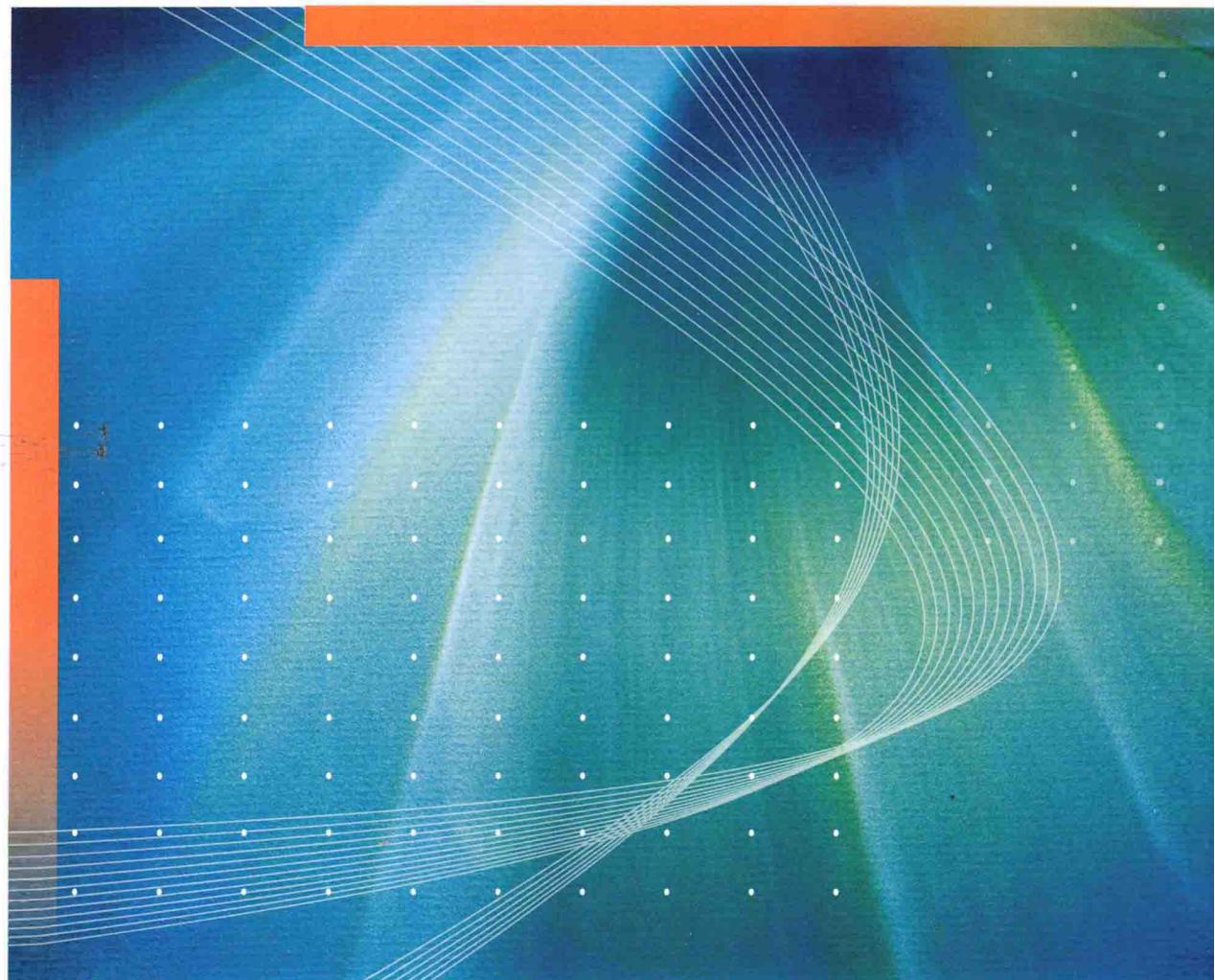


经济学、管理学类研究生教学用书

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



9 787040 329834 >

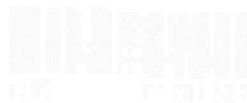
定价 59.00 元

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

GAOJI JILIAO JINGJIXUE JI STATA YINGYONG



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
—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
插图绘制 尹文军

责任编辑 施春花
责任校对 李大鹏

封面设计 赵阳
责任印制 张泽业

版式设计 马敬茹

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



目 录

8.2 自相关的例子	101	11.7 双变量 Probit 模型(选读)	187
8.3 自相关的检验	101	11.8 部分可观测的双变量 Probit 模型(选读)	189
8.4 自相关的处理	103	习题	190
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	12.1 多项 Logit 与多项 Probit	192
9.2 无关变量	117	12.2 条件 Logit 模型	193
9.3 建模策略：“由小到大”还是“由大 到小”	118	12.3 混合 Logit 模型	193
9.4 解释变量个数的选择	118	12.4 嵌套 Logit	205
9.5 对函数形式的检验	120	习题	208
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	14.1 断尾回归	223
10.2 工具变量法作为一种矩估计	138	14.2 零断尾泊松回归与负二项回归	226
10.3 二阶段最小二乘法	140	14.3 随机前沿模型(选读)	228
10.4 有关工具变量的检验	141	14.4 偶然断尾与样本选择	235
10.5 GMM 的假定	146	14.5 归并回归	238
10.6 GMM 的推导	147	14.6 归并数据的两部分模型	243
10.7 GMM 的大样本性质	148	14.7 含内生解释变量的 Tobit 模型 (选读)	246
10.8 如何获得工具变量	151	习题	248
10.9 MLE 也是 GMM	152	附录	248
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	15.1 面板数据的特点	250
11.2 二值选择模型	169	15.2 面板数据的估计策略	251
11.3 二值选择模型的微观基础	177	15.3 混合回归	252
11.4 二值选择模型中的异方差问题	178	15.4 个体固定效应模型	252
11.5 稀有事件偏差(选读)	179	15.5 时间固定效应	253
11.6 含内生变量的 Probit 模型 (选读)	183	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	19. 1 蒙特卡罗法的思想与用途	346
第 16 章 长面板与动态面板	272	19. 2 蒙特卡罗法实例:模拟中心极限定理	347
16. 1 长面板的估计策略	272	19. 3 蒙特卡罗法实例:服从卡方分布的扰动项	348
16. 2 面板校正标准误	272	19. 4 蒙特卡罗积分	349
16. 3 仅解决组内自相关的 FGLS	274	19. 5 最大模拟似然法与模拟矩估计	350
16. 4 全面 FGLS	278	19. 6 自助法的思想与用途	351
16. 5 组间异方差的检验	279	19. 7 自助法的分类	352
16. 6 组内自相关的检验	280	19. 8 使用自助法估计标准误	352
16. 7 组间同期相关的检验	281	19. 9 使用自助法进行区间估计	353
16. 8 变系数模型	283	19. 10 使用自助法进行假设检验	353
16. 9 面板工具变量法	287	19. 11 自助法的一致性(选读)	354
16. 10 豪斯曼 - 泰勒估计量(选读)	288	19. 12 异方差情况下的自助法	354
16. 11 动态面板	289	19. 13 面板数据与时间序列的自助法	355
16. 12 动态面板的 Stata 命令及实例	291	19. 14 自助法的 Stata 命令	355
16. 13 偏差校正 LSDV 法	300	19. 15 使用自助法进行稳健的豪斯曼检验	356
16. 14 重复截面数据与组群分析	301	习题	358
习题	302	附录	358
第 17 章 非线性面板	303	第 20 章 平稳时间序列	361
17. 1 面板二值选择模型	303	20. 1 时间序列的数字特征	361
17. 2 面板二值选择模型的随机效应估计	304	20. 2 自回归模型	362
17. 3 面板二值选择模型的固定效应估计	305	20. 3 移动平均模型	364
17. 4 面板二值选择模型的 Stata 实例	307	20. 4 ARMA	364
17. 5 面板泊松回归	313	20. 5 自回归分布滞后模型	365
17. 6 面板负二项回归	314	20. 6 ARMA 模型的 Stata 命令及实例	366
17. 7 面板计数模型的 Stata 实例	315	20. 7 误差修正模型	371
17. 8 面板 Tobit	325	20. 8 MA(∞) 与滞后算子	372
17. 9 面板随机前沿模型	327	20. 9 向量自回归过程	375
习题	332	20. 10 VAR 的脉冲响应函数	377
第 18 章 随机实验与自然实验	334	20. 11 预测误差的方差分解	380
18. 1 实验数据	334	20. 12 格兰杰因果检验	381
18. 2 理想的随机实验	335	20. 13 面板格兰杰因果检验	381
18. 3 引入更多的解释变量	335	20. 14 VAR 的 Stata 命令及实例	381
18. 4 随机实验执行过程中可能出现的问题	336	20. 15 季节调整	399
18. 5 自然实验	337	习题	407
18. 6 双重差分法	339	第 21 章 单位根与协整	409
18. 7 三重差分法	343	21. 1 非平稳序列	409
18. 8 观测数据的处理效应	344	21. 2 ARMA 的平稳性	410
习题	345	21. 3 VAR 的平稳性	411
第 19 章 蒙特卡罗法与自助法	346	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	25.1 非线性最小二乘法 ······	503
21.6 单位根检验的 Stata 实例 ······	418	25.2 非线性回归的 Stata 命令及实例 ······	504
21.7 面板单位根检验 ······	422	25.3 门限回归 ······	505
21.8 协整的思想与初步检验 ······	432	25.4 面板数据的门限回归 ······	507
21.9 Beveridge-Nelson 分解公式 ······	433	25.5 门限回归的计算机操作 ······	508
21.10 协整的定义与最大似然估计 ······	434	习题 ······	508
21.11 协整分析的 Stata 实例 ······	437	第 26 章 分位数回归 ······	509
习题 ······	445	26.1 为什么需要分位数回归 ······	509
附录 ······	445	26.2 总体分位数 ······	509
第 22 章 自回归条件异方差模型 ······	447	26.3 样本分位数 ······	510
22.1 条件异方差模型的例子 ······	447	26.4 分位数回归的估计方法 ······	512
22.2 ARCH 模型的性质 ······	448	26.5 分位数回归的 Stata 命令及实例 ······	513
22.3 ARCH 模型的 MLE 估计 ······	449	习题 ······	517
22.4 GARCH 模型 ······	450	第 27 章 非参数与半参数估计 ······	518
22.5 何时使用 ARCH 或 GARCH 模型 ······	451	27.1 为什么需要非参数与半参数估计 ······	518
22.6 ARCH 与 GARCH 模型的扩展 ······	451	27.2 对密度函数的非参数估计 ······	518
22.7 ARCH 与 GARCH 的 Stata 命令及实例 ······	453	27.3 核密度估计的性质 ······	520
22.8 多维 GARCH 模型(选读) ······	460	27.4 最优带宽 ······	521
习题 ······	467	27.5 多元密度函数的核估计 ······	523
第 23 章 似不相关回归 ······	468	27.6 非参数核回归 ······	523
23.1 单一方程估计与系统估计 ······	468	27.7 多元核回归 ······	525
23.2 似不相关回归的假定 ······	468	27.8 k 近邻回归 ······	525
23.3 SUR 的 FGLS 估计 ······	470	27.9 局部线性回归 ······	526
23.4 SUR 的假设检验 ······	471	27.10 非参数估计的 Stata 命令及实例 ······	526
23.5 似不相关回归的 Stata 命令及实例 ······	471	27.11 半参数估计 ······	530
23.6 变系数面板数据的 SUR 估计 ······	475	习题 ······	533
习题 ······	478	附录 ······	533
附录 ······	479	第 28 章 处理效应 ······	537
第 24 章 联立方程模型 ······	482	28.1 处理效应与选择难题 ······	537
24.1 联立方程模型的结构式与简化式 ······	482	28.2 通过随机分组解决选择难题 ······	539
24.2 联立方程模型的识别 ······	483	28.3 依可测变量选择 ······	539
24.3 单一方程估计法 ······	486	28.4 匹配估计量的思想 ······	540
24.4 三阶段最小二乘法 ······	487	28.5 倾向得分匹配 ······	542
24.5 三阶段最小二乘法的 Stata 实例 ······	489	28.6 倾向得分匹配的 Stata 实例 ······	545
24.6 结构 VAR ······	493	28.7 偏差校正匹配估计量 ······	555
24.7 SVAR 的 Stata 实例 ······	496	28.8 双重差分倾向得分匹配 ······	557
习题 ······	502	28.9 断点回归的思想 ······	559
第 25 章 非线性回归与门限回归 ······	503	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。

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

$$\ln W_i = \alpha + \beta S_i + \varepsilon_i \quad (1.1)$$

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

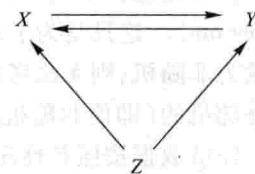


图 1.1 可能的因果关系

^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

$$\ln W_i = \alpha + \beta S_i + \gamma S_i^2 + \varepsilon_i \quad (1.2)$$

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 $P(A)$ ^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 ，则在出太阳的前提下降雨的“条件概率”(conditional probability)为

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.1)$$

其中，“ \cap ”表示事件的交集(intersection)，故 $P(A \cap B)$ 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。

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 $P(A|B) = P(A)$ ，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， $P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A)$ ，故

$$P(A \cap B) = P(A)P(B) \quad (2.2)$$

也可以将此式作为独立事件的定义。

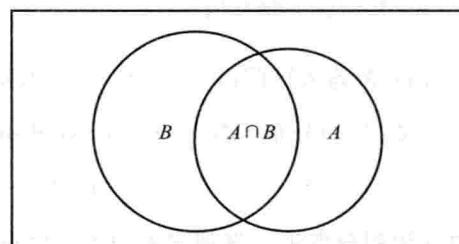


图2.1 条件概率示意图

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 $\{B_1, B_2, \dots, B_n\}$ ($n \geq 2$) 两两互不相容, $P(B_i) > 0$ ($\forall i = 1, \dots, n$), 且 $B_1 \cup B_2 \cup \dots \cup B_n$ 为必然事件 (即在 B_1, B_2, \dots, 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 $\{B_1, B_2, \dots, B_n\}$ 是否有任何关系),

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i) \quad (2.3)$$

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

$$\begin{array}{ccccccc} X & x_1 & x_2 & \cdots & x_k & \cdots \\ p & p_1 & p_2 & \cdots & p_k & \cdots \end{array} \quad (2.4)$$

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,

$$(i) f(x) \geq 0, \forall x;$$

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

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

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

$$F(x) \equiv P(-\infty < X \leq x) = \int_{-\infty}^x f(t) dt \quad (2.5)$$

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:

$$(i) f(x, y) \geq 0, \forall x, y;$$

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

$$(iii) (X, Y) \text{ 落入平面某区域 } D \text{ 的概率为 } P\{(X, Y) \in D\} = \iint_D f(x, y) dx dy.$$

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, \dots, X_n) 可以由联合密度函数 $f(x_1, x_2, \dots, x_n)$ 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) :

$$f_x(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad (2.6)$$

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:

$$f_y(y) = \int_{-\infty}^{+\infty} f(x, y) dx \quad (2.7)$$

定义二维随机向量 (X, Y) 的累积分布函数为

$$F(x, y) \equiv P(-\infty < X \leq x; -\infty < Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(t, s) dt ds \quad (2.8)$$

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 $\varepsilon \rightarrow 0^+$, 可以证明条件密度函数为(参见附录)

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

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 $p_k \equiv P(X=x_k)$ 的离散型随机变量 X , 其“期望”(expectation)^①为

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

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

$$E(X) \equiv \mu \equiv \int_{-\infty}^{+\infty} xf(x) dx \quad (2.11)$$

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 $E(X+Y) = E(X) + E(Y)$, $E(kX) = kE(X)$, 其中 k 为任意常数。

定义 随机变量 X 的“方差”(variance)为

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

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

$$\begin{aligned}\text{证明: } \text{Var}(X) &\equiv E[X - E(X)]^2 = E\{X^2 - 2E(X)X + [E(X)]^2\} \\ &= E[X^2] - 2[E(X)]^2 + [E(X)]^2 = E[X^2] - [E(X)]^2\end{aligned}$$

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

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

如果当随机变量 X 的取值大于(小于)其期望 $E(X)$ 时,随机变量 Y 的取值也倾向于大于(小于)其期望值 $E(Y)$,则 $\text{Cov}(X, Y) > 0$,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 $E(X)$ 时,随机变量 Y 的取值反而倾向于小于(大于)其期望值 $E(Y)$,则 $\text{Cov}(X, Y) < 0$,二者存在负相关。如果 $\text{Cov}(X, Y) = 0$,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:

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

协方差的运算也满足线性性,可以证明(参见习题)

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z) \quad (2.15)$$

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

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

可以证明, $-1 \leq \rho \leq 1$ 。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 $E(X)$ (即期望),二阶原点矩为 $E(X^2)$,三阶原点矩为 $E(X^3)$,四阶原点矩为 $E(X^4)$,等等。

定义 二阶中心矩为 $E[X - E(X)]^2$ (即方差),三阶中心矩为 $E[X - E(X)]^3$,四阶中心矩为 $E[X - E(X)]^4$,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 $E[(X - \mu)/\sigma]^3$ 。

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 $E[(X - \mu)/\sigma]^4$ 。

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 $E[(X-\mu)/\sigma]^4 - 3$ 。

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 $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$ 为“矩”(moment)。

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

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

在上式中,由于 y 已被积分积掉,故 $E(Y|x)$ 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

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

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

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 A (symmetric),如果对于任意 n 维非零列向量 c ,都有二次型 $c'A c \geq 0$,则称 A 为半正定矩阵。

定义 对于 $n \times n$ 对称矩阵 A ,如果对于任意 n 维非零列向量 c ,都有二次型 $c'A c > 0$,则称 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。从几何意义上讲,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 D , $D'D$ 为半正定矩阵。

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

$$c'(D'D)c = (\underbrace{c'D'}_{\text{平方和}})(Dc) = \underbrace{\overline{(Dc)}'Dc}_{\text{平方和}} \geq 0 \quad (2.19)$$

因此, $D'D$ 为半正定矩阵。

定义 设 $X = (X_1 X_2 \cdots X_n)'$ 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:

$$\text{Cov}(X) \equiv \text{Var}(X) \equiv E[(X - E(X))(X - E(X))']$$

$$= E \left[\begin{pmatrix} X_1 - E(X_1) \\ \vdots \\ X_n - E(X_n) \end{pmatrix} (X_1 - E(X_1))' \cdots (X_n - E(X_n))' \right]$$

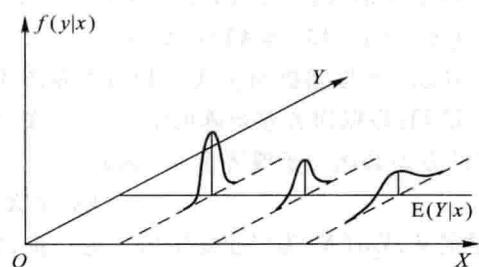


图2.2 条件期望与条件方差示意图

$$\begin{aligned}
 &= E \left(\begin{array}{cccccc} [X_1 - E(X_1)]^2 & \cdots & [X_1 - E(X_1)][X_n - E(X_n)] & & & \\ \vdots & & \vdots & & & \\ [X_1 - E(X_1)][X_n - E(X_n)] & \cdots & [X_n - E(X_n)]^2 & & & \end{array} \right) \quad (2.20) \\
 &= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}
 \end{aligned}$$

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

假设 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) $E(AX) = AE(X)$; (期望算子的线性性)
- (ii) $\text{Var}(X) = E(XX') - E(X)[E(X)]'$; (一维随机变量公式的推广)
- (iii) $\text{Var}(AX) = A\text{Var}(X)A'$.

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

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

$$\text{Var}(c'X) = c'\text{Var}(X)c \geq 0 \quad (2.21)$$

根据定义, $\text{Var}(X)$ 为半正定矩阵。进一步,如果希望任意线性组合 $c'X$ 的方差为正数(不考虑方差为 0 的退化情形),则一般假设 $\text{Var}(X)$ 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 $X = (X_1 X_2 \cdots X_n)'$ 为 n 维随机向量, $Y = (Y_1 Y_2 \cdots Y_m)'$ 为 m 维随机向量,则这两个随机向量之间的协方差矩阵为

$$\text{Cov}(X, Y) \equiv E[(X - E(X))(Y - E(Y))'] = E(XY') - E(X)E(Y') \quad (2.22)$$

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),

$$E(Y) = E_x[E(Y|x)] \quad (2.23)$$

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

$$\text{证明: LHS} = E(Y) = \int_{-\infty}^{+\infty} y f_y(y) dy$$

$$\begin{aligned}
 \text{RHS} &= E_x \left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_x(x)} dy \right] = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_x(x)} dy \right] f_x(x) dx \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x,y) dx dy = \int_{-\infty}^{+\infty} y \left[\int_{-\infty}^{+\infty} f(x,y) dx \right] dy = \text{LHS}
 \end{aligned}$$

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:

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

推而广之,对于任意函数 $g(\cdot)$,可以得到

$$E[g(Y)] = E_x E[g(Y)|x] \quad (2.25)$$

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 $\text{Cov}(X,Y) = 0$ 。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 $E(Y|x)$ 存在。如果 $E(Y|x)$ 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 $E(Y|x) = E(Y)$ (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 $E(Y|x)$ 不依赖于 X ,故 $E_x[E(Y|x)] = E(Y|x)$ 。根据迭代期望定律, $E(Y) = E_x[E(Y|x)] = E(Y|x)$ 。

(2) 假设 $E(Y|x) = E(Y)$,则显然 $E(Y|x)$ 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 $E(Y|x)$ 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 $\text{Cov}(X,Y) = 0$ 。

证明: $\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))]$ (协方差的定义)

$$= E_x E_y [(X - E(X))(Y - E(Y))|x] \quad (\text{迭代期望定律})$$

$$= E_x [(X - E(X)) E_y (Y - E(Y)|x)] \quad (\text{将 } X - E(X) \text{ 视为常数提出})$$

$$= E_x [(X - E(X)) (E(Y|x) - E(Y))] \quad (\text{期望算子的线性性})$$

$$= E_x [(X - E(X)) \cdot 0] = 0 \quad (\text{均值独立的定义})$$

总之,“相互独立” \Rightarrow “均值独立” \Rightarrow “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

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

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 $Z = \frac{X - \mu}{\sigma}$, 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 $\mathbf{X} = (X_1 X_2 \cdots X_n)'$ 的联合密度函数为

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

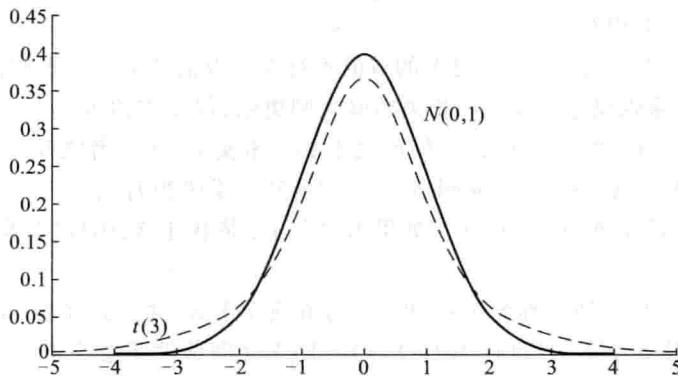


图 2.3 $N(0,1)$ 与 $t(3)$ 的概率密度

则称 \mathbf{X} 服从期望为 $\boldsymbol{\mu}$ 、协方差矩阵为 Σ 的 n 维正态分布, 记为 $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ 。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, \dots, X_n , 如果任意线性组合 $k_1 X_1 + \dots + k_n X_n$ (其中 k_1, \dots, 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, \dots, 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, \dots, Y_m 的任意线性组合 $k_1 Y_1 + \dots + k_m Y_m$, 其中 k_1, \dots, k_m 不全为 0。由于 Y_1, \dots, Y_m 分别是 $(X_1 X_2 \cdots X_n)$ 的线性函数, 故 $k_1 Y_1 + \dots + k_m Y_m$ 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 $k_1 Y_1 + \dots + k_m Y_m$ 服从一维正态分布。根据 k_1, \dots, 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, \dots, X_n 相互独立” 与 “ X_1, X_2, \dots, 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 $\{Z_1, \dots, Z_k\}$ 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。

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

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。

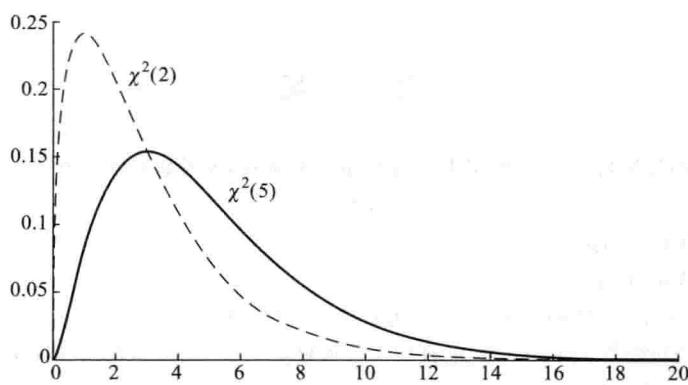


图 2.4 χ^2 分布的概率密度

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 $\frac{Z}{\sqrt{Y/k}} \sim t(k)$, 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 $k \rightarrow \infty$ 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 $\frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2)$ 。 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。

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。

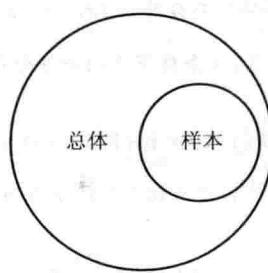


图 2.5 总体与样本

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

- (1) $E(AX) = A\mu$ ；
- (2) $\text{Var}(X) = E(XX') - \mu\mu'$ ；
- (3) $\text{Var}(AX) = A\text{Var}(X)A'$ 。

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

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 X 为 $n \times K$ 常数矩阵， $n > K$ ，且矩阵 X 满列秩，证明 $X'X$ 为正定矩阵。（提示：基于 $X'X$ 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 $f(y|x) = \frac{f(x,y)}{f_x(x)}$

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 $F_x(x)$ 。

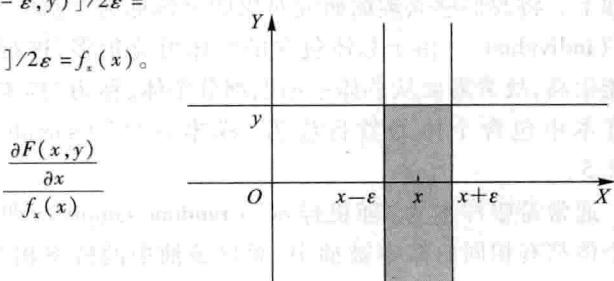
为了求解在 $X=x$ 条件下 Y 的条件密度函数，首先考虑在 $X \in [x-\varepsilon, x+\varepsilon]$ 的条件下， Y 的条件分布函数，参见图2.6。

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

若使 $\varepsilon \rightarrow 0^+$ ，则分子为 $\lim_{\varepsilon \rightarrow 0^+} [F(x+\varepsilon, y) - F(x-\varepsilon, y)]/2\varepsilon =$

$\frac{\partial F(x, y)}{\partial x}$ ，而分母为 $\lim_{\varepsilon \rightarrow 0^+} [F_x(x+\varepsilon) - F_x(x-\varepsilon)]/2\varepsilon = f_x(x)$ 。

故在 $X=x$ 条件下 Y 的条件分布函数为



而条件密度函数为条件分布函数的导数，故

$$f(y|x) = \lim_{\varepsilon \rightarrow 0^+} \frac{\partial}{\partial y} P\{Y \leq y | X \in [x-\varepsilon, x+\varepsilon]\} = \frac{\partial}{\partial y} \frac{\partial F(x, y)}{\partial x} = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{f(x, y)}{f_x(x)}$$

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

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

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($k = 1, \dots, K$), 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 $x_{i1} \equiv 1, \forall i$ 。 $\beta_1, \beta_2, \dots, \beta_K$ 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 $\frac{\partial E(y_i)}{\partial x_{i1}} = \beta_1$ 。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 $\frac{E(y_i)}{x_{ik}} = \beta_k + \gamma x_m$, 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 $\mathbf{x}_i \equiv (x_{i1} \ x_{i2} \ \cdots \ x_{iK})'$, $\boldsymbol{\beta} \equiv (\beta_1 \ \beta_2 \ \cdots \ \beta_K)'$, 则方程(3.1)为

$$\mathbf{y}_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n) \quad (3.2)$$

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

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

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.4)$$

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)

$$E(\varepsilon_i | \mathbf{X}) = E(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n) = 0 \quad (i = 1, \dots, n) \quad (3.5)$$

^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 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 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 $E(\varepsilon_i | X) = c$, 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 $E(\varepsilon_i | X) = 0$ 并不会带来过多限制, 因为如果 $E(\varepsilon_i | X) = c \neq 0$, 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 $E(\varepsilon_i) = 0$, 即扰动项的无条件期望为 0。

证明: 根据迭代期望定律, $E(\varepsilon_i) = E_x E(\varepsilon_i | X) = E_x(0) = 0$ 。

$$= 0$$

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

命题 解释变量与扰动项正交。

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

$$= 0$$

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

如果不满足此条件, 则 β “不可识别” (unidentified), 因为 X 中某个或多个变量为多余。在后面对将看到, 根据 OLS 估计, $\hat{\beta} = (X'X)^{-1}X'y$ 。如果 X 满列秩, 则对称矩阵 $X'X$ 为正定矩阵 (参见第 2 章习题), 故 $(X'X)^{-1}$ 存在; 反之, 则 $(X'X)^{-1}$ 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,

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

其中, 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 I_n 成正比。一方面, 在球型扰动项的假设下, 协方差矩阵 $\text{Var}(\varepsilon | X)$ 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 $\text{Var}(\varepsilon | X)$ 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 $\{x_{i1}, x_{i2}, \dots, x_{ik}\}$: 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 x 与 y 的内积为 0, 即 $x'y = 0$ 。

^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 β , 对于 β 的一个任意假想值 (hypothetical value) $\tilde{\beta}$ (读为 beta tilde), 记第 i 个数据的拟合误差(即残差, residual)为 $e_i = y_i - \mathbf{x}' \tilde{\beta}$ 。类比方程(3.4)可知, 残差向量 $e \equiv (e_1 \ e_2 \ \cdots \ e_n)' = \mathbf{y} - \mathbf{X} \tilde{\beta}$ 。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 $\tilde{\beta}$ 。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

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

其中, $(\mathbf{y}' \mathbf{X} \tilde{\beta})' = \tilde{\beta}' \mathbf{X}' \mathbf{y}$ (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 $2 \mathbf{y}' \mathbf{X} \tilde{\beta}$ 。目标函数 $\text{SSR}(\tilde{\beta})$ 实际上是 $\tilde{\beta}$ 的二次函数 (二次型), 参见图 3.1。

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

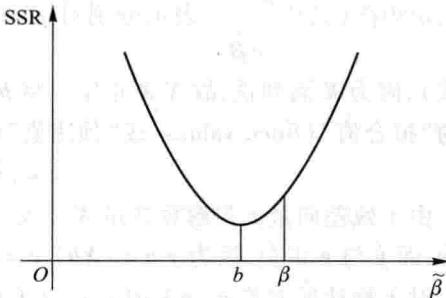


图 3.1 参数的假想值 $\tilde{\beta}$ 、真实值 β 与 OLS 估计值 b

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

这个规则类似于对一次函数求导。从上式可知, 对向量 $\tilde{\beta}$ 求导, 就是分别对 $\tilde{\beta}$ 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 A 为 K 阶对称矩阵, 同理可证 (但较烦琐):

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

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 $(A + B)' = A' + B'$, 以及 $(AB)' = B'A'$ 。

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

其中,由于 $y'y$ 不包含 $\tilde{\beta}$ (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 \hat{b} 满足

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

$$X'y - (X'X)\hat{b} = \mathbf{0} \quad (\text{移项})$$

$$\underbrace{X'(y - X\hat{b})}_{=\epsilon} = \mathbf{0} \quad (\text{向左提取共同的矩阵因子 } X')$$

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

$$\hat{b} = (X'X)^{-1}X'y \quad (3.10)$$

$$\text{二阶条件要求黑塞矩阵 (Hessian)} \frac{\partial^2(\text{SSR})}{\partial \tilde{\beta} \partial \tilde{\beta}'} \equiv \frac{\partial \left(\frac{\partial \text{SSR}}{\partial \tilde{\beta}} \right)}{\partial \tilde{\beta}'} = \begin{pmatrix} \frac{\partial^2 \text{SSR}}{\partial^2 \tilde{\beta}_1} & \cdots & \frac{\partial^2 \text{SSR}}{\partial \tilde{\beta}_1 \partial \tilde{\beta}_K} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \text{SSR}}{\partial \tilde{\beta}_K \partial \tilde{\beta}_1} & \cdots & \frac{\partial^2 \text{SSR}}{\partial^2 \tilde{\beta}_K} \end{pmatrix} = 2X'X$$

为正定矩阵(其中 $\frac{\partial(\cdot)}{\partial \tilde{\beta}}$ 表示分别对 $\tilde{\beta}$ 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 X 满列秩,故 $X'X$ 正定。将 \hat{b} 替代方程“ $y = X\beta + \epsilon$ ”中的 β 并令 $\epsilon = \mathbf{0}$,可得被解释变量的“拟合值”(fitted values)或“预测值”(predicted values)

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

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

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

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

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 $\{e_1, e_2, \dots, e_n\}$ 必须满足 K 个正规方程 $X'\epsilon = \mathbf{0}$,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 $E(s^2) = \sigma^2$ 。当然,如果样本容量 n 很大($n \rightarrow \infty$),则 $\frac{n - K}{n} \rightarrow 1$,

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 $\{\epsilon_1, \epsilon_2, \dots, \epsilon_n\}$ 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{y} 是 y 向超平面 X 的投影(projection),因为 e 与 X 正交。

由于 $\hat{y} \equiv Xb = X \underbrace{(X'X)^{-1}X'y}_{=b} \equiv Py$, 故 $P \equiv$

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

对于矩阵 P 与 M ,可以证明以下性质(参见习题):

- $PX = X$; (自己的投影还是自己)
- $Pe = 0$; (垂直于 X 的向量 e 投影于 X 则退化为一个点)
- $MX = 0$; (自己对自己投影,其残差为 0)
- P 与 M 都是对称阵;
- $P^2 = P$; (再次投影的效果等于一次投影)
- $M^2 = M$ 。 (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 ε 的函数:

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

进一步,可以把残差平方和也写为 ε 的函数:

$$SSR = \varepsilon' \varepsilon = (M\varepsilon)' M\varepsilon = \varepsilon' M' M\varepsilon = \varepsilon' M^2 \varepsilon = \varepsilon' M\varepsilon \quad (3.14)$$

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

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

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

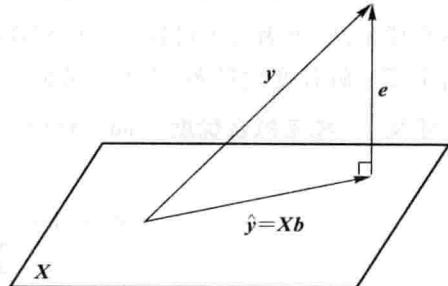


图 3.2 最小二乘法的正交性

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

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 $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2$ 。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) \bar{R}^2 为

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

\bar{R}^2 的一个缺点是它可能为负数。无论 R^2 还是 \bar{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:

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

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

$$R_{uc}^2 = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}} \quad (3.19)$$

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, $R_{uc}^2 = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$ (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 $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 为 \mathbf{y} 的线性组合。

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

证明:“抽样误差”(sampling error)为

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

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

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

将上式移项,即可得 $E(\mathbf{b} | \mathbf{X}) = \boldsymbol{\beta}$ 。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 $E(\mathbf{b}) = \boldsymbol{\beta}$ 。

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

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

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

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

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

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

$$\hat{\boldsymbol{\beta}} = \mathbf{Cy} = (\mathbf{D} + \mathbf{A})\mathbf{y} = \mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \hat{\boldsymbol{b}} = \mathbf{DX}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \hat{\boldsymbol{b}} \quad (3.20)$$

利用 $\hat{\boldsymbol{\beta}}$ 的无偏性可得,

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

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

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

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

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

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) &= \text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}) = \text{Var}[(\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon} | \mathbf{X}] = (\mathbf{D} + \mathbf{A})\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X})(\mathbf{D} + \mathbf{A})' \\ &= \sigma^2 (\mathbf{D} + \mathbf{A})(\mathbf{D}' + \mathbf{A}') = \sigma^2 (\underbrace{\mathbf{DD}'}_{=0} + \underbrace{\mathbf{AD}'}_{=0} + \underbrace{\mathbf{DA}'}_{=0} + \mathbf{AA}') \\ &= \sigma^2 [\mathbf{DD}' + (\mathbf{X}'\mathbf{X})^{-1}] \quad (\text{因为 } \mathbf{DA}' = \mathbf{DX}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}) \end{aligned}$$

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

由于 \mathbf{DD}' 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: $E(s^2 | \mathbf{X}) = \sigma^2$ 。

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

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

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

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

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

因此,对协方差阵 $\text{Var}(\mathbf{b} | \mathbf{X})$ 的无偏估计为 $s^2(\mathbf{X}'\mathbf{X})^{-1}$,在 Stata 中记为“VCE”(Variance-Covariance Matrix Estimated)。Stata 所汇报的协方差矩阵就是根据 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ 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 \mathbf{X} 的情况下, $\boldsymbol{\varepsilon} | \mathbf{X}$ 的条件分布为正态,即 $\boldsymbol{\varepsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 。

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 $H_0: \beta_k = \bar{\beta}_k$,其中 $\bar{\beta}_k$ 为给定的常数。通常 $\bar{\beta}_k = 0$,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”) $H_1: \beta_k \neq \bar{\beta}_k$ 。

直观来说,如果未知参数 β_k 的估计值 b_k 离 $\bar{\beta}_k$ 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 $\boldsymbol{\epsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, 而 $\mathbf{b} - \boldsymbol{\beta} = \mathbf{A}\boldsymbol{\epsilon}$ 为 $\boldsymbol{\epsilon}$ 的线性函数, 故 $(\mathbf{b} - \boldsymbol{\beta}) | \mathbf{X}$ 服从正态分布。更进一步, $E(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X}) = \mathbf{0}$, $\text{Var}(\mathbf{b} | \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, 故 $(\mathbf{b} - \boldsymbol{\beta}) | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$ 。因此, 在原假设“ $H_0: \beta_k = \bar{\beta}_k$ ”成立的情况下, 其第 k 个分量 $(b_k - \bar{\beta}_k) | \mathbf{X} \sim N(0, \sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1})$, 其中 $(\mathbf{X}'\mathbf{X})_{kk}^{-1}$ 为矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 的第 (k, k) 个元素, 而 $\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}$ 为 b_k 的方差。如果 σ^2 已知, 则统计量 $z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \sim N(0, 1)$ 。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“ $H_0: \beta_k = \bar{\beta}_k$ ”也成立的情况下, t 统计量 $t_k \equiv \frac{b_k - \bar{\beta}_k}{\text{SE}(b_k)} \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \sim t(n - K)$ 。其中, $\text{SE}(b_k)$ 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 $\frac{Z}{\sqrt{Y/k}} \sim t(k)$, 其中 k 为自由度。因此, 将统计量 t_k 往这个方向变形。

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

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

已知 $z_k \sim N(0, 1)$, 下面将证明:

- (1) $q | \mathbf{X} \sim \chi^2(n - K)$;
- (2) $z_k | \mathbf{X}$ 与 $q | \mathbf{X}$ 相互独立,

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

$$(1) q \equiv \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\mathbf{e}'\mathbf{M}\mathbf{e}}{\sigma^2} = \frac{\mathbf{e}'}{\sigma} \mathbf{M} \frac{\mathbf{e}}{\sigma} \quad (\text{二次型})$$

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

(2) $z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$ 是 \mathbf{b} 的函数, 而 q 是 \mathbf{e} 的函数。因此, 为了证明 z_k 与 q 相互独立, 只要证明 \mathbf{b} 与 \mathbf{e} 相互独立即可。由于 $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\epsilon}$ 与 $\mathbf{e} = \mathbf{M}\boldsymbol{\epsilon}$ 都是正态扰动项 $\boldsymbol{\epsilon}$ 的线性函数, 故 (\mathbf{b}, \mathbf{e}) 的联合分布也是正态分布(参见第 2 章命题)。因此, 要证明 \mathbf{b} 与 \mathbf{e} 相互独立, 只要证明 $\text{Cov}(\mathbf{b}, \mathbf{e}) = \mathbf{0}$ 。

^① 如果 $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_n)$ (这意味着 \mathbf{x} 的各分量相互独立), 且 \mathbf{A} 为幂等矩阵, 则二次型 $\mathbf{x}'\mathbf{Ax}$ 服从自由度为 $\text{rank}(\mathbf{A})$ 的 χ^2 分布。如果 $\mathbf{A} = \mathbf{I}_n$, 则 $\mathbf{x}'\mathbf{x} \sim \chi^2(n)$, 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。

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

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

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

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

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。

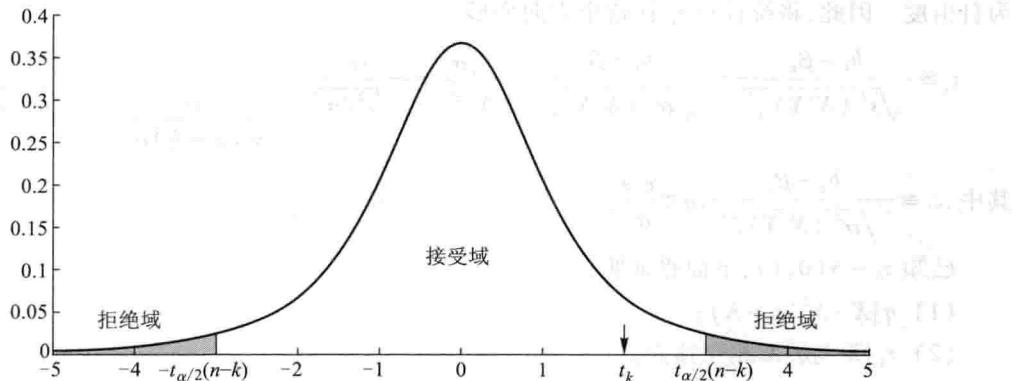


图 3.3 t 检验

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 $\text{P}(t > |t_k|) \times 2$, 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 $\frac{b_k - \beta_k}{SE(b_k)} \sim t(n - K)$, 故

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

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

即置信区间为 $[b_k - t_{\alpha/2} SE(b_k), b_k + t_{\alpha/2} SE(b_k)]$, 以点估计 b_k 为中心, 区间半径为 $t_{\alpha/2} SE(b_k)$ 。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 $P(\text{reject } H_0 | H_0)$ 。

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 $P(\text{accept } H_0 | H_1)$ 。

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“ $1 - P(\text{accept } H_0 | H_1)$ ”。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“ $H_0 : \beta_2 = \cdots = \beta_K = 0$ ”(β_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:

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

其中, \mathbf{r} 为 m 维列向量, \mathbf{R} 为 $m \times K$ 矩阵, $\text{rank}(\mathbf{R}) = m$, 即 \mathbf{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

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

直观来看,由于 \mathbf{b} 是 $\boldsymbol{\beta}$ 的估计量,故如果 H_0 成立,则 $(\mathbf{R}\mathbf{b} - \mathbf{r})$ 应该比较接近于 $\mathbf{0}$ 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“ $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ”也成立的情况下,则F统计量

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

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

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

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

- (1) $w | X \sim \chi^2(m)$;
- (2) $q | X \sim \chi^2(n - K)$; (已在 t 统计量定理中证明)
- (3) $w | X$ 与 $q | X$ 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 $\mathbf{v} \equiv \mathbf{R}\mathbf{b} - \mathbf{r}$ 。在 H_0 成立的情况下, $\mathbf{v} \equiv \mathbf{R}\mathbf{b} - \mathbf{r} = \mathbf{R}\mathbf{b} - \mathbf{R}\boldsymbol{\beta} = \mathbf{R}(\mathbf{b} - \boldsymbol{\beta})$ 。由于 \mathbf{b} 为正态分布, 故 $\mathbf{v} | X$ 为 m 维正态分布, 且 $E(\mathbf{v} | X) = \mathbf{0}$, 其方差为

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

根据数理统计知识知道^①,

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

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

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

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

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 \mathbf{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 $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(m)$ 。

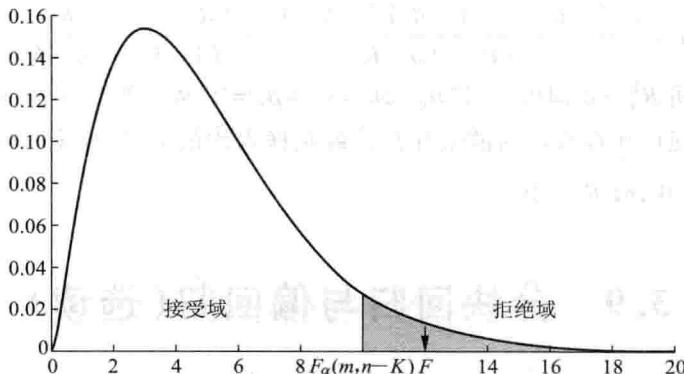


图 3.4 F 检验

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:

$$\begin{aligned} & \min_{\tilde{\beta}} \text{SSR}(\tilde{\beta}) \\ & \text{s. t. } \mathbf{R}\tilde{\beta} = \mathbf{r} \end{aligned} \quad (3.26)$$

其基本思想是,如果“ $H_0 : \mathbf{R}\beta = \mathbf{r}$ ”正确,则加上此约束应该不会使残差平方和 $\text{SSR}(\tilde{\beta})$ 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):

$$F = \frac{(\mathbf{e}^* \mathbf{e}^* - \mathbf{e}' \mathbf{e}) / m}{\mathbf{e}' \mathbf{e} / (n - K)} \quad (3.27)$$

其中, \mathbf{e} 为无约束的残差向量, \mathbf{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 $\frac{R^2 / (K - 1)}{(1 - R^2) / (n - K)}$ 。

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

$$F = \frac{(\mathbf{e}^* \mathbf{e}^* - \mathbf{e}' \mathbf{e}) / (K - 1)}{\mathbf{e}' \mathbf{e} / (n - K)} = \frac{\frac{(\mathbf{e}^* \mathbf{e}^* - \mathbf{e}' \mathbf{e})}{(K - 1)}}{\frac{\mathbf{e}' \mathbf{e}}{(n - K)}} \quad (3.28)$$

其中, \mathbf{e} 为无约束 OLS 的残差向量,而 \mathbf{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 $\frac{\mathbf{e}' \mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - R^2$, $\frac{\mathbf{e}^* \mathbf{e}^*}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - R_*^2$,故

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

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 $\boldsymbol{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 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 $\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2)$, 分别对应于两组解释变量。则多元回归模型可以写为,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (3.30)$$

其中, 回归系数也相应地分组, 即 $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ 。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 \mathbf{X}_2 的边际影响, 考虑以下回归。首先, 把 \mathbf{y} 对 \mathbf{X}_1 进行回归, 将所得的残差记为 \mathbf{e}_1 , 即 \mathbf{y} 中不能由 \mathbf{X}_1 解释的部分; 其次, 把 \mathbf{X}_2 中的每个变量分别对 \mathbf{X}_1 进行回归, 将所得的残差列为一个残差矩阵 \mathbf{e}_2 (其中的每列均为残差向量), 即 \mathbf{X}_2 中不能由 \mathbf{X}_1 解释的部分; 最后, 将 \mathbf{e}_1 对 \mathbf{e}_2 进行回归, 即考察“ \mathbf{X}_2 中与 \mathbf{X}_1 无关的部分”对“ \mathbf{y} 中与 \mathbf{X}_1 无关的部分”的解释力。在最后这一步回归中, \mathbf{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 $(\mathbf{X}'\mathbf{X})\boldsymbol{b} = \mathbf{X}'\mathbf{y}$ 可以写为,

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

$$\begin{pmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_1 \mathbf{y} \\ \mathbf{X}'_2 \mathbf{y} \end{pmatrix} \quad (3.32)$$

$$\begin{cases} \mathbf{X}'_1 \mathbf{X}_1 \boldsymbol{b}_1 + \mathbf{X}'_1 \mathbf{X}_2 \boldsymbol{b}_2 = \mathbf{X}'_1 \mathbf{y} \\ \mathbf{X}'_2 \mathbf{X}_1 \boldsymbol{b}_1 + \mathbf{X}'_2 \mathbf{X}_2 \boldsymbol{b}_2 = \mathbf{X}'_2 \mathbf{y} \end{cases} \quad (3.33)$$

由此方程组的第一个方程可得,

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

从上式可知, 如果 \mathbf{X}_1 与 \mathbf{X}_2 正交, 即 $\mathbf{X}'_1 \mathbf{X}_2 = 0$, 则 $\boldsymbol{b}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$, 相当于将 \mathbf{y} 单独对 \mathbf{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 $\mathbf{X} = \mathbf{1}_{n \times 1} \equiv (1 \ 1 \ \cdots \ 1)'$, 故最小二乘估计为 $\boldsymbol{b}_1^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{y} = n^{-1} \sum_{i=1}^n y_i = \bar{y}$ 。

$$\begin{aligned}
 b_2 &= [\mathbf{X}'_2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1)\mathbf{X}_2]^{-1}[\mathbf{X}'_2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1)\mathbf{y}_2] \\
 &\equiv [\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2]^{-1}[\mathbf{X}'_2\mathbf{M}_1\mathbf{y}_2] \\
 &= [\mathbf{X}'_2\mathbf{M}'_1\mathbf{M}_1\mathbf{X}_2]^{-1}[\mathbf{X}'_2\mathbf{M}'_1\mathbf{M}_1\mathbf{y}_2] \\
 &= (\mathbf{e}'_2\mathbf{e}_2)^{-1}\mathbf{e}'_2\mathbf{e}_1
 \end{aligned} \tag{3.35}$$

其中, $\mathbf{M}_1 \equiv \mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$ 为以 \mathbf{X}_1 为解释变量的消灭矩阵 (\mathbf{M}_1 为对称幂等矩阵, 故 $\mathbf{M}'_1\mathbf{M}_1 = \mathbf{M}_1$)。因此, $\mathbf{M}_1\mathbf{X}_2$ 即为把 \mathbf{X}_2 对 \mathbf{X}_1 进行回归所得的残差矩阵 \mathbf{e}_2 , 而 $\mathbf{M}_1\mathbf{y}_2$ 为把 \mathbf{y} 对 \mathbf{X}_1 进行回归所得的残差 \mathbf{e}_1 。由上式可知, b_2 正是把 \mathbf{e}_1 对 \mathbf{e}_2 进行回归所得的回归系数。上述过程也被称为“过滤掉 \mathbf{X}_1 的影响”(partialing out or netting out the effects of \mathbf{X}_1)。

显然, 如果 \mathbf{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 \mathbf{X}_1 影响的 z ”对“滤去 \mathbf{X}_1 影响的 \mathbf{y} ”的作用。此时, z 对 \mathbf{X}_1 回归所得的 \mathbf{e}_2 为残差向量。将两个残差向量 $(\mathbf{e}_2, \mathbf{e}_1)$ 画成散点图, 则可以直观地看出滤去 \mathbf{X}_1 后 z 对于 \mathbf{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 \mathbf{e}_1 对 \mathbf{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 \mathbf{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 \mathbf{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 \mathbf{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

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

显然, 我们可以用 $\hat{y}_0 = \mathbf{x}'_0\mathbf{b}$ 来对 y_0 作点预测, 其中 \mathbf{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

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

由于 \mathbf{b} 是 $\boldsymbol{\beta}$ 的无偏估计, 故 $E(\hat{y}_0 - y_0) = \mathbf{x}'_0E(\mathbf{b} - \boldsymbol{\beta}) - E(\varepsilon_0) = 0$, 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

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

这个方差反映的是, 由于抽样误差($\mathbf{b} - \boldsymbol{\beta}$)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 $\text{Var}(\hat{y}_0) = \text{Var}(\mathbf{x}'_0\boldsymbol{\beta}) = 0$ 。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。

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

其中,假设 ε_0 与 \mathbf{b} 不相关(因为估计 \mathbf{b} 没有用到 ε_0 的信息),故 $\text{Cov}[\mathbf{x}'_0(\mathbf{b} - \boldsymbol{\beta}), \varepsilon_0] = 0$ 。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 $\sigma^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0$ (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 $\hat{y}_0 - y_0 \sim N(0, \sigma^2 + \sigma^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0)$ 。由于 σ^2 未知,可以用 s^2 来估计,得到下面的 t 统计量:

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

因此, y_0 的置信度为 $(1 - \alpha)$ 的置信区间为

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

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 $P = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$, 消灭矩阵为 $M = I_n - P$ 。证明: $P\mathbf{X} = \mathbf{X}$, $P\varepsilon = \mathbf{0}$, $M\mathbf{X} = \mathbf{0}$, $P' = P$, $M' = M$, $P^2 = P$, $M^2 = M$ 。

3.2 假定一元回归方程为 $y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$ ($i = 1, \dots, n$)。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 $y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i$ ($i = 1, \dots, n$)。

证明:

(1) 残差和 $\sum_{i=1}^n e_i = 0$, $\sum_{i=1}^n x_{ik} e_i = 0$ ($k = 2, \dots, K$); (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 $\bar{y} = \hat{\bar{y}}$;

(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, $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2$ 。[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, $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{[\sum_{i=1}^n (y_i - \bar{y})^2] \cdot [\sum_{i=1}^n (\hat{y}_i - \bar{y})^2]}$

3.5 证明 $R_{ue}^2 = \frac{\mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}}{\mathbf{y}' \mathbf{y}}$ 。

3.6 假设 n 阶对称矩阵 A 半正定。证明: 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 $\text{Var}(\hat{\beta} | \mathbf{X}) - \text{Var}(\mathbf{b} | \mathbf{X})$ 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 \mathbf{b} 的方差? 如果存在,此方差最小能取什么值?

3.8 证明: $\text{Var}(s^2 | \mathbf{X}) = \frac{2\sigma^4}{n - K}$ 。[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：

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

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

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

$$L(\tilde{\beta}, \lambda) = (y - X\tilde{\beta})'(y - X\tilde{\beta}) - \lambda'(r - R\tilde{\beta}) \quad (3.42)$$

其中， λ 为 m 维拉格朗日乘子列向量。一阶条件为

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

$$\frac{\partial L(\tilde{\beta}, \lambda)}{\partial \lambda} = -(r - Rb^*) = \mathbf{0}_{m \times 1} \quad (3.44)$$

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

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

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

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

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

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

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 $(b^* - b)$ ，是 $(r - Rb)$ 的线性函数，而 $(r - Rb)$ 衡量的是无约束 OLS 估计量 b 偏离约束条件 $R\tilde{\beta} = r$ 的程度。如果 b 恰好满足这些约束，则 $b^* = b$ 。

记 e 为无约束的残差向量， e^* 为有约束的残差向量，则

$$e^* = y - Xb^* = y - Xb - X(b^* - b) = e - X(b^* - b) \quad (3.48)$$

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

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

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

将上式代入方程(3.49)可得

$$e^* ' e^* - e'e = (Rb - r)' [R(X'X)^{-1}R']^{-1} (Rb - r) \quad (3.51)$$

根据 F 检验定理，F 统计量为 $F = \frac{(Rb - r)' [R(X'X)^{-1}R']^{-1} (Rb - r) / m}{e'e / (n - K)}$ 。根据方程(3.51)，F 统计量可以

写为 $F = \frac{(e^* ' e^* - e'e) / m}{e'e / (n - K)}$ ，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			
<hr/>				
variable name	storage type	display format	value 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
<hr/>				
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“ $q \geq 10000$ ”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,“ \geq ”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),“ \leq ”(小于等于),“ \neq ”(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“ $q \geq 10000$ ”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“ $q \geq 10000$ ”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* 0.0000	
pk	-0.1845* 0.0263	0.1254 0.1328	1.0000

结果显示,pl与pf的相关系数为0.331,且在5%水平上显著不为零(p值为0.0000);pk与pl的相关系数为-0.1845,且在5%水平上显著(p值为0.0263);pk与pf的相关系数为0.1254,但在5%水平上不显著(p值为0.1328)。

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1000),可输入以下命令,其运行结果如图4.2:

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

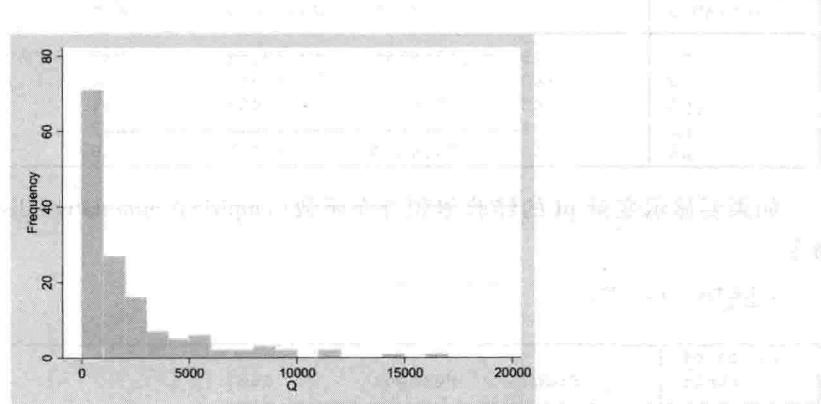


图4.2 直方图

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1000,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。

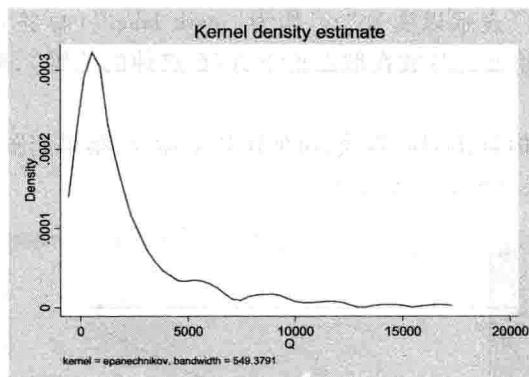


图 4.3 核密度图

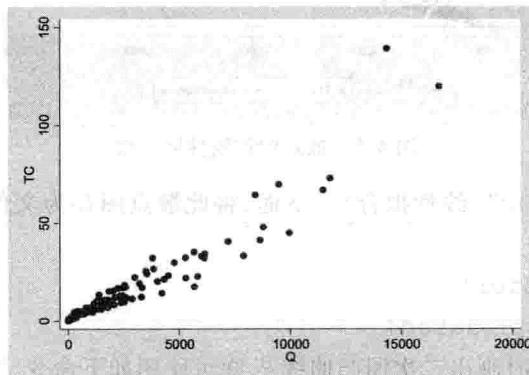


图 4.4 散点图

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)
```

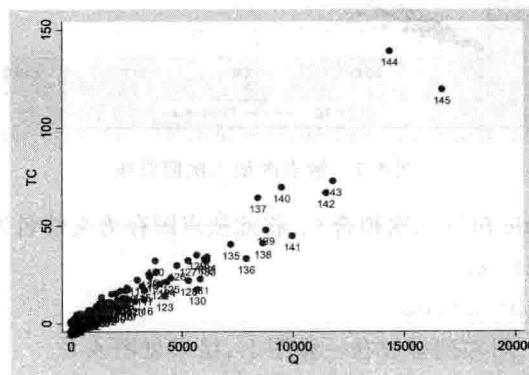


图 4.5 带标签的散点图

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)
```

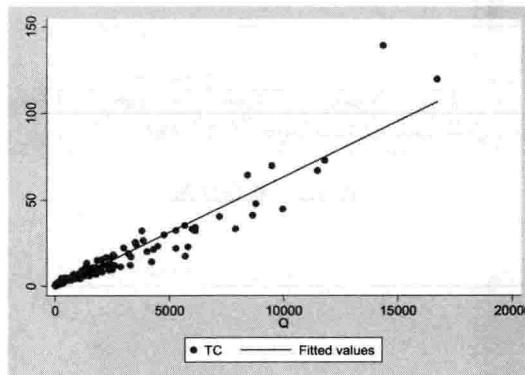


图 4.6 散点图加线性回归线

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)
```

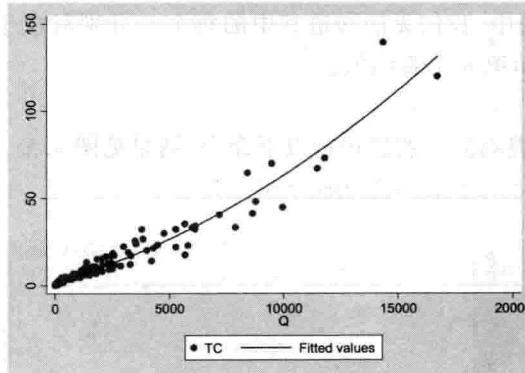


图 4.7 散点图加二次回归线

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

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

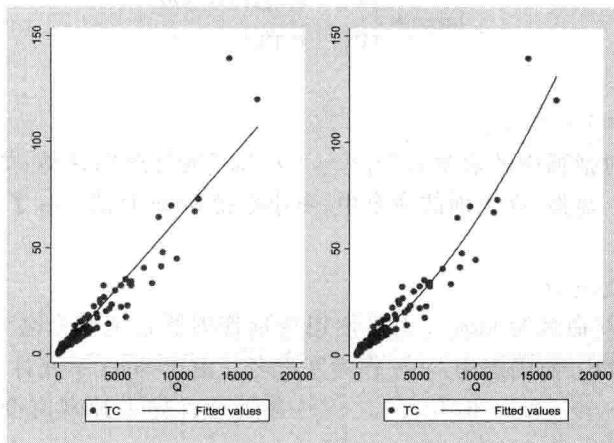


图 4.8 散点图加二次回归线

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:

$$Q_i = A_i L_i^{\alpha_1} K_i^{\alpha_2} F_i^{\alpha_3} \quad (4.1)$$

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:

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

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,

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

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 “ $q \geq 10000$ ” 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

$$\text{large} = \begin{cases} 1, & \text{若 } q \geq 10\,000 \\ 0, & \text{其他} \end{cases} \quad (4.4)$$

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“ $q \geq 6\,000$ ”,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($q \geq 10\,000$)直接替换为新变量($q \geq 6\,000$)。

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 lnc1,lnq2,⋯,lnq30,而只想使用其中的前 15 个变量,则可以用 lnc1—lnc15 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 lnc,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 $\ln 2$,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			
Total	291.066823	144	2.02129738	R-squared = 0.9260			
				Adj R-squared = 0.9239			
				Root MSE = .39227			
				t	P> t	[95% Conf. Interval]	
lntc	Coef.	Std. Err.					
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 $R^2 = 0.9260$, “Adj R-squared”显示 $R^2 = 0.9239$ 。检验整个方程显著性的 F 统计量之 p 值(Prob > F)为 0.0000, 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起示范的作用。

```
. reg lntc linq 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 linq 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 $r = 1.387129 > 1$ ，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 linq = 1
```

这个命令检验的原假设为，变量 linq 的系数等于 1。 F 检验的结果为

```
( 1) linq = 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) linq = 1
( 2) lnpl + lnpk + lnpf = 1
      F( 2,    140) =   128.15
      Prob > F =     0.0000
```

由于 p 值 = 0.0000，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(0.1874)，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 linq 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(0.8334),故我们无法拒绝这个原假设。

在进行回归后,如果要画变量 $\ln q$ 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
```

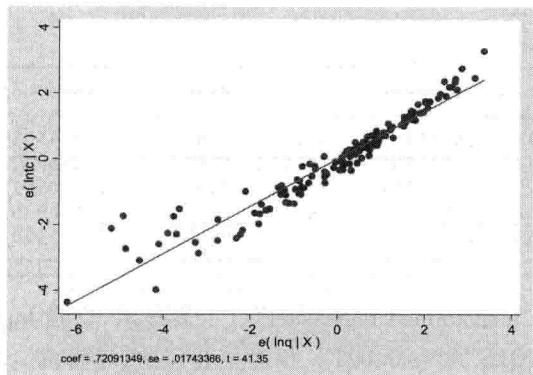


图 4.9 $\ln q$ 的新增变量图

从图4.9可以看出, $\ln q$ 对 $\ln c$ 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
```

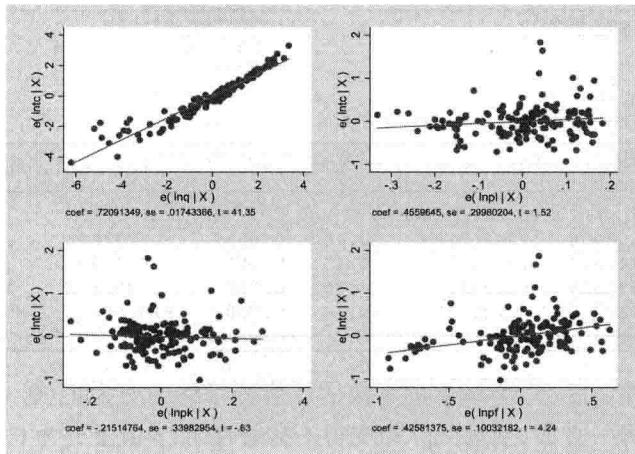


图 4.10 所有变量的新增散点图

从图4.10可以看出,解释变量 $\ln pl, \ln pk, \ln pf$ 与被解释变量 $\ln c$ 的偏关系都没有 $\ln q$ 与 $\ln c$ 的偏关系那么紧密(统计显著性更弱些),而且 $\ln pk$ 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 $\ln pl, \ln pk$ 与 $\ln cc$ 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 lnpl 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 lnpl 的系数估计值终于变为正(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
Total	56422.8385	144	391.825267	R-squared = 0.9073
				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(l1) = -465.7241671687837
    e(l1_0) = -638.1285147616334
    e(rank) = 2

macros:
    e(cmdline) : "regress tc q"
    e(title) : "Linear regression"
    e(marginsok) : "XB default"
    e(vce) : "ols"
    e(devar) : "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 $n \geq 30$,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 $n \rightarrow \infty$ 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 $\{a_n\}_{n=1}^{\infty} = \{a_1, a_2, a_3, \dots\}$ “收敛”(converges)于常数 a ,记为 $\lim_{n \rightarrow \infty} a_n = a$ 或 $a_n \rightarrow a$,如果 $\forall \varepsilon > 0$,存在 $N > 0$,只要 $n > N$,就有 $|a_n - a| < \varepsilon$,即 $\{a_{N+1}, a_{N+2}, \dots\}$ 均落入区间 $(a - \varepsilon, a + \varepsilon)$ 内,参见图 5.1。

2. 随机序列的收敛

定义 随机序列 $\{x_n\}_{n=1}^{\infty} = \{x_1, x_2, x_3, \dots\}$ “依概率收敛”(converges in probability)于常数 a ,记为 $\text{plim}_{n \rightarrow \infty} x_n = a$,

或 $x_n \xrightarrow{P} a$,如果 $\forall \varepsilon > 0$,当 $n \rightarrow \infty$ 时,都有 $\lim_{n \rightarrow \infty} P(|x_n - a| > \varepsilon) = 0$ 。这意味着,任意给定 $\varepsilon > 0$,当

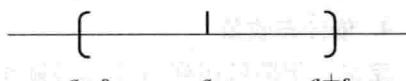


图 5.1 确定性序列的收敛

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。

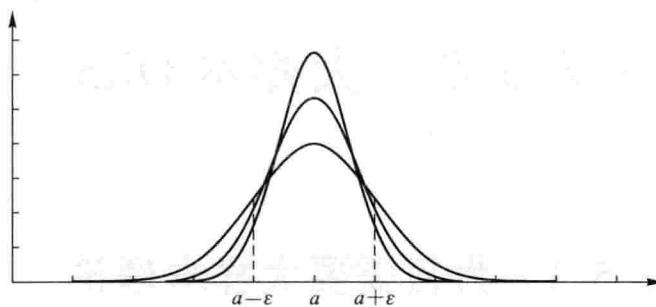


图 5.2 随机序列的收敛

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 $\underset{n \rightarrow \infty}{\text{plim}} g(x_n) = g(\underset{n \rightarrow \infty}{\text{plim}} x_n)$ 。

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 $\underset{n \rightarrow \infty}{\text{plim}}$ 与连续函数 $g(\cdot)$ 可交换运算次序。在这一点上, 概率极限 $\underset{n \rightarrow \infty}{\text{plim}}$ 与普通极限 $\underset{n \rightarrow \infty}{\lim}$ 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, $E(x^2) \neq [E(x)]^2$ 。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 $\text{plim} s^2 = \sigma^2$, 则 $\text{plim} s = \text{plim} (s^2)^{1/2} = (\text{plim} s^2)^{1/2} = (\sigma^2)^{1/2} = \sigma$ (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 $\underset{n \rightarrow \infty}{\lim} E(x_n) = a$, $\underset{n \rightarrow \infty}{\lim} \text{Var}(x_n) = 0$ 。

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 $\underset{n \rightarrow \infty}{\text{plim}} x_n = a$, 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 $\underset{n \rightarrow \infty}{\lim} F_n(c) = F(c)$, 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“ $x_n \xrightarrow{p} x$ ” \Leftrightarrow “ $x_n - x \xrightarrow{p} 0$ ” \Rightarrow “ $x_n \xrightarrow{d} x$ ”; 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)。

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

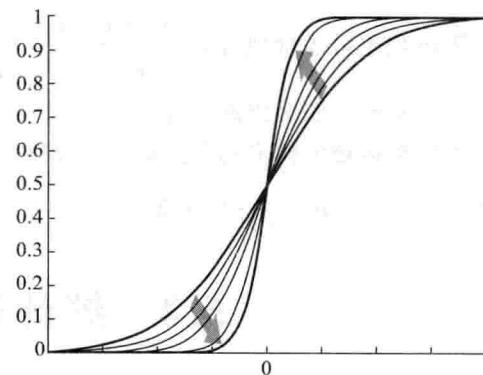


图 5.3 依分布收敛

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 $E(x_1) = \mu$, $\text{Var}(x_1) = \sigma^2$ 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 $E(\bar{x}_n) = \mu$, 而 $\text{Var}(\bar{x}_n) = \text{Var}\left(\frac{x_1 + \dots + x_n}{n}\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \rightarrow 0$ 。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 $E(x_1) = \mu$, $\text{Var}(x_1) = \sigma^2$ 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 $\sqrt{n} \rightarrow \infty$, 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 $\frac{\sigma^2}{n} \rightarrow 0$, 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 $\{\mathbf{x}_n\}_{n=1}^\infty$ 为独立同分布的随机向量序列,且 $E(\mathbf{x}_1) = \boldsymbol{\mu}$, $Var(\mathbf{x}_1) = \boldsymbol{\Sigma}$ 存在,则 $\sqrt{n}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$ 。

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 $MSE(\hat{\beta}) \equiv E[(\hat{\beta} - \beta)^2]$ 。

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 $Bias(\hat{\beta}) \equiv E(\hat{\beta}) - \beta$ 。

定义 如果偏差 $Bias(\hat{\beta}) = 0$,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

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

证明: $MSE(\hat{\beta}) \equiv E[(\hat{\beta} - \beta)^2] = E\{[\hat{\beta} - E(\hat{\beta}) + E(\hat{\beta}) - \beta]^2\}$ (加、减 $E(\hat{\beta})$)

$$= E[\hat{\beta} - E(\hat{\beta})]^2 + 2E\{[\hat{\beta} - E(\hat{\beta})][E(\hat{\beta}) - \beta]\} + E[E(\hat{\beta}) - \beta]^2$$

$$= Var(\hat{\beta}) + 2E\{[\hat{\beta} - E(\hat{\beta})][E(\hat{\beta}) - \beta]\} + [Bias(\hat{\beta})]^2$$

而上式中的交叉项为

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

在多维情况下,也有类似的结论,

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

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 $\lim_{n \rightarrow \infty} \hat{\beta}_n = \boldsymbol{\beta}$,则估计量 $\hat{\beta}_n$ 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, $\hat{\beta}_n$ 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. 漐近正态分布与渐近方差

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

(asymptotically normally distributed), 而称 Σ 为“渐近方差”(asymptotic variance), 记为 $\text{Avar}(\hat{\beta}_n)$ 。直观上, 可以近似地认为 $\hat{\beta}_n \xrightarrow{d} N(\beta, \Sigma/n)$ 。由于 $\sqrt{n}(\hat{\beta}_n - \beta)$ 收敛到一个非退化的分布, 故 $(\hat{\beta}_n - \beta)$ 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 $\hat{\beta}_n$ 与 $\tilde{\beta}_n$ 都是 β 的渐近正态估计量, 其渐近方差分别为 Σ 与 V 。如果 $(V - \Sigma)$ 为半正定矩阵, 则称 $\hat{\beta}_n$ 比 $\tilde{\beta}_n$ “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。

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

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。

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

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。

$$(3) \quad \text{随机向量 } x_n \xrightarrow{d} x, \text{ 随机矩阵 } A_n \xrightarrow{p} A, A_n x_n \text{ 可以相乘} \Rightarrow A_n x_n \xrightarrow{d} Ax.$$

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

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

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

$$\Rightarrow \text{二次型 } x_n' A_n^{-1} x_n \xrightarrow{d} x' A^{-1} x.$$

注: 在极限处, 随机矩阵 A_n 退化为常数矩阵 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 $\{\pi_{1978}, \pi_{1979}, \dots, \pi_{2007}\}$, 参见图 5.4。

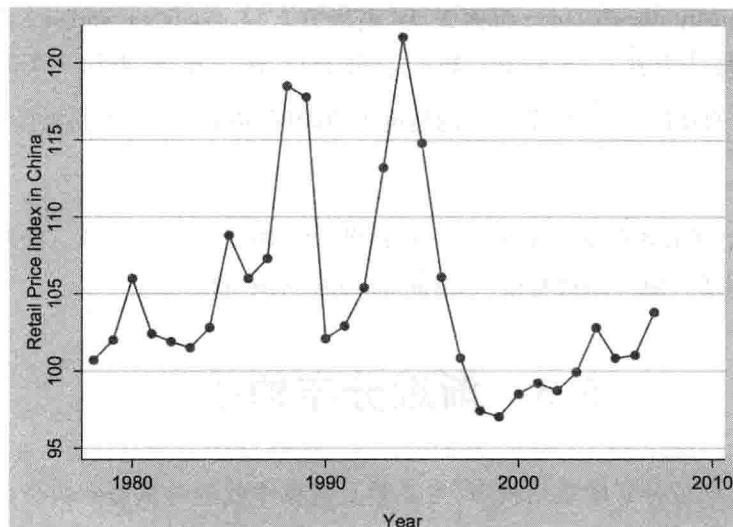


图 5.4 中国零售物价环比指数,1978—2007

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 $E(\pi_{1978})$ 与 $\text{Var}(\pi_{1978})$ 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 $\bar{\pi} = \frac{1}{30} \sum_{t=1978}^{2007} \pi_t$ 作为 $E(\pi_t)$ 的估计量。

从理论上讲,可以把随机过程 $\{x_1, x_2, \dots, x_t, \dots\}$ 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 $\{t_1, t_2, \dots, t_m\}$,随机向量 $\{x_{t_1}, x_{t_2}, \dots, x_{t_m}\}$ 的联合分布等于随机向量 $\{x_{t_1+k}, x_{t_2+k}, \dots, x_{t_m+k}\}$ 的联合分布,其中 k 为任意整数。也就是说,将 $\{x_{t_1}, x_{t_2}, \dots, x_{t_m}\}$ 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, $\{x_{t_1}, x_{t_2}, \dots, x_{t_m}\}$ 的联合分布仅依赖于 $\{t_1, t_2, \dots, t_m\}$ 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 $\{x_t\}_{t=1}^{\infty} = \{x_1, x_1, x_1, \dots\}$ (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)),

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

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, $y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_t$ 。故当 $t \rightarrow \infty$ 时, $\text{Var}(y_t) = t\sigma_{\varepsilon}^2 \rightarrow \infty$, 其中 $\sigma_{\varepsilon}^2 \equiv \text{Var}(\varepsilon_t)$, 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 $\text{Var}(y_t) = \rho^2 \text{Var}(y_{t-1}) + \sigma_{\varepsilon}^2$ 。这是一阶线性差

分方程。容易看出,由于 $\rho^2 < 1$,故 $\text{Var}(y_t)$ 将收敛于 $\frac{\sigma_e^2}{1-\rho^2}$ ^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)。

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 $E(x_t)$ 不依赖于 t ,而且 $\text{Cov}(x_t, x_{t+k})$ 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 $\text{Cov}(x_t, x_{t+k})$ 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 $E(x_t)=0$,而且 $\text{Cov}(x_t, x_{t+k})=0, \forall k \neq 0$ 。

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 $\{x_t\}_{t=1}^\infty$,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 x 即可)。显然,如果 $\{x_t\}_{t=1}^\infty$ 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 $x_t \equiv \begin{pmatrix} \varepsilon_t \\ \varepsilon_1 \end{pmatrix}$ 。显然, $\{x_t\}_{t=1}^\infty = \left\{ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_1 \end{pmatrix}, \begin{pmatrix} \varepsilon_2 \\ \varepsilon_1 \end{pmatrix}, \begin{pmatrix} \varepsilon_3 \\ \varepsilon_1 \end{pmatrix}, \dots \right\}$ 的第一个分量 $\{\varepsilon_t\}_{t=1}^\infty$ 与第二个分量 $\{\varepsilon_1, \varepsilon_1, \varepsilon_1, \dots\}$ 都是平稳过程(参见上文的例子)。然而, $x_1 \equiv \begin{pmatrix} \varepsilon_1 \\ \varepsilon_1 \end{pmatrix}$ 的联合分布却与 $x_2 \equiv \begin{pmatrix} \varepsilon_2 \\ \varepsilon_1 \end{pmatrix}$ 的联合分布不同,故 $\{x_t\}_{t=1}^\infty$ 不是平稳过程。进一步,由于 $\text{Var}(x_1) = \begin{pmatrix} \sigma_e^2 & \sigma_e^2 \\ \sigma_e^2 & \sigma_e^2 \end{pmatrix} \neq \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & \sigma_e^2 \end{pmatrix} = \text{Var}(x_2)$, 故 $\{x_t\}_{t=1}^\infty$ 也不是弱平稳过程。

2. 演近独立性

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“演近独立”。

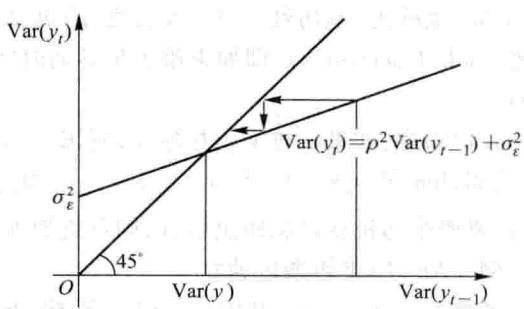


图 5.5 平稳一阶自回归过程的方差收敛

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

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, $\lim_{n \rightarrow \infty} [E(x_t x_{t+n}) - E(x_t)E(x_{t+n})] = 0$, 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,

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

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

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

以此类推, 当时间间隔为 j 时, $\text{Cov}(y_t, y_{t-j}) = \rho^j \sigma_y^2$ 。由于 $|\rho| < 1$, 故当 $j \rightarrow \infty$ 时, $\text{Cov}(y_t, y_{t-j}) \rightarrow 0$ 。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,

$E(x_i x'_i)_{K \times K}$ 可以由 $\frac{1}{n} \sum_{i=1}^n x_i x'_i$ 一致地估计, 其中 $(x_i x'_i)_{K \times K}$ 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 $E(x_i | x_{i-1}, \dots, x_1) = x_{i-1}$, $\forall i \geq 2$ 。

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

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 $E(p_{t+1} | p_t, \dots, p_1) = p_t$ 。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“马领缰”。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 $E(x_i | x_{i-1}, \dots, x_1) = 0, \forall i \geq 2$ 。

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

命题 对鞅序列进行一阶差分, 就得到鞅差分序列。

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

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

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 $\{g_i\}_{i=1}^{\infty}$ 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 $Cov(g_i) = E(g_i g_i') = \Sigma^{\circledR}$, 记 $\bar{g} \equiv \frac{1}{n} \sum_{i=1}^n g_i$, 则 $\sqrt{n}\bar{g} \xrightarrow{d} N(\mathbf{0}, \Sigma)$ 。

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

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

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

($K+1$) 维随机过程 $\{y_i, \mathbf{x}_i\}$ 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 $\{y_i, \mathbf{x}_i\}$ 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^①正交, 即 $E(x_{ik} \varepsilon_i) = 0, \forall i, k$ 。由于回归方程通常有常数项, 故总可以假设 $E(\varepsilon_i) = 0$, 故 $Cov(x_{ik}, \varepsilon_i) = E(x_{ik} \varepsilon_i) - E(x_{ik})E(\varepsilon_i) = 0 - 0 = 0$ 。这意味着 \mathbf{x}_i 与 ε_i 不相关, 仿佛在 ε_i 产生之前, \mathbf{x}_i 便已经确定, 故名“前定解释变量”。定义如下列向量:

$$\mathbf{g}_i \equiv \mathbf{x}_i \varepsilon_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} \varepsilon_i \quad (5.7)$$

① 因为 $E(g_i) = \mathbf{0}$, 故 $Cov(g_i) = E(g_i g_i')$ 。

② 也就是在同一方程中的扰动项。

显然, $E(\mathbf{g}_i) = E(\mathbf{x}_i \varepsilon_i) = \mathbf{0}$ 。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 $E(\mathbf{x}_i \mathbf{x}'_i)$ 为非退化矩阵, 即其逆矩阵 $[E(\mathbf{x}_i \mathbf{x}'_i)]^{-1}$ 存在。这个条件保证了在大样本下, $(\mathbf{X}' \mathbf{X})^{-1}$ 存在。

假定 5.5 \mathbf{g}_i 为鞅差分序列, 且其协方差矩阵 $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}'_i) = E(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}'_i)$ 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 \mathbf{b} 写成 $\{y_i, \mathbf{x}_i\}$ 的函数。

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

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

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

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

其中, 每个 $\mathbf{x}_i y_i$ 都是一个 $K \times 1$ 的向量。

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

$$\mathbf{b} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}' \mathbf{y}}{n} = \left(\frac{\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i}{n} \right)^{-1} \left(\frac{\sum_{i=1}^n \mathbf{x}_i y_i}{n} \right) = \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \quad (5.8)$$

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

定理 (OLS 估计量的大样本性质)

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

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

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

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

证明:(1) 抽样误差可以写为

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

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

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。

假设 $y_i = \alpha + \beta x_i + \varepsilon_i$, 而且 $\operatorname{Cov}(x_i, \varepsilon_i) > 0$ 。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 $\operatorname{Cov}(x_i, \varepsilon_i) < 0$, 则 $\hat{\beta}$ 将倾向于低估 β 。增大样本容量

($n \rightarrow \infty$)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 $\operatorname{Cov}(x_i, \varepsilon_i) \neq 0$ 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 $\mathbf{b} - \boldsymbol{\beta} = \mathbf{S}_{xx}^{-1} \bar{\mathbf{g}}$, 故 $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \mathbf{S}_{xx}^{-1}(\sqrt{n}\bar{\mathbf{g}})$ 。根据假定 5.5 及鞅差分序列的中心极限定理, $\sqrt{n}\bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$, 其中 $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}_i') = E(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$ 。由于 $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \mathbf{S}_{xx}^{-1}(\sqrt{n}\bar{\mathbf{g}})$ 是 $\sqrt{n}\bar{\mathbf{g}}$ 的线性组合, 故 $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \operatorname{Avar}(\mathbf{b}))$ 。由于 $\mathbf{S}_{xx} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \xrightarrow{P} E(\mathbf{x}_i \mathbf{x}_i')$, 故 $\operatorname{Avar}(\mathbf{b}) = [E(\mathbf{x}_i \mathbf{x}_i')]^{-1} \mathbf{S} [E(\mathbf{x}_i \mathbf{x}_i')]^{-1}$ (此处使用了公式 $\operatorname{Var}(\mathbf{AY}) = \mathbf{A} \operatorname{Var}(\mathbf{Y}) \mathbf{A}^T$), 其中 $[E(\mathbf{x}_i \mathbf{x}_i')]^{-1}$ 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 $\hat{\mathbf{S}} \xrightarrow{P} \mathbf{S}$, 已知 $\mathbf{S}_{xx}^{-1} \xrightarrow{P} [E(\mathbf{x}_i \mathbf{x}_i')]^{-1}$, 故估计量 $\widehat{\operatorname{Avar}}(\mathbf{b}) \equiv \mathbf{S}_{xx}^{-1} \hat{\mathbf{S}} \mathbf{S}_{xx}^{-1} \xrightarrow{P} [E(\mathbf{x}_i \mathbf{x}_i')]^{-1} \mathbf{S} [E(\mathbf{x}_i \mathbf{x}_i')]^{-1}$, 是 $\operatorname{Avar}(\mathbf{b})$ 的一致估计量。由于 $\mathbf{S}_{xx}^{-1} \hat{\mathbf{S}} \mathbf{S}_{xx}^{-1}$ 的形式为两个 \mathbf{S}_{xx}^{-1} (“两片面包”)夹着一个 $\hat{\mathbf{S}}$ (“中间的菜”), 故被称为“夹心估计量”或“三明治估计量”(sandwich estimator)。

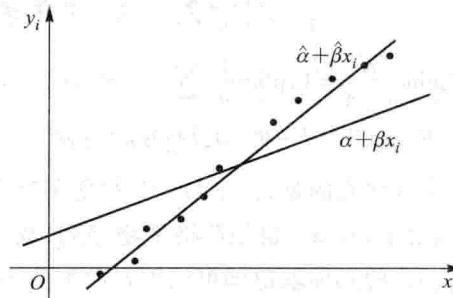


图 5.6 扰动项与解释变量
相关导致不一致估计

^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 $S = E(\varepsilon_i^2 | \mathbf{x}_i)$ 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) $E[(x_{ik}x_{ij})^2]$ 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, $\hat{S} = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}'_i$ 是 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 $E(\varepsilon_i^2) = \sigma^2$ 的一致估计量。

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

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

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

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

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

应该注意的是, 由于 $\{\varepsilon_i\}$ 为严格平稳序列, 故无条件方差 $E(\varepsilon_i^2)$ 是一个常数, 不依赖于 i 。进一步, 由于 $\{\varepsilon_i, \mathbf{x}_i\}$ 也是严格平稳序列, 故条件方差 $E(\varepsilon_i^2 | \mathbf{x}_i)$ 的函数形式也不依赖于 i ^②; 但是, $E(\varepsilon_i^2 | \mathbf{x}_i)$ 的具体取值却可以因 \mathbf{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, $\sqrt{n}(b_k - \bar{\beta}_k) \xrightarrow{d} N(0, \text{Avar}(b_k))$, 其中 b_k 为 OLS 估计量 \mathbf{b} 的第 k 个元素, 而 $\text{Avar}(b_k)$ 为矩阵 $\text{Avar}(\mathbf{b})$ 的第 (k, k) 个元素。另一方面, $\widehat{\text{Avar}}(b_k)$ 是对 $\text{Avar}(b_k)$ 的一致估计, 即 $\widehat{\text{Avar}}(b_k) \xrightarrow{P} \text{Avar}(b_k)$ 。定义 t 统计量为

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

其中, $\text{SE}^*(b_k) \equiv \sqrt{\frac{1}{n} \widehat{\text{Avar}}(b_k)} = \sqrt{\frac{1}{n} (\mathbf{S}_{XX}^{-1} \hat{\mathbf{S}} \mathbf{S}_{XX}^{-1})_{kk}}$ 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, $E(\varepsilon_i^2 | \mathbf{x}_i) = \mathbf{x}'_i \mathbf{x}_i = \sum_{k=1}^K x_{ik}^2$, $\forall i = 1, \dots, n$ 。

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

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

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

$$\widehat{\text{Avar}}(\mathbf{b}) = S_{XX}^{-1} (s^2 S_{XX}) S_{XX}^{-1} = s^2 S_{XX}^{-1} = s^2 \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} = ns^2 (\mathbf{X}' \mathbf{X})^{-1} \quad (5.11)$$

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

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: $H_0: \underbrace{\mathbf{R}}_{m \times K} \underbrace{\boldsymbol{\beta}}_{K \times 1} = \underbrace{\mathbf{r}}_{m \times 1}$, 其中 \mathbf{R} 满行秩。

根据沃尔德检验原理,考察 $(\mathbf{R}\mathbf{b} - \mathbf{r})$ 的大小,譬如其二次型 $(\mathbf{R}\mathbf{b} - \mathbf{r})'(\mathbf{R}\mathbf{b} - \mathbf{r})$ 。在 H_0 成立的情况下,统计量 $W \equiv n(\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R} \widehat{\text{Avar}}(\mathbf{b}) \mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \xrightarrow{d} \chi^2(m)$ 。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

$$W \equiv [\sqrt{n}(\mathbf{R}\mathbf{b} - \mathbf{r})]'[\mathbf{R} \widehat{\text{Avar}}(\mathbf{b}) \mathbf{R}']^{-1}[\sqrt{n}(\mathbf{R}\mathbf{b} - \mathbf{r})] \xrightarrow{d} \chi^2(m) \quad (5.13)$$

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

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

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

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

注:由于 \mathbf{R} 满行秩,且 $\text{Avar}(\mathbf{b})$ 为正定矩阵,故 \mathbf{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 \mathbf{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 $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(m)$ 。

```
.use nerlove.dta,clear
.reg lntc lnq lnpl lnpk lnpf
```

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

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	Robust 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 $\lim_{n \rightarrow \infty} x_n = a$ ，则 $\text{plim}_{n \rightarrow \infty} x_n = a$ 。
- 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 $\theta = \log(\beta)$ ，证明 $\hat{\theta} = \log(\hat{\beta})$ 是 θ 的一致估计量。
- 5.4 鞍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 $\{x_i \varepsilon_i\}$ 是鞍差分序列吗？

(4) 序列 $\{x_i \varepsilon_i\}$ 是严格平稳序列吗？

5.6 当 $n \rightarrow \infty$ 时，是否 $\text{SE}^*(b_k) \xrightarrow{P} 0$ ？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 $\lim_{n \rightarrow \infty} E(x_n) = a, \lim_{n \rightarrow \infty} \text{Var}(x_n) = 0$ 。根据切比雪夫不等式，对

于任意 $\varepsilon > 0$, 都有

$$\text{P}(|x_n - \text{E}(x_n)| \geq \varepsilon) \leq \frac{\text{Var}(x_n)}{\varepsilon^2}$$

当 $n \rightarrow \infty$ 时, 对此不等式两边同时取极限可得,

$$\lim_{n \rightarrow \infty} \text{P}(|x_n - a| \geq \varepsilon) = \lim_{n \rightarrow \infty} \text{P}(|x_n - \text{E}(x_n)| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\text{Var}(x_n)}{\varepsilon^2} = 0$$

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 演近独立过程的定义

如果对于任意两个有界函数 $f: \mathbf{R}^{k+1} \rightarrow \mathbf{R}$ 与 $g: \mathbf{R}^{l+1} \rightarrow \mathbf{R}$, 都有

$$\lim_{n \rightarrow \infty} |\mathbb{E}[f(x_1, \dots, x_{i+k})g(x_{i+n}, \dots, x_{i+l+n})] - [\mathbb{E}[f(x_1, \dots, x_{i+k})] \cdot \mathbb{E}[g(x_{i+n}, \dots, x_{i+l+n})]]| = 0$$

则称随机过程 $\{x_i\}_{i=1}^\infty$ 是一个“演近独立过程”。

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 $n \rightarrow \infty$ 时, $mF \xrightarrow{d} \chi^2(m)$ 。

证明: 因为 $F \sim F(m, n-K)$, 故可设 $F = \frac{\chi^2(m)/m}{\chi^2(n-K)/(n-K)}$ 。

根据 χ^2 分布的性质, $\mathbb{E}[\chi^2(n-K)] = n-K$, 而 $\text{Var}[\chi^2(n-K)] = 2(n-K)$, 故 F 统计量分母的期望值为

$$\mathbb{E}[\chi^2(n-K)/(n-K)] = 1, \text{ 而 } F \text{ 统计量分母的方差为 } \text{Var}[\chi^2(n-K)/(n-K)] = \frac{2(n-K)}{(n-K)^2} = \frac{2}{n-K} \rightarrow 0$$

(当 $n \rightarrow \infty$ 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 “ $H_0: \alpha(\beta) = \mathbf{0}$ ”, 其中 $\alpha(\cdot)$ 为向量函数 (vector-valued function) 且其一阶导数连续。假设 $\alpha(\cdot)$ 的维度是 m , 即共有 m 个非线性约束。记 $A(\beta) = \frac{\partial \alpha(\beta)}{\partial \beta'}$ 为雅可比矩阵 (Jacobian), 即由 $\alpha(\beta)$ 的一阶偏导数组成的 $m \times K$ 矩阵, 其中 K 是 β 的维度。假设矩阵 $A(\beta)$ 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 $\{x_n\}$ 是一个 K 维随机向量序列, 满足 $x_n \xrightarrow{P} \beta$, 并且 $\sqrt{n}(x_n - \beta) \xrightarrow{d} z$, 其中 z 为某随机变量; 假设向量函数 $\alpha(\cdot): \mathbf{R}^K \rightarrow \mathbf{R}^m$ 有连续的一阶导数, 且记其在 β 处的雅可比矩阵为 $A(\beta) \equiv \underbrace{\frac{\partial \alpha(\beta)}{\partial \beta'}}_{m \times K}$, 则

$$\sqrt{n}[\alpha(x_n) - \alpha(\beta)] \xrightarrow{d} A(\beta)z$$

证明: 由于 $x_n \xrightarrow{P} \beta$, 故 $\alpha(x_n) \xrightarrow{P} \alpha(\beta)$, 因此, $\alpha(x_n) - \alpha(\beta) \xrightarrow{P} \mathbf{0}$ 。根据微分中值定理, 存在一个位于 x_n 与 β 之间的向量 y_n (参见图 5.7), 使得

$$\alpha(x_n) - \alpha(\beta) = \frac{\partial \alpha(y_n)}{\partial y'_n}(x_n - \beta) = \underbrace{A(y_n)}_{m \times K} \underbrace{(x_n - \beta)}_{K \times 1}$$

上式两边同时乘以 \sqrt{n} 可得,

$$\sqrt{n}[\alpha(x_n) - \alpha(\beta)] = A(y_n)\sqrt{n}(x_n - \beta)$$

由于 y_n 介于 x_n 与 β 之间, 而 $x_n \xrightarrow{P} \beta$, 故 $y_n \xrightarrow{P} \beta$ 。由于 $A(\cdot)$ 为连续函数, 故 $A(y_n) \xrightarrow{P} A(\beta)$ 。而根据题设, $\sqrt{n}(x_n - \beta) \xrightarrow{d} z$ 。因此,

$$\sqrt{n}[\alpha(x_n) - \alpha(\beta)] \xrightarrow{d} A(\beta)z$$

介绍了 Delta 法之后, 就可以证明如下命题。

命题 在上述原假设 “ $H_0: \mathbf{a}(\boldsymbol{\beta}) = \mathbf{0}$ ” 成立的前提下, 统计量
 $W = n \cdot \mathbf{a}(\mathbf{b})' [\mathbf{A}(\mathbf{b}) \widehat{\text{Avar}}(\mathbf{b}) \mathbf{A}(\mathbf{b})']^{-1} \mathbf{a}(\mathbf{b}) \xrightarrow{d} \chi^2(m)$ (5.14)

证明: 由于 $\sqrt{n}[\mathbf{b} - \boldsymbol{\beta}] \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\mathbf{b}))$, 根据 Delta 法,

$$\sqrt{n}[\mathbf{a}(\mathbf{b}) - \mathbf{a}(\boldsymbol{\beta})] \xrightarrow{d} \mathbf{c}$$

其中 $\mathbf{c} \sim N(\mathbf{0}, \mathbf{A}(\boldsymbol{\beta}) \text{Avar}(\mathbf{b}) \mathbf{A}(\boldsymbol{\beta})')$ 。

根据原假设 $H_0: \mathbf{a}(\boldsymbol{\beta}) = \mathbf{0}$, 故 $\sqrt{n}\mathbf{a}(\mathbf{b}) \xrightarrow{d} \mathbf{c}$, 其中

$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{A}(\boldsymbol{\beta}) \text{Avar}(\mathbf{b}) \mathbf{A}(\boldsymbol{\beta})')$$

由于 $\mathbf{A}(\mathbf{b}) \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta})$, $\widehat{\text{Avar}}(\mathbf{b}) \xrightarrow{p} \text{Avar}(\mathbf{b})$, 故

$$\mathbf{A}(\mathbf{b}) \widehat{\text{Avar}}(\mathbf{b}) \mathbf{A}(\mathbf{b})' \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}) \text{Avar}(\mathbf{b}) \mathbf{A}(\boldsymbol{\beta}) = \text{Var}(\mathbf{c})$$

由于 $\mathbf{A}(\boldsymbol{\beta})$ 满行秩, 而 $\text{Avar}(\mathbf{b})$ 为正定矩阵, 故矩阵 $\text{Var}(\mathbf{c})$ 可逆。因此,

$$W = n \cdot \mathbf{a}(\mathbf{b})' [\mathbf{A}(\mathbf{b}) \widehat{\text{Avar}}(\mathbf{b}) \mathbf{A}(\mathbf{b})']^{-1} \mathbf{a}(\mathbf{b})$$

$$= \sqrt{n}\mathbf{a}(\mathbf{b})' [\mathbf{A}(\mathbf{b}) \widehat{\text{Avar}}(\mathbf{b}) \mathbf{A}(\mathbf{b})']^{-1} \sqrt{n}\mathbf{a}(\mathbf{b}) \xrightarrow{d} \mathbf{c}' \text{Var}(\mathbf{c})^{-1} \mathbf{c} \sim \chi^2(m)$$

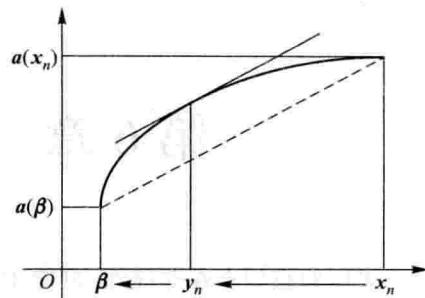


图 5.7 微分中值定理示意图

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 \mathbf{y} 的概率密度函数为 $f(\mathbf{y}; \boldsymbol{\theta})$, 其中 $\boldsymbol{\theta}$ 为 K 维未知参数向量, $\boldsymbol{\theta} \in \Theta$, 其中 Θ 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 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 来估计 $\boldsymbol{\theta}$ 。假设 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 为独立同分布 (iid), 则样本数据的联合密度函数为 $f(\mathbf{y}_1; \boldsymbol{\theta})f(\mathbf{y}_2; \boldsymbol{\theta}) \cdots f(\mathbf{y}_n; \boldsymbol{\theta})$ 。

在抽样之前, $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 被视为随机向量。抽样之后, $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 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 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 给定的情况下, 未知参数 $\boldsymbol{\theta}$ 的函数。定义“似然函数”(likelihood function) 为

$$L(\boldsymbol{\theta}; \mathbf{y}_1, \dots, \mathbf{y}_n) = \prod_{i=1}^n f(\mathbf{y}_i; \boldsymbol{\theta}) \quad (6.1)$$

由此可知, 似然函数与联合密度函数完全相等, 只是 $\boldsymbol{\theta}$ 与 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 的角色互换, 即把 $\boldsymbol{\theta}$ 作为自变量, 而视 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 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

$$\ln L(\boldsymbol{\theta}; \mathbf{y}_1, \dots, \mathbf{y}_n) = \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta}) \quad (6.2)$$

“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 $\hat{\boldsymbol{\theta}}_{ML}$, 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:

$$\max_{\boldsymbol{\theta} \in \Theta} \ln L(\boldsymbol{\theta}; \mathbf{y}) \quad (6.3)$$

在数学上, 常把最大似然估计量 $\hat{\boldsymbol{\theta}}_{ML}$ 写为

$$\hat{\boldsymbol{\theta}}_{ML} \equiv \operatorname{argmax}_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; \mathbf{y}) \quad (6.4)$$

其中, “ argmax ”(即 argument of the maximum) 表示能使 $\ln L(\boldsymbol{\theta}; \mathbf{y})$ 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

$$s(\boldsymbol{\theta}; \mathbf{y}) \equiv \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \equiv \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_K} \end{pmatrix} = \mathbf{0} \quad (6.5)$$

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) $s(\boldsymbol{\theta};\mathbf{y})$ 为**0**。这实际上是一个由 K 个未知参数($\theta_1, \theta_2, \dots, \theta_K$)、 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 $s(\boldsymbol{\theta};\mathbf{y})$ 本身是 \mathbf{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**0**。

命题(得分函数的期望为0**)** 如果似然函数正确(correctly specified),则 $E[s(\boldsymbol{\theta}_0;\mathbf{y})] = \mathbf{0}$,其中 $s(\boldsymbol{\theta};\mathbf{y})$ 表示得分函数 $s(\boldsymbol{\theta};\mathbf{y})$ 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 $y_i \sim N(\theta_0, 1)$, $i = 1, \dots, n$ 。则样本数据的对数似然函数为

$$L(\boldsymbol{\theta}) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2 \quad (6.6)$$

其得分函数为

$$s(\boldsymbol{\theta}) = \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n (y_i - \theta) \quad (6.7)$$

故得分函数的期望值为

$$E[s(\boldsymbol{\theta})] = E\left[\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \sum_{i=1}^n [E(y_i) - \theta] = \sum_{i=1}^n [\theta_0 - \theta] \quad (6.8)$$

在 $\theta = \theta_0$ 处

$$E[s(\boldsymbol{\theta})] \Big|_{\theta=\theta_0} = \sum_{i=1}^n [\theta_0 - \theta] \Big|_{\theta=\theta_0} = 0 \quad (6.9)$$

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

$$s(\boldsymbol{\theta};\mathbf{y}) = \frac{\partial \ln L(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}} = \frac{\partial \sum_{i=1}^n \ln f(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial \ln f(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n s_i(\boldsymbol{\theta}; y_i) \quad (6.10)$$

其中, $s_i(\boldsymbol{\theta}; y_i) \equiv \frac{\partial \ln f(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 为第*i*个观测值对得分函数的贡献。回到上面的例子,则 $s(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n (y_i - \theta)$,而 $s_i(\boldsymbol{\theta}, y_i) = (y_i - \theta)$ 。二阶条件要求,对数似然函数的黑赛矩阵(Hessian matrix) $\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ ^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

$$\mathbf{H}(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \frac{\partial^2 \ln L(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \mathbf{H}_i(\boldsymbol{\theta}; y_i) \quad (6.11)$$

其中, $\mathbf{H}_i(\boldsymbol{\theta}; y_i)$ 为第*i*个观测值对黑赛矩阵的贡献。在上文的一维例子中, $\mathbf{H}(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \theta} \sum_{i=1}^n (y_i - \theta) = -n$,而 $\mathbf{H}_i(\boldsymbol{\theta}; y_i) = \frac{\partial s_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial (y_i - \theta)}{\partial \boldsymbol{\theta}} = -1$ 。

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 $L(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(2-\mu)^2}{2\sigma^2}\right\}$ 。显然, 似然函数在 $\hat{\mu} = 2$ 处取最大值,

参见图 6.1。

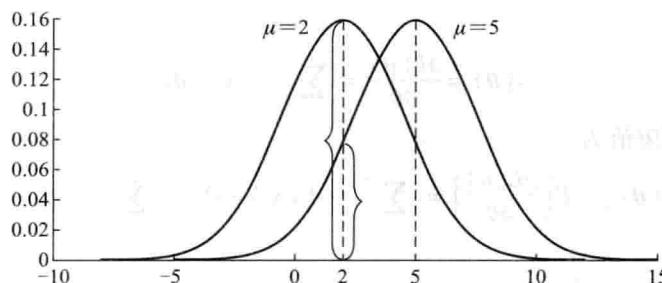


图 6.1 选择参数使观测到样本的可能性最大

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)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6.12)$$

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 $\boldsymbol{\varepsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 。因此, 被解释变量的条件分布为 $\mathbf{y} | \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, 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)

$$f(\mathbf{y} | \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \quad (6.13)$$

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

$$\ln L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad (6.14)$$

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 $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ 最小。而这正是最小二乘法的目标函数 $\mathbf{e}'\mathbf{e}$ 。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。

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (6.15)$$

在第二步,对数似然函数变为 $-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\mathbf{e}'\mathbf{e}$,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

$$-\frac{n}{2}\frac{1}{\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\mathbf{e}'\mathbf{e} = 0 \quad (6.16)$$

求解 σ^2 的MLE估计量为

$$\hat{\sigma}_{\text{ML}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n} \neq \hat{\sigma}_{\text{OLS}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} \equiv s^2 \quad (6.17)$$

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 $\hat{\sigma}_{\text{ML}}^2$ 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($0.1, \dots, 0.9$),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 θ 为多维,或者我们对 θ 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 $\{x_0, x_1, x_2, x_3, \dots\}$,其递推公式为^①

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (6.18)$$

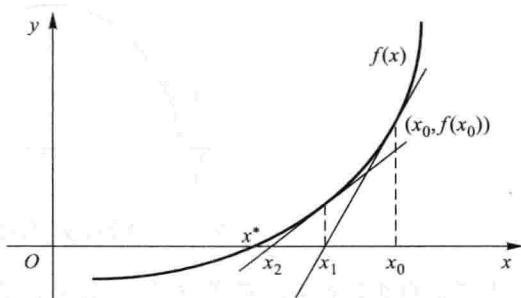


图6.2 高斯-牛顿法

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 $f(\boldsymbol{x})$ 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 $f(\boldsymbol{x})$ 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塞矩阵之期望值(对 \mathbf{y} 求期望)的负数,即

$$\mathbf{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \quad (6.19)$$

在一维情形下, $-\frac{\partial^2 \ln L}{\partial \theta^2}$ 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, $-\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ 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 $\mathbf{I}(\boldsymbol{\theta})$ 即表示平均曲率(对 \mathbf{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 $\mathbf{I}(\boldsymbol{\theta})$ 包含了 $\boldsymbol{\theta}$ 是否容易估计的信息,故称为“信息矩阵”。

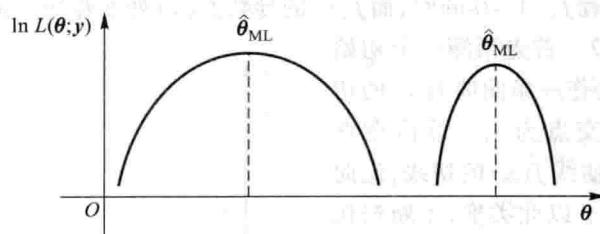


图6.3 平坦(左)与陡峭(右)的对数似然函数

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,

$$\mathbf{I}(\boldsymbol{\theta}_0) \equiv -\mathbb{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = \mathbb{E}\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right] = \mathbb{E}[s(\boldsymbol{\theta}_0; \mathbf{y}) s(\boldsymbol{\theta}_0; \mathbf{y})'] \quad (6.20)$$

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 $\mathbf{I}(\boldsymbol{\theta}_0)$ 就是得分函数的协方差

矩阵 $\text{Var}[s(\boldsymbol{\theta}_0; \mathbf{y})]$ 。

证明：

$$\begin{aligned}\text{Var}[s(\boldsymbol{\theta}_0; \mathbf{y})] &= E[s(\boldsymbol{\theta}_0; \mathbf{y})s(\boldsymbol{\theta}_0; \mathbf{y})'] - \underbrace{E[s(\boldsymbol{\theta}_0; \mathbf{y})]}_{=0} \underbrace{E[s(\boldsymbol{\theta}_0; \mathbf{y})']}_{=0} \\ &= E[s(\boldsymbol{\theta}_0; \mathbf{y})s(\boldsymbol{\theta}_0; \mathbf{y})'] \\ &= I(\boldsymbol{\theta}_0)\end{aligned}\quad (6.21)$$

其中, $E[s(\boldsymbol{\theta}_0; \mathbf{y})] = \mathbf{0}$, 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 $[I(\boldsymbol{\theta}_0)]^{-1}$, 即 $\text{Var}(\hat{\boldsymbol{\theta}}) \geq [I(\boldsymbol{\theta}_0)]^{-1}$ 。其中, $[I(\boldsymbol{\theta}_0)]^{-1}$ 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 $I(\boldsymbol{\theta}_0)$ 越大, 则 $[I(\boldsymbol{\theta}_0)]^{-1}$ 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)

$$[I(\boldsymbol{\theta}_0)]^{-1} = \begin{pmatrix} \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & 2\sigma^4/n \end{pmatrix} \quad (6.22)$$

其中, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta} \ \sigma^2)'$ 。由于 $\text{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, 故 $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{OLS}$ 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 $[I(\boldsymbol{\theta}_0)]^{-1}$, 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 $\lim_{n \rightarrow \infty} \hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{\theta}_0$ 。

(2) 渐近有效性, 即渐近协方差矩阵 $\text{Avar}(\hat{\boldsymbol{\theta}}_{ML}) = n [I(\boldsymbol{\theta}_0)]^{-1}$, 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 $[I(\boldsymbol{\theta}_0)]^{-1}$ 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 $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$, 对数似然函数的三阶导数存在且有界, 信息矩阵 $I(\boldsymbol{\theta})$ 在真实参数 $\boldsymbol{\theta}$ 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) 演近正态, 即 $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1})$, 可以近似地认为 $\hat{\boldsymbol{\theta}}_{\text{ML}} \xrightarrow{d} N(\boldsymbol{\theta}_0, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1})$ 。这表明, $(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0)$ 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 $\ln L(\boldsymbol{\theta}; \mathbf{y}_1, \dots, \mathbf{y}_n) = \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta})$ 除以 n , 并定义

$$Q_n(\boldsymbol{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta}) \quad (6.23)$$

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 $\boldsymbol{\theta}$, $\{\ln f(\mathbf{y}_i; \boldsymbol{\theta}), i = 1, \dots, n\}$ 为 iid, 故根据大数定律, 其样本均值将收敛到总体均值, 即

$$Q_n(\boldsymbol{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta}) \xrightarrow{P} E[\ln f(\mathbf{y}; \boldsymbol{\theta})] \equiv Q(\boldsymbol{\theta}) \quad (6.24)$$

其中, $Q(\boldsymbol{\theta}) \equiv E[\ln f(\mathbf{y}; \boldsymbol{\theta})]$ 只依赖于 $\boldsymbol{\theta}$, 因为期望算子 $E(\cdot)$ 已将样本数据 \mathbf{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 $\frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}$ 。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 $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

$$E\left[\ln \frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}\right] < \ln \left\{ E\left[\frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}\right]\right\} \quad (6.25)$$

其中, 如果 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, 则随机变量 $\frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}$ 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,

$$E\left[\frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}\right] = \int \left[\frac{f(\mathbf{y}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta}_0)}\right] f(\mathbf{y}; \boldsymbol{\theta}_0) d\mathbf{y} = \int f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} = 1 \quad (6.26)$$

其中, 由于 $f(\mathbf{y}; \boldsymbol{\theta})$ 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

$$E[\ln f(\mathbf{y}; \boldsymbol{\theta})] - E[\ln f(\mathbf{y}; \boldsymbol{\theta}_0)] < 0 \quad (6.27)$$

由于 $Q(\boldsymbol{\theta}) \equiv E[\ln f(\mathbf{y}; \boldsymbol{\theta})]$, 故上式表明, 对于任意 $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, 都有

$$Q(\boldsymbol{\theta}) < Q(\boldsymbol{\theta}_0) \quad (6.28)$$

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 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 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 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 也越來越接近 $Q(\boldsymbol{\theta})$ 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

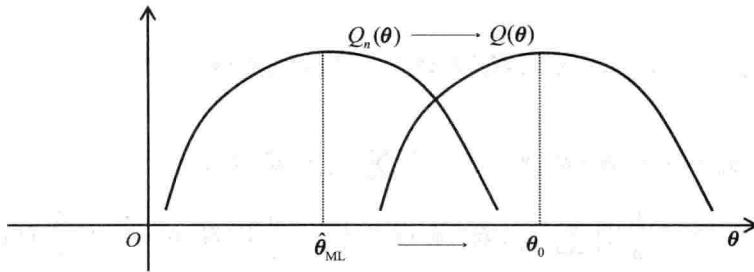


图 6.4 MLE 估计量的一致性示意图

(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 $\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) \xrightarrow{d} N(\mathbf{0}, n[\mathbf{I}(\theta_0)]^{-1})$, 也就证明了(2)渐近有效性, 即 $\text{Avar}(\hat{\theta}_{\text{ML}}) = n[\mathbf{I}(\theta_0)]^{-1}$ 。根据 MLE 的一阶条件

$$\mathbf{s}(\hat{\theta}_{\text{ML}}; \mathbf{y}) = \frac{\partial \ln L(\hat{\theta}_{\text{ML}}; \mathbf{y})}{\partial \theta} = \mathbf{0} \quad (6.29)$$

即 $\hat{\theta}_{\text{ML}}$ 满足 MLE 的一阶条件。利用微分中值定理 (Mean Value Theorem), 将 $s(\hat{\theta}_{\text{ML}}; \mathbf{y})$ 在 $\theta = \theta_0$ 处进行泰勒展开可得

$$s(\hat{\theta}_{\text{ML}}; \mathbf{y}) = s(\theta_0; \mathbf{y}) + H(\theta^*; \mathbf{y})(\hat{\theta}_{\text{ML}} - \theta_0) = \mathbf{0} \quad (6.30)$$

其中, θ^* 介于 θ_0 与 $\hat{\theta}_{\text{ML}}$ 之间, 即存在 $\lambda \in (0, 1)$, 使得 $\theta^* = \lambda \theta_0 + (1 - \lambda) \hat{\theta}_{\text{ML}}$; 而 $H(\theta^*; \mathbf{y}) \equiv \frac{\partial^2 \ln L(\theta^*; \mathbf{y})}{\partial \theta \partial \theta'}$ 为在 $\theta = \theta^*$ 处的黑赛矩阵。将方程(6.30)移项, 然后两边同乘以 \sqrt{n} 可得

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) = [-H(\theta^*; \mathbf{y})]^{-1} \sqrt{n}s(\theta_0; \mathbf{y}) \quad (6.31)$$

根据 MLE 的一致性, $\underset{n \rightarrow \infty}{\text{plim}} \hat{\theta}_{\text{ML}} = \theta_0$; 而 θ^* 夹在 θ_0 与 $\hat{\theta}_{\text{ML}}$ 之间, 故 $\underset{n \rightarrow \infty}{\text{plim}} \theta^* = \theta_0$ 。由此可知

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) \xrightarrow{d} [-H(\theta_0; \mathbf{y})]^{-1} \sqrt{n}s(\theta_0; \mathbf{y}) \quad (6.32)$$

其中, $s(\theta_0; \mathbf{y}) = \sum_{i=1}^n s_i(\theta_0; \mathbf{y}_i)$, 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 $H(\theta_0; \mathbf{y}) = \sum_{i=1}^n H_i(\theta_0; \mathbf{y}_i)$, 参见方程(6.11)。因此, 方程(6.32)可以写为

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) \xrightarrow{d} \left[-\frac{1}{n} \sum_{i=1}^n H_i(\theta_0; \mathbf{y}_i) \right]^{-1} \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n s_i(\theta_0; \mathbf{y}_i) \right) \quad (6.33)$$

由于样本为 iid, 故 $\{H_i(\theta_0; \mathbf{y}_i)\}_{i=1}^n$ 与 $\{s_i(\theta_0; \mathbf{y}_i)\}_{i=1}^n$ 都是 iid。因此, 一方面, 根据大数定律, $-\frac{1}{n} \sum_{i=1}^n H_i(\theta_0; \mathbf{y}_i) \xrightarrow{P} -E[H_i(\theta_0; \mathbf{y}_i)] = A_0$ 。另一方面, 由于 $E[s_i(\theta; \mathbf{y}_i)] = \mathbf{0}$ (得分函数的期望为 $\mathbf{0}$), 故根据中心极限定理, $\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n s_i(\theta_0; \mathbf{y}_i) \right] \xrightarrow{d} N(\mathbf{0}, B_0)$, 其中 $B_0 \equiv \text{Var}(s_i)$ 。由此可知

$$\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_0) \xrightarrow{d} N(\mathbf{0}, A_0^{-1} B_0 A_0^{-1}) \quad (6.34)$$

其中, 渐近方差协方差矩阵为 $A_0^{-1} B_0 A_0^{-1}$, 表现为夹心估计量 (sandwich estimator)。下面分别求

\mathbf{A}_0 与 \mathbf{B}_0 。

由于 $\{\mathbf{H}_i(\boldsymbol{\theta}_0; \mathbf{y}_i)\}_{i=1}^n$ 是 iid, 故 $E[\mathbf{H}_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = E[\mathbf{H}_j(\boldsymbol{\theta}_0; \mathbf{y}_j)]$, $\forall i, j$ (iid 随机变量的期望值相同)。因此

$$\begin{aligned}\mathbf{A}_0 &\equiv -E[\mathbf{H}_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = -\frac{1}{n} \sum_{i=1}^n E[\mathbf{H}_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] \\ &= -\frac{1}{n} E\left\{\sum_{i=1}^n [\mathbf{H}_i(\boldsymbol{\theta}_0; \mathbf{y}_i)]\right\} = -\frac{1}{n} E[\mathbf{H}(\boldsymbol{\theta}_0; \mathbf{y})] = \frac{1}{n} \mathbf{I}(\boldsymbol{\theta}_0)\end{aligned}\quad (6.35)$$

另一方面, $\{s_i(\boldsymbol{\theta}; \mathbf{y}_i)\}_{i=1}^n$ 也是 iid, 故 $\text{Var}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \text{Var}[s_j(\boldsymbol{\theta}_0; \mathbf{y}_j)]$, $\forall i, j$ (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

$$\begin{aligned}I(\boldsymbol{\theta}_0) &= \text{Var}[s(\boldsymbol{\theta}_0; \mathbf{y})] = \text{Var}\left[\sum_{i=1}^n s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)\right] \\ &= \sum_{i=1}^n \text{Var}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = n \text{Var}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)]\end{aligned}\quad (6.36)$$

因此

$$\mathbf{B}_0 = \text{Var}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \frac{1}{n} I(\boldsymbol{\theta}_0) = \mathbf{A}_0 \quad (6.37)$$

将 \mathbf{A}_0 与 \mathbf{B}_0 的表达式(6.37)代入方程(6.34)可知,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}_0^{-1}) = N(\mathbf{0}, n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}) \quad (6.38)$$

由于 $\mathbf{A}_0 = \mathbf{B}_0$, 故渐近方差的表达式简化为 $n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ 。

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 $\boldsymbol{\alpha} \equiv g(\boldsymbol{\theta})$, 则对 $\boldsymbol{\alpha}$ 的最大似然估计就是 $\hat{\boldsymbol{\alpha}}_{\text{ML}} = g(\hat{\boldsymbol{\theta}}_{\text{ML}})$, 其中 $g(\cdot)$ 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 $(\hat{\mu}_{\text{ML}}^2 + \hat{\sigma}_{\text{ML}}^2)$ 。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 $E(\hat{\theta}_{\text{MLE}}) = \theta$ (无偏估计), 但对于一般的函数 $g(\cdot)$ 来说, $E[g(\hat{\theta}_{\text{MLE}})] \neq g[E(\theta)] = g(\theta)$, 故在参数变换后, $g(\hat{\theta}_{\text{MLE}})$ 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

$$\text{Avar}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = n \left\{ -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]\right\}^{-1} \quad (6.39)$$

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 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 替代 $\boldsymbol{\theta}_0$ 可得

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n \left\{ -\mathbb{E} \left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\text{ML}}; \mathbf{y})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right] \right\}^{-1} \quad (6.40)$$

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 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 替代 $\boldsymbol{\theta}_0$ 后,干脆将期望算子忽略掉,即 $\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n \left[-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\text{ML}}; \mathbf{y})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right]^{-1}$ 。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 $\sum_{i=1}^n \hat{s}_i \hat{s}_i'$ 来估计 $I(\boldsymbol{\theta}_0)$,即

$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n \left(\sum_{i=1}^n \hat{s}_i \hat{s}_i' \right)^{-1}$,其中 $\hat{s}_i \equiv \frac{\partial \ln f(\mathbf{y}_i; \hat{\boldsymbol{\theta}}_{\text{ML}})}{\partial \boldsymbol{\theta}}$ 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

$$W \equiv (\hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_0)' [\text{Var}(\hat{\boldsymbol{\beta}}_u)]^{-1} (\hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(K) \quad (6.41)$$

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 “ $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ” 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 “ $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ” 或非线性假设 “ $H_0: h(\boldsymbol{\beta}) = \mathbf{0}$ ”,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 Θ 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。

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

$$\text{LR} \equiv -2\ln \left[\frac{L(\hat{\boldsymbol{\beta}}_R)}{L(\hat{\boldsymbol{\beta}}_U)} \right] = 2[\ln L(\hat{\boldsymbol{\beta}}_U) - \ln L(\hat{\boldsymbol{\beta}}_R)] \xrightarrow{d} \chi^2(K) \quad (6.42)$$

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 $F = \frac{(e^{*\prime} e^* - e'e)/(K-1)}{e'e/(n-K)}$, 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:

$$\begin{aligned} & \max_{\tilde{\boldsymbol{\beta}}} \ln L(\tilde{\boldsymbol{\beta}}) \\ & \text{s. t. } \tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 \end{aligned} \quad (6.43)$$

引入拉格朗日乘子函数,

$$\max_{\tilde{\boldsymbol{\beta}}, \lambda} \ln L(\tilde{\boldsymbol{\beta}}) - \lambda'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad (6.44)$$

其中, λ 为拉格朗日乘子向量。如果 $\hat{\lambda} \approx \mathbf{0}$, 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, $\hat{\lambda} = \frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}}$, 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

$$\text{LM} \equiv \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \right)' [I(\hat{\boldsymbol{\beta}}_R)]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \right) \xrightarrow{d} \chi^2(K) \quad (6.45)$$

其中, $I(\hat{\boldsymbol{\beta}}_R)$ 为信息矩阵在 $\hat{\boldsymbol{\beta}}_R$ 处的取值。由于 $\frac{\partial \ln L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 被称为“得分函数”(score function), 故这个检验也称为“得分检验”(score test); 而 $I(\hat{\boldsymbol{\beta}}_R)$ 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, $\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_U)}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$, 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 $\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \approx \mathbf{0}$, 而 LM 统计量反映的就是此接近程度。

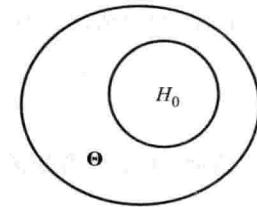


图 6.5 无约束与有约束的参数空间

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。

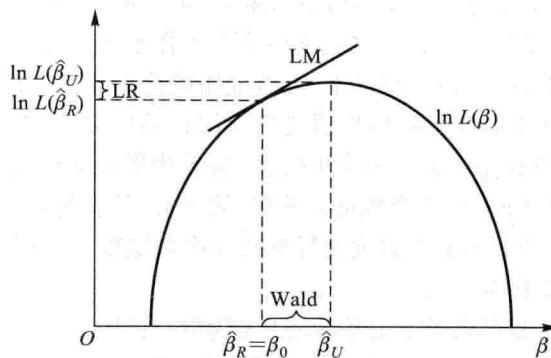


图 6.6 三类渐近等价的统计检验

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_{ue}^2 , 其中 R_{ue}^2 为“非中心 R^2 ”(参见第 3 章)。显然, $\frac{\partial \ln L(\hat{\beta}_R)}{\partial \tilde{\beta}} = \sum_{i=1}^n s_i$ 可以用

$\sum_{i=1}^n \hat{s}_i$ 来估计, 其中 $\hat{s}_i = s_i(\hat{\beta}_R)$ 为第 i 个观测值对得分函数的贡献在 $\hat{\beta}_R$ 处的取值。另一方面, 根据 BHHH 法, $I(\hat{\beta}_R)$ 可以用 $\sum_{i=1}^n \hat{s}_i \hat{s}'_i$ 来估计。因此, LM 统计量可以写为

$$LM = \left(\sum_{i=1}^n \hat{s}'_i \right) \left(\sum_{i=1}^n \hat{s}_i \hat{s}'_i \right)^{-1} \left(\sum_{i=1}^n \hat{s}_i \right) \quad (6.46)$$

定义 $n \times K$ 矩阵 $S = \begin{pmatrix} \hat{s}'_1 \\ \hat{s}'_2 \\ \vdots \\ \hat{s}'_n \end{pmatrix} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{1K} \\ \hat{s}_{21} & \hat{s}_{22} & \cdots & \hat{s}_{2K} \\ \vdots & \vdots & & \vdots \\ \hat{s}_{n1} & \hat{s}_{n2} & \cdots & \hat{s}_{nK} \end{pmatrix}$, 常数向量 $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, 则

$$\sum_{i=1}^n \hat{s}'_i = (1 \ 1 \ \cdots \ 1) \begin{pmatrix} \hat{s}'_1 \\ \hat{s}'_2 \\ \vdots \\ \hat{s}'_n \end{pmatrix} = \mathbf{1}' S, \quad \sum_{i=1}^n \hat{s}_i \hat{s}'_i = (\hat{s}_1 \ \hat{s}_2 \ \cdots \ \hat{s}_n) \begin{pmatrix} \hat{s}'_1 \\ \hat{s}'_2 \\ \vdots \\ \hat{s}'_n \end{pmatrix} = S' S, \text{ 故}$$

$$LM = \mathbf{1}' S (S' S)^{-1} S' \mathbf{1} \quad (6.47)$$

为此, 可以进行如下辅助回归^①:

$$\mathbf{1} \xrightarrow{\text{OLS}} \hat{s}' \gamma + v_i \quad (6.48)$$

其中, 被解释变量为常数向量 $\mathbf{1}$ (相当于线性回归中的 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 S (相当于线性回归中的 X)。根据第 3 章, 有

$$R_{ue}^2 = \frac{\mathbf{y}' X (X' X)^{-1} X' \mathbf{y}}{\mathbf{y}' \mathbf{y}} = \frac{\mathbf{1}' S (S' S)^{-1} S' \mathbf{1}}{\mathbf{1}' \mathbf{1}} = \frac{LM}{n} \quad (6.49)$$

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 $LM = nR_{ue}^2$ 。

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 $Wald \geq LR \geq LM$, 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 $f(y; \boldsymbol{\theta}) = \frac{p(y)e^{r(\boldsymbol{\theta})}}{q(\boldsymbol{\theta})}$ 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 $E(y|x)$ 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 $\hat{\boldsymbol{\theta}}_{QML}$ 的渐近方差也通常不再是 $n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ 。

假设正确的对数似然函数为 $\ln L(\boldsymbol{\theta}; \mathbf{y})$, 而被误设为 $\ln L^*(\boldsymbol{\theta}; \mathbf{y})$, 称为“准对数似然函数”(pseudo log likelihood function)。最大化 $\ln L^*(\boldsymbol{\theta}; \mathbf{y})$ 的结果即 QMLE 估计量

$$\hat{\boldsymbol{\theta}}_{QML} \equiv \arg \max \ln L^*(\boldsymbol{\theta}; \mathbf{y}) \quad (6.50)$$

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 $\hat{\theta}_{QML} \xrightarrow{P} \theta^*$, 其中 θ^* 称为“准真实值”(pseudo-true value), 但通常 $\theta^* \neq \theta_0$ 。对于 $\hat{\theta}_{QML}$ 的大样本分布, 使用类似于 MLE 的推导可证明

$$\sqrt{n}(\hat{\theta}_{ML} - \theta^*) \xrightarrow{d} N(\mathbf{0}, A_0^{*-1} B_0^* A_0^{*-1}) \quad (6.51)$$

其中, A_0^* 与 B_0^* 的表达式类似于 A_0 与 B_0 , 但在 θ^* 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 $\ln L^*(\theta; y)$ 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 $A_0^* \neq B_0^*$, 因此公式(6.51)中的夹心估计量 $A_0^{*-1} B_0^* A_0^{*-1}$ 无法进一步简化。基于 $A_0^{*-1} B_0^* A_0^{*-1}$ 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 $\hat{\beta}_{QML}$ 依然是真实参数 β 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, $\hat{\theta}_{QML} \xrightarrow{P} \theta^* \neq \theta_0$, 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” θ^* , 而且通常不知道 θ^* 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看。

- (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 $E[(X - \mu)/\sigma]^3$,峰度为 $E[(X - \mu)/\sigma]^4$,而超额峰度(excess kurtosis)为 $E[(X - \mu)/\sigma]^4 - 3$ 。因此,对于残差 $\{e_1, \dots, e_n\}$,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

$$\left(\frac{1}{n\hat{\sigma}^4} \sum_{i=1}^n e_i^4\right) - 3 \text{(注意 } \bar{e} = 0\text{)}。在扰动项服从正态分布的原假设下,这两个统计量服从正态分$$

布。较常用的“雅克-贝拉检验”(Jarque and Bera, 1980, 简记 JB)使用了它们的平方之加权平均作为检验统计量:

$$JB \equiv \frac{n}{6} \left[\left(\frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n e_i^3 \right)^2 + \frac{1}{4} \left(\frac{1}{n\hat{\sigma}^4} \sum_{i=1}^n e_i^4 - 3 \right)^2 \right] \xrightarrow{d} \chi^2(2) \quad (6.52)$$

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“`m1`”的命令,可自行定义似然函数来执行最大似然估计,详见“`help m1`”。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)

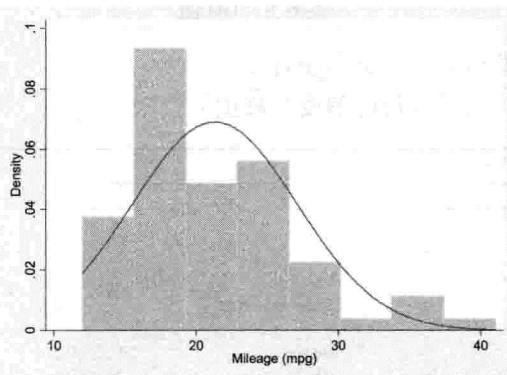


图 6.7 直方图与正态密度

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)

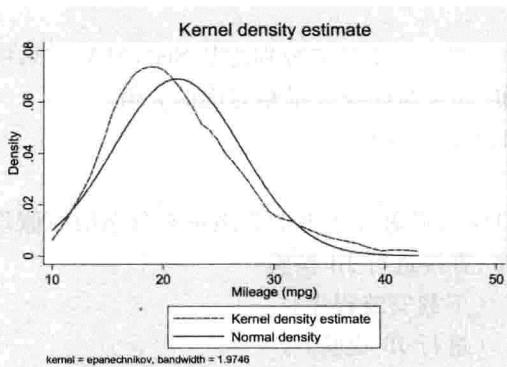


图 6.8 核密度与正态密度

. qnorm mpg (画变量 mpg 的 QQ 图,结果如图 6.9 所示)

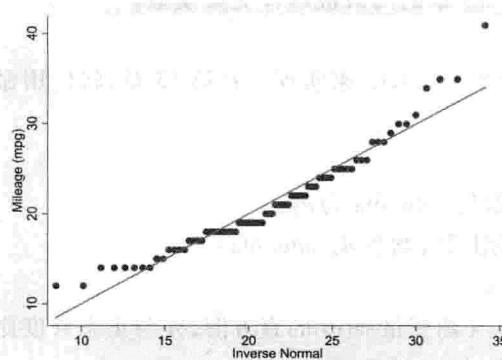


图 6.9 QQ 图

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					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint	
				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(" - ")”。

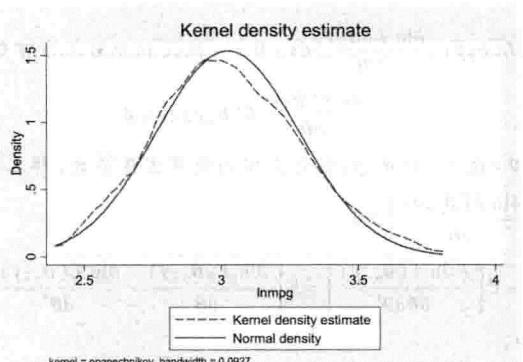


图 6.10 取对数后的核密度与正态密度

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)

Variable	Obs	Skewness/Kurtosis tests for Normality			joint	
		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 $P(y=k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $k=0,1,2,\dots$ 。对于随机样本 $\{y_1, y_2, \dots, y_n\}$, 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 $\{y_1, y_2, \dots, y_n\}$, 求 θ 的 MLE 估计量。

- 6.3 假设样本 $\{y_1, y_2, \dots, y_n\}$ 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 $E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$ 。

证明: 因为似然函数 $L(\boldsymbol{\theta}; \mathbf{y})$ 是概率密度函数, 故似然函数的积分为 1, 即

$$\int L(\boldsymbol{\theta}; \mathbf{y}) d\mathbf{y} = 1$$

$$\int \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] d\mathbf{y} = 1 \quad (\text{引入对数似然函数})$$

方程两边对 $\boldsymbol{\theta}$ 求导可得

$$\int \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} d\mathbf{y} = \mathbf{0} \quad (\text{假设积分与求导可交换次序})$$

$$\int \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y}) d\mathbf{y} = \mathbf{0}$$

由于似然函数正确, 故在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处, $L(\boldsymbol{\theta}_0; \mathbf{y})$ 就是真实的概率密度函数(样本数据 \mathbf{y} 来自于此总体)。因此, 在上式中令 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 即可得, $E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$ 。

- A6.2 证明在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处, $-E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right]$ 。

证明: 根据附录 A6.1 可知,

$$\int \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] d\mathbf{y} = \mathbf{0}$$

该方程两边对 $\boldsymbol{\theta}'$ 求导可得(假设积分与求导可交换次序)

$$\int \left\{ \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] + \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}'} \right\} d\mathbf{y} = \mathbf{0}$$

移项可得

$$- \int \left\{ \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] \right\} d\mathbf{y} = \int \left\{ \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \exp[\ln L(\boldsymbol{\theta}; \mathbf{y})] \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}'} \right\} d\mathbf{y}$$

直接使用似然函数(而不用对数似然函数), 可得

$$- \int \left\{ \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} L(\boldsymbol{\theta}; \mathbf{y}) \right\} d\mathbf{y} = \int \left\{ \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}'} L(\boldsymbol{\theta}; \mathbf{y}) \right\} d\mathbf{y}$$

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

$$-E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = E\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right]$$

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, $\text{Var}(\hat{\boldsymbol{\theta}}) \geq [I(\boldsymbol{\theta}_0)]^{-1}$ 。为简单起见,只证明一维情形,多维情形可类似证明。

由于 $\hat{\theta}(y)$ 是 θ_0 的无偏估计(其中 y 为样本数据),故

$$\theta_0 = E[\hat{\theta}(y)] = \int \hat{\theta}(y) L(\theta_0; y) dy \quad (\text{随机变量函数 } \hat{\theta}(y) \text{ 的期望定义})$$

将上式两边同时对 θ_0 求导可得,

$$\begin{aligned} 1 &= \int \hat{\theta}(y) \frac{\partial L(\theta_0; y)}{\partial \theta} dy = \int \hat{\theta}(y) \frac{\partial \ln L(\theta_0; y)}{\partial \theta} \cdot L(\theta_0; y) dy \\ &= E\left[\hat{\theta}(y) \frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right] \quad (\text{期望的定义}) \\ &= \text{Cov}\left[\hat{\theta}(y), \frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right] \quad (\text{根据附录 A6.1, } E\left[\frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right] = 0) \\ &\leq \text{Var}[\hat{\theta}(y)] \cdot \text{Var}\left[\frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right] \quad (\text{相关系数小于或等于 1}) \\ &= \text{Var}[\hat{\theta}(y)] \cdot E\left[\frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right]^2 \quad (\text{根据信息矩阵等式}) \end{aligned}$$

$$\text{因此, } \text{Var}[\hat{\theta}(y)] \geq \left[E\left(\frac{\partial \ln L(\theta_0; y)}{\partial \theta}\right)^2\right]^{-1} = I(\theta_0)^{-1}.$$

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 $\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ 。为了求导方便,记 $\gamma \equiv \sigma^2$, 则 $\ln L(\boldsymbol{\beta}, \gamma) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \gamma - \frac{1}{2\gamma} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ 。因此,一阶导数为

$$\begin{aligned} \frac{\partial \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta}} &= \frac{1}{\gamma} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \frac{\partial \ln L(\boldsymbol{\beta}, \gamma)}{\partial \gamma} &= -\frac{n}{2\gamma} + \frac{1}{2\gamma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

二阶导数为

$$\begin{aligned} \frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\frac{1}{\gamma} \mathbf{X}' \mathbf{X} \\ \frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \gamma} &= -\frac{1}{\gamma^2} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = -\frac{1}{\gamma^2} \mathbf{X}' \boldsymbol{\epsilon} \\ \frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \gamma^2} &= \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} \boldsymbol{\epsilon}' \boldsymbol{\epsilon} \end{aligned}$$

根据严格外生性 $E(\boldsymbol{\epsilon} | \mathbf{X}) = 0$ 可知, $E(\mathbf{X}' \boldsymbol{\epsilon}) = 0$, 故 $E\left(\frac{\partial \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \gamma}\right) = -\frac{1}{\gamma^2} E(\mathbf{X}' \boldsymbol{\epsilon}) = 0$ 。另外, $E(\boldsymbol{\epsilon}' \boldsymbol{\epsilon}) = E\left(\sum_{i=1}^n \boldsymbol{\epsilon}_i^2\right) = n\sigma^2 = n\gamma$, 故 $E\left(\frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \gamma^2}\right) = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} E(\boldsymbol{\epsilon}' \boldsymbol{\epsilon}) = \frac{n}{2\gamma^2} - \frac{n}{\gamma^2} = -\frac{n}{2\gamma^2}$ 。将 $\gamma \equiv \sigma^2$ 代回表达式可得,

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = -E\left(\begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \frac{\partial \ln L(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \gamma} \\ \frac{\partial \ln L(\boldsymbol{\beta}, \gamma)}{\partial \gamma \partial \boldsymbol{\beta}} & \frac{\partial^2 \ln L(\boldsymbol{\beta}, \gamma)}{\partial \gamma^2} \end{pmatrix}\right) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}$$

因此,信息矩阵的逆矩阵为

$$I(\boldsymbol{\theta})^{-1} = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{pmatrix}$$

A6.5 证明最大似然估计量的不变性。

证明：将 α 视为对原参数 $\boldsymbol{\theta}$ 的参数变换 (reparameterization)，则可以将似然函数写为 α 函数。记 $L(\boldsymbol{\theta}; \mathbf{y})$ 为以 $\boldsymbol{\theta}$ 为参数的似然函数，其最大似然估计为 $\hat{\boldsymbol{\theta}}_{ML}$ 。定义 $\hat{\alpha}_{ML} = g(\hat{\boldsymbol{\theta}}_{ML})$ ，需要证明 $\hat{\alpha}_{ML}$ 为对 α 的最大似然估计。

定义以 α 为参数的似然函数为

$$\begin{aligned} L_g(\alpha; \mathbf{y}) &\equiv \sup_{\boldsymbol{\theta}: g(\boldsymbol{\theta})=\alpha} L(\boldsymbol{\theta}; \mathbf{y}) && \text{(可能多个 } \boldsymbol{\theta} \text{ 满足 “} g(\boldsymbol{\theta}) = \alpha \text{”, 择其最大者)} \\ &\leq \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y}) && \text{(去掉约束条件 “} g(\boldsymbol{\theta}) = \alpha \text{”, 不会减小最大值)} \\ &= L(\hat{\boldsymbol{\theta}}_{ML}; \mathbf{y}) && \text{(根据最大似然估计量的定义)} \\ &= \sup_{\boldsymbol{\theta}: g(\boldsymbol{\theta})=g(\hat{\boldsymbol{\theta}}_{ML})=\hat{\alpha}_{ML}} L(\boldsymbol{\theta}; \mathbf{y}) && \text{(由于 } \hat{\boldsymbol{\theta}}_{ML} \in \{\boldsymbol{\theta}: g(\boldsymbol{\theta})=g(\hat{\boldsymbol{\theta}}_{ML})\}, \text{ 故可加约束条件)} \\ &= L_g(\hat{\alpha}_{ML}; \mathbf{y}) && \text{(根据以 } \alpha \text{ 为参数的似然函数的定义)} \end{aligned}$$

因此， $L_g(\alpha; \mathbf{y}) \leq L_g(\hat{\alpha}_{ML}; \mathbf{y})$, $\forall \alpha$, 即 $L_g(\alpha; \mathbf{y})$ 在 $\alpha = \hat{\alpha}_{ML}$ 处达到最大值，故 $\hat{\alpha}_{ML}$ 是对 α 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 $\sum_{i=1}^n \hat{s}_i \hat{s}_i'$ 来估计信息矩阵 $I(\boldsymbol{\theta})$ 。

$$\begin{aligned} \text{因为 } \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} &= \frac{\partial \left[\sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta}) \right]}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial \ln f(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \equiv \sum_{i=1}^n s_i, \\ \text{所以 } \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} &= \sum_{i=1}^n \frac{\partial^2 \ln f(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}。 \text{ 因此,} \\ I(\boldsymbol{\theta}_0) &\equiv -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] && \text{(信息矩阵的定义)} \\ &= -E\left[\sum_{i=1}^n \frac{\partial^2 \ln f(\mathbf{y}_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] && \text{(样本信息矩阵等于每个观测值信息矩阵之和)} \\ &= \sum_{i=1}^n \left[-E\left(\frac{\partial^2 \ln f(\mathbf{y}_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \right] && \text{(求和与期望算子交换次序)} \\ &= \sum_{i=1}^n \left[E\left(\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y}_i)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y}_i)}{\partial \boldsymbol{\theta}'}\right) \right] && \text{(信息矩阵等式)} \\ &= \sum_{i=1}^n E(s_i s_i') && (s_i \text{ 的定义}) \\ &= nE(s_i s_i') && (\text{iid 的假设}) \end{aligned}$$

由于 $\sum_{i=1}^n \hat{s}_i \hat{s}_i' = n \left(\frac{1}{n} \sum_{i=1}^n \hat{s}_i \hat{s}_i' \right)$, 而 $\frac{1}{n} \sum_{i=1}^n \hat{s}_i \hat{s}_i' \xrightarrow{P} E(s_i s_i')$, 故 $\sum_{i=1}^n \hat{s}_i \hat{s}_i'$ 是 $nE(s_i s_i')$ 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 $\text{Var}(\varepsilon_i | \mathbf{X})$ 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 $\text{Var}(\mathbf{b} | \mathbf{X})$ 的表达式不再是 $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$,因为 $\text{Var}(\varepsilon | \mathbf{X}) \neq \sigma^2 \mathbf{I}$ 。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 $\text{Var}(\varepsilon_i | \mathbf{X})$ 是某解释变量 x_i 的增函数,即 x_i 越大则 $\text{Var}(\varepsilon_i | \mathbf{X})$ 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)。

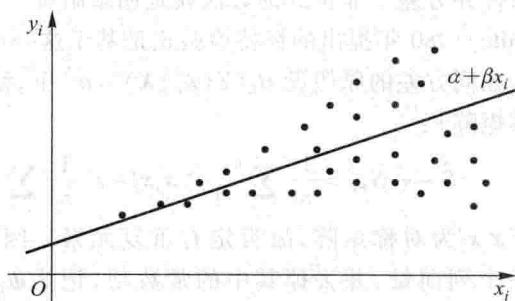


图 7.1 异方差的一种情形

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, $\text{Var}(\varepsilon_i | Y_i)$ 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 $H_0: E(\varepsilon_i^2 | X) = \sigma^2$ 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:

$$\hat{S} - s^2 S_{XX} = \frac{1}{n} \sum_{i=1}^n e_i^2 x_i x_i' - s^2 \frac{1}{n} \sum_{i=1}^n x_i x_i' = \frac{1}{n} \sum_{i=1}^n (e_i^2 - s^2) x_i x_i' \xrightarrow{P} \mathbf{0}_{K \times K} \quad (7.1)$$

由于 $x_i x_i'$ 为对称矩阵,故肯定有重复元素。因此,将随机矩阵 $x_i x_i'$ 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 ψ_i 。 ψ_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 $x_i = (1 \ x_{i2} \ x_{i3})'$, 则 $x_i x_i' = \begin{pmatrix} 1 & x_{i2} & x_{i3} \\ x_{i2} & x_{i2}^2 & x_{i2} x_{i3} \\ x_{i3} & x_{i2} x_{i3} & x_{i3}^2 \end{pmatrix}$, 故 $\psi_i =$

$(x_{i2} \ x_{i3} \ x_{i2}^2 \ x_{i3}^2 \ x_{i2} x_{i3})'$, $m=5$ 。为了得到一维的检验统计量。定义

$$\mathbf{c}_n \equiv \frac{1}{n} \sum_{i=1}^n (e_i^2 - s^2) \psi_i \xrightarrow{P} \mathbf{0}_{m \times 1} \quad (7.2)$$

在一定的条件下可以使用中心极限定理,则 $\sqrt{n} \mathbf{c}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{B})$, 其中, \mathbf{B} 为 \mathbf{c}_n 的渐近方差。假设 $\hat{\mathbf{B}}$ 为 \mathbf{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。

$$n \mathbf{c}' \hat{\mathbf{B}}^{-1} \mathbf{c} \xrightarrow{d} \chi^2(m) \quad (7.3)$$

其中, m 为 \mathbf{c}_n (也就是 ψ_i) 的维度。在实际操作上,一般进行如下的辅助回归:

$$e_i^2 \xrightarrow{\text{OLS}} \text{常数} + \psi_i' \gamma \quad (7.4)$$

并检验 ψ_i 中所有变量的系数 γ 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 $= \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$ 。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 $(K-1)F = \frac{(n-K)R^2}{(1-R^2)} \xrightarrow{d} \chi^2(K-1)$ 。在原假设成立的情况下,当 $n \rightarrow \infty$ 时, $n-K \rightarrow n^{①}$, $(1-R^2) \rightarrow 1^{②}$,故 $(K-1)F \rightarrow nR^2$,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:

$$H_0: E(\varepsilon_i^2 | x_2, \dots, x_K) = \sigma^2 \quad (7.5)$$

如果 H_0 不成立,则条件方差 $E(\varepsilon_i^2 | x_2, \dots, x_K)$ 是 (x_2, \dots, 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:

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + u_i \quad (7.6)$$

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

$$H_0: \delta_2 = \cdots = \delta_K = 0 \quad (7.7)$$

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:

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \quad (7.8)$$

仍然使用 nR^2 统计量:

$$nR^2 \xrightarrow{d} \chi^2(K-1) \quad (7.9)$$

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 $(n-K)/n \rightarrow 1$ 。

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 $R^2 \rightarrow 0$,因此 $1 - R^2 \rightarrow 1$ 。

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 $\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 \mathbf{V}(\mathbf{X}) \neq \sigma^2 \mathbf{I}_n$, 其中 $\mathbf{V}(\mathbf{X})$ 为对称正定矩阵且已知,可能依赖于 \mathbf{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 $\mathbf{V}_{n \times n}$, 存在非退化矩阵 $\mathbf{C}_{n \times n}$, 使得 $\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}$ 。

证明: 参见附录。该命题的直观含义是,在一维情况下,“ \mathbf{V} 正定”即要求 \mathbf{V} 为正数,故 $\frac{1}{\mathbf{V}}$ 也是正数,可以分解为 $\frac{1}{\sqrt{\mathbf{V}}} \cdot \frac{1}{\sqrt{\mathbf{V}}}$;但如果 \mathbf{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 \mathbf{C} 不唯一,但这并不影响 GLS 的最终结果。

将原回归模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 两边同时左乘矩阵 \mathbf{C} 可得

$$\mathbf{C}\mathbf{y} = \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} \quad (7.10)$$

定义以下变量转换:

$$\tilde{\mathbf{y}} \equiv \mathbf{C}\mathbf{y}, \quad \tilde{\mathbf{X}} \equiv \mathbf{C}\mathbf{X}, \quad \tilde{\boldsymbol{\varepsilon}} \equiv \mathbf{C}\boldsymbol{\varepsilon} \quad (7.11)$$

则可以将模型写为

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \quad (7.12)$$

容易验证,变换后的回归模型仍然满足严格外生性,因为

$$\mathbb{E}(\tilde{\boldsymbol{\varepsilon}} | \tilde{\mathbf{X}}) = \mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon} | \mathbf{C}\mathbf{X}) = \mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{C}\mathbb{E}(\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{0} \quad (7.13)$$

其中,由于 \mathbf{C} 非退化,故 $\mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon} | \mathbf{C}\mathbf{X}) = \mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon} | \mathbf{X})$, 因为“ $\mathbf{C}\mathbf{X} = \mathbf{Z} \Leftrightarrow \mathbf{X} = \mathbf{C}^{-1}\mathbf{Z}$ ”。而且,球型扰动项的假定也得到满足,因为

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\varepsilon}} | \tilde{\mathbf{X}}) &= \mathbb{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' | \mathbf{X}) = \mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon}\mathbf{C}'\mathbf{C}' | \mathbf{X}) = \mathbf{C}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})\mathbf{C}' \\ &= \sigma^2 \mathbf{C}\mathbf{V}\mathbf{C}' = \sigma^2 (\mathbf{V}^{-1})^{-1} \mathbf{C}' = \sigma^2 \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1} \mathbf{C}' \\ &= \sigma^2 \mathbf{C}\mathbf{C}^{-1}(\mathbf{C}')^{-1} \mathbf{C}' = \sigma^2 \mathbf{I}_n \end{aligned} \quad (7.14)$$

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:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = [(\mathbf{C}\mathbf{X})'(\mathbf{C}\mathbf{X})]^{-1}(\mathbf{C}\mathbf{X})'\mathbf{C}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned} \quad (7.15)$$

因此,虽然 \mathbf{C} 不唯一,但 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ 却是唯一的,因为 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ 不依赖于 \mathbf{C} 。由于高斯-马尔可夫定理成立,故 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ 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 \mathbf{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 $\mathbf{V}(\mathbf{X})$ 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 $E(\varepsilon_i^2 | \mathbf{x}_i) = \text{Var}(\varepsilon_i | \mathbf{x}_i) = \sigma^2 v_i(\mathbf{X})$, 即

$$\mathbf{V} = \begin{pmatrix} v_1 & & 0 \\ & v_2 & \\ & & \ddots \\ 0 & & v_n \end{pmatrix}, \quad \mathbf{V}^{-1} = \begin{pmatrix} 1/v_1 & & 0 \\ & 1/v_2 & \\ & & \ddots \\ 0 & & 1/v_n \end{pmatrix} \quad (7.16)$$

由于 $\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}$, 可知

$$\mathbf{C} = \mathbf{C}' = \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & & \ddots \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \quad (7.17)$$

$$\tilde{\mathbf{y}} = \mathbf{C}\mathbf{y} = \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & & \ddots \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_1/\sqrt{v_1} \\ y_2/\sqrt{v_2} \\ \vdots \\ y_n/\sqrt{v_n} \end{pmatrix} \quad (7.18)$$

$$\tilde{\mathbf{X}} = \mathbf{C}\mathbf{X} = \begin{pmatrix} 1/\sqrt{v_1} & & 0 \\ & 1/\sqrt{v_2} & \\ & & \ddots \\ 0 & & 1/\sqrt{v_n} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ x_{21} & \cdots & x_{2K} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nK} \end{pmatrix} \quad (7.19)$$

$$= \begin{pmatrix} x_{11}/\sqrt{v_1} & \cdots & x_{1K}/\sqrt{v_1} \\ x_{21}/\sqrt{v_2} & \cdots & x_{2K}/\sqrt{v_2} \\ \vdots & & \vdots \\ x_{n1}/\sqrt{v_n} & \cdots & x_{nK}/\sqrt{v_n} \end{pmatrix}$$

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

$$\frac{y_i}{\sqrt{v_i}} = \beta_1 \frac{x_{i1}}{\sqrt{v_i}} + \beta_2 \frac{x_{i2}}{\sqrt{v_i}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{v_i}} + \frac{\varepsilon_i}{\sqrt{v_i}} \quad (7.20)$$

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”:

$$\min_{\tilde{\beta}} \text{SSR} = \sum_{i=1}^n \left(\frac{e_i}{\sqrt{v_i}} \right)^2 = \sum_{i=1}^n \frac{e_i^2}{v_i} \quad (7.21)$$

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 $V(X)$,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

$$\hat{\beta}_{\text{FGLS}} = (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} y \quad (7.22)$$

其中, \hat{V} 是 V 的一致估计。FGLS 的困难在于 $V(X)$ 中包含太多的参数,待估计参数个数可能多达 $[n(n+1)/2]$ ($V(X)$ 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 $e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + \text{error}_i$ 就可以获得对 σ_i^2 的估计值 $\hat{\sigma}_i^2$ 。在此回归时,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:

$$e_i^2 = \sigma^2 \exp(\delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK}) v_i \quad (7.23)$$

其中, v_i 为乘积形式的扰动项。取对数后可得

$$\ln e_i^2 = (\ln \sigma^2 + \delta_1) + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + \ln v_i \quad (7.24)$$

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), $\hat{\beta}_{\text{FGLS}} = (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} y$,而 \hat{V} 是数据 (y, X) 的非线性函数,故 $\hat{\beta}_{\text{FGLS}}$ 是 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{V} 是 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 $\text{Var}(\epsilon_i | x_i)$,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 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
Model	269.524728	4	67.3811819	F(4, 140) =	437.90
Residual	21.5420958	140	.153872113	Prob > F =	0.0000
Total	291.066823	144	2.02129738	R-squared =	0.9260
				Adj R-squared =	0.9239
				Root MSE =	.39227
<hr/>					
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	-.3870089 .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)

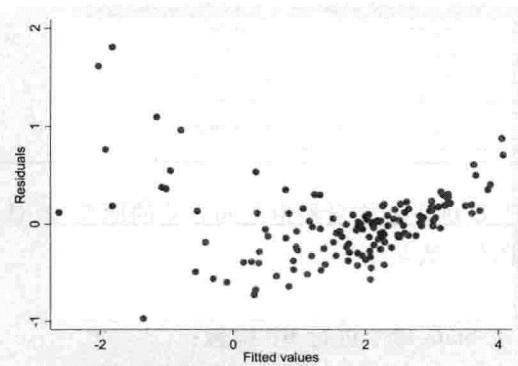


图 7.2 残差与拟合值的散点图

^① 指谁都可以用。

从图 7.2 可以大致看出,当 \hat{y} (即 Intc 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)

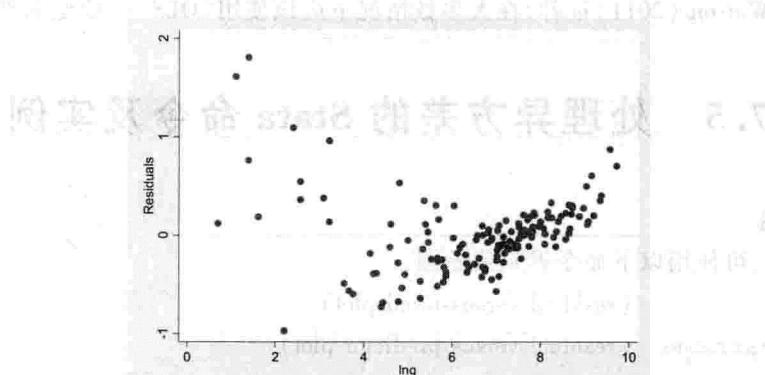


图 7.3 残差与解释变量 \lnq 的散点图

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 0.0000,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 \hat{y})

`estat hettest,rhs` (使用方程右边的解释变量,而不是 \hat{y})

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 0.0000, 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		
Model	2065.53636	1	2065.53636	F(1, 144) = 419.95		
Residual	708.275258	144	4.91857818	Prob > F = 0.0000		
Total	2773.81162	145	19.1297353	R-squared = 0.7447		
				Adj R-squared = 0.7429		
				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 $\ln q$ 可以解释 $\ln e_i^2$ 近 75% 的变动, 残差平方的变动与 $\ln q$ 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		
Model	173.069988	4	43.2674971	F(4, 140) = 895.03		
Residual	6.76790874	140	.048342205	Prob > F = 0.0000		
Total	179.837897	144	1.24887428	R-squared = 0.9624		
				Adj R-squared = 0.9613		
				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, $\ln pk$ 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 $\ln pl$ 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 $\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 \mathbf{V} \neq \sigma^2 \mathbf{I}_n$ ，其中 \mathbf{V} 为对称正定矩阵且已知。

(1) 计算 $\text{Var}(\mathbf{b} | \mathbf{X})$ ，其中 \mathbf{b} 为 OLS 估计量。

(2) 计算 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}} | \mathbf{X})$ ，其中 $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ 为 GLS 估计量。

(3) 矩阵 $\text{Var}(\mathbf{b} | \mathbf{X})$ 与 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}} | \mathbf{X})$ 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)：

$$\text{lprice}_i = \beta_1 + \beta_2 \text{lnox}_i + \beta_3 \text{ldist}_i + \beta_4 \text{rooms}_i + \beta_5 \text{stratio}_i + \varepsilon_i \quad (7.25)$$

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 $V_{n \times n}$ ，存在非退化矩阵 $C_{n \times n}$ ，使得 $V^{-1} = C'C$ 。

证明：由于 V 为实对称矩阵，故 V 可对角化，即存在正交矩阵 B ，满足 $B^{-1} = B'$ ，使得 $B^{-1}VB = A$ ，其中 A 为对

角矩阵。由于 V 正定, 故 A 的主对角线元素(即 V 的特征值)均为正, 而 A^{-1} 的主对角线元素也都为正。 B 中的向量为 V 的特征向量(B 不唯一):

$$B^{-1}VB = A \Rightarrow V = BA B^{-1} \quad (\text{左乘 } B, \text{ 右乘 } B^{-1})$$

$$V^{-1} = (B^{-1})^{-1}A^{-1}B^{-1} = BA^{-1}B^{-1} = (BA^{-1/2})(A^{-1/2}B') \equiv CC'$$

其中, $A^{-1/2}$ 为将对角矩阵 A^{-1} 中的每个元素开平方, 而 $C \equiv BA^{-1/2}$ 。由于 B 与 $BA^{-1/2}$ 都是非退化矩阵, 故 $C \equiv BA^{-1/2}$ 也非退化, 即 C^{-1} 存在。

由于 V^{-1} 为对称矩阵, 故 $V^{-1} = (V^{-1})' = (CC')' = C'C$ 。

注意: 由于 B 不唯一, 故 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 $E(\varepsilon_i \varepsilon_j | \mathbf{X}) \neq 0$, 即扰动项的协方差阵 $\text{Var}(\varepsilon | \mathbf{X})$ 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 $\text{Var}(\mathbf{b} | \mathbf{X})$ 的表达式不再是 $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$, 因为 $\text{Var}(\varepsilon | \mathbf{X}) \neq \sigma^2 \mathbf{I}$, 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 $E(\varepsilon_i \varepsilon_j | \mathbf{X}) > 0$, 参见图 8.1。

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 估计忽略了扰动项相关所包含的信息, 故不是最有效的估计方法。

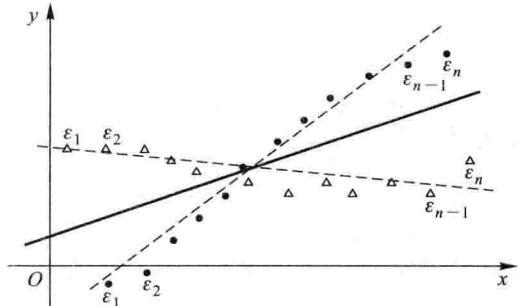


图 8.1 自相关的后果

^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 $\mathbf{g}_i = \mathbf{x}_i \varepsilon_i$ 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:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + u_t \quad (8.1)$$

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:

$$e_t \xrightarrow{\text{OLS}} x_{it}, \dots, x_{ik}, e_{t-1}, \dots, e_{t-p} \quad (t = p+1, \dots, n) \quad (8.2)$$

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:

$$(n-p)R^2 \xrightarrow{d} \chi^2(p) \quad (8.3)$$

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 e 中因滞后而缺失的项用其期望值 $E(e) = 0$ 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

$$\hat{\rho}_j \equiv \frac{\sum_{t=j+1}^n e_t e_{t-j}}{\sum_{t=1}^n e_t^2} \quad (j = 1, 2, \dots, p) \quad (8.4)$$

如果原假设 $H_0: \rho_1 = \dots = \rho_p = 0$ 成立,则 $\hat{\rho}_j \xrightarrow{p} 0, \sqrt{n}\hat{\rho}_j \xrightarrow{d}$ 正态分布, $j = 1, 2, \dots, p$ 。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

$$Q_{BP} \equiv n \sum_{j=1}^p \hat{\rho}_j^2 \xrightarrow{d} \chi^2(p) \quad (8.5)$$

经过改进的“Ljung-Box Q 统计量”(Ljung and Box, 1979)为

$$Q_{LB} \equiv n(n+2) \sum_{j=1}^p \frac{\hat{\rho}_j^2}{n-j} \xrightarrow{d} \chi^2(p) \quad (8.6)$$

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 $p = \min\{\lfloor n/2 \rfloor - 2, 40\}$,其中 $\lfloor n/2 \rfloor$ 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 $|x_{it}, \dots, x_{ik}|$,则由于样本残差 e_t 是 $|x_{it}, \dots, x_{ik}|$ 的函数,故可以将 $|x_{it}, \dots, x_{ik}|$ 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 $|e_{t-1}, \dots, e_{t-p}|$ 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 $|x_{it}, \dots, x_{ik}|$,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 作解释变量,需要知道 $e_0, e_{-1}, \dots, e_{-p+1}$,但我们并没有这些数据,故会损失 p 个样本值。

^③ 即令 $e_0 = e_{-1} = \dots = e_{-p+1} = 0$ 。

^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。

$$\begin{aligned} DW \equiv d \equiv & \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n e_t^2 - 2 \sum_{t=2}^n e_t e_{t-1} + \sum_{t=2}^n e_{t-1}^2}{\sum_{t=1}^n e_t^2} \\ & \approx 2 - 2 \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} = 2(1 - \hat{\rho}_1) \end{aligned} \quad (8.7)$$

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 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 $\mathbf{g}_i \equiv \mathbf{x}_i \varepsilon_i$ 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, \mathbf{g}_i 为鞅差分序列, 故

$$E(\mathbf{g}_i | \mathbf{g}_{i-1}, \dots, \mathbf{g}_1) = E(\mathbf{x}_i \varepsilon_i | \mathbf{g}_{i-1}, \dots, \mathbf{g}_1) = 0$$

因为模型含有截距项, 故向量 $\mathbf{g}_i \equiv \mathbf{x}_i \varepsilon_i$ 的第一个元素为 ε_i 。因此, $E(\varepsilon_i | \mathbf{g}_{i-1}, \dots, \mathbf{g}_1) = 0$ 。由于 $\{\varepsilon_{i-1}, \dots, \varepsilon_1\} \subset \{\mathbf{g}_{i-1}, \dots, \mathbf{g}_1\}$ (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

$$\begin{aligned} E(\varepsilon_i | \varepsilon_{i-1}, \dots, \varepsilon_1) &= E_{\mathbf{g}_{i-1}, \dots, \mathbf{g}_1} [E(\varepsilon_i | \varepsilon_{i-1}, \dots, \varepsilon_1) | \mathbf{g}_{i-1}, \dots, \mathbf{g}_1] \\ &= E_{\mathbf{g}_{i-1}, \dots, \mathbf{g}_1} [\underbrace{E(\varepsilon_i | \mathbf{g}_{i-1}, \dots, \mathbf{g}_1)}_{=0}] = 0 \end{aligned} \quad (8.8)$$

因此, ε_i 均值独立于 $(\varepsilon_{i-1}, \dots, \varepsilon_1)$, 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 $\mathbf{S}_{xx}^{-1} \hat{\mathbf{S}} \mathbf{S}_{xx}^{-1}$ 为夹心估计量, 其中 $\mathbf{S}_{xx} \equiv \frac{1}{n} \mathbf{X}' \mathbf{X}$, $\hat{\mathbf{S}} \equiv$

$\frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i'$ 。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 $\mathbf{S}_{xx}^{-1} \hat{\mathbf{Q}} \mathbf{S}_{xx}^{-1}$ (两侧

“面包”仍为 \mathbf{S}_{xx}^{-1} , 但中间的“菜”变为 $\hat{\mathbf{Q}}$), 其中

$$\hat{Q} = \hat{S} + \frac{1}{n} \sum_{j=1}^p \sum_{t=j+1}^n \left(1 - \frac{j}{p+1} \right) e_t e_{t-j}' (\mathbf{x}_t \mathbf{x}_{t-j}' + \mathbf{x}_{t-j} \mathbf{x}_t') \quad (8.9)$$

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n \quad (8.10)$$

则 β_1 的 OLS 估计量为

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) [\beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (8.11)$$

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

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (8.12)$$

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

$$\text{Var}(\hat{\beta}_1) = \frac{\text{Var}\left(\frac{1}{n} \sum_{i=1}^n v_i\right)}{(\sigma_x^2)^2} \quad (8.13)$$

首先考虑 $n=2$ 的最简单情形, 则上式的分子为

$$\begin{aligned} \text{Var}\left(\frac{1}{n} \sum_{i=1}^n v_i\right) &= \text{Var}\left[\frac{1}{2}(v_1 + v_2)\right] = \frac{1}{4} [\text{Var}(v_1) + \text{Var}(v_2) + 2\text{Cov}(v_1, v_2)] \\ &= \frac{1}{2}\sigma_v^2 + \frac{1}{2}\rho_1\sigma_v^2 = \frac{1}{2}\sigma_v^2(1 + \rho_1) \equiv \frac{1}{2}\sigma_v^2 f_2 \end{aligned} \quad (8.14)$$

其中, $\sigma_v^2 \equiv \text{Var}(v_i)$, $\rho_1 \equiv \text{corr}(v_1, v_2)$ 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 $\text{Var}\left(\frac{1}{n} \sum_{i=1}^n v_i\right) = \frac{1}{2}\sigma_v^2$, 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 $\text{Var}\left(\frac{1}{n} \sum_{i=1}^n v_i\right) =$

$\frac{1}{n}\sigma_v^2 f_n$, 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

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_v^2}{n(\sigma_x^2)^2} \cdot f_n \quad (8.15)$$

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 $\text{Var}[(v_1 + v_2 + \dots + v_n)/n]$ 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 $(\rho_1, \dots, \rho_{n-1})$ 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 x_{ij} , 残差为 e_{ij} , 然后定义 $u_j \equiv \sum_{i=1}^{M_j} e_{ij} x_{ij}$, 则聚类稳健的协方差矩阵可以写为

$$\frac{N-1}{N-K} \frac{M}{M-1} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{j=1}^M u_j' u_j \right) (\mathbf{X}'\mathbf{X})^{-1} \quad (8.16)$$

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” 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($M \rightarrow \infty$); 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 $\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X})$ 。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:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1, \quad u_t \text{ 为白噪声} \quad (8.17)$$

记扰动项的 j 阶协方差 $\rho_j \equiv \text{Cov}(\varepsilon_t, \varepsilon_{t-j} | \mathbf{X})$, 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。

$$\text{Var}(\boldsymbol{\varepsilon}|X) = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{n-2} \\ \vdots & \vdots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_0 \end{pmatrix} \quad (8.18)$$

容易证明(参见第5章), $\rho_0 = \sigma^2 = \text{Var}(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}$,其中 $\sigma_u^2 \equiv \text{Var}(u_t)$ 。

而 $\rho_1 = \rho\sigma^2$,故 $\frac{\rho_1}{\rho_0} = \frac{\rho\sigma^2}{\sigma^2} = \rho$ 为一阶自相关系数; $\rho_2 = \rho^2\sigma^2, \dots, \rho_{n-1} = \rho^{n-1}\sigma^2$,故

$$\text{Var}(\boldsymbol{\varepsilon}|X) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix} \equiv \sigma^2 V \quad (8.19)$$

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, $e_t = \hat{\rho}e_{t-1} + \text{error}_t$ 。也可以通过残差一阶自相关系数 $\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$,或 $\hat{\rho} = 1 - \frac{\text{DW}}{2}$ 来

估计 ρ 。知道 ρ 后,就可以得到 $\text{Var}(\boldsymbol{\varepsilon}|X) = \sigma^2 V$,并将 V 的逆矩阵分解为 $V^{-1} = \mathbf{C}'\mathbf{C}$ 。可以证明:

$$\mathbf{C} = \frac{1}{\sqrt{1-\rho^2}} \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ 0 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \quad (8.20)$$

以 $\sqrt{1-\rho^2}\mathbf{C}$ 左乘原模型,并定义 $\bar{\mathbf{y}} = \sqrt{1-\rho^2}\mathbf{C}\mathbf{y}, \bar{\mathbf{X}} = \sqrt{1-\rho^2}\mathbf{C}\mathbf{X}, \bar{\boldsymbol{\varepsilon}} = \sqrt{1-\rho^2}\mathbf{C}\boldsymbol{\varepsilon}$,则变换后的扰动项 $\bar{\boldsymbol{\varepsilon}}$ 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):

$$\bar{\mathbf{y}} = \sqrt{1-\rho^2}\mathbf{C}\mathbf{y} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ 0 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix} \quad (8.21)$$

$$\bar{\mathbf{X}} = \sqrt{1-\rho^2}\mathbf{C}\mathbf{X} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ 0 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ x_{21} & \cdots & x_{2K} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nK} \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{1-\rho^2}x_{11} & \cdots & \sqrt{1-\rho^2}x_{1K} \\ x_{21}-\rho x_{11} & \cdots & x_{2K}-\rho x_{1K} \\ \vdots & & \vdots \\ x_{n1}-\rho x_{n-1,1} & \cdots & x_{nK}-\rho x_{n-1,K} \end{pmatrix} \quad (8.22)$$

$$\tilde{\boldsymbol{\varepsilon}} = \sqrt{1-\rho^2} \mathbf{C} \boldsymbol{\varepsilon} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ 0 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} = \begin{pmatrix} \sqrt{1-\rho^2} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 - \rho \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n - \rho \boldsymbol{\varepsilon}_{n-1} \end{pmatrix} \quad (8.23)$$

可以写出每个观测值(个体)的回归方程:

$$\begin{aligned} \sqrt{1-\rho^2} y_1 &= \sqrt{1-\rho^2} \beta_1 + \sqrt{1-\rho^2} \beta_2 x_{12} + \cdots + \sqrt{1-\rho^2} \beta_K x_{1K} + \tilde{\boldsymbol{\varepsilon}}_1 \\ y_2 - \rho y_1 &= (1-\rho) \beta_1 + \beta_2 (x_{22} - \rho x_{12}) + \cdots + \beta_K (x_{2K} - \rho x_{1K}) + \tilde{\boldsymbol{\varepsilon}}_2 \\ &\dots \\ y_n - \rho y_{n-1} &= (1-\rho) \beta_1 + \beta_2 (x_{n2} - \rho x_{n-1,2}) + \cdots + \beta_K (x_{nK} - \rho x_{n-1,K}) + \tilde{\boldsymbol{\varepsilon}}_n \end{aligned} \quad (8.24)$$

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

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \boldsymbol{\varepsilon}_t \quad (8.25)$$

将上式滞后一期,然后方程两边同时乘以 ρ 得

$$\rho y_{t-1} = \rho \beta_1 + \rho \beta_2 x_{t-1,2} + \cdots + \rho \beta_K x_{t-1,K} + \rho \boldsymbol{\varepsilon}_{t-1} \quad (8.26)$$

将方程(8.25)减去方程(8.26)可得

$$y_t - \rho y_{t-1} = (1-\rho) \beta_1 + \beta_2 (x_{t2} - \rho x_{t-1,2}) + \cdots + \beta_K (x_{tK} - \rho x_{t-1,K}) + \boldsymbol{\varepsilon}_t - \rho \boldsymbol{\varepsilon}_{t-1} \quad (8.27)$$

其中, $t=2, \dots, K$ 。显然,新扰动项 $\boldsymbol{\varepsilon}_t - \rho \boldsymbol{\varepsilon}_{t-1} = u_t$ 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 β 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 $(\boldsymbol{\varepsilon}_t - \rho \boldsymbol{\varepsilon}_{t-1})$,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 $y_t = \rho y_{t-1} + \mathbf{x}'\beta + \varepsilon_t$ 。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 $y_t = \mathbf{x}'\beta + [\underbrace{\rho y_{t-1} + \varepsilon_t}_{=v_t}]$, 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 $D2.x_t = \Delta(\Delta x_t) = \Delta(x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$ (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：

$$\text{consumption}_t = \beta_0 + \beta_1 \text{temp}_t + \beta_2 \text{price}_t + \beta_3 \text{income}_t + \varepsilon_t \quad (8.28)$$

进行 OLS 回归，得

```
. reg consumption temp price income
```

^① 此例来自 Verbeek(2004)。

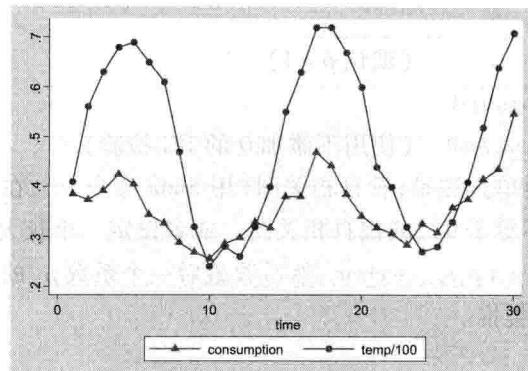


图 8.2 冰淇淋消费与气温的时间趋势

Source	SS	df	MS	Number of obs = 30			
Model	.090250523	3	.030083508	F(3, 26) =	22.17		
Residual	.035272835	26	.001356647	Prob > F =	0.0000		
				R-squared =	0.7190		
Total	.125523358	29	.004328392	Adj R-squared =	0.6866		
				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。

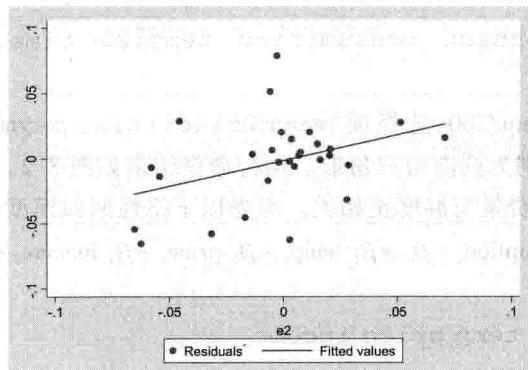


图 8.3 残差与残差滞后的散点图

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
```

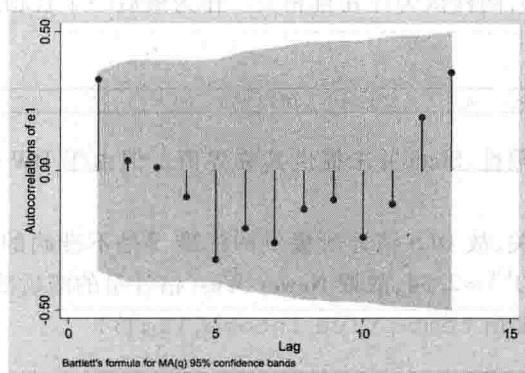


图 8.4 自相关图

```
. pac e1
```

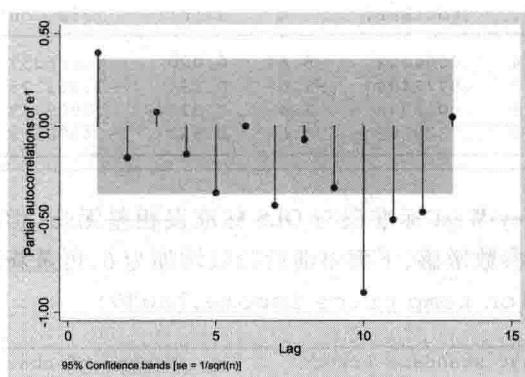


图 8.5 偏自相关图

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 0.0396, 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			
<hr/>					
consumption	Newey-West	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			
<hr/>					
consumption	Newey-West	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
Model	.047040596	3	.015680199	F(3, 25)	=	15.40
Residual	.025451894	25	.001018076	Prob > F	=	0.0000
Total	.072492491	28	.002589018	R-squared	=	0.6489
				Adj R-squared	=	0.6068
				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
Model	.04494596	3	.014981987	F(3, 26)	=	14.35
Residual	.027154354	26	.001044398	Prob > F	=	0.0000
Total	.072100315	29	.002486218	R-squared	=	0.6234
				Adj R-squared	=	0.5799
				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(-0.0008)。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 Prob > F = 0.0000 R-squared = 0.8285 Adj R-squared = 0.7999 Root MSE = .02987		
consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
temp	.0053321	.0006704	7.95	0.000	.0039484	.0067158
--.	-.0022039	.0007307	-3.02	0.006	-.0037119	-.0006959
L1.						
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(-0.0022);而当期气温仍然显著地为正(0.0053)。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 $n \rightarrow \infty$ 时, Box-Pierce Q 与 Ljung-Box Q 统计量之差依概率收敛于 0。[提示: 定义 $x_n = ((\sqrt{n}\hat{\rho}_1)^2 (\sqrt{n}\hat{\rho}_2)^2 \cdots (\sqrt{n}\hat{\rho}_p)^2)',$ 并把两个统计量之差写成 $a_n'x_n$ 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 $\tilde{\epsilon} = (\sqrt{1-\rho^2} \epsilon_1 (\epsilon_2 - \rho\epsilon_1) \cdots (\epsilon_n - \rho\epsilon_{n-1}))'$, 直接证明 $\tilde{\epsilon}$ 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 $\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + u_t$, 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:

$$\Delta rs_t = \alpha + \beta \Delta r20_{t-1} + \epsilon_t$$

(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

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + \varepsilon_i \quad (9.1)$$

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

$$y_i = x'_{i1}\beta_1 + u_i \quad (9.2)$$

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 $\text{Cov}(x_{i1}, x_{i2}) = 0$ 。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 $\text{Cov}(x_{i1}, x_{i2}) \neq 0$ 。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 $E(y|x_1)$ 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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma q + \varepsilon \quad (9.3)$$

其中, q 为不可观测的遗漏变量。假定 $\text{Cov}(x_k, \varepsilon) = 0, \forall k$, 但遗漏变量 q 与某解释变量 x_m 相关 ($1 \leq m \leq K$), 即 $\text{Cov}(x_m, q) \neq 0$, 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

$$q = \delta_0 + \delta_1 z + v, \quad \text{Cov}(z, v) = 0 \quad (9.4)$$

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 $\text{Cov}(z, \varepsilon) = 0$ 。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 $\text{Cov}(x_k, v) = 0, \forall k$ 。将 q 的表达式代入原模型(9.3)可得

$$y = (\beta_0 + \gamma \delta_0) + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma \delta_1 z + (\gamma v + \varepsilon) \quad (9.5)$$

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,

$$\text{Cov}(x_k, \gamma v + \varepsilon) = \gamma \underbrace{\text{Cov}(x_k, v)}_{\text{condition 2}} + \underbrace{\text{Cov}(x_k, \varepsilon)}_{\text{assumption}} = 0 + 0 = 0 \quad (\forall k) \quad (9.6)$$

$$\text{Cov}(z, \gamma v + \varepsilon) = \gamma \underbrace{\text{Cov}(z, v)}_{\text{assumption}} + \underbrace{\text{Cov}(z, \varepsilon)}_{\text{condition 1}} = 0 + 0 = 0 \quad (9.7)$$

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

$$y_i = x'_{i1}\beta_1 + \varepsilon_i \quad (9.8)$$

其中, $\text{Cov}(x_{i1}, \varepsilon_i) = 0$ 。而实际估计的模型为

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + (\underbrace{\varepsilon_i - x'_{i2}\beta_2}_{=0}) \quad (9.9)$$

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 $\text{Cov}(x_{i2}, \varepsilon_i) = 0$ 。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 $\underset{n \rightarrow \infty}{\text{plim}} \hat{\beta}_1 = \beta_1$, $\underset{n \rightarrow \infty}{\text{plim}} \hat{\beta}_2 = \beta_2 = 0$ 。

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:

$$\min_K \text{AIC} \equiv \ln(e'e/n) + \frac{2}{n}K \quad (9.10)$$

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罚。

^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:

$$\min_K \text{BIC} \equiv \ln(\mathbf{e}'\mathbf{e}/n) + \frac{\ln n}{n}K \quad (9.11)$$

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:

$$\min_K \text{HQIC} \equiv \ln(\mathbf{e}'\mathbf{e}/n) + \frac{\ln[\ln(n)]}{n}K \quad (9.12)$$

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),

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1, \dots, T \quad (9.13)$$

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 $T \rightarrow \infty$ 时, $\Pr(\hat{p} < p) \rightarrow 0$, $\Pr(\hat{p} = p) \rightarrow 1$, $\Pr(\hat{p} > p) \rightarrow 0$ 。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 $\Pr(\hat{p} < p) \rightarrow 0$,但 $\Pr(\hat{p} > p) \rightarrow c > 0$ 。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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma x_1^2 + \delta x_2 x_3 + \varepsilon \quad (9.14)$$

分别计算各变量的边际效应如下,

$$\frac{E(y)}{\partial x_1} = \beta_1 + 2\gamma x_1, \quad \frac{E(y)}{\partial x_2} = \beta_2 + \delta x_3, \quad \frac{E(y)}{\partial x_3} = \beta_3 + \delta x_2 \quad (9.15)$$

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 $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ 。回归后可得拟合值 $\hat{y} = \mathbf{x}'\mathbf{b}$ 。既然 \hat{y} 是解释变量 \mathbf{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:

$$y = \mathbf{x}'\boldsymbol{\beta} + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \delta_4 \hat{y}^4 + \mu \quad (9.16)$$

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 $\hat{y}^2, \hat{y}^3, \hat{y}^4$ 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:

$$y = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{error} \quad (9.17)$$

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
Model	277.574775	2	138.787388	F(2, 142) = 1460.70
Residual	13.4920481	142	.095014423	Prob > F = 0.0000
Total	291.066823	144	2.02129738	R-squared = 0.9536
				Adj R-squared = 0.9530
				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 $\ln q_2$):

```
. g lnq2 = lnq^2
```

```
. reg lntc lnq lnq2 lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs = 145		
Model	278.630831	5	55.7261661	F(5, 139) =	622.86	
Residual	12.4359927	139	.089467573	Prob > F =	0.0000	
Total	291.066823	144	2.02129738	R-squared =	0.9573	
				Adj R-squared =	0.9557	
				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“ $0.12 + 0.1 \ln q$ ”(对 $[0.12 \ln q + 0.05 (\ln q)^2]$ 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		
Model	278.638903	2	139.319451	F(2, 142) =	1591.85	
Residual	12.4279206	142	.087520568	Prob > F =	0.0000	
Total	291.066823	144	2.02129738	R-squared =	0.9567	
				Adj R-squared =	0.9567	
				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):

$$\begin{aligned} \ln c = & \beta_0 + \beta_1 \ln q + \beta_2 \ln p_l + \beta_3 \ln p_k + \beta_4 \ln p_f + \delta_1 (\ln q)^2 + \delta_2 (\ln p_l)^2 \\ & + \delta_3 (\ln p_k)^2 + \delta_4 (\ln p_f)^2 + \gamma_{12} \ln q \cdot \ln p_l + \gamma_{13} \ln q \cdot \ln p_k \\ & + \gamma_{14} \ln q \cdot \ln p_f + \gamma_{23} \ln p_l \cdot \ln p_k + \gamma_{24} \ln p_l \cdot \ln p_f + \gamma_{34} \ln p_k \cdot \ln p_f + \varepsilon \end{aligned} \quad (9.18)$$

上式包含了解释变量 $\ln q, \ln p_l, \ln p_k, \ln p_f$ 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 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, $(X'X)^{-1}$ 不存在,总体参数 β 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 $\{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_K\}$ 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 $(X'X)$ 变得几乎不可逆,故从某种意义上来说, $(X'X)^{-1}$ 变得很“大”,致使方差 $\text{Var}(b|X) = \sigma^2 (X'X)^{-1}$ 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 X 中的元素轻微变化,就可能引起 $(X'X)^{-1}$ 很大的变化,进而导致 OLS 估计值 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

$$\text{Var}(b_k|X) = \frac{\sigma^2}{(1 - R_k^2) S_{kk}} \quad (9.19)$$

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。

$$VIF_k \equiv \frac{1}{1 - R_k^2} \quad (9.20)$$

则 $\text{Var}(b_k | X) = VIF_k \cdot (\sigma^2 / S_{kk})$ 。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 $\max\{VIF_1, \dots, VIF_K\}$, 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 lnpf lnpl lnpk lnq
. 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 $P = X(X'X)^{-1}X'$ 的第 i 个主对角线元素来表示:

$$\text{lev}_i \equiv x_i'(X'X)^{-1}x_i \quad (9.21)$$

所有观测数据的影响力 lev_i 满足: (i) $0 \leq \text{lev}_i \leq 1$, ($i = 1, \dots, n$); (ii) $\sum_{i=1}^n \text{lev}_i = K$ (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 $\mathbf{b}^{(i)}$ 为去掉第 i 个观测数据后的 OLS 估计值,可以证明:

$$\mathbf{b} - \mathbf{b}^{(i)} = \left(\frac{1}{1 - \text{lev}_i} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i e_i \quad (9.22)$$

因此, lev_i 越大则 $(\mathbf{b} - \mathbf{b}^{(i)})$ 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。

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

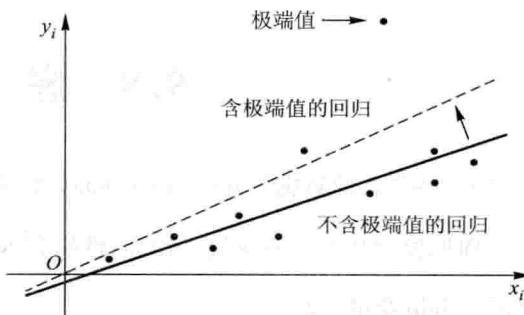


图 9.1 极端观测值对回归系数的影响

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 \mathbf{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 $(1 \cdots 1)'$ (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:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1950, \dots, 2000 \quad (9.23)$$

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:

$$D = \begin{cases} 1, & \text{若 } t \geq 1978 \\ 0, & \text{其他} \end{cases} \quad (9.24)$$

考虑以下两种情况。

(1) 仅仅引入虚拟变量本身

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t \quad (9.25)$$

显然,该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + \beta x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases} \quad (9.26)$$

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。

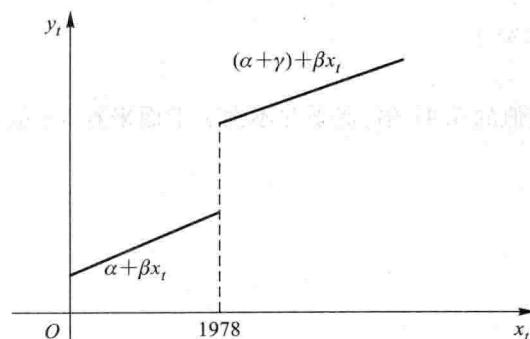


图9.2 仅引入虚拟变量的效果

(2) 引入虚拟变量,以及虚拟变量与解释变量的“互动项”(interaction term)

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \quad (9.27)$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + (\beta + \delta) x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases} \quad (9.28)$$

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。

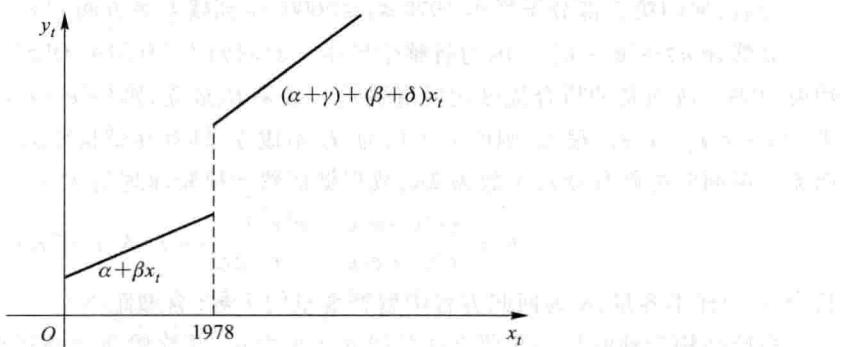


图 9.3 引入虚拟变量及其互动项的效果

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 $1950 \leq t < 1978$, 第 2 个时期为 $1978 \leq t \leq 2000$, 则两个时期对应的回归方程可以分别记为

$$\mathbf{y}^1 = \mathbf{X}^1 \boldsymbol{\beta}^1 + \boldsymbol{\varepsilon}^1 \quad (9.29)$$

$$\mathbf{y}^2 = \mathbf{X}^2 \boldsymbol{\beta}^2 + \boldsymbol{\varepsilon}^2 \quad (9.30)$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9.31)$$

其中, $\mathbf{y} = \begin{pmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}^1 \\ \boldsymbol{\varepsilon}^2 \end{pmatrix}$ 。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 $1950 \leq t \leq 2000$,得到残差平方和 $\mathbf{e}'\mathbf{e}$ 。

其次,回归第1部分子样本 $1950 \leq t < 1978$,得到残差平方和 $\mathbf{e}'\mathbf{e}_1$ 。

最后,回归第2部分子样本 $1978 \leq t \leq 2000$,得到残差平方和 $\mathbf{e}'\mathbf{e}_2$ 。

显然, $\mathbf{e}'\mathbf{e} \geq \mathbf{e}'\mathbf{e}_1 + \mathbf{e}'\mathbf{e}_2$,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 $(\mathbf{e}'\mathbf{e} - \mathbf{e}'\mathbf{e}_1 - \mathbf{e}'\mathbf{e}_2)$ 应该比较小。如果 $(\mathbf{e}'\mathbf{e} - \mathbf{e}'\mathbf{e}_1 - \mathbf{e}'\mathbf{e}_2)$ 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

$$F = \frac{(\mathbf{e}'\mathbf{e} - \mathbf{e}'\mathbf{e}_1 - \mathbf{e}'\mathbf{e}_2)/K}{(\mathbf{e}'\mathbf{e}_1 + \mathbf{e}'\mathbf{e}_2)/(n-2K)} \sim F(K, n-2K) \quad (9.32)$$

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:

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \quad (9.33)$$

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 ($\tau_0 \leq t \leq \tau_1$) 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)
```

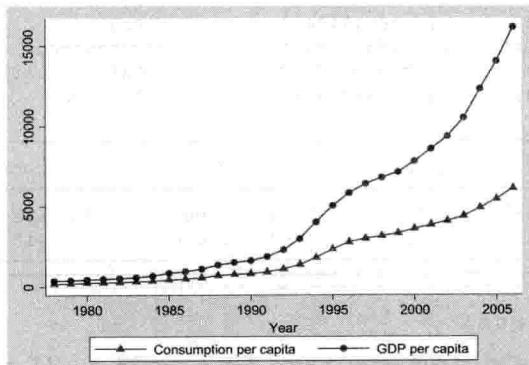


图 9.4 居民人均消费与人均国内总产值时间趋势

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:

$$c_t = \alpha + \beta y_t + \varepsilon_t \quad (9.34)$$

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			
Model	92575330.6	1	92575330.6	F(1, 27) = 2441.51			
Residual	1023766.1	27	37917.2631	Prob > F = 0.0000			
Total	93599096.7	28	3342824.88	R-squared = 0.9891			
				Adj R-squared = 0.9887			
				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			
Model	829144.945	1	829144.945	F(1, 12) = 4381.66			
Residual	2270.7688	12	189.230733	Prob > F = 0.0000			
Total	831415.714	13	63955.0549	R-squared = 0.9973			
				Adj R-squared = 0.9970			
				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			
Model	28749474.3	1	28749474.3	F(1, 13) = 764.58			
Residual	488822.141	13	37601.7031	Prob > F = 0.0000			
Total	29238296.4	14	2088449.74	R-squared = 0.9833			
				Adj R-squared = 0.9820			
				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, $e'e = ssr$, $e'_1e_1 = ssr1$, $e'_2e_2 = ssr2$, $K = 2$, $n = 29$, $n - 2K = 25$, 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			
Total	93599096.7	28	3342824.88	R-squared = 0.9948			
				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 0.0001, 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	Robust						
	Coef.	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

$$\hat{x}_t = \frac{x_{t-1} + x_{t+1}}{2} \quad (9.35)$$

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

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0 \quad (9.36)$$

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 $\exp(\ln \hat{y})$ 。

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)
```

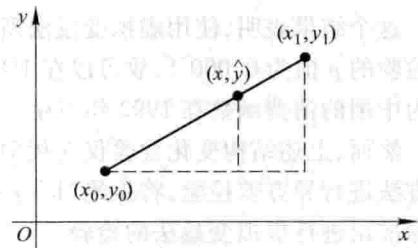


图9.5 线性插值示意图

由于人均 GDP 有指数增长趋势,故更好的做法是,先对 y_1 取对数,进行线性插值,再取反对数,并将结果记为 y_3 。

```
. 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 y_2 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 y_3 (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 X 中某列的数值是另一列的很多倍,这可能使计算机在对 $(X'X)^{-1}$ 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) $\Pr(\hat{p} < p) \rightarrow 0$; (2) $\Pr(\hat{p} > p) \rightarrow 0$; 故 $\Pr(\hat{p} = p) \rightarrow 1$ 。对于更一般的情形 $0 \leq p \leq p_{\max}$,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 $BIC(0) < BIC(1)$, 即 $BIC(0) - BIC(1) < 0$ 。记相应的残差平方和为 $SSR(0)$ 与 $SSR(1)$, 则

$$\begin{aligned} BIC(0) - BIC(1) &= \left[\ln\left(\frac{SSR(0)}{T}\right) + \frac{\ln T}{T} \right] - \left[\ln\left(\frac{SSR(1)}{T}\right) + \frac{2\ln T}{T} \right] \\ &= \ln\left(\frac{SSR(0)}{T}\right) - \ln\left(\frac{SSR(1)}{T}\right) - \frac{\ln T}{T} \end{aligned} \quad (9.37)$$

当 $\hat{p}=0$ 时, $y_t = \hat{\beta}_0 + e_t$, 故 $\frac{SSR(0)}{T} = \frac{T-1}{T} \cdot s_y^2 \xrightarrow{p} \sigma_e^2$; 另一方面, 当 $\hat{p}=1$ 时, $y_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1} + e_t$, 故 $\frac{SSR(1)}{T} = \frac{T-1}{T} \cdot s_y^2 \xrightarrow{p} \sigma_e^2$ 。根据微积分的洛必达法则可知, $\frac{\ln T}{T} \rightarrow 0$ 。因此,

$$BIC(0) - BIC(1) \xrightarrow{p} (\ln \sigma_y^2 - \ln \sigma_e^2) > 0 \quad (9.38)$$

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, $\Pr[BIC(0) - BIC(1) < 0] < 0$, 故 $\Pr(\hat{p}=0) \rightarrow 0$ 。

(2) 要选择 $\hat{p}=2$, 必须有 $BIC(2) < BIC(1)$, 即 $BIC(2) - BIC(1) < 0$ 。

$$\begin{aligned} T[BIC(2) - BIC(1)] &= T\left\{\left[\ln\left(\frac{SSR(2)}{T}\right) + \frac{3\ln T}{T} \right] - \left[\ln\left(\frac{SSR(1)}{T}\right) + \frac{2\ln T}{T} \right]\right\} \\ &= T\ln\left(\frac{SSR(2)}{SSR(1)}\right) + \ln T \\ &= -T\ln\left(1 + \frac{F}{T-2}\right) + \ln T \end{aligned} \quad (9.39)$$

其中, $F \equiv \frac{SSR(1) - SSR(2)}{SSR(2)/(T-2)}$ 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,

$$\begin{aligned} \Pr[BIC(2) - BIC(1) < 0] &= \Pr[T(BIC(2) - BIC(1)) < 0] \\ &= \Pr[-T\ln\left(1 + \frac{F}{T-2}\right) + \ln T < 0] \\ &= \Pr[T\ln\left(1 + \frac{F}{T-2}\right) > \ln T] \end{aligned} \quad (9.40)$$

由于 $\ln(1+a) \approx a$ (当 $a \rightarrow 0$), 故 $\ln\left(1 + \frac{F}{T-2}\right) \approx \frac{F}{T-2}$ (由于 F 的渐近分布给定, 故 F 的取值有限, 而 T 不断增大, 故 $\frac{F}{T-2}$ 很小, 此近似成立), 因此, $T\ln\left(1 + \frac{F}{T-2}\right) \approx T \cdot \frac{F}{T-2} \approx F$ 。综合上面的结果, 有

$$\Pr[BIC(2) - BIC(1) < 0] \longrightarrow \Pr[F > \ln T] \rightarrow 0 \quad (9.41)$$

因此, $\Pr(\hat{p}=2) \rightarrow 0$ 。

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 $\Pr(\hat{p}=0) \rightarrow 0$ 。

在 A9.1(2) 的证明中, 将 $\ln T$ 替换为 $\ln(\ln T)$, 则

$$\Pr[BIC(2) - BIC(1) < 0] \longrightarrow \Pr[F > \ln(\ln T)] \rightarrow 0 \quad (9.42)$$

因此, $\Pr(\hat{p}=2) \rightarrow 0$ 。

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 $\Pr(\hat{p}=0) \rightarrow 0$ 。

在 A9.1(2) 的证明中, 将 $\ln T$ 替换为 2, 则

$$\Pr[BIC(2) - BIC(1) < 0] \longrightarrow \Pr[F > 2] > 0 \quad (9.43)$$

如果扰动项 ε_t 同方差, 则 F 的渐近分布为 $\chi^2(1)$, 故 $\Pr[F > 2] \rightarrow \Pr[\chi^2(1) > 2] = 0.16$ 。在一般情况下, $\Pr(\hat{p}=2) \rightarrow c > 0$ 。因此, $\Pr(\hat{p}=p) \rightarrow d < 1$ 。

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

$$\begin{cases} q_t^d = \alpha_0 + \alpha_1 p_t + u_t & (\text{需求}) \\ q_t^s = \beta_0 + \beta_1 p_t + v_t & (\text{供给}) \\ q_t^d = q_t^s & (\text{均衡}) \end{cases} \quad (10.1)$$

令 $q_t \equiv q_t^d = q_t^s$, 可得

$$\begin{cases} q_t = \alpha_0 + \alpha_1 p_t + u_t \\ q_t = \beta_0 + \beta_1 p_t + v_t \end{cases} \quad (10.2)$$

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。

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:

$$\begin{cases} p_t = p_t(u_t, v_t) = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_t - u_t}{\alpha_1 - \beta_1} \\ q_t = q_t(u_t, v_t) = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_t - \beta_1 u_t}{\alpha_1 - \beta_1} \end{cases} \quad (10.3)$$

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。

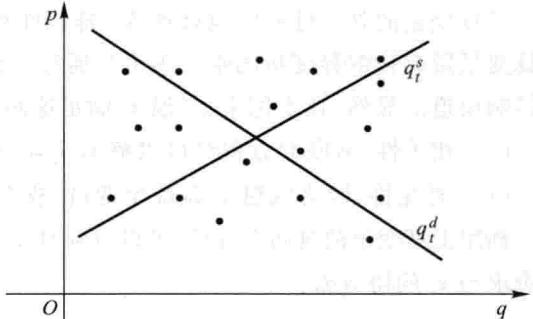


图 10.1 需求与供给决定市场均衡

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因素就是工具变量。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:

$$q_t^s = \beta_0 + \beta_1 p_t + \beta_2 x_t + v_t \quad (10.4)$$

假定气温 x_t 是个前定变量^③,与两个扰动项都不相关,即 $\text{Cov}(x_t, u_t) = 0, \text{Cov}(x_t, v_t) = 0$ 。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 $\text{Cov}(x_t, p_t) \neq 0$ 。
- (ii) 外生性:工具变量与扰动项不相关,即 $\text{Cov}(x_t, u_t) = 0$ 。

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 $\text{Cov}(x_t, p_t) \neq 0$ 。
- (ii) 外生性:因为气温 x_t 是前定变量,故 $\text{Cov}(x_t, u_t) = 0$ 。

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:

$$\begin{aligned} \text{Cov}(q_t, x_t) &= \text{Cov}(\alpha_0 + \alpha_1 p_t + u_t, x_t) \\ &= \alpha_1 \text{Cov}(p_t, x_t) + \underbrace{\text{Cov}(u_t, x_t)}_{=0} = \alpha_1 \text{Cov}(p_t, x_t) \end{aligned} \quad (10.5)$$

根据工具变量的相关性, $\text{Cov}(p_t, x_t) \neq 0$,可以把上式两边同除以 $\text{Cov}(p_t, x_t)$,得

$$\alpha_1 = \frac{\text{Cov}(q_t, x_t)}{\text{Cov}(p_t, x_t)} \quad (10.6)$$

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。

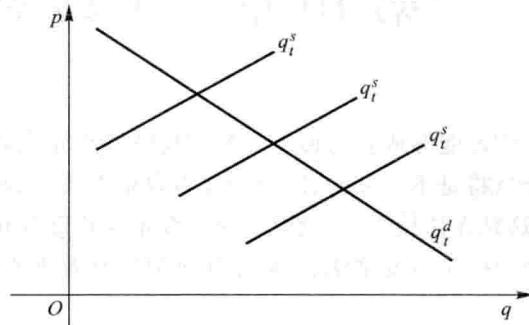


图 10.2 稳定的需求与变动的供给

$$\hat{\alpha}_{1,IV} = \frac{\widehat{\text{Cov}(q_t, x_t)}}{\widehat{\text{Cov}(p_t, x_t)}} \xrightarrow{p} \frac{\text{Cov}(q_t, x_t)}{\text{Cov}(p_t, x_t)} = \alpha_1 \quad (10.7)$$

从上式也可以看出,如果工具变量与内生变量无关,即 $\text{Cov}(x_t, p_t) = 0$,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 $\text{Cov}(x_t, p_t) \approx 0$,则会导致估计量 $\hat{\alpha}_{1,IV}$ 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; Basman, 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 $p_t \xrightarrow{\text{OLS}} x_t$,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 $q_t \xrightarrow{\text{OLS}} \hat{p}_t$ 。

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

$$q_t = \alpha_0 + \alpha_1 \hat{p}_t + \underbrace{[u_t + \alpha_1(p_t - \hat{p}_t)]}_{=\varepsilon_t} \quad (10.8)$$

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

$$\text{Cov}(\hat{p}_t, \varepsilon_t) = \text{Cov}(\hat{p}_t, u_t) + \alpha_1 \text{Cov}(\hat{p}_t, p_t - \hat{p}_t) \quad (10.9)$$

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 $\text{Cov}(x_t, u_t) = 0$ (工具变量的外生性),故上式右边的第一项 $\text{Cov}(\hat{p}_t, u_t) = 0$ 。

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 $\text{Cov}(\hat{p}_t, p_t - \hat{p}_t) = 0$ 。

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

$$\begin{cases} C_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases} \quad (10.10)$$

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

$$y = \alpha + \beta x^* + \varepsilon, \quad \text{Cov}(x^*, \varepsilon) = 0, \quad \beta \neq 0 \quad (10.11)$$

但 x^* 无法精确观测,而只能观测到 x ,二者满足如下关系:

$$x = x^* + u, \quad \text{Cov}(x^*, u) = 0, \quad \text{Cov}(u, \varepsilon) = 0 \quad (10.12)$$

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 $\hat{\alpha}_{1,IV}$,但(10.6)式无法定义,而且 $\hat{\alpha}_{1,IV}$ 也不是 α_1 的一致估计。

$$y = \alpha + \beta x + (\varepsilon - \beta u) \quad (10.13)$$

可以证明,新扰动项($\varepsilon - \beta u$)与解释变量 x 存在相关性:

$$\begin{aligned} \text{Cov}(x, \varepsilon - \beta u) &= \text{Cov}(x^* + u, \varepsilon - \beta u) \\ &= \underbrace{\text{Cov}(x^*, \varepsilon)}_{=0} - \beta \underbrace{\text{Cov}(x^*, u)}_{=0} + \underbrace{\text{Cov}(u, \varepsilon)}_{=0} - \beta \text{Cov}(u, u) \quad (10.14) \\ &= -\beta \text{Var}(u) \neq 0 \end{aligned}$$

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

$$\begin{aligned} \hat{\beta} &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \\ &\stackrel{p}{\longrightarrow} \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} = \frac{\text{Cov}(x_i^* + u, \alpha + \beta x^* + \varepsilon)}{\text{Var}(x_i^* + u)} \\ &= \frac{\beta \text{Var}(x_i^*)}{\text{Var}(x_i^*) + \text{Var}(u)} = \beta \cdot \frac{1}{1 + (\sigma_u^2 / \sigma_{x^*}^2)} \quad (10.15) \end{aligned}$$

在上式中,由于 σ_u^2 与 $\sigma_{x^*}^2$ 一定为正,故 $0 < \frac{1}{1 + (\sigma_u^2 / \sigma_{x^*}^2)} < 1$ 。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, $\frac{1}{1 + (\sigma_u^2 / \sigma_{x^*}^2)} \rightarrow 0$,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

$$y^* = \beta x + \varepsilon, \quad \text{Cov}(x, \varepsilon) = 0, \quad \beta \neq 0 \quad (10.16)$$

但 y^* 无法精确观测,而只能观测到 y ,二者满足如下关系:

$$y = y^* + v \quad (10.17)$$

其中, v 为测量误差。将方程(10.17)代入(10.16)可得

$$y = \beta x + (\varepsilon + v) \quad (10.18)$$

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 $\text{Cov}(x, v) = 0$,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: $E(x) = \mu$

二阶原点矩: $E(x^2) = \text{Var}(x) + [E(x)]^2 = \sigma^2 + \mu^2$

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：

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n x_i = \hat{\mu} \\ \frac{1}{n} \sum_{i=1}^n x_i^2 = \hat{\mu}^2 + \hat{\sigma}^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases} \quad (10.19)$$

其中, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 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 \mathbf{x} 的函数 $f(\mathbf{x})$ 的期望 $E[f(\mathbf{x})]$ 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:

$$\begin{aligned} E[\mathbf{x}_i \varepsilon_i] &= \mathbf{0} \Rightarrow E[\mathbf{x}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0} \quad (\text{代入 } \varepsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}) \\ &\Rightarrow E(\mathbf{x}_i y_i) = E(\mathbf{x}_i \mathbf{x}'_i) \boldsymbol{\beta} \quad (\text{展开、移项}) \\ &\Rightarrow \boldsymbol{\beta} = [E(\mathbf{x}_i \mathbf{x}'_i)]^{-1} E(\mathbf{x}_i y_i) \quad (\text{假设 } [E(\mathbf{x}_i \mathbf{x}'_i)]^{-1} \text{ 存在, 求解 } \boldsymbol{\beta}) \end{aligned}$$

以样本矩替代上式中的总体矩,即可得到矩估计:

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}} \quad (10.20)$$

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

$$y_i = \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{i,K-1} + \beta_K x_{iK} + \varepsilon_i \quad (10.21)$$

其中,只有最后一个解释变量 x_{iK} 为内生变量,即 $\text{Cov}(x_{iK}, \varepsilon_i) \neq 0$,因此 OLS 是不一致的。假设有一个有效工具变量 w 满足 $\text{Cov}(x_{iK}, w_i) \neq 0$ (相关性),以及 $\text{Cov}(w_i, \varepsilon_i) = 0$ (外生性)。由于 x_1, \dots, 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 $\mathbf{x}_i \equiv (x_{i1} \cdots x_{i,K-1} x_{iK})'$,则原模型为 $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$ 。记工具向量 $\mathbf{z}_i \equiv (z_{i1} \cdots z_{i,K-1} z_{iK})' \equiv (x_{i1} \cdots x_{i,K-1} w_i)'$ 。定义 $\mathbf{g}_i \equiv \mathbf{z}_i \varepsilon_i$ 。由于工具向量与扰动项正交,故 $E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 为“总体矩条件”或“正交条件”(orthogonality condition)。由此可得

$$\begin{aligned} E(\mathbf{z}_i \varepsilon_i) &= \mathbf{0} \Rightarrow E[\mathbf{z}_i(y_i - \mathbf{x}'_i \boldsymbol{\beta})] = \mathbf{0} \quad (\text{代入 } \varepsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}) \\ &\Rightarrow E(\mathbf{z}_i y_i) = [E(\mathbf{z}_i \mathbf{x}'_i)] \boldsymbol{\beta} \quad (\text{展开、移项}) \\ &\Rightarrow \boldsymbol{\beta} = [E(\mathbf{z}_i \mathbf{x}'_i)]^{-1} E(\mathbf{z}_i y_i) \quad (\text{假设 } [E(\mathbf{z}_i \mathbf{x}'_i)]^{-1} \text{ 存在}) \end{aligned}$$

以样本矩代替上式中的总体矩,即可得到工具变量估计量,

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i y_i \right) = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y} \quad (10.22)$$

其中, $\mathbf{Z} \equiv (z_1 \cdots z_{n-1} z_n)'$ 。显然,OLS 也是一种工具变量法。这是因为,如果 \mathbf{x}_i 全部是前定变量,则可以将自己作为工具变量,即 $\mathbf{z}_i = \mathbf{x}_i$, $\mathbf{Z} = \mathbf{X}$ 。因此, $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}}$ 。

工具变量法具有怎样的大样本性质呢?

命题 如果秩条件“ $\text{rank}[E(\mathbf{z}_i \mathbf{x}'_i)] = K$ ”成立(即方阵 $E(\mathbf{z}_i \mathbf{x}'_i)$ 满秩),则在一定的正则条件下, $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 是 $\boldsymbol{\beta}$ 的一致估计,且 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 服从渐近正态分布。

证明: 因为 $\text{rank}[E(\mathbf{z}_i \mathbf{x}'_i)] = K$, 故 $[E(\mathbf{z}_i \mathbf{x}'_i)]^{-1}$ 存在。因此, 抽样误差

$$\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' (\mathbf{X} \boldsymbol{\beta} + \varepsilon) - \boldsymbol{\beta} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \varepsilon$$

$$\begin{aligned}
 &= \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \boldsymbol{\varepsilon}_i \right) \\
 &= S_{ZX}^{-1} \bar{\mathbf{g}} \xrightarrow{P} [E(\mathbf{z}_i \mathbf{x}'_i)]^{-1} \underbrace{E(\mathbf{g}_i)}_{=0} = \mathbf{0}
 \end{aligned} \tag{10.23}$$

其中, $S_{ZX} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}'_i$, $\bar{\mathbf{g}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \boldsymbol{\varepsilon}_i$ 。根据与“大样本 OLS 理论”(第 5 章)类似的假定与推导, 可以证明, $\sqrt{n} \bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$, 其中 $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}'_i) = E(\boldsymbol{\varepsilon}_i^2 \mathbf{z}_i \mathbf{z}'_i)$ 。进一步, 工具变量估计量 $\hat{\boldsymbol{\beta}}_{IV}$ 演近地服从正态分布, 即 $\sqrt{n} (\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\boldsymbol{\beta}}_{IV}))$, 其中演近方差矩阵 $\text{Avar}(\hat{\boldsymbol{\beta}}_{IV}) = [E(\mathbf{z}_i \mathbf{x}'_i)]^{-1} \mathbf{S} [E(\mathbf{x}_i \mathbf{z}'_i)]^{-1}$ 。

秩条件 $\text{rank}[E(\mathbf{z}_i \mathbf{x}'_i)] = K$ 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, $K = 2$, $\mathbf{x}_i = (1 \ x_i)'$, $\mathbf{z}_i = (1 \ w_i)'$, 则 $E(\mathbf{z}_i \mathbf{x}'_i) = E\left[\begin{pmatrix} 1 \\ w_i \end{pmatrix} (1 \ x_i)\right] = E\left[\begin{pmatrix} 1 & x_i \\ w_i & w_i x_i \end{pmatrix}\right] = \begin{bmatrix} 1 & E(x_i) \\ E(w_i) & E(w_i x_i) \end{bmatrix}$, 因此,

$$\begin{aligned}
 \text{rank}[E(\mathbf{z}_i \mathbf{x}'_i)] = K = 2 &\Leftrightarrow \text{行列式} \begin{vmatrix} 1 & E(x_i) \\ E(w_i) & E(w_i x_i) \end{vmatrix} \neq 0 \\
 &\Leftrightarrow E(w_i x_i) - E(w_i) E(x_i) \neq 0 \\
 &\Leftrightarrow \text{Cov}(w_i, x_i) \neq 0, \text{即 } w_i \text{ 与 } x_i \text{ 相关。}
 \end{aligned}$$

阶条件: 显然, 满足秩条件的必要条件是 \mathbf{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, $\mathbf{Z}'\mathbf{X}$ 不是方阵, $(\mathbf{Z}'\mathbf{X})^{-1}$ 不存在, 也就无法定义工具变量估计量 $\hat{\boldsymbol{\beta}}_{IV}$ 。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。有效的方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 x_1, \dots, x_K 分别对所有 L 个工具变量 $\{z_1, z_2, \dots, z_L\}$ 作 OLS 回归, 其中第 k 个解释变量 $x_k \equiv (x_{1k} \cdots x_{nk})'_{n \times 1}, k = 1, \dots, K$ (不同于第 3 章对第 i 个观测数据 x_i 的定义)。得到拟合值

$$\hat{x}_1 = P x_1, \quad \hat{x}_2 = P x_2, \quad \dots, \quad \hat{x}_K = P x_K \quad (10.24)$$

其中, $P \equiv Z(Z'Z)^{-1}Z'$ 为对应于 Z 的投影矩阵。写成矩阵形式, 可以定义

$$\hat{X} \equiv (\hat{x}_1 \ \hat{x}_2 \ \cdots \hat{x}_K)' = P(x_1 \ x_2 \ \cdots \ x_K)' = P X = Z[(Z'Z)^{-1}Z'X] \quad (10.25)$$

第二阶段(使用此外生部分进行回归)

由于 \hat{X} 是 $\{z_1, z_2, \dots, z_L\}$ 的线性组合(参见第一阶段回归), 故 \hat{X} 恰好包含 K 个工具变量。使用 \hat{X} 为工具变量对原模型 $y = X\beta + \epsilon$ 进行工具变量法估计:

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (\hat{X}'\hat{X})^{-1}\hat{X}'y \quad (10.26)$$

上式的最后一个等号能成立, 是由于 $\hat{X}'\hat{X} = (P X)'(P X) = X'P'P X = X'P'X = \hat{X}'X$, 其中, 投影矩阵 P 为对称幂等矩阵, 即 $P' = P, P^2 = P$ 。因此, 可以将 $\hat{\beta}_{IV}$ 视为把 y 对 \hat{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 $e_2 \equiv y - \hat{X}\hat{\beta}_{2SLS}$, 而原方程残差却是 $e \equiv y - X\hat{\beta}_{2SLS}$ 。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 $\hat{\beta}_{IV}$ 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, $\hat{\beta}_{IV}$ 的协方差矩阵估计量为 $\widehat{\text{Var}}(\hat{\beta}_{IV}) = s^2 (\hat{X}'\hat{X})^{-1}$, 其中 $s^2 \equiv \frac{e'e}{n-K}$, 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 $\widehat{\text{Var}}(\hat{\beta}_{IV}) = (\hat{X}'\hat{X})^{-1} \left(\sum_{i=1}^n e_i^2 \hat{x}_i \hat{x}_i' \right) (\hat{X}'\hat{X})^{-1}$ 。

将 $\hat{X} = Z(Z'Z)^{-1}Z'X$ 代入方程(10.26), 可得到 2SLS 的最终表达式:

$$\hat{\beta}_{2SLS} = (X'P X)^{-1}X'P y = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \quad (10.27)$$

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 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, 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 $\text{rank}[\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)] = K$ (满列秩), 其中 $\mathbf{z}_i = (\mathbf{z}_{i1} \cdots \mathbf{z}_{iL})'$ (L 个工具变量), $\mathbf{x}_i = (\mathbf{x}_{i1} \cdots \mathbf{x}_{iK})'$ (K 个解释变量), \mathbf{z}_i 与 \mathbf{x}_i 可有重叠元素, 且 $L \geq K$ (阶条件)。如果矩阵 $\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)$ 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“ $H_0 : \text{rank}[\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)] = K - 1$ ”, 而替代假设为“ $H_1 : \text{rank}[\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)] = K$ ”。

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 \mathbf{z} 与内生解释变量 \mathbf{x} 完全不相关, 则无法使用工具变量法, 因为 $[\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)]^{-1}$ 不存在^①。如果 \mathbf{z} 与 \mathbf{x} 仅仅微弱地相关, 则可大致认为 $[\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)]^{-1}$ 很大, 导致工具变量法估计量的渐近方差 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = [\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)]^{-1} \mathbf{S} [\mathbf{E}(\mathbf{z}_i \mathbf{x}'_i)]^{-1}$ 变得很大。直观上, 由于 \mathbf{z} 中仅包含很少与 \mathbf{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

$$y = \mathbf{x}'_1 \boldsymbol{\beta}_1 + x_2 \boldsymbol{\beta}_2 + \varepsilon \quad (10.28)$$

其中, 只有 x_2 为内生解释变量。记工具变量为 $(\mathbf{x}_1 \mathbf{z}_2)$, 其中 \mathbf{z}_2 为方程外的工具变量。在 2SLS 的第一阶段回归中, $x_2 \xrightarrow{\text{OLS}} \mathbf{x}_1, \mathbf{z}_2$, 其 R^2 包含了内生变量 x_2 与工具变量 \mathbf{z}_2 相关性的信息, 但也可能由于 x_2 与 \mathbf{x}_1 的相关性所造成。为此, 应该使用滤去 \mathbf{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 \mathbf{x}_1 回归, $x_2 \xrightarrow{\text{OLS}} \mathbf{x}_1$, 记其残差为 e_{x_2} , 代表 x_2 中不能由 \mathbf{x}_1 解释的部分; 其次, 把 \mathbf{z}_2 对 \mathbf{x}_1 回归, $\mathbf{z}_2 \xrightarrow{\text{OLS}} \mathbf{x}_1$, 记其残差为 e_{z_2} , 代表 \mathbf{z}_2 中不能由 \mathbf{x}_1 解释的部分; 最后, 对两个残差进行回归, 即 $e_{x_2} \xrightarrow{\text{OLS}} e_{z_2}$, 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, $x_2 = \mathbf{x}'_1 \boldsymbol{\gamma}_1 + \mathbf{z}'_2 \boldsymbol{\gamma}_2 + \text{error}$, 检验原假设 “ $H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}$ ”(即工具变量 \mathbf{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 \mathbf{z} 与 \mathbf{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 forceonrobust
```

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“forceonrobust”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 $\{x_1, \dots, x_{K-r}\}$ 为外生解释变量,而后 r 个解释变量 $\{x_{K-r+1}, \dots, x_K\}$ 为内生解释变量。假设共有 m 个方程外的工具变量 $\{z_1, \dots, z_m\}$,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:

$$e_{i,IV} = \gamma_1 x_{i1} + \dots + \gamma_{K-r} x_{i,K-r} + \delta_1 z_{i1} + \dots + \delta_m z_{im} + error_i \quad (10.29)$$

将工具变量法的残差 $e_{i,IV}$ 视为对扰动项 ε_i 的估计,则“扰动项 ε 与工具变量 $\{z_1, \dots, z_m\}$ 无关”的原假设可以写为“ $H_0: \delta_1 = \dots = \delta_m = 0$ ”。记此辅助回归的可决系数为 R^2 ,则 Sargan 统计量为

$$nR^2 \xrightarrow{d} \chi^2(m-r) \quad (10.30)$$

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 $\hat{\beta}_{IV}$ 都将收敛到相同的真实参数 β 。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 $\hat{\beta}_{IV}$ 与 $\hat{\beta}_{OLS}$ 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 β , 因此 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ (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 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 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 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 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

$$(\hat{\beta}_{IV} - \hat{\beta}_{OLS})' D^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \xrightarrow{d} \chi^2(r) \quad (10.31)$$

其中, $D \equiv \widehat{\text{Var}}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$, 而 D^{-1} 为 D 的广义逆矩阵(因为 D 不一定可逆; 但如果 D 正定, 则可逆)^①。当 D 可逆时, 广义逆矩阵就是一般的逆矩阵, 即 $D^{-1} = D^{-1}$ 。 r 为内生解释变量的个数(不包括外生解释变量)。容易证明

$$\text{Var}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = \text{Var}(\hat{\beta}_{IV}) + \text{Var}(\hat{\beta}_{OLS}) - 2\text{Cov}(\hat{\beta}_{IV}, \hat{\beta}_{OLS}) \quad (10.32)$$

但上式中的 $\text{Cov}(\hat{\beta}_{IV}, \hat{\beta}_{OLS})$ 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 $\text{Cov}(\hat{\beta}_{IV}, \hat{\beta}_{OLS}) = \text{Var}(\hat{\beta}_{OLS})$ (参见附录), 故

$$D = \widehat{\text{Var}}(\hat{\beta}_{IV}) - \widehat{\text{Var}}(\hat{\beta}_{OLS}) \quad (10.33)$$

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 $[\widehat{\text{Var}}(\hat{\beta}_{IV}) - \widehat{\text{Var}}(\hat{\beta}_{OLS})]$ 为正定矩阵。选择项“constant”表示 $\hat{\beta}_{IV}$ 与 $\hat{\beta}_{OLS}$ 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 $D \equiv \widehat{\text{Var}}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$, 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, $y = x'_1 \beta_1 + x_2 \beta_2 + \varepsilon$, 其中 x_2 为唯一的内生解释变量。记工具变量为 $z = (x_1 \ z_2)$, 其中 z_2 为方程外的工具变量。考虑 2SLS 的第一阶段回归, 即 $x_2 = x'_1 \gamma + z' \delta + v$ 。由于工具变量 z 与 ε 不相关, 故

^① D (不一定为方阵)的广义逆矩阵 D^{-1} 需要满足四个条件, 即 $DD^{-1}D = D$, $D^{-1}DD^{-1} = D^{-1}$, DD^{-1} 为对称矩阵, $D^{-1}D$ 为对称矩阵, 参见 Poirier(1995, p. 630)。

^② 参见 Durbin(1954), Wu(1973) 与 Hausman(1978)。

$$E(x_2 \varepsilon) = E[(x_1' \gamma + z' \delta + v) \varepsilon] = \underbrace{E(x_1' \gamma \varepsilon)}_{=0} + \underbrace{E(z' \delta \varepsilon)}_{=0} + E(v \varepsilon) = E(v \varepsilon) \quad (10.34)$$

这意味着

“ x_2 为内生变量” $\Leftrightarrow E(x_2 \varepsilon) \neq 0 \Leftrightarrow E(v \varepsilon) \neq 0$ ”

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:

$$\varepsilon = \rho v + \xi \quad (10.35)$$

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

$$y = \mathbf{x}' \boldsymbol{\beta}_1 + x_2 \beta_2 + \rho v + \xi \quad (10.36)$$

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:

$$y = \mathbf{x}' \boldsymbol{\beta}_1 + x_2 \beta_2 + \rho \hat{v} + \text{error} \quad (10.37)$$

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 $y = \mathbf{x}' \boldsymbol{\beta}_1 + \mathbf{x}' \boldsymbol{\beta}_2 + \varepsilon$, 则在第一阶段回归中可以得到与内生解释变量 x_2 相对应的多个残差 \hat{v} , 进行以下辅助回归:

$$y = \mathbf{x}' \boldsymbol{\beta}_1 + \mathbf{x}' \boldsymbol{\beta}_2 + \hat{v}' \boldsymbol{\rho} + \text{error} \quad (10.38)$$

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)

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n) \quad (10.39)$$

其中, $\mathbf{x}_i = (x_{i1} x_{i2} \cdots x_{iK})'$ 为第 i 个观测数据。

假定 10.2 漐近独立的平稳过程

记 L 维工具变量为 \mathbf{z}_i (可能与 \mathbf{x}_i 有重叠部分), \mathbf{w}_i 由 $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}$ 中不重复的变量构成且不含常数项。随机过程 $\{\mathbf{w}_i\}$ 为漐近独立的平稳过程。

假定 10.3 工具变量的正交性

所有工具变量 \mathbf{z}_i 均为“前定”, 即与同期扰动项正交。定义 L 维列向量 $\mathbf{g}_i = \mathbf{z}_i \varepsilon_i^{\top}$, 则 $E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 。

假定 10.4 秩条件

① 此处 \mathbf{g}_i 的定义与第 5 章“大样本 OLS”中不同。

$L \times K$ 维矩阵 $E(z_i x'_i)$ 满列秩, 即 $\text{rank}[E(z_i x'_i)] = K$ 。记 $\Sigma_{zx} \equiv E(z_i z'_i)$ 。

假定 10.5 $\{g_i\}$ 为鞅差分序列, 其协方差矩阵 $S \equiv E(g_i g'_i) = E(\varepsilon_i^2 z_i z'_i)$ 为非退化矩阵。

假定 10.6 四阶矩 $E[(x_{ik} z_{ij})^2]$ 存在且有限, $\forall i, j, k$ (finite fourth moments)。

10.6 GMM 的推导

与总体矩条件 $E(g_i) = E(z_i \varepsilon_i) = \mathbf{0}$ 相对应的样本矩条件为

$$g_n(\hat{\beta}) \equiv \frac{1}{n} \sum_{i=1}^n z_i (y_i - x'_i \hat{\beta}) = \mathbf{0} \quad (10.40)$$

将上式看成一个联立方程组, 则未知数 $\hat{\beta}$ 共有 K 个, 而方程个数为 L 个 (z_i 的维度)。如果 $L < K$, 为不可识别, 则 $\hat{\beta}$ 有无穷多解。如果 $L = K$, 为恰好识别, 则 $\hat{\beta}$ 有唯一解, 即 $\hat{\beta}_{IV}$ 。如果 $L > K$, 为过度识别, 则 $\hat{\beta}$ 无解。此时传统的矩估计法行不通。既然无法找到 $\hat{\beta}$ 使得 $g_n(\hat{\beta}) = \mathbf{0}$, 脑筋急转弯一下, 总可以找到 $\hat{\beta}$, 使得向量 $g_n(\hat{\beta})$ 尽可能地接近 $\mathbf{0}$, 比如, 使二次型 $(g_n(\hat{\beta}))' (g_n(\hat{\beta}))$ 最小。更一般地, 可以用一个“权重矩阵”(weighting matrix) W 来构成二次型。假设 \hat{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{W} 来表示), 而且 $\lim_{n \rightarrow \infty} \hat{W} = W$, 其中 W 为非随机的对称正定矩阵。定义最小化的目标函数为

$$\min_{\hat{\beta}} J(\hat{\beta}, \hat{W}) \equiv n(g_n(\hat{\beta}))' \hat{W} (g_n(\hat{\beta})) \quad (10.41)$$

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) :

$$\hat{\beta}_{GMM}(\hat{W}) \equiv \underset{\hat{\beta}}{\operatorname{argmin}} J(\hat{\beta}, \hat{W}) \quad (10.42)$$

其中, “argmin”(argument of the minimum) 表示能使 $J(\hat{\beta}, \hat{W})$ 最小化的 $\hat{\beta}$ 的取值。显然, GMM 估计量取决于权重矩阵 \hat{W} 。对 \hat{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{W} 使得 $\hat{\beta}_{GMM}$ 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 $S = E(g_i g'_i)$ 的主对角线元素), 是一个比较紧的约束, 故会通过 \hat{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), $g_n(\hat{\beta})$ 是 $\hat{\beta}$ 的一次函数, 故 $J(\hat{\beta}, \hat{W})$ 是 $\hat{\beta}$ 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) :

$$\hat{\beta}_{GMM}(\hat{W}) = (S'_{zx} \hat{W} S_{zx})^{-1} S'_{zx} \hat{W} S_{zy} \quad (10.43)$$

其中, $S_{zx} \equiv \frac{1}{n} \sum_{i=1}^n z_i x'_i$, $S_{zy} \equiv \frac{1}{n} \sum_{i=1}^n z_i y_i$ 。秩条件 $\text{rank}[E(z_i x'_i)] = K$ 及 \hat{W} 为正定矩阵保证了在大样本下, $(S'_{zx} \hat{W} S_{zx})^{-1}$ 存在。

在恰好识别的情况下, S_{zx} 为方阵, 则 GMM 还原为普通的工具变量法, 因为

$$\hat{\beta}_{GMM}(\hat{W}) = S_{zx}^{-1} \underbrace{\hat{W}^{-1} S'_{zx}^{-1} S'_{zx} \hat{W} S_{zy}}_{=I} = S_{zx}^{-1} S_{zy} = \hat{\beta}_{IV} \quad (10.44)$$

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) ($\hat{\beta}_{\text{GMM}}$ 为一致估计) 在假定 10.1 ~ 10.4 之下, $\underset{n \rightarrow \infty}{\text{plim}} \hat{\beta}_{\text{GMM}}(\hat{W}) = \beta$ 。
- (2) ($\hat{\beta}_{\text{GMM}}$ 为渐近正态) 如果假定 10.3 (即 $E(g_i) = \mathbf{0}$) 强化为假定 10.5 (即 $\{g_i\}$ 为鞅差分序列), 则 $\sqrt{n}(\hat{\beta}_{\text{GMM}} - \beta) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\beta}_{\text{GMM}}))$, 其中渐近协方差矩阵

$$\begin{aligned}\text{Avar}(\hat{\beta}_{\text{GMM}}) &= (\Sigma'_{ZX} W \Sigma_{ZX})^{-1} \Sigma'_{ZX} W S W \Sigma_{ZX} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}, \\ S &= E(g_i g'_i) = E(\varepsilon_i^2 z_i z'_i), \Sigma_{ZX} \equiv E(z_i x'_i)\end{aligned}$$

- (3) (Avar($\hat{\beta}_{\text{GMM}}$)的一致估计量) 如果 \hat{S} 是 S 的一致估计量, 则在假定 10.2 下, $\text{Avar}(\hat{\beta}_{\text{GMM}})$ 的一致估计量为

$$\widehat{\text{Avar}(\hat{\beta}_{\text{GMM}})} = (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \hat{S} \hat{W} S_{ZX} (S'_{ZX} \hat{W} S_{ZX})^{-1}$$

证明: (1) 抽样误差可以写为

$$\begin{aligned}\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta &= (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \left(\frac{1}{n} \sum_{i=1}^n z_i y_i \right) - \beta \\ &= (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \left(\frac{1}{n} \sum_{i=1}^n z_i (x'_i \beta + \varepsilon_i) \right) - \beta \\ &= (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \left(S_{ZX} \beta + \frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i \right) - \beta \\ &= (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \bar{g}\end{aligned}\tag{10.45}$$

其中, $\bar{g} \equiv \frac{1}{n} \sum_{i=1}^n g_i$, $g_i \equiv z_i \varepsilon_i$ 。由于 $(S'_{ZX} \hat{W} S_{ZX})^{-1} \xrightarrow{P} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$, $S'_{ZX} \hat{W} \xrightarrow{P} \Sigma'_{ZX} W$, 而 $\bar{g} \xrightarrow{P} E(g_i) = E(z_i \varepsilon_i) = \mathbf{0}$ 。因此, $\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta \xrightarrow{P} \mathbf{0}$ 。从这个证明可以看出, 保证 GMM 一致性的最重要条件仍是 $E(z_i \varepsilon_i) = \mathbf{0}$, 即工具变量与扰动项正交。

- (2) 由于抽样误差 $\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta = (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} \bar{g}$, 故 $\sqrt{n}(\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta) = (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} (\sqrt{n} \bar{g})$ 。根据假定 10.5 及鞅差分序列的中心极限定理, $\sqrt{n} \bar{g} \xrightarrow{d} N(\mathbf{0}, S)$, 其中 $S \equiv E(g_i g'_i) = E(\varepsilon_i^2 z_i z'_i)$ 。由于 $\sqrt{n}(\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta)$ 是 $\sqrt{n} \bar{g}$ 的线性组合, 故 $\sqrt{n}(\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\beta}_{\text{GMM}}))$ 。由于 $(S'_{ZX} \hat{W} S_{ZX})^{-1} \xrightarrow{P} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$, $S'_{ZX} \hat{W} \xrightarrow{P} \Sigma'_{ZX} W$, 故 $\text{Avar}(\hat{\beta}_{\text{GMM}}) = (\Sigma'_{ZX} W \Sigma_{ZX})^{-1} \Sigma'_{ZX} W S W \Sigma_{ZX} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$, 其中 $(\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$ 为对称矩阵。

- (3) 由于 $\hat{S} \xrightarrow{P} S$, 而且 $S_{ZX} \xrightarrow{P} \Sigma_{ZX}$, $\hat{W} \xrightarrow{P} W$, 故估计量 $\widehat{\text{Avar}(\hat{\beta}_{\text{GMM}})} = \underbrace{(S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W}}_{\text{面包}} \hat{S} \hat{W} S_{ZX} (S'_{ZX} \hat{W} S_{ZX})^{-1}$ 是 $\text{Avar}(\hat{\beta}_{\text{GMM}})$ 的一致估计量。从形式上看, 这也是一个夹心估计量(只是

$\hat{S} \hat{W} S_{ZX} (S'_{ZX} \hat{W} S_{ZX})^{-1}$ 是 $\text{Avar}(\hat{\beta}_{\text{GMM}})$ 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 β 的任何一致估计量 $\hat{\beta}$, 定义残差 $e_i \equiv y_i - \mathbf{x}'_i \hat{\beta}$, 则 $s^2 \equiv \frac{1}{n} \sum_{i=1}^n e_i^2$ 是 $\sigma^2 \equiv E(\varepsilon_i^2)$ 的一致估计,而且 $\hat{S} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}'_i$ 是 $S \equiv E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}'_i)$ 的一致估计。

命题 使 $Avar(\hat{\beta}_{GMM})$ 最小化的“最优权重矩阵”(optimal weighting matrix)为 $\hat{W} = \hat{S}^{-1}$ 。

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{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{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{S}^{-1} ,首先必须估计 \hat{S} 。由于 2SLS 也是一致的,故用 2SLS 的残差来计算 $\hat{S} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}'_i$ 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 $\hat{S} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}'_i$ 。

第二步:最小化 $J(\hat{\beta}, \hat{S}^{-1})$,得到 $\hat{\beta}_{GMM}(\hat{S}^{-1})$ 。

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{S} ,然后再求 $\hat{\beta}_{GMM}(\hat{S}^{-1})$,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 $E(\varepsilon_i^2 | \mathbf{z}_i) = \sigma^2 > 0$ (条件同方差^①),则根据迭代期望定律

$$S \equiv E(\mathbf{z}_i \mathbf{z}'_i \varepsilon_i^2) = E_{\mathbf{z}_i} E(\mathbf{z}_i \mathbf{z}'_i \varepsilon_i^2 | \mathbf{z}_i) = E_{\mathbf{z}_i} [\mathbf{z}_i \mathbf{z}'_i E(\varepsilon_i^2 | \mathbf{z}_i)] = \sigma^2 E(\mathbf{z}_i \mathbf{z}'_i) \quad (10.46)$$

因此, $\tilde{S} \equiv s^2 S_{zz}$ 是 S 的一致估计量,其中 $S_{zz} \equiv \frac{1}{n} \mathbf{Z}' \mathbf{Z}$ 。使用 $\tilde{S}^{-1} = (s^2 S_{zz})^{-1}$ 为最优权重矩阵,则最优 GMM 估计量为

$$\begin{aligned} \hat{\beta}_{GMM}(\tilde{S}^{-1}) &= (\mathbf{S}'_{zx} (s^2 S_{zz})^{-1} \mathbf{S}_{zx})^{-1} \mathbf{S}'_{zx} (s^2 S_{zz})^{-1} \mathbf{S}_{zy} \\ &= (\mathbf{S}'_{zx} \mathbf{S}_{zz}^{-1} \mathbf{S}_{zx})^{-1} \mathbf{S}'_{zx} \mathbf{S}_{zz}^{-1} \mathbf{S}_{zy} \end{aligned} \quad (10.47)$$

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 $\hat{\beta}_{GMM}$ 的取值。

由于 $S_{zx} \equiv \frac{1}{n} \mathbf{Z}' \mathbf{X}$, $S_{zz} \equiv \frac{1}{n} \mathbf{Z}' \mathbf{Z}$, $S_{zy} \equiv \frac{1}{n} \mathbf{Z}' \mathbf{y}$, 故

$$\begin{aligned} \hat{\beta}_{GMM}(\tilde{S}^{-1}) &= \left(\frac{1}{n} \mathbf{X}' \mathbf{Z} \cdot n(\mathbf{Z}' \mathbf{Z})^{-1} \cdot \frac{1}{n} \mathbf{Z}' \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}' \mathbf{Z} \cdot n(\mathbf{Z}' \mathbf{Z})^{-1} \frac{1}{n} \mathbf{Z}' \mathbf{y} \\ &= (\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} = \hat{\beta}_{2SLS} \end{aligned} \quad (10.48)$$

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{S}^{-1} ,而在条件同方差假定下,可以直接令 $\hat{S}^{-1} = S_{zz}^{-1}$ 。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 \mathbf{z}_i (而非解释变量 \mathbf{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 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1}) = 0$ 。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 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1})$ 应该离 0 不远。如果 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1})$ 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

$$J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1}) \xrightarrow{d} \chi^2(L - K) \quad (10.49)$$

其中, $(L - K)$ 为过度识别约束的个数,因为在估计 $\hat{\beta}_{\text{GMM}}$ 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 z_i 中,已知前 L_1 个工具变量 z_{i1} 满足正交性,而怀疑后 $(L - L_1)$ 个工具变量 z_{i2} 不满足正交性,即要检验原假设“ $H_0: E(z_{i2}\varepsilon_i) = 0$ ”。进行此检验的前提条件是 $L_1 \geq K$,保证即使仅用前 L_1 个工具变量 z_{i1} 进行估计,该模型也至少为恰好识别。如果 $L_1 \geq K$,则可分别用所有 L 个工具变量 z_i 或前 L_1 个工具变量 z_{i1} 进行 GMM 估计(假设仅使前 L_1 个工具变量 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 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“ $H_0: E(z_{i2}\varepsilon_i) = 0$ ”。可以证明^①,

$$C \equiv J - J_1 \xrightarrow{d} \chi^2(L - L_1) \quad (10.50)$$

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{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 z_{i2} 进行 GMM 估计时,最优权重矩阵 \hat{S}_{i1}^{-1} 应来自于使用 z_i 进行 GMM 估计时最优权重矩阵 \hat{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)。

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

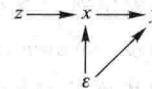


图 10.3 工具变量示意图

(箭头表示相关,无箭头

表示不相关)

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 $E[s(\theta_0; \mathbf{y})] = \mathbf{0}$, 其中 $s(\theta_0; \mathbf{y}) \equiv \frac{\partial \ln L(\theta_0; \mathbf{y})}{\partial \theta}$ 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 $s_i(\boldsymbol{\theta}_0; \mathbf{y}_i) = \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y}_i)}{\partial \boldsymbol{\theta}}$ 也满足这个性质, 即

$$\mathbb{E}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \mathbf{0} \quad (10.51)$$

将上式视为总体矩条件, 则其对应的样本矩条件为

$$\frac{1}{n} \sum_{i=1}^n s_i(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \mathbf{0} \quad (10.52)$$

去掉上式中的 $\frac{1}{n}$, 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“ $\mathbb{E}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \mathbf{0}$ ”正确, 则 GMM 估计量(也就是 MLE)就是一致的。显然, 矩条件“ $\mathbb{E}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \mathbf{0}$ ”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“ $\mathbb{E}[s_i(\boldsymbol{\theta}_0; \mathbf{y}_i)] = \mathbf{0}$ ”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

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma^2}\right\} \quad (10.53)$$

其对数似然函数及偏导数为

$$\ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_i) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma^2} \quad (10.54)$$

$$s_i(\boldsymbol{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}_i)}{\partial \boldsymbol{\beta}} = \frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \mathbf{x}_i}{\sigma^2} \quad (10.55)$$

$$\mathbb{E}[s_i(\boldsymbol{\beta})] = \frac{\mathbb{E}[(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \mathbf{x}_i]}{\sigma^2} = \mathbf{0} \quad (10.56)$$

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
<hr/>					
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
<hr/>					
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
<hr/>					
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						
	Robust					
	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						
	Robust					
	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
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lw						
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 0.0000),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 Centered R2 = 0.2002 Uncentered R2 = 0.9955 Root MSE = .3834						
Total (centered) SS = 139.2861498 Total (uncentered) SS = 24652.24662 Residual SS = 111.39959						
<hr/>						
lw	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
iq	-0.0115468	.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
<hr/>						
Underidentification test (Kleibergen-Paap rk LM statistic): 33.294 Chi-sq(4) P-val = 0.0000						
<hr/>						
Weak identification test (Cragg-Donald 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						
<hr/>						
Source: Stock-Yogo (2005). Reproduced by permission. NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.						
<hr/>						
Hansen J statistic (overidentification test of all instruments): 51.545 Chi-sq(3) P-val = 0.0000						
<hr/>						
-orthog- option: Hansen J statistic (eqn. excluding suspect orthog. conditions): 0.116 Chi-sq(1) P-val = 0.7333						
<hr/>						
C statistic (exogeneity/orthogonality of suspect instruments): 51.429 Chi-sq(2) P-val = 0.0000						
<hr/>						
Instruments tested: mrt age						
<hr/>						
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

iq	Robust					
	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

lw	Robust					
	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 force nonrobust
```

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)				
2SLS Size of nominal 5% Wald test	10%	15%	20%	25%	
LIML Size of nominal 5% Wald test	19.93	11.59	8.75	7.25	
	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
Total (centered) SS	=	139.2861498		F(6, 751) =	61.10	
Total (uncentered) SS	=	24652.24662		Prob > F =	0.0000	
Residual SS	=	100.6291971		Centered R2 =	0.2775	
				Uncentered R2 =	0.9959	
				Root MSE =	.3644	
lw	Coef.	Robust 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
<u>Underidentification test (Kleibergen-Paap rk LM statistic):</u>						24.223
				Chi-sq(2) P-val =		0.0000
<u>-redundant- option:</u>						
<u>IV redundancy test (LM test of redundancy of specified instruments):</u>						22.222
				Chi-sq(1) P-val =		0.0000
<u>Instruments tested: kww</u>						
<u>Weak identification test (Cragg-Donald Wald F statistic):</u>						14.906
						(Kleibergen-Paap rk Wald F statistic): 13.403
<u>Stock-Yogo weak ID test critical values: 10% maximal IV size</u>						19.93
				15% maximal IV size		11.59
				20% maximal IV size		8.75
				25% maximal IV size		7.25
<u>Source: Stock-Yogo (2005). Reproduced by permission.</u>						
<u>NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.</u>						
<u>Hansen J statistic (overidentification test of all instruments):</u>						0.151
				Chi-sq(1) P-val =		0.6972
<u>Instrumented: iq</u>						
<u>Included instruments: s expr tenure rns smsa</u>						
<u>Excluded instruments: med kww</u>						

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						
Total (centered) SS	=	139.2861498			Number of obs =	758
Total (uncentered) SS	=	24652.24662			F(6, 751) =	61.94
Residual SS	=	100.6291971			Prob > F =	0.0000
					Centered R2 =	0.2775
					Uncentered R2 =	0.9959
					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):						
					Chi-sq(2)	P-val = 0.0000
Weak identification test (Cragg-Donald Wald F statistic):						
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):						
					Chi-sq(1)	P-val = 0.130
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-B) Difference	sqrt(diag(V_b-V_B)) S.E.
	(b) iv	(B) ols		
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 H_0 and H_a ; obtained from ivregress
 B = inconsistent under H_a , efficient under H_0 ; obtained from regress

Test: H_0 : difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(1) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 3.84 \\ \text{Prob>chi2} &= 0.0499 \\ (V_b-V_B) &\text{ is not positive definite} \end{aligned}$$

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 endogenous^①

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																																																						
Total (centered) SS	=	139.2861498			Number of obs =	758																																																
Total (uncentered) SS	=	24652.24662			F(6, 751) =	61.10																																																
Residual SS	=	100.6291971			Prob > F =	0.0000																																																
					Centered R2 =	0.2775																																																
					Uncentered R2 =	0.9959																																																
					Root MSE =	.3644																																																
<hr/>																																																						
<table border="1"> <thead> <tr> <th>lw</th><th>Coef.</th><th>Robust Std. Err.</th><th>z</th><th>P> z </th><th>[95% Conf. Interval]</th></tr> </thead> <tbody> <tr> <td>iq</td><td>.0139284</td><td>.0060393</td><td>2.31</td><td>0.021</td><td>.0020916 .0257653</td></tr> <tr> <td>s</td><td>.0607803</td><td>.0189505</td><td>3.21</td><td>0.001</td><td>.023638 .0979227</td></tr> <tr> <td>expr</td><td>.0433237</td><td>.0074118</td><td>5.85</td><td>0.000</td><td>.0287968 .0578505</td></tr> <tr> <td>tenure</td><td>.0296442</td><td>.008317</td><td>3.56</td><td>0.000</td><td>.0133432 .0459452</td></tr> <tr> <td>rns</td><td>-.0435271</td><td>.0344779</td><td>-1.26</td><td>0.207</td><td>-.1111026 .0240483</td></tr> <tr> <td>smsa</td><td>.1272224</td><td>.0297414</td><td>4.28</td><td>0.000</td><td>.0689303 .1855146</td></tr> <tr> <td>cons</td><td>3.218043</td><td>.3983683</td><td>8.08</td><td>0.000</td><td>2.437256 3.998831</td></tr> </tbody> </table>							lw	Coef.	Robust 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
lw	Coef.	Robust 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																																																	
<hr/>																																																						
Underidentification test (Kleibergen-Paap rk LM statistic):																																																						
Chi-sq(2) P-val =																																																						
<hr/>																																																						
Weak identification test (Cragg-Donald Wald F statistic):																																																						
(Kleibergen-Paap rk Wald F statistic):																																																						
Stock-Yogo weak ID test critical values:																																																						
10% maximal IV size																																																						
15% maximal IV size																																																						
20% maximal IV size																																																						
25% maximal IV size																																																						
<hr/>																																																						
Source: Stock-Yogo (2005). Reproduced by permission.																																																						
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.																																																						
<hr/>																																																						
Hansen J statistic (overidentification test of all instruments):																																																						
Chi-sq(1) P-val =																																																						
<hr/>																																																						
-endog- option:																																																						
Endogeneity test of endogenous regressors:																																																						
Chi-sq(1) P-val =																																																						
<hr/>																																																						
Regressors tested: iq																																																						
<hr/>																																																						
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
						Root MSE = .36499
GMM weight matrix: Robust						
lw	Robust					
	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					
	Root MSE = .36499					
GMM weight matrix: Robust						
lw	Robust					
	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_iq	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***
	.0381189***	.03934425***	.04332367***	.04334159***	.04311171***	.04311006***
expr	.03561456***	.03420896***	.02964421***	.02962365***	.02997643***	.02997521***
	-.08407974***	-.07453249**	-.04352713	-.04338751	-.04451599	-.04451145
tenure						
	.13966664***	.13673691***	.12722244***	.1271796***	.12673682***	.12673991***
rns						
	.00327916***	.01392844**	.01397639**	.01408883**	.01409011**	
smsa						
	.00113212***	.00603931***	.00606812***	.00603567***	.00603575***	
iq						
_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 mttitle 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R²(如果使用选择项“pr2”则显示准R²),选择项“mttitle”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*** (0.00621)	0.0928*** (0.00698)	0.0608*** (0.0190)	0.0606*** (0.0190)	0.0604*** (0.0190)	0.0604*** (0.0190)
expr	0.0381*** (0.00661)	0.0393*** (0.00666)	0.0433*** (0.00741)	0.0433*** (0.00742)	0.0431*** (0.00741)	0.0431*** (0.00741)
tenure	0.0356*** (0.00800)	0.0342*** (0.00790)	0.0296*** (0.00832)	0.0296*** (0.00832)	0.0300*** (0.00827)	0.0300*** (0.00827)
rns	-0.0841*** (0.0295)	-0.0745** (0.0300)	-0.0435 (0.0345)	-0.0434 (0.0345)	-0.0445 (0.0344)	-0.0445 (0.0344)
smsa	0.140*** (0.0281)	0.137*** (0.0278)	0.127*** (0.0297)	0.127*** (0.0298)	0.127*** (0.0298)	0.127*** (0.0298)
iq		0.00328*** (0.00113)	0.0139** (0.00604)	-0.0140** (0.00607)	0.0141** (0.00604)	0.0141** (0.00604)
_cons	4.104*** (0.0877)	3.895*** (0.116)	3.218*** (0.398)	3.215*** (0.400)	3.207*** (0.398)	3.207*** (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
mttitle 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 $\hat{\beta}_{IV}$ 还是 β 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

$$\begin{cases} C_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases} \quad (10.57)$$

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 $\hat{\alpha}_{1,IV} = \bar{C}/\bar{Y}$,其中 $\bar{C} = \frac{1}{n} \sum_{t=1}^n C_t$, $\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$ (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 $y_{n \times 1} = X_{n \times K} \beta_{K \times 1} + \varepsilon_{n \times 1}$,而 $Z_{n \times L}$ 为由工具变量构成的矩阵(每列都是一个工具变量), $L \geq K$ 。将原模型两边同时左乘 Z' 可得, $Z'y = Z'X\beta + Z'\varepsilon$ 。记 $S = \text{Var}(Z'\varepsilon)$,而 \hat{S} 为 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 $\text{Cov}(\hat{\beta}_{IV}, \hat{\beta}_{OLS}) = \text{Var}(\hat{\beta}_{OLS})$ 。

证明:为了简单起见,只考虑一维的情形,即 $\hat{\beta}_{IV}$ 与 $\hat{\beta}_{OLS}$ 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 $\hat{q} = \hat{\beta}_{IV} - \hat{\beta}_{OLS}$ 。由于 $\text{Cov}(\hat{\beta}_{IV}, \hat{\beta}_{OLS}) = \text{Cov}(\hat{\beta}_{OLS} + \hat{q}, \hat{\beta}_{OLS}) = \text{Var}(\hat{\beta}_{OLS}) + \text{Cov}(\hat{q}, \hat{\beta}_{OLS})$,故只需证明 $\text{Cov}(\hat{q}, \hat{\beta}_{OLS}) = 0$ 即可。

其次,在 H_0 成立的情况下, $\hat{\beta}_{IV}$ 与 $\hat{\beta}_{OLS}$ 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

$$\text{Var}(\hat{\beta}) = \text{Var}(\hat{\beta}_{OLS}) + \lambda^2 \text{Var}(\hat{q}) + 2\lambda \text{Cov}(\hat{q}, \hat{\beta}_{OLS}) \geq \text{Var}(\hat{\beta}_{OLS}) \quad (10.58)$$

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,

$$\lambda^2 \text{Var}(\hat{q}) + 2\lambda \text{Cov}(\hat{q}, \hat{\beta}_{OLS}) \geq 0 \quad (10.59)$$

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 $\lambda = -\frac{\text{Cov}(\hat{q}, \hat{\beta}_{OLS})}{\text{Var}(\hat{q})} < 0$, 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 $\lambda = -\frac{\text{Cov}(\hat{q}, \hat{\beta}_{OLS})}{\text{Var}(\hat{q})} > 0$, 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 $\hat{\beta}_{GMM}(\hat{W}) = (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} S_{ZY}$

证明: GMM 估计量的最小化目标函数为

$$\begin{aligned} J(\hat{\beta}, \hat{W}) &= n(S_{ZY} - S_{ZX}\hat{\beta})' \hat{W} (S_{ZY} - S_{ZX}\hat{\beta}) = n(S'_{ZY} - \hat{\beta}' S'_{ZX}) \hat{W} (S_{ZY} - S_{ZX}\hat{\beta}) \\ &= n(S'_{ZY} \hat{W} - \hat{\beta}' S'_{ZX} \hat{W}) (S_{ZY} - S_{ZX}\hat{\beta}) \\ &= n(S'_{ZY} \hat{W} S_{ZY} - \hat{\beta}' S'_{ZX} \hat{W} S_{ZY} - S'_{ZY} \hat{W} S_{ZX}\hat{\beta} + \hat{\beta}' S'_{ZX} \hat{W} S_{ZX}\hat{\beta}) \\ &= n(S'_{ZY} \hat{W} S_{ZY} - 2\hat{\beta}' S'_{ZX} \hat{W} S_{ZY} + \hat{\beta}' S'_{ZX} \hat{W} S_{ZX}\hat{\beta}) \end{aligned} \quad (10.60)$$

其中, $(\hat{\beta}' S'_{ZX} \hat{W} S_{ZY})' = S'_{ZY} \hat{W} S_{ZX}\hat{\beta}$, 且都是一维标量(因为目标函数是一维标量), 故 $\hat{\beta}' S'_{ZX} \hat{W} S_{ZY} = S'_{ZY} \hat{W} S_{ZX}\hat{\beta}$, 可以合并在一起。

使用向量微分法则(参见第3章), 对 $\hat{\beta}$ 求导, 可以得到最小化的一阶条件,

$$\frac{\partial J(\hat{\beta}, \hat{W})}{\partial \hat{\beta}} = n(-2S'_{ZX} \hat{W} S_{ZY} + 2S'_{ZX} \hat{W} S_{ZX}\hat{\beta}) = 0 \quad (10.61)$$

方程两边同时消去 $2n$, 移项可得,

$$S'_{ZX} \hat{W} S_{ZX}\hat{\beta} = S'_{ZX} \hat{W} S_{ZY} \quad (10.62)$$

如果 $(S'_{ZX} \hat{W} S_{ZX})^{-1}$ 存在, 则 $\hat{\beta}_{GMM}(\hat{W}) = (S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W} S_{ZY}$, 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 \mathbf{x} 中。

最简单的模型为“线性概率模型”(Linear Probability Model,简记 LPM):

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n) \quad (11.1)$$

对 $\boldsymbol{\beta}$ 的一致估计要求 $\text{Cov}(\mathbf{x}_i, \varepsilon_i) = 0$ 。然而,由于 $\varepsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$,故 $\varepsilon_i = 1 - \mathbf{x}'_i \boldsymbol{\beta}$ 或 $\varepsilon_i = -\mathbf{x}'_i \boldsymbol{\beta}$,因此 ε_i 必然与 \mathbf{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 $\text{Var}(\varepsilon_i) = \text{Var}(\mathbf{x}'_i \boldsymbol{\beta})$,故扰动项 ε_i 的方差依赖于 \mathbf{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 \mathbf{x} 的情况下,考虑 y 的两点分布概率:

$$\begin{cases} P(y=1|x) = F(x, \beta) \\ P(y=0|x) = 1 - F(x, \beta) \end{cases} \quad (11.2)$$

此函数 $F(x, \beta)$ 也称为“连接函数”(link function), 因为它将解释变量 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 $F(x, \beta)$ (比如, 某随机变量的累积分布函数), 可以保证 $0 \leq \hat{y} \leq 1$, 并将 \hat{y} 理解为“ $y=1$ ”发生的概率, 因为

$$E(y|x) = 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x) = P(y=1|x) \quad (11.3)$$

如果 $F(x, \beta)$ 为标准正态的累积分布函数(cdf), 则

$$P(y=1|x) = F(x, \beta) = \Phi(x'\beta) \equiv \int_{-\infty}^{x'\beta} \phi(t) dt \quad (11.4)$$

该模型被称为“Probit”。如果 $F(x, \beta)$ 为“逻辑分布”(logistic distribution)的累积分布函数, 则

$$P(y=1|x) = F(x, \beta) = \Lambda(x'\beta) \equiv \frac{\exp(x'\beta)}{1 + \exp(x'\beta)} \quad (11.5)$$

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

$$f(y_i|x_i, \beta) = \begin{cases} \Lambda(x'_i\beta), & \text{若 } y_i = 1 \\ 1 - \Lambda(x'_i\beta), & \text{若 } y_i = 0 \end{cases} \quad (11.6)$$

将其更紧凑地写为

$$f(y_i|x_i, \beta) = [\Lambda(x'_i\beta)]^{y_i} [1 - \Lambda(x'_i\beta)]^{1-y_i} \quad (11.7)$$

取对数可得

$$\ln f(y_i|x_i, \beta) = y_i \ln[\Lambda(x'_i\beta)] + (1 - y_i) \ln[1 - \Lambda(x'_i\beta)] \quad (11.8)$$

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

$$\ln L(\beta|y, x) = \sum_{i=1}^n y_i \ln[\Lambda(x'_i\beta)] + \sum_{i=1}^n (1 - y_i) \ln[1 - \Lambda(x'_i\beta)] \quad (11.9)$$

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 $\hat{\beta}_{MLE}$ 并非边际效应(marginal effects)。以 Probit 为例,

$$\frac{\partial P(y=1|x)}{\partial x_k} = \frac{\partial P(y=1|x)}{\partial (x'\beta)} \cdot \frac{\partial (x'\beta)}{\partial x_k} = \phi(x'\beta) \cdot \beta_k \quad (11.10)$$

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

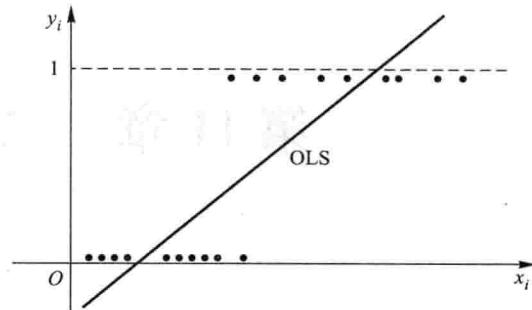


图 11.1 OLS 与二值选择模型

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 $x = \bar{x}$ 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 x^* , 在 $x = x^*$ 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 $\hat{\beta}_{MLE}$ 并非边际效应, 那么它究竟有什么含义呢? 对于 Logit 模型, 记 $p \equiv P(y=1|x)$, 则 $1-p = P(y=0|x)$ 。由于 $p = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$, $1-p = \frac{1}{1+\exp(x'\beta)}$, 故

$$\frac{p}{1-p} = \exp(x'\beta) \quad (11.11)$$

$$\ln\left(\frac{p}{1-p}\right) = x'\beta \quad (11.12)$$

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)

$$\frac{\frac{p^*}{1-p^*}}{\frac{p}{1-p}} = \frac{\exp[\beta_1 + \beta_2 x_2 + \cdots + \beta_j(x_j+1) + \cdots + \beta_K x_K]}{\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_K x_K)} = \exp(\hat{\beta}_j) \quad (11.13)$$

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 $\hat{\beta}_{MLE}$ 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

$$\text{准 } R^2 = \frac{\ln L_0 - \ln L_1}{\ln L_0} \quad (11.14)$$

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, $0 \geq \ln L_1 \geq \ln L_0$, 而 $0 \leq \text{准 } R^2 \leq 1$, 参见图 11.2。由于“准 R^2 ”可以写为 $\frac{\ln L_1 - \ln L_0}{\ln L_{\max} - \ln L_0}$, 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。

判断拟合优度的另一方法是计算“正确预测的百分比”(percent correctly predicted)。如果发生概率的预测值 $\hat{y} \geq 0.5$, 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 $E(y|x) = F(x, \beta)$ 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 $E(y|x) = F(x, \beta)$ 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 $E(y|x) \neq F(x, \beta)$), 则 Probit 与 Logit 模型并不能得到对系数 β 的一致估计, 使用稳健标准误也就没有太大的意义(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 \hat{y} , 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 $x_1 = 0$ 处的边际效应)

`margins, dydx(x1)` (计算解释变量 x_1 的平均边际效应)

`margins, eyex(*)` (计算平均弹性, 其中两个“e”均指 elasticity)

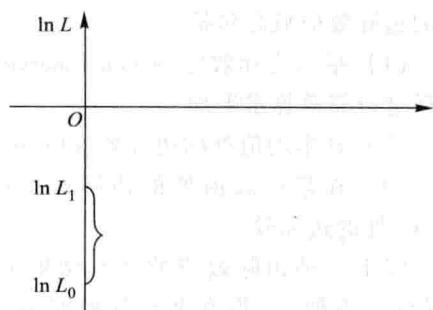


图 11.2 准 R^2 的计算

^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:

$$\text{work}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{married}_i + \beta_3 \text{children}_i + \beta_4 \text{education}_i + \varepsilon_i \quad (11.15)$$

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						
	Coef.	Robust 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						
	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				
	Delta-method					
	dy/dx	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)					
<hr/>									
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						
<hr/>									
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					
Classified	True				
	D	~D	Total		
+	1177	361	1538		
-	166	296	462		
Total	1343	657	2000		
<hr/>					
Classified + if predicted Pr(D) >= .5					
True D defined as work != 0					
<hr/>					
Sensitivity	Pr(+ D)	87.64%			
Specificity	Pr(- ~D)	45.05%			
Positive predictive value	Pr(D +)	76.53%			
Negative predictive value	Pr(~D -)	64.07%			
<hr/>					
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%			
<hr/>					
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				
						Pseudo R2	=	0.1882				
Log pseudolikelihood = -1027.9144												
(Std. Err. adjusted for 40 clusters in age)												
work	Coef.	Robust 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
						Pseudo R2	=	0.1889
Log likelihood = -1027.0616								
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			
Classified	True		Total
	D	~D	
+	1177	361	1538
-	166	296	462
Total	1343	657	2000

Classified + if predicted $\Pr(D) \geq .5$
 True D defined as work != 0

Sensitivity	$\Pr(+ D)$	87.64%
Specificity	$\Pr(- \sim D)$	45.05%
Positive predictive value	$\Pr(D +)$	76.53%
Negative predictive value	$\Pr(\sim D -)$	64.07%
False + rate for true ~D	$\Pr(+ \sim D)$	54.95%
False - rate for true D	$\Pr(- D)$	12.36%
False + rate for classified +	$\Pr(\sim D +)$	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

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \quad (11.16)$$

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

$$y = \begin{cases} 1, & \text{若 } y^* > 0 \\ 0, & \text{若 } y^* \leq 0 \end{cases} \quad (11.17)$$

因此,

$$\Pr(y=1 | \mathbf{x}) = \Pr(y^* > 0 | \mathbf{x}) = \Pr(\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 | \mathbf{x}) = \Pr(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} | \mathbf{x}) \quad (11.18)$$

假设 $\varepsilon \sim N(0, \sigma^2)$ 或服从逻辑分布, 则

$$\Pr(y=1 | \mathbf{x}) = \Pr(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} | \mathbf{x}) = \Pr(\varepsilon < \mathbf{x}'\boldsymbol{\beta}) = F_\varepsilon(\mathbf{x}'\boldsymbol{\beta}) \quad (11.19)$$

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 $k > 0$, $\Pr(\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0) = \Pr(k\mathbf{x}'\boldsymbol{\beta} + k\varepsilon > 0)$ 。记扰动项方差 $\sigma^2 \equiv \text{Var}(\varepsilon)$, 则 $\text{Var}(k\varepsilon) = k^2\sigma^2$, 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 $U_a \leq U_b$, 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 $U_a = \mathbf{x}'\boldsymbol{\beta}_a + \varepsilon_a$, $U_b = \mathbf{x}'\boldsymbol{\beta}_b + \varepsilon_b$, 则

$$\begin{aligned} P(y = 1 | \mathbf{x}) &= P(U_a > U_b | \mathbf{x}) \\ &= P(\mathbf{x}'\boldsymbol{\beta}_a + \varepsilon_a > \mathbf{x}'\boldsymbol{\beta}_b + \varepsilon_b | \mathbf{x}) \\ &= P[\mathbf{x}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + (\varepsilon_a - \varepsilon_b) > 0 | \mathbf{x}] \end{aligned} \quad (11.20)$$

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 $\text{Var}(\varepsilon_a - \varepsilon_b)$ 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 $F(\varepsilon) = \exp\{-e^{-\varepsilon}\}$, 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

$$P(y_i = 1 | \mathbf{x}_i) = \Phi(\mathbf{x}'_i \boldsymbol{\beta} / \sigma) \quad (11.21)$$

其中, 扰动项的标准差 $\sigma = 1$ 。而“异方差”的替代假设 H_1 为

$$P(y_i = 1 | \mathbf{x}_i) = \Phi(\mathbf{x}'_i \boldsymbol{\beta} / \sigma_i) \quad (11.22)$$

其中, $\sigma_i^2 \equiv \text{Var}(\varepsilon_i)$ 。进一步, 假设 σ_i^2 依赖于外生变量 $\mathbf{z} \equiv (z_1, \dots, z_m)$

$$\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\delta}) \quad (11.23)$$

其中, \mathbf{z} 可以与解释变量 \mathbf{x} 有重叠部分, 或者包括 \mathbf{x} , 但不包括常数项^②。对方程 (11.23) 两边取对数可得

$$\ln \sigma_i^2 = \mathbf{z}'_i \boldsymbol{\delta} \quad (11.24)$$

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 $\mathbf{z} \equiv (z_1, \dots, z_m)$ 。在 Stata 的输出结果中, 将汇报对 $H_0: \boldsymbol{\delta} = \mathbf{0}$ 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 $H_0: \boldsymbol{\delta} = \mathbf{0}$, 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 $\{x_1, \dots, x_n\}$, 则其最大值 $\max|x_1, \dots, x_n|$ 可视为一个顺序统计量 (order statistic)。当 $n \rightarrow \infty$ 时, $\max|x_1, \dots, x_n|$ 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
								Prob > chi2 = 0.0123
Log likelihood = -1026.19								
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:

$$P(y=1|x) = \Lambda(\beta_0 + \beta_1 x) \equiv \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad (11.25)$$

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。

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 $x|y=0$ 的密度函数, 而实线表示条件分布 $x|y=1$ 的密度函数。由于样本中

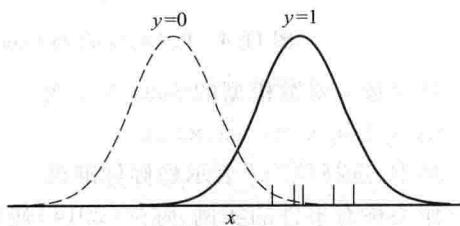


图 11.3 稀有事件偏差示意图

有很多“ $y = 0$ ”的观测数据(未在图中显示),故很容易估计 $x | y = 0$ 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 $x | y = 1$ 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 $x | y = 0$ 的右尾(right tail)可以得到很好的估计,而 $x | y = 1$ 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 $\text{bias}(\hat{\beta})$ 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 $[\hat{\beta} - \text{bias}(\hat{\beta})]$ 。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

$$p = P(y = 1 | x) = F(x, \beta) = 1 - \exp\{-e^{x\beta}\} \quad (11.26)$$

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, $x'\beta = \ln[-\ln(1-p)]$,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 $x'\beta$ 。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。

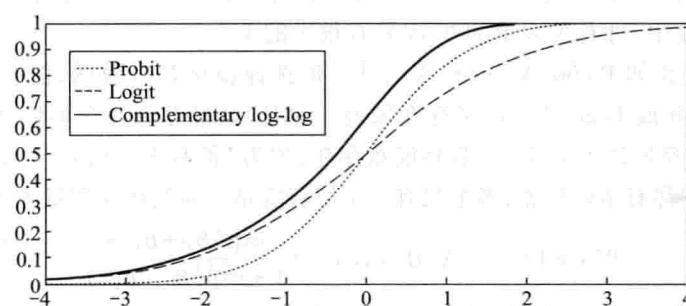


图 11.4 Probit, Logit 与 Complementary log-log 模型的累积分布函数比较

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			
		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					
	Delta-method	dy/dx	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						
conquered	Robust					Number of obs = 212
	Coef.	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						
Prob > chi2 = 0.0000						
Log pseudolikelihood = -21.549974						

conquered	Robust					[95% Conf. Interval]
	Coef.	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					[95% Conf. Interval]
	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,

$$y_{1i}^* = \mathbf{x}'_i \boldsymbol{\alpha} + \beta y_{2i} + u_i \quad (11.27)$$

$$y_{2i} = \mathbf{x}'_i \boldsymbol{\gamma}_1 + \mathbf{z}'_i \boldsymbol{\gamma}_2 + v_i \quad (11.28)$$

$$y_{1i} = \mathbf{1}(y_{1i}^* > 0) \quad (11.29)$$

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

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \sigma_v \\ \rho \sigma_v & \sigma_v^2 \end{pmatrix}\right) \quad (11.30)$$

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 \mathbf{x}_i 与 \mathbf{z}_i ,故在方程(11.27)中, \mathbf{x}_i 为外生解释变量。而且, \mathbf{z}_i 可作为方程(11.27)中内生变量 y_{2i} 的工具变量,因为 \mathbf{z}_i 与内生变量 y_{2i} 相关(参见方程(11.28)),且 \mathbf{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 $\rho=0$,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 \mathbf{x}_i 与 \mathbf{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 $f(y_{1i}, y_{2i} | \mathbf{x}_i, \mathbf{z}_i)$ 分解为 $f(y_{1i} | y_{2i}, \mathbf{x}_i, \mathbf{z}_i) f(y_{2i} | \mathbf{x}_i, \mathbf{z}_i)$,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

$$u_i = \delta v_i + \varepsilon_i \quad (11.31)$$

其中, $\delta \equiv \frac{\text{Cov}(u_i, v_i)}{\text{Var}(v_i)}$,而扰动项 ε_i 独立于 v_i (故也独立于 y_{2i}),也独立于 \mathbf{x}_i 。而且, ε_i 也服从正态

分布,期望为 $E(\varepsilon_i) = 0$,而方差为

$$\begin{aligned} \text{Var}(\varepsilon_i) &= \text{Var}(u_i - \delta v_i) = 1 - \frac{\text{Cov}^2(u_i, v_i)}{\text{Var}^2(v_i)} \cdot \text{Var}(v_i) \\ &= 1 - \frac{\text{Cov}^2(u_i, v_i)}{\text{Var}(v_i)} = 1 - \rho^2 \end{aligned} \quad (11.32)$$

其中, $\text{Var}(u_i) = 1$ 。将方程(11.31)代入方程(11.27)可得

$$y_{1i}^* = \mathbf{x}'_i \boldsymbol{\alpha} + \beta y_{2i} + \delta v_i + \varepsilon_i \quad (11.33)$$

其中, $\varepsilon_i \sim N(0, 1 - \rho^2)$, 而且独立于 \mathbf{x}_i , y_{2i} 与 v_i 。为了把 ε_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

$$\frac{y_{1i}^*}{\sqrt{1 - \rho^2}} = \mathbf{x}'_i \frac{\boldsymbol{\alpha}}{\sqrt{1 - \rho^2}} + \frac{\beta}{\sqrt{1 - \rho^2}} y_{2i} + \frac{\delta}{\sqrt{1 - \rho^2}} v_i + \frac{\varepsilon_i}{\sqrt{1 - \rho^2}} \quad (11.34)$$

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 $0 < \sqrt{1 - \rho^2} \leq 1$, 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)。