_4+y_6)/2=7.5$。类似地,个体 3 的匹配结果也是{4,6}。

再次,考虑个体 4 的匹配。由于个体 4 属于处理组($D_4=1$),故在控制组(即个体 1—3)中寻找最佳匹配。由于 $x_4=3$,在控制组中没有完全相同的匹配,而最近匹配为 $x_1=2$ 与 $x_2=4$,故个体 4 的匹配结果为{1,2}。因此,$\hat{y}_{04}=(y_1+y_2)/2=7.5$,而 $\hat{y}_{14}=y_4=9$。类似地,可以得到个体 5—7 的匹配结果,参见表 28.1。

---

① 比如,产生服从均匀分布的一组随机数,然后根据此组随机数进行样本排序。为了保证结果的可重复性,在产生随机数之前,应确定随机数的"种子"(seed)。

② 正如非参数估计一般存在偏差,因为它使用了附近邻域的信息,参见第 27 章。

③ 此例来自 Abadie et al(2004)。

最后，在 Stata 中分别计算 ATE（考虑整个样本的匹配结果），ATT（只考虑参加者的匹配结果）以及 ATU（只考虑未参加者的匹配结果）。由于此例子很简单，故可手工计算。

```
. dis "ATE = "
((8 -7) + (7.5 -8) + (7.5 -6) + (9 -7.5) + (8 -7) + (6 -7.5) + (5 -7))/7
   ATE = .14285714
. dis "ATT = "((9 -7.5) + (8 -7) + (6 -7.5) + (5 -7))/4
   ATT = -.25
. dis "ATU = "((8 -7) + (7.5 -8) + (7.5 -6))/3
   ATU = .66666667
```

28.5 倾向得分匹配

更一般地，$\boldsymbol{x}_i$ 可能包括多个变量，比如 $\boldsymbol{x}_i$ 为 K 维向量。此时，如果直接使用 $\boldsymbol{x}_i$ 进行匹配，则意味着要在高维度空间进行匹配，可能遇到数据稀疏的问题，即很难找到与 $\boldsymbol{x}_i$ 相近的 $\boldsymbol{x}_j$ 与之匹配。为此，一般使用某函数 $f(\boldsymbol{x}_i)$，将 K 维向量 $\boldsymbol{x}_i$ 的信息压缩到一维，进而根据 $f(\boldsymbol{x}_i)$ 进行匹配。方法之一为使用向量范数（vector norm），即在向量空间（vector space）定义的距离函数。考虑 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的相似度或距离，定义“马氏距离”（Mahalanobis distance）为

$$d(i,j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)' \hat{\sum}_{x}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{28.10}$$

其中，二次型矩阵 $\hat{\sum}_{x}^{-1}$ 为 $\boldsymbol{x}$ 的样本协方差矩阵之逆矩阵，它的作用相当于“权重矩阵”（weighting matrix）[①]。使用马氏距离进行匹配，被称为“马氏匹配”（Mahalanobis matching）。有时，也使用主对角元素为各变量方差的对角矩阵之逆矩阵作为权重矩阵。通过协变量的某个距离函数进行匹配，统称为“协变量匹配”（covariate matching）。

马氏匹配的缺点是，如果 $\boldsymbol{x}$ 包括的变量较多或样本容量不够大，则不容易找到好的匹配，比如，尽管个体 j 与个体 i 的（相对）马氏距离最近，但绝对距离可能依然很远。为此，统计学家 Rosenbaum and Rubin（1983）提出使用“倾向得分”（propensity score，简记 p-score）来度量距离。

定义 个体 i 的倾向得分为，在给定 $\boldsymbol{x}_i$ 的情况下，个体 i 进入处理组的条件概率，即 $p(\boldsymbol{x}_i) \equiv \mathrm{P}(D_i = 1 \mid \boldsymbol{x} = \boldsymbol{x}_i)$，或简记 $p(\boldsymbol{x})$（省略下标 i）。

在使用样本数据估计 $p(\boldsymbol{x})$ 时，可使用参数估计（比如，probit 或 logit）或非参数估计（参见第 27 章），而最流行的方法为 logit。使用倾向得分来度量个体之间距离的好处在于，它不仅是一维变量，而且取值介于[0,1]之间。比如，即使 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 距离很远，但仍可能 $p(\boldsymbol{x}_i) \approx p(\boldsymbol{x}_j)$。

使用倾向得分作为距离函数进行匹配，称为“倾向得分匹配”（Propensity Score Matching，简记 PSM）。PSM 的理论依据在于，如果可忽略性假定成立，则只需在给定 $p(\boldsymbol{x})$ 的情况下，(y_{0i}, y_{1i}) 就独立于 D_i。

命题（倾向得分定理） $(y_0, y_1) \perp D \mid \boldsymbol{x} \Rightarrow (y_0, y_1) \perp D \mid p(\boldsymbol{x})$

证明： 由于 D 为虚拟变量，故只需证明 $\mathrm{P}[D = 1 \mid y_0, y_1, p(\boldsymbol{x})]$ 与 y_0, y_1 无关即可。

---

① 权重矩阵的作用相当于 GMM 估计量目标函数中的二次型，参见第 10 章。

$$
\begin{aligned}
&\mathrm{P}[D=1\mid y_0,y_1,p(\boldsymbol{x})]\\
&=\mathrm{E}[D\mid y_0,y_1,p(\boldsymbol{x})]\\
&=\mathrm{E}_{y_0,y_1,\boldsymbol{x}}[\mathrm{E}(D\mid y_0,y_1,\boldsymbol{x})\mid y_0,y_1,p(\boldsymbol{x})] \quad \text{(迭代期望定律)}\\
&=\mathrm{E}_{y_0,y_1,\boldsymbol{x}}[\mathrm{E}(D\mid \boldsymbol{x})\mid y_0,y_1,p(\boldsymbol{x})] \quad \text{(可忽略性假定)}\\
&=\mathrm{E}_{y_0,y_1,\boldsymbol{x}}[p(\boldsymbol{x})\mid y_0,y_1,p(\boldsymbol{x})]\\
&=p(\boldsymbol{x})
\end{aligned}
\tag{28.11}
$$

此证明的倒数第二行使用了以下关系式：

$$\mathrm{E}(D\mid \boldsymbol{x})=1\cdot\mathrm{P}(D=1\mid \boldsymbol{x})+0\cdot\mathrm{P}(D=0\mid \boldsymbol{x})=p(\boldsymbol{x}) \tag{28.12}$$

当然，为了能够进行匹配，需要在 $\boldsymbol{x}$ 的每个可能取值上都同时存在处理组与控制组的个体。这就是下面的“重叠假定”(overlap assumption)或“匹配假定”(matching assumption)。

假定 28.3 重叠假定。对于 $\boldsymbol{x}$ 的任何可能取值，都有 $0<p(\boldsymbol{x})<1$。

此假定意味着处理组与控制组这两个子样本存在重叠，故名“重叠假定”；另外，它又是进行匹配的前提，故也称“匹配假定”。它保证了处理组与控制组的倾向得分取值范围有相同的部分(common support)，参见图 28.1。如果假定 28.3 不成立，则意味着可能存在某些 $\boldsymbol{x}$，使得 $p(\boldsymbol{x})=1$，即这些个体都属于处理组，无法找到与之匹配的控制组个体；另一方面，也可能存在某些 $\boldsymbol{x}$，使得 $p(\boldsymbol{x})=0$，即这些个体都属于控制组，无法找到与其匹配的处理组个体。

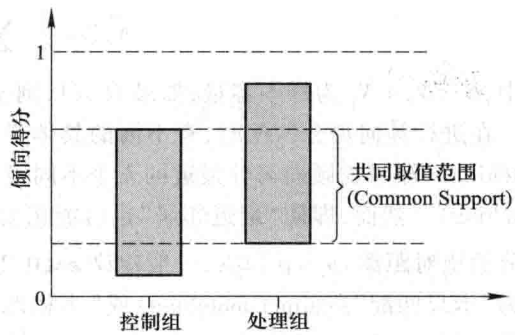


图 28.1 倾向得分的共同取值范围

在进行匹配时，为了提高匹配质量，通常仅保留倾向得分重叠部分的个体(尽管这样做会损失样本容量)。具体来说，如果某处理组个体的倾向得分高于控制组倾向得分的最大值或低于控制组倾向得分的最小值，则去掉该处理组个体(在使用 Stata 命令 psmatch2 时，可用选择项 common 来实现)。如果倾向得分的共同取值范围太小，则会导致偏差。

通过倾向得分匹配计算平均处理效应的一般步骤如下。

(1) 选择协变量 $\boldsymbol{x}_i$。尽量将可能影响 (y_{0i},y_{1i}) 与 D_i 的相关变量包括进来，以保证可忽略性假设得到满足。如果协变量 $\boldsymbol{x}_i$ 选择不当或太少，导致可忽略性假设不满足，将引起偏差。

(2) 估计倾向得分，一般使用 logit 回归。Rosenbaum and Rubin(1985)建议使用形式灵活的 logit 模型，比如包括 $\boldsymbol{x}_i$ 的高次项与互动项。

(3) 进行倾向得分匹配。如果倾向得分估计得较准确，则应使得 $\boldsymbol{x}_i$ 在匹配后的处理组与控制组之间分布较均匀，比如，匹配后的处理组均值 $\bar{\boldsymbol{x}}_{\text{treat}}$ 与控制组均值 $\bar{\boldsymbol{x}}_{\text{control}}$ 较接近；这个过程在统计学上称为“数据平衡”(data balancing)。但 $\bar{\boldsymbol{x}}_{\text{treat}}$ 与 $\bar{\boldsymbol{x}}_{\text{control}}$ 的差距显然与计量单位有关，故一般针对 $\boldsymbol{x}$ 的每个分量 x 考察如下“标准化差距”(standardized differences)或“标准化偏差”(standardized bias)：

$$\frac{|\bar{x}_{\text{treat}}-\bar{x}_{\text{contrl}}|}{\sqrt{(s^2_{x,\text{treat}}+s^2_{x,\text{control}})/2}} \tag{28.13}$$

其中，$s^2_{x,\text{treat}}$ 与 $s^2_{x,\text{control}}$ 分别为处理组与控制组变量 x 的样本方差。一般要求此标准化差距不超过

10%；如果超过，则应回到第(2)步、甚至第(1)步，重新估计倾向得分；或者改变具体的匹配方法(参见下文)。

(4) 根据匹配后样本(matched sample)计算平均处理效应。参加者平均处理效应(ATT)估计量的一般表达式为

$$\widehat{\mathrm{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} (y_i - \hat{y}_{0i}) \tag{28.14}$$

其中，$N_1 = \sum_i D_i$ 为处理组个体数，而 $\sum_{i:D_i=1}$ 表示仅对处理组个体进行加总。类似地，也可为控制组的每位个体 j 寻找处理组的相应匹配。未参加者平均处理效应(ATU)估计量的一般表达式为

$$\widehat{\mathrm{ATU}} = \frac{1}{N_0} \sum_{j:D_j=0} (\hat{y}_{1j} - y_j) \tag{28.15}$$

其中，$N_0 = \sum_j (1 - D_j)$ 为控制组个体数，而 $\sum_{j:D_j=0}$ 表示仅对控制组个体进行加总。整个样本(包括参加者与未参加者)的平均处理效应(ATE)估计量的一般表达式为

$$\widehat{\mathrm{ATE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_{1i} - \hat{y}_{0i}) \tag{28.16}$$

其中，$N = N_0 + N_1$ 为样本容量；如果 $D_i = 1$，则 $\hat{y}_{1i} = y_i$；如果 $D_i = 0$，则 $\hat{y}_{0i} = y_i$。

在进行倾向得分匹配时，有不同的具体方法。**方法之一**为“k 近邻匹配”(k-nearest neighbor matching)，即寻找倾向得分最近的 k 个不同组个体。如果 $k = 1$，则为“一对一匹配”(one-to-one matching)。然而，即使“最近邻居”也可能相去甚远，从而失去可比性。为此，**方法之二**限制倾向得分的绝对距离 $|p_i - p_j| \leqslant \varepsilon$，一般建议 $\varepsilon \leqslant 0.25\hat{\sigma}_{\mathrm{pscore}}$，其中 $\hat{\sigma}_{\mathrm{pscore}}$ 为倾向得分的样本标准差；这被称为“卡尺匹配”(caliper matching)或“半径匹配”(radius matching)。**方法之三**为“卡尺内最近邻匹配”(nearest-neighbor matching within caliper)，即在给定的卡尺 ε 范围寻找最近匹配，此法较为流行。

以上三种方法本质上都是近邻匹配法，其匹配结果为最近的部分个体，然后进行简单算术平均。另一类匹配方式为整体匹配法，每位个体的匹配结果为不同组的全部个体(但通常去掉在 common support 之外的个体)，只是根据个体距离不同给予不同的权重(近者权重大，远者权重小，超出一定范围权重可为0)。比如，在使用表达式(28.14)估计 ATT 时，$\hat{y}_{0i}$ 的估计量为

$$\hat{y}_{0i} = \sum_{j:D_j=0} w(i,j) y_j \tag{28.17}$$

其中，$w(i,j)$ 为适用于配对 (i,j) 的权重。如果使用核函数来计算权重 $w(i,j)$，则为**方法之四**“核匹配”(kernel matching)(Heckman et al,1997,1998)，其权重表达式为

$$w(i,j) = \frac{K[(\boldsymbol{x}_j - \boldsymbol{x}_i)/h]}{\sum_{k:D_k=0} K[(\boldsymbol{x}_k - \boldsymbol{x}_i)/h]} \tag{28.18}$$

其中，h 为指定带宽(bandwidth)，$K(\cdot)$ 为核函数。将表达式(28.18)代入方程(28.17)，可将 $\hat{y}_{0i}$ 视为“核回归估计量”(kernel regression estimator)，参见第27章。如果不进行核回归，而使用局部线性回归来估计 $w(i,j)$，则为**方法之五**“局部线性回归匹配”(local linear regression matching)。**方法之六**使用更为光滑的“三次样条”(cubic spline)来估计 $w(i,j)$，称为“样条匹配”(spline matching)。

在实际进行匹配时，究竟应使用以上哪种具体方法或参数(比如，k 近邻匹配的 k 取值，是否放回，如何处理并列)，目前文献中尚无明确指南。一般认为，不存在适用于一切情形的绝对好方法，只能根据具体数据来选择匹配方法。比如，如果控制组个体并不多(N_0 较小)，则应进行有

放回的匹配。又比如,如果存在较多具有可比性的控制组个体,则可考虑一对多匹配或核匹配,以提高匹配效率。在实践中,一般建议尝试不同的匹配方法,然后比较其结果(类似于敏感度分析);如果不同方法的结果相似,则说明结果是稳健的,不依赖于具体方法;反之,如果存在较大差异,则应考察造成此差异的原因。

在一定意义上,匹配估计量可视为一种再抽样方法(resampling)。因此,在方法论上,PSM 试图通过匹配再抽样的方法使得观测数据尽可能地接近随机实验数据,其思想可以追溯到费舍尔(Ronald Fischer)提出的随机实验设计。然而,尽管 PSM 可能在很大程度上减少观测数据的偏差,但它本身也有如下局限性:

(1) PSM 通常要求比较大的样本容量以得到高质量的匹配。

(2) PSM 要求处理组与控制组的倾向得分有较大的共同取值范围(common support);否则,将丢失较多观测值,导致剩下的样本不具有代表性。

(3) PSM 只控制了可测变量的影响,如果存在依不可测变量选择(selection on unobservable),仍会带来"隐性偏差"(hidden bias)。

28.6 倾向得分匹配的 Stata 实例

倾向得分匹配可通过下载非官方命令 psmatch2 来实现:

```
ssc install psmatch2,replace
```

其中,选择项"replace"表示以该命令的最新版本替代计算机中可能已有的旧版命令。由于该命令仍在不断更新中,故建议使用选择项"replace"。

该命令的一般格式为

```
psmatch2 D x1 x2 x3,outcome(y) logit ties ate common odds pscore(varname) quietly
```

其中,"D"为处理变量(treatment variable),"x1 x2 x3"为协变量,"outcome(y)"用来指定变量"y"为结果变量(outcome variable)。选择项"logit"表示使用 logit 来估计倾向得分,默认方法为 probit。选择项"ties"表示包括所有倾向得分相同的并列个体,默认按照数据排序选择其中一位个体。选择项"ate"表示同时汇报 ATE,ATU 与 ATT,默认仅汇报 ATT。选择项"common"表示仅对共同取值范围(common support)内个体进行匹配,默认对所有个体进行匹配。选择项"odds"表示使用几率比(odds ratio,即 $p/(1-p)$)进行匹配;默认使用倾向得分 p 进行匹配。选择项"pscore(varname)"用来指定某变量作为倾向得分,默认通过"x1 x2 x3"来估计倾向得分。选择项"quietly"表示不汇报对倾向得分的估计过程。

针对不同的匹配方法,命令 psmatch2 提供了一系列选择项。

(1) psmatch2 D x1 x2 x3,outcome(y) neighbor(k) noreplacement

选择项"neighbor(k)"表示进行 k 近邻匹配(k 为正整数);默认 $k=1$,即一对一匹配。选择项"noreplacement"表示无放回匹配,默认为有放回;该选项只能用于一对一匹配。

(2) psmatch2 D x1 x2 x3,outcome(y) radius caliper(real)

选择项"radius"表示进行半径匹配(也称卡尺匹配),其中"caliper(real)"用来指定卡尺 ε,必须为正实数。

(3) psmatch2 D x1 x2 x3,outcome(y) neighbor(k) caliper(real)

选择项“neighbor(k)”与“caliper(real)”表示进行卡尺内的 k 近邻匹配。

(4) psmatch2 D x1 x2 x3,outcome(y) kernel kerneltype(type) bwidth(real)

选择项“kernel”表示进行核匹配,其中“kerneltype(type)”用来指定核函数,默认使用二次核(epan kernel),“bwidth(real)”用来指定带宽,默认带宽为 0.06。

(5) psmatch2 D x1 x2 x3,outcome(y) llr kerneltype(type) bwidth(real)

选择项“llr”表示进行局部线性回归匹配,其中“kerneltype(type)”用来指定核函数,默认使用三三核(tricubic kernel),“bwidth(real)”用来指定带宽,默认带宽为 0.8。

(6) psmatch2 D x1 x2 x3,outcome(y) spline

选择项“spline”表示进行样条匹配。

(7) psmatch2 D x1 x2 x3,outcome(y) mahal(varlist) ai(m)

选择项“mahal(varlist)”表示进行马氏匹配,并指定用于计算马氏距离的协变量。选择项“ai(m)”表示使用由 Abadie and Imbens(2006)提出的异方差稳健标准误,该选项仅适用于使用马氏距离的 k 近邻匹配;其中 m 须为正整数,表示用于计算稳健标准误的近邻个数(一般可让 $m=k$)。使用此命令进行马氏匹配时,无法使用选择项“ties”或“common”。

命令 psmatch2 还带有以下两个“估计后命令”(post-estimation commands),分别用来检验匹配后数据是否平衡,以及画图显示倾向得分的共同取值范围。

pstest x1 x2 x3,both graph

此命令将显示变量“x1 x2 x3”在匹配后是否平衡,选择项“both”表示同时显示匹配前的数据平衡情况,默认仅显示匹配后情形。选择项“graph”表示图示各变量匹配前后的平衡情况。

psgraph,bin(#)

此命令将画直方图,显示倾向得分的共同取值范围,选择项“bin(#)”用来指定直方图的分组数,默认为 20 组(处理组与控制组各分为 10 组)。

下面以数据集 ldw_exper. dta 为例进行演示。该数据集由 Dehejia and Wahba(1999)构建,为 Abadie et al(2004)所使用,原始数据来自 Lalonde(1986)。该数据集包括以下变量:结果变量 re78(1978 年实际收入),处理变量 t(是否参加就业培训),协变量 age(年龄),educ(教育年限),black(是否黑人),hisp(是否拉丁裔),married(是否已婚),re74(1974 年实际收入),re75(1975 年实际收入),u74(1974 年是否失业),以及 u75(1975 年是否失业)。

. use ldw_exper.dta,clear

作为参照,首先进行一元回归。

. reg re78 t,r

Linear regression

					Number of obs =	445
					F(1, 443) =	7.15
					Prob > F =	0.0078
					R-squared =	0.0178
					Root MSE =	6.5795

re78	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
t	1.794343	.6708247	2.67	0.008	.475949	3.112737
_cons	4.554802	.3402038	13.39	0.000	3.886188	5.223416

上表显示,在未控制任何协变量的情况下,平均处理效应为 1.794,即参加就业培训平均能使 1978 年实际收入提高 1 794 美元(变量 re78 的单位为千美元),且在 1% 水平上显著。由于可能存在选择偏差,此结果并不可信。而且,R^2很低,仅为 0.017 8(即是否参加就业培训仅能解释 1978 年实际收入 1.78% 的变动)。

下面,直接引入协变量,进行更为可信的多元回归。

```
. reg re78 t age educ black hisp married re74 re75 u74 u75,r
```

```
Linear regression                               Number of obs =     445
                                                F( 10,    434) =    2.53
                                                Prob > F      =  0.0057
                                                R-squared     =  0.0582
                                                Root MSE      =  6.5093
```

re78	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
t	1.672042	.6617972	2.53	0.012	.3713161	2.972768
age	.0536677	.040388	1.33	0.185	-.0257127	.133048
educ	.4029471	.1610925	2.50	0.013	.0863287	.7195655
black	-2.039466	1.038581	-1.96	0.050	-4.080739	.0018068
hisp	.4246486	1.427471	0.30	0.766	-2.380968	3.230265
married	-.1466618	.8640396	-0.17	0.865	-1.844884	1.551561
re74	.1235727	.127147	0.97	0.332	-.1263278	.3734731
re75	.0194585	.14063	0.14	0.890	-.2569421	.2958591
u74	1.380999	1.554643	0.89	0.375	-1.674566	4.436564
u75	-1.071817	1.408301	-0.76	0.447	-3.839755	1.696121
_cons	.2214288	2.824293	0.08	0.938	-5.329565	5.772422

上表显示,加入协变量后,平均处理效应降为 1.672(变化不大),且显著性水平接近 1%(p 值为 1.2%)。在协变量中,除了 educ(教育年限)与 black(是否黑人)在 5% 水平上显著外,其余协变量均不显著。

下面进行倾向得分匹配。为此,先将数据随机排序。

```
. set seed 10101
. gen ranorder = runiform()
. sort ranorder
```

下面进行一对一匹配。由于样本容量并不十分大,进行有放回匹配,且允许并列(如果进行无放回匹配,将损失约 1/5 样本)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75,outcome
(re78) n(1) ate ties logit common
```

```
Logistic regression                                   Number of obs   =        445
                                                      LR chi2(9)      =      11.70
                                                      Prob > chi2     =     0.2308
Log likelihood = -296.25026                           Pseudo R2       =     0.0194

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0142619   .0142116     1.00   0.316    -.0135923    .0421162
        educ |   .0499776   .0564116     0.89   0.376     -.060587    .1605423
       black |   -.347664   .3606532    -0.96   0.335    -1.054531    .3592032
        hisp |  -.9284851    .50661    -1.83   0.067    -1.921422    .0644521
     married |   .1760431   .2748817     0.64   0.522    -.3627151    .7148012
        re74 |  -.0339278   .0292559    -1.16   0.246    -.0912683    .0234127
        re75 |     .01221   .0471351     0.26   0.796    -.0801731    .1045932
        u74 |  -.1516037   .3716369    -0.41   0.683    -.8799987    .5767913
        u75 |  -.3719486    .317728    -1.17   0.242    -.9946841    .2507869
      _cons |  -.4736308   .8244205    -0.57   0.566    -2.089465    1.142204
------------------------------------------------------------------------------
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.

------------------------------------------------------------------------------
    Variable |     Sample |    Treated     Controls   Difference          S.E.   T-stat
-------------+------------+----------------------------------------------------------
        re78 |  Unmatched |  6.34914538   4.55480228   1.79434311   .632853552     2.84
             |        ATT |  6.40495818   4.99436488    1.4105933   .839875971     1.68
             |        ATU |  4.52683013   6.15618973    1.6293596            .        .
             |        ATE |                           1.53668776            .        .
------------------------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

  psmatch2:  |  psmatch2: Common
  Treatment  |       support
assignment   | Off suppo  On suppor |     Total
-------------+----------------------+----------
  Untreated  |        11        249 |       260
    Treated  |         2        183 |       185
-------------+----------------------+----------
       Total |        13        432 |       445
```

上表上部显示了 logit 回归的结果。上表中部显示，ATT 估计值为 1.411，对应的 t 值为 1.68，小于 1.96 的临界值，故不显著。ATE 与 ATU 的估计值与 ATT 类似，但不汇报标准误。其中，“Unmatched”汇报匹配前样本估计的结果，与前面一元回归的结果完全一样。上表下部汇报观测值是否在共同取值范围中。在总共 445 个观测值中，控制组（Untreated）共有 11 个不在共同取值范围中（off support），处理组（Treated）共有 2 个不在共同取值范围中（off support），其余 432 个观测值均在共同取值范围中（on support）。

上表中部的 Note 显示，所汇报的标准误并未考虑倾向得分为估计所得的事实（即假设倾向得分为真实值，然后推导标准误）[①]；此标准误的另一假设为同方差，也可能不成立。为此，可以考虑使用自助法来得到标准误（参见第 19 章），尽管自助标准误也未必正确（Abadie and Imbens, 2008）。

```
. set seed 10101
. bootstrap r(att) r(atu) r(ate), reps(500): psmatch2 t age educ black
hisp married re74 re75 u74 u75, outcome(re78) n(1) ate ties logit common
```

① 这个理论问题正在得到解决，参见 Abadie and Imbens(2012)。

```
Bootstrap results                                 Number of obs      =       445
                                                  Replications       =       500

      command:  psmatch2 t age educ black hisp married re74 re75 u74 u75,
                    outcome(re78) n(1) ate ties logit common
        _bs_1:  r(att)
        _bs_2:  r(atu)
        _bs_3:  r(ate)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
_bs_1	1.410593	.9303588	1.52	0.129	-.4128764	3.234063
_bs_2	1.62936	.991854	1.64	0.100	-.3146385	3.573358
_bs_3	1.536688	.8059366	1.91	0.057	-.042919	3.116295

上表显示，ATT 的自助标准误为 0.93。上表还汇报了前面不曾给出的 ATU 与 ATE 标准误。根据这些自助标准误可知，ATE 与 ATU 在 10% 水平上显著（ATE 的 p 值为 0.057，接近 5%），而 ATT 并不显著。

下面，使用命令 pstest 来考察此匹配结果是否较好地平衡了数据。

```
.quietly psmatch2 t age educ black hisp married re74 re75 u74 u75,
outcome(re78) n(1) ate ties logit common
. pstest age educ black hisp married re74 re75 u74 u75,both graph
```

Variable	Unmatched Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	25.816	25.054	10.7		1.12	0.265
	Matched	25.781	25.383	5.6	47.7	-0.49	0.626
educ	Unmatched	10.346	10.088	14.1		1.50	0.135
	Matched	10.322	10.415	-5.1	63.9	-0.23	0.816
black	Unmatched	.84324	.82692	4.4		0.45	0.649
	Matched	.85246	.86339	-2.9	33.0	1.02	0.307
hisp	Unmatched	.05946	.10769	-17.5		-1.78	0.076
	Matched	.06011	.04372	5.9	66.0	0.02	0.982
married	Unmatched	.18919	.15385	9.4		0.98	0.327
	Matched	.18579	.19126	-1.4	84.5	-1.91	0.057
re74	Unmatched	2.0956	2.107	-0.2		-0.02	0.982
	Matched	2.0672	1.9222	2.7	-1166.6	-1.00	0.317
re75	Unmatched	1.5321	1.2669	8.4		0.87	0.382
	Matched	1.5299	1.6446	-3.6	56.7	-2.42	0.016
u74	Unmatched	.70811	.75	-9.4		-0.98	0.326
	Matched	.71038	.75956	-11.1	-17.4	0.78	0.436
u75	Unmatched	.6	.68462	-17.7		-1.85	0.065
	Matched	.60656	.63388	-5.7	67.7	2.07	0.039

上表显示,匹配后(Matched)大多数变量的标准化偏差(% bias)小于10%,只是变量u74的偏差为11.1%,似乎可以接受;而且大多数t检验的结果不拒绝处理组与控制组无系统差异的原假设(re75与u75为例外)。对比匹配前(Unmatched)的结果,大多数变量的标准化偏差均大幅缩小,但变量re74与u74的偏差反而有所增加。

Summary of the distribution of the abs(bias)

BEFORE MATCHING

Percentiles		Smallest		
1%	.2159919	.2159919		
5%	.2159919	4.388661		
10%	.2159919	8.386325	Obs	9
25%	8.386325	9.36407	Sum of Wgt.	9
50%	9.414047		Mean	10.19509
		Largest	Std. Dev.	5.726406
75%	14.12198	10.72771		
90%	17.68094	14.12198	Variance	32.79173
95%	17.68094	17.45611	Skewness	-.2354408
99%	17.68094	17.68094	Kurtosis	2.252342

AFTER MATCHING

Percentiles		Smallest		
1%	1.447804	1.447804		
5%	1.447804	2.735669		
10%	1.447804	2.93891	Obs	9
25%	2.93891	3.62947	Sum of Wgt.	9
50%	5.094981		Mean	4.906021
		Largest	Std. Dev.	2.787843
75%	5.709197	5.613233		
90%	11.05192	5.709197	Variance	7.772069
95%	11.05192	5.933004	Skewness	1.057563
99%	11.05192	11.05192	Kurtosis	3.751284

Sample	Pseudo R2	LR chi2	p>chi2	MeanBias	MedBias
Raw	0.019	11.75	0.227	10.2	9.4
Matched	0.018	10.65	0.301	4.9	5.1

上表主要显示了匹配前后偏差绝对值的分布特征。各变量标准化偏差的匹配前后变化,参见图28.2。从图28.2可以直观地看出,大多数变量的标准化偏差在匹配后缩小了。

下面,画条形图来显示倾向得分的共同取值范围。

```
. psgraph
```

此命令的输出结果参见图28.3。

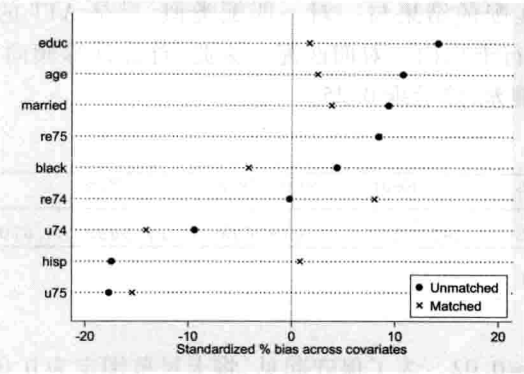


图 28.2 各变量的标准化偏差图示

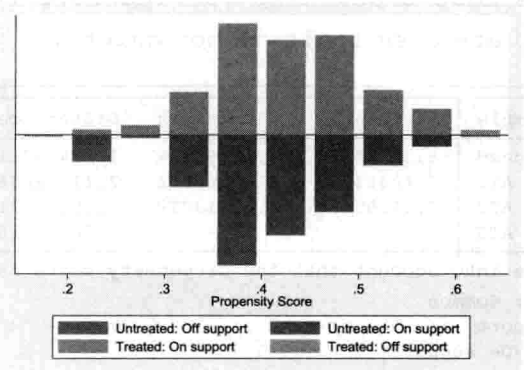


图 28.3 倾向得分的共同取值范围

从图 28.3 可以直观地看出,大多数观测值均在共同取值范围内(on support),故在进行倾向得分匹配时仅会损失少量样本。

下面,进行 k 近邻匹配,并令 $k=4$。为节省空间,使用选择项"quietly"略去对倾向得分估计结果的汇报。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75,outcome
(re78) n(4) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.32359171	2.08136646	.727464025	2.86
	ATU	4.52683013	5.80969564	1.28286551	.	.
	ATE			1.62111939	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support Off suppo	On suppor	Total
Untreated	11	249	260
Treated	2	183	185
Total	13	432	445

从上表可知,一对四匹配的结果与一对一匹配类似,只是 ATT 的估计值有较大差异,且在5%水平上显著。下面,进行卡尺内一对四匹配。为此,首先计算倾向得分(由 Stata 自动生成并记录于变量_pscore)的标准差,然后乘 0.25。

```
. sum _pscore
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_pscore	445	.4157303	.0791695	.1556892	.6102826

```
. dis 0.25 * r(sd)
. 01979237
```

由此可知,$0.25\hat{\sigma}_{\text{pscore}} \approx 0.02$。为了保守起见,将卡尺范围定为 0.01,这意味着对倾向得分相差1%的观测值进行一对四匹配。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome
(re78) n(4) cal(0.01) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.29263811	2.11232006	.729511502	2.90
	ATU	4.51395698	5.70213019	1.18817321	.	.
	ATE			1.58423615	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support Off suppo	On suppor	Total
Untreated	16	244	260
Treated	2	183	185
Total	18	427	445

上表显示,卡尺内一对四匹配的结果与简单的一对四匹配比较接近,这说明大多数一对四匹配均发生在卡尺 0.01 的范围内,不存在太远的“近邻”。下面,进行半径(卡尺)匹配。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome
(re78) radius cal(0.01) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.54598746	1.85897071	.704721737	2.64
	ATU	4.51395698	6.14730708	1.63335009	.	.
	ATE			1.73004464	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support Off suppo	On suppor	Total
Untreated	16	244	260
Treated	2	183	185
Total	18	427	445

此匹配结果依然类似。下面,进行核匹配(使用默认的核函数与带宽)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome
(re78) kernel ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.58467178	1.8202864	.685942087	2.65
	ATU	4.52683013	6.13018211	1.60335198	.	.
	ATE			1.69524781	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	11	249	260
Treated	2	183	185
Total	13	432	445

结果依然类似。然后,进行局部线性回归匹配(使用默认的核函数与带宽)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75,outcome
(re78) llr ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.51340325	1.89155492	.	.
	ATU	4.52683013	6.27518361	1.74835348	.	.
	ATE			1.8090152	.	.

Note: Sample S.E.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	11	249	260
Treated	2	183	185
Total	13	432	445

上表未汇报 ATT 的标准误;为此,使用自助法以得到自助标准误(费时较长)。

```
. set seed 10101
. bootstrap r(att) r(atu) r(ate),reps(500): psmatch2 t age educ black
hisp married re74 re75 u74 u75, outcome(re78) llr ate ties logit
common quietly
```

Bootstrap results — Number of obs = 445; Replications = 500

command: psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78) llr ate ties logit common quietly

_bs_1: r(att)
_bs_2: r(atu)
_bs_3: r(ate)

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
_bs_1	1.891555	.7049953	2.68	0.007	.5097895	3.27332
_bs_2	1.748353	.7959967	2.20	0.028	.1882286	3.308478
_bs_3	1.809015	.6743811	2.68	0.007	.4872525	3.130778

根据上表的自助标准误,对平均处理效应的三种度量均至少在5%水平上显著。

下面,进行样条匹配(同样使用自助法计算标准误)。为此,需要先安装一个非官方命令spline。

```
. net install snp7_1.pkg  (下载安装命令 spline)①
. set seed 10101
. bootstrap r(att) r(atu) r(ate),reps(500): psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78) spline ate ties logit common quietly
```

```
Bootstrap results                                 Number of obs      =       445
                                                  Replications       =       500

      command:  psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78)
                  spline ate ties logit common quietly
        _bs_1:  r(att)
        _bs_2:  r(atu)
        _bs_3:  r(ate)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
_bs_1	1.749681	.6590988	2.65	0.008	.457871	3.041491
_bs_2	1.576453	.7084574	2.23	0.026	.187902	2.965004
_bs_3	1.649834	.6532842	2.53	0.012	.3694209	2.930248

估计结果仍然类似。总之,以上各种倾向得分匹配的结果表明,参加就业培训的平均处理效应为正,不仅在经济上显著(平均可使1978年实际收入增加近两千美元),而且在统计上显著(大多数在5%水平上显著)。

下面,进行马氏匹配,并使用Adabie and Imbens(2006)提供的异方差稳健标准误。

```
. psmatch2 t,outcome(re78) mahal(age educ black hisp married re74 re75 u74 u75) n(4) ai(4) ate
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.34914538	4.42361842	1.92552697	.707891623	2.72
	ATU	4.55480228	6.50818167	1.95337939	.902160108	2.17
	ATE			1.94180029	.785347215	2.47

Note: Sample S.E.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	260	260
Treated	185	185
Total	445	445

① 或输入命令"findit snp7_1"寻找下载地址。

上表显示，无论是平均处理效应的估计值还是显著性，马氏匹配的结果与倾向得分匹配类似；这也说明了以上结果的稳健性。

28.7 偏差校正匹配估计量

由于在倾向得分匹配第一阶段估计倾向得分时存在不确定性（可使用 probit，logit 或非参数估计；即使确定用 logit，模型的具体设定仍取决于研究者），Abadie and Imbens（2002，2004，2006，2011）又重新回到更简单的马氏距离，进行有放回且允许并列（ties）的 k 近邻匹配。这样，研究者在计算匹配估计量时，只需做少量的主观决定。

更重要的是，由于非精确匹配（inexact matching）一般存在偏差，Abadie and Imbens 提出了偏差校正的方法，通过回归的方法来估计偏差，然后得到“偏差校正匹配估计量”（bias-corrected matching estimator）。另外，Abadie and Imbens 还通过在处理组或控制组内部进行二次匹配，来得到在异方差条件下也成立的稳健标准误。

偏差校正匹配估计量可通过非官方命令 nnmatch 来实现，其下载方法为

```
ssc install nnmatch,replace     (下载安装命令 nnmatch)
```

该命令的基本句式为

```
nnmatch y D x1 x2 x3,metric(maha) tc(att) tc(atc) m(k) robust(#) biasadj
(bias|varlist) pop
```

其中，选择项“metric(maha)”表示使用马氏距离，即权重矩阵为样本协方差矩阵的逆矩阵；默认权重矩阵是主对角线元素为各变量样本方差的对角矩阵之逆矩阵（称为 inverse variance）。选择项“tc(att)”表示估计 ATT，选择项“tc(atc)”表示估计 ATU（此处 c 表示 control，是 untreated 的同义词），默认值为 tc(ate)，即估计 ATE。选择项“m(k)”表示进行 k 近邻匹配，默认值为 $k=1$。选择项“robust(#)”表示计算异方差稳健的标准误，其中#须为正整数，表示用于计算稳健标准误的近邻个数（一般可让# $=k$）。选择项“biasadj(bias)”表示根据原来的协变量进行偏差校正，默认不进行偏差校正；也可以通过选择项“biasadj(varlist)”来指定用于偏差校正的变量名单。选择项“pop”表示估计“总体平均处理效应”（Population Average Treatment Effects，简记 PATE），其估计值与一般的“样本平均处理效应”（Sample Average Treatment Effects，简记 SATE）相同，只是标准误略微不同；默认估计 SATE。

下面以数据集 ldw_exper. dta 为例。首先，通过一对四匹配来估计 ATT，不做偏差校正，但使用异方差稳健标准误。

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att)
m(4)robust(4)
```

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att) m(4) robust(4)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
                                            Number of matches,
                                              robust std. err. (h) =         4

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |    1.994622   .7526339     2.65   0.008     .5194864    3.469757
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married re74 re75 u74 u75
```

上表显示，权重矩阵为默认的“inverse variance”，即主对角线元素为各变量样本方差的对角矩阵之逆矩阵。ATT 的估计值为 1.995，且在 1% 水平上显著。

其次，重复以上命令，但进行偏差校正。

. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att) m(4) robust(4) bias(bias)

```
Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =       445
                                            Number of matches  (m) =         4
                                            Number of matches,
                                              robust std. err. (h) =         4

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |    1.838424   .7526339     2.44   0.015      .363289     3.31356
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married re74 re75 u74 u75
Bias-adj variables:  age educ black hisp married re74 re75 u74 u75
```

上表显示，经过偏差校正后，ATT 的估计值减少为 1.838，且仅在 5% 水平上显著（p 值为 1.5%）。最后，以样本协方差矩阵的逆矩阵为权重矩阵，使用马氏距离进行匹配。

. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att) m(4) robust(4) bias(bias) metric(maha)

```
Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: Mahalanobis               Number of obs          =       445
                                            Number of matches  (m) =         4
                                            Number of matches,
                                              robust std. err. (h) =         4

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |    1.796419   .7489237     2.40   0.016     .3285557    3.264283
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married re74 re75 u74 u75
Bias-adj variables:  age educ black hisp married re74 re75 u74 u75
```

由上表可知，无论使用哪种加权矩阵，对估计结果影响不大。

28.8 双重差分倾向得分匹配

以上介绍的各种匹配估计量均依赖于可忽略性假定，即依可测变量选择；故不适用于依不可测变量选择的情形。对于观测数据，如果怀疑存在依不可测变量选择，则大致有以下几种处理方法。

（1）尽量使用更多的相关可测变量，以满足可忽略性假定，然后使用匹配估计量。显然，如果$\boldsymbol{x}_i$中包括的变量太少，则不太可能满足可忽略性假定。

（2）如果影响处理变量D_i的不可观测变量不随时间而变，而且有面板数据，则可使用“双重差分倾向得分匹配估计量”（differences-in-differences PSM estimator）。

（3）使用断点回归法，特别是模糊断点回归，参见本章第9～12节。

（4）使用工具变量法，比如，Imbens and Angrist(1994)提出的局部平均处理效应IV估计量，以及本章第13节的处理效应模型。工具变量法的最大局限在于，通常很难找到有效的工具变量，因为此工具变量既要影响个体选择参与项目的决定D_i，但又必须与是否参加项目的潜在结果(y_{1t}, y_{0i})无关。

（5）根据依可测变量选择的影响来估计依不可测变量选择的影响，参见Altonji et al(2005)。

本节主要关注第(2)种方法，即双重差分PSM，由Heckman et al(1997,1998)所提出。假设有两期面板数据，记实验前的时期为t'，实验后的时期为t。在时期t'，实验还未发生，故所有个体（无论处理组还是控制组）的潜在结果均可记为$y_{0t'}$。在时期t，实验已经发生，故可能有两种潜在结果，分别记为y_{1t}（如果参与实验）与y_{0t}（如果未参与实验）。

双重差分PSM成立的前提为以下均值可忽略性假定：

$$\mathrm{E}(y_{0t}-y_{0t'}\mid \boldsymbol{x}, D=1)=\mathrm{E}(y_{0t}-y_{0t'}\mid \boldsymbol{x}, D=0) \tag{28.19}$$

如果假定(28.19)成立，则可一致地估计ATT：

$$\widehat{ATT}=\frac{1}{N_1}\sum\nolimits_{i:i\in I_1\cap S_p}\left[(y_{1ti}-y_{0t'i})-\sum\nolimits_{j:j\in I_0\cap S_p} w(i,j)(y_{0tj}-y_{0t'j})\right] \tag{28.20}$$

其中，S_p为共同取值范围的集合（common support），$I_1=\{i: D_i=1\}$（处理组的集合），$I_0=\{i: D_i=0\}$（控制组的集合），N_1为集合$I_1\cap S_p$所包含的处理组个体数，而$w(i,j)$为对应于配对(i,j)的权重，可通过核匹配或局部线性回归匹配的方法来确定（参见本章第5节）。表达式(28.20)的方括弧内第一项$(y_{1ti}-y_{0t'i})$为处理组个体i实验前后的变化，而第二项中的$(y_{0tj}-y_{0t'j})$则为控制组个体j的前后变化。概括起来，双重差分PSM法的步骤如下。

第一步，根据处理变量D_i与协变量$\boldsymbol{x}_i$估计倾向得分。

第二步，对于处理组的每位个体i，确定与其匹配的全部控制组个体（即确定集合S_p）。

第三步，对于处理组的每位个体i，计算其结果变量的前后变化$(y_{1ti}-y_{0t'i})$。

第四步，对于处理组的每位个体i，计算与其匹配的全部控制组个体的前后变化$(y_{0tj}-y_{0t'j})$，其中$j\in I_0\cap S_p$。

第五步，针对$(y_{1ti}-y_{0t'i})$与$(y_{0tj}-y_{0t'j})$，根据公式(28.20)进行倾向得分核匹配或局部线性回归匹配，即得到$\widehat{ATT}$。

双重差分PSM法的优点在于它可以控制不可观测（unobservable）但不随时间变化（time invariant）的组间差异，比如处理组与控制组分别来自两个不同的区域，或处理组与控制组使用了不同的调查问卷。

在Stata中操作双重差分PSM法，既可手工进行（手工计算每位个体的前后变化，然后使用

命令 psmatch2),也可使用非官方命令 diff 来自动进行(该命令可用于估计一般的双重差分估计量,参见第 18 章)。

该命令的下载方法为

```
ssc install diff,replace
```

使用命令 diff 进行双重差分 PSM 估计的基本句式为

```
diff outcome_var,treat(varname) period(varname) id(varname) kernel
ktype(kernel) cov(varlist) report logit support test
```

其中,"outcome_var"为结果变量,必选项"treat(varname)"用来指定处理变量,必选项"period(varname)"用来指定实验期虚拟变量(实验期 = 1,非实验期 = 0)。必选项"id(varname)"用来指定个体 ID(这是进行匹配的前提)。必选项"kernel"表示进行基于倾向得分的核匹配(命令 diff 不提供其他匹配方法),选择项"ktype(kernel)"用于指定核函数,默认为二次核。必选项"cov(varlist)"用来指定用于估计倾向得分的协变量,选择项"report"表示汇报对倾向得分的估计结果。选择项"logit"表示使用 Logit 估计倾向得分,默认为 Probit。选择项"support"表示仅使用共同取值范围(common support)内的观测值进行匹配。选择项"test"表示检验在倾向得分匹配后,各变量在实验组与控制组的分布是否平衡(balancing test)。

下面以数据集 cardkrueger1994.dta 为例(参见第 18 章)。

```
. use cardkrueger1994,clear
. diff fte,t(treated) p(t)kernel id(id) logit cov(bk kfc roys) report
support
```

```
KERNEL PROPENSITY SCORE DIFFERENCE-IN-DIFFERENCES
ESTIMATION ON THE COMMON SUPPORT
Report - Propensity score estimation with the logit command:
Iteration 0:   log likelihood = -198.21978
Iteration 1:   log likelihood = -196.77862
Iteration 2:   log likelihood =  -196.7636
Iteration 3:   log likelihood =  -196.7636
Logistic regression                             Number of obs   =        404
                                                LR chi2(3)      =       2.91
                                                Prob > chi2     =     0.4053
Log likelihood =  -196.7636                     Pseudo R2       =     0.0073
------------------------------------------------------------------------------
     treated |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bk |   .3108387   .3561643     0.87   0.383    -.3872306    1.008908
         kfc |   .6814511   .4335455     1.57   0.116    -.1682824    1.531185
        roys |    .520356   .4011747     1.30   0.195     -.265932    1.306644
       _cons |   1.05315    .2998708     3.51   0.000      .465414    1.640886
------------------------------------------------------------------------------
ATENTION: _pscore is estimated at the baseline
Number of observations in the DIFF-IN-DIFF: 795
            Baseline        Follow-up
   Control: 78              76              154
   Treated: 326             315             641
            404             391
R-square:  0.02844
                    DIFFERENCE IN DIFFERENCES ESTIMATION
                    ------------ BASE LINE ------------  ---------- FOLLOW UP ----------
Outcome Variable    Control     Treated     Diff(BL)     Control     Treated     Diff(FU)    DIFF-IN-DIFF
fte                 21.656      17.065      -4.591       18.914      17.499      -1.414      3.177
Std. Error          0.573       1.096       1.237        0.578       1.115       1.256       1.763
t                   37.77       17.47       -3.71        16.91       17.17       -2.06       1.80
P>|t|               0.000       0.000       0.000***     0.000       0.000       0.260       0.072*
* Means and Standard Errors are estimated by linear regression
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

上表上部的 Logit 回归结果表明，各品牌虚拟变量对于处理变量的解释力很弱（准 R^2 仅为 0.007），均不显著。这暗示此数据集并不适合进行倾向得分匹配。故本例纯粹出于演示的目的。上表下部最后一列显示，平均处理效应的系数估计值为 3.177，但仅在 10% 水平上显著。下面检验进行倾向得分匹配后，是否使得各变量在处理组与控制组的分布变得平衡。

```
. diff fte,t(treated) p(t) kernel id(id) logit cov(bk kfc roys) support
test
```

```
TWO-SAMPLE T TEST
TEST ON THE COMMON SUPPORT

Number of observations (baseline): 404
            Baseline        Follow-up
   Control: 78              -           78
   Treated: 326             -           326
            404             -

t-test at period = 0:
```

Weighted Variable(s)	Mean Control	Mean Treated	Diff.	|t|	Pr(|T|>|t|)
fte	21.656	17.065	-4.591	3.22	0.0014***
bk	0.618	0.408	-0.210	3.55	0.0004***
kfc	0.104	0.209	0.104	2.60	0.0097***
roys	0.183	0.252	0.068	1.42	0.1570

```
*** p<0.01; ** p<0.05; * p<0.1
Attention: option kernel weighs variables in cov(varlist)
Means and t-test are estimated by linear regression
```

从上表结果可知，进行匹配后，有两个协变量（bk 与 kfc）的均值在处理组与控制组之间依然存在显著差异。这再次验证此数据集并不适用双重差分 PSM 法。一个显然的原因是，数据集中的协变量太少了。对于这个数据集而言，使用一般的双重差分法即可。

28.9 断点回归的思想

依可测变量选择的一种特殊情形是，有时处理变量 D_i 完全由某连续变量 x_i 是否超过某断点（cutoff point）所决定。据以进行分组的变量 x_i 称为“分组变量”（assignment variable，forcing variable 或 running variable）。比如，考察上大学对工资收入的影响，并假设上大学与否（D_i）完全取决于由高考成绩 x_i 是否超过 500 分：

$$D_i = \begin{cases} 1 & \text{若 } x_i \geqslant 500 \\ 0 & \text{若 } x_i < 500 \end{cases} \tag{28.21}$$

记不上大学与上大学的两种潜在结果分别为 (y_{0i}, y_{1i})。由于 D_i 是 x_i 的确定性函数，故在给定 x_i 的情况下，可将 D_i 视为常数，不可能与任何变量有关系，因此 D_i 独立于 (y_{0i}, y_{1i})，满足可忽略性假定。但此时，并不能使用倾向得分匹配法，因为重叠假定完全不满足，对于所有处理组成员，都有 $x_i \geqslant 500$；而所有控制组成员都有 $x_i < 500$，二者完全没有交集！因此，匹配估计量此路不通，需另辟蹊径。

显然，处理变量 D_i 为 x_i 的函数，记为 $D(x_i)$。由于函数 $D(x_i)$ 在 $x = 500$ 处存在一个断点（discontinuity），这提供了估计 D_i 对 y_i 因果效应的机会。对于高考成绩为 498，499，500，或 501

的考生，可以认为他们在各方面（包括可观测变量与不可观测变量）都没有系统差异。他们高考成绩的细微差异只是由于“上帝之手”随机抽样的结果（考试成绩本身含随机因素）[①]，导致成绩为500或501的考生上大学（进入处理组），而成绩为498或499的考生落榜（进入控制组）。因此，由于制度原因，仿佛对高考成绩在小邻域$[500-\varepsilon, 500+\varepsilon]$之间的考生进行了随机分组，故可视为准实验（quasi experiment）。由于存在随机分组，故可一致地估计在$x=500$附近的局部平均处理效应（Local Average Treatment Effect，简记LATE）[②]，即

$$\begin{aligned}\text{LATE} &\equiv \mathrm{E}(y_{1i}-y_{0i}\mid x=500)\\ &=\mathrm{E}(y_{1i}\mid x=500)-\mathrm{E}(y_{0i}\mid x=500)\\ &=\lim_{x\downarrow 500}\mathrm{E}(y_{1i}\mid x)-\lim_{x\uparrow 500}\mathrm{E}(y_{0i}\mid x)\end{aligned}\tag{28.22}$$

其中，$\lim\limits_{x\downarrow 500}$与$\lim\limits_{x\uparrow 500}$分别表示从500的右侧与左侧取极限（即右极限与左极限）。在上式最后一步推导中，假设条件期望函数$\mathrm{E}(y_{1i}\mid x)$与$\mathrm{E}(y_{0i}\mid x)$为连续函数[③]，故其极限值等于函数取值。

更一般地，断点可以是某常数c，而分组规则为

$$D_i=\begin{cases}1 & \text{若 } x_i\geqslant c\\ 0 & \text{若 } x_i<c\end{cases}\tag{28.23}$$

假设在实验前（pretreatment），结果变量y_i与x_i之间存在如下线性关系：

$$y_i=\alpha+\beta x_i+\varepsilon_i\quad(i=1,\cdots,n)\tag{28.24}$$

不失一般性，假设$D_i=\mathbf{1}(x_i\geqslant c)$的处理效应为正，则$y_i$与$x_i$之间的线性关系在$x=c$处就存在一个向上跳跃（jump）的断点，参见图28.4。由于在$x=c$附近，个体在各方面均无系统差别，故造成条件期望函数$\mathrm{E}(y_i\mid x)$在此跳跃的唯一原因只可能是D_i的处理效应。基于此逻辑，可将此跳跃视为在$x=c$处D_i对y_i的因果效应。

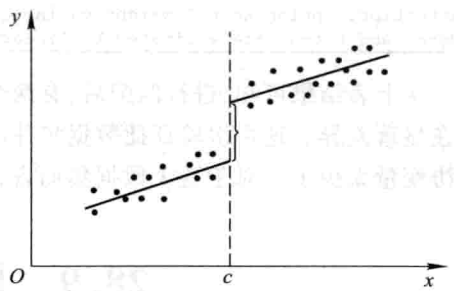


图28.4 断点回归示意图

我们知道，在方程中引入虚拟变量的效果就是在不同的子样本中产生不同的截距项（参见第9章）。因此，为了估计此跳跃，可将方程(28.24)改写为：

$$y_i=\alpha+\beta(x_i-c)+\delta D_i+\gamma(x_i-c)D_i+\varepsilon_i\quad(i=1,\cdots,n)\tag{28.25}$$

在上式中，变量(x_i-c)为x_i的标准化，使得(x_i-c)的断点为0。引入互动项$\gamma(x_i-c)D_i$是为了允许在断点两侧的回归线斜率可以不同[④]。对方程(28.25)进行OLS回归，所得$\hat{\delta}$就是在$x=c$处局部平均处理效应（LATE）的估计量。由于此回归存在一个断点，故称为“断点回归”（Regression Discontinuity，简记RD）或“断点回归设计”（Regression Discontinuity Design，简记

---

① 如果考生能够完全控制分组变量x的取值（比如通过自身努力），则断点回归将失效。但在一般情况下，考生无法精确地控制成绩（imprecise control over the assignment variable）。因此，在断点附近的考生，成绩大于或小于断点的概率大约都是二分之一，形成局部的随机分组（local randomization）。

② 此LATE不同于Imbens and Angrist(1994)基于工具变量所定义的LATE。

③ 严格来说，只需要在$x=c$这一点连续即可。

④ 方程(28.25)等价于使用断点两侧的数据分别进行回归，然后计算两侧截距项之差。如果断点两侧的回归线斜率不同，但未包括互动项，即相当于强迫两侧的斜率相同。这会导致断点右（左）侧的观测值影响对左（右）侧截距项的估计（我们不希望有此影响），从而引起偏差。

RDD)。需要注意的是,在有互动项的情况下,如果在方程(28.25)使用 x_i 而非标准化变量($x_i - c$),则 $\hat{\delta}$ 虽然度量断点两侧回归线的截距之差,但并不等于这两条回归线在 $x = c$ 处的跳跃距离①。

由于在断点附近仿佛存在随机分组,故一般认为断点回归是内部有效性(internal validity)比较强的一种准实验。在某种意义上,断点回归可视为"局部随机实验"(local randomized experiment);而且,可通过考察协变量在断点两侧的分布是否有显著差异来检验此随机性。另一方面,断点回归仅推断在断点处的因果关系,并不一定能推广到其他样本值,故外部有效性(external validity)受局限。

断点回归由 Thistlewaite and Campbell(1960)首次使用,但直到 1990 年代末才引起经济学家的重视。Hahn et al(2001)提供了断点回归的计量经济学理论基础。目前,断点回归在教育经济学、劳动经济学、健康经济学、政治经济学(Political Economy)以及区域经济学等领域的应用仍方兴未艾。参见 Imbens and Lemieux(2008),Van Der Klaauw(2008)以及 Lee and Lemieux(2010)的文献综述。

例 Thistlewaite and Campbell(1960)使用断点回归研究奖学金对于未来学业成就的影响。由于奖学金由学习成绩决定,故成绩刚好达到获奖标准与差一点达到的学生具有可比性。

例 Angrist and Lavy(1999)在研究班级规模对成绩的影响时,利用以色列教育系统的一项制度进行断点回归;该制度限定班级规模的上限为 40 名学生,一旦超过 40 名学生(比如 41 名学生),则该班级将被一分为二。参见习题。

28.10 精确断点回归

断点回归可分为两种类型。一种类型是上节介绍的"精确断点回归"(Sharp Regression Discontinuity,简记 SRD),其特征是在断点 $x = c$ 处,个体得到处理的概率从 0 跳跃为 1。另一种类型为"模糊断点回归"(Fuzzy Regression Discontinuity,简记 FRD),其特征是在断点 $x = c$ 处,个体得到处理的概率从 a 跳跃为 b,其中 $0 < a < b < 1$。本节关注精确断点回归,下节介绍模糊断点回归。

使用方程(28.25)来估计精确断点回归,存在两个问题。首先,如果回归函数包含高次项,比如二次项 $(x - c)^2$,则会导致遗漏变量偏差。其次,既然断点回归是局部的随机实验,则原则上只应使用断点附近的观测值,但方程(28.25)却使用了整个样本。为了解决这两个问题,可在方程(28.25)中引入高次项(比如二次项),并限定 x 的取值范围为 $(c - h, c + h)$:

$$\begin{aligned} y_i = {} & \alpha + \beta_1 (x_i - c) + \delta D_i + \gamma_1 (x_i - c) D_i \\ & + \beta_2 (x_i - c)^2 + \gamma_2 (x_i - c)^2 D_i + \varepsilon_i \quad (c - h < x < c + h) \end{aligned} \tag{28.26}$$

其中,$\hat{\delta}$ 为对 LATE 的估计量,并可使用稳健标准误来控制可能存在的异方差。但上式并未确定 h 的取值,而且仍然依赖于具体的函数形式。为此,研究者开始转向非参数回归。与前面的参数回归相比,非参数回归的优点在于不依赖于具体的函数形式,而且可以通过最小化均方误差

① 存在互动项意味着,断点两侧的回归线斜率不同,并非平行线,故在断点处的跳跃距离并不等于二者的截距项之差。

(MSE)来选择最优带宽 h。直观来看,h 越小,则偏差(bias)越小,但离 $x=c$ 很近的点可能很少,导致方差变大;反之,h 越大,则方差越小,但由于包含了离 $x=c$ 较远的点导致偏差变大。

最简单的非参数方法就是比较 y 在两个区间$(c-h,c)$与$[c,c+h)$的均值。但这种方法缺乏效率,且要求在这两个区间有较多观测值。另一种非参数方法为核回归(kernel regression),即以核函数计算权重,对带宽 h 范围内的观测值进行加权平均(参见第27章)。但核回归的边界性质并不理想,而我们关心的恰恰是回归函数在端点的取值。为此,一般推荐使用局部线性回归(local linear regression),即最小化如下目标函数:

$$\min_{|\alpha,\beta,\delta,\gamma|}\sum_{i=1}^{n}K[(x_i-c)/h][y_i-\alpha-\beta(x_i-c)-\delta D_i-\gamma(x_i-c)D_i]^2 \tag{28.27}$$

其中,$K(\cdot)$为核函数。局部线性回归的实质是,在一个小邻域$(c-h,c+h)$内进行加权最小二乘法估计,此权重由核函数来计算,离 c 越近的点权重越大。针对断点回归,较常用的核函数为三角核(triangular kernel)与矩形核(rectangular kernel,即均匀核)。如果使用矩形核,则为标准OLS回归,等价于上文的参数回归。此估计量也称为"局部沃尔德估计量"(local Wald estimator)。

下面考察最优带宽的选择。记 $m_1(x)\equiv \mathrm{E}(y_1|x)$,$m_0(x)\equiv \mathrm{E}(y_0|x)$,则 $\delta=m_1(c)-m_0(c)$,$\hat{\delta}=\hat{m}_1(c)-\hat{m}_0(c)$。Imbens and Kalyanaraman(2009)提出通过最小化两个回归函数在断点处的均方误差来选择最优带宽:

$$\min_{h}\mathrm{E}\{[\hat{m}_1(c)-m_1(c)]^2+[\hat{m}_0(c)-m_0(c)]^2\} \tag{28.28}$$

另外,也可在方程(28.26)或(28.27)加入影响结果变量 y_i 的其他协变量$\boldsymbol{w}_i$,可通过 Stata 命令 `rd` 的选择项"`cov(varlist)`"来实现。由于断点回归可视为局部随机实验,故是否包括协变量$\boldsymbol{w}_i$并不影响断点回归估计量的一致性。加入协变量$\boldsymbol{w}_i$的好处在于,如果这些协变量对于被解释变量 y_i 有解释力,则可以减少扰动项方差,使得估计更为准确。然而,如果所加入的协变量为内生变量,与扰动项相关,则反而会干扰对 LATE 的估计。

另外,如果协变量$\boldsymbol{w}_i$在 $x=c$ 处的条件密度函数也存在跳跃,则不宜将 $\hat{\delta}$ 全部归功于该项目的处理效应。事实上,断点回归的一个隐含假设是,协变量$\boldsymbol{w}_i$的条件密度在 $x=c$ 处连续。为了检验此假设,可将$\boldsymbol{w}_i$中每个变量作为被解释变量,进行断点回归,考察其分布是否在 $x=c$ 处有跳跃;这可通过 Stata 命令 `rd` 的选择项"`x(varlist)`"来实现。

在进行断点回归时,还应注意可能存在"内生分组"(endogenous sorting):如果个体事先知道分组规则,并可通过自身努力而完全控制分组变量(complete manipulation),则可自行选择进入处理组或控制组,导致在断点附近的内生分组而非随机分组,引起断点回归失效。另一方面,如果个体事先不清楚分组规则,或只能部分地控制分组变量(partial manipulation),则一般不存在此担忧。对于内在分组的可能性,可从理论上讨论,也可根据数据进行检验。假设存在内生分组,则个体将自行选择进入断点两侧,导致在断点两侧的分布不均匀,即分组变量 x 的密度函数 $f(x)$在断点 $x=c$ 处不连续,出现左极限不等于右极限的情形。为此,McCrary(2008)提出检验以下原假设:

$$H_0:\theta\equiv \ln\lim_{x\downarrow c}f(x)-\ln\lim_{x\uparrow c}f(x)\equiv \ln f^{+}-\ln f^{-}=0 \tag{28.29}$$

通过计算 $\hat{\theta}$ 及其标准误,即可检验密度函数 $f(x)$是否 $x=c$ 处连续。另外,根据同样的逻辑,内生分组也可能使得协变量$\boldsymbol{w}_i$在 $x=c$ 两侧分布不均匀;故检验协变量$\boldsymbol{w}_i$的条件密度在 $x=c$ 处是否连续也有帮助。

由于断点回归在操作上存在不同选择,故在实践中,一般建议同时汇报以下各种情形,以保证结果的稳健性。

(1) 分别汇报三角核与矩形核的局部线性回归结果(后者等价于线性参数回归);
(2) 分别汇报使用不同带宽的结果(比如,最优带宽及其二分之一或两倍带宽);
(3) 分别汇报包含协变量与不包含协变量的情形。
(4) 进行模型设定检验,包括检验分组变量与协变量的条件密度是否在断点处连续。

28.11 模糊断点回归

模糊断点回归的特征是,在断点 $x=c$ 处,个体得到处理的概率从 a 跳跃为 b,其中 $0<a<b<1$,参见图 28.5。这意味着,即使 $x>c$,也不一定得到处理,只不过得到处理的概率在 $x=c$ 处有一个不连续的跳跃。显然,所谓"模糊断点回归",其断点并不模糊(断点很明确地在 $x=c$ 处),只不过分组变量 x 跨过断点 c 的后果并非泾渭分明,只是得到处理的概率存在跳跃。在某种意义上,精确断点回归可视为模糊断点回归的特例或极限情形。

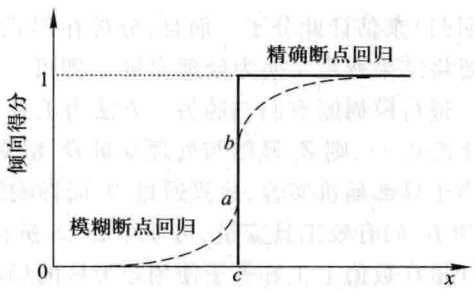


图 28.5 精确断点回归与模糊断点回归

回到高考录取上大学的例子。事实上,高考成绩上线并不能完全保证上大学,因为能否上大学还取决于填报的志愿,甚至有些上线考生放弃上大学的机会;另一方面,即使成绩未上线,但也可能因某种特长而得到加分,从而得到上大学的机会。这表明,分数线并不完全决定上大学。然而,上大学的概率确实在分数线的位置上有一个不连续的跳跃。

在模糊断点的情况下,处理变量 D 并不完全由分组变量 x 所决定。一般来说,影响处理变量 x 的其他因素也会影响结果变量 y,导致在方程(28.25)或(28.27)中处理变量 D 与扰动项 ε 相关,故 OLS 估计量不一致。比如,虽然成绩上线却因志愿不妥而落榜者多有较深实力,而这种不可观测的实力可以影响结果变量 y。

为了在模糊断点的情况下识别平均处理效应,需要引入以下条件独立假定。

假定 28.4 给定 x,则 (y_1-y_0) 独立于 D,即 $(y_{1i}-y_{0i}) \perp D_i \mid x_i$。

此假定意味着,在给定分组变量 x 的情况下,D 可以与 y_0 相关,但不能与参加项目的收益 $(y_{1i}-y_{0i})$ 相关。由于 $y=y_0+D(y_1-y_0)$,故

$$\begin{aligned}\mathrm{E}(y\mid x) &= \mathrm{E}(y_0\mid x)+\mathrm{E}[D(y_1-y_0)\mid x] \\ &= \mathrm{E}(y_0\mid x)+\mathrm{E}(D\mid x)\cdot\mathrm{E}[(y_1-y_0)\mid x]\end{aligned} \tag{28.30}$$

其中,$\mathrm{E}[(y_1-y_0)\mid x]$ 是我们想要估计的平均处理效应,而 $\mathrm{E}(D\mid x)$ 为倾向得分。在上式的第二步使用了条件独立假定。对上式两边从 c 的右边取极限可得

$$\lim_{x\downarrow c}\mathrm{E}(y\mid x)=\lim_{x\downarrow c}\mathrm{E}(y_0\mid x)+\lim_{x\downarrow c}\mathrm{E}(D\mid x)\cdot\lim_{x\downarrow c}\mathrm{E}[(y_1-y_0)\mid x] \tag{28.31}$$

同理,对上式两边从 c 的左边取极限可得

$$\lim_{x\uparrow c}\mathrm{E}(y\mid x)=\lim_{x\uparrow c}\mathrm{E}(y_0\mid x)+\lim_{x\uparrow c}\mathrm{E}(D\mid x)\cdot\lim_{x\uparrow c}\mathrm{E}[(y_1-y_0)\mid x] \tag{28.32}$$

假设函数 $\mathrm{E}(D\mid x)$,$\mathrm{E}(y_0\mid x)$ 与 $\mathrm{E}(y_1\mid x)$ 在 $x=c$ 处连续,则其左极限等于右极限,也等于其函数值,故 $\lim_{x\downarrow c}\mathrm{E}(y_0\mid x)=\lim_{x\uparrow c}\mathrm{E}(y_0\mid x)$,而且 $\lim_{x\downarrow c}\mathrm{E}[(y_1-y_0)\mid x]=\lim_{x\uparrow c}\mathrm{E}[(y_1-y_0)\mid x]=$

$\mathrm{E}[(y_1-y_0)\mid x=c]$。因此，将方程(28.31)减去(28.32)可得

$$\lim_{x\downarrow c}\mathrm{E}(y\mid x)-\lim_{x\uparrow c}\mathrm{E}(y\mid x)=\left[\lim_{x\downarrow c}\mathrm{E}(D\mid x)-\lim_{x\uparrow c}\mathrm{E}(D\mid x)\right]\cdot\mathrm{E}[(y_1-y_0)\mid x=c] \tag{28.33}$$

根据模糊断点回归的定义可知，$\lim_{x\downarrow c}\mathrm{E}(D\mid x)-\lim_{x\uparrow c}\mathrm{E}(D\mid x)=b-a\neq 0$，故可将其作为分母：

$$\mathrm{LATE}\equiv\mathrm{E}[(y_1-y_0)\mid x=c]=\frac{\lim_{x\downarrow c}\mathrm{E}(y\mid x)-\lim_{x\uparrow c}\mathrm{E}(y\mid x)}{\lim_{x\downarrow c}\mathrm{E}(D\mid x)-\lim_{x\uparrow c}\mathrm{E}(D\mid x)} \tag{28.34}$$

显然，上式的分子就是精确断点回归的LATE，而分母为得到处理的概率(即倾向得分)在断点 c 处的跳跃($b-a$)。表达式(28.34)是精确断点回归表达式(28.22)的推广，因为在精确断点的情况下，$b-a=1$，将其代入(28.34)式即可得到(28.22)式。

由于表达式(28.34)的分子就是精确断点回归的LATE，故可用精确断点回归(比如，局部线性回归)来估计此分子。而且，分母在形式上与分子完全一样，故也可用精确断点回归来估计，只要将结果变量 y 换为处理变量 D 即可。

进行模糊断点回归的另一方法为工具变量法。定义 $Z_i=\mathbf{1}(x_i\geqslant c)$(即分组变量是否大于或等于断点 c)，则 Z_i 显然与处理变量 D_i 相关，满足相关性。另一方面，$Z_i=\mathbf{1}(x_i\geqslant c)$ 在断点 c 附近相当于局部随机实验，故只通过 D_i 而影响结果变量 y_i，与扰动项 ε_i 不相关，满足外生性。因此，Z_i 为 D_i 的有效工具变量，可使用2SLS进行估计。可以证明，如果使用相同的带宽 h，则此2SLS估计量在数值上正好等于使用矩形核的局部线性回归估计量。

以上介绍的断点回归均假设在断点附近仿佛存在局部随机分组。但如果分组变量为年龄(时间)或地理区域，则这种解释一般行不通，被Lee and Lemieux(2010)称为“非随机断点设计”(Nonrandomized discontinuity design)。比如，以年龄65岁为分界线，年满65岁即可获得退休金。此时，分组变量为时间，是个确定性过程，个体无法控制。此时，须考虑以下三种可能性。首先，年满65岁是否使得个体有资格参加其他项目，从而通过其他渠道影响结果变量。其次，虽然年满65岁即可获得退休金，但退休金的效应可能需要几年后才能体现(即可能存在动态效应)。最后，由于个体可以预见65岁以后将得到退休金，故可能在65岁之前就调整其经济行为。对于这些可能性，都应进行具体分析，才能得到令人信服的结论。

另一种非随机断点设计使用地理区域作为分组变量，以某种区域分界线作为断点，进行“地理断点回归”(geographic RD)。比如，Black(1999)通过比较在学区分界线两侧的房价来测算居民对高质量小学教育的支付意愿(willingness to pay)。由于个体一般可以选择住在学区分界线的哪一侧，故很难视为局部随机分组。此时，需要说明的是，在分界线两侧，除了处理变量不同外(一侧的学生去一所学校，而另一侧的学生去另一所学校)，在其他方面均几乎没有差别。为了保证分界线两侧的可比性，Black(1999)剔除了分界线为主要街道或高速公路的部分分界线(主要街道或高速公路两侧的社区可能有较大差别，尽管距离很近)。

例(强制矿工制度的长期经济影响) 在1573—1812年期间，西班牙殖民者在秘鲁与玻利维亚实行了一种称为“mining mita”的强制矿工征用制度。该制度规定，在离矿山较近的一定区域内，每个土著社区须提供其成年男性人口的七分之一作为强制矿工。为了研究此制度的长期经济影响，Dell(2010)使用断点回归来比较此区域分界线两侧的当代家庭消费与儿童发育不良比例。为了保证分界线两侧具有可比性，该研究剔除了分界线一侧为平原而另一侧为安第斯山脉的部分分界线。

28.12 断点回归的 Stata 实例

断点回归可通过非官方 Stata 命令 rd 来实现：

```
ssc install rd,replace  (下载安装命令 rd)
```

该命令使用局部线性回归来估计断点回归模型，其基本句式为

```
rd y D x, z0(real) strineq mbw(numlist) graph bdep oxline kernel(rectangle) cov(varlist) x(varlist)
```

其中，“y”为结果变量，“D”为处理变量，而“x”为分组变量。选择项“z0(real)”用来指定断点位置，默认值为“z0(0)”，即断点为原点。如果省略处理变量 D，则默认为精确断点回归，并根据分组变量 x 来计算处理变量，即如果 x 大于或等于断点 z0，则 D 取值为 1；反之，D 取值为 0。选择项“strineq”表示根据严格不等式（strong inequality）来计算处理变量，即如果 x 大于断点 z0，则 D 取值为 1；反之，D 取值为 0。

选择项“mbw(numlist)”用来指定最优带宽的倍数，默认值为“mbw(50 100 200)”，即根据最优带宽的 0.5，1 与 2 倍进行局部线性回归，其中 100 对应于根据 Imbens and Kalyanaraman (2009) 计算的最优带宽。选择项“graph”表示根据所选的每一带宽，画出其局部线性回归图。选择项“bdep”表示通过画图来考察断点回归估计量对带宽的依赖性（bandwidth dependence），而选择项“oxline”表示在此图的默认带宽（即最优带宽）上画一条直线，以便识别。选择项“kernel(rectangle)”表示使用矩形核（即均匀核），默认使用三角核。选择项“cov(varlist)”用来指定加入局部线性回归的协变量。选择项“x(varlist)”表示检验这些协变量是否在断点处有跳跃（估计跳跃值及其显著性）。

下面以命令 rd 自带的数据集 votex.dta 为例演示断点回归的操作。该数据集用于考察美国国会选区如果有一名民主党众议员对该选区联邦支出（federal expenditure within a Congressional district）的影响。传统上，民主党倾向于大政府，故一个选区如果有民主党众议员，则该议员可能为该选区争取更多的联邦支出。然而，直接对二者进行回归可能存在遗漏变量问题或双向因果关系。为此，使用该民主党候选人的得票比例作为分组变量，以 0.5 作为断点（在两党政治中，得票比例大于或等于 0.5 则当选，反之落选），进行断点回归。数据集 votex.dta 的主要变量包括：结果变量 lne（选区联邦开支的对数），分组变量 d（民主党候选人得票比例减去 0.5），处理变量 win（民主党候选人当选），以及一系列协变量。

```
. use votex.dta,clear
```

先来看一下数据集中的变量。

```
. d
```

```
  obs:           349                          102nd Congress
 vars:            19                          4 Feb 2013 15:33
 size:        37,692
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
fips            byte   %8.0g       fips       State code
district        byte   %8.0g                  Congr district
d               double %10.0g                 Dem vote share minus .5
win             byte   %9.0g                  Dem Won Race
lne             float  %9.0g                  Log fed expenditure in district
i               byte   %9.0g                  Incumbent
votingpop       long   %12.0g                 Voting Age Population
votpop          double %10.0g                 Voting Age Population Share
populatn        long   %12.0g                 Population
black           double %12.0g                 Black Population Share
blucllr         double %12.0g                 Blue-collar Population Share
farmer          double %12.0g                 Farmer Population Share
fedwrkr         double %12.0g                 Fed Worker Population Share
forborn         double %12.0g                 Foreign Born Population Share
manuf           double %12.0g                 Manufactur Population Share
unemplyd        double %12.0g                 Unemp Population Share
union           float  %9.0g                  Unionized Population Share
urban           double %12.0g                 Urban Population Share
veterans        double %12.0g                 Veteran Population Share
-------------------------------------------------------------------------------
Sorted by:  fips  district
```

首先,使用最优带宽以及默认的三角核进行精确断点回归,并画图(参见图 28.6)。

```
. rd lne d,gr mbw(100)
```

```
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is d
 Treatment variable X_T unspecified
 Outcome variable y is lne

Command used for graph: lpoly; Kernel used: triangle (default)
Bandwidth: .29287776; loc Wald Estimate: -.07739553
Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889

从上表可知,局部沃尔德估计值(local Wald estimate)为负,且很不显著;说明拥有民主党众议员的选区并不能吸引更多的联邦开支。图 28.6 也显示,条件期望函数 $E(\text{lne}|d)$ 只在断点 $d=0$ 处稍微向下跳跃。

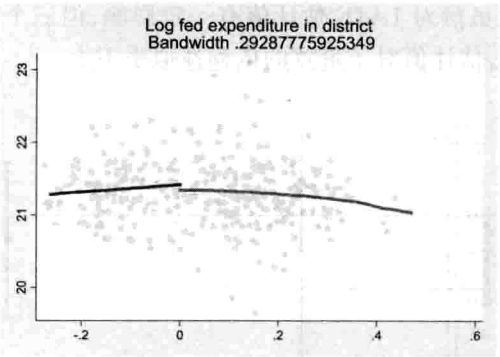


图 28.6 使用最优带宽与三角核的断点回归图

其次,加入协变量宽重复上述估计,但省略画图。

```
. rd lne d,mbw(100) cov(i votpop black blucllr farmer fedwrkr forborn
manuf unemplyd union urban veterans)
```

```
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is d
 Treatment variable X_T unspecified
 Outcome variable y is lne

Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	.0543733	.0921634	0.59	0.555	-.1262636	.2350102

上表显示,LATE 估计值变为正,但依然很不显著。

再次,去掉协变量,但同时估计三种带宽,并画出估计值对带宽的依赖性(参见图 28.7)。

```
. rd lne d,gr bdep oxline
```

```
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is d
 Treatment variable X_T unspecified
 Outcome variable y is lne

Command used for graph: lpoly; Kernel used: triangle (default)
Bandwidth: .29287776; loc Wald Estimate: -.07739553
Bandwidth: .14643888; loc Wald Estimate: -.09491495
Bandwidth: .58575552; loc Wald Estimate: -.0543086
Estimating for bandwidth .29287775925349
Estimating for bandwidth .146438879626745
Estimating for bandwidth .58575551850698
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889
lwald50	-.0949149	.1454442	-0.65	0.514	-.3799804	.1901505
lwald200	-.0543086	.0911788	-0.60	0.551	-.2330157	.1243985

从上表可知，改变带宽虽然对 LATE 估计值有一定影响，但三个估计值均为负，且依然不显著。从图 28.7 也可以看出，估计值对于带宽的依赖性似乎不大。

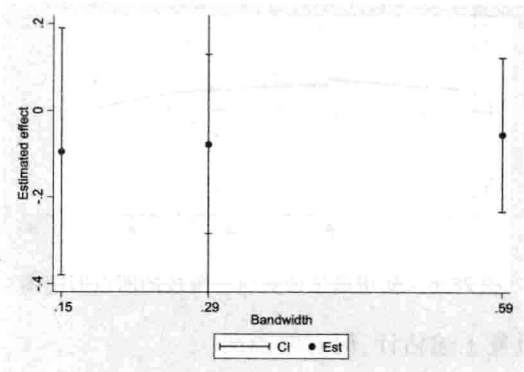


图 28.7 断点回归估计值对带宽的依赖性

进行断点回归后，还需要对其设定进行检验。下面，检验协变量在断点处的条件密度是否存在跳跃。

```
. rd lne d,mbw(100) x(i votpop black blucllr farmer fedwrkr forborn
manuf unemplyd union urban veterans)
```

```
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is d
 Treatment variable X_T unspecified
 Outcome variable y is lne

Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
i	-.0044941	.1208008	-0.04	0.970	-.2412592	.2322711
votpop	-.0082128	.0062347	-1.32	0.188	-.0204326	.0040071
black	-.0036113	.020048	-0.18	0.857	-.0429046	.0356821
blucllr	.0026193	.0057316	0.46	0.648	-.0086144	.013853
farmer	-.0078737	.0037566	-2.10	0.036	-.0152366	-.0005109
fedwrkr	.0001617	.0037584	0.04	0.966	-.0072046	.0075281
forborn	-.015235	.0120682	-1.26	0.207	-.0388882	.0084183
manuf	.0147223	.0100352	1.47	0.142	-.0049463	.0343908
unemplyd	-.0007393	.0019069	-0.39	0.698	-.0044769	.0029982
union	-2.25e-06	3.66e-06	-0.61	0.540	-9.43e-06	4.94e-06
urban	.0370978	.0559882	0.66	0.508	-.072637	.1468326
veterans	.0015796	.0036205	0.44	0.663	-.0055164	.0086756
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889

上表显示，除了变量 farmer（农民占人口比例）外，所有协变量的条件密度函数在断点处都是连续的。下面，使用 McCrary（2008）的方法检验分组变量的密度函数是否在断点处不连续。McCrary 检验分两步进行。第一步，将分组变量在断点 c 两侧尽量等距离细分，画很不光滑的直方图（a very undersmoothed histogram），记组距（bin size）为 b，记每组的中心位置为变量 X_j =

$\left\{\cdots, c-\frac{3b}{2}, c-\frac{b}{2}, c+\frac{b}{2}, c+\frac{3b}{2}, \cdots\right\}$。然后计算每组的标准化频率(normalized cellsize),即频数除以 nb(n 为样本容量),记为 Y_j。第二步,使用三角核,将 Y_j 对 X_j 进行局部线性回归(参见第 27 章);针对分组变量的取值 $r_0 = \{\cdots, c-2b, c-b, c+b, c+2b, \cdots\}$,可得密度函数估计值 $\hat{f}(r_0)$(记为 fhat)及标准误 $\mathrm{SE}[\hat{f}(r_0)]$(记为 se_fhat)。通过方程(28.29)可计算 $\hat{\theta} = \widehat{\ln f^+} - \widehat{\ln f^-}$(称为"log difference in height"),然后检验密度函数是否在 c 处连续。

McCrary 检验可通过非官方 Stata 命令 DCdensity 来实现(其中,DC 表示 Discontinuity),下载地址为 http://emlab.berkeley.edu/~jmccrary/DCdensity/。将该网页的"DCdensity.ado"文件下载到文件夹"\\ado\\plus"即可(在 Stata 中输入命令 sysdir 即可显示此文件夹的位置)。

命令 DCdensity 的基本句式为

```
. DCdensity assign_var,breakpoint(#) generate(Xj Yj r0 fhat se_fhat) graphname(filename)
```

其中,"assign_var"为分组变量,必选项"breakpoint(#)"用来指定断点位置,必选项"generate(Xj Yj r0 fhat se_fhat)"用来指定输出变量名,而选择项"graphname(filename)"用来指定密度函数图的文件名。

下面,将该命令应用于此数据集的分组变量 d,画图结果参见图 28.8。

```
. DCdensity d,breakpoint(0) generate(Xj Yj r0 fhat se_fhat) graphname(rd.eps)
```

```
Using default bin size calculation, bin size = .017174107
Using default bandwidth calculation, bandwidth = .10514868

Discontinuity estimate (log difference in height): -.429396753
                                                   (.444361558)
Performing LLR smoothing.
45 iterations will be performed
....
Exporting graph as rd.eps
(note: file rd.eps not found)
(file rd.eps written in EPS format)
```

从上表可知,$\hat{\theta} = -0.43$,而标准误为 0.44,故可接受密度函数在 c 处连续的原假设。从图 28.8 也可以看出,断点两侧密度函数估计值的置信区间有很大部分重叠,故断点两侧的密度函

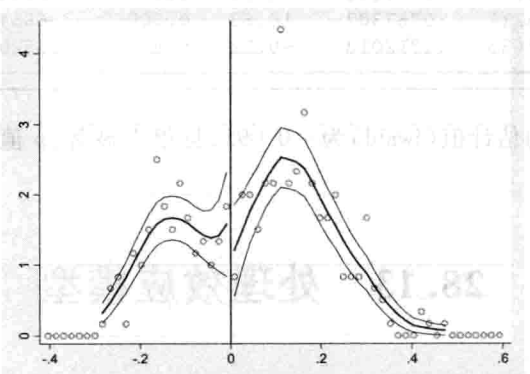


图 28.8 分组变量密度函数在断点处的连续性

数不存在显著差异。

以上为精确断点回归。纯粹为了演示模糊断点回归,下面随机生成一个新的处理变量ranwin。变量ranwin不由分组变量d完全决定,但与原来的处理变量win高度相关(win完全由d决定)。

```
. set seed 10101
. gen ranwin = cond(uniform() < .1,1 - win,win)
```

其中,"uniform()"表示生成一个在[0,1)区间服从均匀分布的随机变量,而命令"cond(uniform() < .1,1 - win,win)"表示,如果此随机变量小于0.1,则变量ranwin = 1 - win;否则,变量ranwin = win。因此,ranwin也是虚拟变量,且与win高度相关。

下面,看一下ranwin与win的分布情况。

```
. tab ranwin win
```

ranwin	Dem Won Race 0	1	Total
0	118	20	138
1	13	198	211
Total	131	218	349

从上表可知,在大多数情况下,ranwin与win的取值相同。下面使用最优带宽与默认的三角核进行模糊断点回归。

```
. rd lne ranwin d,mbw(100)
```

```
Three variables specified; jump in treatment
at Z=0 will be estimated. Local Wald Estimate
is the ratio of jump in outcome to jump in treatment.

 Assignment variable Z is d
 Treatment variable X_T is ranwin
 Outcome variable y is lne

Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
numer	-.0773955	.1051192	-0.74	0.462	-.2834254	.1286343
denom	.8158211	.0767307	10.63	0.000	.6654317	.9662105
lwald	-.0948683	.1312013	-0.72	0.470	-.3520182	.1622816

从上表可知,LATE的估计值(lwald)为-0.095,且很不显著(p值为0.47)。读者可自行尝试其他带宽或矩形核。

28.13 处理效应模型

解决依不可测变量选择问题的另一方法是遵循Heckman(1979)样本选择模型的传统,直接对处理变量D_i进行结构建模。为此,Maddala(1983)提出了以下"处理效应模型"(treatment

effects model)：

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma D_i + \varepsilon_i \tag{28.35}$$

假设处理变量由以下“处理方程”(treatment equation)所决定：

$$D_i = \mathbf{1}(\boldsymbol{z}_i'\boldsymbol{\delta} + u_i) \tag{28.36}$$

其中，$\mathbf{1}(\cdot)$为示性函数(indicator function)。$\boldsymbol{z}_i$可以与$\boldsymbol{x}_i$有重叠的变量，但$\boldsymbol{z}_i$中至少有一个变量，比如z_{1i}，不在$\boldsymbol{x}_i$中。进一步，假设$\mathrm{Cov}(z_{1i}, \varepsilon_i) = 0$，即虽然$z_{1i}$影响个体是否参与项目$D_i$，但并不直接影响结果变量$y_i$(只通过$D_i$间接影响$y_i$)；故可将$z_{1i}$视为$D_i$的工具变量。假设扰动项$(\varepsilon_i, u_i)$服从二维正态分布：

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & 1 \end{pmatrix}\right] \tag{28.37}$$

其中，ρ为(ε_i, u_i)的相关系数，而u_i的方差被标准化为1(因为u_i是Probit模型的扰动项，参见第11章)。我们允许$\rho \neq 0$，这正是模型内生性的来源。反之，如果$\rho = 0$，则不存在内生性，可直接用OLS得到对方程(28.35)的一致估计。对于参加者而言，y_i的条件期望为

$$\begin{aligned} \mathrm{E}(y_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid \boldsymbol{z}_i'\boldsymbol{\delta} + u_i > 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid u_i > -\boldsymbol{z}_i'\boldsymbol{\delta}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \rho\sigma_\varepsilon \lambda(-\boldsymbol{z}_i'\boldsymbol{\delta}) \end{aligned} \tag{28.38}$$

其中，$\lambda(\cdot)$为反米尔斯函数，即$\lambda(c) \equiv \dfrac{\phi(c)}{1 - \Phi(c)}$。在上式推导的最后一步，用到了偶然断尾的条件期望公式(参见第14章)。类似地，未参加者的条件期望为

$$\begin{aligned} \mathrm{E}(y_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid \boldsymbol{z}_i'\boldsymbol{\delta} + u_i \leqslant 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid u_i \leqslant -\boldsymbol{z}_i'\boldsymbol{\delta}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} - \rho\sigma_\varepsilon \lambda(\boldsymbol{z}_i'\boldsymbol{\delta}) \end{aligned} \tag{28.39}$$

将方程(28.38)减去方程(28.39)，可得参加者与未参加者的条件期望之差：

$$\mathrm{E}(y_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) - \mathrm{E}(y_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) = \gamma + \rho\sigma_\varepsilon[\lambda(-\boldsymbol{z}_i'\boldsymbol{\delta}) + \lambda(\boldsymbol{z}_i'\boldsymbol{\delta})] \tag{28.40}$$

显然，如果直接比较处理组与控制组的平均收益y_i，将遗漏上式右边第二项$\rho\sigma_\varepsilon[\lambda(-\boldsymbol{z}_i'\delta) + \lambda(\boldsymbol{z}_i'\boldsymbol{\delta})]$，导致不一致的估计(除非$\rho = 0$)。为了将处理组与控制组放在一起进行回归，定义个体i的风险(hazard)为[①]

$$\lambda_i = \begin{cases} \lambda(-\boldsymbol{z}_i'\boldsymbol{\delta}) & \text{若 } D_i = 1 \\ -\lambda(\boldsymbol{z}_i'\boldsymbol{\delta}) & \text{若 } D_i = 0 \end{cases} \tag{28.41}$$

这样，可以将方程(28.38)与(28.39)合并为一个方程(对于参加者与未参加者都适用)：

$$\mathrm{E}(y_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma D_i + \rho\sigma_\varepsilon \lambda_i \tag{28.42}$$

为此，可进行类似于Heckit的两步法估计。

第一步：用Probit估计方程$\mathrm{P}(D_i = 1 \mid \boldsymbol{z}_i) = \Phi(\boldsymbol{z}_i'\boldsymbol{\delta})$，得到估计值$\hat{\boldsymbol{\delta}}$，计算$\hat{\lambda}_i$。

第二步：用OLS回归$y_i \xrightarrow{\text{OLS}} \boldsymbol{x}_i, D_i, \hat{\lambda}_i$，得到估计值$\hat{\boldsymbol{\beta}}, \hat{\gamma}, \widehat{\rho\sigma_\varepsilon}$。

---

① “风险函数”也称为反米尔斯函数。

在 Stata 中，称 λ_i 的系数估计值 $\widehat{\rho\sigma_\varepsilon}$ 为 lamda。两步法的优点是计算方便；其缺点在于，第一步的估计误差被带入第二步中，导致效率损失。更有效率的做法是，使用最大似然估计法（MLE），同时估计所有模型参数。需要注意的是，上述处理效应模型依赖于对结构方程的正确设定，如果模型设定有误，z_{1i} 不是有效工具变量（比如 z_{1i} 与 ε_i 相关），或扰动项不服从正态分布，都会导致不一致的估计。

处理效应模型的 Stata 命令为

```
treatreg y x1 x2 x3,treat(D = z1 z2 z3) twostep first
```

其中，“y”为结果变量，“x1 x2 x3”为直接影响 y 的自变量。“treat(D = z1 z2 z3)”表示处理方程，其中“D”为处理变量，“z1 z2 z3”为影响 D 的变量。选择项“twostep”表示使用两步法，默认使用 MLE；选择项“first”表示汇报第一阶段 Probit 回归的结果。

下面以数据集“labor. dta”为例来说明处理效应模型的 Stata 操作[①]。被解释变量为 ww（妻子工资），wa（妻子年龄），cit（是否住在大城市），以及 we（妻子受教育年限）；其他变量包括 wmed（妻子母亲的受教育年限），以及 wfed（妻子父亲的受教育年限）。我们希望知道妻子是否上大学对其工资的影响。

```
. use labor.dta,clear
```

假设受教育年限超过 12 年即为受过大学教育，生成以下虚拟变量：

```
. gen wc = (we >12)
```

首先，作为参照，进行 OLS 回归。

```
. reg ww wa cit wc,r
```

```
Linear regression                               Number of obs =     250
                                                F(  3,   246) =    3.29
                                                Prob > F      =  0.0212
                                                R-squared     =  0.0555
                                                Root MSE      =    2.54

------------------------------------------------------------------------------
             |               Robust
          ww |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wa |  -.0104985   .0204216    -0.51   0.608    -.050722     .029725
         cit |   .1278922   .3225258     0.40   0.692    -.507372     .7631564
          wc |   1.332192   .4321052     3.08   0.002     .4810939    2.183289
       _cons |   2.278337   .8543208     2.67   0.008     .5956206    3.961054
------------------------------------------------------------------------------
```

由上表可知，变量 wc（妻子是否上大学）的系数为 1.33，且在 1% 水平上显著。然而，妻子是否上大学显然为内生虚拟变量。一种可能的情形是，（不可观测的）妻子能力越高，则越可能上大学，也越可能获得高工资，故 OLS 可能高估了上大学的作用。

下面，假设 wmed（妻子母亲的受教育年限）与 wfed（妻子父亲的受教育年限）都是影响 wc（妻子是否上大学）的外生因素，进行两步法估计。

```
. treatreg ww wa cit,treat(wc = wmed wfed) two first
```

① 此数据集由 Stata 提供，截取自 Berndt(1996)。

```
Probit regression                                 Number of obs   =        250
                                                  LR chi2(2)      =      51.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -121.31755                       Pseudo R2       =     0.1763

------------------------------------------------------------------------------
          wc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wmed |   .1198888   .0319862     3.75   0.000     .0571971    .1825806
        wfed |   .0960764   .0290583     3.31   0.001     .0391233    .1530295
       _cons |  -2.631496   .3308389    -7.95   0.000    -3.279928   -1.983063
------------------------------------------------------------------------------

Treatment-effects model -- two-step estimates     Number of obs   =        250

                                                  Wald chi2(3)    =       3.67
                                                  Prob > chi2     =     0.2998

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ww           |
          wa |  -.0111623    .020152    -0.55   0.580    -.0506594    .0283348
         cit |   .1276102     .33619     0.38   0.704      -.53131    .7865305
          wc |   1.257995    .8007428     1.57   0.116    -.3114319    2.827422
       _cons |   2.327482    .9610271     2.42   0.015     .4439031     4.21106
-------------+----------------------------------------------------------------
wc           |
        wmed |   .1198888   .0319862     3.75   0.000     .0571971    .1825806
        wfed |   .0960764   .0290583     3.31   0.001     .0391233    .1530295
       _cons |  -2.631496   .3308389    -7.95   0.000    -3.279928   -1.983063
-------------+----------------------------------------------------------------
hazard       |
      lambda |   .0548738   .5283928     0.10   0.917    -.9807571    1.090505
-------------+----------------------------------------------------------------
         rho |    0.02178
       sigma |  2.5198211
------------------------------------------------------------------------------
```

上表上部显示，变量 wmed 与 wfed 在第一阶段 Probit 回归中均有显著的正作用。上表下部显示，尽管变量 wc 的系数估计值仍为 1.26（仅略小于 OLS 估计值），却不显著（p 值为 11.6%）。下面，进行更有效率的 MLE 估计。

```
. treatreg ww wa cit,treat(wc = wmed wfed) nolog
```

Treatment-effects model -- MLE　　Number of obs = 250
　　Wald chi2(3) = 4.11
Log likelihood = -707.07215　　Prob > chi2 = 0.2501

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ww						
wa	-.0110424	.0199652	-0.55	0.580	-.0501735	.0280887
cit	.127636	.3361938	0.38	0.704	-.5312917	.7865638
wc	1.271327	.7412951	1.72	0.086	-.1815842	2.724239
_cons	2.318638	.9397573	2.47	0.014	.4767478	4.160529
wc						
wmed	.1198055	.0320056	3.74	0.000	.0570757	.1825352
wfed	.0961886	.0290868	3.31	0.001	.0391795	.1531977
_cons	-2.631876	.3309128	-7.95	0.000	-3.280453	-1.983299
/athrho	.0178668	.1899898	0.09	0.925	-.3545063	.3902399
/lnsigma	.9241584	.0447455	20.65	0.000	.8364588	1.011858
rho	.0178649	.1899291			-.3403659	.371567
sigma	2.519747	.1127473			2.308179	2.750707
lambda	.0450149	.4786442			-.8931105	.9831404

LR test of indep. eqns. (rho = 0):　chi2(1) = 0.01　Prob > chi2 = 0.9251

上表显示,MLE 的估计结果与两步法类似。然而,上表底部的似然比检验结果却没有拒绝原假设"$H_0:\rho=0$"(p 值高达 0.93)。从表面上看,这意味着不存在内生性,可直接进行 OLS 估计。然而,也可能存在其他模型设定误差,比如忽略了妻子是否进行劳动力市场的内生选择(对于 40% 的观测值,ww = 0,这意味着妻子无工作),或者忽略了高次项或互动项。

习　题

28.1 使用数据集 jtrain3. dta 估计参加就业培训(train)对 1978 年实际收入(re78)的处理效应。该数据集中的变量解释类似于本章正文使用的数据集 ldw_exper. dta。

(1) 把 re78 对 train 进行一元回归。

(2) 加入控制变量进行多元回归。

(3) 进行一对一的倾向得分匹配估计。

(4) 进行 k 近邻倾向得分匹配,令 $k=4$。

(5) 进行卡尺内一对四倾向得分匹配。

(6) 进行倾向得分核匹配。

(7) 进行倾向得分线性回归匹配。

(8) 进行马氏匹配。

(9) 计算偏差校正的匹配估计量。

28.2 使用数据集 angrist. dta[①],利用断点回归估计班级规模对阅读成绩的影响。详见 Angrist and Lavy (1999)。

① 该数据集的下载地址为 http://www. ats. ucla. edu/stat/stata/examples/methods_matter/chapter9/angrist. dta。

第 29 章　空间计量经济学

29.1　地理学第一定律

许多经济数据都涉及一定的空间位置。比如，研究全国各省的国内生产总值、投资、贸易、研发等数据。本书此前各章很少关注各省经济之间的互动，通常假设各省的变量相互独立。但常识告诉我们，各省经济有着广泛的联系，而且距离越近的省份联系越密切。根据 Tobler(1970)，“所有事物都与其他事物相关联，但较近的事物比较远的事物更关联”(Everything is related to everything else, but near things are more related than distant things)。这被称为“地理学第一定律”(First Law of Geography)。

事实上，各省之间的距离信息并不难获得，比如是否相邻，直线距离或运输距离；只是此前一直未利用这些信息。将各省的变量数据，再加上各省的位置信息，即可得到“空间数据”(spatial data 或 areal data)。所谓空间数据，就是在原来的横截面或面板数据上，加上横截面单位的位置信息(或相互距离)。研究如何处理空间数据的计量经济学分支，称为“空间计量经济学”(spatial econometrics)。空间计量经济学的最大特色在于充分考虑横截面单位之间的空间依赖性(spatial dependence)。更一般地，空间效应(spatial effects)包括空间依赖性与“空间异质性”(spatial heterogeneity)。由于标准的计量经济学也考虑横截面单位之间的异质性(比如异方差)，故空间计量经济学的关注重点为空间依赖性。

空间计量经济学诞生于 20 世纪 70 年代。近年来，空间计量经济学蓬勃发展并进入主流，可归功于两方面。首先，由于 GIS(地理信息系统)的发展，空间数据或包含地理信息的数据(geo-referenced data)日益增多。其次，在经济理论方面，人们越来越关注经济行为人之间的互动，而不仅仅停留于代表性厂商或个人。比如，在考察同伴效应(peer effect)，相邻效应(neighborhood effect)，溢出效应(spillover effect)或网络效应(network effect)时，都需要明确地考虑空间因素。克鲁格曼(Paul Krugman)也因为倡导新贸易理论与新经济地理(new economic geography)而获得 2008 年诺贝尔经济学奖(Krugman, 1991)。

有关空间计量经济学的专著包括 Cliff and Ord(1973, 1981)，Anselin(1988)，Arbia(2006)，LeSage and Pace(2009)，以及沈体雁等(2010)的中文教材。Anselin(2010)提供了较新的文献综述。

29.2　空间权重矩阵

进行空间计量分析的前提是度量区域之间的空间距离。记来自 n 个区域的空间数据为 $\{x_i\}_{i=1}^n$，下标 i 表示区域 i。记区域 i 与区域 j 之间的距离为 w_{ij}，则可定义“空间权重矩阵”

(spatial weighting matrix)如下:

$$\boldsymbol{W}=\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix} \tag{29.1}$$

其中,主对角线上元素 $w_{11}=\cdots=w_{nn}=0$(同一区域的距离为0)。显然,空间权重矩阵 $\boldsymbol{W}$ 为对称矩阵。最常用的距离函数为"相邻"(contiguity),即如果区域 i 与区域 j 有共同的边界,则 $w_{ij}=1$;反之,则 $w_{ij}=0$。比照(国际)象棋中不同棋子的行走路线,相邻关系可分为以下几种,参见图29.1。

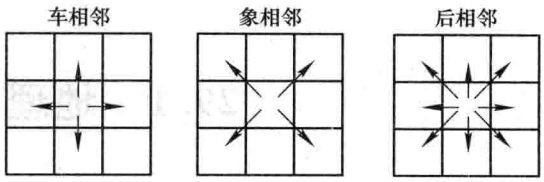


图29.1 常用相邻关系

(1) 车相邻(rook contiguity):两个相邻的区域有共同的边。

(2) 象相邻(bishop contiguity):两个相邻的区域有共同的顶点,但没有共同的边。

(3) 后相邻(queen contiguity):两个相邻的区域有共同的边或顶点。

在实践中,为了区分"边"与"点",常须设定一个最小距离,在此距离以下为点,而在此距离以上为边。究竟使用车、象或后相邻,取决于具体情况。比如,区域 i 与区域 j 仅在一点相交(象相邻),但有一条主要高速公路通过此点连接两区域,则不宜使用车相邻。

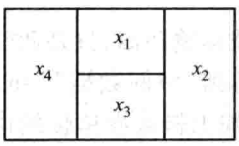


图29.2 假想的四个区域

举一个简单例子,假设有如下四个区域,其变量取值分别为 $\boldsymbol{x}=(x_1\ x_2\ x_3\ x_4)'$,参见图29.2。

针对图29.2中的四个区域,其空间权重矩阵为:

$$\boldsymbol{W}=\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \tag{29.2}$$

矩阵(29.2)的第一行表示,区域1与其余三个区域均相邻;第二行表示,区域2与区域1、区域3相邻,但不与区域4相邻;以此类推。空间权重矩阵考虑的是一阶邻居,还可以考虑二阶邻居,即邻居的邻居,可用矩阵 $\boldsymbol{W}^2$ 来表示。需要注意的是,矩阵 $\boldsymbol{W}^2$ 的主对角线上元素一般不再为0,这意味着邻居的邻居也包括自己。在实践中,有时对空间权重矩阵进行"行标准化"(row standardization),即将矩阵中的每个元素(记为 $\tilde{w}_{ij}$)除以其所在行元素之和,以保证每行元素之和为1:

$$w_{ij}\equiv\frac{\tilde{w}_{ij}}{\sum_j \tilde{w}_{ij}} \tag{29.3}$$

当然,如果区域 i 为孤岛,与其他区域均不相邻,则上式分母为0,并不适用。此时,可将分母改为 $\max\left(1,\sum_j \tilde{w}_{ij}\right)$ ①。不包含孤岛的行标准化矩阵也称为"行随机矩阵"(row-stochastic matrix),因为它的所有元素均介于0与1之间,且每行元素之和为1,在形式上与离散型概率分

① 样本中如果包含孤岛,可能导致程序计算困难,故在实践中一般去掉孤岛。比如,在研究美国各州的空间数据时,常去掉夏威夷州与阿拉斯加州。

布一样。将(29.2)式的空间权重矩阵行标准化可得(仍记为 $\boldsymbol{W}$):

$$\boldsymbol{W}=\begin{pmatrix}0 & 1/3 & 1/3 & 1/3\\ 1/2 & 0 & 1/2 & 0\\ 1/3 & 1/3 & 0 & 1/3\\ 1/2 & 0 & 1/2 & 0\end{pmatrix} \tag{29.4}$$

行标准化的好处在于,如果将行标准化矩阵 $\boldsymbol{W}$ 乘 $\boldsymbol{x}$,则可得到每个区域邻居的平均值。比如,在上例中:

$$\boldsymbol{Wx}=\begin{pmatrix}0 & 1/3 & 1/3 & 1/3\\ 1/2 & 0 & 1/2 & 0\\ 1/3 & 1/3 & 0 & 1/3\\ 1/2 & 0 & 1/2 & 0\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\\ x_4\end{pmatrix}=\begin{bmatrix}(x_2+x_3+x_4)/3\\ (x_1+x_3)/2\\ (x_1+x_2+x_4)/3\\ (x_1+x_3)/2\end{bmatrix} \tag{29.5}$$

比如,区域 1 的邻居为区域 2,3 和 4,而上式右边第一行元素正好为 $(x_2+x_3+x_4)/3$,即区域 1 邻居的平均值;以此类推。比照时间序列中时间滞后(time lag)的概念,$\boldsymbol{Wx}$ 也被称为 $\boldsymbol{x}$ 的"空间滞后"(spatial lag),即 $\boldsymbol{x}$ 邻居的平均取值。

显然,行标准化之后的空间权重矩阵一般不再是对称矩阵,参见矩阵(29.4),这是它的缺陷之一。另外,由于每行元素之和均为 1,这意味着区域 i 所受其邻居的影响之和一定等于区域 j 所受其邻居的影响之和(任意 $i\neq j$);此假定可能过强,这是行标准化的另一局限。

定义相邻关系的另一方法基于区域间的距离。记区域 i 与区域 j 的距离为 d_{ij},可定义空间权重如下:

$$w_{ij}=\begin{cases}1 \text{ 若 } d_{ij}<d\\ 0 \text{ 若 } d_{ij}\geqslant d\end{cases} \tag{29.6}$$

其中,d 为事先给定的距离临界值。另外,也可以不用相邻关系,而直接以距离之倒数(inverse distance)作为空间权重:

$$w_{ij}=\frac{1}{d_{ij}} \tag{29.7}$$

在上式中,距离 d_{ij} 既可以是地理距离,比如直线距离或大圆距离(great circle distance);也可以是基于运输成本或旅行时间的经济距离;甚至社交网络中的距离。

例 林光平等(2005)使用基于地理相邻关系的简单权重矩阵 $\boldsymbol{W}$ 来研究我国 28 个省市在 1978—2002 年期间实际人均 GDP 的收敛情况。但相邻地区经济上的相互关系并不完全相同。例如,河北省虽然在地理上与北京、天津、山西、内蒙古、山东、河南相邻,但很明显河北省与北京、天津的经济密切程度高于其他各省。为此,林光平等(2005)使用地区间人均 GDP 的差额作为测度地区间"经济距离"的指标,并引入经济空间权重矩阵 $\boldsymbol{W}^*=\boldsymbol{W}\times\boldsymbol{E}$,其中矩阵 $\boldsymbol{E}$ 的主对角线元素均为 0,而非主对角线的 (i,j) 元素为 $E_{ij}=\dfrac{1}{|\bar{Y}_i-\bar{Y}_j|}(i\neq j)$,$\bar{Y}_i$ 为地区 i 在样本期间的人均实际 GDP 平均值。结果发现,将经济距离引入空间权重矩阵能更好地拟合我国地区经济的发展状况。

29.3 空间自相关

在确定是否使用空间计量方法时，首先要考察数据是否存在空间依赖性。如果不存在，则使用标准的计量方法即可；如果存在，则可使用空间计量方法。比照时间序列(time series)，空间数据有时也称为“空间序列”(spatial series)。这是因为，时间序列可视为在时间轴上分布的随机过程，而空间数据(序列)则为在空间分布的随机过程。时间序列的一个重要特性是可能存在自相关，特别是一阶自相关。而对于空间序列，自相关的情形则更为复杂；因为时间序列只可能在一个方向上相关(过去影响现在，但现在无法影响过去)，而空间序列则可以在多个方向上相关，而且可以互相影响(x_i 影响 x_j，而 x_j 也影响 x_i)。

“空间自相关”(spatial autocorrelation)可理解为位置相近的区域具有相似的变量取值。如果高值与高值聚集在一起，低值与低值聚集在一起，则为“正空间自相关”(positive spatial autocorrelation)；反之，如果高值与低值相邻，则为“负空间自相关”(negative spatial autocorrelation)；后者较少见。如果高值与低值完全随机地分布，则不存在空间自相关。

考虑空间序列 $\{x_i\}_{i=1}^n$。基于空间自相关的复杂性，文献中提出了一系列度量空间自相关的方法，其中最为流行的是“莫兰指数 I”(Moran's I)(Moran，1950)：

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{29.8}$$

其中，$S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ 为样本方差，w_{ij} 为空间权重矩阵的 (i,j) 元素(用来度量区域 i 与区域 j 之间的距离)，而 $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ 为所有空间权重之和。如果空间权重矩阵为行标准化，则 $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = n$。此时，莫兰指数 I 可写为：

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{29.9}$$

莫兰指数 I 的取值一般介于 -1 到 1 之间，大于 0 表示正自相关，即高值与高值相邻、低值与低值相邻；小于 0 表示负自相关，即高值与低值相邻。一般来说，正自相关比负自相关更为常见。如果莫兰指数 I 接近于 0，则表明空间分布是随机的，不存在空间自相关。莫兰指数 I 可视为观测值与其空间滞后(spatial lag)的相关系数。如果将观测值与其空间滞后画成散点图，称为“莫兰散点图”(Moran scatterplot)，则莫兰指数 I 就是该散点图回归线的斜率。

为了进行严格检验，须导出莫兰指数 I 的渐近分布。考虑原假设“H_0：$\mathrm{Cov}(x_i, x_j) = 0, \forall i \neq j$”(即不存在空间自相关)。在此原假设下，可以证明莫兰指数 I 的期望值为

$$\mathrm{E}(I) = \frac{-1}{n-1} \tag{29.10}$$

莫兰指数 I 的方差表达式更为复杂，记为 $\mathrm{Var}(I)$。可以证明，标准化的莫兰指数 I 服从渐近标准正态分布：

$$I^* \equiv \frac{I - \mathrm{E}(I)}{\sqrt{\mathrm{Var}(I)}} \xrightarrow{d} N(0,1) \tag{29.11}$$

因此,可使用标准正态的临界值进行检验。在使用莫兰指数 I 检验空间自相关时,须注意两个问题。问题之一,莫兰指数 I 取决于空间矩阵 $\boldsymbol{W}$,如果空间矩阵设定不正确,则可能导致错误的结果。问题之二,莫兰指数 I 的核心成分为 $(x_i-\bar{x})(x_j-\bar{x})$,其隐含假设是 $\{x_i\}_{i=1}^n$ 的期望值为常数(constant mean),不存在任何趋势(trend)。如果存在趋势,则可能导致检验结果出现偏差。为了解决问题一,须仔细选择合适的空间矩阵,或使用不同的空间矩阵以考察结果的稳健性。为了解决问题二,可引入协变量,通过回归的方法去掉趋势,然后对残差项进行莫兰指数 I 检验。

以上的莫兰指数 I 也被称为"全局莫兰指数 I"(global Moran's I),因为它考察的是整个空间序列 $\{x_i\}_{i=1}^n$ 的空间集聚情况。如果想知道某区域 i 附近的空间集聚情况,则可使用"局部莫兰指数 I"(local Moran's I):

$$I_i = \frac{(x_i-\bar{x})}{S^2}\sum_{j=1}^{n} w_{ij}(x_j-\bar{x}) \tag{29.12}$$

局部莫兰指数 I 的含义与全局莫兰指数 I 相似。正的 I_i 表示区域 i 的高(低)值被周围的高(低)值所包围;负的 I_i 则表示区域 i 的高(低)值被周围的低(高)值所包围。莫兰指数 I 并非唯一的空间自相关指标,另一常用指标为"吉尔里指数 C"(Geary's C)(Geary,1954),也称为"吉尔里相邻比率"(Geary's Contiguity Ratio):

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-x_j)^2}{2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]} \tag{29.13}$$

与莫兰指数 I 不同,吉尔里指数 C 的核心成分为 $(x_i-x_j)^2$。吉尔里指数 C 的取值一般介于 0 到 2 之间(2 不是严格上界),大于 1 表示负相关,等于 1 表示不相关,而小于 1 表示正相关。因此,吉尔里指数 C 与莫兰指数 I 呈反向变动;一般认为,前者比后者对于局部空间自相关更为敏感。在不存在空间自相关的原假设下,可以证明吉尔里指数 C 的期望值为 1,而方差的表达式较复杂,记为 Var(C)。可以证明,标准化的吉尔里指数 C 服从渐近标准正态分布:

$$C^* \equiv \frac{C-1}{\sqrt{\mathrm{Var}(C)}} \xrightarrow{d} N(0,1) \tag{29.14}$$

因此,可使用标准化的吉尔里指数 C^* 检验空间自相关。然而,莫兰指数 I 与吉尔里指数 C 的共同缺点在于,即无法分别"热点"(hot spot)与"冷点"(cold spot)区域。所谓热点区域,即高值与高值聚集的区域;而冷点区域则是低值与低值聚集的区域。热点区域与冷点区域都表现为正自相关。为此,Getis and Ord(1992)提出了以下"Getis - Ord 指数 G":

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_i x_j} \tag{29.15}$$

其中,$x_i>0, \forall i$;而 w_{ij} 来自非标准化的对称空间权重矩阵,且所有元素均为 0 或 1。显然,如果样本中高值聚集在一起,则 G 较大;如果低值聚集在一起,则 G 较小。在无空间自相关的原假设下,可以证明,$E(G) = \dfrac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{ij}}{n(n-1)}$。如果 G 值大于此期望值,则表示存在热点区域;如果 G 值小于此期望值,则表示存在冷点区域。类似地,标准化的 G 也服从渐近标准正态分布:

$$G^* \equiv \frac{G-E(G)}{\sqrt{\mathrm{Var}(G)}} \xrightarrow{d} N(0,1) \tag{29.16}$$

如果 $G^*>1.96$,则可在 5% 水平上拒绝无空间自相关的原假设,认为存在空间正自相关,且

存在热点区域。反之，如果 $G^* < -1.96$，则可在5%水平上拒绝无空间自相关的原假设，认为存在空间正自相关，且存在冷点区域。如果要考察某区域 i 是否为热点或冷点，则可使用“局部 Getis - Ord 指数 G”：

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j} \tag{29.17}$$

需要指出的是，以上各种空间自相关指标仅提供是否存在空间效应的初步检验，深入检验有赖于建立正式的空间计量模型①。有关空间自相关的指标计算与假设检验可通过下载非官方 Stata 命令来实现（或通过命令“findit spatreg”查找并下载）：

```
. net install sg162.pkg
```

此命令将下载安装以 spat 开头的系列命令：spatwmat（用于定义空间权重矩阵）；spatgsa（进行全局空间自相关检验，“gsa”表示“global spatial autocorrelation”）；spatlsa（进行局部空间自相关检验，“lsa”表示“local spatial autocorrelation”）；spatcorr（考察空间自相关指标对距离临界值 d 的依赖性，参见方程(29.6)）；spatdiag（针对 OLS 回归结果，诊断是否存在空间效应）；以及 spatreg（估计空间滞后与空间误差模型，参见下两节）。

下面以数据集 columbusdata.dta 与 columbusswm.dta 为例。这两个数据集来自 Anselin(1988)，前者包含美国俄亥俄州哥伦布市（Columbus, Ohio）49个社区的社区编号（id）、犯罪率（crime）、房价（hoval）与家庭收入（income）的数据，而后者为这49个社区基于相邻关系的空间权重矩阵；这些社区的分布参见图29.3。本研究考察房价与家庭收入对犯罪率的作用。显然，各个社区的犯罪率是相关的，故应进行空间计量分析。

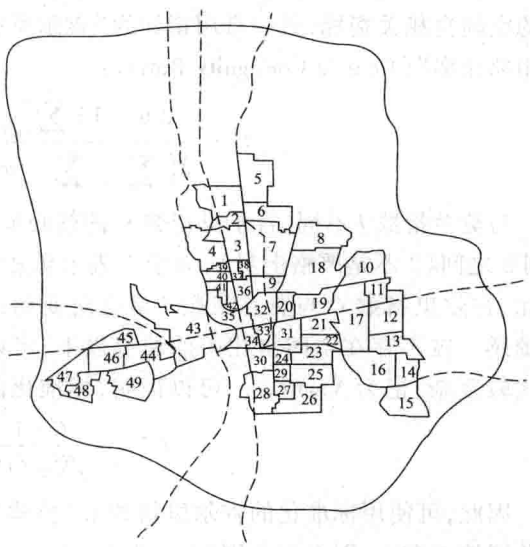


图29.3 美国俄亥俄州哥伦布市的49个社区

运行 spat 系列命令（除 spatcorr 外）的前提是通过命令 spatwmat 来定义空间权重矩阵。为此，将数据集 columbusswm.dta 与 columbusdata.dta 置于 Stata 的当前目录下。如果不知道当前目录在哪里，可通过输入命令 pwd（表示“path of working directory”）来显示当前目录；然后再输入以下命令：

```
. spatwmat using columbusswm.dta, name(W)
```

其中，必选项“name(W)”表示将根据数据集 columbusswm.dta 生成的空间权重矩阵命令为 W。更多选择项参见“help spatwmat”。

```
The following matrix has been created:

1. Imported binary weights matrix W
   Dimension: 49x49
```

① 空间相关系数与空间计量模型的关系正如简单相关系数与多元回归的关系，即简单相关系数仅提供初步证据，而更深入的分析则有赖于多元回归模型。

上表显示，已生成 49 × 49 的空间权重矩阵 $\boldsymbol{W}$，其中元素均为 0 或 1(binary)。如果想看此矩阵，可输入命令“matrix list W”(为节省空间，运行结果从略)。

下面，计算被解释变量 crime 的全局自相关指标及相应检验。

```
. use columbusdata.dta, clear
. spatgsa crime, weights(W) moran geary go twotail
```

其中，必选项“weights(W)”指定空间权重矩阵为 $\boldsymbol{W}$。选择项“moran geary go”分别表示计算莫兰指数 I，吉尔里指数 C 以及 Getis-Ord 指数 G；这三个选择项至少须选一。选择项“twotail”表示进行双边检验，默认为单边检验(即认为只可能存在正空间自相关)。

Measures of global spatial autocorrelation

Weights matrix

Name: W
Type: Imported (binary)
Row-standardized: No

Moran's I

Variables	I	E(I)	sd(I)	z	p-value*
crime	0.521	-0.021	0.087	6.212	0.000

Geary's c

Variables	c	E(c)	sd(c)	z	p-value*
crime	0.584	1.000	0.109	-3.835	0.000

Getis & Ord's G

Variables	G	E(G)	sd(G)	z	p-value*
crime	0.126	0.099	0.006	4.714	0.000

*2-tail test

从上表可知，这三个全局空间自相关指标均强烈拒绝“无空间自相关”的原假设，即认为存在空间自相关。下面计算局部空间自相关指标(为节省空间，仅计算局部莫兰指数 I)。

```
. spatlsa crime, w(W) moran twotail
```

命令 spatlsa 的句型及选择项与 spatgsa 类似。

Measures of local spatial autocorrelation

Weights matrix

Name: W
Type: Imported (binary)
Row-standardized: No

Moran's Ii (Residential burglaries & vehicle thefts pe)

Location	Ii	E(Ii)	sd(Ii)	z	p-value*
1	1.586	-0.063	1.674	0.985	0.325
2	0.019	-0.083	1.912	0.054	0.957
3	0.460	-0.125	2.289	0.255	0.798
4	-7.442	-0.083	1.912	-3.850	0.000
5	1.474	-0.042	1.381	1.097	0.273
6	0.375	-0.083	1.912	0.240	0.810
7	2.429	-0.167	2.581	1.006	0.315
8	-0.363	-0.042	1.381	-0.233	0.816
9	4.409	-0.125	2.289	1.981	0.048
10	0.161	-0.083	1.912	0.128	0.898
11	-0.324	-0.063	1.674	-0.156	0.876
12	2.703	-0.063	1.674	1.652	0.099
13	4.959	-0.083	1.912	2.638	0.008
14	2.495	-0.063	1.674	1.528	0.126
15	0.935	-0.042	1.381	0.707	0.479
16	4.846	-0.104	2.113	2.342	0.019
17	3.414	-0.208	2.815	1.287	0.198
18	0.195	-0.146	2.443	0.140	0.889
19	-0.024	-0.125	2.289	0.044	0.965
20	1.180	-0.104	2.113	0.607	0.544
21	-0.496	-0.083	1.912	-0.216	0.829
22	0.214	-0.083	1.912	0.156	0.876
23	-0.210	-0.146	2.443	-0.026	0.979
24	3.465	-0.125	2.289	1.569	0.117
25	-0.103	-0.104	2.113	0.000	1.000
26	1.090	-0.063	1.674	0.689	0.491
27	0.814	-0.083	1.912	0.470	0.639
28	0.272	-0.083	1.912	0.186	0.853
29	0.035	-0.125	2.289	0.070	0.944
30	2.539	-0.125	2.289	1.164	0.244
31	8.793	-0.188	2.704	3.321	0.001
32	11.771	-0.146	2.443	4.877	0.000
33	10.351	-0.125	2.289	4.577	0.000
34	8.276	-0.104	2.113	3.965	0.000
35	1.600	-0.146	2.443	0.714	0.475
36	9.910	-0.167	2.581	3.904	0.000
37	4.529	-0.146	2.443	1.913	0.056
38	7.290	-0.104	2.113	3.498	0.000
39	0.501	-0.083	1.912	0.306	0.760
40	3.692	-0.125	2.289	1.667	0.095
41	2.400	-0.063	1.674	1.471	0.141
42	2.763	-0.104	2.113	1.357	0.175
43	0.465	-0.063	1.674	0.315	0.753
44	2.114	-0.083	1.912	1.149	0.250
45	2.035	-0.042	1.381	1.503	0.133
46	6.216	-0.104	2.113	2.991	0.003
47	2.304	-0.042	1.381	1.698	0.089
48	2.499	-0.042	1.381	1.840	0.066
49	2.173	-0.042	1.381	1.604	0.109

*2-tail test

上表分别列出了 49 个社区的莫兰指数 I 及检验结果。对于某些社区，可以强烈拒绝“无空间自相关”的原假设，这与全局空间自相关的检验结果相一致。

29.4 空间自回归模型

对于时间序列，最常见的建模方式为一阶自回归 AR(1)。但空间自回归的形式更为复杂，因为空间滞后可以来自不同的方向，而且可以双向。对于空间序列$\{y_i\}_{i=1}^n$(样本容量为 n)，即使假设空间自相关的形式为线性，待估计的参数在理论上最多可达(n^2-n)个(每个区域最多可受 $n-1$ 个区域影响，共有 n 个区域)，显然大大超出了样本容量许可的范围。为此，必须假设空间依赖性服从某种共同的模式，才能简化参数。

作为参照，首先回顾时间序列的一阶自相关模型：

$$y_t=\beta y_{t-1}+\varepsilon_t \quad (t=2,\cdots,n) \tag{29.18}$$

为了示例方便，假设 $n=4$，而 $y_1=\varepsilon_1$。将这 4 个方程叠放在一起可得：

$$y=\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\beta\underbrace{\begin{pmatrix}0&0&0&0\\1&0&0&0\\0&1&0&0\\0&0&1&0\end{pmatrix}}_{\boldsymbol{W}}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix}\equiv\beta\boldsymbol{W}\boldsymbol{y}+\boldsymbol{\varepsilon} \tag{29.19}$$

其中，矩阵 $\boldsymbol{W}$ 的次主对角线元素均为 1，而其他元素都为 0。矩阵 $\boldsymbol{W}$ 可视为“时间滞后矩阵”，它之所以呈现出次对角线元素为 1 而其他元素为 0 的规律性排列，正是由于时间滞后在方向上的单一性。对于空间序列，矩阵 $\boldsymbol{W}$ 则为空间权重矩阵，其元素排列方式表现出多样性(因为空间自相关可以多方向)。为此，引入以下“空间自回归模型”(Spatial Autoregression，简记 SAR)：

$$\boldsymbol{y}=\lambda\boldsymbol{W}\boldsymbol{y}+\boldsymbol{\varepsilon} \tag{29.20}$$

其中，$\boldsymbol{W}$ 为已知的空间权重矩阵(非随机)，而空间依赖性仅由单一参数 λ 来刻画(待估参数大大减少)。λ 度量空间滞后 $\boldsymbol{W}\boldsymbol{y}$ 对 $\boldsymbol{y}$ 的影响，称为“空间自回归系数”(spatial autoregressive parameter)。在形式上，此模型与时间滞后模型(29.19)完全相同，只是空间权重矩阵更为复杂。因此，方程(29.20)也称为“空间滞后模型”(Spatial Lag Model)[①]。直观来看，相邻地区的被解释变量(比如犯罪率)可能相互依赖，并最终形成一个均衡的结果。又比如，假设以地区税收为被解释变量，则不同地区的政府出于相互竞争或博弈的考虑(比如，竞相吸引 FDI)，在制定本地区税收时，会考虑周边地区的税收水平。

作为一个例子，假设 $\boldsymbol{W}$ 为表达式(29.2)，则方程(29.20)可写为：

$$\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\lambda\begin{pmatrix}0&1&1&1\\1&0&1&0\\1&1&0&1\\1&0&1&0\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix}=\lambda\begin{pmatrix}y_2+y_3+y_4\\y_1+y_3\\y_1+y_2+y_4\\y_1+y_3\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix} \tag{29.21}$$

显然，这是一个联立方程系统。如果对(29.21)进行 OLS 估计，将会导致联立方程偏差(simultaneity bias)，参见第 10、24 章。正是由于空间依赖性，导致变量$\{y_i\}_{i=1}^n$之间互相影响，从而产生内生性。更一般地，可在方程(29.20)中加入自变量：

---

① 也称为“混合回归 - 空间自回归模型”(mixed regressive-spatial autoregressive model)。

$$\boldsymbol{y}=\lambda \boldsymbol{W}\boldsymbol{y}+\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{29.22}$$

其中，$\boldsymbol{X}$ 为 $n\times k$ 数据矩阵（参见第 3 章），包括 k 列解释变量；而 $\boldsymbol{\beta}_{k\times 1}$ 为相应系数。方程（29.22）也称为 SAR 模型，而方程（29.20）有时称为“纯 SAR 模型”（Pure SAR）。如果 $\lambda=0$，则方程（29.22）简化为一般的线性回归模型。因此，可通过检验原假设“$H_0:\lambda=0$”来考察是否存在空间效应。对于空间自回归模型（29.22），常使用最大似然估计。首先，假设扰动项 $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$。其次，方程（29.22）可写为

$$\boldsymbol{A}\boldsymbol{y}\equiv(\boldsymbol{I}-\lambda \boldsymbol{W})\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{29.23}$$

其中，$\boldsymbol{A}\equiv\boldsymbol{I}-\lambda\boldsymbol{W}$。由于雅可比行列式 $J\equiv\left|\dfrac{\partial\boldsymbol{\varepsilon}}{\partial\boldsymbol{y}}\right|=\left|\dfrac{\partial(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{\partial\boldsymbol{y}}\right|=\left|\dfrac{\partial\boldsymbol{A}\boldsymbol{y}}{\partial\boldsymbol{y}}\right|=|\boldsymbol{A}'|=|\boldsymbol{A}|$[①]，故根据多维正态的密度函数公式（参见第 2 章），可写出样本的似然函数：

$$L(\boldsymbol{y}\mid\lambda,\sigma^2,\boldsymbol{\beta})=(2\pi\sigma^2)^{-n/2}(\operatorname{abs}|\boldsymbol{A}|)\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \tag{29.24}$$

其中，$\operatorname{abs}|\boldsymbol{A}|$ 表示行列式 $|\boldsymbol{A}|$ 的绝对值。由此可得对数似然函数为

$$\ln L(\boldsymbol{y}\mid\lambda,\sigma^2,\boldsymbol{\beta})=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2+\ln(\operatorname{abs}|\boldsymbol{A}|)-\frac{1}{2\sigma^2}(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \tag{29.25}$$

类似于对古典线性回归模型的 MLE 估计（参见第 6 章），此最大化问题可分两步进行。第一步，在给定 λ 的情况下，选择最优的 $\boldsymbol{\beta},\sigma^2$。第二步，代入第一步的最优 $\boldsymbol{\beta},\sigma^2$，选择最优的 λ。

在第一步，选择 $\boldsymbol{\beta},\sigma^2$ 使得 $\ln L(\boldsymbol{y}\mid\lambda,\sigma^2,\boldsymbol{\beta})$ 最大。由于 $\boldsymbol{\beta}$ 只出现于（29.25）式的最后一项，故这等价于使 $(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ 最小，即 $\boldsymbol{A}\boldsymbol{y}$ 对 $\boldsymbol{X}$ 进行回归：

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A}\boldsymbol{y}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\lambda\boldsymbol{W})\boldsymbol{y}\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}-\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\equiv\hat{\boldsymbol{\beta}}_0-\lambda\hat{\boldsymbol{\beta}}_L\end{aligned} \tag{29.26}$$

其中，$\hat{\boldsymbol{\beta}}_0\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$（$\boldsymbol{y}$ 对 $\boldsymbol{X}$ 的回归系数），而 $\hat{\boldsymbol{\beta}}_L\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$（$\boldsymbol{W}\boldsymbol{y}$ 对 $\boldsymbol{X}$ 的回归系数）。因此，只要知道 λ，即可计算 $\hat{\boldsymbol{\beta}}$。在方程（29.25）中，对 σ^2 求偏导可得 σ^2 的 MLE 估计：

$$\hat{\sigma}^2=\frac{\boldsymbol{e}'\boldsymbol{e}}{n}=\frac{(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{A}\boldsymbol{y})'(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{A}\boldsymbol{y})}{n} \tag{29.27}$$

其中，$\boldsymbol{e}$ 为 $\boldsymbol{A}\boldsymbol{y}$ 对 $\boldsymbol{X}$ 回归的残差向量，$\boldsymbol{M}_{\boldsymbol{X}}\equiv\boldsymbol{I}_n-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 为消灭矩阵（参见第 3 章）。由于 $\boldsymbol{A}\boldsymbol{y}=(\boldsymbol{I}-\lambda\boldsymbol{W})\boldsymbol{y}$，故

$$\boldsymbol{e}=\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{A}\boldsymbol{y}=\boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{I}-\lambda\boldsymbol{W})\boldsymbol{y}=\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}-\lambda\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{W}\boldsymbol{y}\equiv\boldsymbol{e}_0-\lambda\boldsymbol{e}_L \tag{29.28}$$

其中，$\boldsymbol{e}_0\equiv\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$（$\boldsymbol{y}$ 对 $\boldsymbol{X}$ 的回归残差），而 $\boldsymbol{e}_L\equiv\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{W}\boldsymbol{y}$（$\boldsymbol{W}\boldsymbol{y}$ 对 $\boldsymbol{X}$ 的回归残差）。将表达式（29.28）代入（29.27）可得：

$$\hat{\sigma}^2=\frac{\boldsymbol{e}'\boldsymbol{e}}{n}=\frac{(\boldsymbol{e}_0-\lambda\boldsymbol{e}_L)'(\boldsymbol{e}_0-\lambda\boldsymbol{e}_L)}{n} \tag{29.29}$$

因此，只要知道 λ，也可以计算 $\hat{\sigma}^2$。在第二步，将 $\hat{\boldsymbol{\beta}}(\lambda)$，$\hat{\sigma}^2(\lambda)$ 代入对数似然函数（29.25），则可得到“集中对数似然函数”（concentrated log likelihood function），它只是 λ 的函数。然而，由于 λ 出现在行列式 $|\boldsymbol{A}|=|\boldsymbol{I}-\lambda\boldsymbol{W}|$，给计算带来不便。为此，可利用等式 $|\boldsymbol{A}|=\prod_{i=1}^{n}(1-\lambda v_i)$ 来

① 此处的问题是根据 $\boldsymbol{\varepsilon}$ 的概率密度写出 $\boldsymbol{y}$ 的概率密度。在作此变换时，需用到雅可比行列式，参见 Greene（2012，Appendix B. 7. 4），或较为深入的概率统计教材。

计算，其中 $v_1,\cdots,v_n$ 为矩阵 $\boldsymbol{A}$ 特征值(eigenvalues)。另外，为了保证扰动项协方差矩阵为正定，还须限制 λ 的取值为 $\frac{1}{v_{\min}}<\lambda<\frac{1}{v_{\max}}$，其中 $v_{\min}$ 与 $v_{\max}$ 分别为矩阵 $\boldsymbol{A}$ 的最小与最大特征值，而 $v_{\min}$ 一定为负数。

对于 MLE 估计量的渐近协方差矩阵，一般通过信息矩阵来估计，即

$$\boldsymbol{I}(\boldsymbol{\theta})^{-1}\equiv-\left\{\mathrm{E}\left[\frac{\partial^2\ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right\}^{-1} \tag{29.30}$$

其中，$\boldsymbol{\theta}=(\lambda,\sigma^2,\boldsymbol{\beta})$。需要注意的是，对于空间自回归模型(29.22) $\boldsymbol{y}=\lambda\boldsymbol{W}\boldsymbol{y}+\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$，解释变量 $\boldsymbol{X}$ 对 $\boldsymbol{y}$ 的边际效应并非 $\boldsymbol{\beta}$，因为 $\boldsymbol{X}$ 对 $\boldsymbol{y}$ 产生作用后，$\boldsymbol{y}$ 之间还会相互作用(即不同区域的被解释变量互相影响)，直至达到一个新的均衡。由方程(29.23)可知：

$$\boldsymbol{y}=(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \tag{29.31}$$

容易验证，$(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}=\boldsymbol{I}+\lambda\boldsymbol{W}+\lambda^2\boldsymbol{W}^2+\lambda^3\boldsymbol{W}^3+\cdots$。假设 $\boldsymbol{X}$ 中包含 K 个解释变量，并记第 r 个解释变量为 $\boldsymbol{x}_r=(x_{1r},x_{2r}\cdots x_{nr})'$($n\times1$ 列向量)，则 $\boldsymbol{X}\boldsymbol{\beta}=(\boldsymbol{x}_1\cdots\boldsymbol{x}_K)(\beta_1\cdots\beta_K)'=\sum_{r=1}^{K}\beta_r\boldsymbol{x}_r$。因此，方程(29.31)可以写为

$$\boldsymbol{y}=\sum_{r=1}^{K}\beta_r(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{x}_r+(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}\equiv\sum_{r=1}^{K}\boldsymbol{S}_r(\boldsymbol{W})\boldsymbol{x}_r+(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \tag{29.32}$$

其中，$S_r(\boldsymbol{W})\equiv\beta_r(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}$ 为依赖于 β_r 与 $\boldsymbol{W}$ 的 $n\times n$ 矩阵。将方程(29.32)展开来写：

$$\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}S_r(\boldsymbol{W})_{11} & S_r(\boldsymbol{W})_{12} & \cdots & S_r(\boldsymbol{W})_{1n}\\S_r(\boldsymbol{W})_{21} & S_r(\boldsymbol{W})_{22} & \cdots & S_r(\boldsymbol{W})_{2n}\\\vdots & \vdots & & \vdots\\S_r(\boldsymbol{W})_{n1} & S_r(\boldsymbol{W})_{n2} & \cdots & S_r(\boldsymbol{W})_{nn}\end{pmatrix}\begin{pmatrix}x_{1r}\\x_{2r}\\\vdots\\x_{nr}\end{pmatrix}+(\boldsymbol{I}-\lambda\boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \tag{29.33}$$

其中，$S_r(\boldsymbol{W})_{ij}$ 为 $S_r(\boldsymbol{W})$ 的 (i,j) 元素。根据方程(29.33)可知：

$$\frac{\partial y_i}{\partial x_{jr}}=S_r(\boldsymbol{W})_{ij} \tag{29.34}$$

由此可见，区域 j 的变量 x_{jr} 对任意区域 i 的被解释变量都可能有影响，这正是空间计量模型的真谛。特别地，如果 $j=i$，则有

$$\frac{\partial y_i}{\partial x_{ir}}=S_r(\boldsymbol{W})_{ii} \tag{29.35}$$

方程(29.35)表明，区域 i 的变量 x_{ir} 对本区域被解释变量 y_i 的“直接效应”(direct effect)为 $S_r(\boldsymbol{W})_{ii}$，即矩阵 $S_r(\boldsymbol{W})$ 主对角线上的第 i 个元素。因此，如果将矩阵 $S_r(\boldsymbol{W})$ 主对角线上的所有元素进行平均，即可得到变量 x_r 的“平均直接效应”(average direct effect)：

$$\text{平均直接效应}=\frac{1}{n}\mathrm{trace}[S_r(\boldsymbol{W})] \tag{29.36}$$

其中，$\mathrm{trace}[S_r(\boldsymbol{W})]$ 为矩阵 $S_r(\boldsymbol{W})$ 的迹，即主对角线元素之和。另一方面，假设所有区域的变量 x_r 都变化一个单位，其对区域 i 被解释变量 y_i 的“总效应”(total effect)为矩阵 $S_r(\boldsymbol{W})$ 的第 i 行元素之和，即 $\sum_{j=1}^{n}S_r(\boldsymbol{W})_{ij}$。如果将对所有区域的总效应进行平均，则可得到变量 x_r 的“平均总效应”(average total effect)，即矩阵 $S_r(\boldsymbol{W})$ 所有元素的平均：

$$\text{平均总效应}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}S_r(\boldsymbol{W})_{ij}=\frac{1}{n}\boldsymbol{i}_n'S_r(\boldsymbol{W})\boldsymbol{i}_n \tag{29.37}$$

其中，$\boldsymbol{i}_n=(1\cdots1)'$为$n\times1$列向量。最后，可以定义变量$x_r$的“平均间接效应”(average indirect effect)为平均总效应与平均直接效应之差：

$$\text{平均间接效应}=\frac{1}{n}\{\boldsymbol{i}_n'S_r(\boldsymbol{W})\boldsymbol{i}_n-\text{trace}[S_r(\boldsymbol{W})]\} \tag{29.38}$$

29.5 空间杜宾模型

对于空间效应的另一建模方式是，假设区域i的被解释变量y_i依赖于其邻居的自变量：

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{WX\delta}+\boldsymbol{\varepsilon} \tag{29.39}$$

其中，$\boldsymbol{WX\delta}$表示来自邻居自变量的影响，而$\boldsymbol{\delta}$为相应的系数向量。例如，区域i的犯罪率不仅依赖于本区域的警力，还可能依赖于相邻区域的警力。此模型称为“空间杜宾模型”(Spatial Durbin Model)，简记SDM。由于方程(29.39)不存在内生性，故可直接进行OLS估计；只是解释变量$\boldsymbol{X}$与$\boldsymbol{WX}$之间可能存在多重共线性。如果$\boldsymbol{\delta}=\boldsymbol{0}$，则方程(29.39)简化为一般的线性回归模型。

将空间杜宾模型与空间自回归模型相结合，可得

$$\boldsymbol{y}=\lambda\boldsymbol{Wy}+\boldsymbol{X\beta}+\boldsymbol{WX\delta}+\boldsymbol{\varepsilon} \tag{29.40}$$

方程(29.40)有时也称为空间杜宾模型。

29.6 空间误差模型

空间依赖性还可能通过误差项来体现。考虑以下“空间误差模型”(Spatial Errors Model，简记SEM)：

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{u} \tag{29.41}$$

其中，扰动项$\boldsymbol{u}$的生成过程为

$$\boldsymbol{u}=\rho\boldsymbol{Mu}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n) \tag{29.42}$$

其中，$\boldsymbol{M}$为空间权重矩阵。该模型显示，扰动项$\boldsymbol{u}$存在空间依赖性。这意味着，不包含在$\boldsymbol{X}$中但对$\boldsymbol{y}$有影响的遗漏变量存在空间相关性，或者不可观测的随机冲击存在空间相关性。如果$\rho=0$，则简化为一般的线性回归模型。由方程(29.42)可得：

$$\boldsymbol{Bu}\equiv(\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{u}=\boldsymbol{\varepsilon} \tag{29.43}$$

其中，$\boldsymbol{B}\equiv\boldsymbol{I}-\rho\boldsymbol{M}$。对于SEM模型(29.41)，尽管扰动项存在自相关，但由于不存在内生性，故OLS估计是一致的；但由于忽略了扰动项的自相关而损失了效率(参见第8章)。最有效的方法为MLE估计。样本的对数似然函数可写为

$$\ln L(y|\rho,\sigma^2,\beta)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2+\ln(\text{abs}|\boldsymbol{B}|)-\frac{1}{2\sigma^2}(y-\boldsymbol{X\beta})'\boldsymbol{B}'\boldsymbol{B}(y-\boldsymbol{X\beta}) \tag{29.44}$$

在最大化上式时，仍可分两步进行。第一步，通过GLS计算$\hat{\boldsymbol{\beta}}$，即最小化上式最后一项$(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{y}-\boldsymbol{X\beta})$；然后通过一阶条件得到$\hat{\sigma}^2$。第二步，将$\hat{\boldsymbol{\beta}}(\rho)$，$\hat{\sigma}^2(\rho)$代入对数似然函数(29.44)，得到集中对数似然函数(仅为ρ的函数)，求解最优的$\hat{\rho}$。一个技术性问题是，在第一步

中，由于目标函数涉及 $\boldsymbol{B}=\boldsymbol{I}-\rho\boldsymbol{M}$，故得不到 $\hat{\boldsymbol{\beta}}(\rho)$ 与 $\hat{\sigma}^2(\rho)$ 的解析表达式。为此，需要在第一步与第二步之间进行迭代，即给定 $\hat{\rho}^{(1)}$，通过第一步得到 $\hat{\boldsymbol{\beta}}(\hat{\rho}^{(1)})$，$\hat{\sigma}^2(\hat{\rho}^{(1)})$ 的具体取值，再通过第二步得到 $\hat{\rho}^{(2)}$，以此类推，直至 $\hat{\rho}$ 收敛。其中，初始值 $\hat{\rho}^{(1)}$ 可通过在第一步进行 OLS 估计，然后代入第二步而得到。

继续以数据集 columbusdata. dta 为例，演示空间自回归与空间误差模型的 Stata 操作。首先，通过命令 spatdiag 诊断是否存在空间效应。为此，须先进行 OLS 回归。

```
. reg crime hoval income
```

Source	SS	df	MS	Number of obs =	49
				F(2, 46) =	28.39
Model	7423.32674	2	3711.66337	Prob > F =	0.0000
Residual	6014.89281	46	130.758539	R-squared =	0.5524
				Adj R-squared =	0.5329
Total	13438.2195	48	279.962907	Root MSE =	11.435

crime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hoval	-.2739315	.1031987	-2.65	0.011	-.4816597	-.0662033
income	-1.597311	.3341308	-4.78	0.000	-2.269881	-.9247405
_cons	68.61896	4.735486	14.49	0.000	59.08692	78.151

上表显示，房价(hoval)与家庭收入(income)对于犯罪率均有显著的负作用。然而，如果存在空间效应，则 OLS 估计是有偏差的。

```
. spatdiag, weights(W)
```

其中，必选项“weights(W)”指定空间权重矩阵为 $\boldsymbol{W}$。有关如何生成此空间权重矩阵，参见本章第 3 节。

Diagnostic tests for spatial dependence in OLS regression

Fitted model

```
crime = hoval + income
```

Weights matrix

```
Name: W
Type: Imported (binary)
Row-standardized: No
```

Diagnostics

Test	Statistic	df	p-value
Spatial error:			
Moran's I	1.011	1	0.312
Lagrange multiplier	6.804	1	0.009
Robust Lagrange multiplier	1.759	1	0.185
Spatial lag:			
Lagrange multiplier	13.787	1	0.000
Robust Lagrange multiplier	8.741	1	0.003

上表显示，针对空间误差(Spatial error)的三个检验中，有一个拒绝了“无空间自相关”的原假设；而针对空间滞后(Spatial lag)的两个检验均拒绝了此原假设；这些结果再次表明应进行空间计量分析。

下面，使用命令 spatreg 估计空间滞后与空间误差模型。为了运行命令 spatreg，除了需要空间权重矩阵 $\boldsymbol{W}$ 外，还需要先计算矩阵 $\boldsymbol{W}$ 的特征值向量。为此，再次使用命令 spatwmat。

```
. spatwmat using columbusswm.dta, name(W) eigenval(E)
```

其中，选择项“eigenval(E)”表示，计算空间权重矩阵 W 的特征值向量，并记为 E。

```
The following matrices have been created:

1. Imported binary weights matrix W
   Dimension: 49x49

2. Eigenvalues matrix E
   Dimension: 49x1
```

上表显示，特征值向量 $\boldsymbol{E}$ 为 49×1 的列向量。

```
. spatreg crime hoval income,weights(W) eigenval(E) model(lag) nolog
```

其中，必选项“weights(W)”与“eigenval(E)”分别用于指定空间权重矩阵及其特征值向量。选择项“model(lag)”表示估计空间自回归模型(SAR)。另外，还可使用选择项“robust”得到异方差稳健的标准误[①]。

```
Weights matrix
 Name: W
 Type: Imported (binary)
 Row-standardized: No

Spatial lag model                                 Number of obs   =        49
                                                  Variance ratio  =     0.660
                                                  Squared corr.   =     0.660
Log likelihood = -180.65071                       Sigma           =      9.66
```

crime	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
crime						
hoval	-.2516329	.087319	-2.88	0.004	-.422775	-.0804909
income	-1.154463	.3037289	-3.80	0.000	-1.749761	-.5591653
_cons	51.60482	5.888379	8.76	0.000	40.06381	63.14583
rho	.0545427	.0138575	3.94	0.000	.0273825	.0817029

```
Wald test of rho=0:                     chi2(1) =  15.492 (0.000)
Likelihood ratio test of rho=0:         chi2(1) =  13.453 (0.000)
Lagrange multiplier test of rho=0:      chi2(1) =  13.787 (0.000)

Acceptable range for rho: -1.536 < rho < 1.000
```

上表显示，空间自回归系数(记为 rho)的估计值为 0.055，且在 1% 水平上显著，故存在空间自回归效应。对于空间自回归系数的似然比检验与拉格朗日检验也得到同样的结果。SAR 模型的自变量回归系数与 OLS 模型有所不同。下面估计空间误差模型。

① 然而，如果存在异方差，则 MLE 估计量不一致；使用异方差稳健的标准误也就意义不大，参见第 6 章。故此处未使用“robust”选择项。

```
. spatreg crime hoval income, weights(W) eigenval(E) model(error)
nolog
```

其中，选择项“model(error)”表示估计空间误差模型(SEM)。

```
Weights matrix
  Name: W
  Type: Imported (binary)
  Row-standardized: No

Spatial error model                              Number of obs    =        49
                                                 Variance ratio   =     0.402
                                                 Squared corr.    =     0.551
Log likelihood = -182.26823                      Sigma            =      9.98

------------------------------------------------------------------------------
       crime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
crime        |
       hoval |  -.2605768   .0917459    -2.84   0.005    -.4403954   -.0807582
      income |  -1.289794   .3255491    -3.96   0.000    -1.927858   -.6517291
       _cons |   54.45472   6.149929     8.85   0.000     42.40108    66.50836
-------------+----------------------------------------------------------------
      lambda |   .0361054   .0114345     3.16   0.002     .0136943    .0585165
------------------------------------------------------------------------------
Wald test of lambda=0:                   chi2(1) =   9.970 (0.002)
Likelihood ratio test of lambda=0:       chi2(1) =  10.218 (0.001)
Lagrange multiplier test of lambda=0:    chi2(1) =   6.804 (0.009)

Acceptable range for lambda: -1.536 < lambda < 1.000
```

上表显示，误差项的空间自回归系数(记为 lamda)的估计值为 0.036，且在 1% 的水平上显著。对于 lamda 的似然比检验与拉格朗日检验也得到同样的结果。可以看出，SEM 模型对于自变量回归系数的估计结果与 OLS 十分接近。这是因为，如果 SEM 模型正确，则 OLS 也是一致估计量，只不过 MLE 估计量更有效。

29.7 一般的空间计量模型

以上两节分别考虑了空间滞后与空间误差模型，这两种效应也可能同时发生。更为一般的空间计量模型将空间自回归模型(SAR)与空间误差模型(SEM)结合起来。考虑以下模型：

$$\boldsymbol{y} = \lambda \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{29.45}$$

其中，扰动项 $\boldsymbol{u}$ 的生成过程为

$$\boldsymbol{u} = \rho \boldsymbol{M}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \tag{29.46}$$

其中，$\boldsymbol{W}$ 与 $\boldsymbol{M}$ 分别为被解释变量 $\boldsymbol{y}$ 与扰动项 $\boldsymbol{u}$ 的空间权重矩阵，二者可以相等。由方程(29.45)与(29.46)构成的模型称为“带空间自回归误差项的空间自回归模型”(Spatial Autoregressive Model with Spatial Autoregressive Disturbances，简记 SARAR)。SAR 模型，SEM 模型以及 SARAR 模型，也统称为“Cliff - Ord 模型”(Cliff and Ord，1973，1981；Ord，1975)。

显然，SAR 模型与 SEM 模型都是 SARAR 模型的特例，分别对应于 $\rho = 0$ 与 $\lambda = 0$ 的情形。对于 SARAR 模型，也可进行 MLE 估计。可以证明，样本的对数似然函数为

$$\ln L(\boldsymbol{y}\mid\lambda,\rho,\sigma^2,\boldsymbol{\beta}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 + \ln(\text{abs}\mid\boldsymbol{A}\mid) + \ln(\text{abs}\mid\boldsymbol{B}\mid) - \frac{1}{2\sigma^2}(\boldsymbol{Ay}-\boldsymbol{X\beta})'\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{Ay}-\boldsymbol{X\beta}) \tag{29.47}$$

在最大化上式时，仍可分两步进行，并进行迭代，与SEM模型的情形类似。虽然Cliff-Ord模型常用MLE进行估计，但MLE估计量存在三个缺陷。首先，MLE估计量可能计算不方便，特别当空间权重矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{M}$ 的维度较高时。其次，空间计量模型MLE估计量的大样本理论尚不健全。在独立同分布的假设下（但不一定服从正态分布），Lee(2004)导出了SAR模型的QMLE渐近分布，但未涉及SEM模型或更一般的SARAR模型①。再次，如果扰动项 $\boldsymbol{\varepsilon}$ 不服从独立同分布的正态分布，则QMLE估计量可能不一致。比如，Arraiz et al(2010)通过蒙特卡罗模拟证明，在异方差的情况下，QMLE估计量是不一致的。

为此，Kelejian and Prucha(1999,1998,2004,2010)提出利用工具变量，通过GMM来估计SARAR模型，称为"广义空间二段最小二乘法"(Generalized Spatial Two-stage Least Square，简记GS2SLS)。Arraiz et al(2010)的模拟结果表明，此IV估计量在异方差情况下也是一致的；故更为稳健。而且，GS2SLS的计算更为方便，并不会因为样本容量太大而产生困难。

GS2SLS所使用的工具变量为 $\{\boldsymbol{X},\boldsymbol{WX},\cdots,\boldsymbol{W}^q\boldsymbol{X},\boldsymbol{MX},\boldsymbol{MWX}\cdots,\boldsymbol{MW}^q\boldsymbol{X}\}$ 中线性独立的列向量，记此工具变量矩阵(matrix of instrumental variables)为 $\boldsymbol{H}$。蒙特卡罗模拟显示，在很多情况下取 $q=2$ 即可。显然，由于 $\boldsymbol{X}$ 与 $\boldsymbol{y}$ 相关，而且 $\boldsymbol{X}$ 为外生变量；故以上 $\boldsymbol{X}$ 的线性函数为有效工具变量。GS2SLS法通过以下四个步骤来实现。

第一步，使用工具变量矩阵 $\boldsymbol{H}$ 对方程(29.45)进行2SLS估计。所得估计量 $(\tilde{\lambda},\tilde{\boldsymbol{\beta}})$ 虽然是一致的，但并非最有效，因为它没有考虑扰动项 $\boldsymbol{u}$ 存在空间自回归。记所得残差为 $\tilde{\boldsymbol{u}}$。

第二步，将第一步的2SLS残差 $\tilde{\boldsymbol{u}}$ 作为 $\boldsymbol{u}$ 的估计量代入方程(29.46)，然后进行GMM估计，得到估计量 $\tilde{\rho}$。

第三步，使用 $\tilde{\rho}$ 对方程(29.45)进行"空间Cochrane-Orcutt变换"(spatial Cochrane-Orcutt transformation)，以去掉扰动项的空间自相关。对于时间序列 $\{y_t\}$，如果其扰动项存在一阶自回归，比如 $u_t=\rho u_{t-1}+\varepsilon_t$，则可通过做准差分(即Cochrane-Orcutt变换)$(y_t-\rho y_{t-1})$来消除此自相关。对于空间序列，也可以做类似的变换。将方程(29.45)两边同时左乘 $(\boldsymbol{I}-\rho\boldsymbol{M})$ 可得：

$$(\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{y} = \lambda(\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{Wy} + (\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{X\beta} + \underbrace{(\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{u}}_{=\boldsymbol{\varepsilon}} \tag{29.48}$$

其中，根据方程(29.46)可知，新扰动项 $\boldsymbol{\varepsilon}=(\boldsymbol{I}-\rho\boldsymbol{M})\boldsymbol{u}$ 不再存在空间自相关。将第二步估计量 $\tilde{\rho}$ 代入方程(29.48)，再次使用工具变量 $\boldsymbol{H}$ 对变换后的方程(29.48)进行2SLS估计，记所得估计量为 $(\hat{\lambda},\hat{\beta})$，其相应残差为 $\hat{\boldsymbol{u}}$。

第四步，将第三步的GS2SLS残差 $\hat{\boldsymbol{u}}$ 作为 $\boldsymbol{u}$ 的估计量代入方程(29.46)，然后进行GMM估

① 空间计量模型的大样本理论更为复杂。主要困难在于，当样本容量 $n\to\infty$ 时，随着横截面单位增多，空间权重矩阵 $\boldsymbol{W}_n$ 也必然随之改变(此处使用下标 n 强调 $\boldsymbol{W}_n$ 对 n 的依赖性)。为此，必须对 $\boldsymbol{W}_n$ 的渐近行为进行限制。特别地，$n\to\infty$ 可通过两种方式来实现，第一种方式通过增加空间规模来增加区域个数，称为"increasing-domain asymptotic"；而第二种方式在给定空间内通过不断细分区域来增加区域个数，称为"infill asymptotic"或"fixed-domain asymptotic"。Lee(2004)表明，在"infill asymptotic"的情形下，SAR模型的MLE估计量有可能不一致。

计,得到估计量 $\hat{\rho}$。

根据以上四个步骤所得估计量($\hat{\lambda},\hat{\rho},\hat{\boldsymbol{\beta}}$)即为 GS2SLS 估计量。由于该估计量将 2SLS 用于空间数据,而且使用了广义最小二乘法(Cochrane - Orcutt 变换是 GLS 的特例),故称为"广义空间二段最小二乘法"。GS2SLS 法的优点是计算方便且结果稳健(在异方差情况下也成立),但缺点是必须有外生变量 $\boldsymbol{X}$,故不适用于没有外生变量的纯 SAR 模型(参见方程(29.20))。

SARAR 模型可通过非官方 Stata 命令 spreg 来实现①,其下载安装方法为(或通过命令"findit sppack"查找并下载)

```
. net install sppack.pkg
```

此命令将下载一套有关空间计量的命令,由以下三个命令所组成:命令 spmat 用于定义与管理空间权重矩阵,命令 spreg 用于估计 SARAR 模型,命令 spivreg 用于估计含内生解释变量的 SARAR 模型(参见下节)。

继续以数据集 columbusdata. dta 与 columbusswm. dta 为例。首先用命令 spmat 定义空间权重矩阵。

```
. use columbusswm.dta, clear
```

此时,在变量窗口将看到 49 个变量,分别为 a1—a49(即为 49 × 49 权重矩阵的 49 列)。

```
. spmat dta w1 a1 - a49
```

其中,子命令"dta"表示从 Stata 格式的数据集(扩展名为 dta)来生成空间权重矩阵。上述命令表示根据内存中的变量 a1 - a49 来生成空间权重矩阵 w1。更多有关命令 spmat 的用法说明,参见"help spmat"。

```
. use columbusdata.dta, clear
. spreg ml crime hoval income, id(id) dlmat(w1) elmat(w1) nolog
```

其中,"ml"表示使用 MLE 估计 SARAR 模型。必选项"id(id)"表示用来确定横截面单位的变量为 id。选择项"dlmat(w1)"表示以 w1 作为空间自回归方程(29.45)中的空间权重矩阵(即上文的 $\boldsymbol{W}$),而选择项"elmat(w1)"则表示以 w1 作为空间误差方程(29.46)中的空间权重矩阵(即上文的 $\boldsymbol{M}$)。

```
Spatial autoregressive model                    Number of obs   =        49
(Maximum likelihood estimates)                  Wald chi2(2)    =   37.3387
                                                Prob > chi2     =    0.0000
```

crime	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
crime						
hoval	-.2555749	.0947579	-2.70	0.007	-.4412968	-.0698529
income	-1.175428	.3037636	-3.87	0.000	-1.770794	-.5800625
_cons	52.74999	7.156856	7.37	0.000	38.72281	66.77716
lambda						
_cons	.0507629	.0204347	2.48	0.013	.0107117	.0908142
rho						
_cons	.0061692	.0762303	0.08	0.935	-.1432395	.1555779
sigma2						
_cons	93.32433	18.87086	4.95	0.000	56.33813	130.3105

① 此命令有希望日后成为 Stata 的官方命令,因为其作者之一 David Drukker 为 Stata 公司(Stata Corp.)的计量经济学主管(Director of Econometrics)。

上表显示，空间自回归系数 $\lambda = 0.051$，且在 5% 水平上显著。而空间误差系数 ρ 则很不显著，其 p 值高达 0.935。下面以 GS2SLS 法进行估计。

```
. spreg gs2sls crime hoval income, id(id) dlmat(w1) elmat(w1) het nolog
```

其中，子命令“gs2sls”表示使用 GS2SLS 法；选择项“het”表示使用异方差稳健的标准误（默认标准误公式假设同方差）。

```
Spatial autoregressive model                    Number of obs     =        49
(GS2SLS estimates)
```

crime	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
crime						
hoval	-.2226818	.1531299	-1.45	0.146	-.5228109	.0774472
income	-1.153734	.4963804	-2.32	0.020	-2.126622	-.1808466
_cons	48.29302	7.99936	6.04	0.000	32.61457	63.97148
lambda						
_cons	.0652404	.016342	3.99	0.000	.0332105	.0972702
rho						
_cons	-.0303691	.0759236	-0.40	0.689	-.1791768	.1184385

上表显示，GS2SLS 与 MLE 的估计结果十分类似。由于空间误差系数 ρ 很不显著，下面估计不含空间误差的 SAR 模型。

```
. spreg gs2sls crime hoval income, id(id) dlmat(w1) het nolog
```

```
Spatial autoregressive model                    Number of obs     =        49
(GS2SLS estimates)
```

crime	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
crime						
hoval	-.2490946	.1596383	-1.56	0.119	-.5619799	.0637907
income	-1.104052	.5017371	-2.20	0.028	-2.087438	-.1206653
_cons	49.66803	8.125303	6.11	0.000	33.74273	65.59333
lambda						
_cons	.0607515	.0170531	3.56	0.000	.0273281	.0941749

上表显示，SAR 模型的估计结果类似于 SARAR 模型。

29.8 含内生解释变量的 SARAR 模型

使用 GS2SLS 法估计 SARAR 模型的一个优点是，它可以很容易地推广到包含内生解释变量的情形：

$$\boldsymbol{y} = \lambda \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u} \tag{29.49}$$

其中，$\boldsymbol{Z}$ 由内生解释变量所组成，而扰动项 $\boldsymbol{u}$ 的生成过程依然为

$$\boldsymbol{u}=\rho\boldsymbol{M}\boldsymbol{u}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n) \tag{29.50}$$

假设在 $\boldsymbol{X}$ 之外，尚有不在方程中的外生变量 $\boldsymbol{X}_e$。定义 $\boldsymbol{X}_f\equiv(\boldsymbol{X},\boldsymbol{X}_e)$；如果没有不在方程中的外生变量 $\boldsymbol{X}_e$，则 $\boldsymbol{X}_f\equiv\boldsymbol{X}$。定义工具变量矩阵 $\boldsymbol{H}$ 为 $\{\boldsymbol{X}_f,\boldsymbol{W}\boldsymbol{X}_f,\cdots,\boldsymbol{W}^q\boldsymbol{X}_f,\boldsymbol{M}\boldsymbol{X}_f,\boldsymbol{M}\boldsymbol{W}\boldsymbol{X}_f\cdots,\boldsymbol{M}\boldsymbol{W}^q\boldsymbol{X}_f\}$ 中线性独立的列向量，即可进行 GS2SLS 估计。

延续上节的例子。纯粹为了演示的目的，假设 income 为内生变量，而 hoval 为其工具变量，然后使用命令 spivreg 进行估计。

```
. spivreg crime(income = hoval),id(id) dlmat(w1) elmat(w1) het nolog
```

其中，选择项“het”表示使用异方差稳健的标准误。

```
Spatial autoregressive model                    Number of obs     =        49
(GS2SLS estimates)
```

crime	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
crime						
income	-2.654373	.5777431	-4.59	0.000	-3.786729	-1.522018
_cons	66.64373	10.06046	6.62	0.000	46.92558	86.36187
lambda						
_cons	.0370156	.0152704	2.42	0.015	.0070862	.0669451
rho						
_cons	.013882	.0437001	0.32	0.751	-.0717686	.0995325

```
Instrumented:  income
Instruments:   hoval
```

上表显示，空间自回归系数在5%水平上显著，而空间误差系数很不显著。

29.9 空间面板模型

横截面数据的空间自回归模型与空间误差模型可以推广到面板数据中。首先考察面板的空间自回归模型：

$$y_{it}=\rho\boldsymbol{w}_i'\boldsymbol{y}_t+\boldsymbol{x}_{it}'\boldsymbol{\beta}+u_i+\varepsilon_{it}\quad(i=1,\cdots,n;t=1,\cdots,T) \tag{29.51}$$

其中，$\boldsymbol{w}_i'$为空间权重矩阵 $\boldsymbol{W}$ 的第 i 行，$\boldsymbol{w}_i'\boldsymbol{y}_t=\sum_{j=1}^{n}w_{ij}y_{jt}$，$w_{ij}$为空间权重矩阵 $\boldsymbol{W}$ 的(i,j)元素；而 u_i 为区域 i 的个体效应。如果不考虑空间滞后项$\rho\boldsymbol{w}_i'\boldsymbol{y}_t$，则方程(29.51)为标准的静态面板模型。进一步，如果 u_i 与$\boldsymbol{x}_{it}$相关，则为固定效应模型；反之，则为随机效应模型。

对于固定效应的空间自回归模型，可以先做组内离差变换(参见第 15 章)，去掉个体效应 u_i；然后再使用类似于横截面空间自回归模型的 MLE 估计。对于随机效应的空间自回归模型，则可先做广义离差变换(参见第 15 章)，然后再进行 MLE 估计。具体推导参见 Elhorst(2010)与 Lee and Yu(2010)。在决定使用固定效应还是随机效应模型时，可进行通常的豪斯曼检验。

对于空间面板模型的 MLE 估计，可通过非官方命令 xsmle 来实现。此命令可估计很一般的空间面板模型：

$$\begin{cases}y_{it}=\tau y_{i,t-1}+\rho\boldsymbol{w}_i'\boldsymbol{y}_t+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{d}_i'\boldsymbol{X}_t\boldsymbol{\delta}+u_i+\gamma_t+\varepsilon_{it}\\ \varepsilon_{it}=\lambda\boldsymbol{m}_i'\boldsymbol{\varepsilon}_t+v_{it}\end{cases} \tag{29.52}$$

其中，$y_{i,t-1}$为被解释变量y_{it}的一阶滞后（即动态面板；如果不是动态面板，可令$\tau=0$）；$\boldsymbol{d}_i'\boldsymbol{X}_t\boldsymbol{\delta}$表示解释变量的空间滞后，$\boldsymbol{d}_i'$为相应空间权重矩阵$\boldsymbol{D}$的第$i$行；$\gamma_t$为时间效应；而$\boldsymbol{m}_i'$为扰动项空间权重矩阵$\boldsymbol{M}$的第$i$行。

由于模型（29.52）太一般，故通常考虑以下特殊情形（每种情形均以命令 xsmle 所使用的术语来命名）。

（1）如果$\lambda=0$，则为“空间杜宾模型”（Spatial Durbin Model，简记 SDM）。

（2）如果$\lambda=0$且$\boldsymbol{\delta}=\boldsymbol{0}$，则为“空间自回归模型”（Spatial Autoregression Model，简记 SAR）。

（3）如果$\tau=0$且$\boldsymbol{\delta}=\boldsymbol{0}$，则为“空间自相关模型”（Spatial Autocorrelation Model，简记 SAC）；即上文的 SARAR 模型。

（4）如果$\tau=\rho=0$且$\boldsymbol{\delta}=\boldsymbol{0}$，则为“空间误差模型”（Spatial Error Model，简记 SEM）。

命令 xsmle 的句式为

```
xsmle y x1 x2 x3, wmat(name) emat(name) dmat(name) durbin(varlist)
model(sdm) model(sac) model(sem) re fe robust dlag type(ind) type(time)
type(both) noeffects
```

其中，选择项“wmat(name)”表示空间权重矩阵$\boldsymbol{W}$，选择项“emat(name)”表示扰动项空间权重矩阵$\boldsymbol{M}$，选择项“dmat(name)”表示解释变量的空间权重矩阵$\boldsymbol{D}$（默认使用选择项“wmat(name)”所指定的空间权重矩阵）；选择项“durbin(varlist)”用于指定空间滞后的解释变量（默认为所有解释变量）；选择项“model(sar)”表示估计 SAR 模型（默认选择），选择项“model(sdm)”表示估计 SDM 模型，选择项“model(sac)”表示估计 SAC 模型，选择项“model(sem)”表示估计 SEM 模型；选择项“re”表示估计随机效应模型（默认选择），选择项“fe”表示估计固定效应模型；选择项“robust”表示使用聚类稳健的标准误；选择项“dlag”表示加入被解释变量的一阶滞后作为解释变量（即动态面板），默认为静态面板（即$\tau=0$）。选择项“type(ind)”表示只有个体效应u_i（默认选择），选择项“type(time)”表示只有时间效应γ_t，而选择项“type(both)”表示既有个体效应，也有时间效应。选择项“noeffects”表示不显示直接效应、间接效应与总效应。

下面以命令 xsmle 附带的数据集 product.dta 以及空间权重矩阵 usaww.spmat 为例，演示 SDM 模型的估计（SAR 模型是 SDM 模型的特例）。该数据集最初为 Munnell（1990）所使用，包含美国 48 个州 1970－1974 年的如下变量：gsp（州产出，gross state product），pcap（公共资本，public capital），pc（民营资本，private capital），emp（就业量，employment），unemp（失业率，unemployment rate）。

首先下载非官方命令 xsmle[①]：

```
. net install xsmle, all from(http://www.econometrics.it/stata)
```

运行此命令后，数据集 product.dta 以及空间权重矩阵 usaww.spmat 将被存入当前工作路径（path of working directory）。

```
. use product.dta,clear
```

由于产出、资本与就业量存在指数增长趋势，故首先将这些变量取对数。

```
. gen lngsp = log(gsp)
. gen lnpcap = log(pcap)
. gen lnpc = log(pc)
```

① 如果通过命令“ssc install xsmle”下载此命令，则无法下载作为例子的两个附带文件 product.dta 与 usaww.spmat。

```
. gen lnemp = log(emp)
. spmat use usaww using usaww.spmat
```

此命令表示根据文件“usaww.spmat”来生成空间权重矩阵“usaww”，参见“help spmat”。首先使用聚类稳健的标准误估计随机效应的 SDM 模型。

```
. xsmle lngsp lnpcap lnpc lnemp unemp,wmat(usaww) model(sdm) robust nolog
```

SDM with random-effects　　Number of obs = 816

Group variable: state　　Number of groups = 48
Time variable: year　　Panel length = 17

R-sq: within = 0.9460
between = 0.9914
overall = 0.9904

Log-pseudolikelihood = 1504.9528
(Std. Err. adjusted for 48 clusters in state)

lngsp	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Main						
lnpcap	.0432197	.0456273	0.95	0.344	-.0462082	.1326476
lnpc	.2358733	.0818952	2.88	0.004	.0753617	.3963848
lnemp	.7460606	.0796743	9.36	0.000	.5899019	.9022193
unemp	-.0012839	.0029046	-0.44	0.658	-.0069768	.004409
_cons	1.160517	.480234	2.42	0.016	.2192755	2.101758
Wx						
lnpcap	-.0937203	.0885386	-1.06	0.290	-.2672529	.0798123
lnpc	-.0215075	.0728551	-0.30	0.768	-.1643009	.121286
lnemp	-.3895897	.0946901	-4.11	0.000	-.5751789	-.2040006
unemp	-.0036868	.0038042	-0.97	0.332	-.0111428	.0037692
Spatial						
rho	.4873069	.071246	6.84	0.000	.3476672	.6269466
Variance						
lgt_theta	-2.345213	.2519923	-9.31	0.000	-2.839109	-1.851317
sigma_e	.0010299	.0002446	4.21	0.000	.0005504	.0015094
Direct						
lnpcap	.0309365	.0429702	0.72	0.472	-.0532834	.1151565
lnpc	.2540086	.0860114	2.95	0.003	.0854294	.4225877
lnemp	.7422207	.0789719	9.40	0.000	.5874386	.8970029
unemp	-.0018148	.0025186	-0.72	0.471	-.0067512	.0031215
Indirect						
lnpcap	-.1361936	.1765661	-0.77	0.441	-.4822567	.2098696
lnpc	.1419852	.1107685	1.28	0.200	-.0751171	.3590875
lnemp	-.0226692	.1240987	-0.18	0.855	-.2658983	.2205599
unemp	-.0072797	.0057566	-1.26	0.206	-.0185624	.0040029
Total						
lnpcap	-.105257	.1988103	-0.53	0.597	-.4949181	.2844041
lnpc	.3959938	.1065194	3.72	0.000	.1872195	.604768
lnemp	.7195515	.1668882	4.31	0.000	.3924567	1.046646
unemp	-.0090946	.0061702	-1.47	0.140	-.0211879	.0029988

从上表可知，空间自回归系数 ρ(rho)在 1% 水平上显著为正。但变量 lnpcap，lnpc 与 unemp 的空间滞后项并不显著，故使用选择项“durbin(varlist)”将这三个变量去掉。上表底部还显示了各变量的直接、间接与总效应(参见本章第 4 节)；为节省空间，下面使用选择项“noeffects”使其不再显示。

```
. xsmle lngsp lnpcap lnpc lnemp unemp, wmat(usaww) model(sdm) durbin
(lnemp) robust nolog noeffects
```

```
SDM with random-effects                         Number of obs =       816

Group variable: state                           Number of groups =     48
Time variable: year                                 Panel length =     17

R-sq:     within  = 0.9436
          between = 0.9905
          overall = 0.9894

Log-pseudolikelihood =  1493.7191
                                (Std. Err. adjusted for 48 clusters in state)
```

lngsp	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Main						
lnpcap	.0269178	.047324	0.57	0.569	-.0658355	.1196711
lnpc	.2077165	.0691193	3.01	0.003	.0722452	.3431877
lnemp	.7700826	.0818839	9.40	0.000	.6095931	.9305721
unemp	-.0047394	.0027155	-1.75	0.081	-.0100617	.0005829
_cons	.8293648	.2337675	3.55	0.000	.3711889	1.287541
Wx						
lnemp	-.4698969	.076081	-6.18	0.000	-.6190129	-.3207809
Spatial						
rho	.4914654	.0798065	6.16	0.000	.3350474	.6478833
Variance						
lgt_theta	-2.438141	.2559757	-9.52	0.000	-2.939844	-1.936437
sigma_e	.0010468	.0002513	4.16	0.000	.0005542	.0015394

下面进行固定效应的估计。

```
. xsmle lngsp lnpcap lnpc lnemp unemp,wmat(usaww) model(sdm) durbin
(lnemp) robust nolog noeffects fe
```

SDM with spatial fixed-effects　　Number of obs = 816

Group variable: state　　Number of groups = 48
Time variable: year　　Panel length = 17

R-sq: within = 0.9457
between = 0.9856
overall = 0.9845

Mean of fixed-effects = 1.0970

Log-pseudolikelihood = 1650.1738

(Std. Err. adjusted for 48 clusters in state)

lngsp	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Main						
lnpcap	-.024499	.0475762	-0.51	0.607	-.1177468	.0687487
lnpc	.1775734	.0797904	2.23	0.026	.0211871	.3339597
lnemp	.7326918	.0940438	7.79	0.000	.5483694	.9170143
unemp	-.0037328	.0027181	-1.37	0.170	-.0090602	.0015946
Wx						
lnemp	-.3958431	.0931682	-4.25	0.000	-.5784493	-.2132368
Spatial						
rho	.5184863	.0744086	6.97	0.000	.3726481	.6643245
Variance						
sigma2_e	.0009515	.0002346	4.06	0.000	.0004917	.0014113

究竟应使用随机效应还是固定效应模型，须进行豪斯曼检验。为此，需要存储未使用稳健标准误的随机效应与固定效应估计结果。

```
. qui xsmle lngsp lnpcap lnpc lnemp unemp, wmat(usaww) model(sdm) durbin(lnemp) nolog noeffects
. est sto re
. qui xsmle lngsp lnpcap lnpc lnemp unemp, wmat(usaww) model(sdm) durbin(lnemp) nolog noeffects fe
. est sto fe
. hausman fe re
```

	Coefficients (b) fe	(B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
lnpcap	-.024499	.0269178	-.0514168	.0073629
lnpc	.1775734	.2077165	-.0301431	.0077758
lnemp	.7326918	.7700826	-.0373908	.0139211
unemp	-.0037328	-.0047394	.0010066	.

b = consistent under Ho and Ha; obtained from xsmle
B = inconsistent under Ha, efficient under Ho; obtained from xsmle

Test: Ho: difference in coefficients not systematic

chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= -55.17 chi2<0 ==> model fitted on these data fails to meet the asymptotic assumptions of the Hausman test; see suest for a generalized test

上表显示,豪斯曼统计量为负数,故可以接受随机效应的原假设。另外,对空间计量模型的估计,也可以在 Matlab[①] 或 Gauss[②] 中进行。

29.10 空间计量方法的局限性

上文提及,空间计量模型的主要估计方法之一为 MLE,但此 MLE 的大样本理论尚待完善。空间计量方法的另一局限是,它需要研究者主观设定一个非随机的空间权重矩阵(而非根据数据来估计此矩阵),而此空间权重矩阵很可能无法完全反映不同地区之间复杂的相互关系。

处理区域间互动关系的另一方法是直接将地区间的距离作为变量引入回归模型。比如,在国际贸易领域,常使用以下模型来考察两国间贸易量的决定因素:

$$Export_{ij} = \frac{\alpha Y_i^{\beta} Y_j^{\gamma}}{D_{ij}^{\delta}} \tag{29.53}$$

其中,$Export_{ij}$表示 i 国对 j 国的出口额,Y_i,Y_j 分别为 i 国与 j 国的 GDP,而 D_{ij}为 i 国与 j 国之间的地理距离(通常以首都间的距离来衡量)。由于公式(29.53)很像牛顿的万有引力公式,故称为"引力模型"(gravity model),最早由丁伯根(Jan Tinbergen)于 1962 年提出。对方程(29.53)两边取对数,并加上误差项(表示其他遗漏变量与随机扰动的影响):

$$\ln Export_{ij} = \ln \alpha + \beta \ln Y_i + \gamma \ln Y_j - \delta \ln D_{ij} + \varepsilon \tag{29.54}$$

在对引力模型的拓展研究中,经济学家又加入了影响贸易的其他因素,比如地理因素(如岛国、高山)、关税与非关税壁垒、边界效应、文化、宗教、语言、甚至政治制度(比如民主程度),参见余淼杰(2013)。

① 参见 http://www.spatial-econometrics.com/.

② 参见 http://web.pdx.edu/~crkl/WISE/SEAUG.htm.

第30章 久期分析

在实证研究中，有时被解释变量为某种活动持续的时间（duration，spell，time to event），比如病人存活的时间，灯泡报废的时间，失业持续的时间，婚姻持续的时间，罢工持续的时间，战争持续的时间，王朝的寿命，刑满释放犯再次犯罪的时间等。这类数据称为“久期数据”（duration data），其相应的分析方法则称为“久期分析”（duration analysis）。

由于久期分析考察个体从某一状态转换到另一状态所花费的时间，故也称为“转换分析”（transition analysis）或“事件历史分析”（event history analysis）。另外，不同学科对于久期分析的称呼也不同。比如，生物统计领域称其为“生存分析”（survival analysis），运筹学领域称为“报废时间分析”（failure time analysis），人口学领域称为“生命表分析”（life table analysis），而保险领域称为“风险分析”（hazard analysis）。久期分析的许多术语来自生物统计领域。

30.1 久期数据的处理方法

处理久期数据需要特别的方法吗？久期数据通常为横截面数据。假设数据为$\{T_i, \boldsymbol{x}_i\}_{i=1}^n$，其中$T_i$为被解释变量，即个体$i$的持续时间（寿命）；而$\boldsymbol{x}_i$为解释变量。一个自然的想法是用OLS估计以下模型：

$$T_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{30.1}$$

然而，由于持续时间$T_i \geqslant 0$，而由上式得到的预测值$\hat{T}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$有可能为负数，故这是一个不现实的模型。较为现实的建模方法将$\ln T_i$作为被解释变量（假设$T_i > 0$）：

$$\ln T_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{30.2}$$

此对数线性模型的最大问题是，在我们抽样获得观测数据时，通常知道个体已经存活了一段时间。而方程（30.2）却总是站在$T_i = 0$（即病情刚发作或确诊）的角度，无法纳入“个体已经存活了一段时间”这一信息。事实上，我们常常更为关心给定个体已存活了一段时间的条件下，个体在下个时刻死亡的概率，即下文的风险函数。比如，我们关心已经失业三个月的失业者明天找到工作的概率。另外，如果久期数据存在右归并（参见30.3节），或者随时间而变的解释变量$\boldsymbol{x}_{it}$，都很难通过OLS来处理。为此，久期分析常使用基于风险函数的一套特殊方法，形成自成体系的研究领域。

30.2 风险函数

记个体在某种状态中持续的时间（spell）或寿命为$T \geqslant 0$，其一个特定取值记为t。显然，T为随机变量。假设T为连续型随机变量，并记其概率密度函数与累积分布函数分别为$f(t)$与

$F(t)$,其中 $F(t)$ 也被称为“失效函数”(failure function)。考虑“病人”存活期超过 t 的概率,称为“生存函数”(survivor function),其定义为:

$$S(t) \equiv \mathrm{P}(T > t) = 1 - F(t), \quad t \geqslant 0 \tag{30.3}$$

由上式可知,生存函数本质上相当于累积分布函数的“反函数”(reverse cumulative distribution function)。由于累积分布函数 $F(t)$ 单调递增,故生存函数 $S(t)$ 单调递减。假设病人已存活到时刻 t,则病人在 $[t, t+\Delta t)$ 期间($\Delta t > 0$)死亡的概率为:

$$\mathrm{P}(t \leqslant T < t + \Delta t \mid T \geqslant t) = \frac{\mathrm{P}(t \leqslant T < t + \Delta t)}{\mathrm{P}(T \geqslant t)} = \frac{F(t + \Delta t) - F(t)}{S(t)} \tag{30.4}$$

进一步,定义“风险率”(hazard rate)或“风险函数”(hazard function)为病人在时刻 t 的瞬间死亡率:

$$\begin{aligned} \lambda(t) &\equiv \lim_{\Delta t \to 0^+} \frac{\mathrm{P}(t \leqslant T < t + \Delta t \mid T \geqslant t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} = \frac{f(t)}{S(t)} \end{aligned} \tag{30.5}$$

由以上定义与表达式可知,风险函数 $\lambda(t)$ 本质上是在给定存活至时刻 t 条件下的条件密度函数,故也称为“条件死亡率”(conditional failure rate);而 $f(t)$ 为无条件密度函数。如果 $f(t) = \phi(t)$(标准正态的密度函数),则 $\lambda(t)$ 就是反米尔斯比率(IMR,参见第 14 章图 14.2)。风险率的可能取值介于 0(无死亡风险)与∞(必死无疑)之间。在久期分析中,风险函数 $\lambda(t)$ 与生存函数 $S(t)$ 比密度函数 $f(t)$ 与累积分布函数 $F(t)$ 更为方便与常用。

也可以从风险函数 $\lambda(t)$ 出发,反推出生存函数 $S(t)$、累积分布函数 $F(t)$ 以及密度函数 $f(t)$。首先,从方程(30.5)可知,

$$\lambda(t) = -\frac{\mathrm{d}\ln S(t)}{dt} \tag{30.6}$$

由上式可知 $\mathrm{d}\ln S(t) = -\lambda(t)\mathrm{d}t$,然后两边从 0 到 t 作定积分可得:

$$\ln S(t) = -\int_0^t \lambda(u)\mathrm{d}u \tag{30.7}$$

其中,u 为积分变量;$S(0) = 1$(在初始时刻,所有个体都活着)。因此,

$$F(t) = 1 - S(t) = 1 - \exp\left[-\int_0^t \lambda(u)\mathrm{d}u\right] \tag{30.8}$$

$$S(t) = \exp\left[-\int_0^t \lambda(u)\mathrm{d}u\right] \tag{30.9}$$

对方程(30.8)两边求导可得,

$$f(t) = \lambda(t)\exp\left[-\int_0^t \lambda(u)\mathrm{d}u\right] \tag{30.10}$$

为了度量截止时刻 t 的累积总风险,定义“累积风险函数”(cumulative hazard 或 integrated hazard)为:

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = -\ln S(t) \tag{30.11}$$

上式中的第二个等号用到了方程(30.7)。累积风险函数的好处在于,它比风险函数可以更准确地估计。如果知道累积风险函数,则很容易计算生存函数:

$$S(t)=\exp[-\Lambda(t)] \tag{30.12}$$

从方程(30.12)可以看出累积风险函数 $\Lambda(t)$ 的经济含义,它完全决定了至时间 t 为止的存活率或概率。

例 假设 T 服从指数分布(exponential distribution),$f(t)=\lambda e^{-\lambda t}$($\lambda>0$),则 $F(t)=\int_0^t \lambda e^{-\lambda s}ds=\int_0^t e^{-\lambda s}d(\lambda s)=-e^{-\lambda s}\Big|_0^t=1-e^{-\lambda t}$,$S(t)=e^{-\lambda t}$,$\lambda(t)=\frac{f(t)}{S(t)}=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda$,$\Lambda(t)=\lambda t$。因此,指数分布的风险函数为常数,这个性质被称为"无记忆性"(memoryless)。它意味着,瞬间死亡的概率并不依赖于已存活了多久。进一步,可以证明,个体在$(0,t_2)$区间死亡的概率等于在已知个体存活至时刻 t_1 的情况下,其在(t_1,t_1+t_2)区间死亡的概率(参见习题)。另一方面,如果风险函数为常数 λ,则其对应的密度函数为 $f(t)=\lambda\exp\left[-\int_0^t \lambda du\right]=\lambda e^{-\lambda t}$,一定服从指数分布。根据概率统计知识可知,指数分布的期望为 $E(T)=\frac{1}{\lambda}$(风险率的倒数),方差为 $\mathrm{Var}(T)=\frac{1}{\lambda^2}$。指数分布广泛应用于研究电子元器件的寿命问题。

由于指数分布只有一个参数,故如果知道期望,则方差也确定(方差为期望的平方),缺乏灵活性。另外,指数分布的无记忆性有时也不现实。它意味着,一个 20 岁的青年与一个 80 岁的老年不仅瞬间死亡率一样,而且在未来 10 年内死亡的概率也相同。显然,无记忆性的假设太强了。为此,常将指数分布拓展至拥有两个参数的威布尔分布(Weibull distribution)。

定义 如果随机变量 T 的累积分布函数为 $F(t)=1-\exp(-\gamma t^p)$,其中 $\gamma>0$,$p>0$,则称其服从威布尔分布。

显然,威布尔分布的生存函数为 $S(t)=1-F(t)=\exp(-\gamma t^p)$。威布尔分布的密度函数为 $f(t)=F'(t)=\gamma p t^{p-1}\exp(-\gamma t^p)$。因此,威布尔分布的风险函数为

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{\gamma p t^{p-1}\exp(-\gamma t^p)}{\exp(-\gamma t^p)}=\gamma p t^{p-1} \tag{30.13}$$

如果 $p=1$,则威布尔分布的累积分布函数为 $F(t)=1-\exp(-\gamma t)$,就是指数分布的累积分布函数,故指数函数是威布尔分布的一个特例。如果 $p>1$,则风险函数 $\lambda(t)$ 单调递增,这意味着,活得越久则死亡概率越高(或许由于老年化过程),称为"正向久期依赖"(positive duration dependence);反之,如果 $p<1$,则风险函数 $\lambda(t)$ 单调递减(或许新生儿死亡概率最高);被称为"负向久期依赖"(negative duration dependence)。如果 $p=1$(即指数分布的情形),则风险函数 $\lambda(t)$ 为常数(或许死亡由外在的随机因素所造成)。从方程(30.13)可以看出,参数 p 决定了风险函数 $\lambda(t)$ 的形状,称为"形状参数"(shape parameter);而参数 γ 则决定其规模,称为"规模参数"(scale parameter)。

30.3 久期数据的归并问题

久期数据常常存在"右归并"(right censoring)。一个主要原因是,当研究结束时,有些病人可能尚未死亡;或者有些失业者还未找到工作。此时,我们观测到的个体存活时间为从 0 开始直至某归并时间(censoring time)C^*。我们只知道个体的寿命属于区间(C^*,∞),并不知道其具体取值。导致右归并的原因还包括,个体中途退出研究(比如,研究某药物是否能延长病人的寿

命，而病人中途因强烈副作用而提前退出实验），或研究者与个体失去联系，无法继续跟踪调查。

在存在右归并的情形，记个体 i 的真实寿命为 T_i^*（可能不可观测），而归并时间为 C_i^*。实际观测到的 T_i 或为个体寿命 T_i^*，或为归并时间 C_i^*，取决于二者哪个更小：

$$T_i = \min(T_i^*, C_i^*) \tag{30.14}$$

同时，以虚拟变量 d_i 来记录个体 i 的观测记录是否完整：

$$d_i = \mathbf{1}(T_i^* < C_i^*) \tag{30.15}$$

其中，$\mathbf{1}(\cdot)$ 为示性函数，即如果括弧中表达式为真，则取值为 1；反之，则取值为 0。显然，如果 $d_i = 1$，则有完整记录，无归并（uncensored）；如果 $d_i = 0$，则存在右归并。

有时也会出现“左归并”（left censoring），即只知道个体的寿命属于区间 $(0, C^*)$，而不知道其具体取值。另一种归并情形为“区间归并”（interval censoring），即只知道个体的寿命属于区间 $[C_a^*, C_b^*)$，也不知道其具体取值。区间归并常因为数据的离散性而发生，比如，研究者只知道失业者在某一周找到工作，而不知道其具体日期。

为了保证久期分析的有效性，常假设“独立归并”（independent censoring）或“无信息归并”（noninformative censoring），即归并时间 C_i^* 的分布不包含任何有关个体寿命 T_i^* 分布的信息。此时，可将表示归并的虚拟变量 d_i 视为外生变量，不必在意归并究竟如何发生，也无须为“归并机制”（censoring mechanism）建模。

另外，在久期样本中，每一个体开始活动（比如，开始生病或失业）的日历时间（calendar time）可以不同。通常将“风险开始”（onset of risk）的时间标准化为 0 时刻，以此度量的时间在 Stata 中称为“分析时间”（analysis time）。久期分析的被解释变量 T_i 正是以分析时间来计算的。

30.4 描述性分析

在引入解释变量之前，我们常希望进行一些粗略的描述性分析，比如根据样本数据来估计生存函数、累积风险函数与风险函数，看看它们的大致形状。在进行这类分析时，既不需要对数据的概率分布作先验假设，也不需要估计参数，故为非参数方法（参见第 27 章）。通俗地说，这是一种撇开理论（不假设变量的概率分布），让数据自己说话的方法（let data speak for themselves）。

1. 生存函数

根据定义，生存函数 $S(t)$ 为个体存活时间超过时刻 t 的概率。如果不存在归并（censoring），则显然可以定义 $S(t)$ 的估计量为，样本中存活时间超过时刻 t 的个体数目 r 占样本容量 n 的比例，即 $\frac{r}{n}$。但这个简单方法在存在归并的情况下并不适用。此时，一般使用 Kaplan - Meier 估计量（Kaplan and Meier, 1958，简记 KM），它在独立归并（independent censoring）的情况下依然是 $S(t)$ 的一致估计量。

记 $t_1 < t_2 < \cdots < t_j < \cdots < t_K$ 为样本中观测到的死亡时间。记样本中在区间 $[t_{j-1}, t_j)$ 仍存活而面临危险（at risk）的个体数为 n_j。到了时间 t_j，这些 n_j 个体的命运分为三种，即存活、死亡、归并（只知道其死亡时间大于 t_j，但不再有观测数据）。记在时间 t_j 死亡的人数为 m_j（现实数据一般是离散的，故 m_j 可能大于 1）。给定存活至 t_{j-1}，能进一步活至 t_j 的概率（频率）为 $\frac{n_j - m_j}{n_j}$。一般地，活至 t_j 的无条件概率等于活过之前每一个区间的条件概率之乘积。因此，Kaplan - Meier 估

计量为

$$\hat{S}(t) \equiv \prod_{j|t_j \leqslant t} \left(\frac{n_j - m_j}{n_j} \right) \tag{30.16}$$

故 Kaplan - Meier 估计量也称为“连乘估计量”(product limit estimator)。显然,如果不存在归并,则 $n_{j+1} = n_j - m_j$,方程(30.16)中的连乘可以错项相约,故 $\hat{S}(t) = r/n$。

2. 累积风险函数

根据方程(30.11),累积风险函数 $\Lambda(t) = -\ln S(t)$,将 $\hat{S}(t)$ 的表达式(30.16)代入,即可得到 $\Lambda(t)$ 的估计量。但此估计量的小样本性质不如 Nelson(1972)与 Aalen(1978)所提出的 Nelson - Aalen 估计量(简记 NA):

$$\hat{\Lambda}(t) \equiv \sum_{j|t_j \leqslant t} \left(\frac{m_j}{n_j} \right) \tag{30.17}$$

其中的每一项 $\frac{m_j}{n_j}$ 为局部的风险率,而上式则为局部风险率的加总。

3. 风险函数

从上面的讨论可知,可以用 $\hat{\lambda}_j \equiv m_j/n_j$ 作为风险率的估计量。但此估计量为不光滑的阶梯函数(step function)。另一方法是对累积风险函数求导,但 $\hat{\Lambda}(t)$ 也是阶梯函数,并不处处可导。在实践中,一般先通过核密度方法将阶梯形的累积风险函数光滑化[①],然后再以此生成风险函数。在 Stata 中可由命令“`sts graph,hazard`”来直接完成。

30.5 久期模型的最大似然估计

一般来说,(未被归并的)真实持续时间 T^* 的概率分布还可能依赖于某些解释变量 $\boldsymbol{x}$。为此,记 T^* 的密度函数、累积分布函数、生存函数与风险函数为 $f(t|\boldsymbol{x},\boldsymbol{\theta})$,$F(t|\boldsymbol{x},\boldsymbol{\theta})$,$S(t|\boldsymbol{x},\boldsymbol{\theta})$ 与 $\lambda(t|\boldsymbol{x},\boldsymbol{\theta})$,其中 $\boldsymbol{x}$ 为不随时间而变的解释变量(time-invariant covariates)(以后再考虑随时间而变的解释变量),$\boldsymbol{\theta}$ 为待估计的未知参数。比如,T^* 为失业持续的时间,$\boldsymbol{x}$ 包括失业前的教育水平、工作经验、种族、婚姻状况、子女数目,以及政府发放的失业救济金额等;故 T 的分布依赖于 $\boldsymbol{x}$。

下面使用 MLE 来估计未知参数 $\boldsymbol{\theta}$。假设存在右归并,则其 MLE 估计类似于归并回归的 Tobit 模型。对于未被归并的观测值,其对似然函数的贡献为 $f(t|\boldsymbol{x},\boldsymbol{\theta})$。另一方面,对于右归并的观测值,我们只知道其持续时间超过 t,故其对似然函数的贡献为 $\mathrm{P}(T^* > t) = S(t|\boldsymbol{x},\boldsymbol{\theta})$。综合起来,个体 i 对似然函数的贡献可写为

$$f(t_i|\boldsymbol{x}_i,\boldsymbol{\theta})^{d_i} S(t_i|\boldsymbol{x}_i,\boldsymbol{\theta})^{1-d_i} \tag{30.18}$$

其中,d_i 为右归并虚拟变量,参见(30.15),即 $d_i = 1$ 表示无归并,而 $d_i = 0$ 表示右归并。假设样本为 iid,将似然函数取对数,然后加总可得

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} [\, d_i \ln f(t_i|\boldsymbol{x}_i,\boldsymbol{\theta}) + (1-d_i) \ln S(t_i|\boldsymbol{x}_i,\boldsymbol{\theta}) \,] \tag{30.19}$$

上式中的第一项对应于无归并数据,而第二项对应于右归并数据。由于 $f(t) = \lambda(t)S(t)$,故 $\ln f(t) = \ln\lambda(t) + \ln S(t)$,因此方程(30.19)可写为

---

① 有关核密度方法,参见第 27 章。

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} [d_i \ln \lambda(t_i|\boldsymbol{x}_i,\boldsymbol{\theta}) + \ln S(t_i|\boldsymbol{x}_i,\boldsymbol{\theta})] \tag{30.20}$$

如果直接对风险函数$\lambda(t|\boldsymbol{x},\boldsymbol{\theta})$建模，则表达式(30.20)可能更为方便。根据MLE的一般理论，如果似然函数正确，则此MLE估计量为一致、有效估计量，且服从渐近正态分布。然而，如果似然函数不正确，则此MLE估计量通常是不一致的。一个例外是T^*服从指数分布并且不存在归并，则只要条件期望函数(conditional mean function)正确即可保证MLE估计量的一致性。然而，如果存在归并，即使T^*服从指数分布，MLE估计量也依然不一致。因此，以上参数回归模型(parametric regression model)的最大缺点是缺乏稳健性(lack of robustness)，对概率分布的具体假设比较敏感(Tobit模型也有此缺点)。

如果存在其他形式的归并，则应对似然函数进行相应的调整。如果观测值存在左归并，则只知道持续时间小于t，故其对似然函数的贡献为$P(T^* < t) = F(t|\boldsymbol{x},\boldsymbol{\theta})$。如果观测值存在区间归并，则只知道持续时间属于区间$[C_a^*, C_b^*)$，故其对似然函数的贡献为$P(C_a \leqslant T^* < C_b) = S(C_a|\boldsymbol{x},\boldsymbol{\theta}) - S(C_b|\boldsymbol{x},\boldsymbol{\theta})$。

在经济应用中的久期数据常存在区间归并。比如，失业持续时间经常以周或月来计算，而指数分布、威布尔分布等都是连续型分布。此时，通常假设区间归并的影响甚微，而仍将其作为连续数据来处理。

30.6 比例风险模型

称风险函数$\lambda(t;\boldsymbol{x})$为“比例风险”(Proportional Hazard，简记PH)，如果它可以分解为

$$\lambda(t;\boldsymbol{x}) = \lambda_0(t)h(\boldsymbol{x}) \tag{30.21}$$

其中，$h(\cdot) > 0$，而$\lambda_0(t)$被称为“基准风险”(baseline hazard)，依赖于时间t，但不依赖于$\boldsymbol{x}$。基准风险$\lambda_0(t)$对于总体中的每一个体都相同，而个体的风险函数则依据$h(\boldsymbol{x})$与此基准风险$\lambda_0(t)$成正比，故名“比例风险”。通常，令$h(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\beta}}$，则

$$\lambda(t;\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}} \tag{30.22}$$

其中，$e^{\boldsymbol{x}'\boldsymbol{\beta}}$称为“相对风险”(relative hazard)，故$\boldsymbol{x}'\boldsymbol{\beta}$也被称为“对数相对风险”(log-relative hazard)。根据方程(30.22)，如果所有解释变量都等于0，则风险函数$\lambda(t;x)$就等于基准风险$\lambda_0(t)$。对方程(30.22)两边取对数可得

$$\ln \lambda(t;\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} + \ln \lambda_0(t) \tag{30.23}$$

其中，$\boldsymbol{\beta}$可方便地解释为半弹性，即某x边际地增加Δx，将导致风险函数增加百分之$100\boldsymbol{\beta} \cdot \Delta x$。另一解释是，如果$x$增加一单位，则风险率为$\lambda(t;\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}+\beta}$，正好是原来风险率$\lambda(t;\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$的$e^{\beta}$倍。在Stata中，称$\boldsymbol{\beta}$为“系数”(coefficient)，而称$e^{\beta}$为“风险比率”(Hazard Ratio，简记HR)，即x增加一单位，将使新风险率变为原来风险率的e^{β}倍。在估计模型之后，可以选择让Stata汇报风险比率(默认值)或系数。

在方程(30.22)中，如果令基准风险$\lambda_0(t) = e^{a}$，其中a为待估参数(选择指数形式是为了保证$e^{a} > 0$)，则为“指数回归”(exponential regression)，因为指数分布的风险函数为常数。也可以从指数分布的密度函数出发，即$f(t) = \lambda e^{-\lambda t}(\lambda > 0)$，而令$\lambda = e^{a}e^{\boldsymbol{x}'\boldsymbol{\beta}} = e^{a+\boldsymbol{x}'\boldsymbol{\beta}}$，其中$\boldsymbol{x}$不包含常数项(如果有常数项$e^{\beta_0}$，可与$e^{a}$合并；Stata也不汇报常数项)。将指数分布的风险函数λ与生存函数$S(t) = e^{-\lambda t}$代入方程(30.20)可得指数回归的对数似然函数：

$$\ln L(a,\boldsymbol{\beta}) = \sum_{i=1}^{n} [d_i(a+\boldsymbol{x}_i'\boldsymbol{\beta}) - e^{a+\boldsymbol{x}_i'\boldsymbol{\beta}}t_i] \tag{30.24}$$

类似地，如果在方程(30.22)中，令基准风险 $\lambda_0(t) = pt^{p-1}e^a$，其中 $p>0$，a 为待估参数，则为“威布尔回归”(Weibull regression)，因为威布尔分布的风险函数为 γpt^{p-1}(此处令 $\gamma = e^{a+\boldsymbol{x}'\boldsymbol{\beta}}$)。将威布尔分布的风险函数 γpt^{p-1} 与生存函数 $S(t) = \exp(-\gamma t^p)$ 代入方程(30.20)可得威布尔回归的对数似然函数：

$$\ln L(a,p,\boldsymbol{\beta}) = \sum_{i=1}^{n} \{d_i[a+\boldsymbol{x}_i'\boldsymbol{\beta} + \ln p + (p-1)\ln t_i] - e^{a+\boldsymbol{x}_i'\boldsymbol{\beta}}t_i^p\} \tag{30.25}$$

由于当 $p=1$ 时，威布尔分布退化为指数分布，故可通过检验“$H_0: p=1$”(或 $H_0: \ln p = 0$)来确定应使用威布尔回归还是指数回归。

如果在方程(30.22)中，令基准风险 $\lambda_0(t) = e^{a+\gamma t}$，其中 a,γ 为待估参数，则为“冈珀茨回归”(Gompertz regression)，在人口学与保险精算领域应用广泛。与威布尔分布相比，冈珀茨分布的风险函数也是单调函数，即当 $\gamma>0$ 时单调递增，当 $\gamma<0$ 时单调递减，而当 $\gamma=0$ 时为常数(同样退化为指数分布)；二者的不同点在于，冈珀茨分布风险函数的变化率为指数形式($e^{\gamma t}$)，而威布尔分布风险函数的变化率为幂函数形式(t^{p-1})。由于当 $\gamma=0$ 时，冈珀茨分布退化为指数分布，故可通过检验“$H_0: \gamma=0$”来确定应使用冈珀茨回归还是指数回归。

威布尔分布与冈珀茨分布的风险函数都是单调函数，要么单调递增，要么单调递减，这个性质有时也与现实不符。比如，对于人类的死亡率而言，婴儿与老人的死亡率较高，而中青年的死亡率较低，其风险函数并不具有单调性。生物统计学家将人类死亡率的风险函数称为“浴缸风险”(bathtub hazard)，因为它的形状就像一个浴缸。在比例风险模型的框架下，一个解决方法是不设定基准风险 $\lambda_0(t)$ 的函数形式，参见下文的 Cox 模型。另一个解决方法是假设基准风险 $\lambda_0(t)$ 为阶梯函数，即 $\lambda_0(t)$ 在每个小区间内都是待估计的常数(正如指数模型的风险率为常数)，以便更好地拟合实际的风险函数，这被称为“分段固定风险模型”(piecewise constant hazard model 或 piecewise constant exponential model)。具体来说，分段固定风险模型的风险函数可写为

$$\lambda(t;\boldsymbol{x}) = \begin{cases} \lambda_1 e^{\boldsymbol{x}'\boldsymbol{\beta}} & t\in[0,\tau_1) \\ \lambda_2 e^{\boldsymbol{x}'\boldsymbol{\beta}} & t\in[\tau_1,\tau_2) \\ \vdots & \vdots \\ \lambda_M e^{\boldsymbol{x}'\boldsymbol{\beta}} & t\in[\tau_{M-1},\tau_M) \end{cases} \tag{30.26}$$

为了便于估计，上式可写为

$$\lambda(t;\boldsymbol{x}) = \begin{cases} e^{\ln\lambda_1+\boldsymbol{x}'\boldsymbol{\beta}} & t\in[0,\tau_1) \\ e^{\ln\lambda_2+\boldsymbol{x}'\boldsymbol{\beta}} & t\in[\tau_1,\tau_2) \\ \vdots & \vdots \\ e^{\ln\lambda_M+\boldsymbol{x}'\boldsymbol{\beta}} & t\in[\tau_{M-1},\tau_M) \end{cases} \tag{30.27}$$

从方程(30.27)可以看出，分段固定风险模型等价于在每个时间段内分别引入一个截距项。因此，对应于每个时间段，可以定义 M 个虚拟变量，并将 $(M-1)$ 个虚拟变量作为解释变量引入指数模型(假设模型中包括截距项)。例如，Lalive et al(2006)在研究失业持续时间时，将风险函数每四周分为一段(大多数失业者在 60 周内找到工作)。分段固定风险模型的缺点是，需要估计较多参数，故要求样本容量比较大。

对于比例风险模型 $\lambda(t;\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$，无论其基准风险 $\lambda_0(t)$ 如何设定(指数、威布尔、冈珀茨或分段固定)，对于参数 $\boldsymbol{\beta}$ 的经济含义解释都是一样的，即可将 β 解释为 x 对于风险函数的半

弹性，或将 e^{β} 解释为风险比率。

30.7 加速失效时间模型

对于比例风险模型，通常分析的重点是解释变量 $\boldsymbol{x}$ 对于风险函数 $\lambda(t;\boldsymbol{x})$ 的作用，不容易看出 $\boldsymbol{x}$ 对于平均寿命 $\mathrm{E}(T)$ 的影响。为此，考虑直接对 $\ln T$ 建模：

$$\ln T = \boldsymbol{x}'\boldsymbol{\beta} + u \tag{30.28}$$

其中，u 为扰动项。由于 $\ln T$ 可在 $(-\infty,\infty)$ 取值，故 u 可以服从取值于 $(-\infty,\infty)$ 的任何连续型分布。由方程(30.28)可得，$T = \mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}v$，其中 $v = \mathrm{e}^{u}$。记 v 的密度函数、累积分布函数、生存函数与风险函数分别为 $f_v(v)$，$F_v(v)$，$S_v(v)$ 与 $\lambda_v(v)$；记 T 的密度函数、累积分布函数、生存函数与风险函数分别为 $f_T(t)$，$F_T(t)$，$S_T(t)$ 与 $\lambda_T(t)$，则

$$S_T(t) = \mathrm{P}(T>t) = \mathrm{P}[\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}v > t] = \mathrm{P}[v > \mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t] = S_v(\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t) \tag{30.29}$$

因此，

$$f_T(t) = F_T'(t) = [1 - S_T(t)]' = -\frac{\partial S_v[\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t]}{\partial t} = f_v(\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t)\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}} \tag{30.30}$$

由此可知，

$$\lambda_T(t|x) = \frac{f_T(t)}{S_T(t)} = \frac{f_v(\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t)\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}}{S_v(\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t)} = \lambda_v(\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}t)\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}} \tag{30.31}$$

由上式可知，如果 $\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}} > 1$，则意味着对基准风险 $\lambda_v(t)$ 的加速；而如果 $\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}} < 1$，则意味着对基准风险 $\lambda_v(t)$ 的减速。因此，这类模型称为“加速失效时间模型”(Accelerated Failure Time Model，简记 AFT)，尽管 $\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}} < 1$ 意味着减速。

如果方程(30.28)中的扰动项 $u \sim N(0,\sigma^2)$，则称为“对数正态模型”(log-normal model)。如果 u 服从逻辑分布(logistic distribution)，则称为“对数逻辑模型”(log-logistic model)。如果 u 服从广义伽马分布(generalized gamma distribution) $Gamma(\beta_0,\kappa,\sigma)$，则称为“伽马模型”(gamma model)。其中，参数 κ 读为 kappa，在 Stata 中也记为 kappa。对数正态模型与对数逻辑模型的风险函数均为非单调函数(non-monotonic)，适用于风险率首先上升然后下降的情形；而包含三个参数的伽马模型，其风险函数形状更为灵活。如果 $\kappa = 1$，则伽马模型就是威布尔模型；如果 $\kappa = 1$ 而且 $\sigma = 1$，则伽马模型就是指数模型；如果 $\kappa = 0$，则伽马模型就是对数正态模型。因此，对原假设“H_0: $\kappa = 1$”或“H_0: $\kappa = 0$”进行检验，有助于选择模型适用的分布函数。另外，指数模型与威布尔模型也可以写成 AFT 形式，故它们既属于 PH 模型，也属于 AFT 模型；这是指数模型与威布尔模型独有的性质(其他模型，要么为 PH，要么为 AFT)。

对于 AFT 模型，同样可以进行 MLE 估计。但 AFT 模型的参数 $\boldsymbol{\beta}$ 与 PH 模型的参数 $\boldsymbol{\beta}$ 的经济解释不同。在 AFT 模型下，根据方程(30.28)，β 可以解释为当 x 边际增加时，能使平均寿命增加的百分比(即半弹性)。特别地，对于指数模型而言，在 PH 模型下，风险率为 $\lambda = \mathrm{e}^{a+\boldsymbol{x}'\boldsymbol{\beta}}$，而平均寿命为 $\mathrm{E}(T) = \dfrac{1}{\lambda} = \mathrm{e}^{-(a+\boldsymbol{x}'\boldsymbol{\beta})}$，故其 AFT 模型为 $\ln T = -(a+\boldsymbol{x}'\boldsymbol{\beta}) + u$。由此可见，PH 模型下参数 β 与 AFT 模型下的参数 $\boldsymbol{\beta}$ 正好符号相反，而绝对值相等。在 Stata 中进行指数回归时，默认为 PH 模型；但如果加选择项“`time`”，则为 AFT 模型；二者的估计系数绝对值相等，但符号正好相反。

由于 PH 模型与 AFT 模型都对风险函数的形式作了具体假设，故属于“参数回归”(parametric regression)。然而，在进行参数回归时，可供选择的分布函数很多(Stata 支持六种分布函数，即指数分布、威布尔、冈珀茨、对数正态、对数逻辑、伽马分布)。如果进行抉择？对于“嵌套模型”(nested models)，一个模型是另一模型的特例，可以对相关参数进行 Wald 或似然比检验(比如，指数模型是威布尔模型与冈珀茨模型的特例；威布尔模型与对数正态模型是伽马模型的特例)。对于“非嵌套模型”(non-nested models)，不存在这种嵌套关系，一个模型并不是另一模型的特例，则可通过 AIC 信息准则比较模型拟合优度，选择 AIC 最小的那个模型。AIC 准则的定义为

$$\text{AIC} = -2\ln L + 2(K + c) \tag{30.32}$$

其中，$\ln L$ 为对数似然函数，K 为解释变量 $\boldsymbol{x}$ 的维度，c 为概率分布的参数个数(比如，指数分布的 $c=1$；威布尔分布的 $c=2$；伽马分布的 $c=3$)。

30.8 Cox 模型

参数回归对风险函数的具体形式作了假设，如果模型设定正确，则 MLE 是最有效率的。然而，如果风险函数的设定错误，则 MLE 一般将导致不一致的估计。事实上，我们对于风险函数的具体形式常常并无把握。针对参数回归对于分布假定过强的缺点，在比例风险模型(PH)的框架下，Cox(1972,1975)提出了以下广为应用的方法，称为“Cox 模型”或“Cox PH 模型”。

由于比例风险模型 $\lambda(t;\boldsymbol{x}) = \lambda_0(t)\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}$ 在形式上为乘法(multiplicative)，故个体 i 与个体 j 的风险函数之比可以写为

$$\frac{\lambda(t;\boldsymbol{x}_i)}{\lambda(t;\boldsymbol{x}_j)} = \frac{\lambda_0(t)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\lambda_0(t)\mathrm{e}^{\boldsymbol{x}_j'\boldsymbol{\beta}}} = \mathrm{e}^{(\boldsymbol{x}_i-\boldsymbol{x}_j)'\boldsymbol{\beta}} \tag{30.33}$$

显然，个体 i 与个体 j 的风险函数之比不随时间而改变，只与 $(\boldsymbol{x}_i - \boldsymbol{x}_j)$ 有关。这个简单事实使得我们可以不必假设基准风险 $\lambda_0(t)$ 的具体函数形式，而依然得到对 $\boldsymbol{\beta}$ 的估计。由于构成风险函数 $\lambda_0(t)\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}$ 的前半部分 $\lambda_0(t)$ 不需要估计参数(非参数部分)，而后半部分 $\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}$ 需要估计参数(参数部分)，故为半参数回归(参见第 27 章)。

1. 失效时间不重叠的情形(no tied failures)

首先考察失效时间均不相同的情形，即每一个体的“死亡”时间都不一样。为了便于理解，先举一个简单的例子[①]。假设久期样本只包括以下四个观测值：

	subject	t	x
1.	1	2	4
2.	2	3	1
3.	3	6	3
4.	4	12	2

其中，subject 表示个体，t 表示失效时间(failure time)，而 x 为解释变量。记第 j 位个体的失效时间为 t_j，则四位个体的失效时间分别为 $t_1=2, t_2=3, t_3=6, t_4=12$，完全不重叠。记在时间 t_j

① 此例来自 Cleves et al(2004, p. 137)。

面临失效危险的所有个体所构成的集合为“风险集”(risk set)R_j。如果个体已经失效或被归并,则不再面临失效危险。

当 $t=2$ 时,风险集为 $R_1=\{1,2,3,4\}$,实际观测到第 1 位个体失效。

当 $t=3$ 时,风险集为 $R_2=\{2,3,4\}$,实际观测到第 2 位个体失效。

当 $t=6$ 时,风险集为 $R_3=\{3,4\}$,实际观测到第 3 位个体失效。

当 $t=12$ 时,风险集为 $R_4=\{4\}$,实际观测到第 4 位个体失效。

在以上四个失效时间(2,3,6,12),我们假设必然有一位个体会失效,然后计算在此条件下实际失效的那位个体失效的条件概率,分别记为 P_1,P_2,P_3,P_4;则似然函数可写为

$$L(\beta)=P_1P_2P_3P_4 \tag{30.34}$$

即这四个条件概率的乘积。下面分别来计算 P_1,P_2,P_3,P_4。显然,当 $t=12$ 时,风险集只包含第 4 位个体,故第 4 位个体失效的条件概率为 1,即 $P_4=1$。

当 $t=6$ 时,给定第 3、4 位个体必然有一位失效,在此条件下第 3 位个体失效的条件概率为

$$P_3=\frac{\lambda(6|x_3)}{\lambda(6|x_3)+\lambda(6|x_4)}=\frac{\lambda_0(6)e^{x_3\beta}}{\lambda_0(6)e^{x_3\beta}+\lambda_0(6)e^{x_4\beta}}=\frac{e^{x_3\beta}}{e^{x_3\beta}+e^{x_4\beta}} \tag{30.35}$$

其中,$\lambda(6|x_3)$ 表示 $t=6,x=x_3$ 时的风险率。在上式中,由于基准风险 $\lambda_0(6)$ 已被约去,故 P_3 并不依赖于具体的失效时间 $t=6$。

类似地,当 $t=3$ 时,给定第 2、3、4 位个体必然有一位失效,在此条件下第 2 位个体失效的条件概率为

$$P_2=\frac{\lambda(3|x_2)}{\lambda(3|x_2)+\lambda(3|x_3)+\lambda(3|x_4)}=\frac{e^{x_2\beta}}{e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}} \tag{30.36}$$

同样地,当 $t=2$ 时,给定第 1、2、3、4 位个体必然有一位失效,在此条件下第 1 位个体失效的条件概率为

$$P_1=\frac{\lambda(2|x_2)}{\lambda(2|x_1)+\lambda(2|x_2)+\lambda(2|x_3)+\lambda(2|x_4)}=\frac{e^{x_1\beta}}{e^{x_1\beta}+e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}} \tag{30.37}$$

因此,似然函数可写为

$$L(\beta)=\prod_{j=1}^{4}\left(\frac{e^{x_j\beta}}{\sum_{i\in R_j}e^{x_i\beta}}\right) \tag{30.38}$$

其中,R_j 为对应于 t_j(第 j 位个体失效时间)的风险集。更一般地,如果样本中有 L 个失效时间 $\{t_1,\cdots,t_L\}$ 以及多个解释变量 $\boldsymbol{x}$,则似然函数可写为

$$L(\boldsymbol{\beta})=\prod_{j=1}^{L}\left(\frac{e^{\boldsymbol{x}'_j\boldsymbol{\beta}}}{\sum_{i\in R_j}e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}\right) \tag{30.39}$$

上式称为“部分似然函数”(partial likelihood function),因为它并非完整的似然函数。从上面的推导可以看出,部分似然函数只依赖于样本中个体失效时间的排序,而不依赖于个体失效的具体时间。由于后者的信息被忽略了,故它只是“部分似然函数”。该估计量的协方差矩阵可根据 MLE 的一般理论得出。然而,由于同一个体多次出现于不同的风险集中,一般应使用 Lin and Wei(1989)提出的稳健标准误,在 Stata 中可由选择项“r”或“vce(robust)”来实现。

可以证明,基于部分似然函数(30.39)的 MLE 估计量是一致且渐近正态的,尽管其有效性不如基于完整似然函数的 MLE 估计量(即前面的参数回归)。但后者的一致性依赖于对风险函数形式的正确设定(我们对此一般无把握),而前者的效率损失通常不大。这意味着,如果真实

模型为指数回归，则 Cox 模型的估计结果与指数回归接近；如果真实模型为威布尔回归，则 Cox 模型的估计结果与威布尔回归接近；以此类推。由于 Cox 模型的稳健性，它在实践中十分流行，是半参数估计的成功例子。

2. 失效时间有重叠的情形（tied failures）

如果样本中的失效时间有重叠，则部分似然函数的计算更为复杂。延续上面的简单例子，但假设第 3 位与第 2 位个体同时在 $t=3$ 时失效，原始数据变为：

	subject	t	x
1.	1	2	4
2.	2	3	1
3.	3	3	3
4.	4	12	2

当 $t=3$ 时，风险集为 $R_2=\{2,3,4\}$，而实际观测到第 2、3 位个体同时失效。此时，应如何计算在给定风险集中两位个体会失效的条件下，第 2、3 位个体失效的条件概率？通常有以下两种处理方法。

处理方法之一是假设第 2、3 位个体的失效时间其实并不完全相同，称为“精确边际计算法”（exact-marginal calculation，相应的 Stata 选择项为“exactm”）。如果失效时间为连续变量，则第 2、3 位个体同时失效的概率为 0。之所以在样本中存在同时失效的个体，只是由于我们的观测时间不够细致（比如，以周或月为分析时间）。基于这个思想，共有两种可能情形。第一种情形是，第 2 位个体先失效，然后第 3 位个体失效，记此情形的条件概率为 P_{23}；第二种情形是，第 3 位个体先失效，然后第 2 位个体失效，记此情形的条件概率为 P_{32}；而总的条件概率为 $(P_{23}+P_{32})$。

为了简化符号，记 $r_j\equiv e^{x_j\beta}$，根据前面的推导逻辑可知：

$$P_{23}=\frac{r_2}{r_2+r_3+r_4}\cdot\frac{r_3}{r_3+r_4} \tag{30.40}$$

$$P_{32}=\frac{r_3}{r_2+r_3+r_4}\cdot\frac{r_2}{r_2+r_4} \tag{30.41}$$

然而，如果多位个体同时失效，则精确边际法可能计算量过大。比如，有 10 位个体同时失效，则共有 10!=3 628 800 种情形。为此，Breslow（1974）提出了以下近似计算方法（相应的 Stata 选择项“breslow”为默认选项），即在计算方程（30.40）与（30.41）时，不调整每一项分母中的风险集（一直使用最初的风险集）。根据 Breslow 的方法，

$$P_{23}\approx P_{32}\approx\frac{r_2r_3}{(r_2+r_3+r_4)^2} \tag{30.42}$$

Efron（1977）提出了另一近似计算方法（相应的 Stata 选择项为“efron”）。Efron 方法比 Breslow 更为精确，但计算费时更长。当 $t=3$ 时，风险集为 $\{3,4\}$ 或 $\{2,4\}$，对应的条件概率之分母分别为 (r_3+r_4) 或 (r_2+r_4)。如果使用二者的平均数作为分母，可得

$$P_{23}\approx P_{32}\approx\frac{r_2r_3}{(r_2+r_3+r_4)\left(\frac{r_2+r_3}{2}+r_4\right)} \tag{30.43}$$

Breslow 方法与 Efron 方法都是对精确边际法的近似。

处理方法之二是假设第 2、3 位个体确实同时失效，称为“精确部分计算法”（exact-partial

calculation，相应的 Stata 选择项为“exactp”）。给定风险集{2,3,4}中有两位个体同时失效，则共有三种情形，即第 2、3 位个体同时失效，第 2、4 位个体同时失效，以及第 3、4 位个体同时失效。因此，第 2、3 位个体确实同时失效的条件概率为：

$$P_{23}=\frac{r_2r_3}{r_2r_3+r_2r_4+r_3r_4} \tag{30.44}$$

对比表达式(30.44)与(30.40)、(30.41)可知，精确部分法与精确边际法的计算结果并不相等，但一般差别不大。究竟采取哪种方法，取决于将时间理解为连续型变量（精确边际法及其近似）还是离散型变量（精确部分法）。在实践中，如果风险集较大或同时失效的个体较多，则精确法（精确边际法或精确部分法）的计算时间可能很长，故 Stata 的默认方法为 Breslow 近似法（如果风险集不大或同时失效的个体较少，则可采取精确法）。

3. 估计基准风险函数

虽然 Cox 模型不设定基准风险 $\lambda_0(t)$ 的函数形式，但在估计 $\boldsymbol{\beta}$ 之后依然可以得到对 $\lambda_0(t)$ 的估计。首先，根据方程(30.12)，$S(t)=\exp[-\Lambda(t)]$，代入比例风险 $\lambda(t;\boldsymbol{x})=\lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$ 可得

$$\begin{aligned}S(t)&=\exp[-\Lambda(t)]=\exp\left[-\int_0^t\lambda(u;\boldsymbol{x})\mathrm{d}u\right]=\exp\left[-\int_0^t\lambda_0(u)e^{\boldsymbol{x}'\boldsymbol{\beta}}\mathrm{d}u\right]\\&=\exp\left[\left(-\int_0^t\lambda_0(u)\mathrm{d}u\right)(e^{\boldsymbol{x}'\boldsymbol{\beta}})\right]=\left[\exp\left(-\int_0^t\lambda_0(u)\mathrm{d}u\right)\right]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}\equiv S_0(t)^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}\end{aligned} \tag{30.45}$$

其中，$S_0(t)\equiv\exp\left(-\int_0^t\lambda_0(u)\mathrm{d}u\right)$ 为对应于基准风险 $\lambda_0(t)$ 的基准生存函数。根据 Cox 模型，已知 $\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}$；根据 KM 估计量，已知 $S(t)$ 的估计量 $\hat{S}(t)$；故根据方程(30.45)可得基准生存函数 $S_0(t)$ 的估计量 $\hat{S}_0(t)$。其次，累积风险函数可写为

$$\Lambda(t)=\int_0^t\lambda(u)\mathrm{d}u=\int_0^t\lambda_0(u)e^{\boldsymbol{x}'\boldsymbol{\beta}}\mathrm{d}u=e^{\boldsymbol{x}'\boldsymbol{\beta}}\int_0^t\lambda_0(u)\mathrm{d}u\equiv e^{\boldsymbol{x}'\boldsymbol{\beta}}\Lambda_0(t) \tag{30.46}$$

其中，$\Lambda_0(t)\equiv\int_0^t\lambda_0(u)\mathrm{d}u$ 为对应于基准风险 $\lambda_0(t)$ 的基准累积风险函数。从 KM 估计量 $\hat{S}(t)$ 出发，根据 $\Lambda(t)=-\ln S(t)$，可得到 $\Lambda(t)$ 的估计量 $\hat{\Lambda}(t)$。故根据方程(30.46)可以得到 $\hat{\Lambda}_0(t)$。再次，根据关系式 $\Lambda_0(t)\equiv\int_0^t\lambda_0(u)\mathrm{d}u$，可以计算每个失效时间对于 $\hat{\Lambda}_0(t)$ 的边际贡献，即 $\hat{\lambda}_0(t)$。最后，使用核密度方法（参见第 27 章）将此边际贡献光滑化，即得到对 $\lambda_0(t)$ 的估计。

根据对基准风险的估计 $\hat{\lambda}_0(t)$ 以及参数估计 $\hat{\boldsymbol{\beta}}$，可以计算在 $\boldsymbol{x}=\boldsymbol{x}^*$ 处的风险函数：

$$\hat{\lambda}(t;\boldsymbol{x}^*)=\hat{\lambda}_0(t)e^{\boldsymbol{x}^{*\prime}\hat{\boldsymbol{\beta}}} \tag{30.47}$$

通常，令 $\boldsymbol{x}^*=\bar{\boldsymbol{x}}$，即计算在解释变量平均值处的风险函数（对应于样本中解释变量都为平均数的假想个体的风险函数，这是 Stata 的默认取值）；也可以选择其他感兴趣的解释变量取值（特别当 x 为二值变量或离散变量时）。

30.9 比例风险模型的设定检验

比例风险模型（包括 Cox 模型）的重要假设是 $\lambda(t;\boldsymbol{x})=\lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$，如果这个假设不成立，则比例风险模型不能成立。为此，需要对比例风险模型进行设定检验，在 Stata 中称为“PH-

assumption tests"。

1. 对数 - 对数图(log-log plot)

根据比例风险假定 $\lambda(t;\boldsymbol{x})=\lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$,可将累积风险函数写为

$$\Lambda(t)=\int_0^t\lambda(u;\boldsymbol{x})du=\int_0^t\lambda_0(u)e^{\boldsymbol{x}'\boldsymbol{\beta}}du=e^{\boldsymbol{x}'\boldsymbol{\beta}}\int_0^t\lambda_0(u)du\equiv e^{\boldsymbol{x}'\boldsymbol{\beta}}\Lambda_0(t) \tag{30.48}$$

其中,$\Lambda_0(t)$为对应于 $\lambda_0(t)$的累积风险函数。由上式可知,

$$\ln\Lambda(t)=\boldsymbol{x}'\boldsymbol{\beta}+\ln\Lambda_0(t) \tag{30.49}$$

根据方程(30.11),$\Lambda(t)=-\ln S(t)$,将此式代入方程(30.49),且方程两边同乘(-1)可得

$$-\ln[-\ln S(t)]=-\boldsymbol{x}'\boldsymbol{\beta}-\ln[-\ln S_0(t)] \tag{30.50}$$

其中,$S_0(t)$为对应于 $\lambda_0(t)$的生存函数。由于 $0<S(t)<1$,故 $\ln S(t)<0$,$-\ln S(t)>0$。从方程(30.50)可以看出,函数 $-\ln[-\ln S(t)]$的斜率并不依赖于 $\boldsymbol{x}$。因此,当 $\boldsymbol{x}$ 取不同值时(比如,$\boldsymbol{x}$ 为虚拟变量或离散变量),函数 $-\ln[-\ln S(t)]$应为相互平行的曲线,只是截距项($-\boldsymbol{x}'\boldsymbol{\beta}$)不同。据此画出的图被称为"对数 - 对数图"(log-log plot)。如果对数 - 对数图中的曲线相互平行,则支持比例风险假设;如果不同曲线的斜率相差甚远,则意味着比例风险假设不成立。对数 - 对数图的缺点是,如何确定曲线是否平行带有主观性。

2. 观测 - 预测图(observed versus expected plot)

检验比例风险假设的另一图示法为"观测 - 预测图"。针对每一解释变量 x,分别画图。假设 x 为离散变量,首先根据 x 的不同取值水平画其 Kaplan - Meier(KM)生存函数图,即观测图(observed plots)。其次,估计 Cox 模型,以此计算基准生存函数 $S_0(t)$,然后代入 x 的不同取值水平,得到相应的生存函数图,即预测图(expected plot)。最后,给定 x 的一个取值水平,比较其观测图与预测图是否足够接近。如果很接近,则表明变量 x 满足比例风险假设;反之,则比例风险假设不成立。观测 - 预测图的缺点同样是它的主观性。

3. 基于残差的检验(residual - based tests)

对于像久期分析这样的非线性模型,有多种残差的定义。其中,对于检验比例风险假设最有用的是"舍恩菲尔德残差"(Schoenfeld residuals)。对于个体 j 与解释变量 x_k($k=1,\cdots,K$,假设共有 K 个解释变量),可以计算其对应的舍恩菲尔德残差如下:

$$r_{kj}=x_{kj}-\sum_{i\in R_j}\left(x_{ki}\cdot\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\sum_{i\in R_j}e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right) \tag{30.51}$$

其中,R_j 为当个体 j 失效时的风险集。直观来看,舍恩菲尔德残差 r_{kj}即为失效个体的解释变量观测值 x_{kj}减去仍处于风险集中个体的解释变量之加权平均,而权重为"相对风险"(relative hazard)$e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$。如果比例风险假设成立,则舍恩菲尔德残差不应随时间呈现出规律性的变化。针对每个解释变量 x_k,都可以将其舍恩菲尔德残差与时间画图,并考察其斜率是否为 0。进一步,可以把舍恩菲尔德残差对时间回归,然后检验时间的系数是否为 0。

30.10 分层 Cox 模型

如果比例风险假设不满足,处理方法之一为"分层 Cox 模型"(stratified Cox model)。不失一般性,假设变量 sex(性别)不满足比例风险假设,则可将样本中个体分为两组,第一组为男性,第二组为女性;并假设男性与女性的基准风险函数不同,但参数 $\boldsymbol{\beta}$ 都相同(此假设也可以检验)。

因此,个体的风险函数可写为

$$\lambda(t;\boldsymbol{x}_j)=\begin{cases}\lambda_{01}(t)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}, & j\text{为男性}\\ \lambda_{02}(t)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}, & j\text{为女性}\end{cases} \tag{30.52}$$

其中,$\lambda_{01}(t)$为男性组的基准风险,而 $\lambda_{02}(t)$为女性组的基准风险。根据男性组的风险函数 $\lambda_{01}(t)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}$,可以计算男性组的部分似然函数 L_1;类似地,可以计算女性组的部分似然函数 L_2。则整个样本的部分似然函数为 $L=L_1\cdot L_2$。显然,由于变量 sex 不直接出现在似然函数中(因为 sex 不满足比例风险假设,故无法将其直接引入 Cox 模型),故无法估计其效应;这是进行分层分析的代价。

进一步,如果某分层变量有 k 个取值水平,则可将样本中个体分为 k 层或 k 组,然后进行分层估计。更一般地,如果有多个分层变量,则应考虑它们的交叉情形。比如,变量 x 有 3 种可能取值,变量 z 有 2 种可能取值,二者都不满足比例风险假设,故同为分层变量;此时,应将样本分为 6 组,使得每组内的 x,z 取值都相同。在 Stata 中,只需指定分层变量(可以为多个分层变量),其余的计算都可自动进行。

分层 Cox 模型的一个重要假定为,不同组的参数 $\boldsymbol{\beta}$ 都相同。为了检验此假定,可以引入分层变量与留在 Cox 模型中变量的互动项进行分层分析,然后检验所有这些互动项的联合显著性。如果都不显著,则支持"不同组参数 $\boldsymbol{\beta}$ 都相同"的原假设。

30.11 随时间而变的解释变量

如果比例风险假设不满足,处理方法之二为引入随时间而变的解释变量(Time - Varying Covariates,简记 TVC)。比如,在持续失业的过程中,宏观经济状况可能发生变化,从而影响就业概率。引入 TVC 后,需要注意可能导致内生性。一般来说,不由个体控制的变量(比如宏观经济)是外生的;而个体可以影响的变量则可能为内生的。例如,在考察婚姻持续时间时,如果使用孩子的数量作为 TVC 则可能导致内生性问题。由于稳定的婚姻会增加孩子的数量,故根据数据得出的"孩子越多,婚姻越持久"结论并不可靠。下面,假设 TVC 为外生。

引入 TVC 后,可将风险函数写为

$$\lambda(t;\boldsymbol{x}(t))=\lambda_0(t)\mathrm{e}^{\boldsymbol{x}'(t)\boldsymbol{\beta}} \tag{30.53}$$

其中,解释变量 $\boldsymbol{x}(t)$ 可以时间而变,而其系数 $\boldsymbol{\beta}$ 仍为固定。从上式可以看出,一旦解释变量 $\boldsymbol{x}(t)$ 的取值变化,其风险函数也立即发生相应变化。如果要考虑滞后效应,可引入 $\boldsymbol{x}(t-1)$ 作为解释变量。

根据方程(30.53),个体 i 与个体 j 的风险函数之比可写为

$$\frac{\lambda(t;\boldsymbol{x}_i(t))}{\lambda(t;\boldsymbol{x}_j(t))}=\frac{\lambda_0(t)\mathrm{e}^{\boldsymbol{x}_i'(t)\boldsymbol{\beta}}}{\lambda_0(t)\mathrm{e}^{\boldsymbol{x}_j'(t)\boldsymbol{\beta}}}=\mathrm{e}^{[\boldsymbol{x}_i(t)-\boldsymbol{x}_j(t)]'\boldsymbol{\beta}} \tag{30.54}$$

显然,此风险函数之比是时间 t 的函数,故此模型不再满足比例风险假设,称为"扩展 Cox 模型"(extended Cox model)。但这对于基于似然函数(参数回归)或部分似然函数(Cox 模型)的估计并无实质影响。这是因为,现实数据总是离散的(譬如年度数据、月度数据、日度数据),故可将每位个体的单一记录分为几条记录,对应于几个时间段,使得在每个时间段内 TVC 为常数,然后沿用解释变量不随时间而变的方法进行估计。如果已将数据设为此格式,则无论是否包括

TVC 变量，其 Stata 命令并无区别。

作为一个简单例子[①]，假设个体 1 从时间 0 至 T 一直处于失业状态，并在时间 T 找到工作。假设 $0 < t_1 < t_2 < T$。共有两个解释变量，其中 x_1 不随时间而变（假设为虚拟变量，比如性别）；而 $x_2(t)$ 可以随时间而变，在区间 $[0, t_1)$，$[t_1, t_2)$，$[t_2, T)$ 的取值分别为 $x_2(t_1)$，$x_2(t_2)$ 与 $x_2(T)$，故为阶梯函数。则个体 1 的信息由三条记录构成，参见表 30.1。

表 30.1 个体 1 的相关记录

个体 ID	持续时间	x_1	$x_2(t)$	归并虚拟变量 d
1	t_1	1	$x_2(t_1)$	0
1	t_2	1	$x_2(t_2)$	0
1	T	1	$x_2(T)$	1

由表 14.1 可知，个体 1 的完整信息被分为三个部分，对应于三条记录。在第 1、2 条记录，个体 1 均未失效，而以归并结束，故归并虚拟变量为 0；在第 3 条记录，个体 1 在时间 T 失效，故归并虚拟变量为 1。

扩展 Cox 模型的用途之一是用来检验比例风险假设。回到解释变量不随时间而变的情形，可以将比例风险假设解释为风险函数 $\lambda(t;\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$ 的参数 $\boldsymbol{\beta}$ 不随时间而变。考虑最简单的一维情形。假设 β 随时间而变，并将其写为 $\beta + \gamma g(t)$，其中 $g(t)$ 为时间的函数（比如，t 或 $\ln t$），则风险函数可写为

$$\lambda(t;x) = \lambda_0(t)e^{x[\beta+\gamma g(t)]} = \lambda_0(t)e^{x\beta+\gamma x g(t)} \tag{30.55}$$

这相当于是在原模型中引入了一个 TVC，即 $x \cdot g(t)$，由不随时间而变的解释变量 x 与时间的函数 $g(t)$ 的乘积构成，是经人为定义而生成的 TVC。如果 $\gamma = 0$，则方程(30.55)简化为标准 Cox 模型的风险函数。因此，可通过检验原假设“$H_0: \gamma = 0$”来检验比例风险假设。通常选择 $g(t)$ 为 t 或 $\ln t$。在多维情形下，则引入所有解释变量 $\boldsymbol{x}$ 与 $g(t)$ 的乘积，然后检验所有这些乘积项的系数联合显著性。在作此检验时，由于引入了原模型中所没有的 TVC，故需要调整数据格式，将个体的信息根据需要分为多条记录。幸运的是，在 Cox 模型中，Stata 可自动完成这一操作，只要使用选择项“`tvc(var) texp(exp)`”即可，前者用来指定 $\boldsymbol{x}$，而后者用来指定 $g(t)$（详见本章最后一节）。

30.12 不可观测的异质性

在进行久期分析时，样本中的个体可能具有“不可观测的异质性”（unobserved heterogeneity），也称为“弱质”（frailty）。比如，在研究病人的存活时间时，病人的体质或弱质并不能完全观察，但却影响病人的死亡风险。在线性模型中，忽略不可观测的异质性相当于遗漏变量；如果此遗漏变量与解释变量不相关，则不影响估计量的一致性。然而，在久期分析这样的非线性模型中，即使不可观测的异质性与解释变量不相关，仍会导致不一致估计。

一个经典的例子是对于失业持续时间的研究。经验数据表明，失业时间越长，则找到工作的概率越低，即风险率随着时间而下降。如果样本中的每位个体完全一样，则意味着负向久期依赖

① 此例来自 Cameron and Trivedi(2005, p. 598)。

(negative duration dependence)。但事实可能并非如此。假设失业人口可以分为两类,即 F 类(fast,其风险函数恒等于0.4)与 S 类(slow,其风险函数恒等于0.1),这两类各占一半。由于 F 类的风险率高于 S 类,故 F 类平均地比 S 类更快找到工作而退出样本,导致样本中 F 类占比随时间而下降,引起整体的风险率下降。具体来说,假设样本中有200人,其中100人为 F 类,另100人为 S 类。在100位 F 类人中,第1期将有40人找到工作,第2期将有24人找到工作,而第3期将有14.4人找到工作。另一方面,在100位 S 类人中,第1期将有10人找到工作,第2期将有9人找到工作,而第3期将有8.1人找到工作。因此,就整个样本而言,第1期的风险率为 $(40+10)/200=0.25$,第2期的风险率为 $(24+9)/150=0.22$,第3期的风险率为 $(14.4+8.1)/117=0.192$,呈逐渐下降的趋势。显然,整体风险率的逐渐下降,只是由于 F 类与 S 类的结构变化而造成,而个体的风险函数始终为常数。此效应被称为"弱质效应"(frailty effect)。对于威布尔模型,风险函数为 $\lambda(t)=\gamma pt^{p-1}$,其中参数 p 决定风险函数的斜率,表示久期依赖的方向与程度。如果存在不可观测的异质性而被忽略,则久期依赖参数 p 将被系统地低估,得不到一致估计。

解决方法是直接将不可观测的异质性引入模型中。假设个体 j 的风险函数为

$$\lambda(t;\boldsymbol{x}_j,v_j)=\lambda_0(t)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}v_j,\quad v_j>0 \tag{30.56}$$

其中,不可观测的异质性 $v_j>0$ 以乘积的形式进入风险函数。v_j 越大,则个体 j 越弱质,失效的风险越高。$\lambda(t;\boldsymbol{x}_j,v_j)$ 称为"条件风险函数"(conditional hazard function),因为它是给定 v_j 条件下的风险函数。从方程(30.56)出发,可以得到生存函数:

$$S(t;\boldsymbol{x}_j,v_j)=\mathrm{e}^{-\Lambda(t)}=\exp\left[-\int_0^t\lambda_0(u)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}v_j\mathrm{d}u\right]=\left[\exp\left(-\int_0^t\lambda_0(u)\mathrm{e}^{\boldsymbol{x}_j\boldsymbol{\beta}}\mathrm{d}u\right)\right]^{v_j}=[S(t;\boldsymbol{x}_j)]^{v_j} \tag{30.57}$$

其中,$S(t;\boldsymbol{x}_j)$ 为不含异质性的生存函数。由于 v_j 不可观测,故需要假设其概率分布,然后将 v_j 积分掉,得到无条件的风险函数与生存函数。假设 v_j 的概率密度函数为 $g(v_j)$,则无条件风险函数与生存函数分别为

$$\lambda(t;\boldsymbol{x}_j)=\int_0^\infty\lambda(t;\boldsymbol{x}_j,v_j)g(v_j)\mathrm{d}v_j \tag{30.58}$$

$$S(t;\boldsymbol{x}_j)=\int_0^\infty S(t;\boldsymbol{x}_j,v_j)g(v_j)\mathrm{d}v_j \tag{30.59}$$

为了使方程(30.58)与(30.59)有解析解,常假设原模型为威布尔模型(指数模型为其特例),而 v_j 服从 $Gamma(\delta,\delta)$ 分布或"逆正态分布"(Inverse Gaussian,简记IG);这类模型称为"混合模型"(mixture model)。将 v_j 的期望值标准化为1(故 $v_j=1$ 代表平均弱质,average frailty),并记其方差为 θ(在Stata记为theta)。如果 $\theta=0$,则 v_j 退化为常数,不存在异质性;故可通过检验"$H_0:\theta=0$"来判断是否存在异质性。由于不可观测的异质性不再出现在无条件风险函数(30.58)与生存函数(30.59),故可对其进行MLE估计。

另外,还可考虑依某变量的取值将个体分组,然后假设组内个体为同质,而不同组的个体具有异质性,这被称为"共享异质性"(shared frailty),即该异质性为同组内成员所共有。与此对应,上文的个体异质性则被称为"非共享异质性"(unshared frailty)。

30.13 其他久期分析模型

迄今为止,我们假设每位个体一旦离开其持续的状态,就不可能再回去。比如,考察病人存

活时间时，病人一旦去世，就无法复活。这被称为“单一久期”（single-spell）数据。然而，在考察失业持续时间时，失业者找到工作后，还可能再次失业，导致“多重久期”（multiple-spell）数据。多重久期数据的建模难点在于，不同阶段的失业时间可能相关。

另外，我们一直假设个体只可能存在两种状态，比如生或死，失业或就业。现实中还可能出现在多种状态间转换的情形，比如劳动者的状态可以包括失业、全职就业及兼职就业。此时，失业者可能转换到两种不同的状态中，即全职就业或兼职就业。这类模型称为“竞争风险模型”（competing risk model），需要同时考虑多个风险函数。

进一步学习久期分析的资源包括 Cameron and Trivedi（2005，Chapter 17 ~ 19），Cleves et al（2004），Hosmer and Lemeshow（1999），Kleinbaum and Klein（2011），Lancaster（1990）与 van den Berg（2001）。

30.14 久期分析的 Stata 命令及实例

Stata 称久期数据为“生存时间”数据（survival time，简记 st），故所有久期分析的命令都以 st 开头（正如所有面板命令都以 xt 开头）。在进行久期分析之前，首先要告诉 Stata 我们的数据为久期数据，其命令格式为

```
stset duration_var,failure(failvar) id(idvar)
```

其中，“duration_var”表示持续时间，即久期分析的被解释变量；选择项“failure(failvar)”表示归并虚拟变量（失效 =1，归并 =0）；选择项“id(idvar)”用来确定个体 ID（适用于一位个体有多条记录的情形）[①]。有关命令 stset 的更多选择项，参见“help stset”。

运行 stset 命令后，Stata 将自动生成四个新变量_t0，_t，_d，_st，分别表示记录开始的时间，记录结束的时间，归并虚拟变量（失效 =1，归并 =0），以及有效观测值虚拟变量（与当前分析有关的观测值 =1，其他 =0）。Stata 的后续计算将基于此四个变量，而忽略数据集中原有的时间或事件变量。因此，如果中途改变原始数据，则应再次运行 stset 命令，以更新变量_t0，_t，_d，_st 的取值。

```
stdes  （描述久期数据）
stsum  （显示久期数据的特征）
sts graph,survival cumhaz hazard ci by(varlist)  （画图，描述性分析）
```

其中，选择项“survival”为默认选项，表示画 Kaplan - Meier 生存函数；选择项“cumhaz”表示画累积风险函数；选择项“hazard”表示画风险函数；选择项“ci”表示同时画出置信区间；选择项“by(varlist)”表示根据某变量的不同取值，划分子样本，然后分别画图。

如果需要将一条记录分为几条记录，可以用以下命令：

```
stsplit newvar,at(numlist) every(#)  （在指定的时间点分割数据）
stsplit,at(failures)                 （在失效时间分割数据）
```

如果进行参数回归，可使用以下命令：

```
streg x1 x2 x3,r distribution(dist_var) frailty(frailty_dist) shared
```

① 有时同一个体的持续时间被分时段记录在案，对于这种多记录数据（multiple-record data），必须指明 ID 以识别每位个体。

(varname) strata(strata_var) nolog noshow time

其中,"x1 x2 x3"为解释变量。命令 streg 的独特之处在于,不需要指定被解释变量 y,因为在运行 stset 命令时,已经确定了被解释变量。选择项"r"表示使用稳健标准误。选择项"distribution(dist_var)"用来指定概率分布,备选分布包括exponential,weibull,gompertz,loglogistic,lognormal,gamma。选择项"frailty(frailty_dist)"表示使用包含不可观测的异质性模型,备选的异质性分布包括gamma与invgaussian。选择项"shared(varname)"表示界定"共享异质性"(shared frailty)的变量。选择项"strata(strata_var)"表示进行分层分析,并指定分层变量。选择项"nolog"表示不显示 MLE 计算的迭代过程,选择项"noshow"表示不显示有关久期数据设定(stset)的信息(默认显示)。选择项"time"表示估计加速失效模型,该选项只适用于 exponential 和 weibull 分布,因为只有这两者既属 AFT 又属 PH 模型。

如果估计半参数 Cox 模型,可使用以下命令:

stcox x1 x2 x3,r frailty(frailty_dist) shared(varname) strata(strata_var) breslow efron exactm exactp nolog noshow

命令 stcox 的结构与选择项与 streg 大致相同。选择项"breslow,efron,exactm,exactp"分别表示使用 Breslow 方法(默认值)、Efron 方法、精确边际法、精确部分法来处理失效时间重叠(ties)的情形。

使用命令 streg 或 stcox 后,可用以下命令来得到相关函数的直观图形:

stcurve,survival cumhaz hazard alpha1 at(varname1 = # varname2 = #)

其中,选择项"survival"表示画生存函数,选择项"cumhaz"表示画累积风险函数,选择项"hazard"表示画风险函数。在存在不可观测异质性的情况下,选择项"alpha1"表示根据给定异质性的"条件异质性模型"(conditional frailty model)进行画图;默认值为根据"无条件异质性模型"(unconditional frailty model)进行画图。选择项"at(varname1 = # varname2 = #)"表示,在变量 varname1 = #,变量 varname2 = #,而其他变量等于均值处画图;默认值为在所有变量的均值处画图。

检验比例风险假定的命令包括:

stphplot,by(varname)

stcoxkm,by(varname)

estat phtest,detail plot(varname)

其中,命令"stphplot,by(varname)"(表示 st proportional hazard plot)根据变量 varname 画对数 - 对数图。命令"stcoxkm,by(varname)"(表示 st cox Kaplan - Meier)根据变量 varname 画观测 - 预测图;此命令可单独使用,无须先运行 stcox 命令。在运行命令 stcox 之后,可使用命令"estat phtest"进行基于舍恩菲尔德残差的检验,选择项"detail"表示对每个解释变量都分别进行检验(默认为仅进行联合检验),而选择项"plot(varname)"表示将变量 varname 的检验结果画图(这两个选择项不能并用)。

下面以数据集 recid.dta 为例[①]。该数据集用来研究刑满释放犯再次被捕的时间,记为 durat(单位为月)。首先,对在 1977 年 7 月 1 日至 1978 年 6 月 30 日期间释放的罪犯进行随机抽样,然后于 1984 年 4 月考察这些刑满释放犯的再被捕记录。因此,1984 年 4 月为共同的归并日

① 该数据集来自 Wooldridge(2010),最早为 Chung et al(1991)所使用。

期,但由于每位犯人开始自由的时间不尽相同,故归并时间 C_i^* 仍可不同。解释变量包括:workprg(是否参加狱中劳动项目),priors(前科次数),tserved(狱中关押月数),felon(是否犯重罪),alcohol(是否酗酒),drugs(是否有吸毒史),black(是否黑人),married(在狱中时是否已婚),educ(教育年限),age(年龄),其中 workprg 为我们主要关心的变量。另外,归并虚拟变量为 cens(1 = 右归并)。

(1) 首先,调用此数据集,定义为久期数据,并考察数据特征。

```
. use recid.dta,clear
. stset durat,failure(cens =0)
```

其中,由于归并虚拟变量的定义为"1 = 右归并,0 = 失效"(与通常的定义相反),故使用选择项"failure(cens =0)",而不使用"failure(cens)"。

```
     failure event:  cens == 0
obs. time interval:  (0, durat]
 exit on or before:  failure

------------------------------------------------------------------------------
     1445  total obs.
        0  exclusions
------------------------------------------------------------------------------
     1445  obs. remaining, representing
      552  failures in single record/single failure data
    80013  total analysis time at risk, at risk from t =         0
                                earliest observed entry t =         0
                                     last observed exit t =        81
```

上表显示,失效事件以"cens = =0"来定义,在 1 445 个有效观测值中,552 位个体失效,而其他个体被归并。下面,看一下久期数据的基本特征。

```
. stdes
```

```
        failure _d:  cens == 0
  analysis time _t:  durat

                                          |-------------- per subject --------------|
Category                        total        mean         min      median         max
------------------------------------------------------------------------------------
no. of subjects                  1445
no. of records                   1445           1           1           1           1

(first) entry time                               0           0           0           0
(final) exit time                          55.37232           1          71          81

subjects with gap                   0
time on gap if gap                  0
time at risk                    80013     55.37232           1          71          81

failures                          552     .3820069           0           0           1
------------------------------------------------------------------------------------
```

(2) 其次,进行描述性分析,即画生存函数、累积风险函数与风险函数。

```
. sts graph,ci
```

该命令的默认值为画样本生存函数,即 Kaplan-Meier 估计量,结果参见图 30.1。

由于我们关心参加狱中劳动项目的效应,故以 workprg 的取值将全样本一分为二,考察两个子样本的生存函数是否有区别。

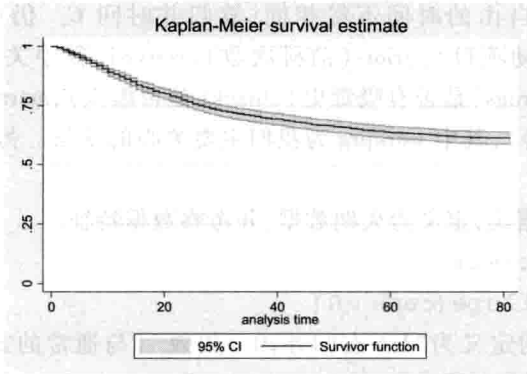


图 30.1 Kaplan-Meier 生存函数估计量

```
. sts graph,by(workprg) plot2opts(lp("-"))
```

其中,选择项"plot2opts(lp("-"))"表示第二个图以虚线表示,便于与第一个图的默认实线相区别。结果参见图 30.2。

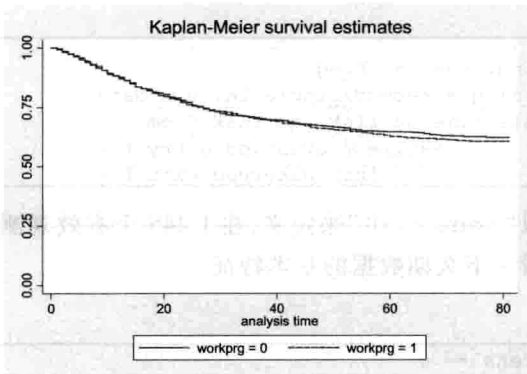


图 30.2 根据 workprg 分组的 Kaplan-Meier 生存函数

从图 30.2 可知,是否参加狱中劳动项目,对于生存函数并无明显影响。这预示着,狱中劳动项目对于罪犯再次被捕的时间可能无显著影响。

```
. sts graph,cumhaz ci
```

其中,选择项"cumhaz"表示画累积风险函数,结果参见图 30.3。

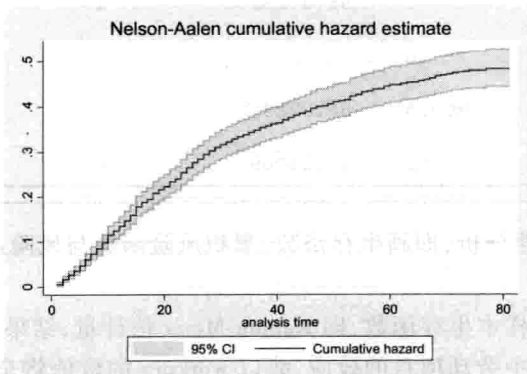


图 30.3 Nelson-Aalen 累积风险函数估计量

```
. sts graph,hazard ci
```

其中,选择项“hazard”表示画风险函数,结果参见图 30.4。

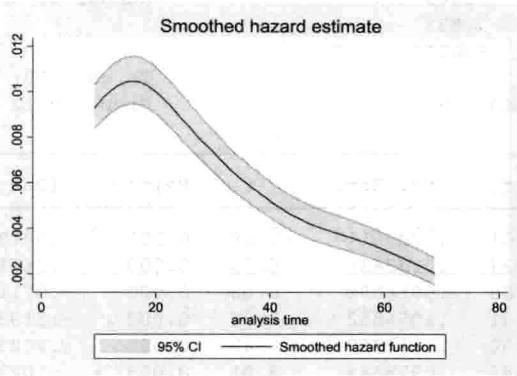


图 30.4 风险函数估计量

从图 30.4 可知,在罪犯出狱后再次被捕的风险率将首先上升,然后才下降。下面,根据 workprg 的取值分组画风险函数,结果参见图 30.5。

```
. sts graph,hazard by(workprg) plot2opts(lp("-"))
```

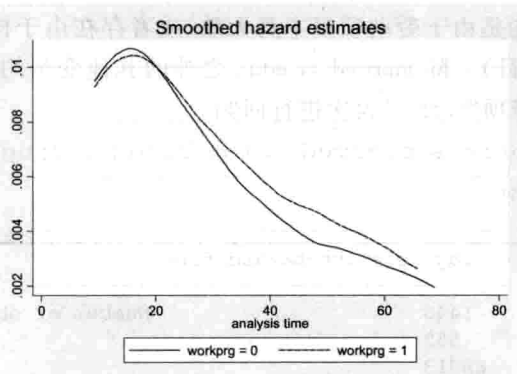


图 30.5 根据 workprg 分组的风险函数

从图 30.5 可知,是否参加狱中劳动项目对于风险函数影响不大。甚至在相当大的区间内,参加劳动项目者的风险率反而更高。当然,由于未引入解释变量,以上分析只是对被解释变量 durat 无条件分布的初步分析。

(3) 进行参数回归。首先进行最简单的单参数指数回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(e) nohr nolog noshow
```

其中,选择项“nohr”表示显示回归系数,不显示风险比率(hazard ratio),默认显示风险比率。

```
Exponential regression -- log relative-hazard form

No. of subjects =         1445                     Number of obs    =      1445
No. of failures =          552
Time at risk    =        80013
                                                   LR chi2(10)      =    180.28
Log likelihood  =   -1649.7531                     Prob > chi2      =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	.0955801	.0905511	1.06	0.291	-.0818968	.2730569
priors	.091337	.0133434	6.85	0.000	.0651845	.1174896
tserved	.0144009	.0016599	8.68	0.000	.0111475	.0176543
felon	-.3122241	.1056552	-2.96	0.003	-.5193045	-.1051437
alcohol	.4676706	.1057752	4.42	0.000	.2603551	.6749862
drugs	.2941648	.0978584	3.01	0.003	.1023658	.4859637
black	.4756749	.0881961	5.39	0.000	.3028137	.6485362
married	-.1519512	.1092578	-1.39	0.164	-.3660925	.0621901
educ	-.0242123	.0194075	-1.25	0.212	-.0622503	.0138256
age	-.0039112	.0005282	-7.40	0.000	-.0049464	-.002876
_cons	-4.169213	.2756276	-15.13	0.000	-4.709433	-3.628993

上表显示，参加狱中劳动项目（workprg = 1）反而会增大再次被捕的风险，尽管统计上并不显著（p 值为 0.291）。这可能是由于劳动项目本身无效，或者存在由于样本选择问题（表现恶劣的罪犯被优先安排参加此项目）。除 married 与 educ 之外的其他变量均很显著。为了方便解释各变量的效应，下面去掉选择项“nohr”再次进行回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(e) nolog noshow
```

```
Exponential regression -- log relative-hazard form

No. of subjects =         1445                     Number of obs    =      1445
No. of failures =          552
Time at risk    =        80013
                                                   LR chi2(10)      =    180.28
Log likelihood  =   -1649.7531                     Prob > chi2      =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	1.100297	.0996331	1.06	0.291	.921367	1.313975
priors	1.095638	.0146195	6.85	0.000	1.067356	1.12467
tserved	1.014505	.001684	8.68	0.000	1.01121	1.017811
felon	.7318175	.0773203	-2.96	0.003	.5949342	.9001952
alcohol	1.596272	.1688459	4.42	0.000	1.297391	1.964006
drugs	1.342005	.1313265	3.01	0.003	1.107789	1.625741
black	1.6091	.1419164	5.39	0.000	1.353662	1.912739
married	.8590302	.0938557	-1.39	0.164	.6934387	1.064165
educ	.9760784	.0189432	-1.25	0.212	.9396476	1.013922
age	.9960964	.0005261	-7.40	0.000	.9950658	.9971282
_cons	.0154644	.0042624	-15.13	0.000	.0090099	.0265429

是否使用选择项“nohr”不影响回归的结果，只是决定汇报系数（Coef.）还是风险比率（Haz. Ratio）而已。例如，上表显示，alcohol 的风险比率约为 1.60，这意味着酗酒者比不酗酒者再次被捕的风险高 60%。由于指数回归的风险函数为常数，假定太强，故下面放松此假定，进行威布尔回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(w) nolog noshow
```

```
Weibull regression -- log relative-hazard form

No. of subjects =         1445                 Number of obs    =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)      =    165.48
Log likelihood  =   -1633.0325                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    workprg |   1.095148   .0992728     1.00   0.316     .9168814    1.308074
     priors |   1.092848    .014683     6.61   0.000     1.064445    1.122008
    tserved |   1.013655   .0017037     8.07   0.000     1.010321       1.017
      felon |   .7412054   .0785485    -2.83   0.005     .6021898    .9123128
    alcohol |   1.564179    .165389     4.23   0.000     1.271406     1.92437
      drugs |   1.325064   .1296765     2.88   0.004     1.093791    1.605237
      black |   1.574149   .1390031     5.14   0.000      1.32398    1.871587
    married |   .8593436   .0938794    -1.39   0.165     .6937084    1.064527
       educ |   .9769709   .0189724    -1.20   0.230     .9404845    1.014873
        age |   .9962823    .000523    -7.09   0.000     .9952577     .997308
      _cons |   .0333035   .0100249   -11.30   0.000     .0184613    .0600781
-------------+----------------------------------------------------------------
       /ln_p |  -.2158398   .0389149    -5.55   0.000    -.2921115   -.1395681
-------------+----------------------------------------------------------------
           p |   .8058644   .0313601                     .7466852    .8697338
         1/p |   1.240904   .0482896                     1.149777    1.339252
------------------------------------------------------------------------------
```

从上表可知，参加劳动项目（workprg）的效应依然为正且不显著。从上表倒数第三行还可知，原假设“H_0: $\ln p = 0$”对应的 p 值为 0.000，故强烈拒绝指数回归，认为应使用威布尔回归。进一步，根据上表倒数第二行，$\hat{p} = 0.806 < 1$，故风险函数随时间而递减，存在负向久期依赖。更直观地，可以通过画图来看风险函数，结果参见图 30.6。

```
. stcurve,hazard
```

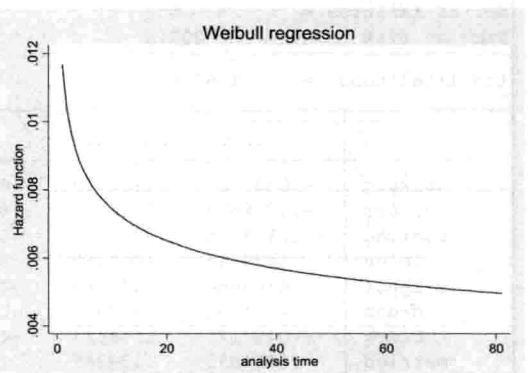


图 30.6 威布尔回归的风险函数

下面，进行指数回归的另一推广，即冈珀茨回归。

```
. streg workprg priors tserved
felon alcohol drugs black married educ age,d(gom) nolog noshow
```

```
Gompertz regression -- log relative-hazard form

No. of subjects =         1445                 Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)     =    154.39
Log likelihood  =   -1595.5837                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    workprg |   1.089068   .0988625     0.94   0.347     .9115602    1.301142
     priors |   1.090805   .0147222     6.44   0.000     1.062328    1.120045
    tserved |   1.012968    .001712     7.62   0.000     1.009618    1.016329
      felon |   .7516398   .0798318    -2.69   0.007     .6103845    .9255845
    alcohol |   1.536163   .1623417     4.06   0.000     1.248769    1.889697
      drugs |   1.313292   .1285117     2.79   0.005     1.084095    1.590945
      black |    1.54159   .1362263     4.90   0.000     1.296434    1.833105
    married |   .8581703   .0937313    -1.40   0.161     .6927928    1.063025
       educ |   .9783779   .0190234    -1.12   0.261     .9417942    1.016383
        age |   .9964432   .0005198    -6.83   0.000      .995425    .9974624
      _cons |   .0270245   .0075192   -12.98   0.000     .0156648    .0466221
-------------+----------------------------------------------------------------
     /gamma |   -.021703    .002192    -9.90   0.000    -.0259992   -.0174068
------------------------------------------------------------------------------
```

上表最后一行显示，原假设“$H_0: \gamma = 0$”的 p 值为 0.000，故强烈拒绝指数回归，认为应选择冈珀茨回归。至于在威布尔回归与冈珀茨回归之间进行选择，则可根据 AIC 准则进行。由于这两个回归的解释变量个数与参数个数均相等，故 AIC 准则等价于最大化对数似然函数，参见公式(30.32)。由于威布尔回归的对数似然值为 -1 633，而冈珀茨回归的对数似然值为 -1 596，故应选择冈珀茨回归。

其他备选的参数回归还包括属于加速失效时间模型(AFT)的对数正态回归与对数逻辑回归。首先进行对数正态回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(logn) nolog noshow
```

```
Lognormal regression -- accelerated failure-time form

No. of subjects =         1445                 Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)     =    166.74
Log likelihood  =    -1597.059                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    workprg |   -.0625714   .1200369    -0.52   0.602    -.2978394    .1726965
     priors |   -.1372528   .0214587    -6.40   0.000     -.179311   -.0951946
    tserved |   -.0193305   .0029779    -6.49   0.000    -.0251671   -.0134939
      felon |    .4439944   .1450865     3.06   0.002     .1596302    .7283586
    alcohol |   -.6349088   .1442165    -4.40   0.000    -.9175681   -.3522496
      drugs |   -.2981599   .1327355    -2.25   0.025    -.5583168   -.0380031
      black |   -.5427175   .1174427    -4.62   0.000     -.772901    -.312534
    married |    .3406835    .139843     2.44   0.015     .0665962    .6147707
       educ |    .0229195   .0253974     0.90   0.367    -.0268584    .0726975
        age |    .0039103   .0006062     6.45   0.000     .0027221    .0050984
      _cons |    4.099386   .3475349    11.80   0.000      3.41823    4.780542
-------------+----------------------------------------------------------------
     /ln_sig |    .5935861   .0344122    17.25   0.000     .5261395    .6610327
-------------+----------------------------------------------------------------
       sigma |    1.810469   .0623022                      1.692386    1.936791
------------------------------------------------------------------------------
```

从上表可知，对数正态回归的系数符号正好与比例风险模型的系数符号相反，因为二者的经济意义不同。加速失效时间模型中的系数表示该变量对平均寿命的半弹性，而比例风险模型中的系数表示对风险率的半弹性。显然，风险率越高，则平均寿命越短；故二者的作用方向正好相反。对数正态回归的对数似然值为 -1 597，略逊于冈珀茨回归。

下面，进行对数逻辑回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(logl) nolog noshow
```

```
Loglogistic regression -- accelerated failure-time form

No. of subjects =         1445                 Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)     =    178.92
Log likelihood  =   -1610.9431                 Prob > chi2     =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	-.0664501	.1195644	-0.56	0.578	-.300792	.1678918
priors	-.148287	.0230937	-6.42	0.000	-.1935498	-.1030243
tserved	-.0191425	.0029906	-6.40	0.000	-.025004	-.013281
felon	.4108081	.1436511	2.86	0.004	.1292572	.6923591
alcohol	-.6325083	.1427802	-4.43	0.000	-.9123523	-.3526643
drugs	-.3493944	.129943	-2.69	0.007	-.6040781	-.0947107
black	-.5612106	.1161105	-4.83	0.000	-.7887831	-.3336381
married	.2708632	.1387666	1.95	0.051	-.0011143	.5428407
educ	.0314568	.0253634	1.24	0.215	-.0182546	.0811682
age	.0045069	.0006628	6.80	0.000	.0032079	.005806
_cons	3.817714	.3509807	10.88	0.000	3.129805	4.505624
/ln_gam	.0347687	.0375148	0.93	0.354	-.0387589	.1082964
gamma	1.03538	.0388421			.9619826	1.114378

从上表可知，对数逻辑回归的对数似然值为 -1 610，也低于冈珀茨回归。最后，进行最具灵活性的三参数广义伽马回归。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(gam) nolog noshow
```

```
Gamma regression -- accelerated failure-time form

No. of subjects =         1445                 Number of obs    =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)      =    144.06
Log likelihood  =   -1588.8622                 Prob > chi2      =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	-.0286748	.1215211	-0.24	0.813	-.2668519	.2095023
priors	-.1403375	.0235349	-5.96	0.000	-.186465	-.0942099
tserved	-.0209601	.0036622	-5.72	0.000	-.0281378	-.0137824
felon	.5244066	.1532514	3.42	0.001	.2240394	.8247738
alcohol	-.6328808	.148593	-4.26	0.000	-.9241178	-.3416439
drugs	-.1892097	.1400017	-1.35	0.177	-.4636079	.0851885
black	-.4965192	.1206089	-4.12	0.000	-.7329084	-.26013
married	.4761969	.1433174	3.32	0.001	.1953	.7570938
educ	.0086858	.0264798	0.33	0.743	-.0432136	.0605852
age	.0029722	.0005984	4.97	0.000	.0017994	.0041451
_cons	3.969667	.3481397	11.40	0.000	3.287326	4.652009
/ln_sig	.7303791	.0375734	19.44	0.000	.6567367	.8040215
/kappa	-.8128316	.2073818	-3.92	0.000	-1.219292	-.4063708
sigma	2.075867	.0779973			1.928489	2.234509

从上表倒数第二行可知，参数 κ 的95%置信区间为(-1.22，-0.41)，故显著地不等于0(对应于对数正态回归)或1(对应于威布尔回归)，故拒绝对数正态回归与威布尔回归，认为应使用广义伽马回归。下面，画出根据广义伽马回归计算的风险函数，结果参见图30.7。

```
. stcurve,hazard
```

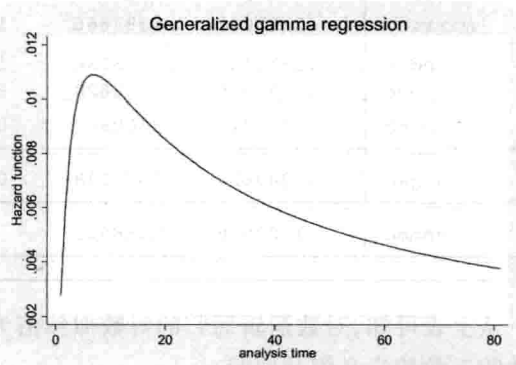


图30.7 广义伽马回归的风险函数

根据图30.7可知，广义伽马模型的风险函数首先上升然后才下降，这与威布尔模型单调递减的风险函数大不相同(参见图30.6)。最后，使用AIC信息准则在广义伽马回归与冈珀茨回归之间进行选择。首先计算广义伽马回归的AIC。

```
. dis -2*e(ll)+2*e(k)
3203.7245
```

其中，e(ll)为对数似然函数，而e(k)为模型中参数(包含解释变量系数)的个数。

然后计算冈珀茨回归的AIC。

```
. qui streg workprg priors tserved felon alcohol drugs black married
educ age,d(gom) nolog noshow
. dis -2*e(ll)+2*e(k)
```

3215.1673

由以上计算可知,广义伽马回归的 AIC 小于冈珀茨回归,在上述所有参数回归中为最小。总之,如果严格按照 AIC 准则来选择参数回归,则应使用广义伽马回归。

(4) 以上分析未考虑每位罪犯的个体异质性,而此异质性可能影响回归结果。为了与 Wooldridge(2010)的结果对比,将不可观测的异质性引入威布尔回归,并假设异质性服从伽马分布。

```
. streg workprg priors tserved felon alcohol drugs black married educ
age,d(w) frailty(gamma) nolog noshow
```

```
Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects =          1445                 Number of obs    =      1445
No. of failures =           552
Time at risk    =         80013
                                                LR chi2(10)      =    143.82
Log likelihood  =   -1584.9172                  Prob > chi2      =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	1.00741	.2053883	0.04	0.971	.6755623	1.502267
priors	1.275214	.0537558	5.77	0.000	1.17409	1.385049
tserved	1.035554	.0072673	4.98	0.000	1.021408	1.049896
felon	.4534124	.1208835	-2.97	0.003	.2688784	.7645939
alcohol	3.233478	.9070623	4.18	0.000	1.865903	5.60339
drugs	1.329452	.2968761	1.28	0.202	.8582091	2.059453
black	2.163173	.4409171	3.79	0.000	1.450748	3.225453
married	.4467732	.1151877	-3.13	0.002	.2695437	.7405342
educ	.9732451	.0436997	-0.60	0.546	.8912559	1.062777
age	.9947974	.0009922	-5.23	0.000	.9928546	.996744
_cons	.0045453	.0032737	-7.49	0.000	.0011079	.0186482
/ln_p	.5352553	.0951206	5.63	0.000	.3488225	.7216882
/ln_the	1.790243	.1788498	10.01	0.000	1.439703	2.140782
p	1.707884	.1624549			1.417398	2.057904
1/p	.5855198	.055695			.4859312	.7055184
theta	5.990906	1.071472			4.219445	8.506084

Likelihood-ratio test of theta=0: chibar2(01) = 96.23 Prob>=chibar2 = 0.000

上表最后一行显示,原假设"$H_0:\theta=0$"(θ 为异质性 v_j 的方差)的 p 值为 0.000,故强烈拒绝无异质性的原假设,认为存在异质性。下面将看到,引入异质性对于风险函数有很大影响。

```
. stcurve,hazard alpha1
```

其中,选择项"hazard alpha1"表示给定异质性 $v_j=1$(即平均异质性)条件下的风险函数,结果参见图 30.8。

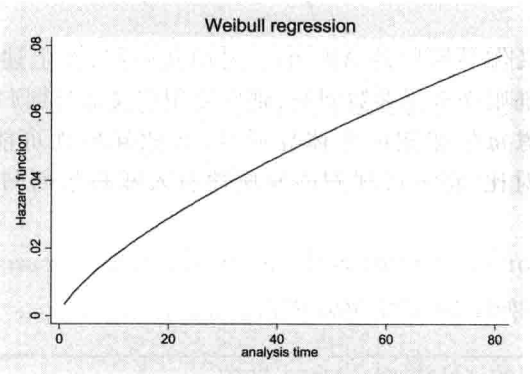


图 30.8 给定异质性的条件威布尔风险函数(Conditional Weibull hazard)

从图 30.8 可知,给定个体异质性,则风险函数单调递增,与不考虑异质性的威布尔风险函数大相径庭(参见图 30.6)。

```
. stcurve,hazard
(option unconditional assumed)
```

如果不含选择项“alpha1”,则默认设置为将个体异质性 v_j 积分平均掉,然后画无条件(unconditional)的风险函数,结果参见图 30.9。此无条件分布称为“伯尔分布”(Burr distribution)。

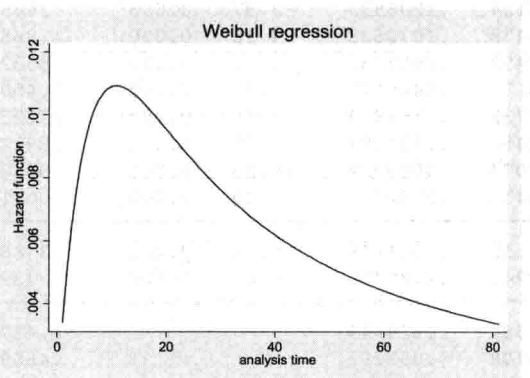


图 30.9 无条件的伯尔风险函数(Unconditional Burr hazard)

从图 30.9 可知,无条件的伯尔风险函数呈驼峰形(hump shape)。无条件的风险函数之所以先升后降,可能与样本中个体结构的变化有关。虽然条件风险函数不断上升,但由于高风险个体首先退出样本,故剩下的个体越来越多地为低风险个体。这两种相反力量的共同作用形成了驼峰形的无条件风险函数。

(5) 由于我们对于参数回归的具体分布始终没有把握,故下面进行半参数 Cox 回归。

```
. stcox workprg priors tserved felon alcohol drugs black married educ
age,r nolog noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects      =         1445                 Number of obs    =      1445
No. of failures      =          552
Time at risk         =        80013
                                                   Wald chi2(10)    =    148.14
Log pseudolikelihood =   -3816.3799                Prob > chi2      =    0.0000
```

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	1.087709	.1026332	0.89	0.373	.9040565	1.308668
priors	1.091542	.0180384	5.30	0.000	1.056754	1.127476
tserved	1.013035	.0018621	7.05	0.000	1.009392	1.016691
felon	.7537191	.0804309	-2.65	0.008	.6114712	.9290583
alcohol	1.538265	.1671288	3.96	0.000	1.243226	1.903321
drugs	1.317338	.1308915	2.77	0.006	1.08423	1.600564
black	1.541249	.1369638	4.87	0.000	1.294882	1.834489
married	.8565166	.0925895	-1.43	0.152	.6929803	1.058646
educ	.9789062	.0188643	-1.11	0.269	.9426224	1.016587
age	.9964248	.0006028	-5.92	0.000	.9952439	.997607

从上表可知，半参数 Cox 回归的结果在性质上与前面的参数回归类似，但不依赖于具体的分布假设，故更为稳健，在实践中十分流行。下面，根据 Cox 模型画风险函数，结果参见图 30.10。

```
. stcurve,hazard
```

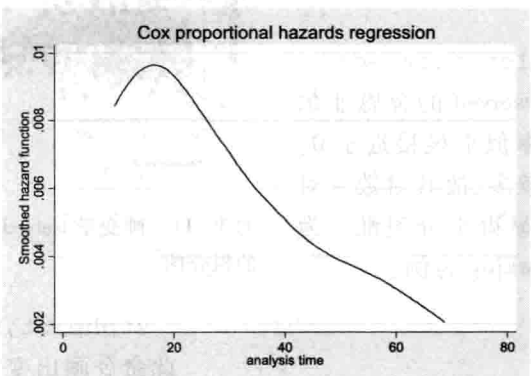


图 30.10 基于 Cox 模型的风险函数

当然，进行 Cox 回归的前提是比例风险假定，为此需进行相应的检验。首先，进行基于舍恩菲尔德残差的比例风险检验。

```
. estat phtest,detail
```

```
Test of proportional-hazards assumption

Time:  Time
```

	rho	chi2	df	Prob>chi2
workprg	0.05246	1.67	1	0.1967
priors	-0.07237	3.42	1	0.0645
tserved	-0.09395	4.14	1	0.0420
felon	0.00497	0.01	1	0.9108
alcohol	-0.04469	1.17	1	0.2802
drugs	-0.01081	0.07	1	0.7963
black	0.04290	1.02	1	0.3128
married	0.03940	0.90	1	0.3430
educ	-0.04669	1.11	1	0.2925
age	0.04627	2.11	1	0.1463
global test		15.26	10	0.1228

```
note: robust variance-covariance matrix used.
```

从上表最后一行可知，整体来看，可以接受比例风险的假定。然而，变量 tserved（狱中关押月数）可能违反比例风险假定，p 值为 0.042。更直观地，画变量 tserved 的残差与时间的拟合图，考察其斜率是否为 0，结果参见 30.11。

```
. estat phtest,plot(tserved)
```

根据图 30.11，变量 tserved 的舍恩菲尔德残差对时间的回归斜率似乎很接近于 0。由于变量 tserved 的取值较多，故其对数 - 对数图与观测 - 预测图均显得十分凌乱。为了演示的目的，以变量 workprg 为例。

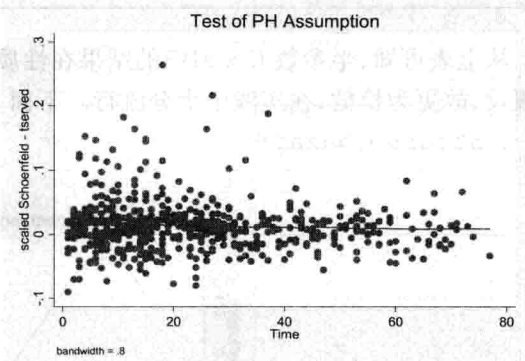


图 30.11 画变量 tserved 的舍恩菲尔德残差与时间的拟合图

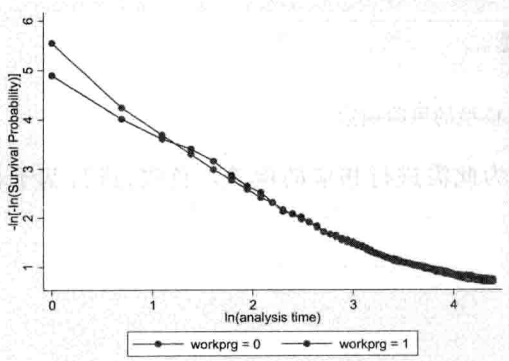


图 30.12 基于变量 workprg 的对数 - 对数图

```
. stphplot,by(workprg)
```

此命令画出变量 workprg 的对数 - 对数图，结果参见图 30.12。

尽管图 30.12 中的两条曲线有相交，二者的斜率依然大致相等，似乎不违背比例风险假定。

```
. stcoxkm,by(workprg)
```

此命令画出变量 workprg 的观测 - 预测图，结果参见图 30.13。

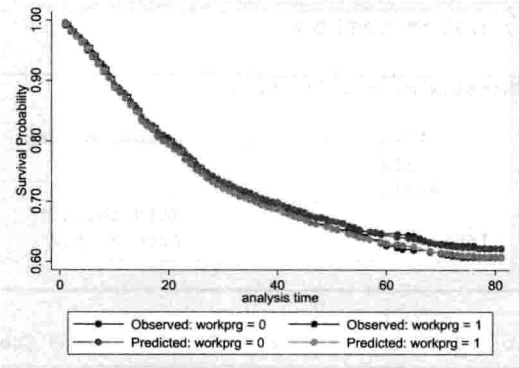


图 30.13 基于变量 workprg 的观测 - 预测图

从图 30.13 可知,生存函数的观测值与预测值相距不远,似乎与比例风险假定相一致。

最后,我们通过引入时变解释变量(TVC)的方法来检验比例风险模型。具体来说,引入变量 tserved 与时间 t 的互动项,进行 Cox 回归,然后检验此互动项的显著性。

```
. stcox workprg priors tserved felon alcohol drugs black married educ
age,tvc(tserved) r nolog noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects      =         1445                 Number of obs    =      1445
No. of failures      =          552
Time at risk         =        80013
                                                   Wald chi2(11)    =    160.50
Log pseudolikelihood =   -3814.8278                Prob > chi2      =    0.0000
```

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
main						
workprg	1.086906	.1021223	0.89	0.375	.9040987	1.306677
priors	1.091509	.0180037	5.31	0.000	1.056787	1.127372
tserved	1.016055	.0023921	6.77	0.000	1.011378	1.020755
felon	.7614949	.0811258	-2.56	0.011	.6179941	.9383173
alcohol	1.533541	.1659029	3.95	0.000	1.240537	1.895749
drugs	1.316638	.1303001	2.78	0.005	1.084495	1.598471
black	1.530788	.1358424	4.80	0.000	1.286409	1.821591
married	.8584437	.0925918	-1.42	0.157	.6948662	1.060529
educ	.9797185	.0187874	-1.07	0.285	.9435792	1.017242
age	.9964386	.0005998	-5.93	0.000	.9952636	.997615
tvc						
tserved	.999833	.0000945	-1.77	0.077	.9996479	1.000018

Note: variables in tvc equation interacted with _t

上表倒数第二行显示,互动项 tserved × t 的 p 值为 0.077,故可在 5% 的水平上认为该互动项不显著,支持比例风险假定。

(6) 上面的分析显示,变量 tserved 可能违反了比例风险假定。解释方法之一为以变量 tserved 分层变量,估计分层 Cox 模型。

```
. stcox workprg priors tserved felon alcohol drugs black married educ
age,strata(tserved) r nolog noshow
```

Stratified Cox regr. -- Breslow method for ties

No. of subjects	= 1445	Number of obs	= 1445
No. of failures	= 552		
Time at risk	= 80013		
		Wald chi2(9)	= 95.27
Log pseudolikelihood	= -1668.4461	Prob > chi2	= 0.0000

_t	Haz. Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	1.007903	.0985513	0.08	0.936	.832127	1.22081
priors	1.091654	.0184789	5.18	0.000	1.05603	1.12848
tserved	1.039889	.	.	.	.	.
felon	.6944598	.0807721	-3.13	0.002	.5528974	.8722676
alcohol	1.524112	.1771505	3.63	0.000	1.213612	1.914051
drugs	1.273412	.1307645	2.35	0.019	1.041263	1.557318
black	1.456758	.1311006	4.18	0.000	1.221192	1.737766
married	.8387165	.0943829	-1.56	0.118	.6727091	1.04569
educ	.9839777	.0194734	-0.82	0.414	.9465412	1.022895
age	.9969553	.0005942	-5.12	0.000	.9957914	.9981205

Stratified by tserved

上表显示，分层 Cox 模型的估计结果与标准 Cox 模型类似，也说明变量 tserved 违反比例风险假定的嫌疑不严重。

习　题

30.1 假设寿命服从指数分布。证明个体在$(0,t_2)$区间死亡的概率等于在已知个体存活至时刻t_1的情况下，其在(t_1,t_1+t_2)区间死亡的概率。

30.2 使用数据集 kennan－1985. dta 进行久期分析。Kennan(1985)使用此数据集研究美国制造业 1968—1976 年的罢工持续时间，主要变量包括 t(罢工持续时间)与 prod(产量对数，已作季节调整并去掉时间趋势)。

30.3 使用数据集 ema1996. dta 分析失业持续时间(spell)的决定因素。McCall(1996)使用此数据集分析失业保险制度(失业者可获得一定上限的兼职收入而不影响领取失业救济)对失业时间的影响。

第 31 章　贝叶斯估计简介

31.1　贝叶斯估计的思想

在统计学中有两派，占主流的是“频率学派”（frequentist school），也称“古典学派”（Classical school），即通常数理统计课程中的那些内容；另外一派则是“贝叶斯学派”（Bayesian school），由 18 世纪英国统计学家贝叶斯（Thomas Bayes）创立。二者的主要区别在于，频率学派假设总体服从某个分布，比如 $f(x;\theta)$，其中 θ 为待估计、未知、给定的参数（或参数向量）；而贝叶斯学派则认为，既然 θ 有不确定性，则应将 θ 本身也视为随机变量（或随机向量），并用概率分布来描述。由于 θ 的这个分布在研究者看到数据之前就有，故称为“先验分布”（prior distribution）。得到样本数据后，可以根据贝叶斯定理（Bayes' Theorem）将先验分布更新（update）为后验分布（posterior distribution），并以后验分布作为统计推断的依据。

由于先验分布的主观性（subjectivity），有些学者对贝叶斯学派持有保留意见。然而，在后面我们将看到，在大样本中，先验分布的作用将变得很小。另外，可以使用“相对不含信息”（relatively uninformative）的先验分布，并对后验分布对于先验分布的依赖性进行“敏感度分析”（sensitivity analysis）。

贝叶斯估计的主要优点包括以下几点。

（1）古典学派一般通过最优化（比如，MLE，OLS）进行参数估计，但有时不易求得最优解。而贝叶斯学派只要反复使用贝叶斯定理即可，不需要进行最优化计算。虽然贝叶斯分析也常常没有解析解，但随着近几十年计算方法的迅猛发展，这已经基本不成问题。

（2）古典学派需要用不同的统计量来估计期望、方差、中位数、分位数等，而贝叶斯学派可以直接得到参数的整个后验分布。从这个后验分布，可以容易地算出其各阶矩。

（3）对于古典学派的统计量，常常不易找出其“精确的有限样本分布”（exact finite-sample distribution），故只能退而求其次，推导大样本渐近分布。而贝叶斯学派一般可以直接计算精确的有限样本分布，不需要渐近理论。

31.2　贝叶斯定理

贝叶斯估计的实质就在于，反复使用贝叶斯定理[①]，将先验分布与样本数据综合为后验分布。对于随机事件 A 与 B，有如下的贝叶斯公式：

---

① 其应用极为广泛，可以说“就靠这一招，打遍天下”。

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{P(B|A)\cdot P(A)}{P(B)} \tag{31.1}$$

其中，第一个等号为条件概率的定义，而第二个等号使用了概率的乘法公式。如果将 $P(A)$ 视为先验概率，将 B 视为样本数据，则贝叶斯公式给出了在看到样本数据 B 后，如何将先验概率 $P(A)$ 更新为后验概率 $P(A|B)$ 的规则。

更一般地，对于随机向量 $\boldsymbol{\theta}$（视为参数）与随机向量 $\boldsymbol{y}$（视为样本数据），根据贝叶斯定理（Bayes' Theorem）可知，

$$f(\boldsymbol{\theta}|\boldsymbol{y})=\frac{f(\boldsymbol{\theta},\boldsymbol{y})}{f(\boldsymbol{y})}=\frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})} \tag{31.2}$$

其中，$f(\boldsymbol{\theta}|\boldsymbol{y})$ 为看到数据 $\boldsymbol{y}$ 之后 $\boldsymbol{\theta}$ 的条件分布密度（即后验分布），$\pi(\boldsymbol{\theta})$ 为参数 $\boldsymbol{\theta}$ 的先验分布密度，$f(\boldsymbol{\theta},\boldsymbol{y})$ 为 $\boldsymbol{\theta}$ 与 $\boldsymbol{y}$ 的联合分布，$f(\boldsymbol{y}|\boldsymbol{\theta})$ 为给定参数 $\boldsymbol{\theta}$ 时 $\boldsymbol{y}$ 的密度函数，而 $f(\boldsymbol{y})$ 为 $\boldsymbol{y}$ 的边缘分布密度。在联合分布 $f(\boldsymbol{\theta},\boldsymbol{y})$ 中将随机参数 $\boldsymbol{\theta}$ 积分掉，就可得到 $\boldsymbol{y}$ 的边缘分布密度：

$$f(\boldsymbol{y})=\int f(\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{\theta}=\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \tag{31.3}$$

在贝叶斯分析中，一般把后验分布 $f(\boldsymbol{\theta}|\boldsymbol{y})$ 记为 $p(\boldsymbol{\theta}|\boldsymbol{y})$（$p$ 表示 posterior），而把 $\boldsymbol{y}$ 的密度函数 $f(\boldsymbol{y}|\boldsymbol{\theta})$ 记为似然函数 $L(\boldsymbol{\theta};\boldsymbol{y})$（将其视为 $\boldsymbol{\theta}$ 的函数，而把数据 $\boldsymbol{y}$ 视为给定）。在公式(31.2)中，分母为边缘分布 $f(\boldsymbol{y})$，不包含 $\boldsymbol{\theta}$，故可以将其视为常数[①]。因此，后验分布与该公式的分子成正比：

$$\propto \bar{h}\,(\theta-\bar{\mu})^2 \tag{31.4}$$

其中，“$\propto$”表示“成正比”。省略常数可以简化后验分布的推导，而被省略的常数可在以后加上[②]。省去常数的密度函数被称为“密度核”（density kernel）。因此，$L(\boldsymbol{\theta};\boldsymbol{y})\pi(\boldsymbol{\theta})$ 就是后验分布 $p(\boldsymbol{\theta}|\boldsymbol{y})$ 的密度核。

31.3 贝叶斯估计的一个例子

记随机样本为 $\boldsymbol{y}=(y_1y_2\cdots y_n)'$，其中 $y_i\sim N(\theta,\sigma^2)$，方差 σ^2 已知，而均值 θ 未知。古典学派通常选择 θ 使似然函数最大化，得到 $\hat{\theta}_{\mathrm{MLE}}=\bar{y}\equiv\frac{1}{n}\sum_{i=1}^{n}y_i$。贝叶斯学派则要额外地设定 θ 的先验分布。为了计算方便，选择先验正态分布[③]，即 $\theta\sim N(\mu,\tau^2)$，其中先验均值 μ 与先验方差 τ^2 为已知常数。如果 τ^2 较大，就表示先验分布的不确定性较大。我们的目标是求出后验分布 $p(\theta|\boldsymbol{y})$。可以把此计算过程分为以下三步。

第一步 写出先验分布密度 $\pi(\theta)$。

在贝叶斯分析中，使用方差的倒数有时更为方便。定义 θ 的“精确度”（precision）为

$$h\equiv 1/\tau^2 \tag{31.5}$$

显然，精确度 h 越大，方差 τ^2 就越小，表明我们对随机变量 θ 知道得越精确。类似地，由于样本均值 $\bar{y}$ 的方差为 (σ^2/n)，记 $\bar{y}$ 的精确度为 $h^*\equiv n/\sigma^2$。θ 的先验密度为

---

① 得到样本数据 $\boldsymbol{y}=(y_1y_2\cdots y_n)'$ 后，$\boldsymbol{y}$ 已经确定，可以视为常数。

② 它本质上是一个积分常数，即保证密度函数的积分为1。

③ 有关如何选择先验分布，参见下文。

$$\pi(\theta)=(2\pi\tau^2)^{-1/2}\exp\{-(\theta-\mu)^2/2\tau^2\}\quad(\text{一元正态密度})$$
$$\propto\exp\{-h(\theta-\mu)^2/2\}\quad(\text{去掉不含}\theta\text{的常数项,代入}h\equiv1/\tau^2)$$

第二步 写出样本数据 $\boldsymbol{y}$ 的联合密度,即似然函数:

$$L(\theta;\boldsymbol{y})=\prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\{-(y_i-\theta)^2/2\sigma^2\}\quad(n\text{个一元正态密度函数相乘})$$
$$=(2\pi\sigma^2)^{-n/2}\exp\left\{-\sum_{i=1}^{n}(y_i-\theta)^2/2\sigma^2\right\}\quad(\text{合并同类项})$$
$$\propto\exp\left[-\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\theta)^2/2\sigma^2\right]\quad(\text{去掉常数项,加减}\bar{y})$$
$$\propto\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\bar{y})(\bar{y}-\theta)+\sum_{i=1}^{n}(\bar{y}-\theta)^2\right]\right\}(\text{展开平方项})$$
$$\propto\exp\left\{-\sum_{i=1}^{n}(\bar{y}-\theta)^2/2\sigma^2\right\}$$
$$\left(\text{去掉常数}\sum_{i=1}^{n}(y_i-\bar{y})^2,\text{而}\sum_{i=1}^{n}(y_i-\bar{y})(\bar{y}-\theta)=(\bar{y}-\theta)\sum_{i=1}^{n}(y_i-\bar{y})=0\right)$$
$$\propto\exp\{-n(\bar{y}-\theta)^2/2\sigma^2\}\quad(n\text{个相同的数相加变为乘法})$$
$$\propto\exp\{-h^*(\bar{y}-\theta)^2/2\}\quad(\text{定义}h^*\equiv n/\sigma^2)$$

注意到,似然函数的形式与先验密度相似,这为后面的计算提供了方便。

第三步 根据贝叶斯定理,写出后验分布的密度核。

$$L(\theta;\boldsymbol{y})\pi(\theta)\propto\exp\{-h^*(\bar{y}-\theta)^2/2\}\cdot\exp\{-h(\theta-\mu)^2/2\}$$
$$\propto\exp\left\{-\frac{1}{2}[h^*(\bar{y}-\theta)^2+h(\theta-\mu)^2]\right\}\quad(\text{指数相加})$$

其中,指数项中的方括弧项可以简化为

$$h^*(\bar{y}-\theta)^2+h(\theta-\mu)^2$$
$$=h^*(\bar{y}^2-2\bar{y}\theta+\theta^2)+h(\theta^2-2\mu\theta+\mu^2)\quad(\text{展开平方项})$$
$$=(h+h^*)\theta^2-2(h\mu+h^*\bar{y})\theta+h^*\bar{y}^2+h\mu^2\quad(\text{根据}\theta\text{合并同类项})$$
$$=\bar{h}\theta^2-2\bar{h}\bar{\mu}\theta+h^*\bar{y}^2+h\mu^2\quad\left(\text{定义}\bar{h}\equiv h+h^*,\bar{\mu}\equiv\frac{h\mu+h^*\bar{y}}{\bar{h}}\right)$$
$$=\bar{h}(\theta^2-2\bar{\mu}\theta)+h^*\bar{y}^2+h\mu^2\quad(\text{提取公因子}\bar{h})$$
$$=\bar{h}(\theta^2-2\bar{\mu}\theta+\bar{\mu}^2)-\bar{h}\bar{\mu}^2+h^*\bar{y}^2+h\mu^2\quad(\text{配方})$$
$$\propto\bar{h}(\theta-\bar{\mu})^2\quad(\text{去掉不含}\theta\text{的常数项})$$

根据以上推导可知,

$$L(\theta;\boldsymbol{y})\pi(\theta)\propto\exp\left\{-\frac{1}{2}[\bar{h}(\theta-\bar{\mu})^2]\right\}\tag{31.6}$$

显然,这还是一个正态分布的密度核,故后验分布仍是正态分布。后验分布的精确度提高为

$$\bar{h}\equiv h+h^*\tag{31.7}$$

即先验精确度 h 与样本精确度 h^* 之和。而后验分布的期望值调整为

$$\bar{\mu}\equiv(h\mu+h^*\bar{y})/\bar{h}\tag{31.8}$$

即先验均值 μ 与样本均值 $\bar{y}$ 之加权平均,权重为各自的精确度。如果先验信息不精确,即 h 较小,则先验均值对于后验均值的影响就较小。方程(31.7)与(31.8)给出了如何将“先验信息”与“样本信息”综合成“后验信息”的规则,即“贝叶斯更新规则”(Bayesian updating rule)。

由于 $\bar{h}=h+h^*=h+\frac{n}{\sigma^2}$,故当 $n\to\infty$ 时,后验精确度 $\bar{h}\to\infty$(后验方差趋于 0);而且样本精确

度对后验精确度的贡献越来越大，即 $h^*/\overline{h}\to 1$。因此，当 $n\to\infty$ 时，后验均值 $\overline{\mu}\to\overline{y}$，即完全由样本数据决定，不受先验分布的影响。直观来看[①]，当 $n\to\infty$ 时，后验分布 $\theta\mid \boldsymbol{y}\xrightarrow{d} N(\overline{y},\sigma^2/n)$。对比古典学派的结果则为 $\overline{y}\xrightarrow{d} N(\theta,\sigma^2/n)$。

总之，当样本容量越来越大时，先验分布所起的作用越来越小。这个结论在一般情况下也成立。回到后验分布密度核的一般表达式：

$$\begin{aligned} p(\boldsymbol{\theta}\mid \boldsymbol{y}) &\propto L(\boldsymbol{\theta};\boldsymbol{y})\pi(\boldsymbol{\theta}) \\ &\propto L(\boldsymbol{\theta};y_1,\cdots,y_n)\pi(\boldsymbol{\theta}) \\ &\propto L(\boldsymbol{\theta};y_1)\cdots L(\boldsymbol{\theta};y_n)\pi(\boldsymbol{\theta}) \quad \text{（假设样本为独立同分布的）} \end{aligned}$$

因此，后验分布密度核的对数为

$$\ln p(\boldsymbol{\theta}\mid \boldsymbol{y})\propto \ln \pi(\boldsymbol{\theta})+\sum_{i=1}^{n}\ln L(\boldsymbol{\theta};y_i) \tag{31.9}$$

当样本容量增大时，对数先验密度 $\ln \pi(\boldsymbol{\theta})$ 始终保持不变，而对数似然函数之和 $\sum_{i=1}^{n}\ln L(\boldsymbol{\theta};y_i)$ 包含的项数越来越多，故越来越以后者为主。这个结论有助于冲淡人们对于贝叶斯估计结果依赖于主观先验分布的顾虑。另外，得到后验分布之后，可以把它视为“修正的先验分布”(revised prior)，并作为未来的先验分布，如此反复，不断更新我们对世界的认识。

31.4 基于后验分布的统计推断

贝叶斯分析的主要结果为后验分布。有了后验分布，就可以进行一系列的统计推断。

1. 边缘后验分布

假设参数 $\boldsymbol{\theta}$ 为向量，$\boldsymbol{\theta}=(\theta_1\cdots\theta_q)'$。在知道 $\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta}\mid \boldsymbol{y})$ 之后，可以求得单个参数 $\theta_k(1\leqslant k\leqslant q)$ 的“边缘后验分布”(marginal posterior)：

$$p(\theta_k\mid \boldsymbol{y})=\int \mathrm{p}(\boldsymbol{\theta}\mid \boldsymbol{y})\mathrm{d}\theta_1\cdots\mathrm{d}\theta_{k-1}\mathrm{d}\theta_{k+1}\cdots\mathrm{d}\theta_q \tag{31.10}$$

即只要把其余的参数 $(\theta_1,\cdots,\theta_{k-1}\theta_{k+1},\cdots,\theta_q)$ 积分掉就行。边缘后验分布常常既不对称也非“单峰”(unimodal)，这与古典学派的统计量不同。

2. 后验分布的各阶矩

根据后验分布，可以计算其各阶矩，比如均值、中位数、方差等。

3. 点估计

在贝叶斯分析中，未知参数 θ 被视为随机变量，而非一个固定的点，故点估计(point estimation)在贝叶斯分析中不那么重要，关注的重点是 θ 的整个后验分布。另一方面，后验均值(posterior mean)或后验中位数(posterior median)常常被作为点估计来汇报。

4. 区间估计

知道参数 θ 的整个后验分布后，很容易找到置信区间或置信区域(confidence interval or region)。但置信度为 $(1-\alpha)$(比如，$\alpha=5\%$)的置信区间并不唯一。一个简单的做法是选择分位数 $\alpha/2$ 与分位数 $(1-\alpha/2)$ 之间的区间。对贝叶斯置信区间的解释与频率派的解释完全不同。

---

① 这只是一种不严格的直观写法，因为当 $n\to\infty$ 时，$\sigma^2/n\to 0$。要得到非退化的渐近分布，则要乘 $\sqrt{n}$。

假设 θ 的95%后验置信区间为(1,3),则可以直接说 θ 落在区间(1,3)的概率为95%。而对于频率派而言,如果 θ 的95%置信区间为(1,3),则只能说,如果进行100次同样的随机抽样,大约有95次,这些随机的置信区间能覆盖真实的参数值 θ。

5. 假设检验

由于贝叶斯分析认为参数不是固定的,故像"$H_0:\theta=\theta_0$"这类的假设检验通常也就没有意义(假设 θ 为连续变量)。但可以检验参数 θ 是否属于参数空间 Θ 的某个子集 Θ_1,比如检验原假设"$H_0:\theta\in\Theta_1\subset\Theta$",而替代假设为"$H_1:\theta\in\Theta_2\subset\Theta$",其中 Θ_1 与 Θ_2 可以有交集,即允许"非嵌套式"(non-nested)检验。在古典学派中,"$H_0:\theta\in\Theta_1$"成立的概率要么为0(伪命题),要么为1(真命题)。而贝叶斯学派则直接计算"$H_0:\theta\in\Theta_1$"成立的后验概率,即

$$\mathrm{P}(H_0\mid\boldsymbol{y})=\frac{\mathrm{P}(\boldsymbol{y}\mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(\boldsymbol{y})} \tag{31.11}$$

其中,$\boldsymbol{y}$ 为样本数据。同样地,替代假设"$H_1:\theta\in\Theta_2$"成立的后验概率为

$$\mathrm{P}(H_1\mid\boldsymbol{y})=\frac{\mathrm{P}(\boldsymbol{y}\mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(\boldsymbol{y})} \tag{31.12}$$

两个假设的后验概率之比称为"后验几率比"(posterior odds ratio):

$$\underbrace{\frac{\mathrm{P}(H_0\mid\boldsymbol{y})}{\mathrm{P}(H_1\mid\boldsymbol{y})}}_{\text{后验几率比}}=\underbrace{\frac{\mathrm{P}(\boldsymbol{y}\mid H_0)}{\mathrm{P}(\boldsymbol{y}\mid H_1)}}_{\text{贝叶斯因子}}\cdot\underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{先验几率比}} \tag{31.13}$$

其中,$B_{01}\equiv\dfrac{\mathrm{P}(\boldsymbol{y}\mid H_0)}{\mathrm{P}(\boldsymbol{y}\mid H_1)}$称为"贝叶斯因子"(Bayes factor),而$\dfrac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}$称为"先验几率比"(prior odds ratio)。上式表明,后验几率比等于贝叶斯因子与先验几率比之乘积。如果后验几率比大于1,则数据显示我们更应相信 H_0 而不是 H_1。总之,贝叶斯假设检验的实质是,计算不同假设的后验概率,然后进行比较。这一点与古典假设检验的小概率事件的"反证法"思想很不相同。

31.5 先验分布的选择

1. 无信息先验分布

与古典学派相比,贝叶斯分析面临的一个主要挑战是如何确定先验分布(specification of the prior)。由于先验分布是贝叶斯分析中主观性的来源,为了减少主观性,一种方法是选择基本不带信息的先验分布,即"无信息先验分布"(uninformative prior),以减少先验分布对后验分布的影响[①]。

如果参数空间 Θ 是有界的,则可以使用均匀分布来作为无信息先验分布,即 $\pi(\boldsymbol{\theta})=c>0$, $\forall\boldsymbol{\theta}\in\Theta$,这被称为"均匀先验分布"(uniform prior)。但如果参数空间 Θ 是无界的,则均匀先验分布就是"非正常密度函数"(improper density),因为 $\int\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=\infty$而不是1。这可能导致后验分布也是非正常密度函数。

均匀先验分布的另一缺点是,在作参数变换(reparameterization)时,不满足不变性(invariance)。比如,假设一维参数 $\theta>0$ 服从均匀先验分布,即 $\pi(\theta)=c>0$。考虑参数变换$\gamma=$

① 另一降低主观性的方法是增大样本容量。

$\ln\theta$,则 $-\infty<\gamma<\infty$。根据随机变量函数的概率分布公式可知,γ 的先验密度为 $\pi^*(\gamma)=\pi(\theta)\left|\dfrac{\mathrm{d}\theta}{\mathrm{d}\gamma}\right|=ce^{\gamma}$,故不再是均匀分布。如此看来,似乎无信息的参数 θ 经过参数转换变成 γ 后,却带有明显的信息了。

为了避免均匀先验分布的非正常密度函数而同时维持很少的信息,可以假设先验分布的方差很大。比如,$\theta\sim N(\mu,\tau^2)$,其中 τ^2 很大。这种先验分布被称为"模糊先验分布"或"弥漫先验分布"(vague or diffuse prior),但它在参数变换下也不满足不变性。

在实践中,较常用的无信息先验分布是 Jeffreys(1946)提出的如下"杰佛里先验分布"(Jeffreys' prior):

$$\pi(\boldsymbol{\theta})\propto|\mathrm{I}(\boldsymbol{\theta})|^{1/2}=\left|-\mathrm{E}\left(\frac{\partial^2\ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\right|^{1/2}\tag{31.14}$$

从上式可知,杰佛里先验分布的密度核是信息矩阵 $\mathrm{I}(\boldsymbol{\theta})$ 的行列式的平方根,其中 $L\equiv L(\boldsymbol{\theta};\boldsymbol{y})$ 为似然函数。杰佛里先验分布的优点是,它具有不变性,即无论作什么参数变换,先验分布的形式不变。下面以一维情形为例来证明其不变性。

证明:对于参数 θ,其杰佛里先验分布为 $\pi(\theta)\propto|\mathrm{I}(\theta)|^{1/2}$。假设参数变换为 $\gamma=h(\theta)$,则需要证明,γ 的杰佛里先验分布为 $\pi^*(\gamma)\propto|\mathrm{I}(\gamma)|^{1/2}$。根据随机变量函数的分布公式,

$$\pi^*(\gamma)=\pi(\theta)\left|\frac{\partial\theta}{\partial\gamma}\right|\propto|\mathrm{I}(\theta)|^{1/2}\left|\frac{\partial\theta}{\partial\gamma}\right|\tag{31.15}$$

因此,只要证明 $|\mathrm{I}(\gamma)|^{1/2}=|\mathrm{I}(\theta)|^{1/2}\left|\dfrac{\partial\theta}{\partial\gamma}\right|$ 即可。为此,计算 $\mathrm{I}(\gamma)=-\mathrm{E}\left[\dfrac{\partial^2\ln L}{\partial\gamma^2}\right]$。

$$\frac{\partial\ln L}{\partial\gamma}=\frac{\partial\ln L}{\partial\theta}\cdot\frac{\partial\theta}{\partial\gamma}\quad(\text{使用求导连锁法则求一阶导数})$$

$$\frac{\partial^2\ln L}{\partial\gamma^2}=\frac{\partial^2\ln L}{\partial\theta^2}\cdot\left(\frac{\partial\theta}{\partial\gamma}\right)^2+\frac{\partial\ln L}{\partial\theta}\cdot\frac{\partial^2\theta}{\partial\gamma^2}\quad(\text{使用求导连锁法则求二阶导数})$$

其中,$\dfrac{\partial\ln L}{\partial\theta}$ 为得分函数,其期望值为 $\mathrm{E}\left[\dfrac{\partial\ln L}{\partial\theta}\right]=0$(参见第6章)。因此,

$$\mathrm{I}(\gamma)=-\mathrm{E}\left[\frac{\partial^2\ln L}{\partial\gamma^2}\right]=-\mathrm{E}\left[\frac{\partial^2\ln L}{\partial\theta^2}\right]\left(\frac{\partial\theta}{\partial\gamma}\right)^2=\mathrm{I}(\theta)\left(\frac{\partial\theta}{\partial\gamma}\right)^2\tag{31.16}$$

对方程(31.16)两边取行列式[①],然后开平方可得

$$|\mathrm{I}(\gamma)|^{1/2}=|\mathrm{I}(\theta)|^{1/2}\left|\frac{\partial\theta}{\partial\gamma}\right|\tag{31.17}$$

因此,$\pi^*(\gamma)\propto|\mathrm{I}(\gamma)|^{1/2}$,故杰佛里先验分布具备不变性。

例 正态分布的杰佛里先验分布。假设 $y\sim N(\mu,\sigma^2)$。

情形一,μ 未知而 σ^2 已知,则 $\mathrm{I}(\mu)=-\mathrm{E}\left[\dfrac{\partial^2\ln L}{\partial\mu^2}\right]=\dfrac{n}{\sigma^2}$(参见习题)。由于 σ^2 已知,故杰佛里先验分布为 $|\mathrm{I}(\mu)|^{1/2}\propto c$,其中 c 为常数。此时,杰佛里先验分布为非正常密度函数。

情形二,σ^2 未知而 μ 已知,则 $\mathrm{I}(\sigma^2)=-\mathrm{E}\left[\dfrac{\partial^2\ln L}{\partial(\sigma^2)^2}\right]=\dfrac{n}{2\sigma^4}$(参见习题),故杰佛里先验分布为 $|\mathrm{I}(\sigma^2)|^{1/2}\propto\sigma^{-2}$。

---

① 常数的行列式就是常数本身。

情形三，μ 与 σ^2 皆未知，则 $\mathrm{I}(\mu,\sigma^2)=-\mathrm{E}\begin{bmatrix}\dfrac{\partial^2 \ln L}{\partial \mu^2} & \dfrac{\partial^2 \ln L}{\partial \mu \partial(\sigma^2)}\\ \dfrac{\partial^2 \ln L}{\partial \mu \partial(\sigma^2)} & \dfrac{\partial^2 \ln L}{\partial(\sigma^2)^2}\end{bmatrix}=\begin{bmatrix}\dfrac{n}{\sigma^2} & 0\\ 0 & \dfrac{n}{2\sigma^4}\end{bmatrix}=\dfrac{n^2}{2\sigma^6}$（参见习题），故杰佛里先验分布为 $|\mathrm{I}(\mu,\sigma^2)|^{1/2}\propto\sigma^{-3}$。

在此例中，在第一种情形下，杰佛里先验分布为均匀分布，但在后两种情形下都不是均匀分布。因此，究竟杰佛里先验分布在何种意义上是无信息先验分布并不清楚。

2. 自然共轭先验分布

为了计算方便，对于不同的样本密度函数（sample density）常常选择合适的先验分布，使得样本密度函数（即似然函数）、先验分布与后验分布都“属于同一族的密度函数”（in the same class of densities），即具有同样的函数形式。这种先验分布被称为“自然共轭先验分布”（natural conjugate prior），而这样的先验分布与样本密度函数被称为“自然共轭对”（natural conjugate pair）。

比如，在前面的例子中，方差已知而均值未知的正态分布，其自然共轭先验分布就是正态分布。此时，由于样本密度函数与先验分布的密度核在形式上都是指数函数，故把它们相乘时，得到的后验分布仍然具有指数函数的形式。

除了计算方便（可能得到解析解），使用共轭先验分布的另一好处是，由于后验分布与先验分布的形式相同，故容易把后验分布作为下一轮估计的先验分布。

常见的共轭先验分布参见 Cameron and Trivedi(2005, p. 428)。

3. 敏感度分析

由于贝叶斯分析的结果依赖于先验分布，而我们对先验分布的具体形式可能并无把握，故常需要进行敏感度分析，即考虑使用不同的先验分布对后验分布的影响有多大；或考虑使用不同的样本密度函数的影响。

31.6 多元回归的贝叶斯分析

考虑以下的多元回归模型：

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \tag{31.18}$$

其中，$\boldsymbol{X}$ 为 $n\times K$ 满列秩数据矩阵，扰动项 $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$。因此，给定 $(\boldsymbol{X},\boldsymbol{\beta},\sigma^2)$，条件分布 $\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta},\sigma^2\sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$。

以无信息的杰佛里先验分布为例。根据前面的例子，对于 $y\sim N(\mu,\sigma^2)$，当 μ 未知而 σ^2 已知时，μ 的杰佛里先验分布为常数；而当 σ^2 未知而 μ 已知时，σ^2 的杰佛里先验分布与 $(1/\sigma^2)$ 成正比。推广到多元回归的情形，则意味着 $\pi(\beta_j)\propto c, j=1,\cdots,K$，而 $\pi(\sigma^2)\propto 1/\sigma^2$。假设 $\boldsymbol{\beta}$ 与 σ^2 相互独立，则 $(\boldsymbol{\beta},\sigma^2)$ 的联合先验分布（joint prior）为

$$\pi(\boldsymbol{\beta},\sigma^2)\propto 1/\sigma^2 \tag{31.19}$$

样本数据的似然函数可以写为

$$L(\boldsymbol{\beta},\sigma^2;\boldsymbol{y},\boldsymbol{X})=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right\} \tag{31.20}$$

定义最小二乘估计量为 $\hat{\boldsymbol{\beta}}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$，残差向量 $\boldsymbol{e}\equiv\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$。由于 $\boldsymbol{y}=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}$，故 $\boldsymbol{y}-\boldsymbol{X\beta}=$

$\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}-\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{e}-\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$。由此可得

$$L(\boldsymbol{\beta},\sigma^2;\boldsymbol{y},\boldsymbol{X})=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{e}'-(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'][\boldsymbol{e}-\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}$$

$$=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{e}'\boldsymbol{e}+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}\ (\text{因为 } \boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0})$$

$$\propto(\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}$$

(其中,样本方差 $s^2\equiv\boldsymbol{e}'\boldsymbol{e}/(n-K)$)

因此,联合后验分布满足

$$p(\boldsymbol{\beta},\sigma^2\mid\boldsymbol{y},\boldsymbol{X})\propto(1/\sigma^2)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}\underbrace{(1/\sigma^2)}_{\text{先验}}$$

$$\propto(1/\sigma^2)^{\frac{n}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}$$

$$\propto(1/\sigma^2)^{\frac{n}{2}+1}\exp\left\{-\frac{(n-K)s^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}$$

$$\propto(1/\sigma^2)^{K/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}$$

$$\times(1/\sigma^2)^{\frac{n-K}{2}+1}\exp\{-(n-K)s^2/2\sigma^2\}\quad(\text{拆成两部分,第二部分不含}\boldsymbol{\beta})$$

从上式可以看出,给定 σ^2,则 $\boldsymbol{\beta}$ 的条件后验分布(conditional posterior)为

$$p(\boldsymbol{\beta}\mid\sigma^2,\boldsymbol{y},\boldsymbol{X})\propto(1/\sigma^2)^{K/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}\qquad(31.21)$$

这是一个 K 维正态分布,其期望为 $\boldsymbol{\beta}$,方差为 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$。要得到 $\boldsymbol{\beta}$ 的边缘后验分布(marginal posterior),则需要将联合后验分布 $p(\boldsymbol{\beta},\sigma^2\mid\boldsymbol{y},\boldsymbol{X})$ 中的 σ^2 积分积掉:

$$p(\boldsymbol{\beta}\mid\boldsymbol{y},\boldsymbol{X})\propto\int_0^\infty(1/\sigma^2)^{\frac{n}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}\mathrm{d}\sigma^2$$

$$\propto\int_\infty^0 z^{\frac{n}{2}+1}\exp\left\{-\frac{z}{2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}(-z^{-2})\mathrm{d}z$$

(作积分变换 $z\equiv1/\sigma^2,\sigma^2=1/z,\mathrm{d}\sigma^2=-(1/z^2)\mathrm{d}z$)

$$\propto\int_0^\infty z^{\frac{n}{2}-1}\exp\left\{-\frac{z}{2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}\mathrm{d}z$$

上面这个表达式在形式上非常接近于 Γ 函数的定义式,即 $\Gamma(c)\equiv\int_0^\infty z^{c-1}\mathrm{e}^{-z}\mathrm{d}z$。可以证明①,

$$\int_0^\infty z^c\mathrm{e}^{-az}\mathrm{d}z=\Gamma(c+1)/a^{c+1}\qquad(31.22)$$

在方程(25.22)中,令 $c=\frac{n}{2}-1$,$a=-\frac{1}{2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]$,可得

$$p(\boldsymbol{\beta}\mid\boldsymbol{y},\boldsymbol{X})\propto\Gamma(n/2)\left\{-\frac{1}{2}[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\right\}^{-n/2}$$

$$\propto[(n-K)s^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]^{-n/2}\qquad(\text{去掉常数项})$$

---

① 作积分变换 $t\equiv az$,则 $\int_0^\infty z^c\mathrm{e}^{-az}\mathrm{d}z=\int_0^\infty(t/a)^c\mathrm{e}^{-t}\frac{1}{a}\mathrm{d}t=\frac{1}{a^{c+1}}\int_0^\infty t^c\mathrm{e}^{-t}\mathrm{d}t=\frac{\Gamma(c+1)}{a^{c+1}}$。

$$\propto \left\{ (n-K)s^2 \left[1 + \frac{1}{(n-K)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' (s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right] \right\}^{-n/2} \quad \text{(提取因子)}$$

$$\propto \left[1 + \frac{1}{(n-K)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' (s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right]^{-\frac{n-K+K}{2}} \quad \text{(去掉常数项)}$$

而这正是 K 维 t 分布的密度核，其中心在 $\hat{\boldsymbol{\beta}}$，自由度为 $(n-K)$，而协方差矩阵为 $s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$。这个结果与古典多元回归类似。单个参数 β_j 的后验分布则为一维 t 分布。

最后，要得到 σ^2 的边缘后验分布(marginal posterior)，只要将联合后验分布 $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X})$ 中的 $\boldsymbol{\beta}$ 积分积掉即可。回到联合后验分布的表达式：

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto (1/\sigma^2)^{K/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' (\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\}$$

$$\times (1/\sigma^2)^{\frac{n-K}{2}+1} \exp\{ -(n-K)s^2/2\sigma^2 \}$$

其中，第二部分不含 $\boldsymbol{\beta}$，而第一部分为 K 维正态分布的密度核，积分为 1，故

$$p(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto (1/\sigma^2)^{\frac{n-K}{2}+1} \exp\{ -(n-K)s^2/2\sigma^2 \} \tag{31.23}$$

这是“Γ 分布的平方根之倒数的密度核”(kernel of an inverted square-root gamma density)。

总之，使用无信息先验分布得到的结果与古典多元回归相类似。有关使用“有信息的先验分布”(informative prior)的结果，参见 Cameron and Trivedi(2005, p. 437)。这些结果启示我们，虽然贝叶斯估计与古典估计在方法论上迥异，有限样本的结论也不同，但在大样本下，却常有异曲同工之妙。

31.7 马尔可夫链蒙特卡罗法

古典学派的核心为点估计。在贝叶斯估计中，则对应于后验均值的计算：

$$\mathrm{E}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \tag{31.24}$$

其中，$p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 为后验密度。但 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 通常并无解析表达式(可能包括复杂的积分)，故难以得到(31.24)式的解析式。为此，常使用蒙特卡罗积分方法(参见第 19 章)。此处的一个技术难题是，如果 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 无解析式，如何从后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 中获得随机样本？

“吉布斯抽样法”(Gibbs sampler)的基本思想是，通过更简单的条件分布来抽样。比如，从上节可知，联合后验分布 $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X})$ 的表达式较复杂，但条件后验分布 $p(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}, \boldsymbol{X})$ 较简单(为 K 维正态)。一般地，考虑从二维联合分布 $f(x, y)$ 进行随机抽样，但由于 $f(x, y)$ 的表达式太复杂而无法直接抽样。另一方面，假设条件分布 $f(x \mid y)$ 与 $f(y \mid x)$ 相对简单，可从中进行一维随机抽样。为此，可使用以下迭代法得到联合分布 $f(x, y)$ 的随机样本。

吉布斯抽样法的步骤：

(1) 选择 x_0(只要在条件分布 $f(x \mid y)$ 的取值范围即可)；

(2) 从条件分布 $f(y \mid x_0)$ 中随机抽取 y_0；

(3) 从条件分布 $f(x \mid y_0)$ 中随机抽取 x_1；

(4) 从条件分布 $f(y \mid x_1)$ 中随机抽取 y_1；

(5) 从条件分布 $f(x \mid y_{t-1})$ 中随机抽取 x_t；

(6) 从条件分布 $f(y \mid x_t)$ 中随机抽取 y_t。

重复以上步骤(5)与(6)几千次之后,即可得到联合分布$f(x,y)$的随机抽样(通常去掉前面几千个抽样值,以避免初始值的影响,称为"burn in")。吉布斯抽样的理论基础在于,当$t\to\infty$时,(x_t,y_t)的极限分布(limiting distribution)就是$f(x,y)$。使用类似方法,可以得到更高维分布(比如$f(x,y,z)$)的随机样本。

然而,吉布斯抽样并非真正的随机抽样,因为每次抽样都是上次抽样的函数,只不过上次抽样也是随机的。这种性质表明,吉布斯抽样所得到的序列其实是马尔可夫链。

定义:考虑时间为离散的随机过程$\{x_t\}_0^\infty$。如果对于任何t,都有$P(x_{t+1}\leqslant x\mid x_t,x_{t-1},\cdots,x_0)=P(x_{t+1}\leqslant x\mid x_t)$,则称$\{x_t\}$为"马尔可夫链"(Markov chain)。

显然,在给定整个历史的情况下,x_{t+1}的条件分布仅与x_t有关。换言之,马尔可夫链的未来状态(x_{t+1})仅与现在状态(x_t)有关,而与历史状态$(x_{t-1},\cdots,x_0)$无关。由于吉布斯抽样法使用了蒙特卡罗抽样,而得到的序列又是马尔可夫链,故也称为"马尔可夫链蒙特卡罗法"(Markov Chain Monte Carlo,简记MCMC),在贝叶斯分析中应用广泛。

有关贝叶斯估计的进一步知识,参见Zellner (1971),Poirier (1995),Koop (2003),Lancaster (2004),Cameron and Trivedi(2005)以及Geweke (2005)。遗憾的是,目前还(几乎)没有专门针对贝叶斯估计的Stata官方命令;但读者可搜到少量的非官方Stata命令,比如在Stata中输入命令"`findit Bayesian`"。

习 题

31.1 一名贝叶斯学派的研究者从总体$y\sim N(\mu,100)$中抽了一个样本容量为4的随机样本,得到样本均值为56。该学者得出结论,μ的后验分布为$N(52,100)$。该学者对μ的先验分布是什么?

31.2 (正态分布的杰佛里先验分布)假设$y\sim N(\mu,\sigma^2)$。

(1) 假设μ未知而σ^2已知,证明$I(\mu)=n/\sigma^2$;

(2) 假设σ^2未知而μ已知,证明$I(\sigma^2)=n/2\sigma^4$;

(3) 假设μ与σ^2皆未知,证明$I(\mu,\sigma^2)=n/2\sigma^6$。

第 32 章　如何做规范的实证研究

在本书中，我们已经学了不少的计量经济学知识。到底应该如何使用这些计量知识，以及在做规范的“实证研究”（empirical research）[①]时，一般有哪些标准的步骤与注意事项，这些是本章关注的内容[②]。

有人把计量经济学理解为概率统计的一门分支，将注意力集中于定理的推导与证明。事实上，计量知识只是一种手段，而最终目的是通过计量知识的指导来做好实证研究。比如，计量知识告诉我们，在何种情况下，应该使用何种计量方法，其假设与局限性又是什么，如何检验，等等。

32.1　计量理论与现实数据

计量经济学的理论总是建立于一些理想化的假定基础之上[③]，而现实的经济数据常常或多或少地不符合这个或那个假定。因此，尽管计量理论是可以严格证明的一门科学，但实证研究的实践本身在一定程度上却是一门艺术。

Kennedy（2003，p. 390）认为，“理论计量经济学家”（econometric theorists）与“计量实践者”（applied econometricians）常常缺乏交流，而“不使用真实数据的计量理论”在某种意义上更容易些。其结果是，计量实践者在处理现实数据时有时不知所措，甚至盲目进行。

事实上，无论多么高深的计量方法，如果原始数据的质量有问题，也只能是“垃圾进去，垃圾出来”（garbage in，garbage out）。Griliches（1994）在给美国经济学会做的主席演讲（presidential address）中指出，由于经济学家不够注意数据的来源及产生过程，经常错误地解释数据，导致研究的进展缓慢；如果不提高数据质量，计量理论方面的重大进展将无用武之地。当然，对于大多数从事应用研究的学者而言，主要是使用别人（比如，统计局）提供的数据。即便如此，也应该对数据的质量有一个清醒的判断，并使用相应的计量方法（至少在做出实证研究的结论时，应考虑到数据质量的影响）。

另一方面，对于计量的初学者来说，与真实数据打交道是加深对计量经济学理解的重要途径。只有弄脏你的手，才能学会实证研究（So let's get our hands dirty）。

---

① “实证研究”的另一含义是“positive analysis”，即研究世界实际上是怎样的；与此相对的是“规范研究”（normative analysis），即研究世界应该是怎样的（比如，福利经济学）。本书中的“实证研究”指的是使用数据进行的经验分析，即“empirical research”。

② 本章包括了一些个人见解，不代表实证研究应该这样做或必须这样做，不当之处请老师与同学们指正。

③ 主要是为了推导与证明的方便。

32.2 实证研究的主要步骤

计量经济学是实证研究的工具。因此，只有站在实证研究的框架上才能更好地回答“如何使用计量知识”这一问题。一般来说，规范的实证分析包括以下几个步骤[①]。

1. 提出研究问题

对于刚开始进行研究的人来说，常常不知该如何选择研究课题。研究者通常知道自己想要研究的领域（比如，经济增长），但这还不是一个具体的研究问题（research question）。对实证分析而言，研究问题通常是有关“x 对 y 有何作用”之类的问题。如果想研究“金融发展对经济增长的作用”，就更具体了，此处 x 指的是“金融发展”（可以包含多个变量），而 y 指的是“经济增长”[②]。

研究问题可以来源于理论（比如，检验资产定价模型 CAPM 是否成立），也可以来自对经济现象的观察（比如，媒体的热点问题报道，社会调研），可以研究某个政策的效应（比如，新劳动法对失业率的影响），也可以对文献中已有论文进行改进。

提出好的研究问题，也许你的研究就成功了一半。什么是好的研究问题呢？总的来说，研究问题越具体、越有趣、越新颖、越有可行性，则越好！

（1）**具体**：简单来说，在以上“x 对 y 有何作用”这样的句型中，应能明确 x 与 y 具体是什么。

（2）**有趣**：你研究的问题别人会感兴趣吗？为什么我们要在乎你的问题（Why should we care）？知道你的问题的答案后，能影响人们对世界的某个方面的看法吗？你的研究问题为什么重要？

（3）**新颖**：论文的核心价值在于它的创新性，即做出了文献中所没有的边际贡献。这种边际贡献可以是研究了新的现象、使用了新的（更好的）计量方法或使用了新的数据集。做研究的过程是创造新知识的过程，在本质上不同于学习已有（旧）知识的过程。

（4）**可行**：即使你的研究问题很具体、很有趣、很新颖，如果找不到相应的数据，则不可行。

有了潜在的研究问题后，就可以进行初步的探索性研究（exploratory study），看看它是否满足以上标准。比如，通过查文献，看别人是否已经做过类似的研究，或大致了解数据是否可得。显然，备选的研究问题或想法越多越好，因为能“存活”下来的课题通常不多。

2. 文献回顾

论文贵在有新意。假设你找到了一个具体、有趣而可行的研究问题，那么它究竟有多少新颖性，这就不可避免地需要查阅文献，看看文献中是否已有类似的研究。对于中文论文，可在 CNKI 中搜索；对于英文论文，可在 JSTOR 或 EconLit 中搜索[③]。某些工作论文则可通过百度或 Google 搜索。

对于刚开始进行研究的研究生们而言，似乎自己能够想得到的研究题目，都已经被别人做过了。其实未必如此。要想做出新的边际贡献，当然可以在前人的基础上，继续拓展与改进（改进

---

① 如果是纯粹的理论文章（不使用数据），则不必遵循此步骤。

② 事实上，文献中充斥着“某某因素对经济增长的作用”这类实证论文。

③ 这里仅列出了主要的论文数据库。

计量方法,增加变量,使用新数据等)[①]。也可以完全撇开前人,去研究全新的现象。比如,20世纪70年代末中国农村实行了家庭联产承包责任制改革,这是史无前例的。到了1990年左右,就涌现出一批研究农村改革对中国农业产出影响的论文[②]。又比如,2008年中国开始实施"新劳动法"。假以时日(有了足够的数据后),就可以研究新劳动法对劳动力市场的影响。即使是前人已经研究过的现象,也常常可以用新的眼光、新的视角去观察。重要的是,要有敏锐的观察力,并"像经济学家那样去思考"(Think like an economist)。当然,对于刚起步的研究者,应尽量避免已经被研究得很烂或过于有挑战性的题目。

什么时候开始看文献,即看文献的时机,也很重要。如果从一开始就大量地看文献,则可能被文献所淹没,望洋兴叹,自觉渺小。更好的方法是,只有当自己已经有了一定的想法后,再去系统地看文献。这样,才会知道自己究竟想看什么,也更能带着批判性的眼光去看。

阅读文献的态度十分重要。虚心地从经典论文中汲取营养,才能"站在巨人的肩膀上"(Standing on the shoulders of giants),但仍应带着某种批判性的眼光。事实上,由于经济现象的复杂性,任何论文一定都有缺点(比如,忽略了某些可能重要的因素),也都有可以改进的空间(甚至可能推翻作者的结论),不必太迷信"权威"。既然经济学还不是科学,获得诺贝尔奖的经济学家可能持有相反的观点,那么又哪来的权威呢?重要的是,使用逻辑与实证的方法对不同的观点进行甄别与质疑。如果认为前人所做的研究已经十全十美,你都赞同,那么,你怎么可能做出新的边际贡献呢?

3. 建立理论模型

虽然实证研究可以没有理论模型,但如果有好的理论模型作为基础,则更有说服力。具体来说,回归分析一般只能说明变量之间的相关性,要对变量之间的因果关系做出判断,常常需要依赖于经济理论。因此,即使无法提供完整的理论模型,也应该进行一定的理论分析。最理想的情形是,从理论模型中推导出计量模型(econometric model),即待估计的回归方程。

4. 收集数据

在正式开始研究之前,应大致知道你所需要的数据是存在的,而且可以得到[③]。总的来说,数据要么是别人提供的(比如,统计局),要么是自己收集的(比如,问卷调查)。如果使用别人提供的数据,要注意其数据是如何定义、计算的(统计口径),是否是与你的理论模型中的变量相对应。本章附录中列出了常用的数据来源。

如果确实不知道该从哪里找数据,可以看文献中同类研究的数据来源,并追溯其本源。近年来,《美国经济评论》(*American Economic Review*)与《应用计量经济学杂志》(*Journal of Applied Econometrics*)都在其网站上公开了发表论文中使用的数据集。

对于非电子版的数据,则需要自己耐心地手工输入数据,并注意检查,防止出错。

5. 使用适当的计量方法

有了数据之后,就可以根据数据的类型与特点,来决定使用合适的计量方法。比如,被解释变量为二元变量,则可以使用probit或logit;如果是非负整数,则可以使用泊松回归;如果是面板数据,则可以考虑固定效应、随机效应、时间效应、动态面板等;如果是时间序列,则应先判断是否

---

① 如果你的新贡献只是增加了一两个控制变量,恐怕不足以作为一篇论文发表。

② 包括林毅夫老师的论文,Lin(1992)。

③ 有些时候,虽然数据存在,但可能是保密的内部数据,或获得的代价过高。

含单位根[①],再决定使用相应的计量方法。

对于一般的数据,通常先作 OLS,看看结果,作为一个参照系。作完 OLS 后,可以画残差图(residual plot),大致看看扰动项是否符合经典假定,然后进行严格的检验。如果有所违背(比如,存在异方差、自相关),则做相应的处理。对于时间序列,还可以检验是否存在结构变动(邹检验)。另外,应该对数据的质量进行检验,判断是否存在多重共线性、极端值、弱工具变量等,并做相应的调整。

对于回归函数的具体形式,可以考虑线性、对数(如果解释变量只取正数且有指数增长趋势)、双对数、非线性(边际效应不是常数)等。

由于受数据可获得性的限制,遗漏变量几乎不可避免。因此,很有必要在实证论文中对此进行讨论。不外乎两种情况。第一,存在遗漏变量,但与解释变量不相关(需要说明为什么不相关),故可以不做处理。第二,存在遗漏变量,且与解释变量相关,则必须进行处理,处理方法包括增加控制变量、寻找代理变量、使用工具变量、使用面板数据等。

另一常见问题是内生解释变量。此时,一般需要找到有效的工具变量才能得到一致的估计。

由于面板数据可以在一定程度上克服遗漏变量问题,故比横截面数据或时间序列更有说服力。因此,如果可以获得面板数据,则应尽力争取。

大多数的实证论文都希望说明 $\boldsymbol{x}$ 对 y 的因果作用。而从回归分析的相关关系到因果关系,是一个很大的飞跃。

总之,在这部分应该说明的是,为什么你所使用的计量方法是最恰当的。

32.3 实证论文的结构

实证论文的一般结构如下。

(1) 引言(Introduction)
(2) 文献回顾(Literature review)
(3) 理论框架(Theoretical framework)
(4) 计量模型(Econometric model)
(5) 数据说明(Data description)
(6) 计量方法(Estimation method)
(7) 回归结果(Regression results)
(8) 结论(Conclusion)

对于初学者,要注意观察经典论文的文章结构与风格(包括如何用表格来汇报回归结果),以便模仿。比如,中文论文可以模仿《经济研究》或《经济学(季刊)》,而英文论文则可参照《美国经济评论》等。

① 如果样本容量太小(比如小于 20),则可以不进行单位根检验,因为单位根检验是大样本检验,不适用于小样本。

32.4 计量实践的十诫

Kennedy (2003, p. 390)提出了以下"计量实践的十诫"(The Ten Commandments of Applied Econometrics),作为指导实证研究的十个准则。

1. 运用常识与理论(Use common sense and economic theory)

在进行模型设定时,应尽量使用常识与经济理论。比如,将"人均变量"(如人均消费)与"人均变量"(如人均 GDP)相匹配;使用实际汇率来解释实际进出口;避免使用带时间趋势的变量来解释不带时间趋势的变量。

2. 避免第Ⅲ类错误(Avoid Type Ⅲ Error)

在作统计检验时,要尽量减少第Ⅰ类错误与第Ⅱ类错误。而第Ⅲ类错误则指的是,对错误的问题给出了正确的答案(the right answer to a wrong question)。比如,考虑 FDI 对经济增长的作用。由于 FDI 起作用需要时间,如果仅把当年的增长率对当年的 FDI 进行回归,则可能没有太大意义。比较适当的做法是,考虑期初的 FDI 对随后 5 年(或若干年)经济增长率的作用。

3. 熟悉经济现象(Know the context)

实证研究者应该知道所研究现象的历史、制度与文化背景,以及数据的来源与产生过程。换言之,你应该知道得比计算机更多。比如,如果你研究中国的失业问题,那么你应该知道,中国的失业率指的是"城镇登记失业率",其统计口径与标准教科书中以及西方国家的失业率概念很不相同。

4. 查看数据(Inspect the data)

研究者的一个常见误区是,只知道按计算机键盘,而不去熟悉原始数据(raw data)。查看数据的方法包括,计算数据的主要统计量(summary statistics),比如,均值、最大最小值、标准差、相关系数等;画图;以及增加对数据的感觉(get a feel for the data)。在查看数据的过程中,还要注意观察数据中是否存在不一致的(inconsistent)地方,比如,出现了不可能、不现实或可疑的取值。如果发现,则要进行处理(比如,可能是数据输入错误),这被称为"数据清理"(data cleaning)。

5. 模型应尽量简洁(Keep It Sensibly Simple, KISS)

尽量不要把模型搞得太复杂。"从小到大"(specific-to-general)的建模方法简单易行,但可能偏差较大(因为存在遗漏变量)。而"从大到小"(general-to-specific)的建模方法偏差小,但却不易执行。通常的折中方案是,选择简单而有解释力的模型。

6. 理解计量结果(Use the interocular trauma test)

计算机输出的计量结果常常很多页,包含密密麻麻的表格与数字。如果你尝试了各种不同的计量方法与控制变量(alternative specifications),则结果就会更复杂。如何看这些结果?只能用双眼一个字一个字地看,直到看明白为止。比如,系数的符号对吗,大小合理吗,主要变量在统计上显著吗,与经济理论矛盾吗?如果把你的结果告诉外行,对方会发笑吗?

7. 理解数据挖掘的成本与收益(Understand the costs and benefits of data mining)

不好的"数据挖掘"(data mining)指的是,根据数据的特征来设定计量模型。这种数据挖掘之所以不好,是因为模型设定是由数据产生的,如果再用这个数据去检验由它产生的模型,那么就不是客观的检验。比如,进行回归之后,发现主要变量的系数不显著,有些研究者就想使用"何种方法",能使得它变得显著。事实上,"系数不显著"本身也是一种结果,也可以在论文中汇

报。如果数据本身没有问题，则它或许预示着，实证数据与理论相矛盾，故可能需要对理论进行修正。

好的“数据挖掘”指的是，对数据进行各种“实验”，以期发现数据中的某种规律性，发现模型设定的错误，以此改进理论。事实上，一定程度的数据挖掘是不可避免的，而这两种形式的数据挖掘的界限并不清晰，正如 Heckman(2000)指出，“尽管使用数据来检验受到该数据启发的理论存在严重的问题，但如果拒绝从数据中学习并修改理论，则会导致更严重的问题”①。

8. 权衡理论与现实(Be prepared to compromise)

在一般情况下，计量实践中的问题与理论计量经济学中最接近的情形会有一定的距离。比如，使用的代理变量，是否满足教科书中对代理变量的要求？内生解释变量问题是否严重？单位根检验的结果是否可信？因此，通常需要在现实与理论之间找到一个适当的妥协。

9. 区分统计显著性与经济显著性(Do not confuse statistical significance with meaningful magnitude)

通常，样本容量越大，则估计值的标准误越小，该系数的“统计显著性”(statistical significance)就越高。但如果该系数估计值为 0.001，则可能不具有“经济显著性”(economic significance)，即这个解释变量的变化对被解释变量的影响不大，尽管这种影响的幅度被估计得很精确。

10. 进行敏感度分析(Report a sensitivity analysis)

为了得到论文的结果，研究者通常需要做一系列的假定。问题是，论文的主要结果是否对这些假定很敏感？为此，有必要放松论文的假定，看结果是否稳健(基本不变)，这被称为“敏感度分析”。比如，改变样本区间、函数形式、控制变量等。只有稳健的结果才是有说服力的。

32.5 结 束 语

在很多情况下，经济理论并不能给我们确切的答案。比如，有关“中国的内需不足问题”，学界与媒体已经提出不下十种解释，莫衷一是。究竟哪个(些)因素是导致内需不足的主要原因呢？要真正回答这个问题，则必须仔细地收集数据，进行深入的实证分析。在一定意义上，或许可以说：“理论是灰色的，而实证之树常青。”

有关如何做研究、写论文以及经济学方法论方面的讨论，可进一步参考 Friedman (1953)，Stigler and Becker (1977)，林毅夫(1995，2013)，Varian (2001，2009)，Cochrane (2005)，李子奈(2011)，傅十和、佩雷拉(2012)等。

习 题

32.1 学习实证研究的好方法是复制(replicate)经典论文或最新工作论文的结果。挑选你感兴趣的几篇论文，下载其数据集，并复制其结果。

---

① While there are serious problems in using the data that suggest a theory to test that theory, even more important problems arise from refusing to learn from the data in revising economic models.

附录：常用数据来源[①]

【中国数据】

国家统计局：www. stats. gov. cn
中国人民银行：www. pbc. gov. cn
财政部：www. mof. gov. cn
国研网（国务院发展研究中心）：www. drcnet. com
中经网（国家信息中心）：www. cei. gov. cn
中宏网（国家发改委）：www. macrochina. com. cn
万得（Wind）金融终端：www. wind. com. cn
国泰安 csmar 数据库：www. gtarsc. com
CCER 经济金融研究数据库：www. ccerdata. com
EPS 数据库：www. epsnet. com. cn
CEIC 数据库：www. ceicdata. com
China Data Center（Univ. of Michigan）：http：//chinadatacenter. org/newcdc/
中国调查数据网（CCER and Univ. of Michigan）：www. chinasurveycenter. org
China Health and Nutrition Survey （CHNS）：www. cpc. unc. edu/projects/china
China Human Capital Project （Univ. of Pennsylvanian）：www. ssc. upenn. edu/china/index. htm
经济发展论坛：www. fed. org. cn/
各类统计年鉴。

【国际数据】

Bloomberg Data：http：//about. bloomberg. com/product_data. html
BREAD data in development economics：http：//ipl. econ. duke. edu/bread/index. htm
Economic Time Series Page：www. economagic. com
IMF：www. imf. org
NBER online data：www. nber. com/data_index. html
Panel Study of Income Dynamics：www. umich. edu/ ~ psid
Penn World Table：http：//pwt. econ. upenn. edu/php_site/pwt_index. php
OECD：www. oecd. org
Standard & Poor's Compustat：http：//www. compustat. com
Thomson Datastream：www. datastream. com
US Federal Reserve Bank：www. federalreserve. gov

---

① 某些数据库为收费数据库，可向你所在的大学或单位询问是否已购买。

US Bureau of Labor Statistics:www. bls. gov

US Bureau of Economic Analysis:www. bea. gov

United Nations:www. un. org

World Bank:www. worldbank. org

BVD 财经系列数据库(欧洲):www. bvdinfo. com (包括 Bankscope)

Inter-University Consortium for Political and Social Research:

www. icpsr. umich. edu/icpsrweb/ICPSR

Journal of Applied Econometrics Data Archive:www. econ. queensu. ca/jae

KOF Index of Globalization:http://globalization. kof. ethz. ch/

Daron Acemoglu's datasets:http://economics. mit. edu/faculty/acemoglu/data

Robert Barro's datasets:www. economics. harvard. edu/faculty/barro/data_sets_barro

Charles Jones' datasets:www. stanford. edu/~chadj/datasets. html

其他学者的个人网站。

【历史数据】

Maddison, Angus. The World Economy: Volume 1: A Millennial Perspective and Volume 2: Historical Statistics. Paris:OECD,2001 - 2003.

Economic History Service:http://eh. net/

NBER Macrohistory Database:www. nber. org/databases/macrohistory/contents/index. html

Historical Statistics (Princeton):http://firestone. princeton. edu/econlib/history1. htm

Historical Financial Statistics:www. centerforfinancialstability. org

Integrated Network for Societal Conflict Research (INSCR): www. systemicpeace. org/inscr/inscr. htm

参考书目

[1] 艾春荣,陈小红. 计量经济学:半参数计量经济学方法. 北京:北京大学出版社,2000.

[2] 白仲林. 面板数据的计量经济分析. 天津:南开大学出版社,2008.

[3] 陈海燕. 面板数据模型的检验方法. 北京:经济科学出版社,2012.

[4] 陈家鼎,刘婉如,汪仁宫. 概率统计讲义. 3 版. 北京:高等教育出版社,2004.

[5] 陈强. 气候冲击、王朝周期与游牧民族的征服. 经济学(季刊),2014,即将发表.

[6] 阿德里安·达内尔. 计量经济学辞典. 钱晓明,译. 上海:上海财经大学出版社,2006.

[7] 沃尔特·恩德斯. 应用计量经济学:时间序列分析. 2 版. 杜江,谢志超,译. 北京:高等教育出版社,2006.

[8] 傅十和,佩雷拉. 如何撰写并发表英文经济学论文. 上海:格致出版社、上海人民出版社,2012.

[9] 高铁梅. 计量经济分析方法与建模:Eviews 应用及实例. 2 版. 北京:清华大学出版社,2009.

[10] 洪永淼. 金融计量的新近发展. 经济学季刊,2002 (2).

[11] 洪永淼. 计量经济学的地位、作用和局限. 经济研究,2007(5).

[12] 洪永淼. 高级计量经济学. 赵西亮,吴吉林,译. 北京:高等教育出版社,2011.

[13] 胡咏梅. 计量经济学基础与 STATA 应用. 北京:北京师范大学出版社,2010.

[14] 姜近勇,潘冠中. 金融计量学. 北京:中国财政经济出版社,2011.

[15] 靳云汇,金赛男. 高级计量经济学(上册). 北京:北京大学出版社,2007.

[16] 靳云汇,金赛男. 高级计量经济学(下册). 北京:北京大学出版社,2011.

[17] 劳伦斯·汉密尔顿. 应用 Stata 做统计分析. 5 版. 郭志刚,等,译. 重庆:重庆大学出版社,2008.

[18] 兰草. 截面、面板数据分析与 STATA 应用. 武汉:武汉大学出版社,2012.

[19] 李宝仁. 高级计量经济学. 北京:经济科学出版社,2012.

[20] 李春涛,张璇. 随机模拟与金融数据处理 Stata 教程. 北京:中国金融出版社,2009.

[21] 李宏彬,孟岭生,施新政,吴斌珍. 父母的政治资本如何影响大学生在劳动力市场的表现:基于中国高校应届毕业生就业调查的经验研究. 经济学季刊,2012(3).

[22] 李贤平. 概率论基础. 2 版. 北京:高等教育出版社,1997.

[23] 李雪松. 高级经济计量学. 北京:中国社会科学出版社,2008.

[24] 李子奈. 计量经济学模型方法论. 北京:清华大学出版社,2011.

[25] 李子奈,叶阿忠. 高级应用计量经济学. 北京:清华大学出版社,2012.

[26] 李子奈,叶阿忠. 高等计量经济学. 北京:清华大学出版社,2000.

[27] 李忠民,张世英,袁学民. 经济计量学教程. 2 版. 天津:天津大学出版社,2009.

[28] 林光平. 计算计量经济学. 杨大勇,译. 北京:清华大学出版社,2003.

[29] 林光平,龙志和,吴梅. 我国地区经济收敛的空间计量实证分析:1978—2002 年. 经济学季刊,2005(4).

[30] 林少宫. 微观计量经济学要义:问题与方法探讨. 武汉:华中科技大学出版社,2003.

[31] 林毅夫. 本土化、规范化、国际化——庆祝《经济研究》创刊 40 周年. 经济研究,1995(10).

[32] 林毅夫. 本体与常无:经济学方法论对话. 北京:北京大学出版社,2013.

[33] 陆懋祖. 高等时间序列经济计量学. 上海:上海人民出版社,1999.

[34] 栾惠德. 张晓峒. 季节调整中的春节模型. 经济学(季刊),2007(2).

[35] 赫尔穆特·鲁克波尔,马库斯·克莱茨希. 应用时间序列计量经济学. 易行健,邓可斌,等,译. 北京:机械工业出版社,2008.

[36] 茆诗松,王静龙,濮晓龙. 高等数理统计. 2 版. 北京:高等教育出版社,2006.
[37] 潘省初. 计量经济学中级教程. 北京:清华大学出版社,2009.
[38] 任燕燕. 平行数据模型及其在经济分析中的应用. 北京:经济科学出版社,2006.
[39] 沈体雁,冯等田,孙铁山. 空间计量经济学. 北京:北京大学出版社,2010.
[40] 盛骤,谢式千,潘承毅. 概率论与数理统计. 4 版. 北京:高等教育出版社,2008.
[41] 苏良军. 高等数理统计. 北京:北京大学出版社,2007.
[42] 孙敬水. 中级计量经济学. 上海:上海财经大学出版社,2009.
[43] 童恒庆. 理论计量经济学. 北京:科学出版社,2005.
[44] 王群勇. Stata 在统计与计量分析中的应用. 天津:南开大学出版社,2007.
[45] 王群勇. Stata 使用指南与应用案例. 北京:中国财政经济出版社,2008.
[46] 王少平,杨继生,欧阳志刚. 计量经济学. 北京:高等教育出版社,2011.
[47] 王升. 计量经济学导论. 北京:清华大学出版社,2006.
[48] 王天夫,李博柏. Stata 实用教程. 北京:中国人民大学出版社,2008.
[49] 王文博. 计量经济学. 西安:西安交通大学出版社,2004.
[50] 王燕. 应用时间序列分析. 北京:中国人民大学出版社,2005.
[51] 王志刚. 面板数据模型及其在经济分析中的应用. 北京:经济科学出版社,2008.
[52] 吴可杰,秦宛顺,赵德滋. 经济计量学——理论与方法. 南京:南京大学出版社,1986.
[53] 谢识予,朱弘鑫. 高级计量经济学. 上海:复旦大学出版社,2005.
[54] 杨菊华. 数据管理与模型分析:STATA 软件应用. 北京:中国人民大学出版社,2012.
[55] 叶阿忠. 非参数计量经济学. 天津:南开大学出版社,2003.
[56] 叶阿忠. 非参数和半参数计量经济模型理论. 北京:科学出版社,2008.
[57] 余森杰. 国际贸易学:理论、政策与实证. 北京:北京大学出版社,2013.
[58] 赵国庆. 高级计量经济学——理论与方法. 北京:中国人民大学出版社,2014.
[59] 张车伟,薛欣欣. 国有部门与非国有部门工资差异及人力资本贡献. 经济研究,2008(4).
[60] 张成思. 金融计量学:时间序列分析视角. 大连:东北财经出版社,2008.
[61] 张定胜. 计量经济学. 武汉:武汉大学出版社,2005.
[62] 张鹏伟,李嫣怡. Stata 统计分析与应用. 北京:电子工业出版社,2011.
[63] 张晓峒. 计量经济分析. 修订版. 北京:经济科学出版社,2000.
[64] 周绍杰,张俊森,李宏彬. 中国城市居民的家庭收入、消费和储蓄行为:一个基于组群的实证研究. 经济学季刊,2009,8(4).
[65] 朱建平,胡朝霞,王艺明,林光平. 高级计量经济学导论. 北京:北京大学出版社,2009.
[66] 朱平芳. 现代计量经济学. 上海:上海财经大学出版社,2004.
[67] Aalen, O., 1978. "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, 6, 701 – 726.
[68] Abadie, A., and G. W. Imbens, 2002. "Simple and Bias-corrected Matching Estimators," Technical report, Department of Economics, University of California, Berkeley.
[69] Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens, 2004. "Implementing Matching Estimators for Average Treatment Effects in Stata," *The Stata Journal*, 4, 290 – 311.
[70] Abadie, A., and G. W. Imbens, 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235 – 267.
[71] Abadie, A., and G. W. Imbens, 2008. "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76, 1537 – 1557.
[72] Abadie, A., and G. W. Imbens, 2011. "Bias Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1 – 11.
[73] Abadie, A., and G. W. Imbens, 2012. "Matching on the Estimated Propensity Score," Harvard University working

paper.

[74] Acemoglu, D., S. Johnson, and J. Robinson, 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369 – 1401.

[75] Acemoglu, D., S. Johnson, J. Robinson, and P. Yared, 2008. "Income and Democracy," *American Economic Review*, 98, 808 – 842.

[76] Acemoglu, D., S. Johnson, and J. Robinson, 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply," *American Economic Review*, 102, 3077 – 3110.

[77] Aigner, D., K. Lovell, and P. Schmidt, 1977. "Formulation and Estimation of Stochastic Frontier Production Models," *Journal of Econometrics*, 6, 21 – 37.

[78] Akaike, H., 1974. "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716 – 722.

[79] Albouy, D., 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment," *American Economic Review*, 102, 3059 – 3076.

[80] Altonji, J., T. Elder, and C. Taber, 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151 – 184.

[81] Amemiya, T., 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

[82] Amemiya, T. and T. MaCurdy, 1986. "Instrumental Variable Estimation of An Error Component Model," *Econometrica*, 54, 869 – 880.

[83] Amisano, G., and C. Giannini. 1997. *Topics in Structural VAR Econometrics*. 2nd ed. Heidelberg: Springer.

[84] Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

[85] Anselin, L., 2010. "Thirty Years of Spatial Econometrics," *Papers in Regional Science*, 89, 3 – 25.

[86] Anderson, T. and C. Hsiao, 1981. "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, 76, 598 – 606.

[87] Andrews, D., and E. Zivot, 1992. "Further Evidence on the Great Crash, the Oil Price Shock, and the Unit-Root Hypothesis," *Journal of Business and Economic Statistics*, 10, 251 – 270.

[88] Angrist, J., 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313 – 336.

[89] Angrist, J. and A. Krueger, 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979 – 1014.

[90] Angrist, J., and V. Lavy, 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533 – 575.

[91] Angrist, J. and J. Pischke, 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

[92] Arbia, G., 2006. *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Springer-Verlag, Berlin.

[93] Arellano, M. and S. Bond, 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equation," *Review of Economic Studies*, 58, 277 – 298.

[94] Arellano, M. and O. Bover, 1995. "Another Look at Instrumental Variable Estimation of Error Components Models," *Journal of Econometrics*, 68, 29 – 51.

[95] Arraiz, I., D. M. Drukker, H. H. Kelejian, and I. R. Prucha, 2010. "A Spatial Cliff-Ord-type Model with Heteroskedastic Innovations: Small and Large Sample Results," *Journal of Regional Science*, 50, 592 – 614.

[96] Baba, Y., R. F. Engle, D. F. Kraft, and K. F. Kroner, 1990. "Multivariate Simultaneous Generalized ARCH," *mimeo*, Department of Economics, University of California, San Diego.

[97] Ball, R., and P. Brown, 1968. "An Empirical Evaluation of Accounting Income Numbers," *Journal of Accounting*

Research,6,159 – 178.

[98] Baltagi, B.,2005. *Econometric Analysis of Panel Data*,3rd edition. Chichester, West Sussex, England: John Wiley & Sons Ltd.

[99] Baltagi, B. and S. Khanti-Akom, 1990. "On Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables," *Journal of Applied Econometrics*,5,401 – 406.

[100] Banerjee, A., M. Marcellino, and C. Osbat, 2005. "Testing for PPP: Should We Use Panel Methods?" *Empirical Economics*,30,77 – 91.

[101] Basmann, R. L., 1957. "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation," *Econometrica*,25,77 – 83.

[102] Battese, G. E., and T. J. Coelli, 1992. "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India," *Journal of Productivity Analysis*,3,153 – 169.

[103] Baum, C.,2006. *An Introduction to Modern Econometrics Using Stata*. College Station, Texas: Stata Press.

[104] Baum, C., M. Schaffer, and S. Stillman, 2007. "Enhanced Routines for Instrumental Variables/GMM Estimation and Testing," *The Stata Journal*,7,465 – 506.

[105] Becketti, S.,2013. *Introduction to Time Series Using Stata*. College Station, Texas: Stata Press.

[106] Belrman, J., Z. Hrubec, P. Taubman, and T. Wales, 1980. *Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment, and Schooling*, Amsterdam: North-Holland.

[107] Bernanke, B., 1986. "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy*, North-Holland, Amsterdam.

[108] Berndt, E. R., 1996. *The Practice of Econometrics: Classic and Contemporary*. New York: Addison-Wesley.

[109] Berndt, E., B. Hall, R. Hall and J. Hausman, 1974. "Estimation and Inference in Nonlinear Structural Model," *Annals of Economics and Social Measurement*,3/4,653 – 655.

[110] Black, S., 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*,114,577 – 599.

[111] Blanchard, O., and D. Quah, 1989. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79,655 – 673.

[112] Blundell, R. and S. Bond, 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*,87,115 – 143.

[113] Bollerslev, T., 1986. "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*,31, 307 – 327.

[114] Bollerslev, T., 1990. "Modeling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH Model," *Review of Economics and Statistics*,72,498 – 505.

[115] Bollerslev, T., R. F. Engle, and J. M. Wooldridge, 1988. "A Capital Asset Pricing Model with Time-varying Covariances," *Journal of Political Economy*,96,116 – 131.

[116] Bound, J., D. Jaeger, and R. Baker, 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*,90,443 – 450.

[117] Box, G., G. Jenkins and G. Reinsel, 1994. *Time Series Analysis: Forecasting and Control*,3rd edition. 影印本. 北京:人民邮电出版社,2005.

[118] Box, G. and D. Pierce, 1970. "Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models," *Journal of the American Statistical Association*,65,1509 – 1526.

[119] Breitung, J. 2000. "The Local Power of Some Unit Root Tests for Panel Data," In *Advances in Econometrics, Volume 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, ed. B. H. Baltagi, 161 – 178. Amsterdam: JAI Press.

[120] Breitung, J., and S. Das. 2005. "Panel Unit Root Tests under Cross-sectional Dependence," *Statistica Neerlandica*, 59, 414 – 433.

[121] Bresnow, N. E., 1974. "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89 – 99.

[122] Breusch, T., 1978. "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17, 334 – 355.

[123] Breusch, T. and A. Pagan, 1979. "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica*, 47, 1287 – 1294.

[124] Breusch, T. and A. Pagan, 1980. "The LM Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239 – 254.

[125] Brewer, D., E. Eide, and R. Ehrenberg, "Does It Pay to Attend an Elite Private College? Cross-cohort Evidence on the Effects of College Type on Earnings," *Journal of Human Resources*, 1999, 34, 104 – 123.

[126] Browning, M., A. Deaton and M. Irish, 1985. "A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle," *Econometrica*, 53, 503 – 543.

[127] Bruno, G., 2005. "Estimation and Inference in Dynamic Unbalanced Panel-data Models with a Small Number of Individuals," *The Stata Journal*, 5, 473 – 500.

[128] Buchinsky, M., 1994. "Changes in the U. S. Wage Structure, 1963 – 1987: Application of Quantile Regression," *Econometrica*, 62, 405 – 458.

[129] Buckles, K. and D. Hungerman, 2012. "Season of Birth and Later Outcomes: Old Questions, New Answers," *Review of Economics and Statistics*, forthcoming.

[130] Cameron, A. C. and P. K. Trivedi, 1998. *Regression Analysis of Count Data*. Econometric Society Monograph No. 30, Cambridge University Press, New York, NY.

[131] Cameron, A. C. and P. K. Trivedi, 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press, New York, NY.

[132] Cameron, A. C. and P. K. Trivedi, 2010, *Microeconometrics using Stata*, revised edition. College Station, Texas: Stata Press.

[133] Campbell, J., A. Lo, and C. MacKinlay, 1997. *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

[134] Card, D. and A. Krueger, 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry," *American Economic Review*, 84, 772 – 793.

[135] Casella, G. and R. Berger, 2002. *Statistical Inference*, 2nd edition. Pacific Grove, CA: Duxbury.

[136] Che, Y., Y. Lu, Z. Tao, and P. Wang, 2013, "The Impact of Income on Democracy Revisited," *Journal of Comparative Economics*, forthcoming.

[137] Chen, Q., 2014. "Climate Shocks, State Capacity, and Peasant Uprisings in North China during 25 – 1911 CE," *Economica*, forthcoming.

[138] Cheng, S., and J. S. Long, 2007. "Testing for IIA in the Multinomial Logit Model," *Sociological Methods Research*, 35, 583 – 600.

[139] Choi, I., 2001. "Unit Root Tests for Panel Data," *Journal of International Money and Finance*, 20, 249 – 272.

[140] Chow, G., 1960. "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591 – 605.

[141] Christensen, L. and W. Greene, 1976. "Economies of Scale in U. S. Electric Power Generation," *Journal of Political Economy*, 84, 655 – 676.

[142] Chung, C. -F., P. Schmidt, and A. D. Witte, 1991. "Survival Analysis: A Survey," *Journal of Quantitative Criminology*, 7, 59 – 98.

[143] Cheung, Y. -W., and K. S. Lai, 1995. "Lag Order and Critical Values of a Modified Dickey-Fuller Test," *Oxford*

Bulletin of Economics and Statistics, 57, 411 – 419.

[144] Cleveland, W., 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829 – 836.

[145] Cleves, M. A., W. W. Gould, and R. G. Gutierrez, 2004. *An Introduction to Survival Analysis Using Stata*. Texas: Stata Press.

[146] Cliff, A. D., and J. K. Ord, 1973. *Spatial Autocorrelation*. London: Pion.

[147] Cliff, A. D., and J. K. Ord, 1981. *Spatial Processes, Models and Applications*. London: Pion.

[148] Cochrane, D. and G. Orcutt, 1949. "Application of Least Square Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, 32 – 61.

[149] Cochrane, J., 2005. "Writing Tips for Ph. D. Students," available online at http://www. bus. lsu. edu/hill/writing/cochrane. pdf.

[150] Coelli, T. J. 1995. "Estimators and Hypothesis Tests for A Stochastic Frontier Function: A Monte Carlo Analysis," *Journal of Productivity Analysis*, 6, 247 – 268.

[151] Cox, D. R., 1972. "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society*, *B*, 34, 187 – 220.

[152] Cox, D. R. 1975. "Partial Likelihood," *Biometrika*, 62, 269 – 276.

[153] Cragg, J. G., 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica*, 39, 829 – 844.

[154] Creel M., 2008. *Econometrics*. University Autonoma de Barcelona, mimeo.

[155] Davidson, R. and J. MacKinnon, 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

[156] Davidson, R. and J. MacKinnon, 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

[157] D'Agostino, R. B., A. Balanger, and R. B. D'Agostino, Jr., 1990. "A Suggestion for Using Powerful and Informative Tests of Normality," *American Statistician*, 44, 316 – 321.

[158] Deaton, A., 1985. "Panel Data from Time Series of Cross-Sections," *Journal of Econometrics*, 30, 109 – 126.

[159] Dehejia, R. H., and S. Wahba, 1999. "Causal Effects in Nonexperimental Studies: Reevaluation of the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053 – 1062.

[160] Dell, Melissa, 2010. "The Persistent Effects of Peru's Mining Mita," *Econometrica*, 78, 1863 – 1903.

[161] Dickey, D. and W. Fuller, 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 74, 427 – 431.

[162] Dickey, D. and W. Fuller, 1981. "Likelihood Ratio Tests for Autoregressive Time Series with a Unit Root," *Econometrica*, 49, 1057 – 1072.

[163] DiNardo, J. and J. L. Tobias, 2001. "Nonparametric Density and Regression Estimation," *Journal of Economic Perspectives*, 15, 11 – 28.

[164] Dougherty, Christopher, 2002. *Introduction to Econometrics*, 2nd edition. Oxford, UK: Oxford University Press.

[165] Drukker, D. M., 2002. "Bootstrapping a Conditional Moments Test for Normality after Tobit Estimation," *The Stata Journal*, 2, 125 – 139.

[166] Durbin, J., 1954. "Errors in Variables," *Review of the International Statistical Institute*, 22, 23 – 32.

[167] Durbin, J. and G. Watson, 1950. "Testing for Serial Correlation in Least-Square Regression," *Biometrika*, 37, 409 – 428.

[168] Efron, B., 1977. "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association*, 72, 557 – 565.

[169] Efron, B., 1979. "Bootstrapping Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1 – 26.

[170] Eicker, F., 1967. "Limit Theorems for Regression with Unequal and Dependent Errors," In *Proceedings of the*

Fifth Berkeley Symposium in Mathematical Statistics, Vol. 1. Berkeley: University of California Press.

[171] Elhorst, J. P., 2010. "Spatial Panel Data Models," in Fischer, M. and A. Getis (eds), *Handbook of Applied Spatial Analysis*. Berlin, Heidelberg and New York: Springer, 377 – 407.

[172] Elliot, G., T. Rothenberg and J. Stock, 1996. "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813 – 836.

[173] Engle, R., 1982. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflations," *Econometrica*, 50, 987 – 1008.

[174] Engle, R. F., 2002. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroscedasticity Models," *Journal of Business and Economic Statistics*, 20, 339 – 350.

[175] Engle, R. and C. Granger, 1987. "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251 – 276.

[176] Engle, R. F. and K. F. Kroner, 1995. "Multivariate Simultaneous Generalized GARCH," *Econometric Theory*, 11, 122 – 150.

[177] Engle, R., D. Lilien and R. Robins, 1987. "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model," *Econometrica*, 55, 391 – 407.

[178] Fairlie, R., and R. London, 2012. "The Effects of Home Computers on Educational Outcomes: Evidence from a Field Experiment with Community College Students," *Economic Journal*, 122, 727 – 753.

[179] Fama, E., L. Fisher, M. Jensen, and R. Roll, 1969. "The Adjustment of Stock Prices to New Information," *International Economic Review*, 10, 1 – 21.

[180] Fan, J., 1992. "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998 – 1004.

[181] Frankel, J. and D. Romer, 1999. "Does Trade Cause Growth," *American Economic Review*, 89, 379 – 399.

[182] Frees, E. W., 1995. "Assessing Cross-sectional Correlations in Panel Data," *Journal of Econometrics*, 64, 393 – 414.

[183] Frees, E. W., 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge, UK: Cambridge University Press.

[184] Friedman, M., 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, 32, 675 – 701.

[185] Friedman, M., 1953. "The Methodology of Positive Economics," *Essays in Positive Economics*, Chicago University Press.

[186] Geary, R., 1954. "The Contiguity Ratio and Statistical Mapping," *The Incorporated Statistician*, 5, 115 – 145.

[187] Getis, A., and J. Ord, 1992. "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis*, 24, 189 – 206.

[188] Geweke, J., 2005. *Contemporary Bayesian Econometrics and Statistics*. New York: John Wiley and Sons.

[189] Glosten, L., R. Jagannathan and D. Runkle, 1993. "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 48, 1779 – 1801.

[190] Godfrey, L., 1978. "Testing against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables," *Econometrica*, 46, 1293 – 1302.

[191] Godfrey, L., 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge, UK: Cambridge University Press.

[192] Goldberger, A., 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

[193] Granger, C., 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37, 424 – 438.

[194] Granger, C. and P. Newbold, 1974. "Spurious Regressions in Econometrics," *Journal of Econometrics*,

2,111 – 120.

[195] Greene, W. H., 2000. *Econometric Analysis*, 4th ed. Upper Saddle River, New Jersey: Prentice Hall.

[196] Greene, W. H., 2008. *Econometric Analysis*, 6th ed. Upper Saddle River, New Jersey: Prentice Hall.

[197] Greene, W. H., 2012. *Econometric Analysis*, 7th ed. Upper Saddle River, New Jersey: Prentice Hall.

[198] Greene, W. H., and D. Hensher, 1995. "Multinomial Logit and Discrete Choice Models," in *LIMDEP version 7.0*, ed. W. Greene. Bellport, NY: Econometric Software, Inc. CDA.

[199] Griliches, Z., 1994. "Productivity, R&D and the Data Constraint," *American Economic Review*, 84, 1 – 23.

[200] Griliches, Z., and W. Mason, 1972. "Education, Income, and Ability," *Journal of Political Economy*, 80, S74 – S103.

[201] Groves, T., Y. Hong, J. McMillan, and B. Naughton, 1994. "Autonomy and Incentives in Chinese State Enterprises," *Quarterly Journal of Economics*, 109, 183 – 209.

[202] Guo, S., and M. Fraser, 2009. *Propensity Score Analysis: Statistical Methods and Applications*. SAGE Publications, Inc.

[203] Gutierrez, R. G., S. Carter, and D. M. Drukker. 2001. "sg160: On boundary-value likelihood-ratio tests," *Stata Technical Bulletin* 60: 15 – 18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 269 – 273. College Station, TX: Stata Press.

[204] Hadri, K., 2000. "Testing for Stationarity in Heterogeneous Panel Data," *The Econometrics Journal*, 3, 148 – 161.

[205] Hahn, J., P. Todd, and W. van der Klaauw, 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201 – 209.

[206] Hall, B., Z. Griliches, and J. Hausman, 1986. "Patents and R&D: Is There a Lag?" *International Economic Review*, 27, 265 – 283.

[207] Hamilton, J., 1994. *Time Series Analysis*. Princeton, New Jersey: Princeton University Press.

[208] Hannan, E. and Quin, G., 1979. "The Determination of the Order of an Autoregression," *Journal of Royal Statistical Society*, B, 41, 190 – 195.

[209] Hansen, B., 1999. "Threshold Effects in Non-dynamic Panels: Estimation, Testing and Inference," *Journal of Econometrics*, 93, 345 – 368.

[210] Hansen, B., 2000. "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575 – 603.

[211] Hansen, L., 1982. "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029 – 1054.

[212] Harris, R. D. F., and E. Tzavalis, 1999. "Inference for Unit Roots in Dynamic Panels where the Time Dimension Is Fixed," *Journal of Econometrics*, 91, 201 – 226.

[213] Hartwig, J., 2009. "A Panel Granger-causality Test of Endogenous vs. Exogenous Growth," KOF Working papers 09 – 231, KOF Swiss Economic Institute, ETH Zurich.

[214] Hartwig, J., 2010. "Is Health Capital Formation Good for Long-term Economic Growth? -Panel Granger-causality Evidence for OECD Countries," *Journal of Macroeconomics*, 32, 314 – 325.

[215] Hayashi, F., 2000. *Econometrics*. Princeton, New Jersey: Princeton University Press.

[216] Hausman, J., 1978. "Specification Tests in Econometrics," *Econometrica*, 46, 1251 – 1271.

[217] Hausman, J., and W. Taylor, 1981. "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377 – 1398.

[218] Hausman, J. and D. McFadden, 1984. "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52, 1219 – 1240.

[219] Heckman, J., 1979. "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153 – 161.

[220] Heckman, J., 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth-Century Retrospective,"

Quarterly Journal of Economics, 115, 45 – 97.

[221] Heckman, J. J., H. Ichimura, and P. E. Todd, 1997, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605 – 654.

[222] Heckman, J. J., H. Ichimura, and P. E. Todd, 1998, "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261 – 294.

[223] Hildreth, C. and J. Lu, 1960. "Demand Relations with Autocorrelated Disturbances," Technical Bulletin No. 276, Michigan Stata University.

[224] Holtz-Eakin, D., W. Newey, and H. S. Rosen, 1988. "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371 – 1395.

[225] Horowitz, J., 1997. "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in D. Kreps and K. Wallis (ed.), *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, Vol. 3, 188 – 222. Cambridge, UK: Cambridge University Press.

[226] Horowitz, J., 2001. "The Bootstrap," in J. Heckman and E. Leamer (ed.), *Handbook of Econometrics*, Vol. 5, 3159 – 3228. Amsterdam: North-Holland.

[227] Hosmer, D., and S. Lemeshow, 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley & Sons, Inc.

[228] Hoxby, C., 2000. "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review*, 90, 1209 – 1238.

[229] Hoxby, C., 2007. "Does Competition Among Public Schools Benefit Students and Taxpayers? Reply," *American Economic Review*, 97, 2038 – 2055.

[230] Hsiao, C., 2003. *Analysis of Panel Data*, 2nd edition. Cambridge, UK: Cambridge University Press.

[231] Huber, P., 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, Vol. 1. Berkeley: University of California Press, 1967.

[232] Im, K., M. Pesaran and Y. Shin, 2003. "Testing for Unit Roots in Heterogeneous Panels," *Journal of Econometrics*, 115, 53 – 74.

[233] Imbens, G., and J. Angrist, 1994. "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467 – 476.

[234] Imbens, G., and T. Lemieux, 2008. "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615 – 635.

[235] Imbens, G., and K. Kalyanaraman, 2009. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," NBER Working Paper 14726.

[236] Imbens, G., and J. Wooldridge, 2009. "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5 – 86.

[237] James, C., 1987. "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics*, 19, 217 – 235.

[238] Jarque, C. and A. Bera, 1987. "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55, 163 – 172.

[239] Jeffreys, H., 1946. "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Statistical Society of London*, 186, 453 – 461.

[240] Johansen, S., 1988. "Statistical Analysis of Cointegrated Vectors," *Journal of Economic Dynamics and Control*, 12, 231 – 254.

[241] Johnson, R. and D. Wichern, 2002. *Applied Multivariate Statistical Analysis*, 5th edition. Upper Saddle River, New Jersey: Prentice Hall.

[242] Johnston, J. and J. Dinardo, 1996. *Econometric Methods*, 4th edition. McGraw-Hill/Irvin.

[243] Judson, R. A. and A. L. Owen, 1999. "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists," *Economics Letters*, 65, 9 – 15.

[244] Kaplan, E. L., and P. Meier, P., 1958. "Nonparametric Estimation from Incomplete Observations," *Journal of American Statistical Association*, 53, 457 – 481.

[245] Kelejian, H. H., and I. R. Prucha, 1998. "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances," *Journal of Real Estate Finance and Economics*, 17, 99 – 121.

[246] Kelejian, H. H., and I. R. Prucha, 1999. "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model," *International Economic Review*, 40, 509 – 533.

[247] Kelejian, H. H., and I. R. Prucha, 2004. "Estimation of Simultaneous Systems of Spatially Interrelated Cross Sectional Equations," *Journal of Econometrics*, 118, 27 – 50.

[248] Kelejian, H. H., and I. R. Prucha, 2010. "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances," *Journal of Econometrics*, 140, 53 – 130.

[249] Kennan, J., 1985. "The Duration of Contract Strikes in U. S. Manufacturing," *Journal of Econometrics*, 28, 5 – 28.

[250] Kennedy, P., 2003. *A Guide to Econometrics*. Cambridge, MA: MIT Press.

[251] King, G., and L. Zeng, 2001a. "Explaining Rare Events in International Relations," *International Organization*, 55, 693 – 715.

[252] King, G., and L. Zeng, 2001b. "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137 – 163.

[253] Kiviet, J. F., 1995. "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53 – 78.

[254] Kleinbaum, D., and M. Klein, 2011. *Survival Analysis: A Self-Learning Text*, 3rd edition. New York: Springer.

[255] Koenker, R., 1981. "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics*, 17, 107 – 112.

[256] Koenker, R., 2004. "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91, 74 – 89.

[257] Koenker, R. and G. Bassett, 1978. "Regression Quantiles," *Econometrica*, 46, 107 – 112.

[258] Koenker, R. and Z. Xiao, 2002. "Inference on the Quantile Regression Process," *Econometrica*, 81, 1583 – 1612.

[259] Koenker, R., and Z. Xiao, 2004. "Unit Root Quantile Regression Inference," *Journal of the American Statistical Association*, 99, 775 – 787.

[260] Koenker, R. and Z. Xiao, 2006. "Quantile Autoregression," *Journal of the American Statistical Association*, with discussion and rejoinder, 475, 980 – 1006.

[261] Koop, G., 2003. *Bayesian Econometrics*. New York: John Wiley and Sons.

[262] Krugman, P., 1991. "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, 483 – 499.

[263] Kumbhakar, S., and K. Lovell. *Stochastic Frontier Analysis*. New York: Cambridge University Press, 2000.

[264] Kwiatkowski, D., P. Phillips, P. Schmidt and Y. Shin, 1992. "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159 – 178.

[265] Lalive, R., J. van Ours, and J. Zweimuller, 2006. "How Changes in Financial Incentives Affect the Duration of Unemployment," *Review of Economic Studies*, 73, 1009 – 1038.

[266] Lalonde, R. J., 1986. "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604 – 620.

[267] Lancaster, T., 1990. *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

[268] Lancaster, T., 2004. *An Introduction to Modern Bayesian Inference*. Oxford: Oxford University Press.

[269] Lee, L. F., 2004. "Asymptotic Distributions of Quasi-maximum Likelihood Estimators for Spatial Autoregressive Models," *Econometrica*, 72, 1899 – 1925.

[270] Lee, L. F., and J. Yu, 2010. "Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects," *Journal of Econometrics*, 154, 165 – 185.

[271] Leigh, D., and A. Gill, 1997, "Labor Market Returns to Community Colleges: Evidence from Returning Adults," *Journal of Human Resources*, 32, 334 – 353.

[272] Levin, A., C. Lin and C. Chu, 2002. "Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties," *Journal of Econometrics*, 108, 1 – 24.

[273] Levitt, S., 1997. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 87, 270 – 290.

[274] Levitt, S., 2002. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review*, 92, 1244 – 1250.

[275] Li, H. and L. Zhou, 2005. "Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China," *Journal of Public Economics*, 89, 1743 – 1762.

[276] Lin, D. Y., and L. J. Wei. 1989. "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 84, 1074 – 1078.

[277] Lin, J., 1992. "Rural Reforms and Agricultural Growth in China," *American Economic Review*, 82, 34 – 51.

[278] Liu, R., 1988. "Bootstrap Procedures under Some Non-iid Models," *Annals of Statistics*, 16, 1696 – 1708.

[279] Ljung and Box, 1979. "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 66, 265 – 270.

[280] Long, J. Scott, and Jeremy Freese, 2006. *Regression Models for Categorical Dependent Variables Using Stata*, 2nd edition. College Station, TX: Stata Press.

[281] Lutkepohl, H., 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

[282] Maddala, G. S., 2001. *Introduction to Econometrics*, 3rd Edition. Chichester, West Sussex, England: John Wiley & Sons Ltd.

[283] Maddala, G. S., 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

[284] Mankiw, N. G., D. Romer and D. Weil, 1992. "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 107, 407 – 437.

[285] McCall, B. P., 1996. "Unemployment Insurance Rules, Joblessness, and Part-Time Work," *Econometrica*, 64, 647 – 682.

[286] McCrary, J., 2002. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment," *American Economic Review*, 92, 1236 – 1243.

[287] McCrary, J., 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698 – 714.

[288] McFadden, D., 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers of Econometrics*, P. Zarembka (Ed.), New York, Academic Press.

[289] McFadden, D., 1978. "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, 75 – 96. A. Karlquist, L. Lundquist, F. Snickars, and J. W. Weibull et al. (Eds.), Amsterdam, New York, North-Holland.

[290] Miguel, Edward, Shanker Satyanath, and Ernest Sergenti, 2004. "Economic Shocks and Civil Conflict: An Instrumental Variable Approach," *Journal of Political Economy*, 112, 725 – 753.

[291] Moran, P., 1950. "Notes on Continuous Stochastic Phenomena," *Biometrika*, 37, 17 – 23.

[292] Morgan, S., and C. Winship, 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social*

Research. Cambridge University Press.

[293] Munnell, A. H., 1990. "Why Has Productivity Growth Declined? Productivity and Public Investment," *New England Economic Review*, January/February, 3 – 22.

[294] Munnell, A. H., G. Tootell, L. Browne, and J. McEneaney, 1996. "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, 86, 25 – 53.

[295] Nadaraya, E., 1964. "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141 – 142.

[296] Nelson, W., 1972. "Theory and Application of Hazard Plotting for Censored Failure Data," *Technometrics*, 14, 945 – 965.

[297] Nelson, C. R., and C. I. Plosser, 1982. "Trends versus Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, 10, 139 – 162.

[298] Newey, W. K. 1987. "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 36, 231 – 250.

[299] Newey, W. and D. McFadden, 1994. "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol. IV, edited by R. Engle and D. McFadden, 2112 – 2245, Elsevier Science.

[300] Newey, W. and K. West, 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703 – 708.

[301] Ng, S., and P. Perron, 1995. "Unit Root Tests in ARMA Models with Data-dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90, 268 – 281.

[302] Ng, S., and P. Perron, 2000. "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power," *Econometrica*, 69, 1519 – 1554.

[303] Nerlove, M., 1963. "Returns to Scale in Electricity Supply," in C. Christ (ed.), *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*. Stanford: Stanford University Press.

[304] Nickell, S. J., 1981. "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417 – 1426.

[305] O'Connell, P. G. J., 1998. "The Overvaluation of Purchasing Power Parity," *Journal of International Economics*, 44, 1 – 19.

[306] Ord, J. K., 1975. "Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association*, 70, 120 – 126.

[307] Pesaran, M. H., 2004. "General Diagnostic Tests for Cross Section Dependence in Panels," Cambridge Working Papers in Economics, 0435, University of Cambridge.

[308] Phillips, P., 1986. "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33, 311 – 340.

[309] Phillips, P., 1987. "Time Series Regressions with a Unit Root," *Econometrica*, 55, 277 – 301.

[310] Phillips, P. and P. Perron, 1988. "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335 – 346.

[311] Pindyck, R. S., and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts*. 4th ed. New York: McGraw-Hill.

[312] Poirier, D. J., 1980. "Partial Observability in Bivariate Probit Models," *Journal of Econometrics*, 12, 209 – 217.

[313] Poirier, D. J., 1995. *Intermediate Statistics and Econometrics*. Cambridge, MA: MIT Press.

[314] Powell, J. L., 1984. "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303 – 325.

[315] Prais, S. and C. Winsten, 1954. "Trend Estimation and Serial Correlation," Cowles Commission Discussion Paper No. 383.

[316] Pregibon, D., 1979. "Data Analytic Methods for Generalized Linear Models," PhD dissertation, University of

Toronto.

[317] Pregibon, D., 1980. "Goodness of Link Tests for Generalized Linear Models," *Applied Statistics*, 29, 15 – 24.

[318] Ramsey, J. B., 1969. "Tests for Specification Errors in Classical Linear Least-squares Regression Analysis," *Journal of the Royal Statistical Society, Series B*, 31, 350 – 371.

[319] Rivers, D., and Q. H. Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347 – 366.

[320] Robinson, P., 1988. "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931 – 954.

[321] Romer, C. and D. Romer, 2004. "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055 – 1084.

[322] Roodman, D., 2009. "How to Do xtabond2: An Introduction to Difference and System GMM in Stata," *The Stata Journal*, 9, 86 – 136.

[323] Rosenblatt, M., 1956. "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, 27, 832 – 837.

[324] Rosenbaum, P. R., 2010. *Observational Studies*. Springer.

[325] Rosenbaum, P. R., and D. B. Rubin, 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 70, 41 – 55.

[326] Rosenbaum, P. R. and D. B. Rubin, 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, 39, 33 – 38.

[327] Rosenzweig, M. and K. Wolpin, 2000. "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, 38, 827 – 874.

[328] Rothstein, J., 2007. "Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)," *American Economic Review*, 97, 2026 – 2037.

[329] Rubin, D. B., 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688 – 701.

[330] Schwarz, G., 1978. "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461 – 464.

[331] Schwert, G., 1989. "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 7, 147 – 159.

[332] Shapiro, M., and M. Watson, 1988. "Sources of Business Cycle Fluctuations," *NBER Macroeconomics Annual*, 3, 111 – 156.

[333] Shea, J., 1997. "Instrumental Relevance in Multivariate Linear Models: A Simple Measure," *Review of Economics and Statistics*, 79, 348 – 352.

[334] Shen, Y. and Y. Yao, 2008. "Does Grassroots Democracy Reduce Income Inequality in China?" *Journal of Public Economics*, 92, 2182 – 2198.

[335] Sims, C., 1980. "Macroeconomics and Reality," *Econometrica*, 48, 1 – 48.

[336] Sims, C., 1981. "An Autoregressive Index Model for the U. S. 1948 – 1975", in J. Kmenta & J. B. Ramsey (eds), *Large-Scale Macro-Econometric Models*, North-Holland, Amsterdam, pp. 283 – 327.

[337] Sims, C., 1986. "Are Forecasting Models Usable for Policy Analysis?" *Quarterly Review, Federal Reserve Bank of Minneapolis*, 10, 2 – 16.

[338] Sims, C., 1992, "Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy," *European Economic Review*, 36, 975 – 1011.

[339] Skeels, C. L., and F. Vella, 1999. "A Monte Carlo Investigation of the Sampling Behavior of Conditional Moment Tests in Tobit and Probit Models," *Journal of Econometrics*, 92, 275 – 294.

[340] Small, K., and C. Hsiao, 1985. "Multinomial Logit Specification Tests," *International Economic Review*, 26, 619 – 627.

[341] Stevenson, R. E., 1980. "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics*, 13, 57 - 66.

[342] Stigler, G. and G. Becker, 1977. "De Gustibus Non Est Disputandum," *American Economic Review*, 67, 76 - 90.

[343] Stock, J. and M. Watson, 1993. "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61, 783 - 820.

[344] Stock, J. and M. Watson, 2001. "Vector Autoregressions," *Journal of Economic Perspective*, 15, 101 - 115.

[345] Stock, J. and M. Watson, 2011. *Introduction to Eonometrics*, 3rd edition, Boston: Addison-Wesley.

[346] Stock, J. and M. Yogo, 2005. "Testing for Weak Instruments in Linear IV Regression," in D. Andrew and J. Stock (ed.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80 - 108. Cambridge: Cambridge University Press.

[347] Swamy, P., 1970. "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311 - 323.

[348] Theil, H., 1953. "Repeated Least Squares Applied to Complete Equation System," *The Hague, Central Planning Bureau*, mimeo.

[349] Thistlethwaite, D., and D. T. Campbell, 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309 - 51, 317.

[350] Tobin, J., 1958. "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24 - 36.

[351] Tobler, W., 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography*, 46, 234 - 240.

[352] Tse, Y. K., and A. K. C. Tsui, 2002. "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-varying Correlations," *Journal of Business and Economic Statistics*, 20, 351 - 362.

[353] Tukey, J. W., 1949. "One Degree of Freedom for Non-additivity," *Biometrics*, 5, 232 - 242.

[354] Van den Berg, G., 2001. "Duration Models: Specification, Identification, and Multiple Durations," in *Handbook of Econometrics*, Volume 5, ed. J. J. Heckman and E. Leamer. Amsterdam: North Holland, 3381 - 3460.

[355] Van der Klaauw, W., 2008. "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22, 219 - 245.

[356] Varian, H., 2001. "What I Have Learn about Writing Economics," available online at http://people.ischool.berkeley.edu/~hal/Papers/writing-economics.html.

[357] Varian, H., 2009. "How to Build an Economic Model in Your Spare Time," available online at http://people.ischool.berkeley.edu/~hal/Papers/how.pdf.

[358] Verbeek, M., 2004. *A Guide to Modern Econometrics*, 2nd edition. Chichester, West Sussex, England: John Wiley & Sons Ltd.

[359] Vuong, Q. 1989. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." *Econometrica*, 57, 307 - 334.

[360] Waldman, M., S. Nicholson, and N. Adilov, 2006. "Does Television Cause Autism?" Cornell University, working paper.

[361] Waldman, M., S. Nicholson, N. Adilov, and J. Williams, 2008. "Autism Prevalence and Precipitation Rates in California, Oregon, and Washington Counties," *Archives of Pediatrics & Adolescent Medicine*, 162, 1026 - 1034.

[362] Wang, Q., and N. Wu, 2012. "Menu-driven X-12-ARIMA Seasonal Adjustment in Stata," *The Stata Journal*, 12, 214 - 241.

[363] Watson, G., 1964. "Smooth Regression Analysis," *Sankhya A*, 26, 359 - 372.

[364] Windmeijer, F. 2005. "A Finite Sample Correction for the Variance of Linear Efficient Two-step GMM Estimators," *Journal of Econometrics*, 126, 25 - 51.

[365] White, H., 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817 - 838.

[366] Wooldridge, J., 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

[367] Wooldridge, J., 2003. *Introductory Econometrics: A Modern Approach*, 2nd edition. South-Western, a division of Thomson Learning.

[368] Wooldridge, J., 2009. *Introductory Econometrics: A Modern Approach*, 4th edition. South-Western, a part of Cengage Learning.

[369] Wooldridge, J., 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd edition. Cambridge, MA: MIT Press.

[370] Wu, C., 1986. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261 - 1295.

[371] Wu, D., 1973. "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrics*, 41, 733 - 775.

[372] Xiao, Z., 2012. "Time Series Quantile Regression," *Handbook of Statistics*, Vol. 30, Elsevier.

[373] Zellner, A., 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias," *Journal of the American Statistical Association*, 58, 977 - 992.

[374] Zellner, A., 1971. *Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

[375] Zellner, A., and N. S. Revankar. 1969. "Generalized Production Functions," *Review of Economic Studies*, 36, 241 - 250.

[376] Zellner, A. and H. Theil, 1962. "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, 30, 63 - 68.

数 学 符 号

数学符号	中文名称	英文名称
β	参数 β 的真实值	True value of parameter β
$\hat{\beta}$	参数 β 的估计值	Estimated value of parameter β
$\tilde{\beta}$	参数 β 的假想值	Hypothetical value of parameter β
$\equiv$	定义	Definition
$\boldsymbol{A}\otimes\boldsymbol{B}$	克罗内克尔乘积	Kronecker product
∞	无穷大	Infinity
$\propto$	成正比	Proportional to
$\xrightarrow{d}$	依分布收敛	Converge in distribution
$\xrightarrow{p}$	依概率收敛	Converge in probability
$\lvert x\rvert$	x 的绝对值	Absolute value of x
$\mathbf{1}(x)$	示性函数	Indicator Function
$\exp(x)$	指数函数	Exponential Function
$\boldsymbol{A}'$ 或 $\boldsymbol{A}^{\mathrm{T}}$	矩阵 $\boldsymbol{A}$ 的转置	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	矩阵 $\boldsymbol{A}$ 的逆矩阵	Inverse of matrix $\boldsymbol{A}$
$\lvert\boldsymbol{A}\rvert$	矩阵 $\boldsymbol{A}$ 的行列式	Determinant of matrix $\boldsymbol{A}$
$\mathrm{Rank}(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的秩	Rank of matrix $\boldsymbol{A}$
$\mathrm{Trace}(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的迹	Trace of matrix $\boldsymbol{A}$
Δ	差分算子	Difference operator
L	滞后算子	Lag operator
$\sum_{t=1}^{T}$	求和算子	Summation operator
$\prod_{t=1}^{T}$	求积算子	Product operator
$\int_a^b$	定积分	Definite integral
$\frac{\mathrm{d}y}{\mathrm{d}x}$	y 对 x 的导数	Differentiation of y with respect to x
$\frac{\partial y}{\partial x}$	y 对 x 的偏导数	Partial differentiation of y with respect to x
$f'(x)$	函数 $f(x)$ 的一阶导数	First order differentiation of $f(x)$
$f''(x)$	函数 $f(x)$ 的二阶导数	Second order differentiation of $f(x)$
Var	方差(矩阵)	Variance (Covariance matrix)
Avar	渐近方差(矩阵)	Asymptotic variance (matrix)

续表

数学符号	中文名称	英文名称
Bias	偏差	Bias
Corr	相关系数	Correlation coefficient
Cov	协方差(矩阵)	Covariance (matrix)
E	(数学)期望	(Mathematical) Expectation
max	最大值	maximize
min	最小值	minimize
MSE	均方误差	Mean Squared Error
P(A)	事件 A 发生的概率	Probability of event A
$A \cap B$	事件 A 与 B 的交集	Intersection of event A and B
$A \cup B$	事件 A 与 B 的并集	Union of event A and B
P($A \mid B$)	给定事件 B 发生的条件下,事件 A 发生的概率	Probability of event A, given event B
plim	依概率收敛	Converge in probability
lim	收敛(微积分)	Converge (as in Calculus)
s. t.	受约束	Subject to
vech	半向量化算子	half vectorization

英文缩写

英文缩写	中文名称	英文名称
2SLS	二阶段最小二乘法	Two Stage Least Square
3SLS	三阶段最小二乘法	Three Stage Least Square
ACF	自相关函数	Autocorrelation Function
ADF	ADF 检验	Augmented Dickey-Fuller Test
ADL	自回归分布滞后模型	Autoregressive Distributed Lag Model
AFT	加速失效时间模型	Accelerated Failure Time Model
AIC	赤池信息准则	Akaike Information Criterion
AR(p)	p 阶自回归过程	Autoregressive Process of order p
ARCH	自回归条件异方差过程	Autoregressive Conditional Heteroskedasticity Process
ARCH-M	“ARCH-in-Mean”模型	ARCH-in-Mean Model
ARMA	自回归移动平均过程	Autoregressive Moving Average Process
ARIMA	自回归单整移动平均过程	Autoregressive Integrated Moving Average Process
AS	绝对值可加总	Absolutely Summable
ATE	平均处理效应	Average Treatment Effect
ATT	参与者平均处理效应	Average Treatment Effect on the Treated
ATU	非参与者平均处理效应	Average Treatment Effect on the Untreated
BG	BG 自相关检验	Breusch-Godfrey Test for Autocorrelation
BIC	贝叶斯信息准则	Bayesian Information Criterion
BLUE	最佳线性无偏估计	Best Linear Unbiased Estimator
BP	BP 异方差检验	Breusch-Pagan Test for Heteroskedasticity
BUE	最佳无偏估计	Best Unbiased Estimator
cdf	累积分布函数	cumulative distribution function
CIRF	累积脉冲响应函数	Cumulative Impulse Response Function
CLRM	古典线性回归模型	Classical Linear Regression Model
CO	Cochrane-Orcutt 估计法	Cochrane-Orcutt estimation
CV	交叉核实法	Cross Validation
DF	DF 检验	Dickey-Fuller Test
DF-GLS	DF-GLS 单位根检验	DF-GLS Unit Root Test
DGP	数据生成过程	Data Generating Process
DD	双重差分估计量	Differences-in-differences Estimator
DDD	三重差分估计量	Differences-in-differences-in-differences Estimator

续表

英文缩写	中文名称	英文名称
DPD	动态面板数据	Dynamic Panel Data
DVECH	对角半向量化 GARCH	Diagonal Vech GARCH
DW	DW 统计量	Durbin-Watson Statistic
DWH	德宾 - 吴 - 豪斯曼检验	Durbin-Wu-Hausman Test
ECM	误差修正模型	Error Correction Model
EG-ADF	EG-ADF 检验	EG-ADF Test
EGARCH	指数 GARCH	Exponential GARCH
FE	固定效应模型	Fixed Effects Model
FEVD	预测误差方差分解	Forecast-error Variance Decomposition
FGLS	可行广义最小二乘法	Feasible Generalized Least Square
FIML	全信息最大似然估计法	Full Information Maximum Likelihood Estimation
FWLS	可行加权最小二乘法	Feasible Weighted Least Square
GARCH	广义自回归条件异方差过程	Generalized Autoregressive Conditional Heteroskedasticity Process
GLS	广义最小二乘法	Generalized Least Square
GMM	广义矩估计	Generalized Method of Moments
H_0	原假设、零假设	Null Hypothesis
H_1	替代假设、备择假设	Alternative Hypothesis
HAC	异方差自相关稳健的标准误	Heteroskedasticity and Autocorrelation Consistent Standard Error
HQIC	汉南 - 昆信息准则	Hannan-Quinn Information Criterion
HT	HT 面板单位根检验	Harris and Tzavalis (1999) Panel Unit Root Test
iid	独立同分布	independently identically distributed
ILS	间接最小二乘法	Indirect Least Square
IMR	反米尔斯比率	Inverse Mill's Ratio
IMSE	积分均方误差	Integrated Mean Squared Error
IPS	IPS 面板单位根检验	Im, Pesaran and Shin(2002) Panel Unit Root Test
IRF	脉冲响应函数	Impulse Response Function
IV	工具变量(法)	Instrumental Variable (Estimation)
JB	雅克 - 贝拉正态检验	Jarque-Bera Test for Normality
KPSS	KPSS 平稳性检验	Kwiatkowski, Phillips, Schmidt and Shin (1992) Stationarity Test
LAD	最小绝对离差估计量	Least Absolute Deviation Estimator
LHS	方程左边	Left Hand Side
LIML	有限信息最大似然估计法	Limited Information Maximum Likelihood Estimation
LLC	LLC 面板单位根检验	Levin, Lin and Chu (2002) Panel Unit Root Test

续表

英文缩写	中文名称	英文名称
LM	拉格朗日乘子检验	Lagrange Multiplier Test
Lowess	局部加权散点光滑估计量	Locally weighted scatterplot smoothing
LPM	线性概率模型	Linear Probability Model
LR	似然比检验	Likelihood Ratio Test
LSDV	最小二乘虚拟变量模型	Least Square Dummy Variable Model
LSDVC	偏差校正 LSDV	Biased-corrected LSDV
MA(q)	q 阶移动平均过程	Moving Average Process of Order q
MC	蒙特卡罗方法	Monte Carlo Methods
MCMC	马尔可夫链蒙特卡罗法	Markov Chain Monte Carlo Methods
MDS	鞅差分序列	Martingale Difference Sequence
MGARCH	多维 GARCH	Multivariate GARCH
MLE	最大似然估计法	Maximum Likelihood Estimation
MM	矩估计	Method of Moments
MSE	均方误差	Mean Squared Error
MSL	最大模拟似然法	Maximum Simulated Likelihood
MSM	模拟矩估计	Method of Simulated Moments
NLS	非线性最小二乘法	Nonlinear Least Square
OIM	观测信息矩阵	Observed Information Matrix
OIRF	正交化的脉冲响应函数	Orthogonalized Impulse Response Function
OLS	普通最小二乘法	Ordinary Least Square
OPG/ BHHH	梯度向量外积	Outer Product of Gradients
OS	一可加总	One-summable
PA	总体平均估计量	Population-averaged Estimator
PSM	倾向得分匹配	Propensity Score Matching
PACF	偏自相关函数	Partial Autocorrelation Function
pdf	概率密度函数	probability density function
PH	比例风险模型	Proportional Hazard Model
PL	部分线性模型	Partially Linear Model
PP	Phillips-Perron 单位根检验	Phillips-Perron Unit Root Test
PW	Prais-Winsten 估计法	Prais-Winsten Estimation
QLR	匡特似然比统计量	Quandt Likelihood Ratio
QMLE	准最大似然估计	Quasi Maximum Likelihood Estimation
QQ plot	分位数 - 分位数图	Quantile-quantile plot
QR	分位数回归	Quantile Regression
R^2	拟合优度、可决系数	Goodness of Fit, Coefficient of Determination
$\overline{R}^2$	校正拟合优度	Adjusted Goodness of Fit

续表

英文缩写	中文名称	英文名称
R_{uc}^2	非中心 R^2	Uncentered R^2
RDD	断点回归设计	Regression Discontinuity Design
RE	随机效应模型	Random Effects Model
RHS	方程右边	Right Hand Side
RLS	约束最小二乘法	Restricted Least Square
SAR	空间自回归	Spatial Autoregression
SARAR	带空间自回归误差项的空间自回归模型	Spatial Autoregressive Model with Spatial Autoregressive Disturbances
SIC	施瓦茨信息准则	Schwarz Information Criterion
SSR	残差平方和	Sum of Squared Residuals
SUR	似不相关回归	Seemingly Unrelated Regression Estimation
TARCH	门限 GARCH	Threshold GARCH
VAR	向量自回归	Vector Autoregression
vech	半向量化算子	half vectorization
VIF	方差膨胀系数	Variance Inflation Factor
VMA	向量移动平均过程	Vector Moving Average Process
Wald	沃尔德检验	Wald Test
WLS	加权最小二乘法	Weighted Least Square
WN	白噪声	White Noise
ZINB	零膨胀负二项回归	Zero-inflated Negative Binomial Regression
ZIP	零膨胀泊松回归	Zero-inflated Poisson Regression

[General Information]
书名=高级计量经济学及Stata应用　第2版
SS号=13526050
作者=陈强编著
页数=669
ISBN号=2014.04
原书定价=北京：高等教育出版社
出版社=978-7-04-032983-4
出版日期=
丛书名=
主题词=计量经济学-高等学校-教材-经济计量分析-应用软件
起始页=1

封面
书名
版权
前言
目录
第1章　绪论／1
1.1什么是计量经济学／1
1.2经济数据的特点与类型／2
第2章　概率统计回顾／3
2.1概率与条件概率／3
2.2分布与条件分布／4
2.3随机变量的数字特征／5
2.4迭代期望定律／8
2.5随机变量无关的三个层次概念／9
2.6常用连续型统计分布／9
2.7统计推断的思想／11
习题／12
附录／12
第3章　小样本OLS／13
3.1古典线性回归模型的假定／13
3.2　OLS的代数推导／14
3.3　OLS的几何解释／17
3.4拟合优度／17
3.5　OLS的小样本性质／18
3.6对单个系数的t检验／20
3.7对线性假设的F检验／23
3.8　F统计量的似然比原理表达式／25
3.9分块回归与偏回归（选读）／26
3.10预测／27
习题／28
附录／29
第4章　Stata简介／30
4.1为什么使用Stata／30
4.2　Stata的窗口／30
4.3　Stata操作实例／31
4.4　Stata命令库的更新／46
4.5进一步学习Stata的资源／47
习题／48
第5章　大样本OLS／49
5.1为何需要大样本理论／49
5.2随机收敛／49
5.3大数定律与中心极限定理／51

5.4统计量的大样本性质／52
5.5渐近分布的推导／53
5.6随机过程的性质／53
5.7大样本OLS的假定／57
5.8　OLS的大样本性质／58
5.9线性假设的大样本检验／60
5.10大样本OLS的Stata命令及实例／61
习题／63
附录／63
第6章　最大似然估计法／66
6.1最大似然估计法的定义／66
6.2线性回归模型的最大似然估计／68
6.3最大似然估计的数值解／69
6.4信息矩阵与无偏估计的最小方差／70
6.5最大似然法的大样本性质／71
6.6最大似然估计量的渐近协方差矩阵／74
6.7三类渐近等价的统计检验／75
6.8准最大似然估计法／78
6.9对正态分布假设的检验／80
6.10最大似然估计法的Stata命令及实例／80
习题／84
附录／84
第7章　异方差与GLS／87
7.1异方差的后果／87
7.2异方差的例子／87
7.3异方差的检验／88
7.4异方差的处理／90
7.5处理异方差的Stata命令及实例／93
7.6　Stata命令的批处理／96
习题／98
附录／98
第8章　自相关／100
8.1自相关的后果／100
8.2自相关的例子／101
8.3自相关的检验／101
8.4自相关的处理／103
8.5处理自相关的Stata命令及实例／108
习题／115
第9章　模型设定与数据问题／116
9.1遗漏变量／116
9.2无关变量／117
9.3建模策略：“由小到大”还是“由大到小”／118

9.4解释变量个数的选择／118
9.5对函数形式的检验／120
9.6多重共线性／123
9.7极端数据／124
9.8虚拟变量／126
9.9经济结构变动的检验／127
9.10缺失数据与线性插值／132
9.11变量单位的选择／133
习题／133
附录／133
第10章　工具变量，2SLS与GMM／135
10.1解释变量与扰动项相关的例子／135
10.2工具变量法作为一种矩估计／138
10.3二阶段最小二乘法／140
10.4有关工具变量的检验／141
10.5　GMM的假定／146
10.6　GMM的推导／147
10.7　GMM的大样本性质／148
10.8如何获得工具变量／151
10.9　MLE也是GMM／152
10.10工具变量法的Stata命令及实例／153
习题／167
附录／167
第11章　二值选择模型／169
11.1离散被解释变量的例子／169
11.2二值选择模型／169
11.3二值选择模型的微观基础／177
11.4二值选择模型中的异方差问题／178
11.5稀有事件偏差（选读）／179
11.6含内生变量的Probit模型（选读）／183
11.7双变量Probit模型（选读）／187
11.8部分可观测的双变量Probit模型（选读）／189
习题／190
第12章　多值选择模型／192
12.1多项Logit与多项Probit／192
12.2条件Logit模型／193
12.3混合Logit模型／193
12.4嵌套Logit／205
习题／208
第13章　排序与计数模型／209
13.1排序模型／209
13.2泊松回归／211

13.3负二项回归／213
13.4零膨胀泊松回归与负二项回归／215
13.5计数模型的Stata实例／215
习题／222
第14章　受限被解释变量／223
14.1断尾回归／223
14.2零断尾泊松回归与负二项回归／226
14.3随机前沿模型（选读）／228
14.4偶然断尾与样本选择／235
14.5归并回归／238
14.6归并数据的两部分模型／243
14.7含内生解释变量的Tobit模型（选读）／246
习题／248
附录／248
第15章　短面板／250
15.1面板数据的特点／250
15.2面板数据的估计策略／251
15.3混合回归／252
15.4个体固定效应模型／252
15.5时间固定效应／253
15.6一阶差分法／254
15.7随机效应模型／254
15.8组间估计量／255
15.9拟合优度的度量／255
15.10非平衡面板／256
15.11究竟该用固定效应还是随机效应模型／257
15.12个体时间趋势／257
15.13短面板的Stata命令及实例／258
习题／271
第16章　长面板与动态面板／272
16.1长面板的估计策略／272
16.2面板校正标准误／272
16.3仅解决组内自相关的FGLS／274
16.4全面FGLS／278
16.5组间异方差的检验／279
16.6组内自相关的检验／280
16.7组间同期相关的检验／281
16.8变系数模型／283
16.9面板工具变量法／287
16.10豪斯曼-泰勒估计量（选读）／288
16.11动态面板／289
16.12动态面板的Stata命令及实例／291

16.13偏差校正LSDV法／300
16.14重复截面数据与组群分析／301
习题／302
第17章　非线性面板／303
17.1面板二值选择模型／303
17.2面板二值选择模型的随机效应估计／304
17.3面板二值选择模型的固定效应估计／305
17.4面板二值选择模型的Stata实例／307
17.5面板泊松回归／313
17.6面板负二项回归／314
17.7面板计数模型的Stata实例／315
17.8面板Tobit／325
17.9面板随机前沿模型／327
习题／332
第18章　随机实验与自然实验／334
18.1实验数据／334
18.2理想的随机实验／335
18.3引入更多的解释变量／335
18.4随机实验执行过程中可能出现的问题／336
18.5自然实验／337
18.6双重差分法／339
18.7三重差分法／343
18.8观测数据的处理效应／344
习题／345
第19章　蒙特卡罗法与自助法／346
19.1蒙特卡罗法的思想与用途／346
19.2蒙特卡罗法实例：模拟中心极限定理／347
19.3蒙特卡罗法实例：服从卡方分布的扰动项／348
19.4蒙特卡罗积分／349
19.5最大模拟似然法与模拟矩估计／350
19.6自助法的思想与用途／351
19.7自助法的分类／352
19.8使用自助法估计标准误／352
19.9使用自助法进行区间估计／353
19.10使用自助法进行假设检验／353
19.11自助法的一致性（选读）／354
19.12异方差情况下的自助法／354
19.13面板数据与时间序列的自助法／355
19.14自助法的Stata命令／355
19.15使用自助法进行稳健的豪斯曼检验／356
习题／358
附录／358

第20章　平稳时间序列／361
20.1时间序列的数字特征／361
20.2自回归模型／362
20.3移动平均模型／364
20.4　ARMA／364
20.5自回归分布滞后模型／365
20.6　ARMA模型的Stata命令及实例／366
20.7误差修正模型／371
20.8　MA（∞）与滞后算子／372
20.9向量自回归过程／375
20.10　VAR的脉冲响应函数／377
20.11预测误差的方差分解／380
20.12格兰杰因果检验／381
20.13面板格兰杰因果检验／381
20.14　VAR的Stata命令及实例／381
20.15季节调整／399
习题／407
第21章　单位根与协整／409
21.1非平稳序列／409
21.2　ARMA的平稳性／410
21.3　VAR的平稳性／411
21.4单位根所带来的问题／411
21.5单位根检验与平稳性检验／414
21.6单位根检验的Stata实例／418
21.7面板单位根检验／422
21.8协整的思想与初步检验／432
21.9　Beveridge-Nelson分解公式／433
21.10协整的定义与最大似然估计／434
21.11协整分析的Stata实例／437
习题／445
附录／445
第22章　自回归条件异方差模型／447
22.1条件异方差模型的例子／447
22.2　ARCH模型的性质／448
22.3　ARCH模型的MLE估计／449
22.4　GARCH模型／450
22.5何时使用ARCH或GARCH模型／451
22.6　ARCH与GARCH模型的扩展／451
22.7　ARCH与GARCH的Stata命令及实例／453
22.8多维GARCH模型（选读）／460
习题／467
第23章　似不相关回归／468

23.1单一方程估计与系统估计／468
23.2似不相关回归的假定／468
23.3　SUR的FGLS估计／470
23.4　SUR的假设检验／471
23.5似不相关回归的Stata命令及实例／471
23.6变系数面板数据的SUR估计／475
习题／478
附录／479
第24章　联立方程模型／482
24.1联立方程模型的结构式与简化式／482
24.2联立方程模型的识别／483
24.3单一方程估计法／486
24.4三阶段最小二乘法／487
24.5三阶段最小二乘法的Stata实例／489
24.6结构VAR／493
24.7　SVAR的Stata实例／496
习题／502
第25章　非线性回归与门限回归／503
25.1非线性最小二乘法／503
25.2非线性回归的Stata命令及实例／504
25.3门限回归／505
25.4面板数据的门限回归／507
25.5门限回归的计算机操作／508
习题／508
第26章　分位数回归／509
26.1为什么需要分位数回归／509
26.2总体分位数／509
26.3样本分位数／510
26.4分位数回归的估计方法／512
26.5分位数回归的Stata命令及实例／513
习题／517
第27章　非参数与半参数估计／518
27.1为什么需要非参数与半参数估计／518
27.2对密度函数的非参数估计／518
27.3核密度估计的性质／520
27.4最优带宽／521
27.5多元密度函数的核估计／523
27.6非参数核回归／523
27.7多元核回归／525
27.8　k近邻回归／525
27.9局部线性回归／526
27.10非参数估计的Stata命令及实例／526

27.11半参数估计／530
习题／533
附录／533
第28章　处理效应／537
28.1处理效应与选择难题／537
28.2通过随机分组解决选择难题／539
28.3依可测变量选择／539
28.4匹配估计量的思想／540
28.5倾向得分匹配／542
28.6倾向得分匹配的Stata实例／545
28.7偏差校正匹配估计量／555
28.8双重差分倾向得分匹配／557
28.9断点回归的思想／559
28.10精确断点回归／561
28.11模糊断点回归／563
28.12断点回归的Stata实例／565
28.13处理效应模型／570
习题／574
第29章　空间计量经济学／575
29.1地理学第一定律／575
29.2空间权重矩阵／575
29.3空间自相关／578
29.4空间自回归模型／583
29.5空间杜宾模型／586
29.6空间误差模型／586
29.7一般的空间计量模型／589
29.8含内生解释变量的SARAR模型／592
29.9空间面板模型／593
29.10空间计量方法的局限性／598
第30章　久期分析／599
30.1久期数据的处理方法／599
30.2风险p函数／599
30.3久期数据的归并问题／601
30.4描述性分析／602
30.5久期模型的最大似然估计／603
30.6比例风险模型／604
30.7加速失效时间模型／606
30.8　Cox模型／607
30.9比例风险模型的设定检验／610
30.10分层Cox模型／611
30.11随时间而变的解释变量／612
30.12不可观测的异质性／613

30.13其他久期分析模型／614
30.14久期分析的Stata命令及实例／615
习题／630
第31章　贝叶斯估计简介／631
31.1贝叶斯估计的思想／631
31.2贝叶斯定理／631
31.3贝叶斯估计的一个例子／632
31.4基于后验分布的统计推断／634
31.5先验分布的选择／635
31.6多元回归的贝叶斯分析／637
31.7马尔可夫链蒙特卡罗法／639
习题／640
第32章　如何做规范的实证研究／641
32.1计量理论与现实数据／641
32.2实证研究的主要步骤／642
32.3实证论文的结构／644
32.4计量实践的十诫／645
32.5结束语／646
习题／646
附录：常用数据来源／647
参考书目／649
数学符号／664
英文缩写／666