高等学校经济学类核心课程教材

# 计量经济学及Stata应用

陈 强 编著

高等教育出版社

高等学校经济学类核心课程教材及配套资源

| | |
|---|---|
| □ 政治经济学（第五版）（送教师课件） | 逄锦聚　等 |
| □ 政治经济学（第五版）（送教师课件） | 李建平　等 |
| □ 政治经济学（第四版）（送教师课件） | 谢　地　等 |
| □ 现代政治经济学（送教师课件） | 白永秀　等 |
| □ 政治经济学热点难点争鸣 | 逄锦聚 |
| □ 西方经济学（第三版）（送教师课件） | 厉以宁 |
| □ 微观经济学（西方经济学·第四版）（送教师课件） | 黎诣远 |
| □ 宏观经济学（西方经济学·第四版）（送教师课件） | 黎诣远 |
| □ 西方经济学（第四版）（送教师课件） | 许纯祯　等 |
| □ 西方经济学学习辅导书 | 黎诣远　等 |
| □ 西方经济学课程题解（附答案与练习盘） | 周加来 |
| □ 国际经济学（送教师课件） | 海　闻 |
| □ 国际经济学（第三版）（送教师课件） | 李坤望 |
| □ 国际经济学学习与习题指南 | 张伯伟 |
| □ 金融学（第四版）（送教师课件） | 曹龙骐 |
| □ 金融学案例与分析 | 曹龙骐 |
| □ 计量经济学（第四版）（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　等 |
| ■ 计量经济学及Stata应用（送教师课件） | 陈　强 |
| □ 财政学（第四版）（送教师课件） | 邓子基　等 |
| □ 财政学（第三版）（送教师课件） | 储敏伟　等 |
| □ 财政学（第二版）（送教师课件） | 王国清　等 |
| □ 财政学 | 钟晓敏 |
| □ 财政学原理（送教师课件） | 刘京焕　等 |
| □ 财政学案例 | 刘京焕　等 |
| □ 会计学（第二版）（上、下册）（配学习卡、送教师课件） | 葛家澍　等 |
| □ 会计学（第二版）学习指导书 | 葛家澍　等 |
| □ 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 |

ISBN 978-7-04-042
9 787040 427516
定价 39.00元

高等学校经济学类核心课程教材

# 计量经济学及Stata应用

陈 强 编著

Jiliang Jingjixue ji Stata Yingyong

高等教育出版社·北京

内容简介

本书为既接轨现代计量经济学，又适合中国国情的本科计量经济学教材。在理论体系上，本书充分借鉴最新国际主流教材，以大样本理论为主线，并针对中国学生的知识体系进行编写。本书内容全面，包括横截面数据（多元回归、工具变量法、离散选择）、时间序列（平稳时间序列、单位根、协整），以及面板数据（随机效应、固定效应）等。

本书力图以清晰而生动的语言、较多的插图与经济意义，来直观地解释计量方法。同时结合目前欧美最为流行的 Stata 计量软件，及时地介绍相应的计算机操作与经典实例，为读者提供“一站式”服务。本书还较多地使用计算机模拟（蒙特卡罗法），作为强有力的学习工具。

本书适合高等学校经济管理类及社科类的本科生使用。先修课为微积分、线性代数与概率统计。阅读本书可使读者掌握当代实证研究的精神实质与基本方法，并学会实际处理数据的重要技能，从而为毕业论文乃至读研深造打下良好基础。

**图书在版编目（CIP）数据**

计量经济学及 Stata 应用/陈强编著. --北京：高等教育出版社，2015.7

ISBN 978-7-04-042751-6

Ⅰ. ①计… Ⅱ. ①陈… Ⅲ. ①计量经济学-应用软件-高等学校-教材 Ⅳ. ①F224.0-39

中国版本图书馆 CIP 数据核字（2015）第 101233 号

| | | | |
|---|---|---|---|
| 策划编辑 施春花 | 责任编辑 施春花 | 封面设计 杨立新 | 版式设计 杜微言 |
| 插图绘制 于 博 | 责任校对 陈 杨 | 责任印制 毛斯璐 | |

| | | | |
|---|---|---|---|
| 出版发行 | 高等教育出版社 | 咨询电话 | 400-810-0598 |
| 社 址 | 北京市西城区德外大街 4 号 | 网 址 | http://www.hep.edu.cn |
| 邮政编码 | 100120 | | http://www.hep.com.cn |
| 印 刷 | 国防工业出版社印刷厂 | 网上订购 | http://www.landraco.com |
| 开 本 | 787mm×1092mm 1/16 | | http://www.landraco.com.cn |
| 印 张 | 22.5 | 版 次 | 2015 年 7 月第 1 版 |
| 字 数 | 550 千字 | 印 次 | 2015 年 7 月第 1 次印刷 |
| 购书热线 | 010-58581118 | 定 价 | 39.00 元 |




物 料 号 42751-00

# 前　言

自从1995年我以北大硕士生身份留校，并执教“统计学”与“计量经济学”课程以来，至今20年已匆匆过去了。期间，西方国家的计量经济学突飞猛进，国内的计量经济学教学也取得了长足进步。而我自己也赴美求学多年，并于2008年加盟山东大学经济学院，再次执教“计量经济学”（本科生、硕士生与博士生）课程。

在山东大学的最初几年，我一直使用Stock and Watson的*Introduction to Econometrics*作为本科计量经济学教材，并以Wooldridge的*Introductory Econometrics: A Modern Approach*作为辅助参考，这是目前国际上最为流行的两本本科计量经济学教材。这两本教材的作者均为国际计量经济学大师，分别执教于哈佛大学、普林斯顿大学与密歇根州立大学，因此他们的教材能很好地贴近当代计量经济学的主流方法，对于细节的交代也很清晰到位。

然而，在赞叹之余，也渐渐发现这两本国际畅销教材并不太适合中国的国情。首先，这两本书的英文版均超过800页，教师无法在一学期教完，而学生甚至无法通读一遍。由于计量经济学本身技术性较强，再加上英文的语言障碍，则是难上加难。虽然，这两本教材均有不同的中译本，但面对巨大的翻译工作量，译者们往往无法长时间去润色打磨，不仅难以达到“信达雅”的理想境界，甚至与我们所熟悉的汉语阅读习惯也大相径庭。其次，由于美国大学数学课开设较晚，而这两本教材主要面向美国读者，故在正文尽量回避数学，而把矩阵形式的计量模型放到最后的章节甚至附录，导致不得不用大量的语言来解释计量理论，致使教材篇幅过长。另外，中国学生在学习计量经济学之前，一般已学过微积分、线性代数与概率统计，如果不适当地运用这些数学，则是巨大的资源浪费与效率损失。

如何才能让中国学生更轻松地掌握计量经济学的当代知识？单纯地依赖国际教材及其中译本似乎并非捷径。时下，在中国的高等教育界，国际化是一股轰轰烈烈的潮流。许多人认为，国际化就是采用英文教材，进行双语甚至全英文教学。或许，这适用于一小部分学生，比如全英文实验班，但对于大多数中国学生而言，恐怕会出现水土不服，甚至事倍功半的效果。笔者以为，国际化的一个较高境界其实是本土化，即将国际知识洋为中用，以汉语的形式最快地让绝大部分中国学生直接受益。

为此，在编著畅销研究生教材《高级计量经济学及Stata应用》（高等教育出版社，2010年第1版，2014年第2版）的基础上，我决心为广大中国学生编写一本既通俗易懂、贴近主流，又极具操作性的本科计量经济学教材。本书的主要特色可概括如下：

（1）接轨现代计量经济学。本书较多地借鉴了Baum（2006），Dougherty（2011），Gujarati and Porter（2008），Hill *et al.*（2011），Kennedy（2008），Maddala and Lahiri（2009），Stock and Watson（2012），Studenmund（2010），Wooldridge（2009）等国际流行的本科教材，其中尤以Stock and Watson（2012）与Wooldridge（2009）对本书的影响最深。另外，还借鉴了一些经典的研究生教材，比如Cameron and Trivedi（2005，2010），Greene（2012），Hayashi（2000），Poirier（1995），

Verbeek(2004), Wooldridge (2010),以期高屋建瓴。

(2) 在内容上坚持不灌水。长期以来,出于教学上的权宜之计,国内本科教材常常做一些不现实的假设。比如,假定解释变量为固定、非随机的(fixed regressors),这固然使问题简化便于理解,但现实中的经济变量几乎都是随机的。假设解释变量非随机,便无从讨论解释变量与扰动项的相关性,给后续教学造成莫大障碍。又比如,目前国际计量学界的主流方法已是大样本理论,而多数国内教材仍侧重于小样本理论。小样本理论的严苛假设(比如,严格外生性、正态分布),使得它在现实中很难应用。为此,本书坚持在内容上不灌水,自始至终假设随机解释变量,并以大样本理论为核心,将当代计量的主流方法介绍给学生。试想,如果只给学生灌水的内容,却要求他们去看最新的文献,完成高质量的毕业论文,则无异于赶着未经良好训练的士兵上战场。"工欲善其事,必先利其器",故需将最好的计量工具介绍给学生。

(3) 理论与实践紧密结合。学习计量经济学的学生,既需要了解计量经济学原理,也需要知道如何在计算机上实现,掌握处理数据的实际技能。为此,本书提供一站式服务,在讲解每个估计方法后,随即介绍相应的 Stata 计算机操作(Stata 为目前欧美最为流行的计量软件),并深入分析有趣的经典实例,便于读者迅速掌握相应的理论与操作。本书还较多地使用计算机模拟(蒙特卡罗法),作为强有力的学习工具,直观地呈现计量理论与结果。

(4) 尽量使用清晰、通俗而生动的语言。在某些方面,写作计量经济学本科教材的难度甚至超过研究生教材,因为前者无法像后者那样自由地使用矩阵等数学工具。当然,一味地回避数学显然不可取,因为最精确的理解仍然依赖于数学表达式。另外,对于必不可少的数学公式,则应辅以清晰、通俗而生动的语言,乃至插图,加以直观地解释。为了清晰地表述某个理论或思想,有时需要多番修改提炼,以找到最为直指人心的表达方式。

在本书出版之际,特别感谢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同事与学生们对本书的写作给予了大力支持。常东风副教授、唐明哲副教授、王永副教授、韩青博士、孔建宁博士、薛欣欣博士、博士生张博,以及硕士生李昱璇、刘春雨、卢秋全、毛会贞、孟鸽、孙丰凯、徐艳娴等参加了本书的校对,并提出了很好的修改意见,在此表示衷心感谢(当然,文责自负)。最后,特别感谢高等教育出版社的施春花编辑及其同仁们,为保证本书的高质量,他们付出了辛勤的劳动。

当然,由于笔者学识有限,对于本书的错漏之处,恳请读者及时指出,以便在网上公布勘误表,并在未来的版本中更正。联系邮箱为 qiang2chen2@126.com。

本书用到的所有数据集以及勘误表,均可在我的个人网页下载:http://econ.sdu.edu.cn/tree/content.php?id=52841。

陈强

2015 年 3 月

# 目　录

*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.* ——H. G. Wells

*There are three kinds of lies: lies, damn lies, and statistics.* ——Benjamin Disraeli

# 1. 导　论

## 1.1 什么是计量经济学

"计量经济学"(econometrics[①],也译为"经济计量学"),顾名思义,是运用概率统计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,一方面,由于实验数据的缺乏,计量经济学常常不足以确定经济变量之间的因果关系;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 $X$ 是否导致 $Y$),而非仅仅是相关关系。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。

**例**(相关关系)　你看到街上的人们带雨伞,于是预测今天要下雨。但这只是一种相关关系,因为"人们带伞"并不是造成"下雨"的原因。

**例**(相关关系)　根据与流感相关的海量词条搜索记录,谷歌公司通过分析大数据(big data),可以很快地预测流行病的地域传播(参见本章附录)。但这也只是相关关系,因为上网搜索流感信息并不导致流感的传播。

由以上两例可知,如果我们只对预测感兴趣,则相关关系就足够了。然而,如果为了推断变量之间的因果关系,则计量分析必须建立在经济理论的基础之上,即在理论上存在 $X$ 导致 $Y$ 的作用机制。然而,即使有理论基础,因果关系常常依然不好分辨。首先,可能存在"逆向因果关系"(reverse causality)或"双向因果关系"。

**例**(逆向因果)　FDI(外商直接投资)促进经济增长,但 FDI 也被吸引到快速增长的地区。

**例**(逆向因果)　收入增加引起消费增长,而消费增长也拉动收入增加。

**例**(逆向因果)　经济萧条可能引起内战,但内战也会导致经济停滞。

其次,被遗漏的第三个变量($Z$)也可能对这两个变量($X$,$Y$)同时起作用,参见图 1.1。

**例**(遗漏变量)　某外星人来到地球,发现人类会死亡,十分不解。于是开始在全球广泛观察死亡现象,并收集了大量的数据。结果发现,许多人类躺在医院病床($X$)之后死去($Y$),故推断医院病床是死亡的原因。外星人认为,由于躺在医院病床上,总是发生在死亡之前,故不可能

---

① 其中,"econ"表示经济,"metrics"表示度量或测量,故"econometrics"的字面意思为"economic measurement",即"经济度量"。而计量经济学家则称为"econometrician"。

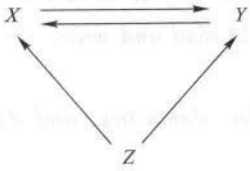



图 1.1 可能的因果关系

存在逆向因果关系。外星人于是将研究报告投稿发表于某顶尖经济学期刊,并在文末给出政策建议"珍爱生命,远离病床"①。在此例中,遗漏的变量 $Z$ 是什么?

**例**(遗漏变量) 考虑决定教育投资回报率(returns to schooling)的因素:

$$\ln w_i = \alpha + \beta s_i + \varepsilon_i \tag{1.1}$$

其中,$\ln w_i$(工资对数)为"被解释变量"(dependent variable)②。$s_i$(schooling,教育年限)为"解释变量"(explanatory variable,regressor)、"自变量"(independent variable)或"协变量"(covariate)。$\varepsilon_i$ 为不可观测(unobservable)的"误差项"(error term)或"随机扰动项"(stochastic disturbance),包括所有除 $s_i$ 以外对 $\ln w_i$ 有影响的因素,以及人类行为的随机性。下标 $i$ 表示第 $i$ 个观测值(即个体 $i$)。截距项 $\alpha$ 与斜率 $\beta$ 为待估参数。其中,$\beta$ 的经济含义为教育投资的回报率,即多上一年学,未来工资能增加百分之几(参见第 4 章)。

如果用数据估计一元回归方程(1.1),其结果一般会显示,对数工资与教育年限显著正相关,而且教育投资回报率 $\beta$ 还较高。然而,一个人的工资收入也与能力有关,但能力一般不能直接观测,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,教育的高回报其实包含了对能力的回报。

进一步,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,需要尽可能多地引入"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 $\beta$。然而,现实中总有某些相关的变量无法观测,即存在"遗漏变量"(omitted variables),而这些遗漏变量统统被纳入到随机扰动项 $\varepsilon_i$ 中了。

随机扰动项 $\varepsilon_i$ 中还可能包含哪些其他因素呢?如果真实模型(true model)为

$$\ln w_i = \alpha + \beta s_i + \gamma s_i^2 + \varepsilon_i \tag{1.2}$$

那么 $\gamma s_i^2$ 也被纳入到扰动项中了(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,一方面,扰动项就像是一个"垃圾桶",所有你不想要、无法把握的东西都往里面扔;另一方面,我们又希望扰动项拥有很好的性质。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"。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 此例来自香港大学商学院周文教授,参见 http://www2.fbe.hku.hk/staff/wzhou/。

② 之所以使用工资对数而不用工资作为被解释变量,是基于劳动经济学(labor economics)的理论模型,而且对工资对数建模也与经验数据更为吻合,参见第 4 章。

## 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),而是自然发生的“观测数据”(observational data);比如,统计局所收集的数据。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[①]。

在有些本科计量教材中,为了简单起见,有时假设解释变量是非随机的、固定的(fixed regressors)。这只是教学法上的权宜之计,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,在本书中,所有变量都是随机的(即使非随机的常数,也可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,2013 年中国各省(直辖市、自治区)的 GDP,参见表 1.1。

**表 1.1 2013 年中国分省(直辖市、自治区)GDP** 单位:亿元

| 省(直辖市、自治区) | GDP | 省(直辖市、自治区) | GDP |
|---|---|---|---|
| 北京 | 19 500.56 | 湖北 | 24 668.49 |
| 天津 | 14 370.16 | 湖南 | 24 501.67 |
| 河北 | 28 301.41 | 广东 | 62 163.97 |
| 山西 | 12 602.24 | 广西 | 14 378.00 |
| 内蒙古 | 16 832.38 | 海南 | 3 146.46 |
| 辽宁 | 27 077.65 | 重庆 | 12 656.69 |
| 吉林 | 12 981.46 | 四川 | 26 260.77 |
| 黑龙江 | 14 382.93 | 贵州 | 8 006.79 |
| 上海 | 21 602.12 | 云南 | 11 720.91 |
| 江苏 | 59 161.75 | 西藏 | 807.67 |
| 浙江 | 37 568.49 | 陕西 | 16 045.21 |
| 安徽 | 19 038.87 | 甘肃 | 6 268.01 |
| 福建 | 21 759.64 | 青海 | 2 101.05 |
| 江西 | 14 338.50 | 宁夏 | 2 565.06 |
| 山东 | 54 684.33 | 新疆 | 8 360.24 |
| 河南 | 32 155.86 | | |

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

(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。

① 你能举出哪些经济数据(变量)不是随机变量吗?

比如,1994—2013 年山东省每年的 GDP,参见表 1.2。

**表 1.2 1994—2013 年山东省 GDP** 单位:亿元

| 年份 | GDP | 年份 | GDP |
|---|---|---|---|
| 1994 | 3 844.5 | 2004 | 15 021.8 |
| 1995 | 4 953.35 | 2005 | 18 366.9 |
| 1996 | 5 883.8 | 2006 | 21 900.2 |
| 1997 | 6 537.07 | 2007 | 25 776.9 |
| 1998 | 7 021.35 | 2008 | 30 933.3 |
| 1999 | 7 493.84 | 2009 | 33 896.6 |
| 2000 | 8 337.47 | 2010 | 39 169.9 |
| 2001 | 9 195.04 | 2011 | 45 361.9 |
| 2002 | 10 275.5 | 2012 | 50 013.2 |
| 2003 | 12 078.2 | 2013 | 54 684.3 |

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

(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,1994—2013 年中国各省(直辖市、自治区)每年的 GDP,参见表 1.3。

**表 1.3 1994—2013 年中国分省(直辖市、自治区)GDP** 单位:亿元

| 省(直辖市、自治区) | 年份 | GDP |
|---|---|---|
| 北京 | 1994 | 1 145.31 |
| 北京 | 1995 | 1 507.69 |
| ⋮ | ⋮ | ⋮ |
| 北京 | 2012 | 17 879.4 |
| 北京 | 2013 | 19 500.56 |
| 天津 | 1994 | 732.89 |
| 天津 | 1995 | 931.97 |
| ⋮ | ⋮ | ⋮ |
| 天津 | 2012 | 12 893.88 |
| 天津 | 2013 | 14 370.16 |
| ⋮ | ⋮ | ⋮ |
| 新疆 | 1994 | 662.32 |
| 新疆 | 1995 | 814.85 |
| ⋮ | ⋮ | ⋮ |
| 新疆 | 2012 | 7 505.31 |
| 新疆 | 2013 | 8 360.24 |

资料来源:国家统计局网站. http://data.stats.gov.cn/workspace/index?m=fsnd

本书介绍的计量经济理论包括以上三种数据类型，并使用国际上最为流行的 Stata 计量软件进行数据处理（Stata 13 版本，2013 年发布）。为此，我们将在第 2 章介绍 Stata 软件。第 3 章将回顾相关数学知识，并引入一些新概念（比如，均值独立、迭代期望定律）。有了这些铺垫之后，第 4 章将正式进入计量经济学的理论部分。

## 附录 A1.1 谷歌如何通过搜索记录预测流感的传播

2009 年 3 月底，一种新流感甲型 H1N1 流感（最初命名为“人感染猪流感”）在墨西哥和美国加利福尼亚州、得克萨斯州爆发，并在全球不断蔓延。这种新型病毒的基因中包含猪流感、禽流感和人流感三种流感病毒的基因片段。截至 2010 年 5 月底，出现疫情的国家和地区达到 214 个，持续一年多的疫情造成约 1.85 万人死亡。①

由于缺乏对抗这种新型流感病毒的疫苗，公共卫生专家所能做的只是减慢其传播速度，而这取决于知道这种流感出现在哪里。在美国，虽然要求医生在发现新型流感病例时告知疾控中心（Centers for Disease Control and Prevention），但人们可能患病多日才去医院，而此信息传到疾控中心也要时间，故通告新流感病例往往有一两周的延迟。然而，对于迅速传播的流行病，信息滞后两周可能带来致命的后果。

能否找到“预测”流行病的更快方法？迈尔－舍恩伯格与库克耶（2013，p. 2－4）在畅销书《大数据时代》介绍谷歌如何通过搜索记录来更快更准地预测流感的传播：

> 在甲型 H1N1 流感爆发的几周前，互联网巨头谷歌公司的工程师们在《自然》杂志上发表了一篇引人注目的论文。它令公共卫生官员们和计算机科学家感到震惊。文中解释了谷歌为什么能够预测冬季流感的传播：不仅是全美范围的传播，而且可以具体到特定的地区和州。谷歌通过观察人们在网上的搜索记录来完成这个预测，而这种方法以前一直是被忽略的。谷歌保存了多年来所有的搜索记录，而且每天都会收到来自全球 30 亿条的搜索指令，如此庞大的数据资源足以支撑和帮助它完成这项工作。
>
> 谷歌公司把 5 000 万条美国人最频繁检索的词条和美国疾控中心在 2003—2008 年间季节性流感传播时期的数据进行了比较。……他们设立的这个系统唯一关注的就是特定检索词条的使用频率与流感在时间和空间上的传播之间的联系。谷歌公司为了测试这些检索词条，总共处理了 4.5 亿个不同的数学模型。在将得出的预测与 2007 年、2008 年美国疾控中心记录的实际流感病例进行对比后，谷歌公司发现，他们的软件发现了 45 条检索词条的组合，将它们用于一个特定的数学模型后，他们的预测与官方数据的相关性高达 97%。和疾控中心一样，他们也能判断出流感是从哪里传播出来的，而且判断非常及时，不会像疾控中心那样要在流感爆发后一两周才可以做到。
>
> 所以，2009 年甲型 H1N1 流感爆发的时候，与习惯性滞后的官方数据相比，谷歌成为了一个更有效、更及时的指示标。公共卫生机构的官员获得了非常有价值的数据信息。

① 详见维基百科：http://zh.wikipedia.org/wiki/。

*Computers are useless. They can only give you answers.* ——Pablo Picasso

# 2. Stata 入门

## 2.1 为什么使用 Stata

Stata 软件因其操作简单且功能强大，成为目前在欧美最流行的统计与计量分析软件，拥有为数众多的用户。Stata 公司也通过定期升级软件，以适应计量经济学的迅猛发展。同时，Stata 软件留有“用户接口”，允许用户自己编写命令与函数，并上传到网上实现共享。因此，对于一些最新的计量方法，可以在线查找和下载由用户编写的 Stata 命令程序（user-written Stata commands）。这些“非官方命令”（也称“外部命令”）的使用方法与官方命令完全相同，使得 Stata 的功能如虎添翼，深受用户的喜爱。

本书使用 Stata 13 版本（2013 年 6 月发布）。对于绝大多数的命令与功能，即使你用更低的 Stata 版本（比如 Stata 11 或 Stata 12），也几乎没有差别。即使你没有任何基础，通常只需要半天时间，看完本章内容并亲自操作一遍，就可以实现 Stata 入门的要求（后续学习可随着本书而逐渐深入）。

## 2.2 Stata 的窗口

安装 Stata 13 后，在安装的文件夹中将出现如图 2.1 所示的 Stata 13 图标（Stata 11 或 Stata 12 的图标大同小异）。

图 2.1 Stata 13 的图标

双击此 Stata 图标，即可打开 Stata。如果想在计算机桌面创建一个开启 Stata软件的快捷方式，可以右键点击 Stata 13 的图标，然后选择“发送到”→“桌面快捷方式”，参见图 2.2。

打开 Stata 后可看到，在最上方有一排“下拉式菜单”（pull-down menu），参见图 2.3。

点击图 2.3 中的任何选项，都会弹出一个“级联式”菜单，每个选项之下还可能有子菜单。在 Stata 中运行单个命令主要有两种方式，其一为点击菜单，其二为在“命令窗口”输入命令（参见下文）。通过菜单执行命令（menu-driven）可能要点击多重菜单，通常最后还要填写一个对话框（dialog），以明确命令的参数，故一般不如在命令窗口直接输入命令更为方便有效。

在菜单之下，为一系列图标，起着快捷键的作用，参见图 2.4。只要将鼠标放在图 2.4 中的任何快捷键上，就会显示其相应的功能。这些快捷键的具体用法，将在下文逐步介绍。

在快捷键图标之下，有五个窗口，参见图 2.5。其中，左边为“历史命令窗口”（Review），记录启动 Stata 后用过的命令。中间的大窗口为“结果窗口”（Results），显示执行 Stata 命令后

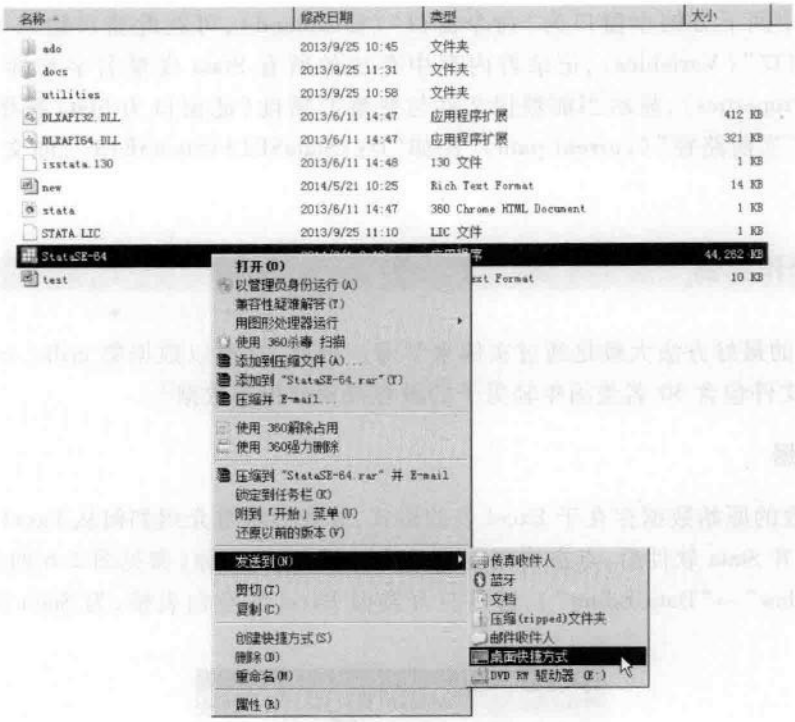



图 2.2 发送 Stata 13 到桌面快捷方式

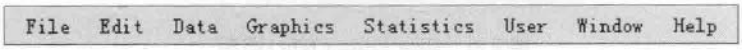



图 2.3 Stata 的下拉式菜单

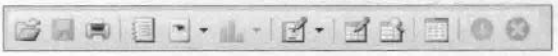


图 2.4 Stata 的快捷键

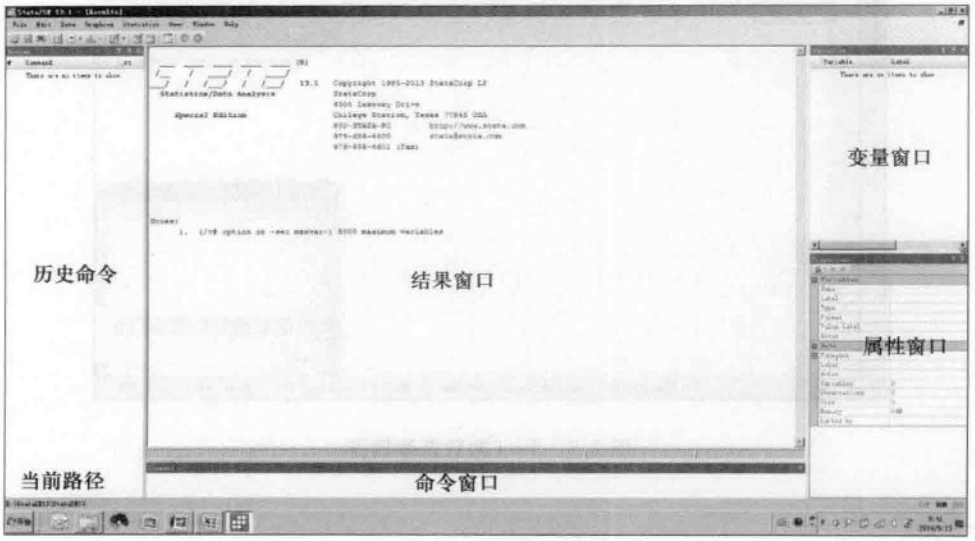



图 2.5 Stata 13 的主要窗口

的输出结果。中间下方的小窗口为“命令窗口”(Command),可在此窗口输入 Stata 命令。右上方为“变量窗口”(Variables),记录着内存中存放的所有 Stata 变量名字与标签。右下方为“属性窗口”(Properties),显示当前数据文件与变量的属性(此窗口为 Stata 新增,并不常用)。在左下角显示“当前路径”(current path),比如“D:\StataSE13\StataSE13”,即文件的默认存储与调用位置。

## 2.3 Stata 操作实例

学习 Stata 的最好方法大概是通过实例来学习。因此,这里以数据集 grilic_small. xls (Excel 文件)为例,该文件包含 30 名美国年轻男子的教育投资回报率数据[①]。

### 1. 导入数据

由于大多数的原始数据存在于 Excel 表的格式,故我们着重介绍如何从 Excel 表将数据导入 Stata。首先,打开 Stata 软件后,点击快捷键 Data Editor(Edit)图标(参见图 2.6 的鼠标位置,也可点击菜单“Window”→“Data Editor”),即可打开类似 Excel 的空白表格,为 Stata 的数据编辑器,参见图 2.7。

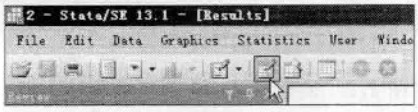



图 2.6 Data Editor(Edit)图标

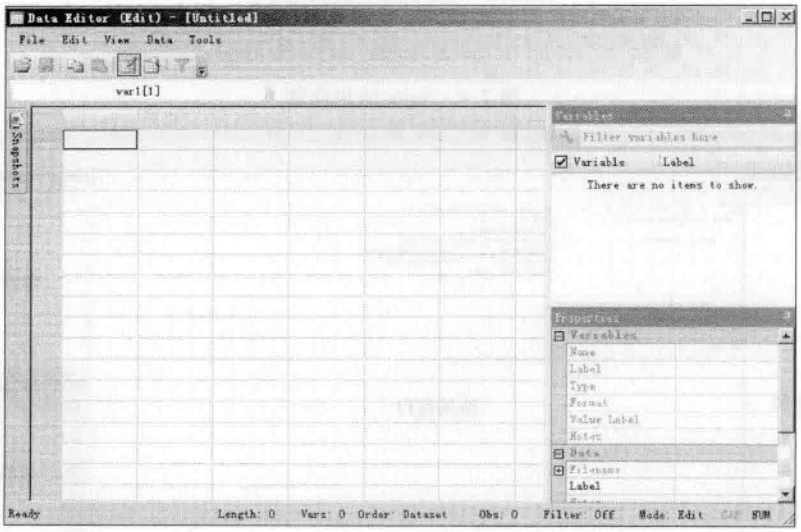
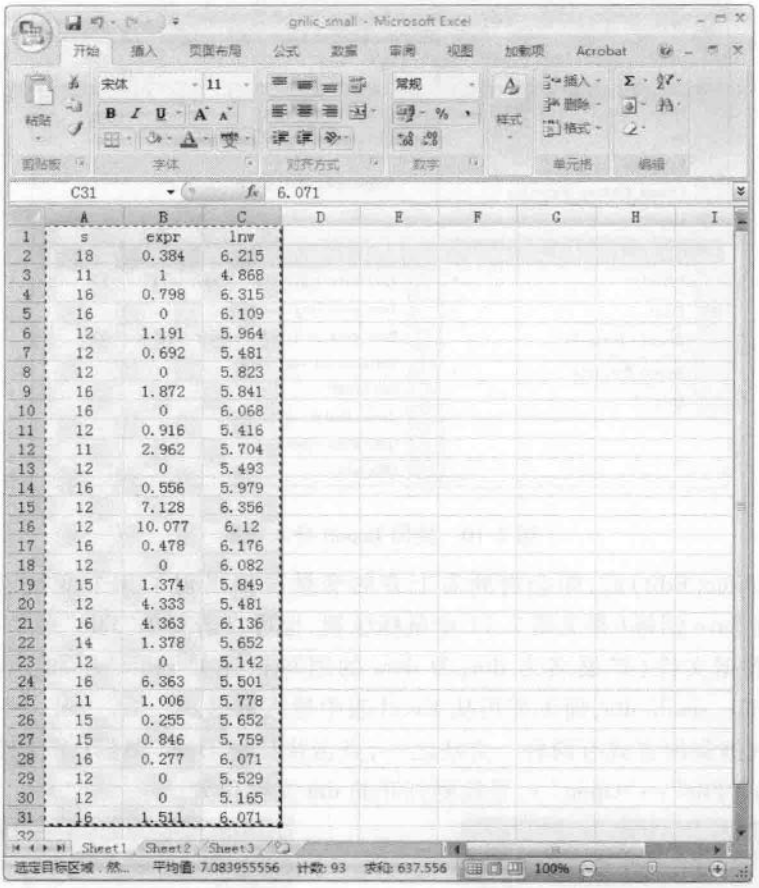



图 2.7 Stata 的数据编辑器

① 此数据集截取自 grilic. dta,样本容量为 758,我们将在第 4 章使用完整的数据集。

其次，用 Excel 打开文件“grilic_small. xls”，会看到如图 2.8 所示 Excel 格式的数据文件。

图 2.8 显示，共有 3 列变量，分别为 $s$（schooling，教育年限），$expr$（experience，工龄）与 ln $w$（lnwage，工资对数）。复制此 Excel 表中所有数据（Ctrl + C），然后粘贴到 Data Editor 中（Ctrl + V）。此时在 Data Editor 中会出现一个对话框，参见图 2.9。



| | A | B | C |
|---|---|---|---|
| 1 | s | expr | lnw |
| 2 | 18 | 0.384 | 6.215 |
| 3 | 11 | 1 | 4.868 |
| 4 | 16 | 0.798 | 6.315 |
| 5 | 16 | 0 | 6.109 |
| 6 | 12 | 1.191 | 5.964 |
| 7 | 12 | 0.692 | 5.481 |
| 8 | 12 | 0 | 5.823 |
| 9 | 16 | 1.872 | 5.841 |
| 10 | 16 | 0 | 6.068 |
| 11 | 12 | 0.916 | 5.416 |
| 12 | 11 | 2.962 | 5.704 |
| 13 | 12 | 0 | 5.493 |
| 14 | 16 | 0.556 | 5.979 |
| 15 | 12 | 7.128 | 6.356 |
| 16 | 12 | 10.077 | 6.12 |
| 17 | 16 | 0.478 | 6.176 |
| 18 | 12 | 0 | 6.082 |
| 19 | 15 | 1.374 | 5.849 |
| 20 | 12 | 4.333 | 5.481 |
| 21 | 16 | 4.363 | 6.136 |
| 22 | 14 | 1.378 | 5.652 |
| 23 | 12 | 0 | 5.142 |
| 24 | 16 | 6.363 | 5.501 |
| 25 | 11 | 1.006 | 5.778 |
| 26 | 15 | 0.255 | 5.652 |
| 27 | 15 | 0.846 | 5.759 |
| 28 | 16 | 0.277 | 6.071 |
| 29 | 12 | 0 | 5.529 |
| 30 | 12 | 0 | 5.165 |
| 31 | 16 | 1.511 | 6.071 |

图 2.8 Excel 表中的数据

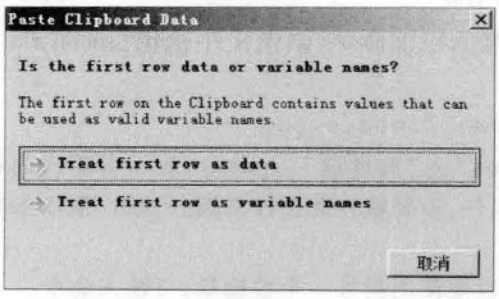



图 2.9 Data Editor 的对话框

此对话框问你“第一行为数据还是变量名”，点击相应的选择即可。对于此数据集，由于第

一行为变量名而非数据，故应选择第二项，即“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，参见图 2.10；但不如直接从 Excel 表中粘贴数据更为方便直观。

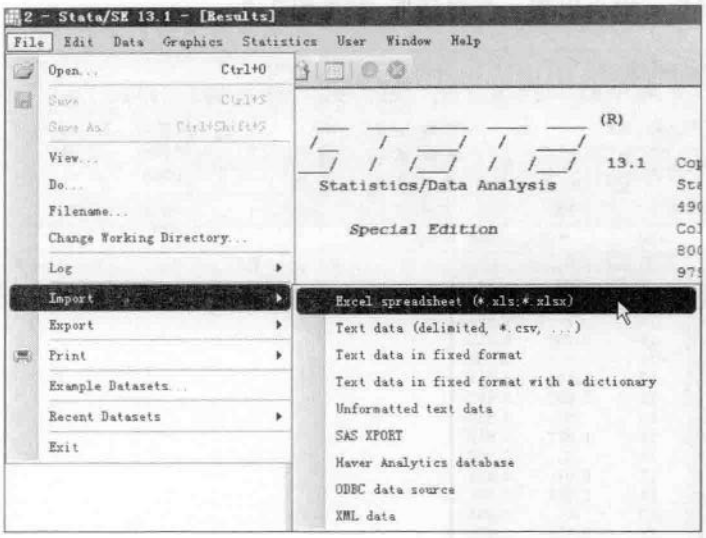



图 2.10 使用 Import 导入数据

关闭 Data Editor(Edit)后，即会看到右上方的变量窗口出现了 3 个变量，分别为 $s$，*expr* 与 $\ln w$。点击快捷键 Save 图标（参见图 2.11 中鼠标位置，也可点击菜单“File”→“Save”），将数据存为 Stata 格式的数据文件（扩展名为 dta，为 data 的缩写），比如 grilic_small. dta。此后，就可用 Stata直接打开 grilic_small. dta，而无需再从 Excel 表中导入数据。

打开 Stata 数据集的方式有两种。方法之一，点击快捷键 Open 图标（参见图 2.12 中鼠标位置，也可点击菜单“File”→“Open”），寻找要打开的 dta 文件位置。

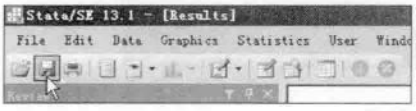



图 2.11 Save 图标


图 2.12 Open 图标

方法之二，在命令窗口输入以下命令（假设文件 grilic_small. dta 在 E 盘的根目录），然后回车（按 Enter 键）：

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

其中，逗号“,”之后的“clear”为“选择项”（option），表示可替代内存中的已有数据。显然，使用命令 use 打开 dta 数据文件，需要输入此文件的路径；故一般不如使用快捷键 Open 图标寻找此文件，再打开更为方便。

如要关闭一个数据集，以便使用另外一个数据集，可输入命令

```
. clear
```

内存中数据将被清空，然后可再打开另一数据集。

### 2. 变量的标签

在变量窗口,变量的“名字”(Name)旁边会显示其“标签”(label)。目前的标签过于简略,缺乏变量的解释信息。点击 Variables Manager 图标(参见图 2.13 中鼠标位置,也可点击菜单“Window”→“Variables Manager”),即可打开变量管理器,然后编辑变量名、标签等。


图 2.13 Variables Manager 图标

比如,将变量 *s* 的标签改为“schooling”,然后点击“Apply”(应用),参见图 2.14。需要注意的是,Stata 严格区分大小写字母(case sensitive)。一般建议变量名使用小写字母,以便于阅读。

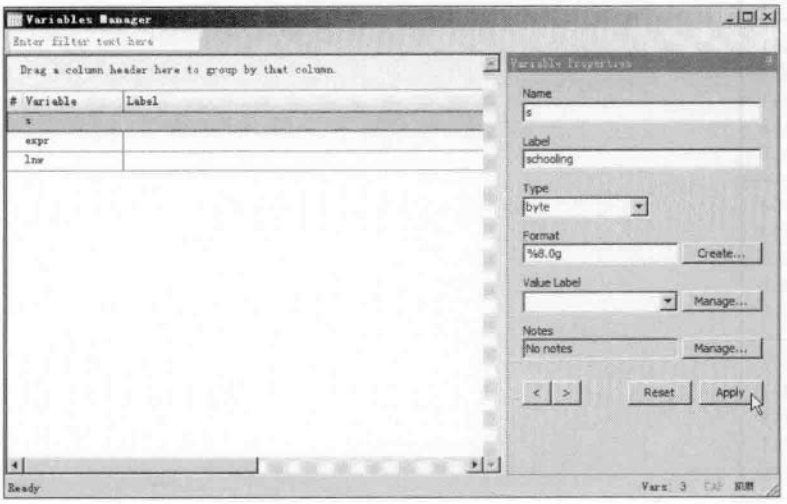



图 2.14 变量管理器的对话框

### 3. 审视数据

如果想看数据集中的变量名称、标签等,可输入命令

```
. describe
```

其中,“describe”的下划线表示,可将该命令简写为“d”。

```
Contains data
  obs:            30
 vars:             3
 size:           270
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
s               byte    %8.0g                 schooling
expr            float   %8.0g
lnwage          float   %8.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

除了变量名称(variable name)与变量标签(variable label)外,上表还显示了变量的存储类型(storage type)与显示格式(display format),初学者可不必理会。如果想看变量 $s$ 与 ln $w$ 的具体数据,可使用命令

```
. list s lnw
```

| | s | lnw |
|---|---|---|
| 1. | 18 | 6.215 |
| 2. | 11 | 4.868 |
| 3. | 16 | 6.315 |
| 4. | 16 | 6.109 |
| 5. | 12 | 5.964 |
| 6. | 12 | 5.481 |
| 7. | 12 | 5.823 |
| 8. | 16 | 5.841 |
| 9. | 16 | 6.068 |
| 10. | 12 | 5.416 |
| 11. | 11 | 5.704 |
| 12. | 12 | 5.493 |
| 13. | 16 | 5.979 |
| 14. | 12 | 6.356 |
| 15. | 12 | 6.12 |
| 16. | 16 | 6.176 |
| 17. | 12 | 6.082 |
| 18. | 15 | 5.849 |
| 19. | 12 | 5.481 |
| 20. | 16 | 6.136 |
| 21. | 14 | 5.652 |
| 22. | 12 | 5.142 |
| 23. | 16 | 5.501 |
| 24. | 11 | 5.778 |
| 25. | 15 | 5.652 |

—more—

在上面的结果中,由于无法在整个屏幕显示所有结果,故在屏幕底端出现一个带下划线的英文字"more",用鼠标单击"more",即可翻看下页的结果。假如结果有多页,则需多次手工点击"more"翻页。如果想连续滚屏显示命令运行结果,可输入命令

```
. set more off
```

如果又想恢复分页显示命令运行结果,可输入命令

```
. set more on
```

如果只想对数据集的一部分子集执行命令,比如只看 $s$ 与 ln $w$ 的前 5 个数据,可使用命令

```
. list s lnw in 1/5
```

| | s | lnw |
|---|---|---|
| 1. | 18 | 6.215 |
| 2. | 11 | 4.868 |
| 3. | 16 | 6.315 |
| 4. | 16 | 6.109 |
| 5. | 12 | 5.964 |

类似地,如果要罗列第 11—15 个观测值,可输入命令

```
. list s lnw in 11/15
```

| | s | lnw |
|---|---|---|
| 11. | 11 | 5.704 |
| 12. | 12 | 5.493 |
| 13. | 16 | 5.979 |
| 14. | 12 | 6.356 |
| 15. | 12 | 6.12 |

也可以通过逻辑关系来定义数据集的子集。比如,要列出所有满足条件"$s \geq 16$"(教育年限为 16 年及以上)的数据,可使用以下命令

```
. list s lnw if s >=16
```

| | s | lnw |
|---|---|---|
| 1. | 18 | 6.215 |
| 3. | 16 | 6.315 |
| 4. | 16 | 6.109 |
| 8. | 16 | 5.841 |
| 9. | 16 | 6.068 |
| 13. | 16 | 5.979 |
| 16. | 16 | 6.176 |
| 20. | 16 | 6.136 |
| 23. | 16 | 5.501 |
| 27. | 16 | 6.071 |
| 30. | 16 | 6.071 |

其中," >= "表示"大于等于"。其他表示关系的逻辑符号为" == "(等于)," > "(大于)," < "(小于)," <= "(小于等于)," ~= "(不等于,也可用"! = "表示)。在 Stata 中(与一般计算机语言一样),一个等号" = "表示"赋值",而两个等号" == "表示"等于"。

查看具体数据的直接方法是,点击 Data Editor (Edit)图标,或者点击该图标右边的 Data Editor (Browse)图标,参见图 2.15 中的鼠标位置。二者的区别在于,后者(Browse)只能看,不能改;而前者(Edit)还可以修改数据。

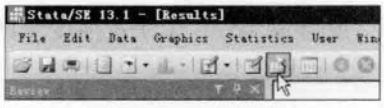



图 2.15 Data Editor(Browse)图标

如果要删除满足"$s \geq 16$"条件的观测值,可输入命令

```
. drop if s >=16
```

反之，如果只想保留满足“$s \geq 16$”条件的观测值，可使用命令

```
. keep if s >=16
```

需要注意的是，删除观测值之后，Stata 并不提供类似于 Microsoft Word 的撤销（undo）命令。故一般建议慎重删除数据，最好先将原始数据备份。

如果想将数据按照变量 $s$ 的升序排列，可输入命令

```
. sort s
. list
```

| | s | expr | lnw |
|---|---|---|---|
| 1. | 11 | 1 | 4.868 |
| 2. | 11 | 1.006 | 5.778 |
| 3. | 11 | 2.962 | 5.704 |
| 4. | 12 | 0 | 6.082 |
| 5. | 12 | 0 | 5.529 |
| 6. | 12 | 0 | 5.823 |
| 7. | 12 | 7.128 | 6.356 |
| 8. | 12 | 0 | 5.493 |
| 9. | 12 | 0 | 5.165 |
| 10. | 12 | 10.077 | 6.12 |
| 11. | 12 | .916 | 5.416 |
| 12. | 12 | 4.333 | 5.481 |
| 13. | 12 | .692 | 5.481 |
| 14. | 12 | 0 | 5.142 |
| 15. | 12 | 1.191 | 5.964 |
| 16. | 14 | 1.378 | 5.652 |
| 17. | 15 | .255 | 5.652 |
| 18. | 15 | .846 | 5.759 |
| 19. | 15 | 1.374 | 5.849 |
| 20. | 16 | 0 | 6.109 |
| 21. | 16 | 6.363 | 5.501 |
| 22. | 16 | 1.511 | 6.071 |
| 23. | 16 | 0 | 6.068 |
| 24. | 16 | .478 | 6.176 |
| 25. | 16 | .277 | 6.071 |
| 26. | 16 | 4.363 | 6.136 |
| 27. | 16 | 1.872 | 5.841 |
| 28. | 16 | .798 | 6.315 |
| 29. | 16 | .556 | 5.979 |
| 30. | 18 | .384 | 6.215 |

但命令 sort 无法按照变量的降序排列。如果想按降序排列，可使用命令 gsort：

```
. gsort -s
. list
```

| | s | expr | lnw |
|---|---|---|---|
| 1. | 18 | .384 | 6.215 |
| 2. | 16 | .556 | 5.979 |
| 3. | 16 | .798 | 6.315 |
| 4. | 16 | 1.872 | 5.841 |
| 5. | 16 | 4.363 | 6.136 |
| 6. | 16 | .277 | 6.071 |
| 7. | 16 | .478 | 6.176 |
| 8. | 16 | 0 | 6.068 |
| 9. | 16 | 1.511 | 6.071 |
| 10. | 16 | 6.363 | 5.501 |
| 11. | 16 | 0 | 6.109 |
| 12. | 15 | 1.374 | 5.849 |
| 13. | 15 | .846 | 5.759 |
| 14. | 15 | .255 | 5.652 |
| 15. | 14 | 1.378 | 5.652 |
| 16. | 12 | 1.191 | 5.964 |
| 17. | 12 | 0 | 5.142 |
| 18. | 12 | .692 | 5.481 |
| 19. | 12 | 4.333 | 5.481 |
| 20. | 12 | .916 | 5.416 |
| 21. | 12 | 10.077 | 6.12 |
| 22. | 12 | 0 | 5.165 |
| 23. | 12 | 0 | 5.493 |
| 24. | 12 | 7.128 | 6.356 |
| 25. | 12 | 0 | 5.823 |
| 26. | 12 | 0 | 5.529 |
| 27. | 12 | 0 | 6.082 |
| 28. | 11 | 2.962 | 5.704 |
| 29. | 11 | 1.006 | 5.778 |
| 30. | 11 | 1 | 4.868 |

**4. 画图**

看数据的最直观方法是画图。比如，想看样本中变量 $s$ 的分布情况，可输入以下命令来画直方图（结果参见图 2.16）：

```
. histogram s, width(1) frequency
```

其中，“histogram”表示直方图，选择项“width(1)”表示将组宽设为 1（否则将使用 Stata 根据样本容量计算的默认分组数），选择项“frequency”表示将纵坐标定为频数（默认使用密度）。从图 2.16 可知，教育年限的分布呈双峰状，受 12 年教育的人数最多（高中毕业），其次为受 16 年教育者（大学毕业）。如果想知道更多有关命令 histogram 的选项与用法，可输入命令

```
. help histogram
```

事实上，对于任何 Stata 命令，只要输入“help command_name”即可查看该命令的“帮助文件”（help file）。初学者应养成经常查看帮助文件的习惯。

如果想考察教育年限与工资对数之间的关系，最直观的方法便是画 $s$ 与 $\ln w$ 之间的散点图，

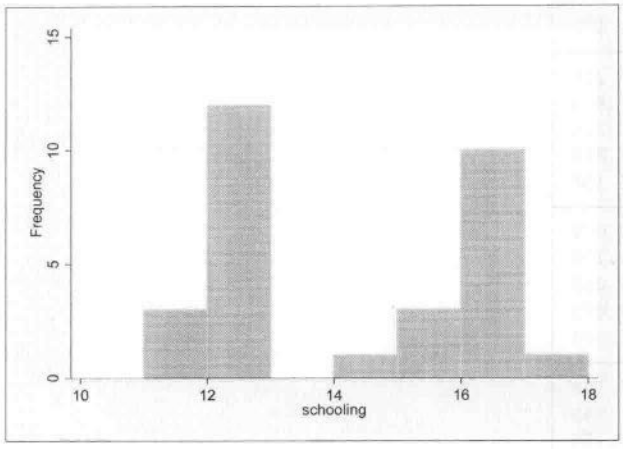



图 2.16 教育年限的直方图

可输入命令(结果参见图 2.17):

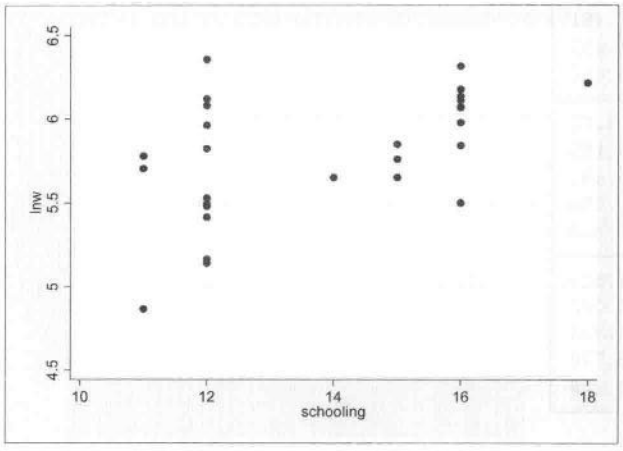



图 2.17 教育年限与工资对数的散点图

```
. scatter lnw s
```

从图 2.17 可知,工资对数与教育年限似乎存在正相关关系。如果想在散点图上标注出每个点对应于哪个观测值,可先定义变量 *n*,表示第 *n* 个观测值:

```
. gen n = _n
```

其中,"_n"表示第 *n* 个观测值。然后以变量 *n* 作为每个点的标签来画散点图,结果参见图 2.18。

```
. scatter lnw s,mlabel(n)
```

其中,选择项"mlabel(n)"表示,以变量 *n* 作为标签(mark label)。

Stata 提供了丰富的作图方法。更多作图方法,参见下拉式菜单"Graphics"(参见图 2.19)。

**5. 统计分析**

Stata 是一款功能强大的统计软件。如果想看变量 *s* 的统计特征,可输入命令

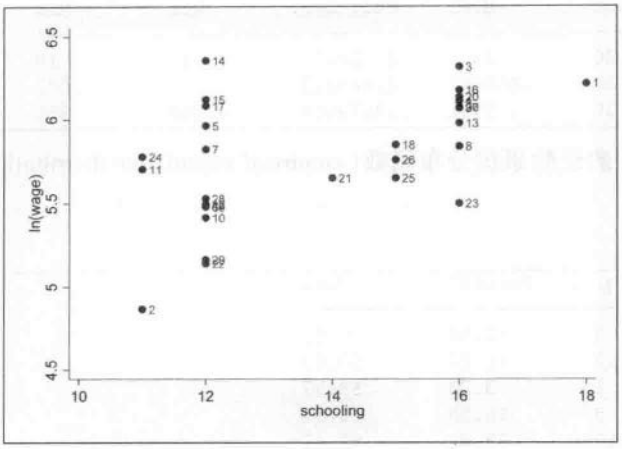



图 2.18 加标签的散点图

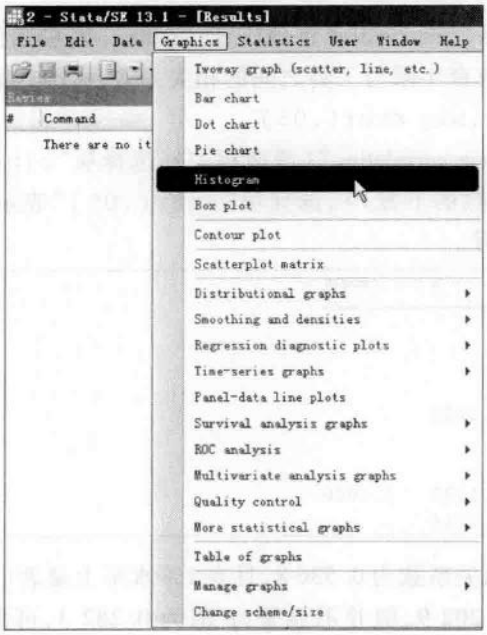



图 2.19 Stata 的作图功能

```
. summarize s
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| s | 30 | 13.8 | 2.139932 | 11 | 18 |

此结果显示了变量 $s$ 的样本容量、平均值、标准差、最小值与最大值。如果不指明变量，则将显示数据集中所有变量的统计指标。

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| s | 30 | 13.8 | 2.139932 | 11 | 18 |
| expr | 30 | 1.658667 | 2.445213 | 0 | 10.077 |
| lnw | 30 | 5.7932 | .3679956 | 4.868 | 6.356 |

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

```
. tabulate s
```

| schooling | Freq. | Percent | Cum. |
|---|---|---|---|
| 11 | 3 | 10.00 | 10.00 |
| 12 | 12 | 40.00 | 50.00 |
| 14 | 1 | 3.33 | 53.33 |
| 15 | 3 | 10.00 | 63.33 |
| 16 | 10 | 33.33 | 96.67 |
| 18 | 1 | 3.33 | 100.00 |
| Total | 30 | 100.00 | |

其中,"Freq"表示频数,"Percent"表示百分比,而"Cum."表示累积百分比。

如果要显示工资对数、教育年限与工龄之间的相关系数,可输入命令

```
. pwcorr lnw s expr,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)[①],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。

| | lnw | s | expr |
|---|---|---|---|
| lnw | 1.0000 | | |
| s | 0.5368* | 1.0000 | |
| | 0.0022 | | |
| expr | 0.2029 | -0.1132 | 1.0000 |
| | 0.2823 | 0.5514 | |

上表显示,ln *w* 与 *s* 的相关系数为 0.536 8,且在 1% 水平上显著(*p* 值为 0.002 2);ln *w* 与 *expr* 的相关系数虽然也达到 0.202 9,但并不显著(*p* 值为 0.282 3,可能因为样本容量较小,仅为 30);*s* 与 *expr* 的相关系数为 -0.113 2,可能因为上学时间长的年轻人,参加工作时间就不长,但此负相关关系也不显著(*p* 值为 0.551 4)。

### 6. 生成新变量

在进行统计与计量分析时,常需根据已有变量生成新变量,比如取对数、平方等。在 Stata 中定义新变量,可通过命令generate 来实现。比如,输入如下命令可定义一个新变量"教育年限的对数":

① 有关显著性水平与 *p* 值,参见第 5 章。

```
. generate lns = log(s)
```

如果需要定义 $s$ 的平方项,可使用命令

```
. gen s2 = s ^2
```

如要生成 $s$ 与 *expr* 的互动项(interaction term),可输入命令

```
. gen exprs = s * expr
```

如果想根据工资对数 ln $w$ 计算工资水平 $w$,可使用命令

```
. gen w = exp(lnw)
```

在计量经济学中,经常使用"虚拟变量"(dummy variable,也称"哑变量"),即取值只能为 0 或 1 的变量,比如性别。假设定义"$s \geqslant 16$"为"受过高等教育",并使用变量 *college* 来表示:

$$college \equiv \begin{cases} 1, & \text{如果} \quad s \geqslant 16 \\ 0, & \text{其他} \end{cases} \tag{2.1}$$

可使用如下命令

```
. gen colleg = (s >=16)
```

其中,括弧"()"表示对括弧中的表达式"s >=16"进行逻辑评估:如果此表达式为真,则取值为 1;如果为假,则取值为 0。在上面命令中,不慎把 *college* 写成 *colleg* 了。可使用如下命令将变量重新命名:

```
. rename colleg college
```

这样,变量 *colleg* 被重新命名为 *college*(也可使用变量管理器重新命名)。

假设想将"受过高等教育"的定义改为"$s \geqslant 15$",但仍用 *college* 作为变量名。方法之一为,先去掉现有变量 *college*,然后再重新定义一次:

```
. drop college
. gen college = (s >=15)
```

方法之二,则只需使用一个命令:

```
. replace college = (s >=15)
```

此命令直接将原变量($s \geqslant 16$)替换为新变量($s \geqslant 15$)。

对于较长的变量名,一一输入变量名比较麻烦。有如下三个简便的方法。方法一,直接在变量窗口双击需要的变量,该变量名就会出现在命令窗口。方法二,如有以下变量 $s1$, $s2$, $s3$, $s4$, $s5$(比如,对教育年限的 5 种度量方法),可用 $s1$—$s5$ 来简略地表示这 5 个变量。方法三,用"*"号来简化变量名的书写。假设想将内存中所有以"$s$"开头的变量都去掉,可输入命令

```
. drop s*
```

这将去掉内存中的 $s1$, $s2$, $s3$, $s4$, $s5$ 变量(删除之后无法恢复,故应慎重使用)。

**7. Stata 的计算器功能**

Stata 也可作为计算器使用,命令格式为"display expression"。比如,计算 ln2,可输入如下命令

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

如果要计算$\sqrt{2}$,则可输入命令

```
. dis 2 ^0.5
1.4142136
```

### 8. 调用命令与终止命令

如果每次都完整地输入整行命令,可能较费时。较有效率的方法是,调用某个曾经使用过的命令,并在此基础上修改。调用旧命令的方法有两种。方法一,把光标放在命令窗口,按键盘上的"Pg Up"键调用上一条命令,而按"Pg Dn"键即可调用下一条命令。

方法二,在历史命令窗口单击旧命令,可将旧命令调入命令窗口,然后进行编辑;如果用鼠标双击旧命令,则将再次执行此旧命令。

有时候,运行某个命令花费时间较长(比如,在进行数值计算时,迭代无法收敛)。如果想中途停止该命令的执行,可点击快捷键 Break 图标(参见图 2.20 中的鼠标位置),或直接在键盘上同时按"Ctrl + Break"。

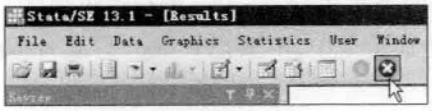



图 2.20 Break 图标

### 9. Stata 的日志

在进行实证研究时,有时会得到很多结果,这些结果虽然显示在屏幕上,但退出 Stata 后结果将丢失[①]。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"来定义"日志文件"(log file),参见图 2.21。

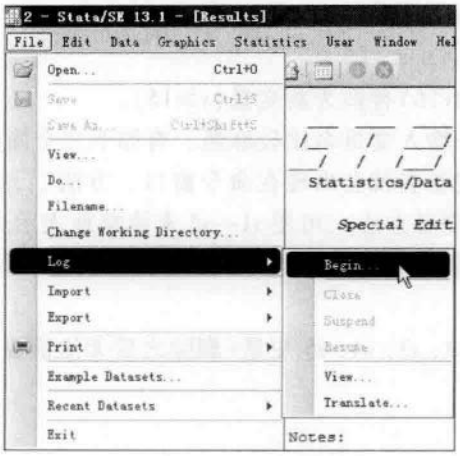



图 2.21 定义日志文件

① 如果 Stata 输出结果太多,则可能无法全部在屏幕显示,即只能看到较新的结果,而最早的结果被"挤出"而不再显示于屏幕。

也可以直接点击快捷键 Log 图标，参见图 2.22 中的鼠标位置。

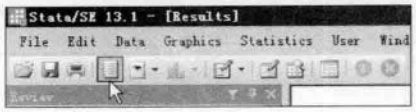



图 2.22 Log 图标

此时，会出现如下对话框，参见图 2.23。只要在对话框中输入日志的文件名，并存储在指定的位置即可。

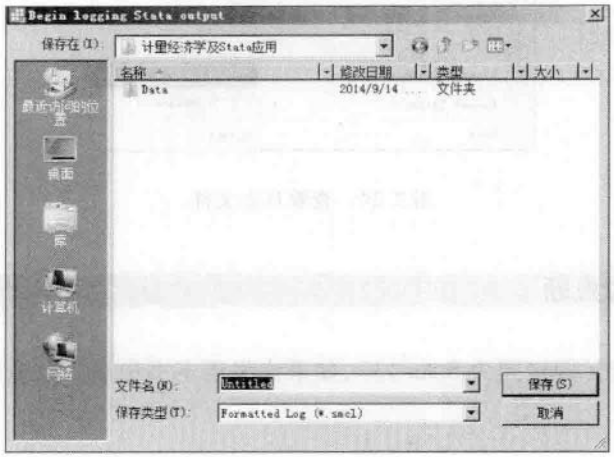



图 2.23 日志文件的对话框

从图 2.23 可知，Stata 日志文件的扩展名为 smcl[①]。也可以直接在命令窗口输入如下命令：

```
. log using today
```

```
      name:  <unnamed>
       log:  D:\StataSE13\StataSE13\today.smcl
  log type:  smcl
 opened on:  15 Sep 2014, 14:00:15
```

这样，在当前路径就会生成一个名为“today. smcl”的日志文件。定义日志文件后，在 Stata 中的所有操作及结果，都将记录在日志中，直至选择退出此日志文件。

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，则可输入命令

```
. log close
```

如果要查看日志文件的内容，可点击菜单“File”→“Log”→“View”，然后寻找想要打开的日志文件，参见图 2.24。

① “smcl”为“Stata Markup and Control Language”的缩写。


图 2.24 查看日志文件

## 2.4 Stata 命令库的更新

由于 Stata 版本不同(即使同为 Stata 13),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(Stata"ado"文件及其他可执行文件)。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,SSC),由 Boston College 维护,网址为 http://ideas. repec. org/s/boc/bocode. html。

从 SSC 下载 Stata 程序的命令为:

```
. ssc install newcommand
```

此时,所有下载与安装过程都将自动完成(包括新命令的帮助文件)。但如果非官方命令不是来自 SSC,则一般需自己手工安装,只要将所有相关文件下载到指定的 Stata 文件夹中即可(通常为 ado\plus\)。如果不清楚应把文件复制到哪个文件夹,可输入以下命令,显示 Stata 的系统路径(system directories):

```
. sysdir
```

你会看到类似于以下的结果(取决于 Stata 的安装位置),

```
   STATA:  D:\StataSE13\StataSE13\
    BASE:  D:\StataSE13\StataSE13\ado\base\
    SITE:  D:\StataSE13\StataSE13\ado\site\
    PLUS:  c:\ado\plus\
PERSONAL:  c:\ado\personal\
OLDPLACE:  c:\ado\
```

将下载的新命令文件复制到 PLUS 所指示的那个文件夹即可(此处为"c:\ado\plus\")。如果想使用某种估计方法,但不知道它是否存在,可输入命令

```
. search keyword
```

此命令将搜索 Stata 帮助文件、Stata 常见问题[①]、Stata 案例[②]、*Stata Journal*[③]，*Stata Technical Bulletin*[④] 等。进一步的搜索可输入以下命令

```
. findit keyword
```

命令 findit 的搜索范围比命令 search 更广，还包括 Stata 的网络资源。事实上，"findit"等价于"search,all"。另外，由于命令 search 的搜索结果较少，故直接在 Stata 结果窗口显示；而命令 findit 的搜索结果较多，故将打开另一页面显示。

## 2.5 进一步学习 Stata 的资源

更多 Stata 知识，将在本书以后章节中逐步介绍。Stata 英文参考书包括 Baum(2006)[⑤]，Cameron and Trivedi(2010)，以及 Stata 出版社(Stata Press)出版的系列书籍。加州大学洛杉矶分校(UCLA)网站(http://www.ats.ucla.edu/stat/stata/)有大量 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)菜单包含了较详细的使用说明，比如，"help histogram"。更高级的学习，可查看 Stata 手册(Stata manuals)[⑦]，这些手册对每个 Stata 命令都进行了详尽的说明。

**习题**

2.1 安装 Stata 软件，并将本章的 Stata 命令与实例操作一遍。

---

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 中译本为克里斯托弗·鲍姆.用 Stata 学计量经济学.北京：中国人民大学出版社，2012。

⑥ 可搜索下载电子版。

⑦ 如果安装的是较完整的 Stata 版本，这些 Stata 手册通常为 PDF 文件，放在文件夹"docs"中。也可在 Stata 官网下载：http://www.stata.com/features/documentation/。

*Probability is the very guide to life.* ——Cicero

*Statistics is the grammar of science.* ——Karl Pearson

# 3. 数学回顾

本章将回顾计量经济学中常用的微积分、线性代数与概率统计的知识，侧重于直观含义的解释。更全面的介绍，请参考你用过的相应数学教材。

## 3.1 微积分

### 1. 导数

对于一元函数 $y=f(x)$，记其**一阶导数**(first derivative)为$\frac{dy}{dx}$或$f'(x)$，其定义为

$$\frac{dy}{dx} \equiv f'(x) \equiv \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \equiv \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{3.1}$$

其中，“≡”表示“定义”。从几何上看，(一阶)导数就是函数 $y=f(x)$ 在 $x$ 处的切线斜率，参见图3.1。一阶导数$f'(x)$仍然是 $x$ 的函数，故可定义$f'(x)$的导数，即**二阶导数**(second derivative)：

$$\frac{d^2y}{dx^2} \equiv f''(x) \equiv \frac{d\left(\frac{dy}{dx}\right)}{dx} \equiv [f'(x)]' \tag{3.2}$$

直观来看，二阶导数表示切线斜率的变化速度，即曲线$f(x)$的弯曲程度，也称“曲率”(curvature)。

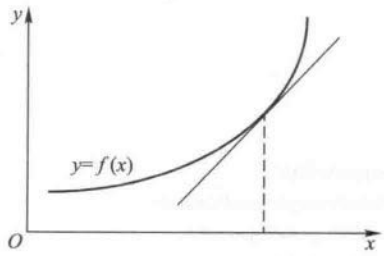



图 3.1 导数的示意图

### 2. 一元最优化

计量经济学中常见的两种估计方法为最小二乘法(参见第4章)与最大似然估计(参见第11章)。二者的本质都是最优化问题(optimization),前者为最小化问题(minimization),而后者为最大化问题(maximization)。为此,考虑以下无约束的一元最大化问题(参见图3.2),

$$\max_{x} f(x) \tag{3.3}$$

从图3.2可知,函数$f(x)$在其"山峰"顶端$x^*$处达到最大值。在$x^*$处,$f(x)$的切线恰好为水平线,故切线斜率为0。这意味着,一元最大化问题的必要条件为

$$f'(x^*) = 0 \tag{3.4}$$

由于此最大化的必要条件涉及一阶导数,故通常称为**一阶条件**(first order condition)。

类似地,考虑无约束的一元最小化问题(参见图3.3),

$$\min_{x} f(x) \tag{3.5}$$

从图3.3可知,最小化问题的一阶条件与最大化问题相同,都要求在最优值$x^*$处的切线斜率为0,即$f'(x^*)=0$。二者的区别仅在于最优化的**二阶条件**(second order condition),即最大化要求二阶导数$f''(x^*)\leqslant 0$,而最小化要求$f''(x^*)\geqslant 0$。在经济学中,一般假设二阶条件满足,故主要关注一阶条件。

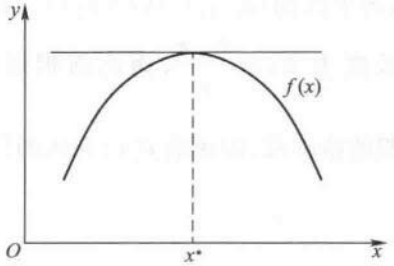



图3.2 最大化的示意图

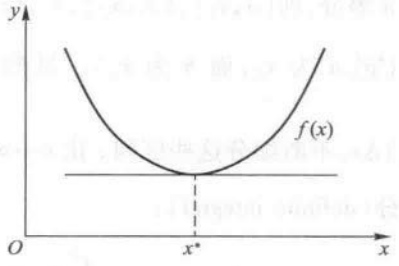



图3.3 最小化的示意图

### 3. 偏导数

对于多元函数$y=f(x_1,x_2,\cdots,x_n)$,定义$y$对于$x_1$的**偏导数**(partial derivative)为

$$\frac{\partial y}{\partial x_1} \equiv \frac{\partial f(x_1,x_2,\cdots,x_n)}{\partial x_1} \equiv \lim_{\Delta x_1 \to 0} \frac{f(x_1+\Delta x_1,x_2,\cdots,x_n)}{\Delta x_1} \tag{3.6}$$

由上式可知,在计算$y$对$x_1$的一阶偏导数时,首先给定$x_2,\cdots,x_n$为常数(可视为参数),则$y=f(x_1,x_2,\cdots,x_n)$可看成$x_1$的一元函数$y=f(x_1,\cdot)$,而$\frac{\partial y}{\partial x_1}$便是此"一元函数"$y=f(x_1,\cdot)$的导数。类似地,可定义$y$对$x_i(i=2,\cdots,n)$的偏导数$\frac{\partial y}{\partial x_i}$。在经济学中,如果$y=f(x_1,x_2,\cdots,x_n)$为

效用函数，则$\frac{\partial y}{\partial x_1}$表示商品 $x_1$ 所能带来的边际效用(marginal utility)。如果 $y=f(x_1,x_2,\cdots,x_n)$ 为生产函数，则$\frac{\partial y}{\partial x_1}$表示生产要素 $x_1$ 所能带来的边际产出(marginal output)。

**4. 多元最优化**

考虑以下无约束的多元最大化问题，

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) \equiv f(x_1,x_2,\cdots,x_n) \tag{3.7}$$

其中，$\boldsymbol{x}\equiv(x_1,x_2,\cdots,x_n)$。其一阶条件要求在最优值 $\boldsymbol{x}^*$ 处，所有偏导数均为 0：

$$\frac{\partial f(\boldsymbol{x}^*)}{\partial x_1}=\frac{\partial f(\boldsymbol{x}^*)}{\partial x_2}=\cdots=\frac{\partial f(\boldsymbol{x}^*)}{\partial x_n}=0 \tag{3.8}$$

多元最小化的一阶条件与此相同。此一阶条件要求在最优值 $\boldsymbol{x}^*$ 处，曲面 $f(\boldsymbol{x})$ 在各个方向的切线斜率都为 0。

**5. 积分**

考虑计算连续函数 $y=f(x)$ 在区间 $[a,b]$ 上的面积，参见图 3.4。作为近似，可将区间 $[a,b]$ 划分为 $n$ 等分，即 $[a,x_1],(x_1,x_2],\cdots,(x_{n-1},b]$，然后从每个区间 $(x_{i-1},x_i]$ $(i=1,\cdots,n)$ 中任取一点 $\xi_i$(记 $a$ 为 $x_0$，而 $b$ 为 $x_n$)。显然，每个区间的长度为 $\Delta x\equiv\frac{b-a}{n}$，而此面积近似等于 $\sum_{i=1}^{n}f(\xi_i)\Delta x$。不断细分这些区间，让 $n\to\infty$，可得到此面积的精确值，即函数 $f(x)$ 在区间 $[a,b]$ 上的**定积分**(definite integral)：

$$\int_a^b f(x)\,\mathrm{d}x \equiv \lim_{n\to\infty}\sum_{i=1}^{n}f(\xi_i)\Delta x \tag{3.9}$$

其中，在极限处，将 $\Delta x$ 记为 $\mathrm{d}x$，而将求和符号 $\sum$(英文为 Summation)记为 $\int$，由大写字母 S 向上拉长而成。由此可见，定积分的实质就是求和(只不过是无穷多项之和)。

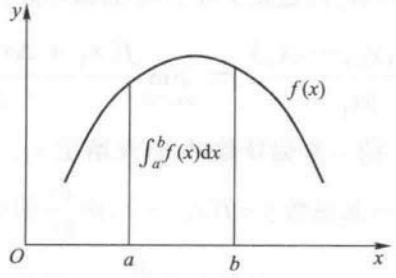



图 3.4 定积分的示意图

## 3.2 线性代数

### 1. 矩阵

将 $m \times n$ 个实数排列成如下矩状的阵形：

$$\boldsymbol{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{3.10}$$

称 $\boldsymbol{A}$ 为 $m \times n$ **矩阵**(matrix)，其中 $m$ 为矩阵 $\boldsymbol{A}$ 的行数(row dimension)，而 $n$ 为矩阵 $\boldsymbol{A}$ 的列数(column dimension)。$\boldsymbol{A}$ 中元素 $a_{ij}$ 表示矩阵 $\boldsymbol{A}$ 的第 $i$ 行、第 $j$ 列元素。比如，$a_{12}$ 为第 1 行、第 2 列的元素，以此类推。矩阵 $\boldsymbol{A}$ 有时也记为 $\boldsymbol{A}_{m\times n}$(以下标强调矩阵的维度)，或 $(a_{ij})_{m\times n}$(以代表性元素 $a_{ij}$ 表示此矩阵)。如果 $\boldsymbol{A}$ 中所有元素都为 0，则称为**零矩阵**(zero matrix)，记为 $\boldsymbol{0}$。零矩阵在矩阵运算中的作用，相当于 0 在数的运算中的作用。

### 2. 方阵

如果 $m = n$，则称 $\boldsymbol{A}$ 为 $n$ 级**方阵**(square matrix)，即

$$\boldsymbol{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{3.11}$$

此时，称 $a_{11}, a_{22}, \cdots, a_{nn}$ 为主对角线上的元素(diagonal elements)，而 $\boldsymbol{A}$ 中的其他元素为非主对角线元素(off-diagonal elements)。如果方阵 $\boldsymbol{A}$ 中的元素满足 $a_{ij} = a_{ji}$(任意 $i, j = 1, \cdots, n$)，则称矩阵 $\boldsymbol{A}$ 为**对称矩阵**(symmetric matrix)。如果方阵 $\boldsymbol{A}$ 的非主对角线元素全部为 0，则称为**对角矩阵**(diagonal matrix)：

$$\boldsymbol{A} \equiv \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \tag{3.12}$$

如果一个 $n$ 级对角矩阵的主对角线元素都为 1，则称为 $n$ 级**单位矩阵**(identity matrix)，记为 $\boldsymbol{I}$ 或 $\boldsymbol{I}_n$(以下标 $n$ 强调其维度)：

$$\boldsymbol{I} \equiv \boldsymbol{I}_n \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n\times n} \tag{3.13}$$

单位矩阵在矩阵运算中的作用,相当于 1 在数的运算中的作用。

### 3. 矩阵的转置

如果将矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$ 的第 1 行变为第 1 列,第 2 行变为第 2 列,……,第 $m$ 行变为第 $m$ 列,则可以得到其**转置矩阵**(transpose),记为 $\boldsymbol{A}'$(英文读为 *A prime*),其维度为 $n \times m$。换言之,矩阵 $\boldsymbol{A}'$ 的 $(i,j)$ 元素 $(\boldsymbol{A}')_{ij}$ 正好是矩阵 $\boldsymbol{A}$ 的 $(j,i)$ 元素 $(\boldsymbol{A})_{ji}$,即

$$(\boldsymbol{A}')_{ij} \equiv (\boldsymbol{A})_{ji} \tag{3.14}$$

如果 $\boldsymbol{A}$ 为对称矩阵,则 $\boldsymbol{A}$ 的转置还是它本身,即 $\boldsymbol{A}' = \boldsymbol{A}$。显然,矩阵转置的转置仍是它本身,即 $(\boldsymbol{A}')' = \boldsymbol{A}$。

### 4. 向量

如果 $m = 1$,则矩阵 $\boldsymbol{A}_{1\times n}$ 为 $n$ 维**行向量**(row vector);如果 $n = 1$,则矩阵 $\boldsymbol{A}_{m\times 1}$ 为 $m$ 维**列向量**(column vector)。显然,向量是矩阵的特例。

考察 $n$ 维列向量 $\boldsymbol{a} = (a_1 \quad a_2 \quad \cdots \quad a_n)'$ 与 $\boldsymbol{b} = (b_1 \quad b_2 \quad \cdots \quad b_n)'$。向量 $\boldsymbol{a}$ 与 $\boldsymbol{b}$ 的**内积**(inner product)或**点乘**(dot product)可定义为

$$\boldsymbol{a}'\boldsymbol{b} \equiv (a_1 \quad a_2 \quad \cdots \quad a_n)\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \equiv a_1b_1 + a_2b_2 + \cdots + a_nb_n = \sum_{i=1}^{n} a_ib_i \tag{3.15}$$

如果 $\boldsymbol{a}'\boldsymbol{b} = 0$,则称向量 $\boldsymbol{a}$ 与 $\boldsymbol{b}$ **正交**(orthogonal),这意味着两个向量在 $n$ 维向量空间中相互垂直(夹角为 90 度),参见图 3.5。

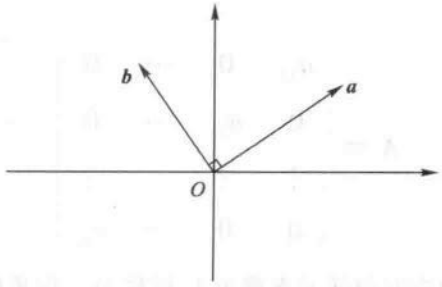



图 3.5 正交的向量

方程(3.15)提示我们,任何形如 $\sum_{i=1}^{n} a_i b_i$ 的乘积求和,都可以很方便地写为向量内积 $\boldsymbol{a}'\boldsymbol{b}$ 的形式。特别地,平方和 $\sum_{i=1}^{n} a_i^2$ 可写为 $\boldsymbol{a}'\boldsymbol{a}$:

$$\boldsymbol{a}'\boldsymbol{a} \equiv (a_1 \quad a_2 \quad \cdots \quad a_n)\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \equiv a_1^2 + a_2^2 + \cdots + a_n^2 = \sum_{i=1}^{n} a_i^2 \tag{3.16}$$

**5. 矩阵的加法**

如果两个矩阵的维度相同(即行数与列数都分别相同),则可以相加。对于 $m \times n$ 矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$,$\boldsymbol{B} = (b_{ij})_{m\times n}$,矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 之和定义为两个矩阵相应元素之和,即

$$\boldsymbol{A} + \boldsymbol{B} \equiv (a_{ij})_{m\times n} + (b_{ij})_{m\times n} \equiv (a_{ij} + b_{ij})_{m\times n} \tag{3.17}$$

容易证明,矩阵的加法满足以下规则:

(1) $\boldsymbol{A} + \boldsymbol{0} = \boldsymbol{A}$ (加上零矩阵不改变矩阵)

(2) $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$ (加法交换律)

(3) $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$ (加法结合律)

(4) $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$ (转置为线性运算)

**6. 矩阵的数乘**

矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$ 与实数 $k$ 的**数乘**(scalar multiplication)定义为此实数 $k$ 与矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$ 每个元素的乘积:

$$k\boldsymbol{A} \equiv k(a_{ij})_{m\times n} \equiv (ka_{ij})_{m\times n} \tag{3.18}$$

**7. 矩阵的乘法**

如果矩阵 $\boldsymbol{A}$ 的列数与矩阵 $\boldsymbol{B}$ 的行数相同,则可以定义**矩阵乘积**(matrix multiplication)$\boldsymbol{A} \times \boldsymbol{B}$,简记 $\boldsymbol{AB}$。假设矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$,矩阵 $\boldsymbol{B} = (b_{ij})_{n\times q}$,则矩阵乘积 $\boldsymbol{AB}$ 的 $(i,j)$ 元素即为矩阵 $\boldsymbol{A}$ 第 $i$ 行与矩阵 $\boldsymbol{B}$ 的第 $j$ 列的内积:

$$(\boldsymbol{AB})_{ij} \equiv (a_{i1} \quad a_{i2} \quad \cdots \quad a_{in})\begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{pmatrix} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{3.19}$$

需要注意的是,矩阵乘法不满足交换律,即一般来说,$\boldsymbol{AB} \neq \boldsymbol{BA}$。而且,只有当矩阵 $\boldsymbol{B}$ 的列数 $q$ 等于矩阵 $\boldsymbol{A}$ 的行数 $m$ 时,$\boldsymbol{B}_{n\times q}\boldsymbol{A}_{m\times n}$ 才有定义。因此,在做矩阵乘法时,需要区分左乘(premultiplication)与右乘(postmultiplication),即 $\boldsymbol{A}$ 左乘 $\boldsymbol{B}$ 为 $\boldsymbol{AB}$,而 $\boldsymbol{A}$ 右乘 $\boldsymbol{B}$ 为 $\boldsymbol{BA}$。

矩阵的乘法满足以下规则：

(1) $\boldsymbol{IA}=\boldsymbol{A},\boldsymbol{AI}=\boldsymbol{A}$　　(乘以单位矩阵不改变矩阵)

(2) $(\boldsymbol{AB})\boldsymbol{C}=\boldsymbol{A}(\boldsymbol{BC})$　　(乘法结合律)

(3) $\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C})=\boldsymbol{AB}+\boldsymbol{AC}$　　(乘法分配律)

(4) $(\boldsymbol{AB})'=\boldsymbol{B}'\boldsymbol{A}',(\boldsymbol{ABC})'=\boldsymbol{C}'\boldsymbol{B}'\boldsymbol{A}'$　　(转置与乘积的混合运算)

### 8. 线性方程组

考虑以下由 $n$ 个方程，$n$ 个未知数构成的线性方程组：

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{3.20}$$

其中，$(x_1 \quad x_2 \quad \cdots \quad x_n)$ 为未知数。根据矩阵乘法的定义，可将上式写为

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{21} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{\boldsymbol{A}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}}_{\boldsymbol{b}} \tag{3.21}$$

记上式中的相应矩阵分别为 $\boldsymbol{A},\boldsymbol{x}$ 与 $\boldsymbol{b}$，可得

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{3.22}$$

直观上，如果可将此方程左边的方阵 $\boldsymbol{A}$“除”到右边去，则可得到 $\boldsymbol{x}$ 的解。为此，引入逆矩阵的概念。

### 9. 逆矩阵

对于 $n$ 级方阵 $\boldsymbol{A}$，如果存在 $n$ 级方阵 $\boldsymbol{B}$，使得 $\boldsymbol{AB}=\boldsymbol{BA}=\boldsymbol{I}_n$（$n$ 级单位矩阵），则称 $\boldsymbol{A}$ 为**可逆矩阵**（invertible matrix）或**非退化矩阵**（nonsingular matrix），而 $\boldsymbol{B}$ 为 $\boldsymbol{A}$ 的**逆矩阵**（inverse matrix），记为 $\boldsymbol{A}^{-1}$。由此定义可知，逆矩阵的逆矩阵还是矩阵本身，即 $(\boldsymbol{A}^{-1})^{-1}=\boldsymbol{A}$。方阵 $\boldsymbol{A}$ 可逆的充分必要条件为其行列式 $|\boldsymbol{A}|\neq 0$。而且，如果 $\boldsymbol{A}$ 可逆，则其逆矩阵 $\boldsymbol{A}^{-1}$ 是唯一的。假设方程(3.22)中的矩阵 $\boldsymbol{A}$ 可逆，则在该方程两边同时左乘其逆矩阵 $\boldsymbol{A}^{-1}$ 可得：

$$\boldsymbol{A}^{-1}\boldsymbol{Ax} = \boldsymbol{A}^{-1}\boldsymbol{b} \Rightarrow \boldsymbol{Ix} = \boldsymbol{A}^{-1}\boldsymbol{b} \Rightarrow \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{3.23}$$

矩阵求逆满足以下规则：

(1) $(\boldsymbol{A}^{-1})'=(\boldsymbol{A}')^{-1}$　　(求逆与转置可交换次序)

(2) $(\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1},(\boldsymbol{ABC})^{-1}=\boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$　　(求逆与乘积的混合运算)

### 10. 矩阵的秩

考虑两个 $n$ 维列向量 $\boldsymbol{a}_1$ 与 $\boldsymbol{a}_2$。如果 $\boldsymbol{a}_1$ 正好是 $\boldsymbol{a}_2$ 的固定倍数，则在向量组 $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ 中，真正含有信息的只是其中的一个向量。更一般地，考虑由 $K$ 个 $n$ 维向量构成的向量组 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$，如果存在 $c_1, c_2, \cdots, c_K$ 不全为零，使得

$$c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_K\boldsymbol{a}_K = \boldsymbol{0} \tag{3.24}$$

则称向量组 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ **线性相关**(linearly dependent)。显然，如果 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ 线性相关，则其中至少有一个向量可写为其他向量的**线性组合**(linear combination)，也称**线性表出**。反之，如果方程(3.24)成立必须 $c_1 = c_2 = \cdots = c_K = 0$，则称 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ **线性无关**(linearly independent)。

进一步，如果 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ 线性相关，但从中去掉一个向量后，就变得线性无关，则 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ 中正好有 $(K-1)$ 个向量真正含有信息，称 $(K-1)$ 为此向量组的秩。更一般地，向量组 $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}$ 的极大线性无关部分组所包含的向量个数，称为该**向量组的秩**(rank)。

对于 $m \times n$ 矩阵 $\boldsymbol{A}$，可将其 $n$ 个列向量看成一个向量组，称此列向量组的秩为矩阵 $\boldsymbol{A}$ 的**列秩**(column rank)。如果矩阵 $\boldsymbol{A}_{m\times n}$ 的列秩正好等于 $n$，则称矩阵 $\boldsymbol{A}$ **满列秩**(full column rank)。类似地，可将矩阵 $\boldsymbol{A}_{m\times n}$ 的 $m$ 个行向量看成一个向量组，称此行向量组的秩为矩阵 $\boldsymbol{A}$ 的**行秩**(row rank)。可以证明，任何矩阵的行秩与列秩一定相等，称为**矩阵的秩**(matrix rank)。

### 11. 二次型

对于 $n$ 维列向量 $\boldsymbol{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)'$，如何度量它与零向量 $\boldsymbol{0}$(即原点)的距离？最简单的方法为欧几里得距离(Euclidean distance)的平方，即内积

$$x_1^2 + x_2^2 + \cdots + x_n^2 = (x_1 \quad x_2 \quad \cdots \quad x_n)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \boldsymbol{x}'\boldsymbol{x} \tag{3.25}$$

注意到上式可以写为

$$(x_1 \quad x_2 \quad \cdots \quad x_n)\underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{\boldsymbol{I}_n}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \boldsymbol{x}'\boldsymbol{I}_n\boldsymbol{x} \tag{3.26}$$

方程(3.26)中的单位矩阵 $\boldsymbol{I}_n$ 相当于给予此内积的每一项相同的权重。更一般地，如果允许不同的权重，则可使用任意对称矩阵 $\boldsymbol{A}$，构成如下**二次型**(quadratic form)：

$$f(x_1,x_2,\cdots,x_n) = (x_1 \quad x_2 \quad \cdots \quad x_n)\underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{\boldsymbol{A}}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \tag{3.27}$$

其中，对称矩阵 $\boldsymbol{A}$ 称为此二次型的矩阵。由此可知，所谓二次型，就是 $x_1,x_2,\cdots,x_n$ 的二次齐次多项式函数：

$$\begin{aligned} f(x_1,x_2,\cdots,x_n) = {} & a_{11}x_1^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1n}x_1x_n \\ & + a_{22}x_2^2 + \cdots + 2a_{2n}x_2x_n \\ & + \cdots\cdots\cdots\cdots \\ & + a_{nn}x_n^2 \end{aligned} \tag{3.28}$$

反之，任意二次型(3.28)，都可以写为 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 的形式，其中 $\boldsymbol{A}$ 为对称矩阵。例如，考虑一般的二维二次型：

$$f(x_1,x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \tag{3.29}$$

则此二次型可写为：

$$f(x_1,x_2) = (x_1x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{3.30}$$

其中，$a_{21} = a_{12}$。显然，如果 $\boldsymbol{x} = \boldsymbol{0}$，则二次型 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = 0$。更有趣的情形是，当 $\boldsymbol{x} \neq \boldsymbol{0}$ 时，二次型 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 如何取值？首先，考虑一维二次型：

$$f(x_1) = a_{11}x_1^2 = x_1'a_{11}x_1 \tag{3.31}$$

在上式中，二次型的矩阵就是常数 $a_{11}$。如果 $a_{11} > 0$，则只要 $x_1 \neq 0$，就有 $f(x_1) = a_{11}x_1^2 > 0$。此时，称此二次型为“正定”(positive definite)，其图形为开口向上的抛物线(参见图 3.6)①。

反之，如果 $a_{11} < 0$，则只要 $x_1 \neq 0$，就有 $f(x_1) = a_{11}x_1^2 < 0$。此时，称此二次型为“负定”(negative definite)，其图形为开口向下的抛物线(参见图 3.7)②。

对于二维的二次型，其取值的确定性则更为复杂。例如，对于 $x_1,x_2$ 不全为 0，二次型 $(x_1^2 + x_2^2)$ 一定为正，故为正定；二次型 $(-x_1^2 - x_2^2)$ 一定为负，故为负定；而二次型 $(x_1^2 - x_2^2)$ 则可正可负，

---

① 生成图 3.6 的 Stata 命令为“twoway function y = x^2,range( -1 1) xline(0) yline(0) lwidth(thick) xtitle(x1) ytitle(f(x1))”。下文将解释此类 Stata 画图命令。

② 生成图 3.7 的 Stata 命令为“twoway function y = -x^2,range( -1 1) xline(0) yline(0) lwidth(thick) xtitle(x1) ytitle(f(x1))”。

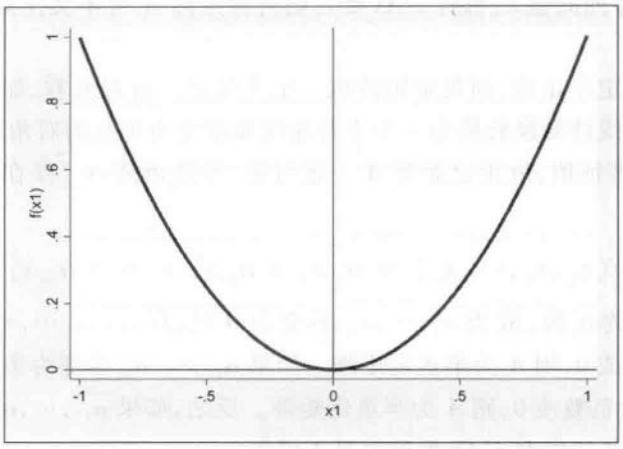



图 3.6 正定的一维二次型

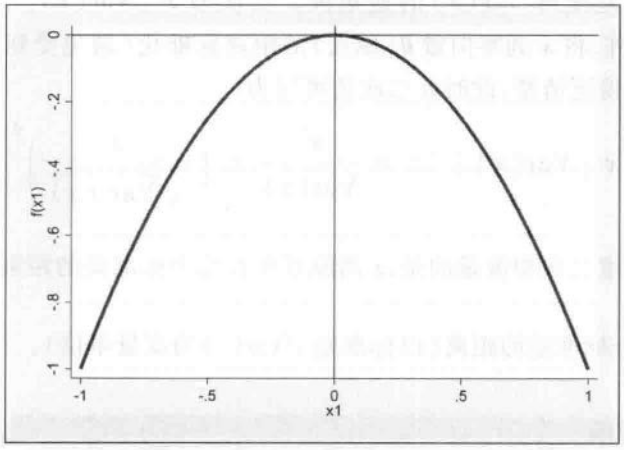



图 3.7 负定的一维二次型

称为“不定”(indefinite)。但这依然没有穷尽所有的情形。考虑以下二次型:

$$f(x_1,x_2) = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2 \tag{3.32}$$

显然,二次型$(x_1+x_2)^2 \geqslant 0$(必然非负);但即使$x_1,x_2$不全为0,也可能出现$(x_1+x_2)^2=0$,只要$x_1=-x_2$;比如,$x_1=1$而$x_2=-1$。此时,称此二次型为“半正定”(positive semidefinite)。另外,二次型$-(x_1+x_2)^2 \leqslant 0$(必然非正);但即使$x_1,x_2$不全为0,也可能出现$(x_1+x_2)^2=0$,只要$x_1=-x_2$。此时,称此二次型为“半负定”(negative semidefinite)。

在一般的$n$维情况下,给定对称矩阵$\boldsymbol{A}$,针对二次型$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$的取值确定性,可引入以下定义。

(1) 对于任意非零列向量$\boldsymbol{x}$,都有$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}>0$,则对称矩阵$\boldsymbol{A}$为**正定矩阵**(positive definite)。

(2) 对于任意非零列向量$\boldsymbol{x}$,都有$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\geqslant 0$,则对称矩阵$\boldsymbol{A}$为**半正定矩阵**(positive semidefinite)。

(3) 对于任意非零列向量$\boldsymbol{x}$,都有$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}<0$,则对称矩阵$\boldsymbol{A}$为**负定矩阵**(negative definite)。

(4) 对于任意非零列向量 $\boldsymbol{x}$,都有 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\leqslant 0$,则对称矩阵 $\boldsymbol{A}$ 为**半负定矩阵**(negative semidefinite)。

显然,正定矩阵一定半正定,而负定矩阵也一定半负定。直观来看,如果对称矩阵 $\boldsymbol{A}$ 为正定矩阵,则该矩阵可通过线性变换转换为一个主对角线元素全为正数的对角矩阵;而这些主对角线元素正好是矩阵 $\boldsymbol{A}$ 的特征值,故正定矩阵 $\boldsymbol{A}$ 一定可逆,即逆矩阵 $\boldsymbol{A}^{-1}$ 存在。线性变换后的正定二次型可写为

$$f(x_1,x_2,\cdots,x_n) = \alpha_{11}x_1^2 + \alpha_{22}x_2^2 + \cdots + \alpha_{nn}x_n^2 \tag{3.33}$$

其中,$\alpha_{11},\cdots,\alpha_{nn}$ 全部为正数,故当 $x_1,\cdots,x_n$ 不全为 0 时,$f(x_1,x_2,\cdots,x_n)$ 必然大于 0。如果 $\alpha_{11},\cdots,\alpha_{nn}$ 全部为正数或 0,则 $\boldsymbol{A}$ 为半正定矩阵。如果 $\alpha_{11},\cdots,\alpha_{nn}$ 全部为负数,则 $\boldsymbol{A}$ 为负定矩阵。如果 $\alpha_{11},\cdots,\alpha_{nn}$ 全部为负数或 0,则 $\boldsymbol{A}$ 为半负定矩阵。反之,如果 $\alpha_{11},\cdots,\alpha_{nn}$ 有正有负,则 $\boldsymbol{A}$ 为不定的(indefinite)矩阵,其二次型 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 的取值可正可负。

在计量经济学中,常使用形如 $\boldsymbol{x}'[\mathrm{Var}(\boldsymbol{x})]^{-1}\boldsymbol{x}$ 的二次型,其中 $\boldsymbol{x}$ 为 $n$ 维随机向量,而 $[\mathrm{Var}(\boldsymbol{x})]^{-1}$ 为其协方差矩阵 $\mathrm{Var}(\boldsymbol{x})$ 的逆矩阵。二次型 $\boldsymbol{x}'[\mathrm{Var}(\boldsymbol{x})]^{-1}\boldsymbol{x}$ 的直观含义是,以 $[\mathrm{Var}(\boldsymbol{x})]^{-1}$ 为权重矩阵,将 $\boldsymbol{x}$ 到零向量 $\boldsymbol{0}$(原点)的距离标准化(避免受到 $\boldsymbol{x}$ 度量单位的影响)。在一维情况下,可以看得更清楚,此时此二次型可写为

$$\boldsymbol{x}'[\mathrm{Var}(\boldsymbol{x})]^{-1}\boldsymbol{x} = \frac{x^2}{\mathrm{Var}(x)} = \left(\frac{x}{\sqrt{\mathrm{Var}\ (x)}}\right)^2 \tag{3.34}$$

从上式可知,此一维二次型度量的是,$x$ 离原点 0 有几个标准差的距离。比如,$\dfrac{x}{\sqrt{\mathrm{Var}(x)}}=2$,则 $x$ 离原点 0 有两个标准差的距离(以标准差 $\sqrt{\mathrm{Var}(x)}$ 为度量单位)。

## 3.3 概率与条件概率

### 1. 概率

假如街上有位老太太问你:“什么是概率”,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 $P(A)$。

### 2. 条件概率

**例** 已知明天会出太阳,则下雨的概率有多大?

记事件“出太阳”为 $B$,则在出太阳的前提条件下,降雨的**条件概率**(conditional probability)为

$$P(A \mid B) \equiv \frac{P(AB)}{P(B)} \tag{3.35}$$

其中，$AB$ 表示事件 $A$ 与 $B$ 同时发生（即交集，也记为 $A \cap B$），故 $P(AB)$ 为“太阳雨”的概率，参见图 3.8。条件概率是计量经济学的重要概念之一。

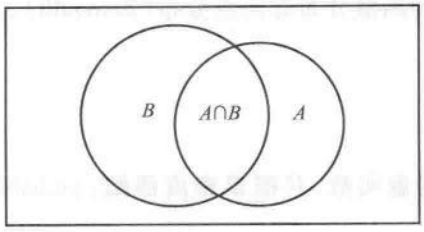



图 3.8 条件概率的示意图

**例** 股市崩盘的可能性为无条件概率；在已知经济已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，$P(A \mid B) = P(A)$，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，$P(A \mid B) \equiv \frac{P(AB)}{P(B)} = P(A)$，故

$$P(AB) = P(A)P(B) \tag{3.36}$$

也可将此式作为独立事件的定义。

**4. 全概率公式**

如果事件组 $\{B_1, B_2, \cdots, B_n\}$ $(n \geqslant 2)$ 两两互不相容，但必有一件事发生，且每件事的发生概率均为正数，则对任何事件 $A$（无论 $A$ 与 $\{B_1, B_2, \cdots, B_n\}$ 是否有任何关系），都有

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \tag{3.37}$$

全概率公式把世界分成了 $n$ 个可能的情形 $\{B_1, B_2, \cdots, B_n\}$，再把每种情况下的条件概率 $P(A \mid B_i)$“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 $P(B_i)$）。该公式有助于理解下文的迭代期望定律。

## 3.4 分布与条件分布

*A random variable is the soul of an observation... An observation is the birth of a random variable.*
——D. W. Watts

**1. 离散型概率分布**

假设随机变量 $X$ 的可能取值为 $\{x_1, x_2, \cdots, x_k, \cdots\}$，其对应概率为 $\{p_1, p_2, \cdots, p_k, \cdots\}$，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

$$\begin{array}{cccccc} X & x_1 & x_2 & \cdots & x_k & \cdots \\ p & p_1 & p_2 & \cdots & p_k & \cdots \end{array} \tag{3.38}$$

其中，$p_k \geqslant 0$，$\sum\limits_k p_k = 1$。常见的离散分布有两点分布（Bernoulli）、二项分布（Binomial）、泊松分布（Poisson）等。

**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：

（1）$f(x) \geqslant 0, \forall x$；

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

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

概率密度函数的示意图参见图 3.9。

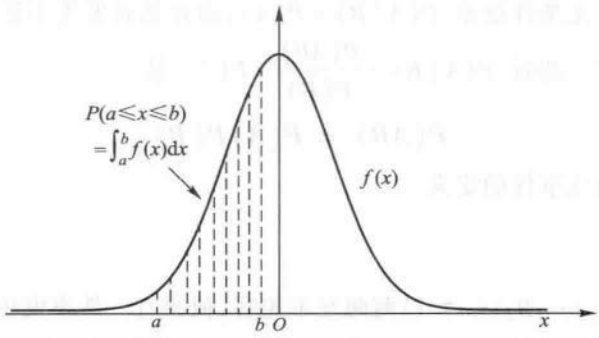



图 3.9 概率密度函数的示意图

定义**累积分布函数**（cumulative distribution function，简记 cdf）：

$$F(x) \equiv P(-\infty < X \leqslant x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \tag{3.39}$$

其中，$t$ 为积分变量。直观来看，$F(x)$ 度量的是，从 $-\infty$ 至 $x$ 为止，概率密度函数 $f(t)$ 曲线下的面积，参见图 3.10。

**3. 多维随机向量的概率分布**

为研究变量间的关系，常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）$f(x,y) \geqslant 0, \forall x,y$；

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

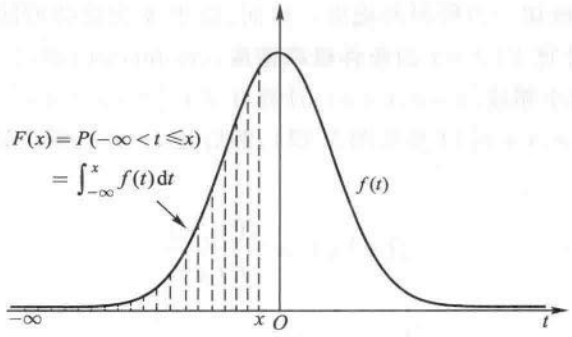



图 3.10 累积分布函数的示意图

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

二维随机向量的联合密度函数就像倒扣的草帽。落入平面某区域 $D$ 的概率就是此草帽下区域 $D$ 之上的体积,参见图 3.11。更一般地,$n$ 维连续型随机向量$(X_1,X_2,\cdots,X_n)$可由联合密度函数$f(x_1,x_2,\cdots,x_n)$来描述。

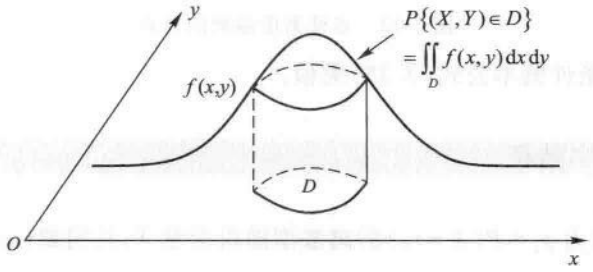



图 3.11 二维联合密度函数的示意图

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

$$f_x(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y \tag{3.40}$$

即给定 $X=x$,把所有 $Y$ 取值的可能性都"加总"起来(积分的本质就是加总)。类似地,可以计算 $Y$ 的(一维)边缘密度函数:

$$f_y(y)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}x \tag{3.41}$$

即给定 $Y=y$,把所有 $X$ 取值的可能性都"加总"起来。定义二维随机向量$(X,Y)$的累积分布函数为:

$$F(x,y)\equiv P(-\infty<X\leqslant x;-\infty<Y\leqslant y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(t,s)\mathrm{d}t\mathrm{d}s \tag{3.42}$$

#### 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 $Y\mid X=x$ 或 $Y\mid x$。对于连续型分布,此条件分布相当于在"草帽"(联合密度函

数)上 $X=x$ 的位置垂直地切一刀所得的截面。然而,由于 $X$ 为连续型随机变量,事件 $\{X=x\}$ 发生的概率为 0。应如何计算 $Y \mid X=x$ 的**条件概率密度**(conditional pdf)?

为此,考虑 $x$ 附近的小邻域 $[x-\varepsilon, x+\varepsilon]$,计算在 $X \in [x-\varepsilon, x+\varepsilon]$ 条件下 $Y$ 的累积分布函数,即 $P\{Y \leqslant y \mid X \in [x-\varepsilon, x+\varepsilon]\}$(参见图 3.12),然后让 $\varepsilon \to 0^+$,则可证明条件密度函数为(陈强,2014,p. 12),

$$f(y \mid x) = \frac{f(x,y)}{f_x(x)} \tag{3.43}$$

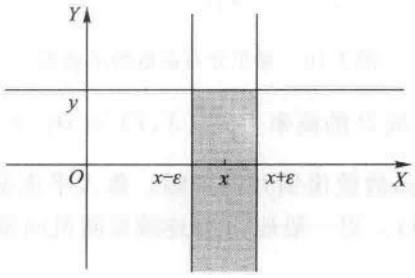



图 3.12 条件密度函数的计算

直观上,此公式与条件概率公式(3.35)类似。

## 3.5 随机变量的数字特征

**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)为

$$\mathrm{E}(X) \equiv \mu \equiv \sum_{k=1}^{\infty} x_k p_k \tag{3.44}$$

由上式可知,期望的直观含义就是对 $x_k$ 进行加权平均,而权重为概率 $p_k$。

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

$$\mathrm{E}(X) \equiv \mu \equiv \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \tag{3.45}$$

直观上,上式也是对 $x$ 进行加权平均,而权重为概率密度 $f(x)$。有时称求期望这种运算为**期望算子**(expectation operator)。容易证明,期望算子满足**线性性**(linearity),即对于任意常数 $k$ 都有

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y), \quad \mathrm{E}(kX) = k\mathrm{E}(X) \tag{3.46}$$

**定义** 随机变量 $X$ 的**方差**(variance)为

$$\mathrm{Var}(X) \equiv \sigma^2 \equiv \mathrm{E}[X - \mathrm{E}(X)]^2 \tag{3.47}$$

方差越大,则随机变量取值的波动幅度越大。称方差的平方根为**标准差**(standard deviation),通常记为 $\sigma$。在计算方差时,常利用以下简便公式(参见习题):

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \tag{3.48}$$

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

**定义** 随机变量 $X$ 与 $Y$ 的**协方差**(covariance)为

$$\mathrm{Cov}(X,Y) \equiv \sigma_{XY} \equiv \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] \tag{3.49}$$

如果当随机变量 $X$ 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时，随机变量 $Y$ 的取值也倾向于大于(小于)其期望值 $\mathrm{E}(Y)$，则 $\mathrm{Cov}(X,Y)>0$，二者存在**正相关**；反之，如果当随机变量 $X$ 的取值大于(小于)其期望 $\mathrm{E}(X)$ 时，随机变量 $Y$ 的取值反而倾向于小于(大于)其期望值 $\mathrm{E}(Y)$，则 $\mathrm{Cov}(X,Y)<0$，二者存在**负相关**。如果 $\mathrm{Cov}(X,Y)=0$，则说明二者**线性不相关**(uncorrelated)，但不一定**相互独立**(independent)，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(参见习题)：

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{3.50}$$

协方差的运算也满足线性性，可以证明(参见习题)：

$$\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z) \tag{3.51}$$

协方差的缺点是，它受 $X$ 与 $Y$ 计量单位的影响。为将其标准化，引入相关系数的定义。

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

$$\rho \equiv \mathrm{Corr}(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{3.52}$$

可以证明，相关系数一定介于 $-1$ 与 $1$ 之间，即 $-1\leqslant\rho\leqslant1$。需要注意的是，如果以上各定义式中的积分不收敛，则随机变量的数字特征可能不存在。比如，自由度为 1 的 $t$ 分布变量，其期望与方差都不存在。更一般地，对于随机变量 $X$，可以定义一系列的数字特征，即各阶**矩**(moment)的概念。

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

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

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

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

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

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

对于正态分布，其峰度为 3。如果随机变量 $X$ 的峰度大于 3(比如 $t$ 分布)，则其密度函数的最高处(山峰)比正态分布更“尖”，而两侧尾部更“厚”，称为“厚尾”(fat tails)。存在厚尾的概率分布更容易在尾部取值，称为**极端值**(outlier)，参见图 3.13。

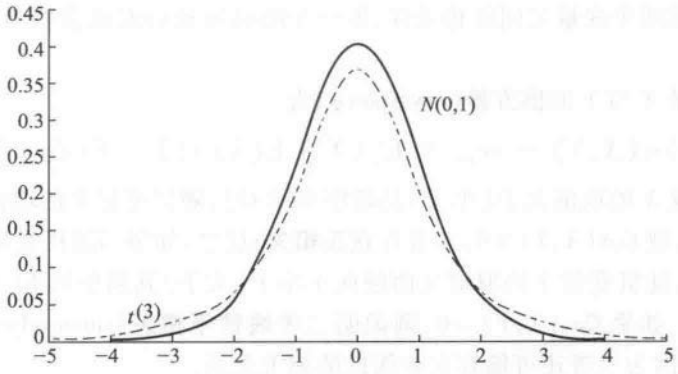



图 3.13 $N(0,1)$与$t(3)$的概率密度

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

由于正态分布的偏度为0,峰度为3,故可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,对于随机变量 $X$ 与任意函数 $g(\cdot)$,称随机变量函数 $g(X)$ 的期望 $\mathrm{E}[g(X)]=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x$ 为**矩**(moment)。

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

$$\mathrm{E}(Y|X=x)\equiv\mathrm{E}(Y|x)=\int_{-\infty}^{+\infty}yf(y|x)\mathrm{d}y \tag{3.53}$$

在上式中,由于 $y$ 已被积分积掉,故 $\mathrm{E}(Y|x)$ 只是 $x$ 的函数,参见图 3.14。

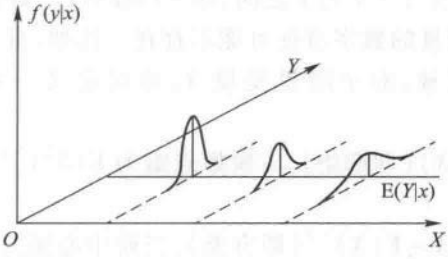



图 3.14 条件期望与条件方差示意图

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

$$\mathrm{Var}(Y|X=x)\equiv\mathrm{Var}(Y|x)=\int_{-\infty}^{+\infty}[y-\mathrm{E}(Y|x)]^2f(y|x)\mathrm{d}y \tag{3.54}$$

同样地,在上式中,$y$ 已被积分积掉,故 $\mathrm{Var}(Y|x)$ 也只是 $x$ 的函数,参见图 3.14。

**定义** 设 $\boldsymbol{X}=(X_1\ X_2\cdots X_n)'$ 为 $n$ 维随机向量,则其**协方差矩阵**(covariance matrix)为 $n\times n$ 的对称矩阵:

$$\mathrm{Var}(\boldsymbol{X})\equiv\mathrm{E}[(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))']$$

$$
\begin{aligned}
&= \mathrm{E}\left[\begin{pmatrix} X_1 - \mathrm{E}(X_1) \\ \vdots \\ X_n - \mathrm{E}(X_n) \end{pmatrix} (X_1 - \mathrm{E}(X_1) \quad \cdots \quad X_n - \mathrm{E}(X_n))\right] \\
&= \mathrm{E}\begin{pmatrix} [X_1 - \mathrm{E}(X_1)]^2 & \cdots & [X_1 - \mathrm{E}(X_1)][X_n - \mathrm{E}(X_n)] \\ \vdots & & \vdots \\ [X_1 - \mathrm{E}(X_1)][X_n - \mathrm{E}(X_n)] & \cdots & [X_n - \mathrm{E}(X_n)]^2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}
\end{aligned}
\tag{3.55}
$$

其中，主对角线元素 $\sigma_{ii} \equiv \mathrm{Var}(X_i)$，非主对角线元素 $\sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j)$。可以证明，协方差矩阵必然为半正定矩阵（positive semidefinite）。在一维情况下，这意味着随机变量的方差必然为非负。

对于随机向量 $\boldsymbol{X}$ 的期望与协方差矩阵的运算，有如下重要法则。假设 $\boldsymbol{A}$ 为 $m \times n$ 常数矩阵（不含随机变量），则可以证明：

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

（2）$\mathrm{Var}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{XX}') - \mathrm{E}(\boldsymbol{X})[\mathrm{E}(\boldsymbol{X})]'$　　（一维公式的推广）

（3）$\mathrm{Var}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}'$　　（夹心估计量）

如果 $\boldsymbol{A}$ 为对称矩阵，则 $\mathrm{Var}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}$，称为**夹心估计量**（sandwich estimator），其中两边的 $\boldsymbol{A}$ 为“面包”，而夹在中间的 $\mathrm{Var}(\boldsymbol{X})$ 为“菜”，在形式上类似于三明治。

以上随机变量的数字特征都可视为“总体矩”（population moment）。在抽取随机样本后，可用样本数据计算相应的“样本矩”（sample moment），作为相应总体矩的估计值。这意味着以“求样本平均值运算”$\left(\frac{1}{n}\sum_{i=1}^{n}(\cdot)\right)$来替代总体矩表达式的期望算子 $\mathrm{E}(\cdot)$。比如，可用样本均值（sample mean）$\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)$来估计总体均值（population mean）或期望 $\mathrm{E}(X)$。

以数据集 grilic. dta 为例。该数据集截取自 Griliches（1976）对教育投资回报率的经典研究，由 Blackburn and Neumark（1992）更新数据，包括 758 名美国年轻男子的数据。此数据集比第 2 章 grilic_small. dta 包含更多的变量与观测值（后者是前者的子集）。

首先，打开此数据集，看一下它所包含的变量情况。

```
. use grilic.dta,clear
. describe
```

```
Contains data from D:\desktop\±¾¿Î³Ì¼ÆÁ¿\2014\¼ÆÁ¿¾¼Ã¼°StataÓ¦ÓÃ\Data\grilic.dta
  obs:           758
 vars:            11                          15 Sep 2014 07:21
 size:        13,644
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
rns             byte    %8.0g                 south = 1
mrt             byte    %8.0g                 married = 1
smsa            byte    %8.0g                 big cities =1
med             byte    %8.0g                 mother's education
iq              int     %8.0g                 IQ
kww             byte    %8.0g                 KWW
age             byte    %8.0g                 age
s               byte    %8.0g                 schooling
expr            float   %9.0g                 experience
tenure          byte    %8.0g                 tenure
lnw             float   %9.0g                 ln(wage)
-------------------------------------------------------------------------------
Sorted by:
```

从上表可知，此数据集的样本容量为758，其中被解释变量为 ln $w$（工资对数），主要解释变量包括 $s$（教育年限）、*expr*（工龄）、*tenure*（在现单位工作年限）、*age*（年龄）、*iq*（智商）、*kww*（在KWW（Knowledge of the World of Work）测试的成绩）、*med*（母亲的教育年限）、*mrt*（婚否）、*smsa*（是否住在大城市）以及 *rns*（是否住在美国南方）。

在上表第一行路径名中有些汉字出现了乱码。如果希望在 Stata 中正确显示汉字，可通过修改结果窗口的背景颜色来实现，即点击菜单“Edit”→“Preferences”→“General Preferences”→“Result Colors”→“Color Scheme”，然后选择“Classic”（将背景颜色设为黑色）即可。

下面看一下各变量的基本统计指标。

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| rns | 758 | .2691293 | .4438001 | 0 | 1 |
| mrt | 758 | .5145119 | .5001194 | 0 | 1 |
| smsa | 758 | .7044855 | .456575 | 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 |
| age | 758 | 21.83509 | 2.981756 | 16 | 30 |
| s | 758 | 13.40501 | 2.231828 | 9 | 18 |
| expr | 758 | 1.735429 | 2.105542 | 0 | 11.444 |
| tenure | 758 | 1.831135 | 1.67363 | 0 | 10 |
| lnw | 758 | 5.686739 | .4289494 | 4.605 | 7.051 |

如果想看 ln $w$ 的更多统计指标，比如偏度、峰度，可加上选择项“detail”：

```
. sum lnw,detail
```

```
                          ln(wage)
-------------------------------------------------------------
      Percentiles      Smallest
 1%       4.804          4.605
 5%       5.011          4.605
10%       5.165          4.654      Obs                 758
25%        5.38          4.718      Sum of Wgt.         758

50%       5.684                     Mean           5.686739
                        Largest     Std. Dev.      .4289494
75%       5.991          6.786
90%       6.252          6.844      Variance       .1839976
95%       6.399          6.869      Skewness       .1744968
99%       6.706          7.051      Kurtosis       2.73237
```

下面，通过画 ln $w$ 的直方图来看其(无条件)分布，结果参见图 3.15。

```
. hist lnw,width(0.1)
(bin=25, start=4.605, width=.1)
```

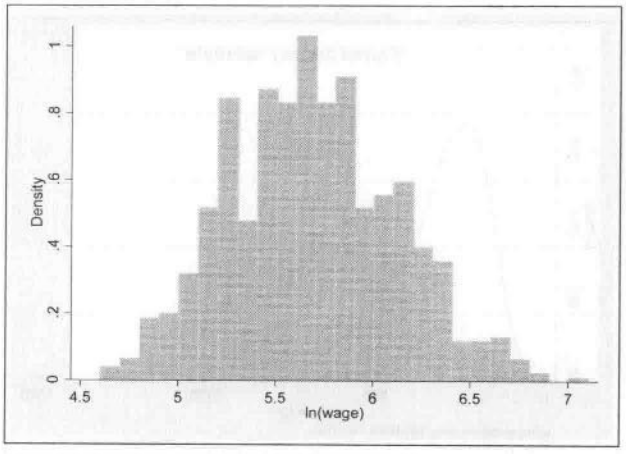



图 3.15 工资对数的直方图

然而，直方图必然是不连续的。如果想得到概率密度函数的连续估计，可输入命令(结果参见图 3.16)

```
. kdensity lnw,normal normop(lpattern(dash))
```

其中，“kdensity”表示核密度估计(kernel density estimation)[①]，选择项“normal”表示画正态分布的密度函数作为对比，而选择项“normop(lpattern(dash))”则指示将正态密度用虚线(dash)来画(其中，normop 表示 normal options；而 lpattern 表示 line pattern)。

从图 3.16 可知，工资对数的分布接近于正态分布，也基本为对称分布。作为对比，下面考察工资水平本身的分布，结果参见图 3.17。

```
. gen wage=exp(lnw)
. kdensity wage
```

① 参见陈强(2014，p.518)。

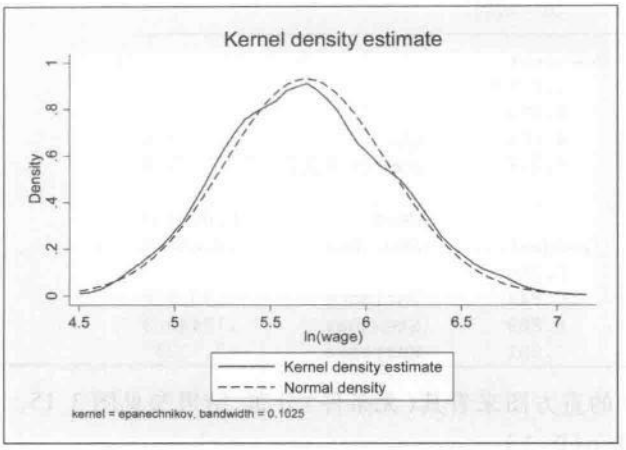



图 3.16 工资对数的核密度估计

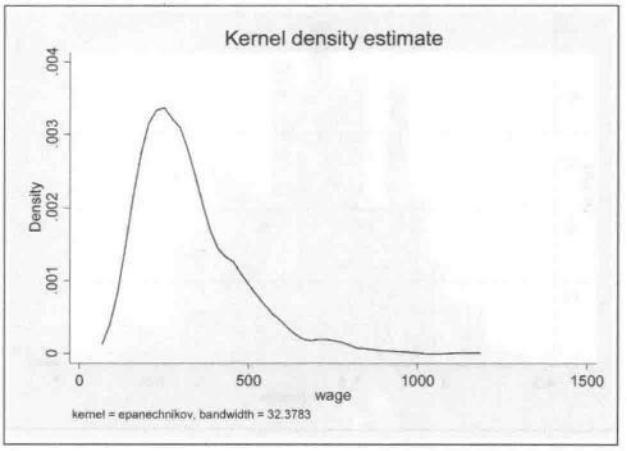



图 3.17 工资的核密度估计

从图 3.17 可知,一方面,工资水平的分布相去正态分布甚远,为非对称分布,在右边存在很长的尾巴,称为“向右偏”。另一方面,工资对数的分布则很接近正态分布,这是使用工资对数作为被解释变量的原因之一。此例也提示我们,对于取值为正的非对称分布,有时可通过取对数使其变得更为对称,也更接近于正态分布。

以上考察的均为无条件分布以及无条件期望等。下面考察给定教育年限情况下的条件分布,比如给定教育年限为 16 年(大学毕业),工资对数的条件密度(结果参见图 3.18)。

```
. kdensity lnw if s ==16
```

为便于比较,下面将 $\ln w$ 的无条件密度与条件密度画在一起(结果参见图 3.19):

```
. twoway kdensity lnw || kdensity lnw if s ==16,lpattern(dash)
```

其中,“||”为分隔符(separator)。分隔符“||”的作用,也可以通过两个括号“( ) ( )”来等价地实现,比如:

```
. twoway (kdensity lnw) (kdensity lnw if s ==16,lpattern(dash))
```

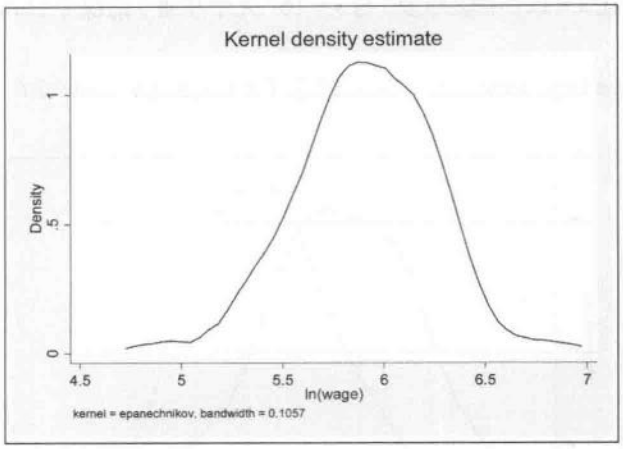



图 3.18 给定 s = 16 的工资对数条件密度

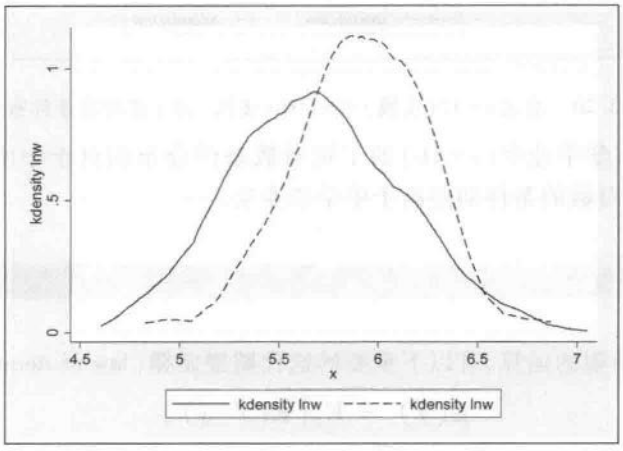



图 3.19 给定 $s = 18$ 的工资对数条件密度

从图 3.19 可清楚看出，给定 $s = 16$ 的工资对数条件密度（图中虚线），明显比工资对数的无条件密度向右移，故条件期望增大，而条件方差则似乎也变小。下面，比较 $\ln w$ 的无条件期望、方差与条件期望、条件方差。

```
. sum lnw
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lnw | 758 | 5.686739 | .4289494 | 4.605 | 7.051 |

```
. sum lnw if s ==16
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lnw | 151 | 5.907338 | .3396442 | 4.828 | 6.869 |

从以上两表可知，条件期望为 5.91，大于无条件期望 5.69；而条件标准差为 0.34，小于无条件标准差 0.43。

进一步，下面比较在 $s=12$（中学毕业）与 $s=16$（大学毕业）情况下，$\ln w$ 的条件密度（结果参见图 3.20）。

```
. twoway (kdensity lnw if s==12) (kdensity lnw if s==16, lpattern(dash))
```

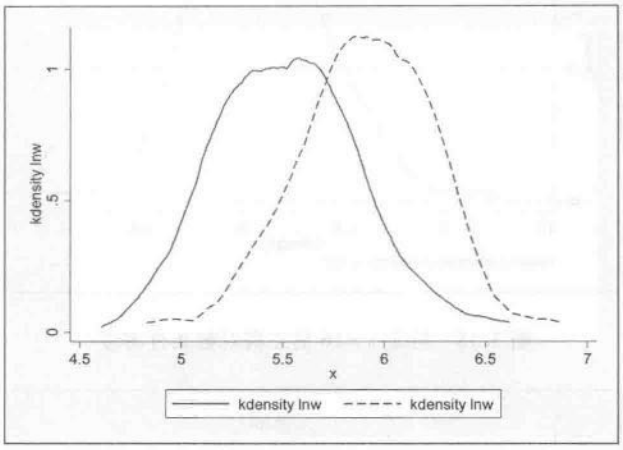



图 3.20 给定 $s=12$（实线）与 $s=16$（虚线）的工资对数条件密度

从图 3.20 可知，大学毕业生（$s=16$）的工资对数条件分布相对于中学毕业生（$s=12$）向右移，故大学毕业生工资对数的条件期望高于中学毕业生。

## 3.6 迭代期望定律

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

$$\mathrm{E}(Y) = \mathrm{E}_X[\mathrm{E}(Y \mid x)] \tag{3.56}$$

上式表明，无条件期望 $\mathrm{E}(Y)$ 等于，给定 $X=x$ 情况下 $Y$ 的条件期望 $\mathrm{E}(Y \mid x)$（仍为 $x$ 的函数），再对 $X$ 求期望。如果 $X$ 为离散随机变量，则根据期望定义，上式可写为：

$$\mathrm{E}(Y) = \sum_i P(X = x_i)\mathrm{E}(Y \mid x_i) \tag{3.57}$$

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（取值可能性）。下面以数据集 grilic. dta 为例，来验证迭代期望定律，即

$$\mathrm{E}(\ln w) = \mathrm{E}_{rns}[\mathrm{E}(\ln w \mid rns)] \tag{3.58}$$

其中，$rns$ 为美国南方居民的虚拟变量，取值为 0 或 1。首先，计算 $rns=0$ 情况下，$\ln w$ 的条件期望：

```
. use grilic.dta, clear
. sum lnw if rns==0
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lnw | 554 | 5.725644 | .4129207 | 4.605 | 7.051 |

由上表可知,北方居民有 554 位,其条件期望 $E(\ln w \mid rns=0)=5.725\,644$。

其次,计算 $rns=1$ 情况下,$\ln w$ 的条件期望:

```
. sum lnw if rns ==1
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lnw | 204 | 5.581083 | .4542189 | 4.718 | 6.844 |

由上表可知,南方居民有 204 位,其条件期望 $E(\ln w \mid rns=1)=5.581\,083$,故美国南方居民的工资略低于北方。下面,以北方与南方居民所占比重作为权重,将北方与南方居民的平均工资对数进行加权平均。

```
. dis 5.725644 *(554 /(554 +204)) +5.581083 *(204 /(554 +204))
5.6867384
```

最后,用命令 summarize 直接计算无条件期望。

```
. sum lnw
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lnw | 758 | 5.686739 | .4289494 | 4.605 | 7.051 |

显然,二者的结果完全相等(忽略计算误差),从而验证了等式(3.58)成立。

下面以离散型变量为例,来严格地证明等式(3.57)。

假设 $X$ 的可能取值为 $x_1,x_2,\cdots,x_i,\cdots$,而 $Y$ 的可能取值为 $y_1,y_2,\cdots,y_j,\cdots$。记 $p_i\equiv P(X=x_i)$,$q_j\equiv P(Y=y_j)$,而 $p_{ij}\equiv P(X=x_i,Y=y_j)$。

**证明**:从等式(3.57)的右边开始证明。

$$
\begin{aligned}
E_X[E(Y\mid x)] &= \sum_i P(X=x_i)E(Y\mid x_i) \quad \text{(期望的定义式)}\\
&= \sum_i P(X=x_i)\Big[\sum_j P(Y=y_j\mid x_i)\cdot y_j\Big] \quad \text{(条件期望的定义式)}\\
&= \sum_i \cancel{P(X=x_i)}\Big[\sum_j \frac{P(X=x_i,Y=y_j)}{\cancel{P(X=x_i)}}\cdot y_j\Big] \quad \text{(条件概率的定义式)}\\
&= \sum_i \Big[\sum_j P(X=x_i,Y=y_j)\cdot y_j\Big] \quad \text{(消去 } P(X=x_i)\text{)}\\
&= \sum_j \Big[\sum_i P(X=x_i,Y=y_j)\cdot y_j\Big] \quad \text{(交换加总的次序)}\\
&= \sum_j \Big[y_j\sum_i P(X=x_i,Y=y_j)\Big] \quad (y_j \text{ 与 } i \text{ 无关,可提出})\\
&= \sum_j y_jP(Y=y_j) \quad \text{(边缘概率与联合概率的关系)}
\end{aligned}
$$

$$= \mathrm{E}(Y) \quad \text{（期望的定义式）}$$

迭代期望定律很像全概率公式：无条件期望等于条件期望之加权平均，权重为条件概率密度。对于连续型变量，可以类似地证明，参见陈强(2014，p. 8)。

将迭代期望定律(3.56)推而广之，对于任意函数 $g(\cdot)$，可以得到

$$\mathrm{E}[g(Y)] = \mathrm{E}_X\mathrm{E}[g(Y) \mid x] \tag{3.59}$$

有时期望算子 $\mathrm{E}_X$ 的下标被省去，需注意对什么变量求期望。

## 3.7 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**(independent)。

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

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

一般来说，条件期望 $\mathrm{E}(Y \mid x)$ 是 $x$ 的函数；而在均值独立的情况下，$\mathrm{E}(Y \mid x)$ 不依赖于 $x$，这表明 $Y$ 与 $x$ 的关系并不紧密，处于某种“无关”的状态。

需要注意的是，均值独立不是一种对称的关系，即“$Y$ 均值独立于 $X$”并不意味着“$X$ 均值独立于 $Y$”。

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

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

(2) 假设 $\mathrm{E}(Y \mid x) = \mathrm{E}(Y)$，则显然 $\mathrm{E}(Y \mid x)$ 不依赖于 $X$，故 $Y$ 均值独立于 $X$。

**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 $Y$。

**定理** 如果 $Y$ 均值独立于 $X$，或 $X$ 均值独立于 $Y$，则 $\mathrm{Cov}(X,Y) = 0$。

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

总之，“相互独立”⇒“均值独立”⇒“线性不相关”；反之，则不然。

## 3.8 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、$\chi^2$ 分布、$t$ 分布与 $F$ 分布等。

**1. 正态分布**

如果随机变量 $X$ 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \tag{3.60}$$

则称 $X$ 服从**正态分布**(normal distribution),记为 $X \sim N(\mu,\sigma^2)$,其中 $\mu$ 为期望,$\sigma^2$ 为方差。将 $X$ 进行标准化,定义 $Z \equiv \frac{X-\mu}{\sigma}$,则 $Z$ 服从**标准正态分布**(standard normal distribution),记为 $Z \sim N(0,1)$,其概率密度函数为

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\{-x^2/2\} \tag{3.61}$$

标准正态分布的概率密度以原点为对称,呈钟形(bell-shaped,参见图 3.21),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)。

在 Stata 中,使用函数 `normalden(x)` 与 `normal(x)` 分别表示标准正态的密度函数 $\phi(x)$ 与累积分布函数 $\Phi(x)$。比如,计算标准正态变量小于 1.96 的概率:

```
. dis normal(1.96)
. 9750021
```

如果要画标准正态的密度函数,可输入如下命令(结果参见图 3.21):

```
. twoway function y = normalden(x),range( -5 5) xline(0) ytitle(概率密度)
```

其中,选择项"`range( -5 5)`"表示在横轴区间(-5, 5)上画此图;默认为"`range(0 1)`",即在(0, 1)区间画图。选择项"`xline(0)`"表示在横轴 x = 0 处画一条直线,而选择项"`ytitle(概率密度)`"表示将纵轴的标签设为"概率密度"。

进一步,正态分布 $N(m,s^2)$ 的密度函数可用 `normalden(x,m,s)` 来表示,其中 $m$ 与 $s$ 分别为期望与标准差。下面,将 $N(0,1)$ 与 $N(1,4)$ 的密度函数画在一起(结果参见图 3.22)。

```
. twoway function y = normalden(x),range( -5 10) || function z = normalden(x,1,2),range( -5 10) lpattern(dash) ytitle(概率密度)
```

其中,选择项"`lpattern(dash)`"表示使用虚线画图。

**多维正态分布**:如果 $n$ 维随机向量 $\boldsymbol{X} = (X_1\ X_2 \cdots\ X_n)'$ 的联合密度函数为

$$f(x_1,\cdots,x_n) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\} \tag{3.62}$$

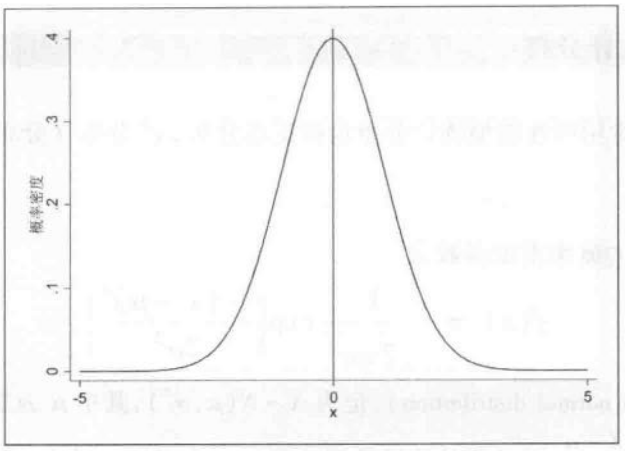



图 3.21 标准正态的概率密度

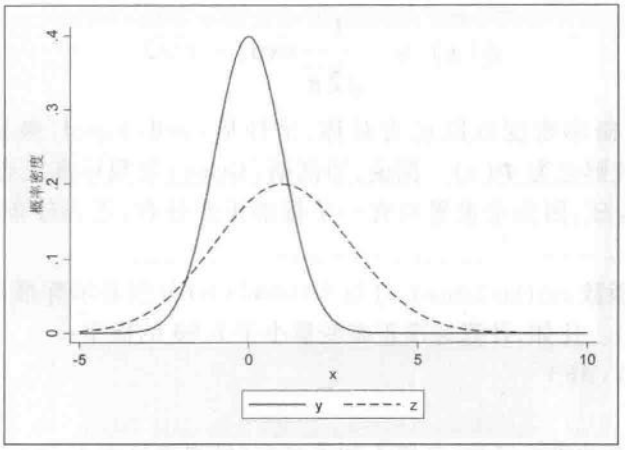



图 3.22 $N(0,1)$与$N(1,4)$的密度函数

则称 $\boldsymbol{X}$ 服从期望为 $\boldsymbol{\mu}$、协方差矩阵为 $\boldsymbol{\Sigma}$ 的 $n$ **维正态分布**，记为 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。在表达式(3.62)中，$(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$为$(\boldsymbol{X}-\boldsymbol{\mu})$的二次型，其二次型矩阵为协方差矩阵的逆矩阵 $\boldsymbol{\Sigma}^{-1}$；$|\boldsymbol{\Sigma}|$为协方差矩阵 $\boldsymbol{\Sigma}$ 的行列式。

多维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。另外，如果$(X_1 \quad X_2 \quad \cdots \quad X_n)$服从 $n$ 维正态分布，则“$X_1, X_2, \cdots, X_n$ 相互独立”与“$X_1, X_2, \cdots, X_n$ 两两不相关”是等价的。利用此性质，有时很容易证明正态变量的独立性。

**2. $\chi^2$ 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 $\chi^2$ 分布。如果$\{Z_1, \cdots, Z_k\}$为独立同分布的标准正态，则其平方和服从自由度为 $k$ 的**卡方分布**，记为

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k) \tag{3.63}$$

其中，参数 $k$ 为**自由度**(degree of freedom)，因为 $\sum_{i=1}^{k} Z_i^2$ 由 $k$ 个相互独立(自由)的随机变量所构成。$\chi^2$ 分布来自标准正态的平方和，故取值为正。可以证明，$\chi^2(k)$ 分布的期望为 $k$，而方差为 $2k$。

在 Stata 中，使用函数 chi2den(k,x)与 chi2(k,x)分别表示自由度为 $k$ 的卡方分布的概率密度与累积分布函数。比如，输入以下命令将 $\chi^2(3)$ 与 $\chi^2(5)$ 的密度函数画在一起(结果参见图 3.23)。

```
. twoway function chi3 = chi2den(3,x),range(0 20) || function chi5 =
chi2den(5,x),range(0 20) lpattern(dash) ytitle(概率密度)
```

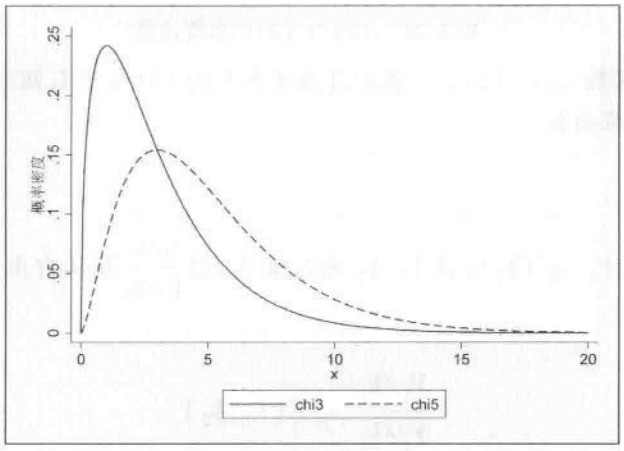



图 3.23 $\chi^2(3)$ 与 $\chi^2(5)$ 的概率密度

**3. $t$ 分布**

假设 $Z \sim N(0,1)$，$Y \sim \chi^2(k)$，且 $Z$ 与 $Y$ 相互独立，则 $\dfrac{Z}{\sqrt{Y/k}}$ 服从自由度为 $k$ 的 $t$ **分布**，记为

$$\frac{Z}{\sqrt{Y/k}} \sim t(k) \tag{3.64}$$

其中，$k$ 为自由度。需要注意的是，如果表达式(3.64)中的分子与分母不相互独立，则一般不服从 $t$ 分布。$t$ 分布也以原点为对称，但与标准正态分布相比，中间的“山峰”更低(但更尖)，而两侧有“厚尾”(fat tails)，参见图 3.13。当自由度 $k \to \infty$ 时，$t$ 分布收敛于标准正态分布。

在 Stata 中，使用函数 tden(k,t)与 t(k,t)分别表示自由度为 $k$ 的 $t$ 分布的概率密度与累积分布函数。比如，使用以下命令将 $t(1)$ 与 $t(5)$ 的密度函数画在一起(结果参见图 3.24)。

```
. twoway function t1 = tden(1,x),range( -5 5) || function t5 = tden(5,
x),range( -5 5) lpattern(dash) ytitle(概率密度)
```

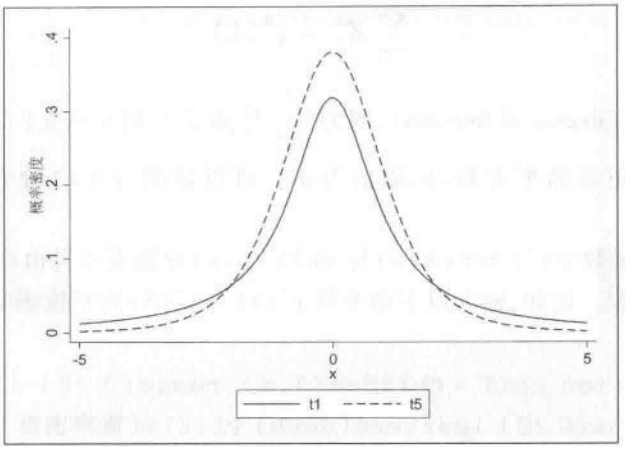



图 3.24 $t(1)$与$t(5)$的密度函数

另外,Stata 还以函数 ttail(k,t)表示自由度为 $k$ 的 $t$ 分布的右侧尾部概率,即 $P(T>t)$,正好是反向的累积分布函数。

**4. $F$ 分布**

假设 $Y_1 \sim \chi^2(k_1)$,$Y_2 \sim \chi^2(k_2)$,且 $Y_1, Y_2$ 相互独立,则$\dfrac{Y_1/k_1}{Y_2/k_2}$服从自由度为 $k_1, k_2$ 的 $F$ 分布,记为

$$\frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2) \tag{3.65}$$

其中 $k_1, k_2$ 为自由度。$F$ 分布的取值也只能为正数,其概率密度的形状与 $\chi^2$ 分布相似。需要注意的是,如果表达式(3.65)中的分子与分母不相互独立,则一般不服从 $F$ 分布。

在 Stata 中,使用函数 Fden(k1,k2,x)与 F(k1,k2,x)分别表示自由度为$(k_1, k_2)$的 $F$ 分布的概率密度与累积分布函数。比如,输入以下命令将 $F(10,20)$与 $F(10,5)$的密度函数画在一起(结果参见图 3.25)。

```
. twoway function F20 = Fden(10,20,x),range(0 5) || function F5 = Fden
(10,5,x),range(0 5) lpattern(dash) ytitle(概率密度)
```

如果想对图 3.25 作进一步的编辑,比如将变量标签改为"$F(10,20)$"与"$F(10,5)$",可在图像上点菜单"File"→"Start Graph Editor",启动 Stata 的图像编辑器,参见图 3.26。

启动图像编辑器之后,直接点击原来的变量标签"F20"与"F5"即可进行编辑,将标签分别改为"F(10, 20)"与"F(10, 5)",结果参见图 3.27。

$F$ 分布与 $t$ 分布存在着密切的关系,因为 $t$ 分布的平方就是 $F$ 分布。

**命题** 如果 $X \sim t(k)$,则 $X^2 \sim F(1,k)$。

**证明**:由于 $X \sim t(k)$,故根据 $t$ 分布的定义,可将 $X$ 写为 $X = \dfrac{Z}{\sqrt{Y/k}} \sim t(k)$,其中 $Z \sim$

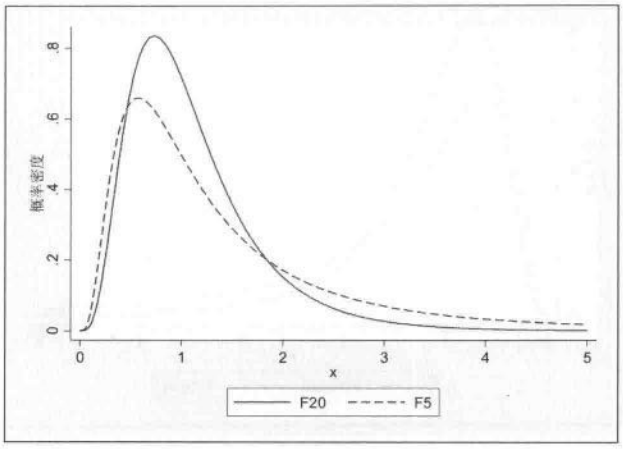



图 3.25 $F(10,20)$ 与 $F(10,5)$ 的概率密度

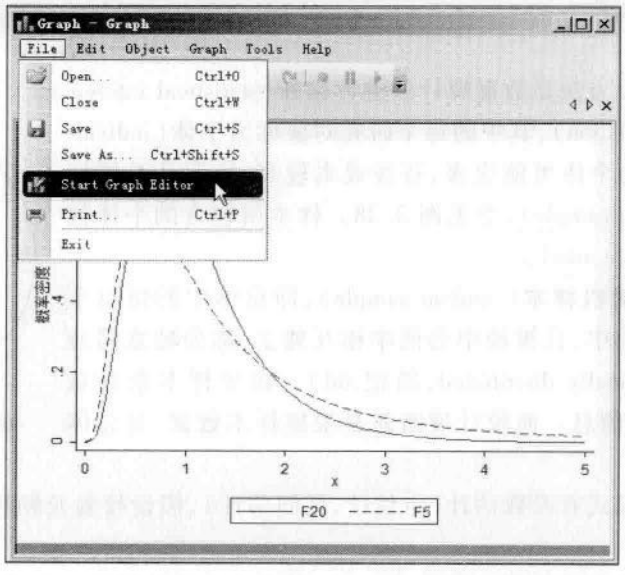



图 3.26 启动 Stata 的图像编辑器

$N(0,1)$，$Y\sim\chi^2(k)$，且 $Z$ 与 $Y$ 相互独立。因此，

$$X^2=\left(\frac{Z}{\sqrt{Y/k}}\right)^2=\frac{Z^2/1}{Y/k}\sim F(1,k) \tag{3.66}$$

其中，由于 $Z\sim N(0,1)$，故 $Z^2\sim\chi^2(1)$；而且，由于 $Z$ 与 $Y$ 相互独立，故 $Z^2$ 也与 $Y$ 相互独立。因此，故根据 $F$ 分布的定义，$X^2$ 服从自由度为 $(1,k)$ 的 $F$ 分布。

更多有关概率分布的 Stata 函数，参见“help density function”。

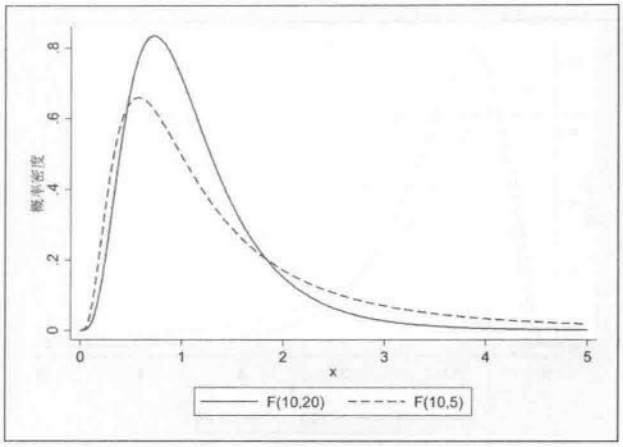



图 3.27 编辑变量标签的结果

## 3.9 统计推断的思想

计量经济学的主要方法是数理统计的**统计推断**(statistical inference)。称我们感兴趣的研究对象全体为**总体**(population),其中的每个研究对象称为**个体**(individual)。由于总体包含的个体可能很多,普查成本较高,故常从总体抽取部分个体,称为**样本**(sample),参见图 3.28。样本所包含的个体数目称为**样本容量**(sample size)。

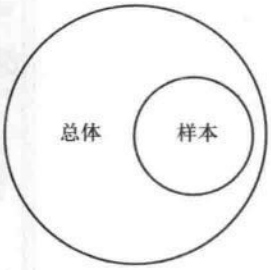



图 3.28 总体与样本

通常希望样本为**随机样本**(random sample),即总体中的每位个体都有相同的概率被抽中,且被抽中的概率相互独立,称为**独立同分布**(independently identically distributed,简记 iid)。由于样本来自总体,故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(点估计、区间估计)、假设检验及预测等,其中点估计为一切统计推断的基础。

假设随机变量 $X$ 的概率密度函数为 $f(x;\theta)$,其中 $\theta$ 为待估参数。为了估计总体参数 $\theta$,从总体中抽取样本容量为 $n$ 的样本数据 $\{x_1,x_2,\cdots,x_n\}$。我们希望根据此样本数据来设计一个性质良好的统计量 $\hat{\theta}(x_1,x_2,\cdots,x_n)$,以此估计 $\theta$。显然,统计量 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 为样本数据 $\{x_1,x_2,\cdots,x_n\}$ 的函数,故仍为随机变量,且随着样本不同而变化。由于使用 $\hat{\theta}$(英文读为"theta hat")来估计 $\theta$,故称 $\hat{\theta}$ 为 $\theta$ 的**估计量**(estimator)。相应地,给定 $\{x_1,x_2,\cdots,x_n\}$ 后,可得到估计量 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 的具体取值,称为**估计值**(estimate)。

比如,$\theta$ 为总体均值,即 $E(X)=\theta$,则一般使用样本均值来估计 $\theta$,即估计量 $\hat{\theta}=\bar{x}\equiv\frac{1}{n}\sum_{i=1}^{n}x_i$。但这并非是唯一的估计量,其他可能的估计量包括第一个观测值 $x_1$,中位数 median$(x_1,x_2,\cdots,$

$x_n)$，最大值 $\max(x_1, x_2, \cdots, x_n)$，最小值 $\min(x_1, x_2, \cdots, x_n)$ 等。

由于潜在的估计量很多（所有样本数据的函数都可视为估计量），故需要有评判估计量优劣的标准。首先，我们希望估计量没有系统性偏差（systematic error），即 $\hat{\theta}$ 不会系统地高估或低估 $\theta$。

**定义** 以估计量 $\hat{\theta}$ 来估计参数 $\theta$，则其**偏差**为 $\text{Bias}(\hat{\theta}) \equiv \text{E}(\hat{\theta}) - \theta$。

**定义** 如果偏差 $\text{Bias}(\hat{\theta}) = 0$，则称 $\hat{\theta}$ 为**无偏估计量**（unbiased estimator）；反之，则称为**有偏估计**（biased estimator），参见图 3.29。

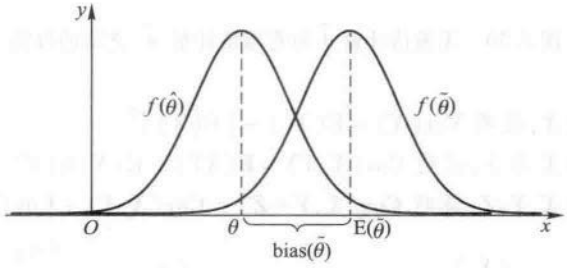



图 3.29 无偏估计量 $\hat{\theta}$ 与有偏估计量 $\tilde{\theta}$ 的概率分布

其次，我们希望抽样误差（sampling error）$(\hat{\theta} - \theta)$ 尽量地小，即 $\hat{\theta}$ 离真实参数 $\theta$ 越近越好。由于 $(\hat{\theta} - \theta)$ 可正可负，故以误差平方（squared error）$(\hat{\theta} - \theta)^2$ 作为度量。但 $\hat{\theta}$ 是随机变量，故引入“均方误差”的概念。

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

在理想的情况下，最优的估计量应在所有估计量中具有最小的均方误差。估计量 $\hat{\theta}$ 的均方误差来源于 $\hat{\theta}$ 的方差与偏差。

**命题** 均方误差可以分解为方差与偏差平方之和，即

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2 \tag{3.67}$$

**证明：**

$$\begin{aligned}\text{MSE}(\hat{\theta}) &\equiv \text{E}[(\hat{\theta} - \theta)^2] = \text{E}\{[\hat{\theta} - \text{E}(\hat{\theta}) + \text{E}(\hat{\theta}) - \theta]^2\} \\ &= \text{E}[\hat{\theta} - \text{E}(\hat{\theta})]^2 + 2\text{E}\{[\hat{\theta} - \text{E}(\hat{\theta})][\text{E}(\hat{\theta}) - \theta]\} + \text{E}[\text{E}(\hat{\theta}) - \theta]^2 \\ &= \text{Var}(\hat{\theta}) + 2\text{E}\{[\hat{\theta} - \text{E}(\hat{\theta})][\text{E}(\hat{\theta}) - \theta]\} + [\text{Bias}(\hat{\theta})]^2\end{aligned}$$

只需证明上式的交叉项为 0 即可：

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

因此，均方误差最小化，可视为在“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 $\hat{\theta}$，如果方差很大，可能不如一个有偏但方差很小的估计量 $\tilde{\theta}$（英文读为 theta tilde），参见图 3.30。

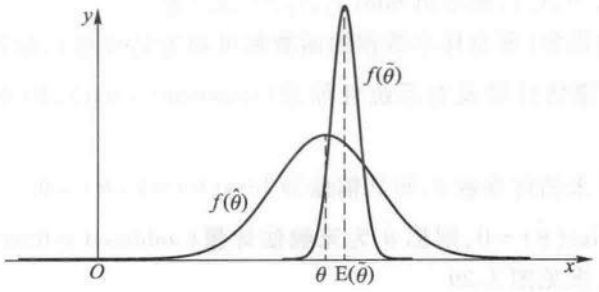



图 3.30 无偏估计量 $\hat{\theta}$ 与有偏估计量 $\tilde{\theta}$ 之间的权衡

**习题**

3.1 对于随机变量 $X$,证明 $\mathrm{Var}(X)=\mathrm{E}(X^2)-[\mathrm{E}(X)]^2$。

3.2 对于随机变量 $X$ 与 $Y$,证明 $\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$。

3.3 对于随机变量 $X,Y,Z$,证明 $\mathrm{Cov}(X,Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z)$。

3.4 二维随机向量 $\boldsymbol{X}=\begin{pmatrix}X_1\\X_2\end{pmatrix}$ 的期望为 $\mathrm{E}(\boldsymbol{X})=\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}$。$\boldsymbol{A}=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\\a_{31}&a_{32}\end{pmatrix}$ 为常数矩阵。证明以下等式。

(1) $\mathrm{E}(\boldsymbol{AX})=\boldsymbol{A\mu}$ (提示:使用期望算子的线性性及矩阵乘法的定义。)

(2) $\mathrm{Var}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{XX}')-\boldsymbol{\mu\mu}'$ (提示:使用协方差矩阵定义,期望与转置算子的线性性。)

(3) $\mathrm{Var}(\boldsymbol{AX})=\boldsymbol{A}\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}'$ (提示:使用协方差矩阵定义,以及(1)的结论。)

3.5 (不相关,但不满足均值独立的例子)假设 $X$ 与 $Z$ 都服从标准正态分布,且相互独立,定义 $Y=X^2+Z$。

(1) 计算 $\mathrm{E}(Y|X)$。该条件期望是否依赖于 $X$?

(2) 计算 $\mathrm{E}(Y)$。条件期望是否等于无条件期望?

(3) 计算 $\mathrm{E}(XY)$。(提示:奇函数在对称区间的积分为 0。)

(4) 证明 $\mathrm{Cov}(X,Y)=0$。

3.6 假设随机变量 $Y$ 服从两点分布,即 $P(Y=1)=p$,而 $P(Y=0)=p$。从 $Y$ 的分布中抽取独立同分布的随机样本 $\{Y_1,\cdots,Y_n\}$。记 $\hat{p}$ 为此样本中成功(即 $Y=1$)的比例。

(1) 证明 $\hat{p}=\overline{Y}\equiv\frac{1}{n}\sum_{i=1}^{n}Y_i$。

(2) 证明估计量 $\hat{p}$ 是 $p$ 的无偏估计。

(3) 证明估计量 $\hat{p}$ 的方差为 $\mathrm{Var}(\hat{p})=\frac{p(1-p)}{n}$。

3.7 假设 $Y_i\sim N(0,\sigma^2)$,且为独立同分布,$i=1,\cdots,n$。

(1) 证明 $\mathrm{E}(Y_i^2/\sigma^2)=1$。(提示:使用公式 $\mathrm{E}(X^2)=\mathrm{Var}(X)+[\mathrm{E}(X)]^2$。)

(2) 证明 $W\equiv\frac{1}{\sigma^2}\sum_{i=1}^{n}Y_i^2$ 服从 $\chi^2(n)$ 分布。

(3) 证明 $E(W)=n$。

(4) 证明 $V \equiv \dfrac{Y_1}{\sqrt{\dfrac{\sum_{i=2}^{n} Y_i^2}{n-1}}}$ 服从 $t(n-1)$ 分布。

*Econometricians have found their Philosophers' Stone; it is called regression analysis and is used for transforming data into significant results.* ——D. F. Hendry

# 4. 一元线性回归

## 4.1 一元线性回归模型

为什么在青少年时期要选择上学？除了满足好奇心、求知欲及个人成长外，一个重要原因是教育能提高未来的收入水平。但如何从理论上解释教育投资的回报率(returns to schooling)？Mincer (1958)率先提出基于效用最大化的理性选择模型[①]，其基本逻辑是：个体选择多上一年学，则需推迟一年挣钱(另外还要交学费)；为了弥补其损失，市场均衡条件要求给予受教育多者更高的未来收入。由此模型，可得工资对数与(受)教育年限的线性关系：

$$\ln w = \alpha + \beta s \tag{4.1}$$

其中，$\ln w$ 为工资对数，$s$ 为教育年限(schooling)，而 $\alpha$ 与 $\beta$ 为参数。其中，$\alpha$ 为截距项，表示当教育年限为 0 时的工资对数水平，因为 $\ln w = \alpha + \beta \cdot 0 = \alpha$。$\beta$ 为斜率，表示教育年限对工资对数的边际效应，即每增加一年教育，将使工资增加百分之几，因为对方程(4.1)两边求导可得

$$\beta = \frac{\mathrm{d}\ln w}{\mathrm{d}s} = \frac{\dfrac{\mathrm{d}w}{w}}{\mathrm{d}s} \approx \frac{\dfrac{\Delta w}{w}}{\Delta s} \tag{4.2}$$

显然，教育年限只是影响工资对数的因素之一，故严格来说，方程(4.1)应写为

$$\ln w = \alpha + \beta s + \text{其他因素} \tag{4.3}$$

将其他因素记为 $\varepsilon$，则有

$$\ln w = \alpha + \beta s + \varepsilon \tag{4.4}$$

方程(4.4)即劳动经济学(labor economics)中著名的明瑟方程(the Mincer equation)的基本形式(Mincer, 1974)。但多上一年学，究竟能使未来收入提高百分之几？这取决于参数 $\beta$ 的取值。而明瑟模型并未提供关于 $\alpha$ 与 $\beta$ 具体取值的信息。一般来说，对于这种定量问题(quantitative question)，只有通过数据才能给出定量回答(quantitative answer)。换言之，需要用计量经济学的方法，通过样本数据来估计未知参数 $\alpha$ 与 $\beta$。

明瑟模型推断工资对数与教育年限存在线性关系，此预言是否与现实数据相符？我们使用数据集 grilic. dta 来考察一下，此数据集包括 758 位美国年轻男子的教育投资回报率数据。为了

① 更准确地说，最大化终生收入的净现值(present value of life earnings)。

得到直观认识,看一下此数据集的变量 $s$ 与 ln $w$ 的前 10 个观测值。

```
. use grilic.dta,clear
. list s lnw in 1/10
```

| | s | lnw |
|---|---|---|
| 1. | 12 | 5.9 |
| 2. | 16 | 5.438 |
| 3. | 14 | 5.71 |
| 4. | 12 | 5.481 |
| 5. | 9 | 5.927 |
| 6. | 9 | 4.804 |
| 7. | 18 | 6.512 |
| 8. | 15 | 5.808 |
| 9. | 12 | 5.737 |
| 10. | 18 | 6.382 |

为了考察工资对数与教育年限的关系,下面画二者的散点图,并在图上画出离这些样本点最近的"回归直线",结果参见图 4.1。

```
. twoway scatter lnw s || lfit lnw s
```

其中,"lfit"表示"linear fit",即线性拟合。由图 4.1 可知,工资对数与教育年限正相关,似乎存在线性关系,在形式上与明瑟方程(4.4)相一致。

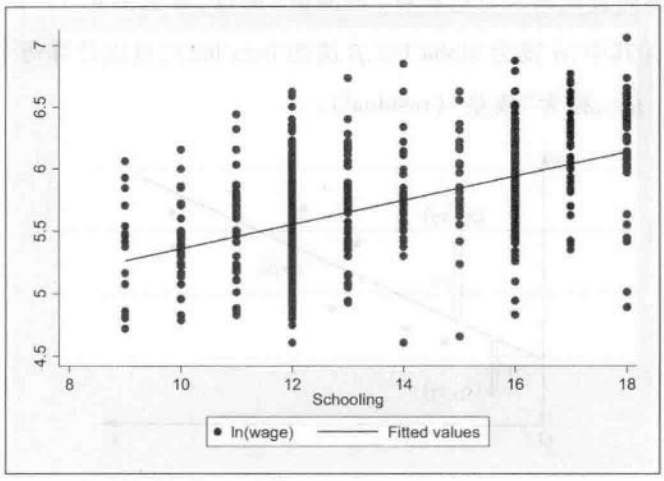



图 4.1 工资对数与教育年限的散点图与线性拟合

更一般地,假设从总体随机抽取了 $n$ 位个体,则一元线性回归模型可写为

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{4.5}$$

其中,$y_i$ 为被解释变量(dependent variable, regressand),$x_i$ 为解释变量(explanatory variable, independent variable, regressor)。$\alpha$ 为截距项(intercept)或常数项(constant),$\beta$ 为斜率(slope),而 $\alpha$ 与 $\beta$ 统称"回归系数"(regression coefficients)或"参数"(parameters),参见图 4.2。$\varepsilon_i$ 为"误差项"(error term)或"扰动项"(disturbance),包括遗漏的其他因素、变量的测量误差、回归函数的设定误差(比如,忽略了非线性项)以及人类行为的内在随机性等。换言之,除 $x_i$ 以外,影响 $y_i$ 的所

有其他因素都在 $\varepsilon_i$ 中。下标 $i$ 表示个体 $i$，比如第 $i$ 个人，第 $i$ 个企业，第 $i$ 个国家等。$i$ 的取值为 $1,\cdots,n$，其中 $n$ 为“样本容量”(sample size)。

方程(4.5)右边的确定性部分为 $\alpha+\beta x_i$，称为**总体回归线**(Population Regression Line)或**总体回归函数**(Population Regression Function，PRF)。方程(4.5)假设总体回归函数为线性，可视为其一阶近似(暂时忽略二次项及高阶项)。

模型 $y_i=\alpha+\beta x_i+\varepsilon_i$ 也称为**数据生成过程**(Data Generation Process，DGP)，参见图 4.2。从数据生成的角度来看，随机变量 $x_i$ 与 $\varepsilon_i$ 首先从相应的概率分布中抽取观测值(observation)，在确定 $x_i$ 与 $\varepsilon_i$ 的取值后，则根据方程 $y_i=\alpha+\beta x_i+\varepsilon_i$ 来生成 $y_i$ 的取值。然而，由于 $\varepsilon_i$ 通常无法观测(unobservable)，故研究者只知道 $(x_i,y_i)$。计量经济学的主要任务之一就是通过数据 $\{x_i,y_i\}_{i=1}^n$ 来获取关于总体参数 $(\alpha,\beta)$ 的信息。

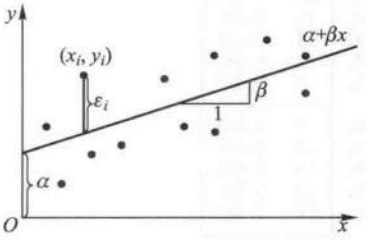


图 4.2 数据生成过程

## 4.2 OLS 估计量的推导

我们的任务是，根据观测值 $\{x_i,y_i\}_{i=1}^n$ 来估计总体回归直线 $\alpha+\beta x_i$。因此，希望在 $(x,y)$ 平面上找到一条直线，使得此直线离所有这些点(观测值)最近，参见图 4.3。在此平面上，任意给定一条直线，$y_i=\hat{\alpha}+\hat{\beta}x_i$(其中，$\hat{\alpha}$ 读为 alpha hat，$\hat{\beta}$ 读为 beta hat)，可以计算每个点(观测值)到这条线的距离，$e_i\equiv y_i-\hat{\alpha}-\hat{\beta}x_i$，称为“残差”(residual)。

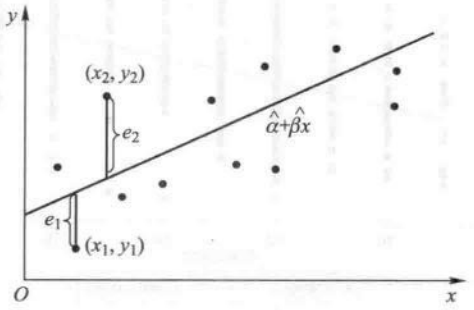


图 4.3 残差平方和最小化

如果直接把残差加起来，即 $\sum_{i=1}^{n}e_i$，则会出现正负相抵的现象。解决方法之一为使用绝对值，即 $\sum_{i=1}^{n}|e_i|=\sum_{i=1}^{n}|y_i-\hat{\alpha}-\hat{\beta}x_i|$。但绝对值不易运算(比如无法微分)，故考虑其平方，$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2$，称为“残差平方和”(Sum of Squared Residuals，SSR；或 Residual Sum of Squares，RSS)。“普通最小二乘法”(Ordinary Least Squares，OLS)就是选择 $\hat{\alpha},\hat{\beta}$，使得残差平方和最小化。在数学上，可将 OLS 的目标函数写为

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{4.6}$$

OLS 是线性回归模型的基本估计方法。根据微积分知识,此最小化问题的一阶条件为

$$\begin{cases} \dfrac{\partial}{\partial \hat{\alpha}} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \dfrac{\partial}{\partial \hat{\beta}} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i) x_i = 0 \end{cases} \tag{4.7}$$

消去方程左边的“-2”可得

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i) x_i = 0 \end{cases} \tag{4.8}$$

对上式各项分别求和,并移项可得

$$\begin{cases} n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{4.9}$$

这是一个有关估计量 $\hat{\alpha},\hat{\beta}$ 的二元一次线性方程组,称为“正规方程组”(normal equations)。从方程组(4.9)的第 1 个方程可得

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4.10}$$

其中,$\bar{y} \equiv \dfrac{1}{n} \sum_{i=1}^{n} y_i$ 为 $y$ 的样本均值,而 $\bar{x} \equiv \dfrac{1}{n} \sum_{i=1}^{n} x_i$ 为 $x$ 的样本均值。将表达式(4.10)代入方程组(4.9)的第 2 个方程可得

$$(\bar{y} - \hat{\beta}\bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{4.11}$$

在上式中,$\hat{\beta}$ 出现在两处。合并同类项,并移项可得

$$\hat{\beta} \left( \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \tag{4.12}$$

使用关系式 $\sum_{i=1}^{n} x_i = n\bar{x}$,求解 $\hat{\beta}$ 可得

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{4.13}$$

上式可写为更为直观的离差形式(参见习题):

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4.14}$$

显然,OLS 估计量要有定义,必须上式的分母 $\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$,这意味着解释变量 $x_i$ 应有所变动,而不能是常数,这是对数据的最基本要求。如果 $x_i$ 没有任何变化,则相同的 $x_i$ 取值将对应于不同的 $y_i$ 取值,故无法估计 $x$ 对 $y$ 的作用,参见图 4.4。

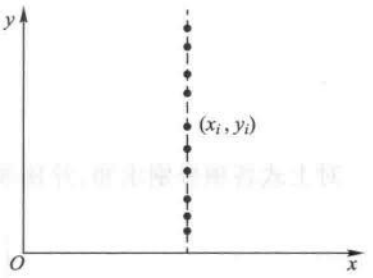


图 4.4 解释变量 $x$ 没有变化的情形

根据方程(4.10)与(4.14),可求解 OLS 估计量 $\hat{\alpha}, \hat{\beta}$,由此得到 $\hat{y} \equiv \hat{\alpha} + \hat{\beta} x_i$,称为**样本回归线**(sample regression line)或**样本回归函数**(Sample Regression Function,SRF),参见图 4.5。从方程(4.10)可知,$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$,即样本回归线一定经过$(\bar{x}, \bar{y})$。

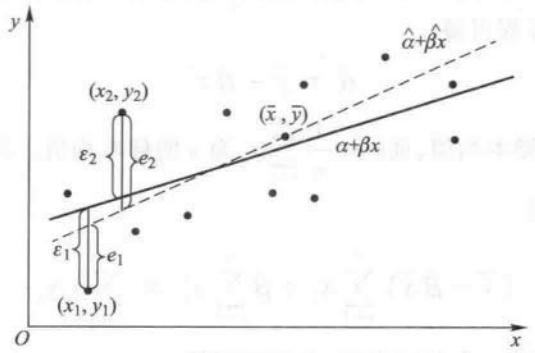


图 4.5 总体回归线与样本回归线

## 4.3 OLS 的正交性

定义被解释变量 $y_i$ 的"拟合值"(fitted value)或"预测值"(predicted value)为

$$\hat{y}_i \equiv \hat{\alpha} + \hat{\beta} x_i \tag{4.15}$$

由此可将残差写为

$$e_i = y - (\hat{\alpha} + \hat{\beta}x_i) = y_i - \hat{y}_i \tag{4.16}$$

根据正规方程组(4.8)可知,

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases} \tag{4.17}$$

将上式写为向量内积的形式:

$$(1 \quad \cdots \quad 1)\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = 0, \quad (x_1 \quad \cdots \quad x_n)\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = 0 \tag{4.18}$$

为了书写方便,分别定义常数向量、残差向量、解释向量以及拟合值向量为

$$\boldsymbol{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}, \quad \boldsymbol{e} \equiv \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} \tag{4.19}$$

则方程(4.18)可写为

$$\boldsymbol{1}'\boldsymbol{e} = 0, \quad \boldsymbol{x}'\boldsymbol{e} = 0 \tag{4.20}$$

由此可知,残差向量 $\boldsymbol{e}$ 与常数向量 $\boldsymbol{1}$ 正交,而且 $\boldsymbol{e}$ 也与解释向量 $\boldsymbol{x}$ 正交。广义地,可以将常数项视为取值都为 1 的解释变量,而 $\alpha$ 为此变量的系数。如此看来,残差向量则与所有解释变量(包括 $\boldsymbol{1}$ 与 $\boldsymbol{x}$)正交。容易证明,残差向量 $\boldsymbol{e}$ 也与拟合值向量 $\hat{\boldsymbol{y}}$ 正交,因为

$$\hat{\boldsymbol{y}}'\boldsymbol{e} \equiv (\hat{y}_1 \quad \cdots \quad \hat{y}_n)\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta}x_i)e_i = \hat{\alpha}\underbrace{\sum_{i=1}^{n} e_i}_{=0} + \hat{\beta}\underbrace{\sum_{i=1}^{n} x_i e_i}_{=0} = 0 \tag{4.21}$$

OLS 残差与解释变量及拟合值的正交性是 OLS 的重要特征,为下面的推导证明提供了方便。比如,考虑方程(4.16),$e_i = y_i - \hat{y}_i$,将两边对 $i$ 加总,并除以 $n$ 可得:

$$0 = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y} - \bar{\hat{y}} \tag{4.22}$$

其中,$\bar{\hat{y}} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$ 为拟合值 $\hat{y}_i$ 的均值。由上式可知,被解释变量的均值恰好等于拟合值的均值,即

$$\bar{y} = \bar{\hat{y}} \tag{4.23}$$

## 4.4 平方和分解公式

从上文可知，被解释变量可分解为相互正交的两个部分，即

$$y_i = \hat{y}_i + e_i \tag{4.24}$$

进一步，如果回归方程有常数项（通常都有常数项），则被解释变量的离差平方和 $\sum_{i=1}^{n}(y_i - \bar{y})^2$（Total Sum of Squares，TSS）可分解为

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{RSS}} \tag{4.25}$$

其中，$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ 为 $y_i$ 的样本均值。方程(4.25)称为“平方和分解公式”，它将 $\sum_{i=1}^{n}(y_i - \bar{y})^2$ 分解为两部分。上式右边第一项为 $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，由于 $\bar{y} = \bar{\hat{y}}$（被解释变量的均值等于拟合值的均值），故可写为 $\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2$，即可由模型解释的部分，称为 Explained Sum of Squares，简记 ESS。上式右边第二项为残差平方和 $\sum_{i=1}^{n} e_i^2$（Residual Sum of Squares，RSS），是模型所无法解释的部分。平方和分解公式能够成立，正是由于 OLS 的正交性。

**证明**：将离差 $(y_i - \bar{y})$ 写为 $(y_i - \hat{y}_i + \hat{y}_i - \bar{y})$，则可将 TSS 写为

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(e_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y})\end{aligned} \tag{4.26}$$

在上式中，只需证明交叉项 $\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = 0$ 即可，而这又由 OLS 的正交性所保证：

$$\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i\hat{y}_i - \sum_{i=1}^{n} e_i = 0 - 0 = 0 \tag{4.27}$$

显然，如果没有常数项，则无法保证 $\sum_{i=1}^{n} e_i = 0$，故平方和分解公式在无常数项的情况下不再成立。

## 4.5 拟合优度

在某种意义上，OLS 的样本回归线为离所有样本点最近的直线。但此最近的直线，究竟离这

些样本点有多近？我们希望有一个绝对的度量，以此衡量样本回归线对数据的拟合优良程度。

在有常数项的情况下，根据平方和分解公式，可将被解释变量的离差平方和分解为模型可以解释与不可解释的部分。显然，如果模型可以解释的部分所占比重越大，则样本回归线的拟合程度越好。

**定义** 拟合优度(goodness of fit) $R^2$ 为

$$R^2 \equiv \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.28}$$

拟合优度 $R^2$ 也称**可决系数**(coefficient of determination)。可以证明(参见习题)，在有常数项的情况下，拟合优度等于被解释变量 $y_i$ 与拟合值 $\hat{y}_i$ 之间相关系数的平方，即 $R^2 = [\mathrm{Corr}(y_i, \hat{y}_i)]^2$，故记为 $R^2$。显然，$0 \leqslant R^2 \leqslant 1$；$R^2$ 越高，则样本回归线对数据的拟合程度越好。

特别地，如果 $R^2 = 1$，则解释变量 $x$ 可以完全解释 $y$ 的变动。此时，根据方程(4.28)，残差平方和 $\sum_{i=1}^{n} e_i^2 = 0$，故所有残差均为 0，因此所有样本点都在样本回归线上。

反之，如果 $R^2 = 0$，则解释变量 $x$ 对于解释 $y$ 没有任何帮助。此时，根据方程(4.28)，$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0$，故对于任何个体 $i$，都有 $\hat{y}_i \equiv \bar{y}$，因此样本回归线为水平线，与 $x$ 轴平行。这意味着 $\hat{\beta} = 0$，故无论 $x$ 如何变动，对 $y$ 都没有影响。

如果 $0 < R^2 < 1$，则为介于以上两种极端的中间情形，即 $x$ 可以解释 $y$ 的一部分，但无法解释其余部分。有关这三种情形的示意图参见图 4.6。需要注意的是，$R^2$ 只是反映拟合程度的好坏，除此之外并无太多意义。评估回归方程是否显著，仍应使用 $F$ 检验(尽管 $R^2$ 与 $F$ 统计量也有联系)。

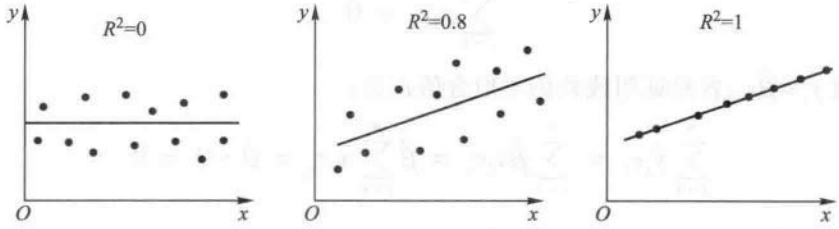



图 4.6 拟合优度的三种情形示意图

## 4.6 无常数项的回归

偶尔也进行无常数项的回归，这或许是由于经济理论的要求[①]，也可能在做模型变换时消去了常数项。由于无常数项的回归必然经过原点，故也称为“经过原点的回归”(regression through

① 比如，研究个人所得税对收入的依赖性。如果收入为 0，则个人所得税也必然为 0，故不应有常数项。

the origin)。此时,一元线性回归模型可写为

$$y_i = \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{4.29}$$

依然进行 OLS 估计,即寻找 $\hat{\beta}$,最小化残差平方和:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2 \tag{4.30}$$

这是一元函数的极值问题,其一阶条件为

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\beta}} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i = 0 \tag{4.31}$$

消去方程左边的"-2"可得

$$\sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i = 0 \tag{4.32}$$

求解 $\hat{\beta}$ 可得

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{4.33}$$

方程(4.33)与有常数项回归的相应表达式(4.14)类似。需要注意的是,如果回归模型无常数项,则平方和分解公式不成立,故不宜使用 $R^2$ 来度量拟合优度。

然而,即使没有常数项,OLS 也仍满足正交性,因为正规方程(组)(4.32)的表达式基本不变。记 $e_i \equiv y_i - \hat{\beta} x_i$,则正规方程(4.32)可写为

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{4.34}$$

记拟合值 $\hat{y}_i \equiv \hat{\beta} x_i$,容易证明残差仍与拟合值正交:

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} \hat{\beta} x_i e_i = \hat{\beta} \sum_{i=1}^{n} x_i e_i = \hat{\beta} \cdot 0 = 0 \tag{4.35}$$

因此,仍可利用 OLS 的正交性将 $\sum_{i=1}^{n} y_i^2$ 分解为:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (\hat{y}_i + e_i)^2 = \sum_{i=1}^{n} \hat{y}_i^2 + 2 \underbrace{\sum_{i=1}^{n} \hat{y}_i e_i}_{=0} + \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 \tag{4.36}$$

其中,$\sum_{i=1}^{n} \hat{y}_i^2$ 为可由模型解释的部分,而 $\sum_{i=1}^{n} e_i^2$ 为模型不可解释的部分。利用上式,可定义**非中心** $R^2$(uncentered $R^2$):

$$R_{uc}^2 \equiv \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{4.37}$$

如果无常数项,Stata 汇报的 $R^2$ 正是 $R_{uc}^2$。显然,$R_{uc}^2$ 与 $R^2$ 的定义不同,二者不具有可比性,但在 Stata 中都称为"R-squared"(仅在无常数项回归时汇报 $R_{uc}^2$)。

## 4.7 一元回归的 Stata 实例

在 Stata 中,进行一元回归的命令为

```
. regress y x,noconstant
```

其中,"y"为被解释变量,"x"为解释变量,选择项"noconstant"表示无常数项(默认有常数项)。

下面使用数据集 grilic.dta,将工资对数($\ln w$)对教育年限($s$)进行一元回归。

```
. use grilic.dta,clear
. reg lnw s
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 1, 756) = | 255.70 |
| Model | 35.2039946 | 1 | 35.2039946 | Prob > F = | 0.0000 |
| Residual | 104.082155 | 756 | .137674809 | R-squared = | 0.2527 |
| | | | | Adj R-squared = | 0.2518 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .37105 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .0966245 | .0060425 | 15.99 | 0.000 | .0847624 | .1084866 |
| _cons | 4.391486 | .0821136 | 53.48 | 0.000 | 4.230288 | 4.552684 |

在上表中,"Coef."表示回归系数(Coefficient),而"_cons"表示常数项(constant)。根据此回归结果,可将样本回归线写为

$$\widehat{\ln w} = 4.391 + 0.097s \tag{4.38}$$

其中,$\widehat{\ln w}$表示被解释变量 $\ln w$ 的拟合值或预测值,而 $\hat{\alpha} = 4.391$,$\hat{\beta} = 0.097$。根据一元回归的结果,教育投资的回报率为 9.7%,即每增加一年教育,平均可提高收入 9.7%。上表左上部显示,TSS(Total)为 139.286 15,其中可解释部分 ESS(Model)为 35.203 994 6,而不可解释部分 RSS(Residual)为 104.082 155。上表右上部显示,$R^2$(R-squared)为 0.252 7,即教育年限约可解释工资对数 25% 的变动。上表的其他统计指标,将在第 5 章介绍。

如果想进行无常数项的回归,可输入如下命令:

```
. reg lnw s,noc
```

| Source | SS | df | MS | Number of obs = 758 |
|---|---|---|---|---|
| | | | | F( 1, 757) =36727.24 |
| Model | 24154.3906 | 1 | 24154.3906 | Prob > F = 0.0000 |
| Residual | 497.855977 | 757 | .657669719 | R-squared = 0.9798 |
| | | | | Adj R-squared = 0.9798 |
| Total | 24652.2466 | 758 | 32.5227528 | Root MSE = .81097 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .4154001 | .0021676 | 191.64 | 0.000 | .4111449 | .4196553 |

从上表可知,无常数项回归的 $R^2$ 高达 0.979 8。但无常数项的 $R^2$ 与有常数项的 $R^2$ 并不可比,而后者更可信(更具经济意义)。上表还显示,教育投资回报率上升为 41.54%,这显然很不合理。另外,在有常数项的回归中,常数项在 1% 水平上显著不为 0,故此例应包括常数项。第 5 章将介绍回归系数的统计显著性(statistical significance)。

## 4.8 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()",可以通过输入命令"return list"来显示,比如:

```
. use grilic.dta,clear
. sum s
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| s | 758 | 13.40501 | 2.231828 | 9 | 18 |

```
. return list
```

```
scalars:
                  r(N) =  758
              r(sum_w) =  758
               r(mean) =  13.40501319261214
                r(Var) =  4.981058057949899
                 r(sd) =  2.231828411403955
               r(min) =  9
               r(max) =  18
               r(sum) =  10161
```

上表列出了在运行命令"sum s"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。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)
.16649207
```

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

```
. reg lnw s
```

```
      Source |       SS       df       MS              Number of obs =     758
-------------+------------------------------           F(  1,   756) =  255.70
       Model |  35.2039946     1  35.2039946            Prob > F      =  0.0000
    Residual |  104.082155   756  .137674809           R-squared     =  0.2527
-------------+------------------------------           Adj R-squared =  0.2518
       Total |  139.28615    757  .183997556           Root MSE      =  .37105

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .0966245   .0060425    15.99   0.000     .0847624    .1084866
       _cons |   4.391486   .0821136    53.48   0.000     4.230288    4.552684
------------------------------------------------------------------------------
```

```
. ereturn list

scalars:
                  e(N) =  758
               e(df_m) =  1
               e(df_r) =  756
                  e(F) =  255.7039662336329
                 e(r2) =  .2527458374860054
               e(rmse) =  .3710455612613365
                e(mss) =  35.20399459202199
                e(rss) =  104.0821552499956
               e(r2_a) =  .2517574060541087
                 e(ll) =  -323.0498302841153
               e(ll_0) =  -433.4714451849197
               e(rank) =  2

macros:
            e(cmdline) : "regress lnw s"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
            e(depvar) : "lnw"
                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]。

① “宏”是 Stata 编程使用的一种缩写方式，它以一个简洁的字符串来代指另一个更为复杂的字符串。

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)”为虚拟变量，即如果观测值在样本中，则取值为 1；反之，则取值为 0。

## 4.9 总体回归函数与样本回归函数：蒙特卡罗模拟

为了更直观地理解总体回归函数(PRF)与样本回归函数的关系(SRF)，下面使用蒙特卡罗法进行模拟。所谓“蒙特卡罗法”(Monte Carlo Methods，MC)[①]，指的是通过计算机模拟，从总体抽取大量随机样本的计算方法。

考虑如下数据生成过程(DGP)或总体回归模型：

$$y_i = 1 + 2x_i + \varepsilon_i \quad (i = 1, \cdots, 30) \tag{4.39}$$

其中，解释变量 $x_i \sim N(3, 2^2)$，扰动项 $\varepsilon_i \sim N(0, 3^2)$，而样本容量为 $n = 30$。我们将从 $N(3, 2^2)$ 随机抽取 30 个解释变量 $x_i$ 的观测值，并从 $N(0, 3^2)$ 随机抽取 30 个扰动项 $\varepsilon_i$ 的观测值；然后，根据总体回归模型(4.39)计算相应的被解释变量 $y_i$。最后，把 $y_i$ 对 $x_i$ 进行回归，得到样本回归函数(SRF)，并与总体回归函数(PRF)进行比较。

具体来说，可在 Stata 命令窗口依次输入如下命令：

```
. clear                          (删除内存中已有数据)
. set obs 30                     (确定随机抽样的样本容量为 30)
. set seed 10101                 (指定随机抽样的“种子”为 10101)
. gen x = rnormal(3,4)           (得到服从 N(3,2^2) 分布的随机样本，记为 x)
. gen e = rnormal(0,9)           (得到服从 N(0,3^2) 分布的随机样本，记为 e)
. gen y = 1 + 2 * x + e          (计算被解释变量 y)
. reg y x                        (把 y 对 x 进行 OLS 回归)
```

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   28.26
       Model |  2362.04948     1  2362.04948            Prob > F      =  0.0000
    Residual |  2340.54765    28  83.5909875            R-squared     =  0.5023
-------------+------------------------------           Adj R-squared =  0.4845
       Total |  4702.59714    29  162.158522            Root MSE      =  9.1428

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    2.355635   .4431423     5.32   0.000     1.447899    3.26337
       _cons |   -1.641879    2.52587    -0.65   0.521    -6.815888   3.532131
------------------------------------------------------------------------------
```

其中，命令“set seed 10101”用来确定随机数的初始值(称为“种子”，可任意设置，此处设为 10101)，以便再次模拟时得到完全一样的结果(有关随机数的产生，参见本章附录)。

由上表的回归结果可知，由于样本容量仅为 30，故存在一定的抽样误差。比如，斜率的真实值为 2，而样本估计值为 2.36；截距项的真实值为 1，而样本估计值为 -1.64，符号相反(但不显著)。

① 此名来源于摩纳哥(在法国附近)的蒙特卡罗赌场(Monte Carlo Casino)，据说这是最早使用此法的一位美国物理学家的叔叔常去的赌场。

更直观地，下面把总体回归函数、散点图与样本回归函数画在一起，结果参见图 4.7。

```
. twoway function PRF = 1 + 2 * x, range( -5 15) || scatter y x || lfit y x,
lpattern(dash)
```

其中，选择项“range( -5 15)”用于指定画图的横轴范围介于 -5 与 15 之间；默认为 0 与 1 之间，即“range(0 1)”。选择项“lpattern(dash)”表示画虚线，默认画实线。

在图 4.7 中，实线为总体回归函数（PRF）；而虚线为样本回归线（SRF），即被解释变量的拟合值（fitted values）。从图 4.7 看，SRF 似乎比较接近于 PRF。显然，如果使用不同的随机数种子再次抽样，将得到不同的 SRF；而 PRF 始终不变。

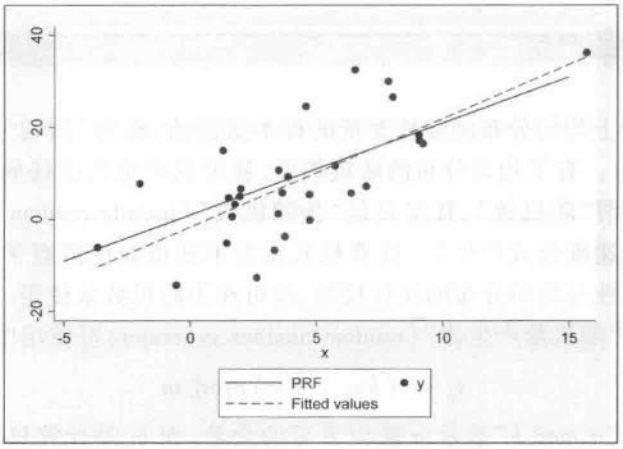



图 4.7　总体与样本回归线的蒙特卡罗模拟

## 附录 A4.1　高尔顿与回归

弗朗西斯·高尔顿（Francis Galton，1822—1911）是英国心理学家、人类学家、统计学家，以及优生学的创始人（参见图 4.8）。在 19 世纪末，高尔顿曾研究子女身高与父母身高的关系，并收集了 928 位子女与其 205 对父母的身高数据（平均每对父母约有 4.5 位子女）。

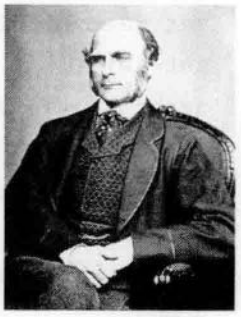


图 4.8　Francis Galton（1822—1911）[①]

经过统计分析，高尔顿发现，虽然高个父母的子女身高平均而言高于矮个父母的子女身高，但特别高个父母的子女通常矮于其父母，而特别矮个父母的子女则通常高于其父母（Galton，1886）。高尔顿称此现象为“回归平庸”（regression towards mediocrity），后人则多称为“回归均值”（regression to the mean）。对于一元回归模型 $y_i = \alpha + \beta x_i + \varepsilon_i$，回归均值相当于 $0 < \beta < 1$ 的情形，故父辈（$x_i$）的优势在子辈（$y_i$）有所削弱。

产生回归均值现象的根本原因在于，特别高个父母的身高是由于偶然现象所造成的（相当于在概率分布的最右边尾部取了极端值），而其子

① 此图来自维基百科：http://en.wikipedia.org/wiki/Francis_Galton。

女们再取到如此大极端值的可能性很低，故一般而言身高反而低于父母。同样道理，伟人的后代通常很难同样伟大，因为造就伟人的诸多偶然因素，很难在其下一代全部再现。2002 年诺贝尔经济学奖得主、心理学家 Daniel Kahneman 曾指出，回归均值可以解释为什么批评似乎能起改善作用，而表扬则适得其反，因为批评一般发生在极糟糕结果后，而表扬则通常出现在巅峰表现之后。

在当代，回归分析(regression analysis)已成为计量经济学的主要方法，主要指用概率统计方法来估计与检验变量之间的(平均)函数关系。在这个意义上，回归均值现象已经不重要(回归系数 $\beta$ 未必小于 1)，但“回归”这一术语则沿用下来。

## 附录 A4.2 随机数的产生

服从在(0,1)区间上均匀分布的随机变量的样本观测值，称为“均匀分布随机数”，简称“随机数”(random number)。有了均匀分布的随机数后，就可以产生几乎任何分布的随机数了。然而，由计算机产生的所谓“随机数”，其实只是“伪随机数”(pseudo random number)，因为它仍然来自确定性的序列(由递推公式产生)。计算机其实并不知道如何掷骰子！尽管如此，这些“伪随机数”都能通过独立性与均匀分布的统计检验，故可作为随机数来使用。

比如，一个简单的“随机数产生器”(random number generator)可使用以下递推公式：

$$x_j = (kx_{j-1} + c) \bmod m \tag{4.40}$$

其中，$k$，$c$ 为常数，算子“$a$ mod $b$”表示 $a$ 除以 $b$ 后的余数，而 $m$ 是计算机所能表示的最大数[①]。这个序列就是介于 0 与 $m$ 之间的整数，而 $r_j \equiv x_j/m$ 就是介于 0 与 1 之间的随机数。初始值 $x_0$，称为种子(seed)，决定了该随机序列的初始值 $r_0 \equiv x_0/m$。为了使随机样本具有可重复性，保证再次抽样或别人抽样也能得到完全一样的样本，需要设定种子。如果不给定种子，则按照“计算机内置钟表”(computer clock)来自动选择种子，这样每次抽样的样本就会不同，模拟的结果也就不能完全复制。

在 Stata 中，使用命令“set seed #”来确定“种子”，例如：

```
set seed 10101          (确定“种子”为 10101)
set obs 30              (确定样本容量为 30)
gen x = runiform()      (产生均匀分布的随机变量 x)
gen x = rnormal()       (产生标准正态分布 N(0,1)的随机样本)
gen x = rnormal(m,s)    (产生正态分布 N(m,s^2)的随机样本)
gen x = rt(m)           (产生自由度为 m 的 t 分布随机样本)
gen x = rchi2(m)        (产生自由度为 m 的 χ^2 分布随机样本)
```

**习题**

4.1 考虑以下消费函数(consumption function)：

$$C_i = \alpha + \beta Y_i + \varepsilon_i \tag{4.41}$$

① 比如，如果计算机使用 32 位数进行计算，则 $m = 2^{32} - 1$。

其中，$C_i$ 为个体 $i$ 的消费开支，而 $Y_i$ 为个体 $i$ 的可支配收入。假设 OLS 回归所得的样本回归线为

$$\hat{C}_i = \hat{\alpha} + \hat{\beta} Y_i \tag{4.42}$$

(1) 斜率 $\hat{\beta}$ 的经济含义是什么？

(2) 截距项 $\hat{\alpha}$ 的经济含义是什么？

(3) 对于个体 $i$，计算其平均消费倾向(average propensity to consume) $C_i/Y_i$。假设 $\hat{\alpha}>0$，则随着个体 $i$ 可支配收入的增加，其平均消费倾向将如何变化？

4.2 假设把 $y$ 对 $x$ 进行回归，样本容量为 30，$\sum_{i=1}^{30} y_i = 150$，$\sum_{i=1}^{30} x_i = 60$。如果截距项的 OLS 估计值为 2，则斜率的 OLS 估计值是多少？

4.3 证明 $\dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$，其中 $\bar{x} \equiv \dfrac{1}{n}\sum_{i=1}^{n} x_i$，$\bar{y} \equiv \dfrac{1}{n}\sum_{i=1}^{n} y_i$。(提示：从等式右边向左边证明。)

4.4 考虑只有常数项的回归：

$$y_i = \alpha + \varepsilon_i \tag{4.43}$$

其中，常数项 $\alpha$ 是唯一的解释变量。推导 $\alpha$ 的 OLS 估计量，并证明此回归的 $R^2$ 等于 0。

4.5 考虑如下线性回归：

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{4.44}$$

其中，假设已知 $\alpha=3$，推导 $\beta$ 的 OLS 估计量。

4.6 考虑有常数项的回归：

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{4.45}$$

证明 $R^2 = [\mathrm{Corr}(y_i, \hat{y}_i)]^2$，其中 $\mathrm{Corr}(y_i, \hat{y}_i) \equiv \dfrac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{\hat{y}})^2}}$。(提示：利用以下性质，即 $\bar{\hat{y}} = \bar{y}$，$y_i = \hat{y}_i + e_i$，以及 OLS 的正交性。)

4.7 数据集 galton. dta 包含 Galton(1886)的原始数据。变量 *parent* 为父母的平均身高(英寸)，而 *child* 为子女身高(英寸)。其中，为平衡身高的性别差异，女性身高(包括母亲与女儿)均乘以 1.08。

(1) 计算变量 *child* 与 *parent* 的基本统计特征。

(2) 将变量 *child* 与 *parent* 的散点图与线性拟合图画在一起。

(3) 考虑以下回归方程：

$$child_i = \alpha + \beta\, parent_i + \varepsilon_i \tag{4.46}$$

其中，随机扰动项 $\varepsilon_i$ 代表哪些因素？

(4) 使用 OLS 估计方程(4.46)并回答：父母身高每增加 1 英寸，子女身高平均将增加多少？父母身高可解释子女身高变动的百分之几？

(5) 定义 *parent_dev* 为父母身高减去父母那一辈人群的平均身高，并定义 *gengap* 为子女身高减去父母身高。将 *gengap* 对 *parent_dev* 进行回归。是否存在“回归均值现象”(参见附录 A4.1)？

4.8 重复本章 4.9 节的蒙特卡罗模拟，但将样本容量从 30 增加到 100。此时，对于截距项与斜率的估计是否更为准确？将总体回归线、样本回归线以及散点图画在一起。

*With four parameters I can fit an elephant, and with five*
*I can make him wiggle his trunk.* ——John von Neumann

# 5. 多元线性回归

## 5.1 二元线性回归

一元回归很可能遗漏了其他因素。比如，在第4章关于教育投资回报率的研究中，将工资对数对教育年限回归。但一般来说，工资还依赖于个人能力，而个人能力未包括在回归方程中，故被纳入扰动项。而且，能力强的人通常上学更久（二者存在正相关），故一元回归所估计的教育回报率事实上也包括了对能力的回报，导致估计出现偏差。其他遗漏变量还包括年龄、工龄、性别、种族、相貌等，其中年龄与工龄可视为“在职培训”（on the job training）的代理变量，而在职培训是增加人力资本（human capital）的另一重要方式。

为此，本章考虑多元回归。首先考察比较简单的二元回归。

$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{5.1}$$

其中，$x_{i1}$ 与 $x_{i2}$ 为解释变量；$\alpha$ 为截距项；$\beta$ 为在给定 $x_2$ 条件下，$x_1$ 对 $y$ 的边际效应（忽略扰动项 $\varepsilon_i$）；而 $\gamma$ 为在给定 $x_1$ 条件下，$x_2$ 对 $y$ 的边际效应。

OLS 估计量的最优化问题仍为残差平方和最小化：

$$\min_{\hat{\alpha},\hat{\beta},\hat{\gamma}} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_{i1} - \hat{\gamma} x_{i2})^2 \tag{5.2}$$

从几何上，这意味着寻找一个回归平面 $\hat{y}_i \equiv \hat{\alpha} + \hat{\beta} x_{1i} + \hat{\gamma} x_{2i}$，即估计参数 $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$，使得所有样本点 $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^{n}$ 离此回归平面最近，参见图 5.1。将表达式(5.2)分别对 $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ 求偏导数，可得此最小化问题的一阶条件，求解可获得 $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ 的 OLS 估计量。

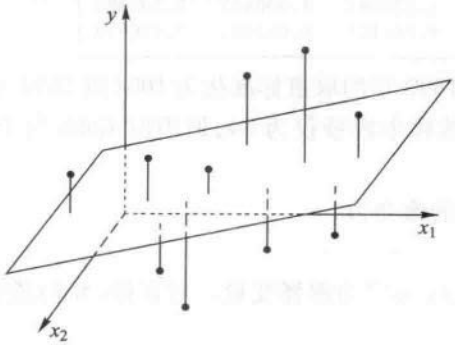



图 5.1 二元线性回归的示意图

**例** （Cobb-Douglas 生产函数）Cobb and Douglas（1928）使用美国 1899—1922 年制造业产出（$y$）、资本（$k$）与劳动力（$l$）的数据，估计如下生产函数：

$$y_t = \alpha k_t^{\beta} l_t^{\gamma} e^{\varepsilon_t} \tag{5.3}$$

其中，$e^{\varepsilon_t}$为乘积形式的扰动项，而下标 $t$ 表示时间（年）。上式似乎为非线性模型，但只要在方程两边同时取对数，即可转换为线性模型：

$$\ln y_t = \ln \alpha + \beta \ln k_t + \gamma \ln l_t + \varepsilon_t \tag{5.4}$$

数据集 cobb_douglas. dta 提供了 Cobb and Douglas（1928）的原始数据。由于样本容量较小，首先看一下数据集中的观测值。

```
. use cobb_douglas.dta, clear
. list
```

| | year | k | l | y | lnk | lnl | lny |
|---|---|---|---|---|---|---|---|
| 1. | 1899 | 100 | 100 | 100 | 4.60517 | 4.60517 | 4.60517 |
| 2. | 1900 | 107 | 105 | 101 | 4.672829 | 4.65396 | 4.61512 |
| 3. | 1901 | 114 | 110 | 112 | 4.736198 | 4.70048 | 4.718499 |
| 4. | 1902 | 122 | 118 | 122 | 4.804021 | 4.770685 | 4.804021 |
| 5. | 1903 | 131 | 123 | 124 | 4.875197 | 4.812184 | 4.820282 |
| 6. | 1904 | 138 | 116 | 122 | 4.927254 | 4.75359 | 4.804021 |
| 7. | 1905 | 149 | 125 | 143 | 5.003946 | 4.828314 | 4.962845 |
| 8. | 1906 | 163 | 133 | 152 | 5.09375 | 4.890349 | 5.02388 |
| 9. | 1907 | 176 | 138 | 151 | 5.170484 | 4.927254 | 5.01728 |
| 10. | 1908 | 185 | 121 | 126 | 5.220356 | 4.795791 | 4.836282 |
| 11. | 1909 | 198 | 140 | 155 | 5.288267 | 4.941642 | 5.043425 |
| 12. | 1910 | 208 | 144 | 159 | 5.337538 | 4.969813 | 5.068904 |
| 13. | 1911 | 216 | 145 | 153 | 5.375278 | 4.976734 | 5.030438 |
| 14. | 1912 | 226 | 152 | 177 | 5.420535 | 5.02388 | 5.17615 |
| 15. | 1913 | 236 | 154 | 184 | 5.463832 | 5.036952 | 5.214936 |
| 16. | 1914 | 244 | 149 | 169 | 5.497168 | 5.003946 | 5.129899 |
| 17. | 1915 | 266 | 154 | 189 | 5.583496 | 5.036952 | 5.241747 |
| 18. | 1916 | 298 | 182 | 225 | 5.697093 | 5.204007 | 5.416101 |
| 19. | 1917 | 335 | 196 | 227 | 5.81413 | 5.278115 | 5.42495 |
| 20. | 1918 | 366 | 200 | 223 | 5.902633 | 5.298317 | 5.407172 |
| 21. | 1919 | 387 | 193 | 218 | 5.958425 | 5.26269 | 5.384495 |
| 22. | 1920 | 407 | 193 | 231 | 6.008813 | 5.26269 | 5.442418 |
| 23. | 1921 | 417 | 147 | 179 | 6.033086 | 4.990433 | 5.187386 |
| 24. | 1922 | 431 | 161 | 240 | 6.066108 | 5.081404 | 5.480639 |

其中，变量 $k$，$l$ 与 $y$ 均将 1899 年的取值标准化为 100（以 1899 年为指数的基期），而 $\ln k$，$\ln l$ 与 $\ln y$ 分别为其对数值。虽然样本容量仅为 24，但当时 Cobb 与 Douglas 在获得这些统计数据时，还颇费周折。

在 Stata 中进行二元回归的命令为

```
. regress y x1 x2
```

其中，“y”为被解释变量，而“x1 x2”为解释变量。对方程（5.4）进行二元回归估计，可输入如下命令

```
. reg lny lnk lnl
```

| Source | SS | df | MS | Number of obs = | 24 |
|---|---|---|---|---|---|
| | | | | F( 2, 21) = | 236.12 |
| Model | 1.59622155 | 2 | .798110773 | Prob > F = | 0.0000 |
| Residual | .070981736 | 21 | .003380083 | R-squared = | 0.9574 |
| | | | | Adj R-squared = | 0.9534 |
| Total | 1.66720328 | 23 | .072487099 | Root MSE = | .05814 |

| lny | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnk | .2330537 | .0635298 | 3.67 | 0.001 | .1009363 | .3651711 |
| lnl | .807278 | .1450762 | 5.56 | 0.000 | .5055755 | 1.108981 |
| _cons | -.1773099 | .4342933 | -0.41 | 0.687 | -1.080472 | .7258525 |

其中，$\ln k$（资本对数）与 $\ln l$（劳动力对数）的系数分别为 0.233 与 0.807，且拟合优度 $R^2$ 高达 0.957。这种形式的生产函数后来被称为“柯布 - 道格拉斯生产函数”（Cobb-Douglas production function）[①]，在经济学中得到广泛应用。

根据上表的回归结果，可得样本回归平面：

$$\widehat{\ln y_t} = -0.177 + 0.233\ln k_t + 0.807\ln l_t \tag{5.5}$$

其中，$\widehat{\ln y_t}$ 为 $\ln y_t$ 的拟合值或预测值。在 Stata 中，做完 OLS 回归后，可用命令 predict 来计算拟合值与残差。

```
. predict lny1
(option xb assumed; fitted values)
```

此命令将 $\ln y$ 的拟合值记为“lny1”。如果要计算残差，并记为 e，可输入命令

```
. predict e,residual
```

其中，选择项“residual”表示计算残差（如果省略此选择项，则默认为计算拟合值）。下面，将 $\ln y$ 及其拟合值、残差同时列表。

```
. list lny lny1 e
```

① 事实上，这种形式的生产函数最早由 Wicksell(1896) 提出，但 Cobb and Douglas(1928) 最早使用此生产函数进行回归分析，遂以 Cobb-Douglas 生产函数而闻名。

| | lny | lny1 | e |
|---|---|---|---|
| 1. | 4.60517 | 4.613595 | -.0084246 |
| 2. | 4.61512 | 4.66875 | -.0536295 |
| 3. | 4.718499 | 4.721073 | -.0025745 |
| 4. | 4.804021 | 4.793554 | .010467 |
| 5. | 4.820282 | 4.843644 | -.0233621 |
| 6. | 4.804021 | 4.808474 | -.0044528 |
| 7. | 4.962845 | 4.88667 | .0761749 |
| 8. | 5.02388 | 4.957679 | .0662019 |
| 9. | 5.01728 | 5.005354 | .0119254 |
| 10. | 4.836282 | 4.91085 | -.074568 |
| 11. | 5.043425 | 5.04442 | -.0009945 |
| 12. | 5.068904 | 5.078644 | -.0097398 |
| 13. | 5.030438 | 5.093026 | -.0625884 |
| 14. | 5.17615 | 5.141634 | .0345157 |
| 15. | 5.214936 | 5.162277 | .0526584 |
| 16. | 5.129899 | 5.143401 | -.0135028 |
| 17. | 5.241747 | 5.190166 | .0515813 |
| 18. | 5.416101 | 5.351499 | .0646015 |
| 19. | 5.42495 | 5.438601 | -.0136506 |
| 20. | 5.407172 | 5.475536 | -.0683641 |
| 21. | 5.384495 | 5.459777 | -.0752818 |
| 22. | 5.442418 | 5.47152 | -.0291027 |
| 23. | 5.187386 | 5.25739 | -.0700038 |
| 24. | 5.480639 | 5.338525 | .142114 |

更直观地，可将产出对数及其拟合值画在一起(结果参见图 5.2)。

```
. line lny lny1 year,lpattern(solid dash)
```

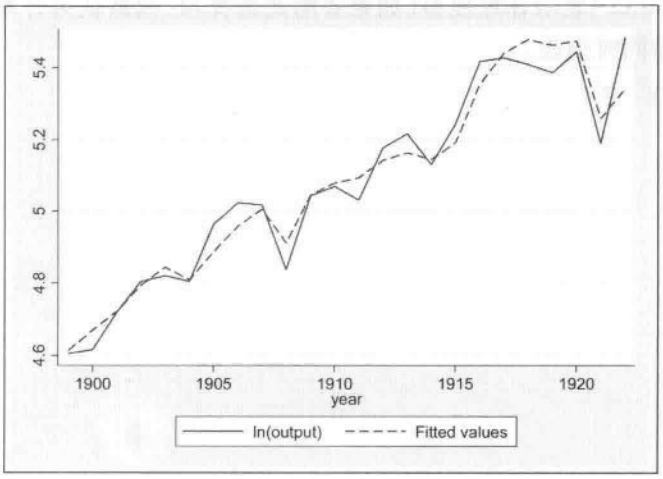



图 5.2 产出对数的实际值与预测值

从图 5.2 可知，产出对数的预测值与实际值相当吻合。

## 5.2 多元线性回归模型

一般的多元线性回归模型可写为

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{5.6}$$

其中，$x_{i1}$为个体 $i$ 的第 1 个解释变量，$x_{i2}$为个体 $i$ 的第 2 个解释变量，以此类推。一般地，$x_{ik}$的第一个下标表示个体 $i$（共有 $n$ 位个体，即样本容量为 $n$），而第二个下标表示第 $k$ 个解释变量（共有 $K$ 个解释变量）。

在绝大多数情况下，回归方程都有常数项，故通常令 $x_{i1} \equiv 1$（恒等于 1），则方程(5.6)简化为

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \tag{5.7}$$

多元线性回归模型可以更方便地以矩阵表示。如果不考虑扰动项 $\varepsilon_i$，此方程右边的主体部分为乘积之和，即 $\sum_{k=1}^{K} \beta_k x_{ik}$，其中 $x_{i1} \equiv 1$。根据线性代数知识，乘积之和可写为两个向量的内积。定义列向量 $\boldsymbol{x}_i \equiv (1\ x_{i2} \cdots x_{iK})'$（包含个体 $i$ 的全部解释变量）；而参数向量 $\boldsymbol{\beta} \equiv (\beta_1\ \beta_2 \cdots \beta_K)'$（包含全部回归系数），则 $\sum_{k=1}^{K} \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}$。故可将原模型(5.7)写为

$$y_i = (1 \quad x_{i2} \quad \cdots \quad x_{iK}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \varepsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \tag{5.8}$$

上式对所有个体 $i$ 都成立（$i = 1, \cdots, n$），故有 $n$ 个形如(5.8)的方程。将所有这 $n$ 个方程都叠放在一起可得

$$\begin{pmatrix} y_1 = \boldsymbol{x}_1' \boldsymbol{\beta} + \varepsilon_1 \\ y_2 = \boldsymbol{x}_2' \boldsymbol{\beta} + \varepsilon_2 \\ \vdots \\ y_n = \boldsymbol{x}_n' \boldsymbol{\beta} + \varepsilon_n \end{pmatrix} \tag{5.9}$$

将共同的参数向量 $\boldsymbol{\beta}$ 向右边提出，经整理可得

$$\boldsymbol{y} \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}}_{\boldsymbol{X}} \boldsymbol{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.10}$$

其中，$\boldsymbol{y} \equiv (y_1 \quad y_2 \quad \cdots \quad y_n)'$为被解释变量构成的列向量，$\boldsymbol{\varepsilon} \equiv (\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n)'$为所有扰动项

构成的列向量，$\boldsymbol{X}$ 为 $n \times K$ 数据矩阵(data matrix)，其第 $i$ 行包含个体 $i$ 的全部解释变量，而第 $k$ 列包含第 $k$ 个解释变量的全部观测值，即

$$\boldsymbol{X} \equiv \begin{pmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nK} \end{pmatrix}_{n \times K} \tag{5.11}$$

## 5.3 OLS 估计量的推导

对于多元回归模型，OLS 估计量的最小化问题为

$$\min_{\hat{\beta}_1, \cdots, \hat{\beta}_K} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_K x_{iK})^2 \tag{5.12}$$

最小二乘法寻找使残差平方和(SSR) $\sum_{i=1}^{n} e_i^2$ 最小的 $(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_K)$。在几何上，一元回归寻找最佳拟合的回归直线，使得观测值 $y_i$ 到该回归直线的距离的平方和最小；二元回归寻找最佳拟合的回归平面；而多元回归则寻找最佳拟合的回归超平面(superplane)。

此最小化问题的一阶条件为

$$\begin{cases} \dfrac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) = 0 \\ \dfrac{\partial}{\partial \hat{\beta}_2} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) x_{i2} = 0 \\ \cdots\cdots\cdots\cdots \\ \dfrac{\partial}{\partial \hat{\beta}_K} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) x_{iK} = 0 \end{cases} \tag{5.13}$$

消去方程左边的“-2”可得

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) = 0 \\ \sum_{i=1}^{n} x_{i2} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) = 0 \\ \cdots\cdots\cdots\cdots \\ \sum_{i=1}^{n} x_{iK} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}) = 0 \end{cases} \tag{5.14}$$

这是一个包含 $K$ 个未知数($\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_K$)与 $K$ 个方程的联立方程组,称为“正规方程组”(normal equations)。满足此正规方程组的 $\hat{\boldsymbol{\beta}}\equiv(\hat{\beta}_1 \quad \hat{\beta}_2 \quad \cdots \quad \hat{\beta}_K)'$ 称为 OLS 估计量(OLS estimator)。

正规方程组可以更方便地用矩阵来表达。由于残差 $e_i\equiv y_i-\hat{\beta}_1-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}$,故正规方程组可写为

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_{i2}e_i = 0 \\ \cdots\cdots\cdots\cdots \\ \sum_{i=1}^{n} x_{iK}e_i = 0 \end{cases} \tag{5.15}$$

上式每一方程都是乘积求和的形式,故可用向量内积表示。比如,第 1 个方程可写为

$$\sum_{i=1}^{n} e_i = (1 \quad 1 \quad \cdots \quad 1)\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = 0 \tag{5.16}$$

而第 2 个方程可写为

$$\sum_{i=1}^{n} x_{i2}e_i = (x_{12} \quad x_{22} \quad \cdots \quad x_{n2})\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = 0 \tag{5.17}$$

以此类推,第 $K$ 个方程可写为

$$\sum_{i=1}^{n} x_{iK}e_i = (x_{1K} \quad x_{2K} \quad \cdots \quad x_{nK})\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = 0 \tag{5.18}$$

由此可知,残差向量 $\boldsymbol{e}\equiv(e_1 \quad e_2 \quad \cdots \quad e_n)'$ 与每个解释变量都正交,这是 OLS 估计量的一大特征。将以上内积以矩阵形式表示可得

$$\underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1K} & x_{2K} & \cdots & x_{nK} \end{pmatrix}}_{\boldsymbol{X}'} \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}_{\boldsymbol{e}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{0} \tag{5.19}$$

其中，$\boldsymbol{X}'$为数据矩阵 $\boldsymbol{X}$ 的转置。因此，正规方程组可简洁地写为

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \tag{5.20}$$

由于 $\boldsymbol{X}'$的第 $k$ 行包含第 $k$ 个解释变量的全部观测值，故根据 $\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0}$ 也可看出，残差向量 $\boldsymbol{e}$ 与每个解释变量都正交。从 $e_i \equiv y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK})$ 出发，可将残差向量写为（参见习题）

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{5.21}$$

将上式代入正规方程组(5.20)可得

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \tag{5.22}$$

乘开来，并移项可知，最小二乘估计量 $\hat{\boldsymbol{\beta}}$ 满足

$$(\boldsymbol{X}'\boldsymbol{X})_{K\times K}\hat{\boldsymbol{\beta}}_{K\times 1} = \boldsymbol{X}'_{K\times n}\boldsymbol{y}_{n\times 1} \tag{5.23}$$

假设$(\boldsymbol{X}'\boldsymbol{X})^{-1}$存在，可求解 OLS 估计量

$$\hat{\boldsymbol{\beta}} \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5.24}$$

这就是多元线性回归的 OLS 估计量。

## 5.4 OLS 的几何解释

定义被解释变量 $y_i$ 的拟合值（fitted value）或预测值（predicted value）为

$$\hat{y}_i \equiv \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK} \quad (i = 1, \cdots, n) \tag{5.25}$$

将所有个体的拟合值写为列向量 $\hat{\boldsymbol{y}}$，并参照方程(5.10)同样的推导可得

$$\hat{\boldsymbol{y}} \equiv (\hat{y}_1 \quad \hat{y}_2 \quad \cdots \quad \hat{y}_n)' = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{5.26}$$

容易证明，拟合值向量与残差向量正交，因为

$$\hat{\boldsymbol{y}}'\boldsymbol{e} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{e} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{e} = \hat{\boldsymbol{\beta}}' \cdot \boldsymbol{0} = 0 \tag{5.27}$$

由于 $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{y}-\hat{\boldsymbol{y}}$，故

$$\boldsymbol{y}=\hat{\boldsymbol{y}}+\boldsymbol{e}$$

因此，被解释变量 $\boldsymbol{y}$ 可分解为相互正交的拟合值 $\hat{\boldsymbol{y}}$ 与残差 $\boldsymbol{e}$ 之和，参见图 5.3。在图 5.3 中，拟合值 $\hat{\boldsymbol{y}}$ 可视为被解释变量 $\boldsymbol{y}$ 向解释变量超平面 $\boldsymbol{X}$ 的投影(projection)。由于拟合值为解释变量的线性组合，即 $\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$，故拟合值向量 $\hat{\boldsymbol{y}}$ 正好在超平面 $\boldsymbol{X}$ 上。而根据 OLS 的正交性，残差向量 $\boldsymbol{e}$ 与 $\hat{\boldsymbol{y}}$ 正交。图 5.3 可视为 OLS 的几何解释。

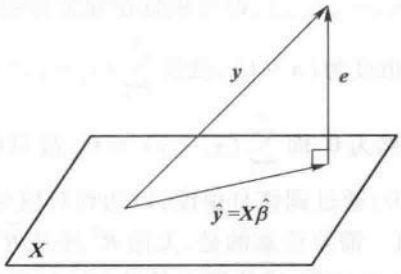



图 5.3 最小二乘法的正交性

## 5.5 拟合优度

对于多元回归，在回归方程有常数项的情况下，由于 OLS 的正交性，平方和分解公式依然成立（证明方法与一元回归相同），故仍可将被解释变量的离差平方和 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 分解如下：

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{TSS}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{ESS}}+\underbrace{\sum_{i=1}^{n}e_i^2}_{\text{RSS}} \tag{5.28}$$

其中，ESS 为模型可解释的部分，而 RSS 为模型不可解释的部分。根据平方和分解公式(5.28)，可定义拟合优度。

**定义** 拟合优度 $R^2$ 为

$$0\leqslant R^2\equiv\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\leqslant 1 \tag{5.29}$$

拟合优度 $R^2$ 的缺点是，如果增加解释变量的数目，则 $R^2$ 只增不减，因为至少可让新增解释变量的系数为 0 而保持 $R^2$ 不变。另外，通过最优地选择新增解释变量的系数(以及已有解释变量的系数)，通常可以提高 $R^2$。为此，引入如下校正拟合优度，对解释变量过多(即模型不够简洁)进行惩罚。

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

$$\overline{R}^2 \equiv 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-K)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)} \tag{5.30}$$

其中，$\sum_{i=1}^{n} e_i^2$ 的自由度(degree of freedom)为$(n-K)$。虽然 $\sum_{i=1}^{n} e_i^2$ 由 $n$ 个随机变量$\{e_1,\cdots,e_n\}$所构成，但$\{e_1,\cdots,e_n\}$受由 $K$ 个方程组成的正规方程组(5.19)的约束，故只有其中$(n-K)$个残差是(自由)独立的。换言之，给定$\{e_1,\cdots,e_{n-K}\}$，即可根据正规方程组求解其余$\{e_{n-K+1},\cdots,e_n\}$。

类似地，$\sum_{i=1}^{n} (y_i - \bar{y})^2$ 的自由度为$(n-1)$。虽然 $\sum_{i=1}^{n} (y_i - \bar{y})^2$ 由 $n$ 个离差$\{(y_1-\bar{y}),\cdots,(y_n-\bar{y})\}$所构成，但这些离差之和必然为0，即 $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$，故只有其中$(n-1)$个离差是(自由)独立的。由此可见，定义式(5.30)通过调整自由度，以达到对模型过于复杂的惩罚。

$\overline{R}^2$ 的缺点是，它可能为负值。需要注意的是，无论 $R^2$ 还是 $\overline{R}^2$，只反映拟合程度的好坏，除此并无太多意义。评估回归方程是否显著，应使用 $F$ 检验($R^2$ 与 $F$ 统计量也有联系)。另外，如果回归模型无常数项，则仍需使用“非中心 $R^2$”(uncentered $R^2$)：

$$R_{uc}^2 \equiv \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{5.31}$$

## 5.6 古典线性回归模型的假定

为了得到 OLS 估计量的良好性质，“古典线性回归模型”(Classical Linear Regression Model)作了如下假定[①]。

**假定 5.1** 线性假定(linearity)。总体模型为

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1,\cdots,n) \tag{5.32}$$

线性假设的含义是每个解释变量对 $y_i$ 的边际效应为常数，比如 $\frac{\partial y_i}{\partial x_{i2}} = \beta_2$(忽略扰动项 $\varepsilon_i$)。如果边际效应可变，可加入平方项(比如 $x_{i2}^2$)或交叉项(比如 $x_{i2}x_{i3}$)。交叉项也称为互动项(interaction term)。

**例**(平方项)：考虑如下回归方程，

$$\ln w_i = \beta_1 + \beta_2 s_i + \beta_3 s_i^2 + \varepsilon_i \tag{5.33}$$

其中，$\ln w_i$ 为工资对数，$s_i$ 为教育年限。则教育年限对工资对数的边际效应为(忽略扰动项)

① 这些假定主要参考了 Wooldridge(2009), 4e, Appendix E.2, p.802。

$$\frac{\partial \ln w_i}{\partial s_i} = \beta_2 + 2\beta_3 s_i \tag{5.34}$$

如果 $\beta_3 < 0$，则存在教育投资回报率递减，即小学的教育回报率高于中学，而中学的教育回报率高于大学（经验数据一般与此相符）。反之，如果 $\beta_3 > 0$，则存在教育投资回报率递增。总之，如果变量的边际效应不是常数，可考虑在回归方程中加入平方项[①]。此时，只要将 $s^2$ 也视为解释变量（根据 $s$ 可算出 $s^2$ 的取值），则仍然符合线性模型的假定。

**例**（互动项）：考虑如下生产函数方程：

$$y_i = \beta_1 + \beta_2 k_i + \beta_3 l_i + \beta_4 k_i \times l_i + \varepsilon_i \tag{5.35}$$

其中，$y$ 为产出，$k$ 为资本，$l$ 为劳动力，而 $k \times l$ 为资本与劳动力的互动项。劳动力的边际产出为 $\frac{\partial y_i}{\partial l_i} = \beta_3 + \beta_4 k_i$（忽略扰动项）。如果 $\beta_4 > 0$，则说明资本与劳动力是互补的，即随着资本上升，劳动力的边际产出也增加。此时，只要将 $k \times l$ 也视为解释变量，则依然符合线性模型的假定。

**例**（函数形式）：经济学中常用的生产函数并非方程（5.35），而是 Cobb-Douglas 生产函数：

$$y_i = e^{\beta_1} k_i^{\beta_2} l_i^{\beta_3} e^{\varepsilon_i} \tag{5.36}$$

其中，$e^{\beta_1}$ 与 $e^{\varepsilon_i}$ 分别为乘积形式的常数项与扰动项。将上式两边同时取对数可得

$$\ln y_i = \beta_1 + \beta_2 \ln k_i + \beta_3 \ln l_i + \varepsilon_i \tag{5.37}$$

其中，$\beta_2$ 为产出的资本弹性，即资本每增加 1%，产出平均增加百分之几；而 $\beta_3$ 为产出的劳动力弹性，即劳动力每增加 1%，产出平均增加百分之几。此时，只要将 $\ln y_i$ 视为被解释变量，而将 $\ln k_i$ 与 $\ln l_i$ 视为解释变量，则仍然符合线性模型的假定。

总之，只要将回归方程中变量的高次项（比如 $x^2$）或函数（比如 $\ln x$）都作为变量来看待，则依然满足线性假定。由此可知，线性假定的本质要求是，回归函数是参数（$\beta_1, \cdots, \beta_K$）的线性函数（linear in parameters）。

**假定 5.2** 严格外生性（strict exogeneity）要求

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

严格外生性意味着，在给定数据矩阵 $\boldsymbol{X}$ 的情况下，扰动项 $\varepsilon_i$ 的条件期望为 0。因此，$\varepsilon_i$ 均值独立于（mean-independent）所有解释变量的观测数据，而不仅仅是同一观测数据 $\boldsymbol{x}_i$ 中的解释变量。这意味着，$\varepsilon_i$ 与所有个体的解释变量都不相关，即 $\mathrm{Cov}(\varepsilon_i, x_{jk}) = 0, \forall j, k$。此假定很强，在第 6 章大样本 OLS 可放松。

事实上，均值独立仅要求 $E(\varepsilon_i \mid \boldsymbol{X}) = c$，其中 $c$ 为常数，不一定为 0。但当回归方程有常数项时，要求 $E(\varepsilon_i \mid \boldsymbol{X}) = 0$ 并不会带来过多限制，因为如果 $E(\varepsilon_i \mid \boldsymbol{X}) = c \neq 0$，总可以把 $c$ 归入常数项。

从 $E(\varepsilon_i \mid \boldsymbol{X}) = 0$ 出发，容易证明扰动项的无条件期望也为 0，因为

$$E(\varepsilon_i) = E_{\boldsymbol{X}} \underbrace{E(\varepsilon_i \mid \boldsymbol{X})}_{=0} = E_{\boldsymbol{X}}(0) = 0 \tag{5.38}$$

① 有时也可加入三次方项，比如 $s^3$，但较少见。

上式使用了迭代期望定律。另外，从 $\mathrm{Cov}(\varepsilon_i, x_{jk}) = 0$ 出发，可以证明扰动项与解释变量“正交”。在线性代数中，如果两个向量的内积为 0，则这两个向量正交。这与在概率统计中，两个随机变量正交的定义有所不同。[①]

**定义** 如果随机变量 $x, y$ 满足 $\mathrm{E}(xy) = 0$，则称 $x, y$ **正交**(orthogonal)。

根据此定义，容易证明解释变量与扰动项正交，因为

$$0 = \mathrm{Cov}(x_{jk}, \varepsilon_i) = \mathrm{E}(x_{jk}\varepsilon_i) - \mathrm{E}(x_{jk}) \underbrace{\mathrm{E}(\varepsilon_i)}_{=0} = \mathrm{E}(x_{jk}\varepsilon_i) \tag{5.39}$$

**假定 5.3** 不存在“严格多重共线性”(strict multicollinearity)，即数据矩阵 $\boldsymbol{X}$ 满列秩(full column rank)。

这意味着，数据矩阵的各列向量为线性无关，即不存在某个解释变量为另一解释变量的倍数，或可由其他解释变量线性表出的情形。换言之，$\boldsymbol{X}$ 中不存在多余的变量。特别地，考虑以下一元回归模型：

$$\ln w_i = \beta_1 + \beta_2 s_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{5.40}$$

其数据矩阵 $\boldsymbol{X}$ 为

$$\boldsymbol{X} = \begin{pmatrix} 1 & s_1 \\ 1 & s_2 \\ \vdots & \vdots \\ 1 & s_n \end{pmatrix} \tag{5.41}$$

数据矩阵 $\boldsymbol{X}$ 满列秩要求解释变量 $s_i$ 不是常数项的固定倍数，即 $s_i$ 应有变动，不能是常数。如果所有个体的教育年限 $s_i$ 都相同，则无法定义 OLS 估计量

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (s_i - \bar{s})(\ln w_i - \overline{\ln w})}{\sum_{i=1}^{n} (s_i - \bar{s})^2} \tag{5.42}$$

其中，$\bar{s}$ 与 $\overline{\ln w}$ 分别为 $s$ 与 $\ln w$ 的样本均值。因为上式分母 $\sum_{i=1}^{n} (s_i - \bar{s})^2 = 0$，故无法估计教育年限对工资对数的作用。

更一般地，对于多元回归，如果 $\boldsymbol{X}$ 满列秩，则 $\boldsymbol{X}'\boldsymbol{X}$ 为正定矩阵(positive definite)，故 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在，可以计算 OLS 估计量 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$。反之，如果 $\boldsymbol{X}$ 不满列秩，则 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在，无法定义 OLS 估计量；此时，称 $\boldsymbol{\beta}$“不可识别”(unidentified)。数据矩阵 $\boldsymbol{X}$ 满列秩只是对数据的最低

---

① 二者也有联系。假设 $\mathrm{E}(xy) = 0$，而 $\{x_i, y_i\}_{i=1}^{n}$ 为来自总体 $(x, y)$ 的 iid 样本，则根据大数定律，当 $n \to \infty$ 时，$\frac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{p} \mathrm{E}(xy) = 0$。有关大数定律，参见第 6 章。

要求。在现实数据中，并不容易出现严格多重共线性。即使出现，Stata 也会自动识别，并去掉多余的变量。

## 5.7 OLS 的小样本性质

显然，OLS 估计量 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 为样本数据的函数，故也是随机变量，其分布称为“抽样分布”(sampling distribution)。在古典线性回归模型的假定 5.1—5.3 之下，OLS 估计量具有以下良好性质。

(1) 线性性。OLS 估计量 $\hat{\boldsymbol{\beta}}$ 为线性估计量(linear estimator)。从 OLS 估计量的表达式 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 可知，$\hat{\boldsymbol{\beta}}$ 可视为 $\boldsymbol{y}$ 的线性组合(将 $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 视为系数矩阵)，故为线性估计量。

(2) 无偏性。$\mathrm{E}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=\boldsymbol{\beta}$，即 $\hat{\boldsymbol{\beta}}$ 不会系统地高估或低估 $\boldsymbol{\beta}$，参见图 5.4。

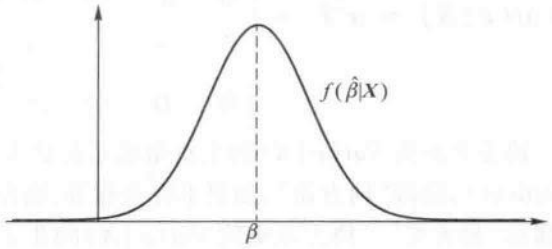



图 5.4 OLS 估计量 $\hat{\beta}$ 的无偏性

**证明：** 抽样误差(sampling error)为

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta}=\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{\boldsymbol{A}}\boldsymbol{\varepsilon}\equiv\boldsymbol{A}\boldsymbol{\varepsilon} \tag{5.43}$$

其中，记 $\boldsymbol{A}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$。给定解释变量 $\boldsymbol{X}$，对上式两边求条件期望，根据严格外生性可得

$$\mathrm{E}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\mid\boldsymbol{X})=\mathrm{E}(\boldsymbol{A}\boldsymbol{\varepsilon}\mid\boldsymbol{X})=\boldsymbol{A}\underbrace{\mathrm{E}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})}_{=\boldsymbol{0}}=\boldsymbol{0} \tag{5.44}$$

移项可得，$\mathrm{E}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=\boldsymbol{\beta}$。在此证明中，严格外生性不可或缺。通过使用迭代期望定律，可进一步证明，无条件期望 $\mathrm{E}(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$，因为

$$\mathrm{E}(\hat{\boldsymbol{\beta}})=\mathrm{E}_{\boldsymbol{X}}\mathrm{E}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=\mathrm{E}_{\boldsymbol{X}}(\boldsymbol{\beta})=\boldsymbol{\beta} \tag{5.45}$$

其中，常数 $\boldsymbol{\beta}$ 的期望仍为它本身。

(3) 估计量 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵。由于 $\boldsymbol{\beta}$ 为常数，故 $\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=\mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\mid\boldsymbol{X})$，因此

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})&=\mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\mid\boldsymbol{X})=\mathrm{Var}(\boldsymbol{A}\boldsymbol{\varepsilon}\mid\boldsymbol{X})\\&=\boldsymbol{A}\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})\boldsymbol{A}'=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned} \tag{5.46}$$

在上式中，使用了协方差矩阵的夹心估计量表达式，而且 $\boldsymbol{A}\equiv(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$，$\boldsymbol{A}'\equiv\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$

(其中,$(\boldsymbol{X}'\boldsymbol{X})^{-1}$为对称矩阵)。此式似乎很复杂,尤其在20世纪计量经济学发展初期,当时计算机还未普及。因此,古典模型对扰动项协方差矩阵 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})$作了最简单的假定,即 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=\sigma^2\boldsymbol{I}_n$,与单位矩阵成正比,称为"球形扰动项"。在球形扰动项的假定下,表达式(5.46)可大大简化:

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned} \tag{5.47}$$

**假定 5.4** 球形扰动项(spherical disturbance),即扰动项满足"同方差"、"无自相关"的性质,故扰动项 $\boldsymbol{\varepsilon}$ 的协方差矩阵可写为

$$\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n = \begin{pmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{pmatrix} \tag{5.48}$$

其中,$\boldsymbol{I}_n$ 为 $n$ 级单位矩阵。协方差矩阵 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})$的主对角线元素都等于 $\sigma^2$,即满足"条件同方差"(conditional homoskedasticity),简称"同方差";如果不完全相等,则存在"条件异方差"(conditional heteroskedasticity),简称"异方差"。协方差矩阵 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})$的非主对角线元素都为0,故不同个体的扰动项之间无"自相关"(autocorrelation)或"序列相关"(serial correlation);反之,则存在自相关。

显然,球形扰动项假定是证明协方差表达式 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$的关键(无偏性则不依赖于球形扰动项)。此表达式虽然很简单,但付出的代价是它只在同方差与无自相关的情况下才成立。如果存在条件异方差,则方差表达式有所不同,应使用"稳健标准误差"(robust standard error),参见第7章。引入球形扰动项假定的另一好处是,可以证明OLS估计量在某种范围内是最有效率的估计量,即方差最小。

(4) 高斯-马尔可夫定理(Gauss-Markov Theorem)。在假定5.1—5.4之下,最小二乘法是最佳线性无偏估计(Best Linear Unbiased Estimator,BLUE),即在所有线性的无偏估计中,最小二乘法的方差最小,参见图5.5。严格来说,记OLS估计量为 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$,而任一线性无偏估计量为 $\tilde{\boldsymbol{\beta}}$,则 $[\mathrm{Var}(\tilde{\boldsymbol{\beta}}\mid\boldsymbol{X})-\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\mid\boldsymbol{X})]$为半正定矩阵。

高斯-马尔可夫定理的核心假设是球形扰动项;如果不满足球形扰动项(比如,存在异方差或自相关),则高斯-马尔可夫定理不成立。此定理的证明需使用较多的矩阵代数,参见陈强(2014,p. 19)。

(5) 对扰动项方差的无偏估计。对于扰动项方差 $\sigma^2=\mathrm{Var}(\varepsilon_i)$,由于$\{\varepsilon_1,\cdots,\varepsilon_n\}$不可观测,将$\{e_1,\cdots,e_n\}$视为其实现值,可得到对 $\sigma^2$ 的估计:

$$s^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n}e_i^2 \tag{5.49}$$

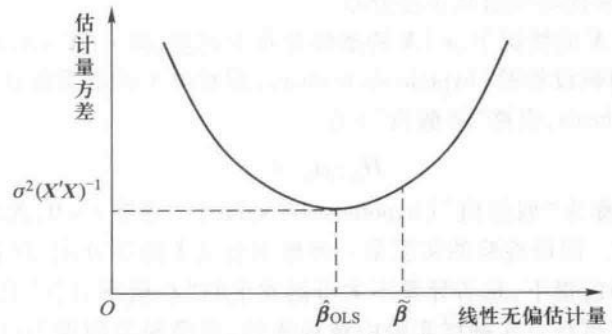



图 5.5 OLS 估计量为 BLUE

其中，$(n-K)$ 为自由度。在上式中，为什么除以 $(n-K)$ 而不除以 $n$？这是因为，虽然共有 $n$ 个残差 $\{e_1, e_2, \cdots, e_n\}$，随机变量 $\{e_1, e_2, \cdots, e_n\}$ 必须满足 $K$ 个正规方程 $\boldsymbol{X'e}=\boldsymbol{0}$，故只有其中 $(n-K)$ 个残差是（自由）独立的。经过自由度校正后，才是“无偏估计”（unbiased estimator），即 $\mathrm{E}(s^2)=\sigma^2$[①]。当然，如果样本容量 $n$ 很大，当 $n\to\infty$ 时，则 $\frac{n-K}{n}\to 1$，是否进行“小样本校正”（small sample adjustment）并无多大差别。

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”，它衡量回归方程扰动项的波动幅度。

因此，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵 $\sigma^2(\boldsymbol{X'X})^{-1}$ 可用 $s^2(\boldsymbol{X'X})^{-1}$ 来估计。特别地，$\hat{\boldsymbol{\beta}}$ 的第 $k$ 个分量 $\hat{\beta}_k$ 的估计方差为 $s^2(\boldsymbol{X'X})^{-1}_{kk}$，其中 $(\boldsymbol{X'X})^{-1}_{kk}$ 表示矩阵 $(\boldsymbol{X'X})^{-1}$ 的 $(k,k)$ 元素，即主对角线上的第 $k$ 个元素。称 $\sqrt{s^2(\boldsymbol{X'X})^{-1}_{kk}}$ 为 OLS 估计量 $\hat{\beta}_k$ 的“标准误差”（standard error），简称“标准误”，记为 $\mathrm{SE}(\hat{\beta}_k)$，即

$$\mathrm{SE}(\hat{\beta}_k) \equiv \sqrt{s^2(\boldsymbol{X'X})^{-1}_{kk}} \tag{5.50}$$

更一般地，称对某统计量的标准差的估计值（estimated standard deviation）为该统计量的“标准误”（standard error），作为对统计量估计误差的度量。通常，得到参数的点估计（point estimate）后，还需给出相应的标准误，才能知道此点估计的准确程度（或不确定性）。

## 5.8 对单个系数的 $t$ 检验

计量经济学中的统计推断（statistical inference）方法，可分为两大类，即“小样本理论”（small sample theory）与“大样本理论”（large sample theory）。无论样本容量是多少，小样本理论都成立，不需要让样本容量 $n\to\infty$，故也称“有限样本理论”（finite sample theory）。反之，大样本理论要求 $n\to\infty$，适用于较大的样本容量，将在第 6 章介绍。本章介绍小样本理论。小样本理论虽然适用于任何样本容量，但代价是不容易推导其统计量的分布，为此需对随机变量的概率分布作很

① 证明参见陈强（2014，p.19）。

强的具体假定，比如要求扰动项服从正态分布。

**假定 5.5** 在给定 $\boldsymbol{X}$ 的情况下，$\boldsymbol{\varepsilon} \mid \boldsymbol{X}$ 的条件分布为正态，即 $\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$。

首先考虑最简单的假设检验(hypothesis testing)，即对单个回归系数 $\beta_k$ 进行检验。需要检验的“原假设”(null hypothesis，也称“零假设”)为

$$H_0: \beta_k = c \tag{5.51}$$

其中，$c$ 为给定常数，也称为“假想值”(hypothesized value)。通常 $c=0$，此时即检验变量 $x_{ik}$ 的系数是否显著地不等于 0。假设检验的实质是一种概率意义上的反证法，即首先假定原假设成立，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”在一次抽样的样本中出现。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，而接受“替代假设”(alternative hypothesis，也称“备择假设”)：

$$H_1: \beta_k \neq c \tag{5.52}$$

替代假设“$H_1: \beta_k \neq c$”也称为“双边替代假设”(two-sided alternative hypothesis)，因为它既包括 $\beta_k > c$，也包括 $\beta_k < c$ 的情形。相应地，这类检验称为“双边检验”(two-sided test)。

直观上，如果未知参数 $\beta_k$ 的估计值 $\hat{\beta}_k$ 离 $c$ 较远，则应倾向于拒绝原假设。使用此原理的这类统计检验称为“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。

由于扰动项 $\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$，而 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\varepsilon}$ 为 $\boldsymbol{\varepsilon}$ 的线性函数(其中 $\boldsymbol{A} \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$)，故 $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \mid \boldsymbol{X}$ 也服从正态分布。进一步，由于 $\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}) = \boldsymbol{0}$，$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$，故

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \mid \boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{5.53}$$

单独考虑上式的第 $k$ 个分量，则有

$$(\hat{\beta}_k - \beta_k) \mid \boldsymbol{X} \sim N(0, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}) \tag{5.54}$$

其中，$(\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}$ 为矩阵 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的 $(k,k)$ 元素(即主对角线上第 $k$ 个元素)，而 $\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}$ 为 $\hat{\beta}_k$ 的方差。在原假设“$H_0: \beta_k = c$”成立的情况下，上式可写为

$$(\hat{\beta}_k - c) \mid \boldsymbol{X} \sim N(0, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}) \tag{5.55}$$

如果 $\sigma^2$ 已知，则标准化的统计量服从标准正态分布：

$$z_k \equiv \frac{\hat{\beta}_k - c}{\sqrt{\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}} \sim N(0,1) \tag{5.56}$$

但通常 $\sigma^2$ 未知，称为“厌恶参数”(nuisance parameter)。虽然我们对 $\sigma^2$ 不感兴趣，但 $\sigma^2$ 却出现在表达式(5.56)中，故名。合格的“检验统计量”(test statistic)必须满足两个条件。首先，它应能够根据样本数据计算出来；其次，它的概率分布是已知的。对于上式的 $z_k$ 统计量，虽然已知其分布为标准正态，但由于不知道 $\sigma^2$，故无法根据数据计算 $z_k$ 统计量的样本观测值。为此，以 $\sigma^2$ 的估计量 $s^2$ 替代 $\sigma^2$，即可得到以下 $t$ 统计量($t$-statistic)。

**定理($t$ 统计量的分布)**① 在假定 5.1—5.5 均满足,且原假设"$H_0: \beta_k = c$"也成立的情况下,$t$ 统计量服从自由度为$(n-K)$的 $t$ 分布:

$$t_k \equiv \frac{\hat{\beta}_k - c}{\mathrm{SE}(\hat{\beta}_k)} \sim t(n-K) \tag{5.57}$$

其中,$\mathrm{SE}(\hat{\beta}_k) \equiv \sqrt{s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}_{kk}}$ 为 $\hat{\beta}_k$ 的标准误。更一般地,$t$ 统计量的通用公式为

$$t \equiv \frac{\text{估计量} - \text{假想值}}{\text{估计量的标准误}} \tag{5.58}$$

从上式可知,$t$ 统计量度量估计量$(\hat{\beta}_k)$离假想值$(c)$的距离,并以估计量的标准误 $\mathrm{SE}(\hat{\beta}_k)$作为距离的度量单位,即此距离为标准误的多少倍。

### 1. $t$ 检验的步骤

第一步:计算 $t$ 统计量,记其具体取值为 $t_k$。如果 $H_0$ 为真,则 $|t_k|$ 很大的概率将很小(为小概率事件),不应在抽样中观测到。因此,如果 $|t_k|$ 很大,则 $H_0$ 较不可信。

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

$$P\{T > t_{\alpha/2}(n-K)\} = P\{T < -t_{\alpha/2}(n-K)\} = \alpha/2 \tag{5.59}$$

其中,随机变量 $T \sim t(n-K)$。上式表明,随机变量 $T > t_{\alpha/2}(n-K)$,或 $T < -t_{\alpha/2}(n-K)$的概率都是 $\alpha/2$,参见图 5.6。在实践中,通常取 $\alpha = 5\%$,则 $\alpha/2 = 2.5\%$。有时也使用 $\alpha = 1\%$ 或 $\alpha = 10\%$。

第三步:如果 $|t_k| \geqslant t_{\alpha/2}(n-K)$,则 $t_k$ 落入"拒绝域"(rejection region),故拒绝 $H_0$。反之,如果 $|t_k| < t_{\alpha/2}(n-K)$,则 $t_k$ 落入"接受域"(acceptance region),故接受 $H_0$。因为拒绝域分布在 $t$ 分布两边,故称为"双边检验"(two-sided test)。

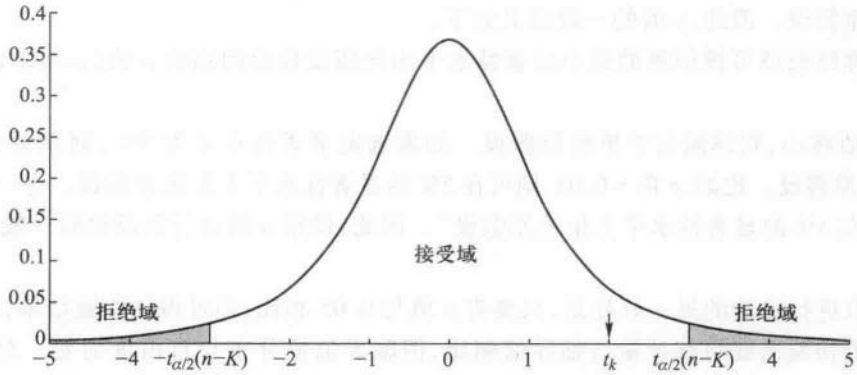



图 5.6 双边 $t$ 检验的临界值与拒绝域

① 证明此定理的大致思路如下。将 $t$ 统计量变形后,证明其分子服从标准正态分布,而分母为卡方分布除以其自由度的开平方,然后使用 OLS 的正交性证明分子与分母相互独立,则根据 $t$ 分布的定义,该统计量服从 $t$ 分布。详见陈强(2014, p. 21)。

### 2. 计算 $p$ 值

假设检验的基本逻辑就是概率意义上的反证法，即如果在一次抽样中看到很不可能发生的小概率事件，则拒绝原假设。至于观测到样本数据的发生概率究竟小到何种程度，可通过下面的 $p$ 值来度量。

在双边 $t$ 检验中，给定 $t$ 统计量的样本观测值 $t_k$，此假设检验问题的 $p$ 值（probability value，即 $p$-value）为

$$p\text{ 值} \equiv P(|T| > |t_k|) \tag{5.60}$$

其中，随机变量 $T \sim t(n-K)$。直观来看，给定 $t$ 统计量 $t_k$，则 $p$ 值衡量比 $|t_k|$ 更大的 $t$ 分布两端的尾部概率，参见图 5.7。

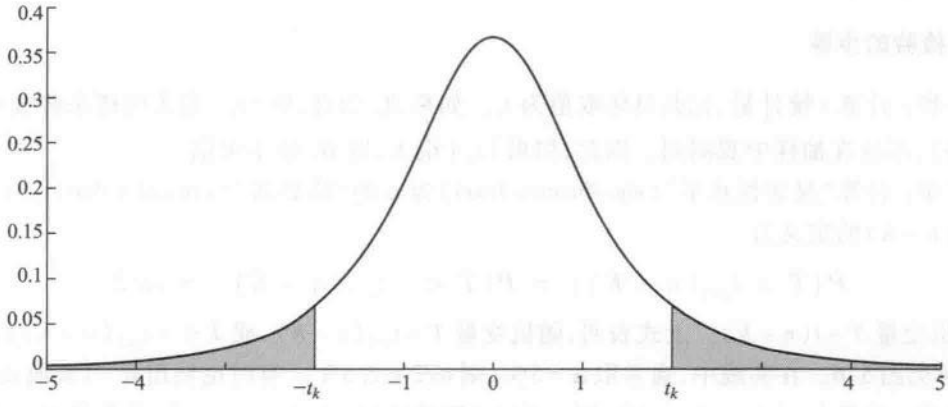



图 5.7 双边 $t$ 检验的 $p$ 值

如果 $p$ 值为 0.05，则正好可以在 5% 的显著性水平上拒绝原假设，但无法在 4.9% 的显著性水平上拒绝原假设。因此，$p$ 值的一般定义如下。

**定义** 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）。

显然，$p$ 值越小，则越倾向于拒绝原假设。如果选定显著性水平为 5%，则只要 $p$ 值比 0.05 小，即可拒绝原假设。比如，$p$ 值 =0.03，则可在 5% 的显著性水平上拒绝原假设。进一步，“$p$ 值 = 0.03”还可“在 3% 的显著性水平上拒绝原假设”。因此，使用 $p$ 值进行假设检验一般比临界值更有信息量。

使用 $p$ 值进行检验的另一好处是，只要将 $p$ 值与 0.05 相比，即可得到检验结果，操作十分简便。而传统的检验需要将统计量与临界值相比，但临界值依分布与自由度而变。总之，当 Stata 直接给出 $p$ 值时，就不需要知道临界值了。由于当代的统计检验一般在计算机中进行（比如 Stata），故本书未在附录提供常规的概率分布表（如需要，可参见标准的概率统计教材）。

### 3. 计算置信区间

有时只进行点估计还不够，还希望进行区间估计，即参数最可能的取值范围。假设“置信

度"(confidence level)为$(1-\alpha)$(比如 $\alpha=5\%$,则 $1-\alpha=95\%$),即要找到"置信区间"(confidence interval),使得该区间覆盖真实参数 $\beta_k$ 的概率为$(1-\alpha)$。

由于 $t_k=\dfrac{\hat{\beta}_k-\beta_k}{\mathrm{SE}(\hat{\beta}_k)}\sim t(n-K)$,故 $t$ 统计量落入接受域的概率为$(1-\alpha)$:

$$P\left\{-t_{\alpha/2}<\frac{\hat{\beta}_k-\beta_k}{\mathrm{SE}(\hat{\beta}_k)}<t_{\alpha/2}\right\}=1-\alpha \tag{5.61}$$

其中,$t_{\alpha/2}$为显著性水平为 $\alpha$ 的临界值。将上式中的不等式变形可得

$$P\{\hat{\beta}_k-t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k)<\beta_k<\hat{\beta}_k+t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k)\}=1-\alpha \tag{5.62}$$

由此可知,$\beta_k$ 的置信区间为

$$[\hat{\beta}_k-t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k),\ \hat{\beta}_k+t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k)] \tag{5.63}$$

此置信区间以点估计 $\hat{\beta}_k$ 为中心,区间半径为 $t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k)$,参见图 5.8。显然,标准误 $\mathrm{SE}(\hat{\beta}_k)$ 越大,则对 $\beta_k$ 的估计越不准确,故置信区间也越宽。本质上,置信区间是随机区间,随着样本不同而不同。如果置信度为95%,抽样 100 次,得到 100 个置信区间,大约 95 个置信区间能覆盖到真实参数 $\beta_k$。

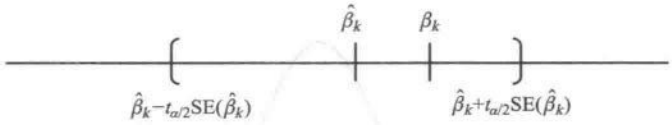



图 5.8 置信区间的示意图

#### 4. 单边检验

假设检验有时也进行单边检验(one-sided test)。不失一般性,考虑原假设为 $H_0:\beta_k=0$,而替代假设为 $H_1:\beta_k>0$,比如,从理论上认为解释变量 $x_k$ 对 $y$ 的作用不可能为负。此时,仍可计算 $t$ 统计量:

$$t_k\equiv\frac{\hat{\beta}_k}{\mathrm{SE}(\hat{\beta}_k)}\sim t(n-K) \tag{5.64}$$

显然,如果此 $t$ 统计量很大,则倾向于拒绝原假设;而如果此 $t$ 统计量很小(比如为负数),则倾向于接受原假设。因此,拒绝域只在概率分布的最右边一侧。给定显著性水平 $\alpha$ 后,需要计算的临界值为 $t_\alpha(n-K)$,使得取值大于此临界值的概率为 $\alpha$:

$$P\{T>t_\alpha(n-K)\}=\alpha \tag{5.65}$$

其中,随机变量 $T\sim t(n-K)$。如果要计算此单边检验的 $p$ 值,则为比统计量 $t_k$ 更大的右侧尾部概率(参见图 5.9):

$$p\text{ 值}\equiv P(T>t_k) \tag{5.66}$$

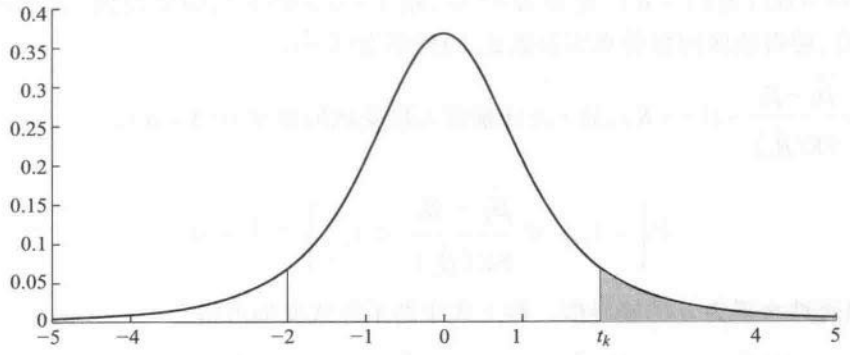



图 5.9 单边右侧 $t$ 检验的 $p$ 值

由于拒绝域只在分布的右侧尾部,故也称"单边右侧检验"(one-sided right-tail test)。

相应地,如果原假设为 $H_0:\beta_k=0$,而替代假设为 $H_1:\beta_k<0$,则为"单边左侧检验"(one-sided left-tail test)。此时,$t$ 统计量越小(比如为负数),则越倾向于拒绝原假设,故拒绝域只在分布的左侧尾部。因此,对于单边左侧检验,计算 $p$ 值的公式为(参见图 5.10):

$$p\text{ 值} \equiv P(T < t_k) \tag{5.67}$$

其中,随机变量 $T \sim t(n-K)$。

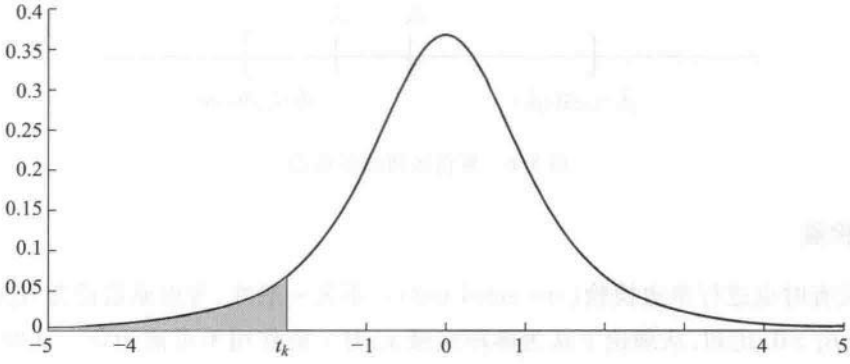



图 5.10 单边左侧 $t$ 检验的 $p$ 值

**5. 第 I 类错误与第 II 类错误**

在进行假设检验时,可能犯以下两类错误。

**定义** **第 I 类错误**(Type Ⅰ error)指的是,虽然原假设为真,但却根据观测数据作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

$$P(\text{拒绝 } H_0 \mid H_0) = P(\text{检验统计量落入拒绝域} \mid H_0) = \alpha \tag{5.68}$$

其中,$\alpha$ 正是此检验的显著性水平。

**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),但却根据观测数据作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

$$P(\text{接受 } H_0 \mid H_1) = P(\text{检验统计量落入接受域} \mid H_1) \tag{5.69}$$

在 $\beta_k$ 可能取值的参数空间中，通常 $H_0$ 仅包含一个点（比如 $\beta_k = 0$），故很容易计算第 I 类错误的发生概率，即 $P(\text{拒绝 } H_0 \mid H_0) = \alpha$。反之，替代假设 $H_1$ 则一般包括许多点（比如 $\beta_k \neq 0$），故不容易计算第 II 类错误的发生概率。而且，第 I 类错误与第 II 类错误存在此消彼长的关系，即如果减少第 I 类错误的发生概率，则第 II 类错误的发生概率必然增加；反之亦然。一般来说，如果要同时减少第 I 类错误与第 II 类错误的发生概率，则必须增加样本容量。因此，在进行假设检验时，一般先指定可接受的发生第 I 类错误的最大概率，即显著性水平 $\alpha$（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

$$\text{功效} = 1 - P(\text{接受 } H_0 \mid H_1) = P(\text{拒绝 } H_0 \mid H_1) \tag{5.70}$$

换言之，功效为在原假设为假的情况下，拒绝原假设的概率。在进行检验时，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，则比较理直气壮，因为知道犯错概率（显著性水平）。反之，如果接受原假设，则比较没有把握，因为通常不知犯错概率（可能较高）。

## 5.9 对线性假设的 $F$ 检验

我们常想知道整个回归方程是否显著，即除常数项以外，所有解释变量的回归系数是否都为 0。这需要检验以下原假设：

$$H_0: \beta_2 = \cdots = \beta_K = 0 \tag{5.71}$$

其中，$\beta_1$ 为常数项。此原假设等价于对 $(K-1)$ 个约束条件进行联合检验（joint test）：

$$H_0: \beta_2 = 0, \quad \beta_3 = 0, \cdots, \quad \beta_K = 0 \tag{5.72}$$

**例** 对于模型 $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$，检验以下两个约束：

$$H_0: \beta_2 = \beta_3, \quad \beta_4 = 0 \tag{5.73}$$

将此原假设的两个约束写成向量形式，经整理可得

$$H_0: \begin{pmatrix} \beta_2 - \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.74}$$

进一步，上式可写为

$$H_0: \underbrace{\begin{pmatrix} \beta_2 - \beta_3 \\ \beta_4 \end{pmatrix}}_{} = \underbrace{\begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{R} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{r} \tag{5.75}$$

更一般地，考虑检验 $m$ 个线性假设是否同时成立：

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

其中，$\boldsymbol{r}$ 为 $m$ 维列向量（$m<K$），$\boldsymbol{R}$ 为 $m\times K$ 维矩阵，而且 $\text{rank}(\boldsymbol{R})=m$，即 $\boldsymbol{R}$ 满行秩，没有多余或自相矛盾的行或方程。在上例中，$\boldsymbol{R}=\begin{pmatrix}0 & 1 & -1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$，而 $\boldsymbol{r}=\begin{pmatrix}0\\0\end{pmatrix}$。

根据沃尔德检验原理，由于 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的估计量，故如果 $H_0$ 成立，则 $(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$ 应比较接近 $\boldsymbol{0}$（零向量）。这种接近程度可用其二次型来衡量，比如

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'[\text{Var}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}) \tag{5.76}$$

其中，$(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$ 的协方差矩阵可写为

$$\begin{aligned}\text{Var}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}) &= \text{Var}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) && \text{（去掉常数，方差不变）}\\ &= \boldsymbol{R}\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{R}' && \text{（夹心估计量的公式）}\\ &= \sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' && \text{（因为 } \text{Var}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\text{）}\end{aligned} \tag{5.77}$$

其中，$\sigma^2$ 可由 $s^2$ 来估计，故有如下定理。

**定理（$F$ 统计量的分布）**[①] 在假定 5.1—5.5 均满足，且原假设 $H_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$ 也成立的情况下，则 $F$ 统计量服从自由度为 $(m,n-K)$ 的 $F$ 分布：

$$F \equiv \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})/m}{s^2} \sim F(m,n-K) \tag{5.78}$$

**$F$ 检验的步骤如下：**

第一步：计算 $F$ 统计量。如果 $H_0$ 为真，则“$F$ 统计量很大”的概率将很小（为小概率事件），不应在一次抽样中观测到。因此，如果 $F$ 统计量很大，则 $H_0$ 较不可信。

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

$$P\{\tilde{F} > F_\alpha(m,n-K)\} = \alpha \tag{5.79}$$

其中，随机变量 $\tilde{F}\sim F(m,n-K)$。上式表明，$\tilde{F}$ 大于临界值 $F_\alpha(m,n-K)$ 的概率恰好为 $\alpha$。

第三步：如果 $F$ 统计量大于临界值 $F_\alpha(m,n-K)$，即落入右边拒绝域，则拒绝 $H_0$；反之，如果 $F$ 统计量小于临界值，即落入接受域，则接受 $H_0$，参见图 5.11。

对于 $F$ 检验，也可使用 $p$ 值来进行。给定 $F$ 统计量的样本观测值，此假设检验问题的 $p$ 值为比 $F$ 统计量更大的 $F$ 分布的右侧尾部概率，即

$$p\text{ 值} \equiv P(\tilde{F} > F) \tag{5.80}$$

---

① 证明参见陈强(2014, p.24)。

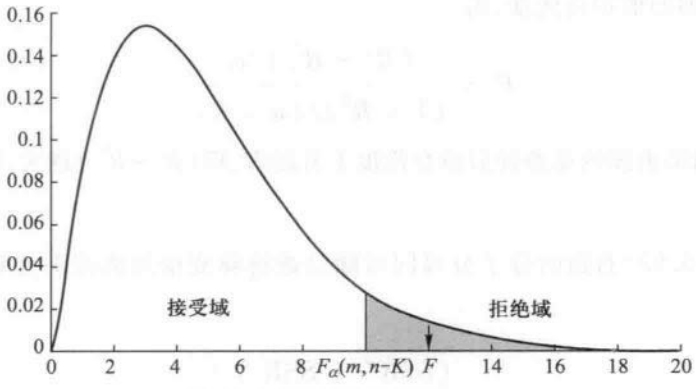



图 5.11 F检验

其中，随机变量$\widetilde{F} \sim F(m,n-K)$，而 $F$ 为 $F$ 统计量的取值。

## 5.10 F统计量的似然比原理表达式

在作假设检验时，如果接受原假设，则可将此原假设作为约束条件，代入最小二乘法的最优化问题。使用约束条件下的最小二乘法，即“约束最小二乘法”（Restricted OLS 或 Constrained OLS），可得到 $F$ 统计量的另一方便表达式。考虑以下约束极值问题：

$$\min_{\hat{\boldsymbol{\beta}}} \text{SSR}(\hat{\boldsymbol{\beta}}) \quad s.t.\ \boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{r} \tag{5.81}$$

其中，$\text{SSR}(\hat{\boldsymbol{\beta}})$ 为残差平方和，是 $\hat{\boldsymbol{\beta}}$ 的函数；而 $\hat{\boldsymbol{\beta}}$ 还需满足约束条件 $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{r}$（$s.t.$ 表示受约束，即 subject to）。换言之，$\hat{\boldsymbol{\beta}}$ 并不能任意取值，而只能在所有满足 $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{r}$ 的子集中，选择使残差平方和 $\text{SSR}(\hat{\boldsymbol{\beta}})$ 最小化的 $\hat{\boldsymbol{\beta}}$。

如果 $H_0: \boldsymbol{R\beta} = \boldsymbol{r}$ 正确，则加上此约束不应使残差平方和增大很多。记无约束回归的残差平方和为 SSR，而有约束回归的残差平方和为 $\text{SSR}^*$。这意味着，在 $H_0$ 正确的情况下，$(\text{SSR}^* - \text{SSR})$ 不应很大。由此可构造如下 $F$ 统计量。通过求解此约束极值问题，可以证明[①]：

$$F = \frac{(\text{SSR}^* - \text{SSR})/m}{\text{SSR}/(n-K)} \tag{5.82}$$

其中，$m$ 为约束条件个数（即矩阵 $\boldsymbol{R}$ 的行数），$n$ 为样本容量，而 $K$ 为参数个数（即 $\boldsymbol{\beta}$ 的维度）。此 $F$ 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，LR）。

$F$ 统计量的似然比表达式(5.82)，也可以通过拟合优度来表示。记 $R^2$ 为无约束回归的拟合

① 参见陈强(2014，p.29)。

优度，而 $R_*^2$ 为约束回归的拟合优度，则

$$F = \frac{(R^2 - R_*^2)/m}{(1 - R^2)/(n - K)} \tag{5.83}$$

从上式可知，如果去掉约束条件后拟合优度上升越多，即 $(R^2 - R_*^2)$ 越大，则越应该拒绝约束条件成立的原假设。

**证明**：在方程(5.82)右边的分子分母同时除以被解释变量的离差平方和 $\text{TSS} \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2$ 可得

$$F = \frac{\left.\dfrac{(\text{SSR}^* - \text{SSR})}{\text{TSS}}\right/ m}{\left.\dfrac{\text{SSR}}{\text{TSS}}\right/ (n - K)} \tag{5.84}$$

由于 $\dfrac{\text{SSR}}{\text{TSS}} = 1 - R^2$，而 $\dfrac{\text{SSR}^*}{\text{TSS}} = 1 - R_*^2$，故

$$F = \frac{[(1 - R_*^2) - (1 - R^2)]/m}{(1 - R^2)/(n - K)} = \frac{(R^2 - R_*^2)/m}{(1 - R^2)/(n - K)} \tag{5.85}$$

作为应用，下面考虑一个特殊情形，即检验整个回归方程的显著性。

**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 $H_0: \beta_2 = \cdots = \beta_K = 0$ 的 $F$ 统计量等于 $\dfrac{R^2/(K-1)}{(1-R^2)/(n-K)}$。

**证明**：对于 $H_0: \beta_2 = \cdots = \beta_K = 0$，共有 $(K-1)$ 个约束，故在表达式(5.83)中，$m = (K-1)$。另外，当原假设成立时，$y_i = \beta_1 + \varepsilon_i$，故约束回归只是对常数项回归，因此 $R_*^2 = 0$（参见第 4 章习题）。将 $m = (K-1)$ 与 $R_*^2 = 0$ 代入表达式(5.85)即得证。

此命题表明了 $F$ 统计量与拟合优度 $R^2$ 之间的关系。但 $R^2$ 并非决定 $F$ 统计量的唯一因素；$F$ 统计量还取决于样本容量 $n$，以及解释变量个数 $K$。

## 5.11 预测

*We have two classes of forecasters: Those who don't know... and those who don't know they don't know.* ——John Kenneth Galbraith

有时也用计量模型进行预测（prediction 或 forecasting），即给定解释向量 $\boldsymbol{x}_0$ 的（未来）取值，预测被解释变量 $y_0$ 的取值。假设计量模型对所有观测值都成立（包括外推到未来的观测值），

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0 \tag{5.86}$$

记 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的 OLS 估计量，对 $y_0$ 作点预测为

$$\hat{y}_0 \equiv \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \tag{5.87}$$

因此,"预测误差"(prediction error)$(\hat{y}_0 - y_0)$可写为

$$\hat{y}_0 - y_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - (\boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0) = \boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0 \tag{5.88}$$

容易证明,$\hat{y}_0$ 为"无偏预测"(unbiased predictor),即用 $\hat{y}_0$ 作为 $y_0$ 的预测值不会系统地高估或低估 $y_0$,因为

$$\mathrm{E}(\hat{y}_0 - y_0) = \boldsymbol{x}_0'\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \mathrm{E}(\varepsilon_0) = 0 \tag{5.89}$$

其中,由于 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计,故 $\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 0$。有时候,仅作点预测是不够的,还希望知道此预测的置信区间。为此,计算预测误差$(\hat{y}_0 - y_0)$的方差为:

$$\begin{aligned}
\mathrm{Var}(\hat{y}_0 - y_0) &= \mathrm{Var}[\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0] \\
&= \mathrm{Var}(\varepsilon_0) + \mathrm{Var}[\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= \sigma^2 + \mathrm{Var}[\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}})] && \text{(去掉常数,方差不变)} \\
&= \sigma^2 + \boldsymbol{x}_0'\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0 && \text{(夹心估计量的公式)} \\
&= \sigma^2 + \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 && \text{(因为 } \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\text{)}
\end{aligned} \tag{5.90}$$

其中,假设 $\varepsilon_0$ 与 $\hat{\boldsymbol{\beta}}$ 不相关(估计 $\hat{\boldsymbol{\beta}}$ 没用到 $\varepsilon_0$ 的信息)。由上式可知,预测误差的方差有两个来源,即抽样误差 $\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$(不能精确知道参数 $\boldsymbol{\beta}$),以及 $y_0$ 本身的不确定性($\varepsilon_0$ 的方差 $\sigma^2$)。如果样本很大,则抽样误差将很小;但扰动项的方差 $\sigma^2$ 始终存在。将方程(5.90)中的 $\sigma^2$ 用 $s^2$ 来替代,并开平方,则可得到预测误差$(\hat{y}_0 - y_0)$的标准误:

$$\mathrm{SE}(\hat{y}_0 - y_0) = s\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}x_0} \tag{5.91}$$

由此可得 $t$ 统计量:

$$\frac{\hat{y}_0 - y_0}{\mathrm{SE}(\hat{y}_0 - y_0)} \sim t(n - K) \tag{5.92}$$

进一步,$y_0$ 的置信度为$(1-\alpha)$的置信区间为

$$(\hat{y}_0 - t_{\alpha/2}\mathrm{SE}(\hat{y}_0 - y_0),\quad \hat{y}_0 + t_{\alpha/2}\mathrm{SE}(\hat{y}_0 - y_0)) \tag{5.93}$$

其中,$t_{\alpha/2}$为显著性水平为 $\alpha$ 的 $t(n-K)$分布的双边检验临界值。由于预测通常有时间维度,故将在第 13 章使用时间序列进行预测。

## 5.12 多元回归的 Stata 实例

在 Stata 中进行多元回归的命令为

```
. regress y x1 x2 x3
```

其中，“y”为被解释变量，而“x1 x2 x3”为解释变量。以数据集 grilic. dta 为例，该数据集包括 758 名美国年轻男子的数据。对以下方程进行多元回归估计：

$$\ln w = \beta_1 + \beta_2 s + \beta_3 expr + \beta_4 tenure + \beta_5 smsa + \beta_6 rns + \varepsilon \tag{5.94}$$

其中，被解释变量为 ln $w$（工资对数），主要解释变量包括 $s$（教育年限）、$expr$（工龄）、$tenure$（在现单位工作年限）、$smsa$（是否住在大城市）以及 $rns$（是否住在美国南方）。为估计方程(5.94)，可输入如下命令

```
. reg lnw s expr tenure smsa rns
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 5, 752) = | 81.75 |
| Model | 49.0478814 | 5 | 9.80957628 | Prob > F = | 0.0000 |
| Residual | 90.2382684 | 752 | .119997697 | R-squared = | 0.3521 |
| | | | | Adj R-squared = | 0.3478 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .34641 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .102643 | .0058488 | 17.55 | 0.000 | .0911611 | .114125 |
| expr | .0381189 | .0063268 | 6.02 | 0.000 | .0256986 | .0505392 |
| tenure | .0356146 | .0077424 | 4.60 | 0.000 | .0204153 | .0508138 |
| smsa | .1396666 | .0280821 | 4.97 | 0.000 | .0845379 | .1947954 |
| rns | -.0840797 | .0287973 | -2.92 | 0.004 | -.1406124 | -.0275471 |
| _cons | 4.103675 | .085097 | 48.22 | 0.000 | 3.936619 | 4.270731 |

其中，“_cons”表示常数项，“R - squared”显示 $R^2 = 0.352\ 1$，“Adj R - squared”显示 $\overline{R}^2 = 0.347\ 8$。表上方的回归结果显示，残差平方和 $\sum_{i=1}^{n} e_i^2 = 90.24$，方程的标准误差（Root MSE）为 $s = 0.346\ 41$。检验整个方程显著性的 $F$ 统计量为 81.75，其对应的 $p$ 值（Prob > F）为 0.000 0，表明这个回归方程整体是高度显著的。

所有解释变量（包括常数项）的回归系数的 $p$ 值（P > | t |）都小于 0.01，故均在 1% 水平上显著，而且符号与理论预期一致。其中，教育年限（s）的系数估计值为 0.103，即教育投资回报率为 10.3%。工龄（*expr*）与在现单位工作年限（*tenure*）的回报率分别为 3.8% 与 3.6%（可视为在职培训的回报率），小于正规教育的回报率。住在大城市的回报率高达 14.0%，甚至高于一年教育的回报率，说明了环境的重要性。变量 *rns* 的系数为 -0.084，表明在给定其他变量的情况下，南方居民的工资比北方居民低 8.4%。常数项的估计值为 4.104，这意味着未受任何教育（$s = 0$）、也无工作经验（$expr = tenure = 0$）、不住在大城市（$smsa = 0$），且身在南方（$rns = 0$）的年轻男子预期工资对数为 4.104。

如果要显示回归系数的协方差矩阵，可输入命令

```
. vce
```

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

Covariance matrix of coefficients of regress model

| e(V) | s | expr | tenure | smsa | rns | _cons |
|---|---|---|---|---|---|---|
| s | .00003421 | | | | | |
| expr | 8.660e-06 | .00004003 | | | | |
| tenure | -3.997e-08 | -.00001107 | .00005994 | | | |
| smsa | -.0000144 | 3.261e-06 | -7.819e-06 | .00078861 | | |
| rns | 8.524e-06 | 7.334e-07 | 7.259e-06 | .00012486 | .00082928 | |
| _cons | -.00046567 | -.00016778 | -.00008646 | -.00038746 | -.00043997 | .0072415 |

上表中的主对角线元素为各回归系数的方差，而非主对角线元素则为相应的协方差。尽管在上述回归中常数项很显著，为演示目的，下面加上选择项“noconstant”，进行无常数项回归：

```
. reg lnw s expr tenure smsa rns,noc
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 5, 753) = | 9902.73 |
| Model | 24282.9531 | 5 | 4856.59061 | Prob > F = | 0.0000 |
| Residual | 369.293555 | 753 | .490429688 | R-squared = | 0.9850 |
| | | | | Adj R-squared = | 0.9849 |
| Total | 24652.2466 | 758 | 32.5227528 | Root MSE = | .70031 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .3665333 | .0041742 | 87.81 | 0.000 | .3583389 | .3747277 |
| expr | .1331991 | .0121535 | 10.96 | 0.000 | .1093403 | .1570578 |
| tenure | .0846129 | .0155168 | 5.45 | 0.000 | .0541515 | .1150743 |
| smsa | .3592339 | .0560206 | 6.41 | 0.000 | .2492588 | .4692089 |
| rns | .1652489 | .0572715 | 2.89 | 0.004 | .0528181 | .2776796 |

从上表可知，根据无常数项回归的估计，教育投资回报率高达每年 36.7%，这显然不合理。由于常数项很显著，故忽略常数项将导致估计偏差，得不到一致估计。即使真实模型不包括常数项，在回归中加入常数项，也不会导致不一致的估计，故危害较小。反之，如果真实模型包括常数项，但在回归时被忽略了，则可能导致严重的估计偏差。因此，一般建议在回归中包括常数项。

如果只对南方居民的子样本进行回归，可使用虚拟变量 *rns*：

```
. reg lnw s expr tenure smsa if rns
```

| Source | SS | df | MS | Number of obs = | 204 |
|---|---|---|---|---|---|
| | | | | F( 4, 199) = | 36.07 |
| Model | 17.603542 | 4 | 4.40088551 | Prob > F = | 0.0000 |
| Residual | 24.2783596 | 199 | .122001807 | R-squared = | 0.4203 |
| | | | | Adj R-squared = | 0.4087 |
| Total | 41.8819016 | 203 | .206314786 | Root MSE = | .34929 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .1198242 | .0113156 | 10.59 | 0.000 | .0975103 | .1421381 |
| expr | .0451903 | .0122572 | 3.69 | 0.000 | .0210197 | .069361 |
| tenure | .0092643 | .0156779 | 0.59 | 0.555 | -.0216518 | .0401804 |
| smsa | .1746563 | .0506762 | 3.45 | 0.001 | .0747251 | .2745876 |
| _cons | 3.806148 | .1586202 | 24.00 | 0.000 | 3.493356 | 4.11894 |

如果只对北方居民的子样本进行回归,可使用命令:

```
. reg lnw s expr tenure smsa if ~rns
```

其中,“~”表示逻辑的“否”(not)运算。

| Source | SS | df | MS | Number of obs = | 554 |
|---|---|---|---|---|---|
| | | | | F( 4, 549) = | 62.45 |
| Model | 29.486457 | 4 | 7.37161426 | Prob > F = | 0.0000 |
| Residual | 64.8019636 | 549 | .118036364 | R-squared = | 0.3127 |
| | | | | Adj R-squared = | 0.3077 |
| Total | 94.2884207 | 553 | .170503473 | Root MSE = | .34356 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .0944787 | .0068365 | 13.82 | 0.000 | .0810498 | .1079076 |
| expr | .0358675 | .0073558 | 4.88 | 0.000 | .0214184 | .0503165 |
| tenure | .0455117 | .0088792 | 5.13 | 0.000 | .0280703 | .0629531 |
| smsa | .1199364 | .0337443 | 3.55 | 0.000 | .0536526 | .1862202 |
| _cons | 4.214014 | .0995796 | 42.32 | 0.000 | 4.018411 | 4.409618 |

根据以上两表的结果,南方居民的教育投资回报率为12.0%,反而高于北方居民9.4%的教育投资回报率。如果只对中学以上(s >=12)的子样本进行回归,可输入命令:

```
. reg lnw s expr tenure smsa rns if s >=12
```

| Source | SS | df | MS | Number of obs = | 679 |
|---|---|---|---|---|---|
| | | | | F( 5, 673) = | 69.81 |
| Model | 41.8750434 | 5 | 8.37500867 | Prob > F = | 0.0000 |
| Residual | 80.7410668 | 673 | .119971867 | R-squared = | 0.3415 |
| | | | | Adj R-squared = | 0.3366 |
| Total | 122.61611 | 678 | .18084972 | Root MSE = | .34637 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .1077261 | .0066792 | 16.13 | 0.000 | .0946115 | .1208408 |
| expr | .0344524 | .0071189 | 4.84 | 0.000 | .0204745 | .0484304 |
| tenure | .0363033 | .0082594 | 4.40 | 0.000 | .0200859 | .0525206 |
| smsa | .1583146 | .0298248 | 5.31 | 0.000 | .0997537 | .2168754 |
| rns | -.074063 | .0308884 | -2.40 | 0.017 | -.1347123 | -.0134137 |
| _cons | 4.015335 | .098159 | 40.91 | 0.000 | 3.8226 | 4.20807 |

如果只对中学以上(S >=12)且在南方居住的子样本进行回归,可输入命令:

```
. reg lnw s expr tenure smsa if s >=12 & rns
```

| Source | SS | df | MS | Number of obs = 174 |
|---|---|---|---|---|
| | | | | F( 4, 169) = 32.17 |
| Model | 15.404067 | 4 | 3.85101675 | Prob > F = 0.0000 |
| Residual | 20.2300414 | 169 | .119704387 | R-squared = 0.4323 |
| | | | | Adj R-squared = 0.4188 |
| Total | 35.6341084 | 173 | .205977505 | Root MSE = .34598 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .1269124 | .0131847 | 9.63 | 0.000 | .1008845 | .1529404 |
| expr | .0226531 | .0156062 | 1.45 | 0.148 | -.0081551 | .0534613 |
| tenure | .0146869 | .0182079 | 0.81 | 0.421 | -.0212573 | .0506312 |
| smsa | .2136309 | .0548788 | 3.89 | 0.000 | .1052947 | .3219671 |
| _cons | 3.699026 | .1873691 | 19.74 | 0.000 | 3.329141 | 4.068912 |

回到最初估计的全样本：

```
. quietly reg lnw s expr tenure smsa rns
```

其中，前缀“quietly”表示不汇报回归结果。如果要计算被解释变量的拟合值($\widehat{\ln w}$)，并将其记为 lnw1，可使用命令：

```
. predict lnw1
```

如果要计算残差，并将其记为 e，可输入命令：

```
. predict e,residual
```

其中，选择项“residual”表示计算残差，默认计算拟合值。

对于回归方程 $\ln w = \beta_1 + \beta_2 s + \beta_3 expr + \beta_4 tenure + \beta_5 smsa + \beta_6 rns + \varepsilon$，考虑检验教育投资回报率是否为 10%，即检验原假设“$H_0: \beta_2 = 0.1$”，可使用命令：

```
. test s =0.1
```

此命令检验的原假设为，变量 $s$ 的系数等于 0.1。

```
( 1)  s = .1

       F(  1,   752) =    0.20
            Prob > F =    0.6515
```

由于 $t$ 分布的平方为 $F$ 分布，故 Stata 统一汇报 $F$ 统计量及其 $p$ 值。上表显示，$p$ 值 = 0.651 5，故无法拒绝原假设。事实上，对于单个系数的检验，手工计算 $t$ 统计量也十分方便。根据公式(5.58)可得

$$t \equiv \frac{\text{估计量} - \text{假想值}}{\text{估计量的标准误}} = \frac{0.102\,643 - 0.1}{0.005\,848\,8} = 0.451\,887\,57 \sim t(n-K)$$

$$= t(752) \tag{5.95}$$

由于默认为双边检验，故可计算此 $t$ 统计量对应的 $p$ 值如下：

```
. dis ttail(752,0.45188757)*2
.65148029
```

其中，“ttail(752,0.45188757)”表示自由度为 752 的 $t$ 分布比 0.451 887 57 更大的右侧尾部概率，正好是反向的累积分布函数。由此可知，手工计算的 $t$ 统计量的 $p$ 值，与 Stata 汇报的 $F$

统计量的 $p$ 值完全相同。

如果要进行单边检验，比如原假设仍为 $H_0:\beta_2=0.1$，而替代假设为 $H_1:\beta_2>0.1$，则拒绝域在 $t$ 分布的右侧尾部。相应的 $t$ 统计量仍为 0.451 887 57，但在计算 $p$ 值时，只需计算大于此 $t$ 统计量的右侧尾部概率即可：

```
. dis ttail(752,0.45188757)
.32574014
```

由于 $p$ 值仍高达 0.325 7，故依然可以接受原假设。总之，如果已知双边检验的 $p$ 值，在做单边检验时（假设 $t$ 统计量的符号与替代假设的方向相同），一般只需将双边检验的 $p$ 值除以 2，即可得到单边检验的 $p$ 值，然后得到单边检验的结果。

作为示例，下面考虑检验 *expr* 与 *tenure* 的系数是否相等，即检验 $H_0:\beta_3=\beta_4$，可输入命令：

```
. test expr = tenure
```

```
( 1)  expr - tenure = 0

      F(  1,   752) =    0.05
           Prob > F =    0.8208
```

由于 $p$ 值 =0.820 8，故可以轻松地接受原假设。为演示目的，考虑检验工龄回报率与现单位年限回报率之和是否等于教育回报率，即 $H_0:\beta_3+\beta_4=\beta_2$，可使用命令：

```
. test expr + tenure = s
```

```
( 1)  - s + expr + tenure = 0

      F(  1,   752) =    8.82
           Prob > F =    0.0031
```

由于 $p$ 值 =0.003 1，故可在 1% 的显著性水平上拒绝原假设，即认为 $\beta_3+\beta_4\neq\beta_2$。

**习题**

5.1 从残差 $e_i\equiv y_i-(\hat{\beta}_1+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_Kx_{iK})$ 出发，证明残差向量 $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$。

5.2 考虑一元回归模型 $y_i=\beta_1+\beta_2x_i+\varepsilon_i$。证明：$\mathrm{E}(\varepsilon_i\mid x_i)=0$ 意味着 $\mathrm{E}(y_i\mid x_i)=\beta_1+\beta_2x_i$。

5.3 考虑只对常数项进行回归，即 $y_i=\beta_1+\varepsilon_i$。写出其数据矩阵 $\boldsymbol{X}$，并根据公式 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 推导 $\beta_1$ 的 OLS 估计量。

5.4 假设数据矩阵为 $\boldsymbol{X}=\begin{pmatrix}1&2\\1&2\\1&2\\1&2\\1&2\end{pmatrix}$。

（1）此数据矩阵是否满列秩？

（2）写出数据矩阵 $\boldsymbol{X}$ 的转置。

（3）计算矩阵 $\boldsymbol{X}'\boldsymbol{X}$，其逆矩阵 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 是否存在？

5.5[①] 数据集 airq.dta 包含 1972 年美国加州 30 个大城市的如下变量：*airq*（空气质量指数，越低越好），*vala*（公司的增加值，千美元），*rain*（降雨量，英寸），*coast*（是否为海岸城市），*density*（人口密度，每平方英里），*income*（人均收入，美元）。

（1）把 *airq* 对其他变量进行 OLS 回归。

（2）检验原假设“平均收入对空气质量没有影响”。

（3）检验经济变量 *density* 与 *income* 的联合显著性。

（4）检验环境变量 *rain* 与 *coast* 的联合显著性。

（5）检验所有解释变量的联合显著性。

5.6 穷国能否赶上富国？由于穷国的资本较少，故资本的边际产出较高。因此，一种理论认为，穷国的经济增长速度应比富国快，并收敛于富国，称为“绝对收敛”（absolute convergence）。另一种观点则认为，只有在控制其他因素（比如人力资本）的情况下，穷国的增长速度才快于富国，称为“条件收敛”（conditional convergence）。使用 Gallup，Sachs and Mellinger（1999）的部分跨国数据集 geodata_short.dta，检验是否存在绝对收敛或条件收敛。该数据集的被解释变量为 *gdpg*6590（1965—1990 年人均 *GDP* 的增长率），而解释变量包括 *lgdp*65（1965 年人均 GDP 的对数），以及 *syr*1965（1965 年平均受中学教育年限的对数）。

（1）以 5% 的显著性水平检验是否存在绝对收敛。

（2）以 5% 的显著性水平检验是否存在条件收敛。

5.7[②] 使用回归模型进行餐馆选址。数据集 Woody3.dta 包含 33 家 Woody's 连锁餐馆的以下变量：*y*（毛销售收入），*competitors*（两英里内直接竞争者的数目），*pop*（三英里内的居民人数），*income*（三英里内的家庭平均收入）。

（1）把 *y* 对其他变量进行多元回归。

（2）评论拟合优度，以及各变量系数的符号与显著性。

（3）解释此回归结果如何有助于为一家新的 Woody's 餐馆选址。

---

① 此例来自 Verbeek（2012）。

② 此例来自 Studenmund（2010）。

*If you need to use asymptotic arguments, do not forget to let the number of observations tend to infinity.* ——Lucien Le Cam

*If you can't get it right as n goes to infinity, you shouldn't be in this business.* ——Clive Granger

# 6. 大样本 OLS

## 6.1 为何需要大样本理论

"大样本理论"(large sample theory),也称"渐近理论"(asymptotic theory),研究当样本容量 $n$ 趋向无穷大时统计量的性质。大样本理论已成为当代计量经济学的主流方法,因此学好本章内容的重要性不言而喻。大样本理论近年来之所以大行其道,其主要原因如下。

(1) 小样本理论的假设过强。首先,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(不相关)。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。这个假定有时太强。

**例** 考虑以下一阶自回归模型(first order autoregression, AR(1)):

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (t = 2, \cdots, T) \tag{6.1}$$

其中,解释变量 $y_{t-1}$ 为被解释变量 $y_t$ 的一阶滞后;而且 $\mathrm{Cov}(y_{t-1}, \varepsilon_t) = 0$。严格外生性要求,解释变量 $y_{t-1}$ 与所有 $\{\varepsilon_2, \cdots, \varepsilon_T\}$ 均不相关。这意味着,$y_t$ 也不与 $\varepsilon_t$ 相关。然而,根据自回归方程(6.1),$\varepsilon_t$ 是 $y_t$ 的一部分,故二者一定相关,因为

$$\begin{aligned} \mathrm{Cov}(y_t, \varepsilon_t) &= \mathrm{Cov}[(\rho y_{t-1} + \varepsilon_t), \varepsilon_t] = \rho \underbrace{\mathrm{Cov}(y_{t-1}, \varepsilon_t)}_{=0} + \mathrm{Var}(\varepsilon_t) \\ &= \mathrm{Var}(\varepsilon_t) > 0 \end{aligned} \tag{6.2}$$

因此,以被解释变量滞后值为解释变量的自回归模型,必然违背严格外生性的假定。另外,大样本理论则只要求解释变量与同期(同方程)的扰动项不相关。

其次,小样本理论假定扰动项为正态分布,而大样本理论无此限制。在很多情况下,我们并无把握经济变量是否服从正态分布。比如,正态分布为对称分布,但许多经济变量的分布并不对称,例如工资收入。即使考虑比较对称的工资对数,由于正态变量的取值范围为 $(-\infty, +\infty)$,而工资对数一般为正数(假设工资大于 1),故也不相符。作为示例,下面将数据集 grilic.dta 的工资与工资对数的核密度图画在一起,结果参见图 6.1。

```
. use grilic.dta,clear
. gen wage = exp(lnw)
```

```
. twoway kdensity wage,xaxis(1) yaxis(1) xvarlab(wage) || kdensity
lnw,xaxis(2) yaxis(2) xvarlab(ln(wage)) lpattern(dash)
```

其中,选择项"xaxis(1) yaxis(1)"与"xaxis(2) yaxis(2)"指定对于变量 *wage* 与 ln *w*分别使用不同的 *x* 轴与 *y* 轴,因为这两个变量的取值范围与概率密度均很不相同;选择项"xvarlab(wage)"与"xvarlab(ln(wage))"将变量 *wage* 与 ln *w* 核密度图的横轴标签分别指定为"*wage*"与"ln(*wage*)"。

从图 6.1 可知,工资的分布与正态分布相去甚远;而即使工资对数,在取值范围为(-∞,+∞)这一点上,严格来说也与正态分布不符。

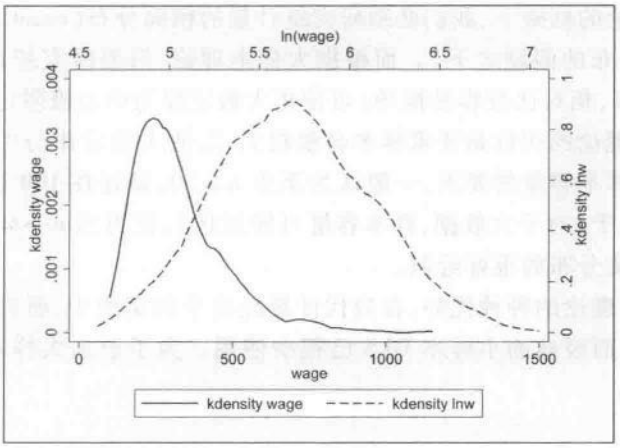



图 6.1 工资与工资对数的分布

事实上,被解释变量的分布可能为各种形状;有时即使取对数也不能使其接近正态分布。继续以数据集 grilic. dta 为例,将教育年限(*s*)与其对数(ln *s*)的核密度图画在一起,结果参见图 6.2。

```
. gen lns = log(s)
. twoway kdensity s,xaxis(1) yaxis(1) xvarlab(s) || kdensity lns,xaxis
(2) yaxis(2) xvarlab(lns) lpattern(dash)
```

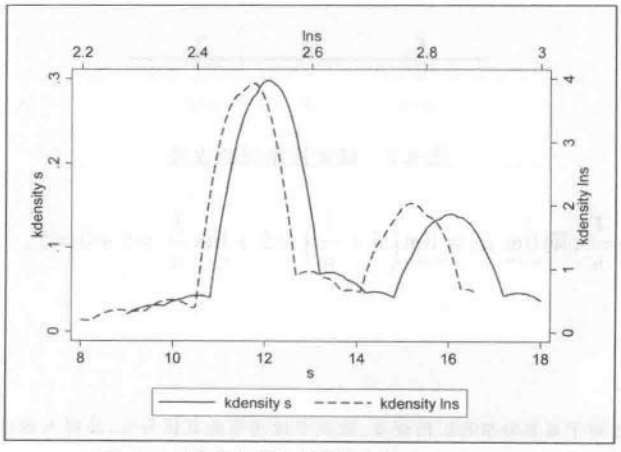



图 6.2 教育年限 *s* 与其对数 ln *s* 的分布

从图 6.2 可知，教育年限的分布呈现“双峰”形状，即多数人为中学或大学毕业。这种双峰形状，即使取对数后，也难以改变。因此，无论教育年限还是其对数的分布，都与“单峰”的正态分布相去甚远。这也说明，通过取对数使得变量的分布接近于正态并非万能（对于像工资那样的单峰右偏分布比较有效）。

对于小样本理论来说，为了进行统计推断（比如，推导 $t$ 统计量与 $F$ 统计量的有限样本分布），必须假设扰动项服从正态分布（故被解释变量也服从正态分布）。由于现实中的被解释变量可能服从各种分布（比如，变量婚否 $mrt$ 为离散的两点分布），故基于正态假设的小样本理论的适用范围受到很大限制。

（2）在小样本理论的框架下，我们必须研究统计量的精确分布（exact distribution），但常常难以推导（即使在正态分布的假设之下）。而根据大样本理论，只要研究统计量的大样本分布，即当 $n\to\infty$ 时的渐近分布，相对比较容易推导（可使用大数定律与中心极限定理）。

（3）使用大样本理论的代价是要求样本容量较大，以便大数定律与中心极限定理可以起作用。大样本理论对于样本容量的要求，一般认为至少 $n\geqslant 30$，最好在 100 以上。[①] 现代的数据集越来越大，经常成百上千（对于大数据，样本容量可能过亿），使得当 $n\to\infty$ 时才严格成立的渐近理论成为对统计量真实分布的很好近似。

总之，由于大样本理论的种种优势，在当代计量经济学的实践中，研究人员所使用的计量方法一般为大样本理论；而经典的小样本 OLS 已很少使用。为了引入大样本理论，下面首先介绍各种随机收敛的概念。

## 6.2 随机收敛

### 1. 确定性序列的收敛

**定义** 确定性序列 $\{a_n\}_{n=1}^{\infty}=\{a_1,a_2,a_3,\cdots\}$ **收敛**（converge）于常数 $a$，记为 $\lim\limits_{n\to\infty}a_n=a$ 或 $a_n\to a$，如果对于任意小的正数 $\varepsilon>0$，都存在 $N>0$，只要 $n>N$，就有 $|a_n-a|<\varepsilon$，即在 $a_N$ 以后的序列 $\{a_{N+1},a_{N+2},\cdots\}$ 均落入区间 $(a-\varepsilon,a+\varepsilon)$ 内，参见图 6.3。

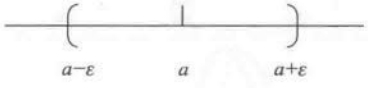



图 6.3 确定性序列的收敛

**例** 假设 $a_n=5+\dfrac{1}{n}$，则 $\lim\limits_{n\to\infty}a_n=\lim\limits_{n\to\infty}\left(5+\dfrac{1}{n}\right)=5+\lim\limits_{n\to\infty}\dfrac{1}{n}=5+0=5$。

① 统计量的大样本分布对于真实分布的近似程度，取决于扰动项的具体分布，故样本容量 30 在某些情况下可能偏小。另外，准确地说，此近似程度依赖于模型的自由度，即样本容量减去待估参数个数（$n-K$）。

**2. 随机序列的收敛**

考虑随机序列$\{x_n\}_{n=1}^{\infty}=\{x_1,x_2,x_3,\cdots\}$，即由随机变量构成的序列，其中每个元素$x_n$都是随机变量，而下标$n$通常表示样本容量。

**定义** 随机序列$\{x_n\}_{n=1}^{\infty}$**依概率收敛**(converge in probability)于常数$a$，记为$\operatorname{plim}_{n\to\infty} x_n=a$，或$x_n\xrightarrow{p}a$，如果对于任意$\varepsilon>0$，当$n\to\infty$时，都有$\lim_{n\to\infty}P(|x_n-a|>\varepsilon)=0$。

这意味着，任意给定很小的正数$\varepsilon>0$，当$n$越来越大时，随机变量$x_n$落在区间$(a-\varepsilon,a+\varepsilon)$之外的概率收敛于0，参见图6.4。换言之，当$n$变大时，$x_n$远离常数$a$的可能性越来越小，变得几乎不可能。由于已将随机事件$(|x_n-a|>\varepsilon)$取概率，故$P(|x_n-a|>\varepsilon)$其实是确定性序列(为概率的具体取值，已无不确定性)，而$\lim_{n\to\infty}P(|x_n-a|>\varepsilon)$只是普通的微积分极限。

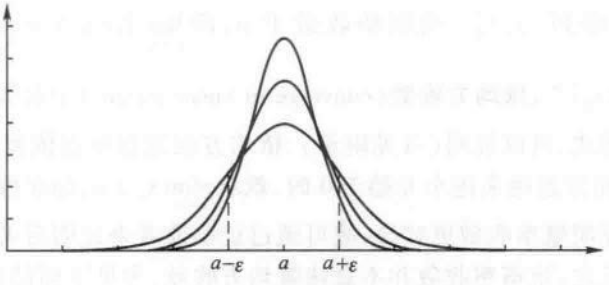



图6.4 随机序列的收敛

**例** 假设$x_n$服从如下两点分布：

$$x_n=\begin{cases}0 & \text{取值概率 } 1-1/n\\ n & \text{取值概率 } 1/n\end{cases}\tag{6.3}$$

显然，随着$n\to\infty$，$x_n$的分布越来越集中于0，而取值为$n$的可能性越来越小(尽管$n$离0越来越远)。因此，根据定义，$\operatorname{plim}_{n\to\infty}x_n=0$。

利用随机变量依概率收敛于常数的概念，可定义随机变量之间的随机收敛，只要随机变量之差依概率收敛于0。

**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。

概率收敛($\operatorname{plim}_{n\to\infty}$)的运算规则类似于微积分中极限($\lim_{n\to\infty}$)的运算。比如，假设$g(\cdot)$为连续函数，则

$$\operatorname{plim}_{n\to\infty}g(x_n)=g\left(\operatorname{plim}_{n\to\infty}x_n\right)\tag{6.4}$$

上式意味着，概率极限$\operatorname{plim}_{n\to\infty}$与连续函数$g(\cdot)$可交换运算次序，即无论先用函数$g(\cdot)$去作用$x_n$，再取概率极限；还是先对$x_n$取概率极限，再用函数$g(\cdot)$去作用，二者的效果是一样的。直观来看，当$x_n$的分布越来越集中于$x^*\equiv\operatorname{plim}_{n\to\infty}x_n$附近时，$g(x_n)$的分布自然也就越来越集中于$g(x^*)$附近。然而，期望算子$\mathrm{E}(\cdot)$却无此性质，因为一般来说，$\mathrm{E}(x^2)\neq[\mathrm{E}(x)]^2$。这正是大样

本理论的方便之处。

**例** 如果$\operatorname{plim}_{n\to\infty} s^2=\sigma^2$(样本方差依概率收敛于总体方差),则样本标准差 $s$ 也依概率收敛于总体标准差 $\sigma$,因为

$$\operatorname{plim}_{n\to\infty} s = \operatorname{plim}_{n\to\infty}\sqrt{s^2} = \sqrt{\operatorname{plim}_{n\to\infty} s^2} = \sqrt{\sigma^2} = \sigma \tag{6.5}$$

其中,"开根号"($\sqrt{\cdot}$)是连续函数,故可与求概率极限的运算交换次序。

对于随机向量序列(即序列中每个元素都是随机向量),也可类似地定义依概率收敛,只要定义其每个分量都依概率收敛即可。比如,随机向量序列$\{\boldsymbol{x}_n\}_{n=1}^{\infty}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots\}$依概率收敛于随机向量 $\boldsymbol{x}$,意味着 $\boldsymbol{x}_n$ 的每个分量都依概率收敛至 $\boldsymbol{x}$ 的相应分量,记为$\operatorname{plim}_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}$。

### 3. 依均方收敛

**定义** 如果随机序列$\{x_n\}_{n=1}^{\infty}$的期望收敛于 $a$,即$\lim_{n\to\infty}\mathrm{E}(x_n)=a$;而方差收敛于 0,即$\lim_{n\to\infty}\mathrm{Var}(x_n)=0$,则称$\{x_n\}_{n=1}^{\infty}$**依均方收敛**(converge in mean square)于常数 $a$,记为 $x_n\xrightarrow{ms}a$。

通过切比雪夫不等式,可以证明(参见附录),依均方收敛意味着依概率收敛。直观上,当 $x_n$ 的均值越来越趋于 $a$,而方差越来越小并趋于 0 时,就有$\operatorname{plim}_{n\to\infty}x_n=a$,即在极限处 $x_n$ 退化为常数 $a$。证明均方收敛通常比证明概率收敛更容易,故可通过证明前者来证明后者,这也是依均方收敛概念的主要用途之一。反之,依概率收敛并不意味着均方收敛,参见下面的反例。

**例** 回到$\{x_n\}$服从两点分布的例子,即 $x_n$ 取值为 0 的概率为 $1-1/n$,而取值为 $n$ 的概率为 $1/n$。虽然 $x_n$ 依概率收敛到 0,但 $x_n$ 并不依均方收敛到 0,因为此序列的期望恒等于 1:

$$\lim_{n\to\infty}\mathrm{E}(x_n) = \lim_{n\to\infty}\left[0\cdot\left(1-\frac{1}{n}\right)+n\cdot\frac{1}{n}\right] = 1 \neq 0 \tag{6.6}$$

直观来看,随着 $n\to\infty$,随机序列 $x_n$ 取值大于 0 的概率越来越小(为 $1/n$),但一旦取值为正数,则很大(等于 $n$),故此序列的期望始终为 1。不难证明,此序列的方差发散,即$\lim_{n\to\infty}\mathrm{Var}(x_n)=\infty$(参见习题)。

### 4. 依分布收敛

**定义** 记随机序列$\{x_n\}_{n=1}^{\infty}$与随机变量 $x$ 的累积分布函数分别为 $F_n(x)$ 与 $F(x)$。如果对于任意给定 $x$,都有$\lim_{n\to\infty}F_n(x)=F(x)$,则称随机序列$\{x_n\}_{n=1}^{\infty}$**依分布收敛**(converge in distribution)于随机变量 $x$,记为 $x_n\xrightarrow{d}x$,并称 $x$ 的分布为 $x_n$ 的**渐近分布**(asymptotic distribution)或**极限分布**(limiting distribution)。

直观上,这意味着,当 $n\to\infty$ 时,$x_n$ 的分布函数(概率密度函数)越来越像 $x$ 的分布函数(概率密度函数)。

**例** 当 $t$ 分布的自由度越来越大时,$t$ 分布依分布收敛于标准正态分布,即当 $k\to\infty$ 时,$t(k)\xrightarrow{d}N(0,1)$。为了直观地显示依分布收敛的过程,下面在 Stata 中画 $N(0,1)$,$t(1)$ 与 $t(5)$ 的累积分布函数,结果参见图 6.5。

```
. twoway function N = normal(x),range( -5 5) || function t1 = t(1,x),
range( -5 5) lpattern(dash) || function t5 = t(5,x),range( -5 5) lpat-
tern(shortdash) ytitle(累积分布函数)
```

其中,选择项"lpattern(shortdash)"表示以短横来画线。

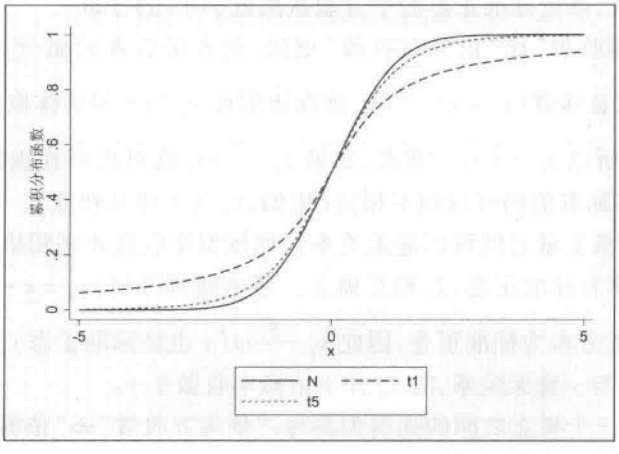



图 6.5 依分布收敛(累积分布函数)

更直观地,可以通过概率密度函数,来考察 $t$ 分布依分布收敛于标准正态的过程,结果参见图 6.6。

```
. twoway function N = normalden(x),range( -5 5) || function t1 = tden(1,
x),range( -5 5) lpattern(dash) || function t5 = tden(5,x),range( -5 5)
lpattern(shortdash) ytitle(概率密度)
```

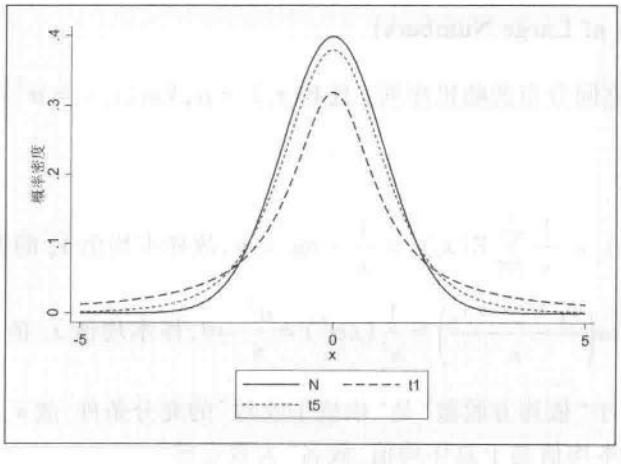



图 6.6 依分布收敛(概率密度函数)

在计量经济学中,许多统计量的大样本分布均为正态分布,故引入如下概念。

**定义** 如果 $x_n \xrightarrow{d} x$,且 $x$ 服从正态分布,则称 $x_n$ 为**渐近正态**(asymptotically normal),即当 $n \to \infty$ 时,$x_n$ 的分布越来越像正态分布。

依分布收敛的运算也很方便。比如,假设 $x_n \xrightarrow{d} x$,而 $g(\cdot)$ 为连续函数,则 $g(x_n)$ 的渐近分

布就是 $g(x)$，即 $g(x_n)\xrightarrow{d}g(x)$。直观上，当 $x_n$ 的分布越来越像 $x$ 的分布时，$g(x_n)$ 的分布自然也越来越像 $g(x)$ 的分布。这为大样本理论的推导提供了方便。

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

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；但反之，则不然。首先，如果 $x_n\xrightarrow{p}x$，则意味着 $(x_n-x)\xrightarrow{p}0$，即在极限处 $x_n$ 与 $x$ 的具体取值并无区别，故二者的概率分布也必然相同，所以 $x_n\xrightarrow{d}x$。其次，如果 $x_n\xrightarrow{d}x$，这只说明在极限处 $x_n$ 与 $x$ 的分布函数相同，但 $x_n$ 与 $x$ 的实际取值仍可以很不相同（比如，$x_n$ 与 $x$ 相互独立）。事实上，依分布收敛只是分布函数的收敛（随机变量之间可以毫无关系），而依概率收敛才是随机变量本身的收敛。

**例** 假设 $x$ 与 $y$ 都为标准正态，且相互独立。考虑随机序列 $\{x_n=x+(1/n)\}_{n=1}^{\infty}$。显然，由于 $1/n\to0$，故 $x_n$ 的渐近分布为标准正态，因此 $x_n\xrightarrow{d}y$（$y$ 也是标准正态）。然而，$x_n$ 却与 $y$ 相互独立，$x_n$ 的具体取值也与 $y$ 毫无关系，故 $x_n$ 并不依概率收敛于 $y$。

总之，随机收敛的三个概念之间的强弱关系为，“依均方收敛”⇒“依概率收敛”⇒“依分布收敛”；反之，此箭头的相反方向则不成立。

## 6.3 大数定律与中心极限定理

大样本理论所依赖的两大工具是概率统计中的大数定律与中心极限定理，但需作一些推广。下面复习概率统计中的这两个核心结论。

### 1. 大数定律（Law of Large Numbers）

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

**证明：** 首先，$\mathrm{E}(\bar{x}_n)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i)=\frac{1}{n}\cdot n\mu=\mu$，故样本均值 $\bar{x}_n$ 的期望仍为 $\mu$。

其次，$\mathrm{Var}(\bar{x}_n)=\mathrm{Var}\left(\frac{x_1+\cdots+x_n}{n}\right)=\frac{1}{n^2}(n\sigma^2)=\frac{\sigma^2}{n}\to0$，样本均值 $\bar{x}_n$ 的方差收敛到 0。因此，$\bar{x}_n$ 依均方收敛于 $\mu$。由于“依均方收敛”是“依概率收敛”的充分条件，故 $\bar{x}_n\xrightarrow{p}\mu$。这意味着，当样本容量 $n$ 很大时，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）

根据大数定律，当 $n\to\infty$ 时，样本均值 $\bar{x}_n$ 依概率收敛到总体均值 $\mu$。但在一般情况下，$\bar{x}_n$ 的具体分布则很难推导。中心极限定理告诉我们，无论原序列 $\{x_n\}_{n=1}^{\infty}$ 服从什么分布，当 $n\to\infty$ 时，样本均值 $\bar{x}_n$ 的渐近分布都为正态分布。换言之，只要样本容量 $n$ 足够大，则 $\bar{x}_n$ 的真实分布将很

接近于正态分布，故可以用正态分布来很好地近似 $\bar{x}_n$ 的真实分布（此真实分布通常无法求解），并以此渐近分布作为统计推断的基础。中心极限定理为大样本理论提供了极大的方便。

**中心极限定理** 假定 $\{x_n\}_{n=1}^{\infty}$ 为独立同分布的随机序列，且 $E(x_1)=\mu$，$Var(x_1)=\sigma^2$ 存在，则

$$\frac{\bar{x}_n-\mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1) \tag{6.7}$$

此定理告诉我们，标准化之后的样本均值（即减去期望，除以标准差）的渐近分布为标准正态。直观上，可视为 $\bar{x}_n \xrightarrow{d} N(\mu,\sigma^2/n)$；但这是不严格的写法，因为 $\bar{x}_n$ 的方差 $\sigma^2/n \to 0$（在极限处，$\bar{x}_n$ 的方差为 0，故退化为常数 $\mu$）。将表达式（6.7）两边同乘 $\sigma$，并将分母的 $\sqrt{1/n}$ 放到分子上，可得中心极限定理的等价表达式：

$$\sqrt{n}(\bar{x}_n-\mu) \xrightarrow{d} N(0,\sigma^2) \tag{6.8}$$

显然，$\sqrt{n}\to\infty$；而根据大数定律，$(\bar{x}_n-\mu)\xrightarrow{p}0$，故上式用 $\sqrt{n}\cdot(\bar{x}_n-\mu)$（即“$\infty\cdot 0$”型）得到非退化的渐近正态分布。表达式（6.8）的好处是，它更容易推广到多维的情形。

**多维的中心极限定理**：假定 $\{\boldsymbol{x}_n\}_{n=1}^{\infty}$ 为独立同分布的随机向量序列，且 $E(\boldsymbol{x}_1)=\boldsymbol{\mu}$，$Var(\boldsymbol{x}_1)=\boldsymbol{\Sigma}$ 存在，则 $\sqrt{n}(\bar{\boldsymbol{x}}_n-\boldsymbol{\mu})\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{\Sigma})$。

## 6.4 使用蒙特卡罗法模拟中心极限定理

中心极限定理是一个非常神奇的结果，但不容易用初等方法证明。下面，我们使用蒙特卡罗法来模拟中心极限定理。作为示例，假设 $x$ 服从在（0，1）上的均匀分布，从此分布随机抽取观测值，样本容量为 30，我们希望用蒙特卡罗法直观地“看到”样本均值 $\bar{x}_{30}$ 的分布，并与正态分布相比较。为此，从（0，1）上的均匀分布抽取 10 000 个样本容量为 30 的随机样本，得到 10 000 个 $\bar{x}_{30}$ 的观测值，然后画其直方图。

为达到此目的，可使用如下 Stata 程序[①]：首先，用命令 `program` 定义一个叫“`onesample`”（可自行命名）的程序，从均匀分布抽取一个样本容量为 30 的随机样本，并计算 $\bar{x}_{30}$；其次，用命令 `simulate` 重复此程序 10 000 次，得到 10 000 个 $\bar{x}_{30}$ 的观测值；最后，用命令 `histogram` 画 $\bar{x}_{30}$ 的直方图。具体来说，可在 Stata 命令窗口依次输入如下命令：

```
. program onesample,rclass        (定义程序 onesample，并以 r()形式储存结果)
    drop _all                     (删去内存中已有数据)
    set obs 30                    (确定随机抽样的样本容量为 30)
    gen x = runiform()            (得到在(0,1)上均匀分布的随机样本)
    sum x                         (使用命令 sum 计算样本均值)
    return scalar mean_sample = r(mean)   (将样本均值记为 mean_sample)
    end                           (程序 onesample 结束)
```

① 此程序得益于 Cameron and Trivedi(2010)。

```
. set more off                                    (指定 Stata 输出结果连续翻页)
. simulate xbar = r(mean_sample), seed(101) reps(10000) nodots: onesample
```

其中,选择项"reps(10000)"表示,命令 simulate 将运行"onesample"程序 10 000 遍,并生成变量 xbar 来记录这 10 000 个样本均值。选择项"seed(101)"用来确定随机数的初始值,以便再次模拟或别人运行此程序时,也能得到完全一样的结果(有关随机数的产生,参见第 4 章附录)。选择项"nodots"表示不显示表示模拟过程的点(默认以一个点表示抽取一个样本)。

```
. hist xbar, normal
```

其中,选择项"normal"表示画出相应的正态分布,结果参见图 6.7。

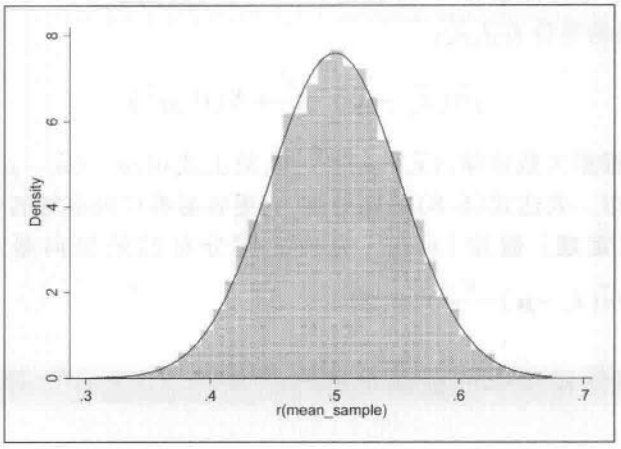



图 6.7 模拟中心极限定理

从图 6.7 可知,虽然样本容量仅为 30,但 $\bar{x}_{30}$ 的分布已经很接近于正态分布。作为练习,可从 $\chi^2(10)$ 中抽取随机样本,重复上面的蒙特卡罗模拟。只要将上面程序中的语句"gen x = runiform()"改为"gen x = rchi2(10)"即可。有关随机数的产生以及如何从常见分布中随机抽样,参见第 4 章附录。

## 6.5 统计量的大样本性质

在大样本理论下,我们关心当样本容量 $n\to\infty$ 时,统计量是否具有良好的大样本性质。

### 1. 一致估计量

**定义** 考虑参数 $\beta$ 的估计量 $\hat{\beta}_n$,其中下标 $n$ 为样本容量(强调 $\hat{\beta}_n$ 对样本容量 $n$ 的依赖)。如果 $\operatorname{plim}_{n\to\infty}\hat{\beta}_n=\beta$,则称 $\hat{\beta}_n$ 是参数 $\beta$ 的**一致估计量**(consistent estimator)。

一致性(consistency)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 $\beta$,参见图 6.8。这是对估计量最基本,也是最重要的要求。如果估计方法不一致,则意味着研究没有太大意义;因为无论样本容量多大,估计量也不会收敛到真实值。

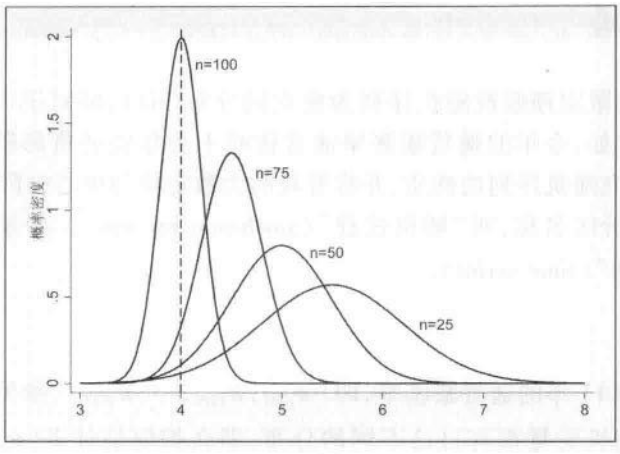



图 6.8 一致估计量的示意图①

在多维情况下，称估计量 $\hat{\boldsymbol{\beta}}_n$ 是参数 $\boldsymbol{\beta}$ 的一致估计量，如果 $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n=\boldsymbol{\beta}$，即 $\hat{\boldsymbol{\beta}}_n$ 的各分量都是 $\boldsymbol{\beta}$ 相应分量的一致估计。

**2. 渐近正态分布与渐近方差**

**定义** 如果 $\sqrt{n}(\hat{\beta}_n-\beta)\xrightarrow{d}N(0,\sigma^2)$，则称 $\hat{\beta}_n$ 为**渐近正态**(asymptotically normal)，称 $\sigma^2$ 为其**渐近方差**(asymptotic variance)，记为 $\operatorname{Avar}(\hat{\beta}_n)$。直观上，可近似认为 $\hat{\beta}_n\xrightarrow{d}N(\beta,\sigma^2/n)$，尽管这是不严格的写法(方差 $\sigma^2/n$ 趋于 0，故为退化的分布)。

在多维情况下，如果 $\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma}$ 为半正定矩阵，则称 $\hat{\boldsymbol{\beta}}_n$ 为渐近正态分布，而称 $\boldsymbol{\Sigma}$ 为 $\hat{\boldsymbol{\beta}}_n$ 的渐近协方差矩阵，记为 $\operatorname{Avar}(\hat{\boldsymbol{\beta}}_n)$。

**3. 渐近有效**

假设 $\hat{\beta}_n$ 与 $\tilde{\beta}_n$ 都是 $\beta$ 的渐近正态估计量。如果 $\operatorname{Avar}(\hat{\beta}_n)\leqslant\operatorname{Avar}(\tilde{\beta}_n)$，则称 $\hat{\beta}_n$ 比 $\tilde{\beta}_n$ **更为渐近有效**(asymptotically more efficient)。这意味着，在大样本下，$\hat{\beta}_n$ 的方差小于 $\tilde{\beta}_n$ 的方差(尽管在小样本下未必如此)。

在多维情况下，假设 $\hat{\boldsymbol{\beta}}_n$ 与 $\tilde{\boldsymbol{\beta}}_n$ 都是 $\boldsymbol{\beta}$ 的渐近正态估计量。如果 $[\operatorname{Avar}(\tilde{\boldsymbol{\beta}}_n)-\operatorname{Avar}(\hat{\boldsymbol{\beta}}_n)]$ 为半正定矩阵，则称 $\hat{\boldsymbol{\beta}}_n$ 比 $\tilde{\boldsymbol{\beta}}_n$ **更为渐近有效**。

---

① 生成此图的 Stata 命令为“twoway function y1 = normalden(x,4,.2),range(3 8) || function y2 = normalden(x,4.5,.3),range(3 8) || function y3 = normalden(x,5,.5),range(3 8) || function y4 = normalden(x,5.5,.7),range(3 8) xline(4)”，然后在图像编辑器(Graph Editor)中进行少量编辑。

## 6.6 随机过程的性质

大数定律与中心极限定理假设随机序列为独立同分布(iid),但对于大多数经济变量而言,此假定可能太强了。比如,今年的通货膨胀率通常依赖于去年的通货膨胀率,二者并非相互独立。为此,我们需要研究随机序列的性质,并将常规的大数定律与中心极限定理进行推广。随机序列$\{x_n\}_{n=1}^{\infty}$有个更好听的名称,叫"随机过程"(stochastic process)。如果下标为时间,则记为$\{x_t\}_{t=1}^{\infty}$,也称"时间序列"(time series)。

### 1. 严格平稳过程

考察中国1978—2013年的通货膨胀率,即$\{\pi_{1978},\pi_{1979},\cdots,\pi_{2013}\}$,参见图6.9。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,那么如何估计$\mathrm{E}(\pi_{1978})$与$\mathrm{Var}(\pi_{1978})$呢?每年通货膨胀率的样本容量仅为1,且历史不能重演(也无法穿越)。如果这36年的通货膨胀率分布都不变,则可将$\overline{\pi}\equiv\frac{1}{36}\sum_{t=1978}^{2013}\pi_t$作为$\mathrm{E}(\pi_t)$的估计量。

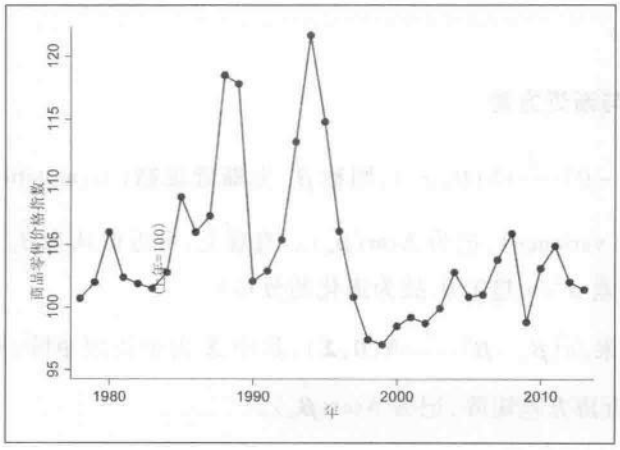



图6.9 中国零售商品价格指数(上年=100)

资料来源:国家统计局网站(http://data.stats.gov.cn/workspace/index?m=hgnd)

基于这样的考虑,我们通常要求随机过程$\{x_t\}_{t=1}^{\infty}$的概率分布不随时间推移而改变。换言之,无论过去、现在还是未来去看此随机过程,它的概率分布性质都一样。这种随机过程称为"严格平稳过程",它要求随机过程的有限维分布不随时间推移而改变。比如,$x_t$的分布与$x_s$的分布相同($\forall t,s$);$(x_1,x_4)$的分布与$(x_2,x_5)$相同(二者均相隔3期);$(x_1,x_2,x_3)$的分布与$(x_5,x_6,x_7)$相同(二者均为连续3期)。

**定义** 随机过程$\{x_t\}_{t=1}^{\infty}$是**严格平稳过程**(strictly stationary process),简称平稳过程,如果对任意$m$个时期的时间集合$\{t_1,t_2,\cdots,t_m\}$,随机向量$\{x_{t_1},x_{t_2},\cdots,x_{t_m}\}$的联合分布等于随机向量$\{x_{t_1+k},x_{t_2+k},\cdots,x_{t_m+k}\}$的联合分布,其中$k$为任意整数。

这意味着,将$\{x_{t_1},x_{t_2},\cdots,x_{t_m}\}$中每个变量的时间下标全部前移或后移$k$期,不会改变其分

布。$\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\}$ 的联合分布仅取决于 $\{t_1, t_2, \cdots, t_m\}$ 各个时期之间的相对距离，而不依赖于其绝对位置。

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

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

**例** 考虑以下一阶自回归过程（AR(1)）：

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (t = 1, \cdots, T) \tag{6.9}$$

其中，$\{\varepsilon_t\}$ 为独立同分布，且 $\mathrm{Cov}(y_{t-1}, \varepsilon_t) = 0$。

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

**证明：** 如果 $\rho = 1$，则 $y_t = y_{t-1} + \varepsilon_t$。因此，$y_1 = y_0 + \varepsilon_1$，而 $y_2 = y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2$，以此类推可知

$$y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t \tag{6.10}$$

因此，给定初始值 $y_0$，当 $t \to \infty$ 时，$\mathrm{Var}(y_t) = t\sigma_\varepsilon^2 \to \infty$，其中 $\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_t)$，即方差越来越大，以至无穷。因此，$\{y_t\}$ 不是平稳过程（平稳过程要求同分布，故方差不变）。由于 $y_t$ 只是在 $y_{t-1}$ 的基础上，加上一个随机扰动项 $\varepsilon_t$，故当 $\rho = 1$ 时，称 $\{y_t\}$ 为“随机游走”（random walk）。

如果 $|\rho| < 1$，则 $\mathrm{Var}(y_t)$ 会收敛到常数。对方程(6.9)两边同时取方差，可得

$$\mathrm{Var}(y_t) = \rho^2 \mathrm{Var}(y_{t-1}) + \sigma_\varepsilon^2 \tag{6.11}$$

记 $z_t \equiv \mathrm{Var}(y_t)$，$z_{t-1} = \mathrm{Var}(y_{t-1})$，则上式可写为

$$z_t = \rho^2 z_{t-1} + \sigma_\varepsilon^2 \tag{6.12}$$

这是确定性的一阶线性差分方程，因为 $z_t \equiv \mathrm{Var}(y_t)$ 为非随机。由于 $\rho^2 < 1$，故 $\mathrm{Var}(y_t)$ 将收敛到一个稳定值，参见图 6.10。在方程(6.12)中，令 $z_t = z_{t-1}$，可求解此收敛的稳定值 $z^*$：

$$z^* = \rho^2 z^* + \sigma_\varepsilon^2 \tag{6.13}$$

将上式移项整理后可得，$z^* = \dfrac{\sigma_\varepsilon^2}{1 - \rho^2}$。这说明，如果忽略序列 $\{y_t\}$ 的前面几项，则可将 $\{y_t\}$ 的方差视为常数。进一步可证明，$\{y_t\}_{t=0}^{\infty}$ 是严格平稳过程[①]。

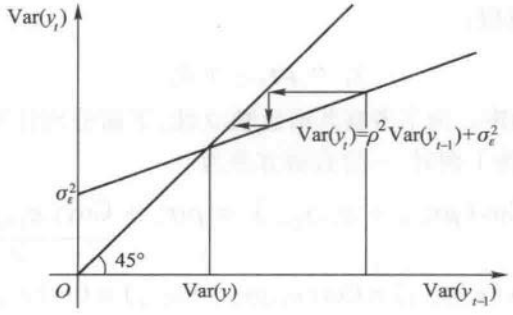



图 6.10 平稳一阶自回归过程的方差收敛

① 参见 Stock and Watson(2012, p. 620)。

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，而不要求整个分布都稳定，故引入以下“弱平稳过程”的概念。

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

对于弱平稳过程，由于$\mathrm{E}(x_t)$不依赖于$t$，故其期望为常数；进一步，由于$\mathrm{Cov}(x_t,x_{t+k})$仅依赖于$k$，如果令$k=0$，则$\mathrm{Cov}(x_t,x_t)=\mathrm{Var}(x_t)$也不依赖于$t$，故弱平稳过程的方差也是常数。

显然，严格平稳过程[①]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)，而概率分布还可能依赖于更高阶的矩。在实践中较常用的弱平稳过程是期望为0，且不存在序列相关的白噪声过程。

**定义** 对于弱平稳过程$\{x_t\}_{t=1}^{\infty}$，如果对于$\forall t$，都有$\mathrm{E}(x_t)=0$，而且$\mathrm{Cov}(x_t,x_{t+k})=0(\forall k\neq 0)$，则称为**白噪声过程**(white noise process)。

需要注意的是，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。

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

### 2. 渐近独立性

“严格平稳过程”(相当于“同分布”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(iid)。但“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。但今年的通胀率与100年前的通胀率或许可近似地视为相互独立，称为**渐近独立**(ergodic，也称**遍历性**[③])，或**弱相依**(weakly dependent)。渐近独立意味着，只要两个随机变量相距足够远，可近似认为它们相互独立。

**例** 相互独立的随机序列是渐近独立的。

**例** AR(1)是否渐近独立？

考虑以下一阶自回归模型：

$$y_t=\rho y_{t-1}+\varepsilon_t \tag{6.14}$$

其中，$|\rho|<1$，而$\varepsilon_t$为白噪声。为了考察其渐近独立性，下面分别计算其各阶“自协方差”(autocovariance)。当时间间隔为1期时，一阶自协方差为

$$\mathrm{Cov}(y_t,y_{t-1})=\mathrm{Cov}(\rho y_{t-1}+\varepsilon_t,y_{t-1})=\rho\sigma_y^2+\underbrace{\mathrm{Cov}(\varepsilon_t,y_{t-1})}_{=0}=\rho\sigma_y^2 \tag{6.15}$$

其中，$\sigma_y^2$为$y$的方差；而$\mathrm{Cov}(\varepsilon_t,y_{t-1})=\mathrm{Cov}(\varepsilon_t,\rho y_{t-2}+\varepsilon_{t-1})=\mathrm{Cov}(\varepsilon_t,\varepsilon_{t-1})=0$，因为$\varepsilon_t$为白噪

---

① 假设严格平稳过程的期望、方差与协方差都存在。

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 “遍历性”这一术语来自物理学，字面意思为“经历所有的物理状态”。但对于计量经济学而言，译为“渐近独立性”更为贴切。

声(无序列相关)。

当时间间隔为 2 期时,原方程(6.14)可写为

$$y_t = \rho y_{t-1} + \varepsilon_t = \rho(\rho y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \rho^2 y_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t \tag{6.16}$$

因此,二阶自协方差为

$$\mathrm{Cov}(y_t, y_{t-2}) = \mathrm{Cov}(\rho^2 y_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t, y_{t-2}) = \rho^2\sigma_y^2 \tag{6.17}$$

以此类推,当时间间隔为 $j$ 期时,

$$\mathrm{Cov}(y_t, y_{t-j}) = \rho^j\sigma_y^2 \tag{6.18}$$

由于 $|\rho| < 1$,故当上式 $j \to \infty$ 时,$\mathrm{Cov}(y_t, y_{t-j}) \to 0$。由此可知,相距越远,则序列 $\{y_t\}$ 的自协方差越小,且在极限处变为 0(不相关),故此 AR(1)模型为渐近独立的过程。

有了严格平稳过程与渐近独立的概念后,可以将大数定律作以下重要推广。

**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 $\mathrm{E}(x_i) = \mu$ 存在,则 $\bar{x}_n \equiv \dfrac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu$,即样本均值 $\bar{x}_n$ 是总体均值 $\mathrm{E}(x_i)$ 的一致估计。

渐近独立定理是对大数定律的重要推广,更适用于经济数据。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),只要此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。类似地,可将中心极限定理作相应的推广;即在一定条件下,中心极限定理也适用于渐近独立的平稳过程①。

**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 $f(\cdot)$,$\{y_i \equiv f(x_i)\}_{i=1}^{\infty}$ 也是渐近独立的严格平稳过程。

根据此命题,则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) $\mathrm{E}[f(x_i)]$,都可以由其对应的"样本矩"(sample moment) $\dfrac{1}{n}\sum_{i=1}^{n} f(x_i)$ 来一致地估计。

**例** 对于渐近独立的平稳过程 $\{x_i\}_{i=1}^{\infty}$,样本方差 $s^2 \equiv \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 是总体方差 $\mathrm{Var}(x) \equiv \mathrm{E}[x - \mathrm{E}(x)]^2$ 的一致估计;而样本协方差 $s_{xy} \equiv \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ 为总体协方差 $\mathrm{Cov}(x,y) \equiv \mathrm{E}[(x - \mathrm{E}(x))(y - \mathrm{E}(y))]$ 的一致估计。

## 6.7 大样本 OLS 的假定②

**假定 6.1** 线性假定

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{6.19}$$

---

① 参见陈强(2014,p.57)或 Hayashi(2000,p.106)。

② 参见 Wooldridge(2009,p.382)与 Hayashi(2000,p.109)。

此假定与小样本 OLS 完全相同。

**假定 6.2** $(K+1)$ 维随机过程 $\{y_i, x_{i1}, \cdots, x_{iK}\}$ 为渐近独立的平稳过程(ergodic stationarity),故适用大数定律与中心极限定理。

**例** 如果样本为随机样本,则 $\{y_i, x_{i1}, \cdots, x_{iK}\}$ 独立同分布,故是渐近独立的平稳过程。

**假定 6.3** 前定解释变量(predetermined regressors)

所有解释变量均为"前定"(predetermined),也称"同期外生"(contemporaneously exogenous),即它们与同期(同方程)的扰动项正交,即 $\mathrm{E}(x_{ik}\varepsilon_i)=0, \forall i,k$。由于 $\mathrm{E}(x_{ik}\varepsilon_i)=0$,故 $x_{ik}$ 与 $\varepsilon_i$ 不相关,仿佛在 $\varepsilon_i$ 产生之前,$x_{ik}$ 已经确定,故名"前定解释变量"。此假定比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),而前定变量仅要求与同期的扰动项不相关。

**假定 6.4** 秩条件(rank condition)

数据矩阵 $\boldsymbol{X}$ 满列秩,即 $\boldsymbol{X}$ 中没有多余(可由其他变量线性表出)的解释变量,故不存在严格多重共线性。

显然,大样本理论的假定 6.1 与 6.4 与小样本理论相同,而假定 6.2 与 6.3 则比小样本理论更为放松;特别地,大样本 OLS 无需假设"严格外生性"与"正态随机扰动项",故具有更大的适用性与稳健性。

## 6.8 OLS 的大样本性质

在假定 6.1—6.4 之下,可以证明 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 具有以下良好的大样本性质。

(1) $\hat{\boldsymbol{\beta}}$ 为一致估计量,即 $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}$。以一元回归为例进行说明。考虑以下模型:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{6.20}$$

其中,$\beta$ 的 OLS 估计量为(参见第 4 章)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.21}$$

不难证明,方程(6.20)的离差形式为(参见习题)

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \tag{6.22}$$

其中,$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,而 $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$。将方程(6.22)代入方程(6.21)可得

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

$$= \frac{\beta \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta + \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \xrightarrow{p} \beta + \frac{\mathrm{Cov}(x_i, \varepsilon_i)}{\mathrm{Var}(x_i)} = \beta \tag{6.23}$$

其中，根据假定 6.3，$\mathrm{Cov}(x_i, \varepsilon_i) = 0$。由此可知，前定解释变量，或扰动项与解释变量同期不相关，是保证 OLS 一致的最重要条件。反之，如果 $\mathrm{Cov}(x_i, \varepsilon_i) \neq 0$，则 $\operatorname{plim}_{n\to\infty} \hat{\beta} = \beta + \frac{\mathrm{Cov}(x_i, \varepsilon_i)}{\mathrm{Var}(x_i)} \neq \beta$。进一步，如果 $\mathrm{Cov}(x_i, \varepsilon_i) > 0$，则 $\operatorname{plim}_{n\to\infty} \hat{\beta} > \beta$。比如，考察教育投资的回报率，$x_i$ 为教育年限，而 $\varepsilon_i$ 为被遗漏的个人能力。显然，$x_i$ 与 $\varepsilon_i$ 正相关（能力高者通常上学更久），故 OLS 估计量将高估教育投资的回报率。

另外，如果 $\mathrm{Cov}(x_i, \varepsilon_i) < 0$，则 $\operatorname{plim}_{n\to\infty} \hat{\beta} < \beta$。比如，考察上医院对健康的作用，$x_i$ 为是否上医院，而 $\varepsilon_i$ 为个人原来的健康状况（被遗漏）。显然，$x_i$ 与 $\varepsilon_i$ 负相关（通常只有健康不佳者才上医院），故 OLS 估计量将低估上医院对健康的正面作用（去医院者的健康往往不如未去医院者）。

更直观地，可通过图示来考察 $\mathrm{Cov}(x_i, \varepsilon_i) \neq 0$ 的后果，参见图 6.11。在图 6.11 中，真实（总体）回归线为 $\alpha + \beta x_i$，而样本回归线为 $\hat{\alpha} + \hat{\beta} x_i$。不失一般性，假设 $\mathrm{Cov}(x_i, \varepsilon_i) > 0$。由于 $x_i$ 与 $\varepsilon_i$ 正相关，故当 $x_i$ 较小时，$\varepsilon_i$ 也倾向于较小；而当 $x_i$ 较大时，$\varepsilon_i$ 也倾向于较大。因此，样本回归线比真实回归线更为陡峭，$\hat{\beta}$ 将高估 $\beta$。反之，如果 $\mathrm{Cov}(x_i, \varepsilon_i) < 0$，则 $\hat{\beta}$ 将低估 $\beta$。增大样本容量

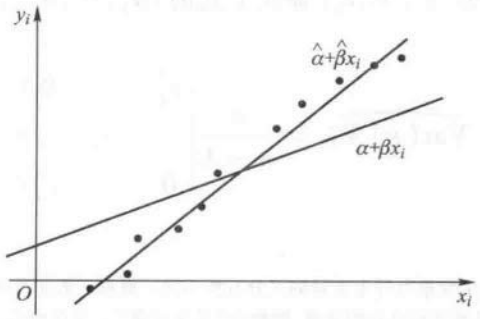



图 6.11 扰动项与解释变量相关导致不一致估计

($n\to\infty$)能使偏差(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。

在计量经济学中，如果解释变量与扰动项相关，即 $\mathrm{Cov}(x_i,\varepsilon_i)\neq 0$，则称此解释变量为“内生解释变量”(endogenous regressor)，简称“内生变量”；反之，则为“外生变量”(exogenous variable)[①]。由于内生变量的存在，致使 OLS 回归出现偏差，统称为“内生性偏差”(endogeneity bias)，或简称“内生性”。

在什么情况下可能出现内生性偏差？如果存在遗漏变量、双向因果关系或解释变量测量误差(measurement errors)，则常会出现解释变量与扰动项同期相关的情形，导致 OLS 不一致。我们将在第 9—12 章探讨相应的解决方法。

(2) $\hat{\boldsymbol{\beta}}$ 服从渐近正态分布，即 $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N\left(\boldsymbol{0},\mathrm{Avar}(\hat{\boldsymbol{\beta}})\right)$，其中 $\mathrm{Avar}(\hat{\boldsymbol{\beta}})$ 为 $\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵。$\hat{\boldsymbol{\beta}}$ 之所以服从渐近正态，是因为在一定条件下，中心极限定理适用于渐近独立的平稳过程。我们将在下文以蒙特卡罗法进行验证。

(3) 由于大样本理论一般不假设球形扰动项，故渐近协方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\beta}})$ 的表达式更为复杂。根据第 5 章公式(5.46)，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵可写为夹心估计量的形式：

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{6.24}$$

其中，$\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})$ 为扰动项的协方差矩阵。如果存在球形扰动项(同方差、无自相关)，则 $\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=\sigma^2\boldsymbol{I}_n$，上式可简化为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{6.25}$$

对于横截面数据，经常存在异方差，但无自相关(比如，各截面单位之间相互独立)。为此，考虑存在条件异方差，但无自相关的情形。此时，扰动项的协方差矩阵可写为

$$\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=\begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_n^2\end{pmatrix} \tag{6.26}$$

其中，$\sigma_1^2,\cdots,\sigma_n^2$ 不全相等。由于无自相关，各扰动项的协方差为 0，故上式非主对角线元素全部为 0。问题是，如何估计上式的 $\{\sigma_1^2,\cdots,\sigma_n^2\}$。由于 $\{\sigma_1^2,\cdots,\sigma_n^2\}$ 为各扰动项的方差，故一个自然想法是，以 OLS 残差平方 $\{e_1^2,\cdots,e_n^2\}$ 替代上式的 $\{\sigma_1^2,\cdots,\sigma_n^2\}$，得到扰动项协方差矩阵的估计量：

$$\widehat{\mathrm{Var}(\boldsymbol{\varepsilon}\mid\boldsymbol{X})}=\frac{n}{n-K}\begin{pmatrix}e_1^2 & & 0\\ & \ddots & \\ 0 & & e_n^2\end{pmatrix} \tag{6.27}$$

---

① 此定义与经济理论中关于内生变量与外生变量的区分有所不同。根据一般的经济理论，在经济系统(模型)内部决定的变量，称为“内生变量”；而在经济系统之外决定的变量，则称为“外生变量”。相对而言，计量经济学对于内生变量与外生变量的定义更为简单而直接。

其中，$\frac{n}{n-K}$为自由度的调整（在大样本下无差别）。将表达式(6.27)代入方程(6.24)，可得如下方差估计量

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid X)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{6.28}$$

当然，上式只在大样本下才成立；而当 $n \to \infty$ 时，对参数 $\hat{\boldsymbol{\beta}}$ 的估计变得无限准确，故 $\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid X)} \to 0$。为此，考虑$\sqrt{n}\hat{\boldsymbol{\beta}}$ 的方差估计量，即 $\hat{\boldsymbol{\beta}}$ 的渐近方差估计量[①]：

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})} = n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{6.29}$$

可以证明[②]，上式为 $\hat{\boldsymbol{\beta}}$ 渐近协方差矩阵的一致估计量，即

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})} \xrightarrow{p} \mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) \tag{6.30}$$

由于表达式(6.29)在推导过程中并未假设“条件同方差”，故它提供了在“条件异方差”情况下也成立的标准误，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)。在形式上，稳健标准误也是夹心估计量。稳健标准误的思想最早由 Eicker(1967)与 Huber(1967)提出，并由 White(1980)严格证明，故也称 White's standard errors，Huber-White standard errors，或 Eicker-Huber-White standard errors。

稳健标准误的表达式(6.29)虽然比较复杂，但对于计算机而言，其计算成本可以忽略(也无需人为记忆)。通过使用迭代期望定律可以证明，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[③]。特别地，考虑同方差的一种极端情形，即 $e_1^2 = e_2^2 = \cdots = e_n^2$(所有残差的绝对值都相等，但符号可以相反)，则

$$\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})} = \frac{n}{n-K}\begin{pmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{pmatrix} = \frac{ne_i^2}{n-K}\boldsymbol{I}_n = \frac{\sum_{i=1}^{n} e_i^2}{n-K}\boldsymbol{I}_n = s^2\boldsymbol{I}_n \tag{6.31}$$

此时，稳健的协方差矩阵可简化为同方差情况下的普通(非稳健)协方差矩阵：

$$\begin{aligned}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(s^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned} \tag{6.32}$$

## 6.9 大样本统计推断

对于渐近独立的平稳过程，如果样本容量足够大，则 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的渐近正态分布是对其

① 根据定义，$\hat{\boldsymbol{\beta}}$ 的渐近方差其实指的是 $\mathrm{Var}(\sqrt{n}\hat{\boldsymbol{\beta}})$ 的概率极限。

② 参见陈强(2014，p.58)。

③ 参见陈强(2014，p.61)。

真实分布的较好近似，故可使用其渐近分布进行大样本假设检验与区间估计。大样本统计推断（large sample inference）的步骤与小样本 OLS 基本相同。

**1. 检验单个系数：$H_0:\beta_k=c$**

考虑检验 $H_0:\beta_k=c$，其中 $c$ 为已知常数。根据大样本理论，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 服从渐近正态分布，即 $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N\left(\boldsymbol{0},\mathrm{Avar}(\hat{\boldsymbol{\beta}})\right)$，其中 $\mathrm{Avar}(\hat{\boldsymbol{\beta}})$ 为渐近协方差矩阵。具体到 $\hat{\boldsymbol{\beta}}$ 的第 $k$ 个元素 $\hat{\beta}_k$，则有

$$\sqrt{n}(\hat{\beta}_k-\beta_k)\xrightarrow{d}N\left(0,\mathrm{Avar}(\hat{\beta}_k)\right) \tag{6.33}$$

其中，$\mathrm{Avar}(\hat{\beta}_k)$ 为 $\hat{\beta}_k$ 的渐近方差，即渐近方差矩阵 $\mathrm{Avar}(\hat{\boldsymbol{\beta}})$ 主对角线上的第 $k$ 个元素。在原假设 $H_0$ 成立的情况下，$\beta_k=c$，故表达式(6.33)可写为

$$\sqrt{n}(\hat{\beta}_k-c)\xrightarrow{d}N\left(0,\mathrm{Avar}(\hat{\beta}_k)\right) \tag{6.34}$$

进一步，记 $\widehat{\mathrm{Avar}(\hat{\beta}_k)}$ 为渐近方差矩阵估计量 $\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}})}$ 主对角线上的第 $k$ 个元素，则 $\widehat{\mathrm{Avar}(\hat{\beta}_k)}$ 是 $\mathrm{Avar}(\hat{\beta}_k)$ 的一致估计量。定义 $t$ 统计量为

$$t_k\equiv\frac{\sqrt{n}(\hat{\beta}_k-c)}{\sqrt{\widehat{\mathrm{Avar}(\hat{\beta}_k)}}}=\frac{\hat{\beta}_k-c}{\sqrt{\frac{1}{n}\widehat{\mathrm{Avar}(\hat{\beta}_k)}}}\equiv\frac{\hat{\beta}_k-c}{\mathrm{SE}^*(b_k)}\xrightarrow{d}N(0,1) \tag{6.35}$$

其中，$\mathrm{SE}^*(\hat{\beta}_k)\equiv\sqrt{\frac{1}{n}\widehat{\mathrm{Avar}(\hat{\beta}_k)}}$ 即为异方差稳健的标准误。之所以这样称呼，是因为在其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 $t_k$ 称为“稳健 $t$ 比值”（robust $t$ ratio），服从渐近标准正态分布，而不是 $t$ 分布。

显然，对于双边检验（即 $H_1:\beta_k\neq c$），则 $|t_k|$ 越大，越倾向于拒绝 $H_0$。比如，对于 5% 的显著性水平，如果 $|t_k|$ 大于临界值 1.96，则可拒绝 $H_0$。也可以通过 $p$ 值进行检验，其方法与小样本理论相同（参见第 5 章）。

**2. 检验线性假设：$H_0:R\beta=r$**

在大样本理论下，对于多个线性假设的联合检验，与小样本理论下的 $F$ 检验类似。考虑检验 $m$ 个线性假设是否同时成立：

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

其中，$\boldsymbol{r}$ 为 $m$ 维列向量（$m<K$），$\boldsymbol{R}$ 为 $m\times K$ 矩阵，且 $\mathrm{rank}(\boldsymbol{R})=m$，即 $\boldsymbol{R}$ 满行秩，没有多余或自相矛盾的行或方程。对于原假设 $H_0:\boldsymbol{R\beta}=\boldsymbol{r}$，根据沃尔德检验原理，可考察 $(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$ 的大小，譬如二次型 $(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$。在 $H_0$ 成立的情况下，可以证明统计量

$$W \equiv n(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}\,\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \xrightarrow{d} \chi^2(m) \tag{6.36}$$

其中，$\boldsymbol{R}\,\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})}\boldsymbol{R}'$为$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$的渐近方差矩阵(使用夹心估计量公式)。

如果统计量 $W$ 大于$\chi^2(m)$的临界值，则拒绝原假设。在表达式(6.36)中，虽然统计量 $W$ 服从$\chi^2$ 分布，而非小样本理论下的 $F$ 分布，但$\chi^2$ 分布与 $F$ 分布在大样本情况下是等价的。事实上，即使在大样本下使用稳健标准误进行假设检验，Stata 也依然汇报 $F$ 统计量及其 $p$ 值。

**命题** 假设统计量 $F \sim F(m,n)$分布，则当 $n\to\infty$ 时，$mF \xrightarrow{d} \chi^2(m)$。

**证明：** 因为 $F \sim F(m,n)$，故可写为 $F = \dfrac{\chi^2(m)/m}{\chi^2(n)/n}$，其中分子与分母相互独立。

根据$\chi^2$ 分布的性质，$\chi^2$ 分布的期望等于自由度，而方差等于自由度的两倍；即 $\text{E}[\chi^2(n)] = n$，而 $\text{Var}[\chi^2(n)] = 2n$。

考察此 $F$ 统计量的分母$\chi^2(n)/n$。其期望为 $\text{E}[\chi^2(n)/n] = n/n = 1$，而方差为 $\text{Var}[\chi^2(n)/n] = 2n/n^2 = 2/n \to 0$(当 $n\to\infty$ 时)。因此，此 $F$ 统计量的分母依均方收敛于 1，故依概率收敛于 1(前者是后者的充分条件)，即$\chi^2(n)/n \xrightarrow{p} 1$。换言之，在大样本下，$\chi^2(n)/n$ 退化为 1，此 $F$ 统计量的性质仅由分子$\chi^2(m)/m$ 决定，故 $F \xrightarrow{d} \chi^2(m)/m$。因此，在大样本下，$mF \xrightarrow{d} \chi^2(m)$。

## 6.10 大样本 OLS 的 Stata 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。

下面以数据集 nerlove. dta 为例，取自 Nerlove(1963)对电力行业规模报酬的经典研究[①]。此数据集包括 1955 年美国 145 家电力企业的横截面数据，主要变量为 $TC$(total cost，总成本)，$Q$(total output，总产量)，$P_L$(price of labor，小时工资率)，$P_K$(user cost of capital，资本的使用成本)与 $P_F$(price of fuel，燃料价格)，以及相应的对数值 ln $TC$，ln $Q$，ln $P_L$，ln $P_K$ 与 ln $P_F$。

假设企业 $i$ 的生产函数为 Cobb-Douglas 函数：

$$Q_i = A_i L_i^{\alpha_1} K_i^{\alpha_2} F_i^{\alpha_3} \tag{6.37}$$

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)的主要目的是确定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[②]：

$$TC_i = \delta_i Q_i^{1/r} (P_L)_i^{\alpha_1/r} (P_K)_i^{\alpha_2/r} (P_F)_i^{\alpha_3/r} \tag{6.38}$$

其中，$\delta_i$ 是 $A_i,\alpha_1,\alpha_2,\alpha_3$ 的函数。取对数后得到如下模型，

① 此例来自 Hayashi(2000)。

② 参见标准的微观经济学教材。

$$\ln TC_i = \beta_1 + \frac{1}{r}\ln Q_i + \frac{\alpha_1}{r}\ln P_{L,i} + \frac{\alpha_2}{r}\ln P_{K,i} + \frac{\alpha_3}{r}\ln P_{F,i} + \varepsilon_i \tag{6.39}$$

这就是 Nerlove(1963)所要估计的主要方程。首先,打开数据集 nerlove. dta,并使用普通标准误对方程(6.39)进行 OLS 估计:

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

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

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

从上表可知,$R^2 = 0.926\ 0$,$\overline{R}^2 = 0.923\ 9$,检验整个方程显著性的 $F$ 统计量高达 437.9,其相应 $p$ 值(Prob > F)为 0.000 0,表明此回归方程高度显著。然而,$\ln P_L$ 与 $\ln P_K$ 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 $\ln P_K$ 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。

由于 $\ln Q$ 的系数为 $1/r$(即规模报酬的倒数),故可估计规模报酬为

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

其中,"_b[lnq]"表示 $\ln q$ 的 OLS 系数估计值。

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

```
. test lnq=1
```

此命令检验的原假设为,变量 $\ln Q$ 的系数等于 1。

```
( 1)  lnq = 1

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

由于 $p$ 值为 0.000 0,故强烈拒绝原假设,认为存在规模报酬递增。

其次,使用稳健标准误重新进行回归。

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

```
Linear regression                               Number of obs =     145
                                                F(  4,   140) =  177.19
                                                Prob > F      =  0.0000
                                                R-squared     =  0.9260
                                                Root MSE      =  .39227
```

| lntc | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | .7209135 | .0325376 | 22.16 | 0.000 | .656585 | .785242 |
| lnpl | .4559645 | .260326 | 1.75 | 0.082 | -.0587139 | .9706429 |
| lnpk | -.2151476 | .3233711 | -0.67 | 0.507 | -.8544698 | .4241745 |
| lnpf | .4258137 | .0740741 | 5.75 | 0.000 | .2793653 | .5722622 |
| _cons | -3.566513 | 1.718304 | -2.08 | 0.040 | -6.963693 | -.1693331 |

对比以上两个回归的结果可知，使用选择项“robust”所得到的 OLS 回归系数完全相同，只是所得到的稳健标准误（Robust Std. Err.）与普通标准误（Std. Err.）不同[①]。对于变量 ln $Q$ 的系数，其稳健标准误（0.033）几乎是普通标准误（0.017）的两倍。另外，其他变量系数的稳健标准误反而比普通标准误有所下降。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 ln $Q$ 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 章介绍。

在 Stata 中使用稳健标准误，即可进行大样本检验。对单个变量系数显著性的检验，可以使用上表中的稳健 $t$ 统计量（服从正态分布）来进行。更直观地，可直接看表中所列的 $p$ 值（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 ln $Q$ 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 0.000 0，故即使用稳健标准误，也仍然强烈拒绝“变量 ln $Q$ 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 $\chi^2$ 分布在大样本下是等价的，参见本章 6.9 节。

## 6.11 大样本理论的蒙特卡罗模拟

为了更直观地理解大样本理论，下面使用蒙特卡罗法来验证。考虑以下数据生成过程（DGP）：

① 不同的软件包所汇报的标准误可能略有不同，原因之一在于是否作“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。

$$y = \alpha + \beta x + \varepsilon, \quad x \sim \chi^2(1), \quad \varepsilon \sim \chi^2(10) - 10 \tag{6.40}$$

其中，$\alpha = 1$，$\beta = 2$，解释变量 $x$ 服从 $\chi^2(1)$ 分布；扰动项 $\varepsilon$ 服从经过位移后的 $\chi^2(10)$ 分布，以保证其期望为 0（卡方分布的期望为其自由度）；$x$ 与 $\varepsilon$ 相互独立。由于小样本理论要求扰动项服从正态分布，这个模型显然不满足小样本理论的假定，但符合大样本理论的要求。

首先，考虑样本容量为 20 的情形，看 OLS 估计量 $\hat{\beta}$ 与真实值 $\beta = 2$ 的差距，以及 $\hat{\beta}$ 的分布能否收敛到正态分布。为此，抽取 10 000 个样本容量为 20 的随机样本，进行回归，得到 10 000 个 $\hat{\beta}$。具体来说，先用命令 program 定义一个叫“chi2data”（可自行命名）的程序进行一次抽样；然后，用命令 simulate 来重复此程序 10 000 次：

```
. program chi2data,rclass                (定义程序 chi2data,以 r( )形式储存结果)
    drop _all                            (删去内存中已有数据)
    set obs 20                           (确定随机抽样的样本容量为 20)
    gen x = rchi2(1)                     (生成服从χ²(1)分布的解释变量)
    gen y = 1 + 2 * x + rchi2(10) - 10   (生成被解释变量)
    reg y x                              (线性回归)
    return scalar b = _b[x]              (存储β̂的估计值)
    end                                  (程序 chi2data 结束)
. set more off                           (指定 Stata 输出结果连续翻页)
. simulate bhat = r(b),reps(10000) seed(10101) nodots:chi2data
```

其中，选择项“reps(10000)”表示通过命令 simulate 将程序“chi2data”模拟 10 000 次。得到 10 000 个 $\hat{\beta}$ 后，可计算其均值与标准差：

```
. sum bhat
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| bhat | 10000 | 1.990334 | .967356 | -3.513781 | 8.522547 |

从上表可知，$\hat{\beta}$ 的样本均值为 1.990，很接近真实值 2，验证了 $\hat{\beta}$ 为 $\beta$ 的无偏估计。但标准（误）差为 0.967，接近于 1，故估计误差较大（因为样本容量仅为 20）。下面，通过直方图来看这 10 000 个 $\hat{\beta}$ 的分布，结果参见图 6.12。

```
. hist bhat,normal
```

其中，选择项“normal”表示同时画相应的正态分布密度图。

从图 6.12 可知，当样本容量为 20 时，$\hat{\beta}$ 的真实分布与正态分布仍有一定差距。

其次，将样本容量增加至 100，仍然抽取 10 000 个随机样本，即在上述程序中将命令“set obs 20”改为“set obs 100”，再次得到 10 000 个 $\hat{\beta}$；然后看 $\hat{\beta}$ 的统计特征。

```
. sum bhat
```

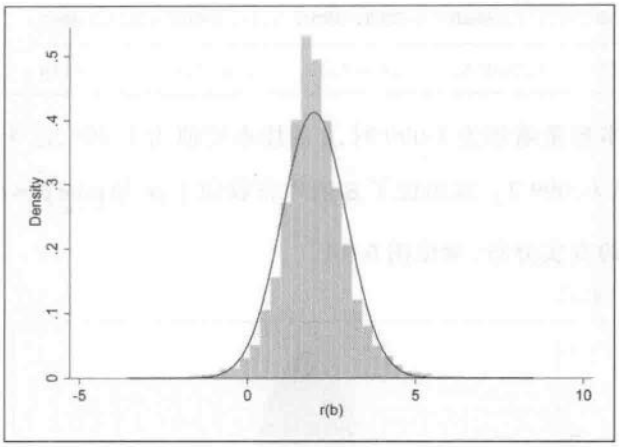



图 6.12 $\hat{\beta}$ 的分布(样本容量为 20)

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| bhat | 10000 | 1.99551 | .3359594 | .7352199 | 3.459108 |

从上表可知,$\hat{\beta}$ 的样本均值为 1.996,更加接近真实值 2;更重要的是,当样本容量从 20 增加到 100 后,$\hat{\beta}$ 的标准(误)差从 0.967 下降到 0.336。下面,画 $\hat{\beta}$ 的直方图,并与正态分布比较,结果参见图 6.13。

```
. hist bhat,normal
```

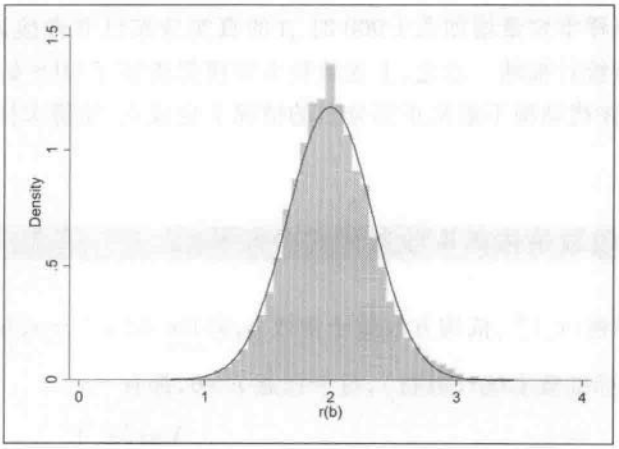



图 6.13 $\hat{\beta}$ 的分布(样本容量为 100)

从图 6.13 可知,当样本容量为 100 时,$\hat{\beta}$ 的真实分布与正态分布已较为接近。为了进一步验证 OLS 估计量 $\hat{\beta}$ 的一致性与渐近正态性,下面将样本容量增加为 1 000,得到 10 000 个 $\hat{\beta}$,再看其统计特征。

```
. sum bhat
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| bhat | 10000 | 1.999062 | .0997443 | 1.620384 | 2.429879 |

从上表可知，当样本容量增加为 1 000 时，$\hat{\beta}$ 的样本均值为 1.999，已十分接近真实值 2；而 $\hat{\beta}$ 的标准(误)差则下降为 0.099 7。这验证了 $\hat{\beta}$ 依均方收敛于 $\beta$，故 $\operatorname{plim}_{n\to\infty}\hat{\beta}=\beta$，即 $\hat{\beta}$ 为一致估计量。下面，通过直方图看 $\hat{\beta}$ 的真实分布，参见图 6.14。

```
. hist bhat,normal
```

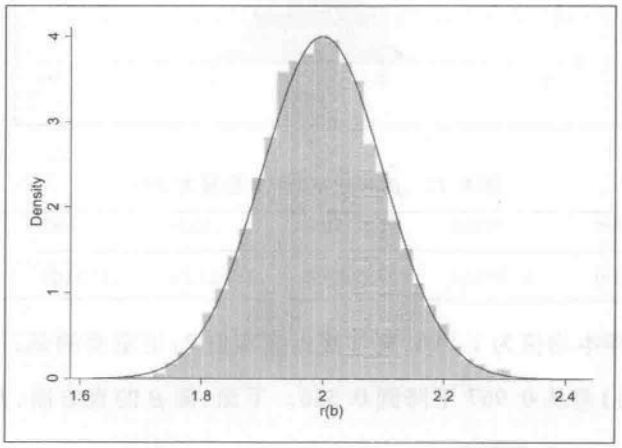



图 6.14 $\hat{\beta}$ 的分布(样本容量为 1 000)

从图 6.14 可知，当样本容量增加为 1 000 时，$\hat{\beta}$ 的真实分布已非常接近正态分布，可以放心地使用大样本理论进行统计推断。总之，上述蒙特卡罗模拟验证了 OLS 估计量的一致性与渐近正态性；这些性质即使在扰动项不服从正态分布的情况下也成立，使得大样本理论具有很大的适用性与稳健性。

## 附录 A6.1 依均方收敛是依概率收敛的充分条件

**证明：** 假设随机序列 $\{x_n\}_{n=1}^{\infty}$ 依均方收敛于常数 $a$，即 $\lim_{n\to\infty}\mathrm{E}(x_n)=a$，$\lim_{n\to\infty}\mathrm{Var}(x_n)=0$。根据切比雪夫不等式(参见标准概率统计教材)，对于任意 $\varepsilon>0$，都有

$$P(|x_n-\mathrm{E}(x_n)|\geqslant\varepsilon)\leqslant\frac{\mathrm{Var}(x_n)}{\varepsilon^2} \tag{6.41}$$

当 $n\to\infty$ 时，对此不等式两边同时取极限可得

$$\lim_{n\to\infty}P(|x_n-a|\geqslant\varepsilon)=\lim_{n\to\infty}P(|x_n-\mathrm{E}(x_n)|\geqslant\varepsilon)\leqslant\lim_{n\to\infty}\frac{\mathrm{Var}(x_n)}{\varepsilon^2}=0 \tag{6.42}$$

根据依概率收敛的定义可知，$\{x_n\}_{n=1}^{\infty}$依概率收敛于 $a$。

**习题**

6.1 假设随机变量 $y$ 的期望为 $\mu$，抽样得到其 iid 随机样本 $\{y_1, \cdots, y_n\}$，记样本均值为 $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$。

(1) $\bar{y}$ 是 $\mu$ 的无偏估计。$\bar{y}^2$ 是否是 $\mu^2$ 的无偏估计？

(2) $\bar{y}$ 是 $\mu$ 的一致估计。$\bar{y}^2$ 是否是 $\mu^2$ 的一致估计？

6.2 考虑随机序列 $\{x_n\}_{n=1}^{\infty}$，其中 $x_n$ 取值为 0 的概率为 $1-(1/n)$，而取值为 $n$ 的概率为 $1/n$。证明此序列的方差发散，即 $\lim_{n\to\infty} \mathrm{Var}(x_n) = \infty$。

6.3 证明线性模型的离差形式，即方程(6.22)。

6.4 当 $n\to\infty$ 时，是否稳健标准误 $\mathrm{SE}^*(\hat{\beta}_k) \xrightarrow{p} 0$？为什么？

6.5 使用数据集 grilic.dta，以稳健标准误估计以下回归方程：

$$\ln w = \beta_1 + \beta_2 s + \beta_3 expr + \beta_4 tenure + \beta_5 smsa + \varepsilon \tag{6.43}$$

其中，$\ln w$ 为工资对数，$s$ 为教育年限，*expr* 为工龄，*tenure* 为在现单位工作年限，而 *smsa* 表示是否住在大城市。另外，变量 *rns* 表示是否住在美国南方。

(1) 使用全样本，估计方程(6.43)。

(2) 使用美国南方的子样本，估计方程(6.43)。

(3) 使用美国北方的子样本，估计方程(6.43)。

(4) 与全样本相比，子样本估计量的标准误有何变化，为什么？

6.6 房屋的价格如何决定？一种理论认为，房价由房屋性能所决定，称为“特征价格法”(hedonic price)。数据集 hprice2a.dta 包含美国波士顿 506 个社区的房屋中位数价格的横截面数据。[①] 考虑以下特征价格回归：

$$lprice_i = \beta_1 + \beta_2 lnox_i + \beta_3 ldist_i + \beta_4 rooms_i + \beta_5 stratio_i + \varepsilon_i \tag{6.44}$$

其中，*lprice* 为房价的对数，*lnox* 为空气污染程度的对数，*ldist* 为社区到就业中心距离的对数，*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 $i$ 表示社区 $i$。

(1) 使用普通标准误进行回归，并评论解释变量系数的符号、统计显著性及经济意义。

(2) 使用稳健标准误进行回归，稳健标准误与普通标准误差别大吗？

(3) 使用稳健标准误，以 5% 的显著性水平，检验 $H_0: \beta_3 = \beta_5$。

(4) 使用稳健标准误，以 5% 的显著性水平，分别检验 $H_0: \beta_4 = 0.31$ 与 $H_0: \beta_4 = 0.30$。

---

① 此数据集来自 Baum(2006)。

# 7. 异 方 差

古典线性回归模型假设的是一种理想状态(参见第 5 章)。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,我们将逐步放松古典模型的各项假定。

## 7.1 异方差的后果

"条件异方差"(conditional heteroskedasticity),简称"异方差"(heteroskedasticity),是违背球形扰动项假设的一种情形,即条件方差 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{X})$ 依赖于 $i$,而不是常数 $\sigma^2$。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,在证明这些性质时,并未用到"同方差"的假定。

(2) OLS 估计量方差 $\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})$ 的表达式不再是 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$,因为 $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) \neq \sigma^2\boldsymbol{I}$。因此,使用普通标准误的 $t$ 检验、$F$ 检验失效。

(3) 高斯 - 马尔可夫定理不再成立,OLS 不再是 BLUE(最佳线性无偏估计)。在存在异方差的情况下,本章将要介绍的"加权最小二乘法"才是 BLUE。为了直观地理解为何 OLS 不再是 BLUE,考虑一元回归 $y_i = \alpha + \beta x_i + \varepsilon_i$,并假设 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{X})$ 是解释变量 $x_i$ 的增函数,即 $x_i$ 越大则 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{X})$ 越大,参见图 7.1。

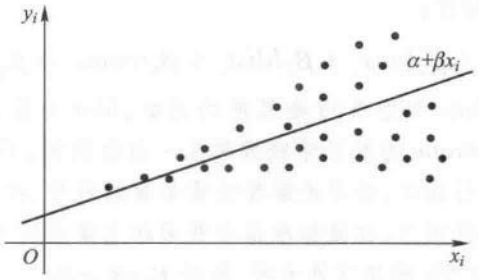



图 7.1 异方差示意图

显然,OLS 回归线在 $x_i$ 较小时可以较精确地估计,而在 $x_i$ 较大时则难以准确估计。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。"加权最小二乘法"(Weighted Least Square, 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。

需要强调的是,计量经济学所指的"异方差"一般都是"条件异方差",而非"无条件异方差"。比如,大样本 OLS 理论要求样本数据为平稳过程,而平稳过程的方差不变。这是否意味

着,大样本 OLS 理论已经假设了同方差?这里关键要区分无条件方差(unconditional variance)与条件方差(conditional variance)。

以一元回归模型 $y_i = \alpha + \beta x_i + \varepsilon_i$ 为例,假设 $\{x_i, y_i\}$ 为平稳过程,则 $\varepsilon_i = y_i - \alpha - \beta x_i$ 也是平稳过程[①],故其无条件方差 $\mathrm{Var}(\varepsilon_i) = \sigma^2$ 为常数,不随 $i$ 而变。进一步,所有个体的条件方差函数 $\mathrm{Var}(\varepsilon_i \mid x_1, \cdots, x_n)$ 在函数形式上也完全相同,比如,$\mathrm{Var}(\varepsilon_i \mid x_1, \cdots, x_n) = x_i^2$。然而,此条件方差函数的具体取值却依赖于 $x_i$,故仍可存在条件异方差。比如,$\mathrm{Var}(\varepsilon_1 \mid x_1, \cdots, x_n) = x_1^2$,$\mathrm{Var}(\varepsilon_2 \mid x_1, \cdots, x_n) = x_2^2$,以此类推。

## 7.2 异方差的例子

(1) 考虑以下消费函数:

$$c_i = \alpha + \beta y_i + \varepsilon_i \tag{7.1}$$

其中,$c_i$ 为消费,$y_i$ 为收入。一般来说,一方面,富人的消费计划较有弹性,而穷人的消费多为必需品,很少变动。另一方面,富人的消费支出可能更难测量,故包含较多测量误差。因此,$\mathrm{Var}(\varepsilon_i \mid y_i)$ 可能随 $y_i$ 的上升而变大。

(2) 企业的投资、销售与利润: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,则可能存在异方差。

(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,而第二组为打工族的收入,则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通常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,则人口较多的省份其方差较小,方差与人口数成反比。

## 7.3 异方差的检验

### 1. 画残差图(residual plot)

由于残差可视为扰动项的实现值,故可通过残差的波动来大致考察是否存在异方差。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),或"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。

### 2. BP 检验(Breusch and Pagan,1979)

判断异方差的严格方法仍需通过统计检验。假设回归模型为

① 平稳过程的线性组合依然是平稳过程。

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \tag{7.2}$$

记 $\boldsymbol{x}_i = (1\ x_{i2} \cdots\ x_{iK})$。假设样本数据为 iid，则 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{X}) = \mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i)$。因此，“条件同方差”的原假设为

$$H_0: \mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i) = \sigma^2 \tag{7.3}$$

由于 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i) = \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i) - [\underbrace{\mathrm{E}(\varepsilon_i \mid \boldsymbol{x}_i)}_{=0}]^2 = \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i)$，故可将原假设写为

$$H_0: \mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i) = \sigma^2 \tag{7.4}$$

如果 $H_0$ 不成立，则条件方差 $\mathrm{E}(\varepsilon_i^2 \mid \boldsymbol{x}_i)$ 是 $\boldsymbol{x}_i$ 的函数，称为“条件方差函数”(conditional variance function)。假设此条件方差函数为线性函数：

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + u_i \tag{7.5}$$

则原假设可简化为

$$H_0: \delta_2 = \cdots = \delta_K = 0 \tag{7.6}$$

在方程(7.5)中，由于扰动项 $\varepsilon_i$ 不可观测，故使用残差平方 $e_i^2$ 替代之，进行以下辅助回归(auxiliary regression)[①]：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \tag{7.7}$$

记此辅助回归的拟合优度为 $R^2$。显然，$R^2$ 越高，则此辅助回归方程越显著，越可以拒绝 $H_0$: $\delta_2 = \cdots = \delta_K = 0$。Breusch and Pagan(1979) 使用 *LM* 统计量，进行 *LM* 检验(Lagrange Multiplier Test，参见第 11 章)：

$$LM = nR^2 \xrightarrow{d} \chi^2(K-1) \tag{7.8}$$

如果 *LM* 统计量大于 $\chi^2(K-1)$ 的临界值，则拒绝同方差的原假设。为什么 *LM* 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。

首先，对于辅助回归(7.7)，检验原假设 $H_0: \delta_2 = \cdots = \delta_K = 0$ 的 *F* 统计量为(参见第 5 章)

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K) \tag{7.9}$$

其次，在大样本情况下，*F* 分布与 $\chi^2$ 分布是等价的(参见第 6 章)，即 $(K-1)F = \dfrac{(n-K)R^2}{(1-R^2)} \xrightarrow{d} \chi^2(K-1)$。进一步，在 $H_0: \delta_2 = \cdots = \delta_K = 0$ 成立的情况下，辅助回归方程(7.7)仅对常数项回归，故当 $n \to \infty$ 时，$R^2 \xrightarrow{p} 0$，而 $(1-R^2) \xrightarrow{p} 1$。因此，

$$(K-1)F = \frac{(n-K)R^2}{1-R^2} \xrightarrow{p} (n-K)R^2 \tag{7.10}$$

显然，在大样本下，$(n-K)R^2$ 与 $nR^2$ 并无差别，故 *LM* 检验与 *F* 检验渐近等价。

如果认为异方差主要依赖于被解释变量的拟合值，可将辅助回归(7.7)改为

---

① 辅助回归的含义是，我们通常对此回归本身的结果并不直接感兴趣，但可通过此回归计算所需的统计量。

$$e_i^2 = \delta_1 + \delta_2 \hat{y}_i + error_i \tag{7.11}$$

其中，$\hat{y}_i$ 为回归方程(7.2)的拟合值；然后检验 $H_0: \delta_2 = 0$(可使用 $F$ 或 $LM$ 统计量)。

Breusch and Pagan(1979)的最初检验假设扰动项 $\varepsilon_i$ 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。

**3. 怀特检验(White,1980)**

BP 检验假设条件方差函数为线性函数，只是对条件方差函数的一阶近似，可能忽略了高次项。为此，怀特检验(White,1980)在 BP 检验的辅助回归(7.7)中加入所有的二次项(含平方项与交叉项)。

不失一般性，考虑以下二元回归：

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \tag{7.12}$$

其中，除常数项外，只有两个解释变量 $x_{i2}$ 与 $x_{i3}$，故二次项包括 $x_{i2}^2$，$x_{i3}^2$ 与 $x_{i2}x_{i3}$。因此，怀特检验的辅助回归为

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + \delta_4 x_{i2}^2 + \delta_5 x_{i3}^2 + \delta_6 x_{i2} x_{i3} + error_i \tag{7.13}$$

其中，$e_i^2$ 为回归方程(7.12)的残差平方。然后，对原假设 $H_0: \delta_2 = \cdots = \delta_6 = 0$ 进行 $F$ 检验或 $LM$ 检验。怀特检验的优点是，它可以检验任何形式的异方差，因为根据泰勒展开式(Taylor expansion)，二次函数可以很好地逼近任何光滑函数。怀特检验的缺点是，如果解释变量较多，则解释变量的二次项(含交叉项)将更多，在辅助回归中将损失较多样本容量。

## 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(OLS 仍然无偏、一致且渐近正态)，但使用在异方差情况下也成立的稳健标准误。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，只要使用稳健标准误，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 更有效率的方法，即下文的加权最小二乘法。

**2. 加权最小二乘法(WLS)**

由于方差较小的观测值包含的信息量较大，故对于异方差的另一处理方法是，给予方差较小的观测值较大的权重，然后进行加权最小二乘法估计。对于存在异方差的数据，WLS 的基本思想是，通过变量转换，使得变换后的模型满足球形扰动项的假定(变为同方差)，然后进行 OLS 估计，即为最有效率的 BLUE。

考虑线性回归模型：

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \tag{7.14}$$

其中，假定 $\mathrm{Var}(\varepsilon_i \mid \boldsymbol{x}_i) \equiv \sigma_i^2 = \sigma^2 v_i$，而且个体 $i$ 的异方差因子 $\{v_i\}_{i=1}^n$ 为已知。在上式两边同时乘以权重 $1/\sqrt{v_i}$（个体 $i$ 的标准差倒数）可得

$$\frac{y_i}{\sqrt{v_i}} = \beta_1 \frac{1}{\sqrt{v_i}} + \beta_2 \frac{x_{i2}}{\sqrt{v_i}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{v_i}} + \frac{\varepsilon_i}{\sqrt{v_i}} \tag{7.15}$$

在上式中，新扰动项 $\varepsilon_i/\sqrt{v_i}$ 不再存在异方差，因为

$$\mathrm{Var}(\varepsilon_i/\sqrt{v_i}) = \frac{1}{v_i}\mathrm{Var}(\varepsilon_i) = \frac{\sigma^2 v_i}{v_i} = \sigma^2 \tag{7.16}$$

对方程(7.15)进行 OLS 回归，即为 WLS。加权之后的回归方程满足球形扰动项的假定，故是 BLUE。从方程(7.15)可知，也可将 WLS 定义为最小化“加权的残差平方和”，即

$$\min \sum_{i=1}^{n} (e_i/\sqrt{v_i})^2 = \sum_{i=1}^{n} \frac{e_i^2}{v_i} \tag{7.17}$$

从这个角度来看，权重为 $1/v_i$（即方差的倒数），在 Stata 中也是这样约定的。需要注意的是，加权最小二乘法的 $R^2$ 通常没有太大意义，因为它衡量的是变换之后的解释变量（$x_{ik}/\sqrt{v_i}$）对变换之后的被解释变量（$y_i/\sqrt{v_i}$）的解释力，而我们一般对此没有太大兴趣。

### 3. 可行加权最小二乘法（FWLS）

使用 WLS 虽然可得到 BLUE 估计，但前提是，必须确切地知道每位个体的方差，即 $\{\sigma_i^2\}_{i=1}^n$。而在实践中，我们通常不知道 $\{\sigma_i^2\}_{i=1}^n$，故 WLS 事实上“不可行”（infeasible）。解决方法是先用样本数据估计 $\{\sigma_i^2\}_{i=1}^n$，然后再使用 WLS，称为“可行加权最小二乘法”（Feasible WLS，FWLS）。

在作 BP 检验时，进行如下辅助回归：

$$e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \tag{7.18}$$

其中，$e_i^2$ 为原方程(7.14)的残差平方。通过此辅助回归的拟合值，即可获得 $\sigma_i^2$ 的估计值：

$$\hat{\sigma}_i^2 = \hat{\delta}_1 + \hat{\delta}_2 x_{i2} + \cdots + \hat{\delta}_K x_{iK} \tag{7.19}$$

然而，上式可能出现 $\hat{\sigma}_i^2 < 0$ 的情形，而方差不能为负。为保证 $\hat{\sigma}_i^2$ 始终为正，一般假设条件方差函数为对数形式：

$$\ln e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_K x_{iK} + error_i \tag{7.20}$$

对此方程进行 OLS 回归，可得 $\ln e_i^2$ 的预测值，记为 $\ln \hat{\sigma}_i^2$；进而得到拟合值 $\hat{\sigma}_i^2 = \exp(\ln \hat{\sigma}_i^2)$（一定为正数），然后以 $1/\hat{\sigma}_i^2$ 为权重对原方程(7.14)进行 WLS 估计，记此估计量为 $\hat{\beta}_{\mathrm{FWLS}}$。

### 4. 究竟使用“OLS + 稳健标准误”还是 FWLS

在理论上，WLS 是 BLUE。但实践中使用的 FWLS 并非线性估计，因为权重 $1/\hat{\sigma}_i^2$ 也是 $\boldsymbol{y}$ 的

函数。而且，由于$\hat{\boldsymbol{\beta}}_{\text{FWLS}}$是$\boldsymbol{y}$的非线性函数，故一般来说是有偏的。因此，$\hat{\boldsymbol{\beta}}_{\text{FWLS}}$甚至无资格参加 BLUE 的评选。FWLS 的优点主要体现在大样本理论中。如果$\hat{\sigma}_i^2$是$\sigma_i^2$的一致估计，则 FWLS 一致，且在大样本下比 OLS 更有效率。FWLS 的缺点是必须估计条件方差函数$\hat{\sigma}_i^2(\boldsymbol{x}_i)$，而通常并不知道条件方差函数的具体形式①。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，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。在 Stata 中的操作也十分简单，只要在命令 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 更有效率。因此，必须在稳健性与有效性之间做选择。前者相当于“万金油”（指谁都适用），而后者相当于“特效药”。由于“病情”通常难以诊断（无法判断条件异方差的具体形式），故“特效药”也可能失效，甚至起反作用。如果对$\sigma_i^2$估计不准确，则 FWLS 即使在大样本下也不是 BLUE，其估计效率可能还不如 OLS。因此，Stock and Watson(2012)推荐，在大多数情况下应使用“OLS + 稳健标准误”。

Wooldridge(2009)指出，如果确实存在严重的异方差，则可通过使用 FWLS 来提高估计效率。如果对于条件异方差函数的具体形式没有把握，即不知道经过加权处理之后的新扰动项$\varepsilon_i/\sqrt{v_i}$是否同方差，可在进行 WLS 回归时仍使用异方差稳健的标准误，以保证 FWLS 标准误的有效性。

另外，如果被解释变量取值为正，有时将被解释变量取对数，可以缓解异方差问题。

## 7.5 处理异方差的 Stata 命令及实例

下面以数据集 nerlove. dta 为例（参见第 6 章），演示如何在 Stata 中处理异方差。此数据集包括以下变量：*tc*（总成本），*q*（总产量），*pl*（工资率），*pk*（资本的使用成本）与 *pf*（燃料价格），以及相应的对数值 ln *tc*，ln *q*，ln *pl*，ln *pk* 与 ln *pf*。

### 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)
```

首先，打开数据集 nerlove. dta，并以 OLS 估计对数形式的成本函数：

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

① 此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（应放入哪些变量?），可能的函数形式就更多了。

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

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

为初步考察是否存在异方差，下面画残差与拟合值的散点图，结果参见图 7.2。

```
. rvfplot
```

从图 7.2 可大致看出，当总成本（ln *tc* 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 ln *q* 的散点图，结果参见图 7.3。

```
. rvpplot lnq
```

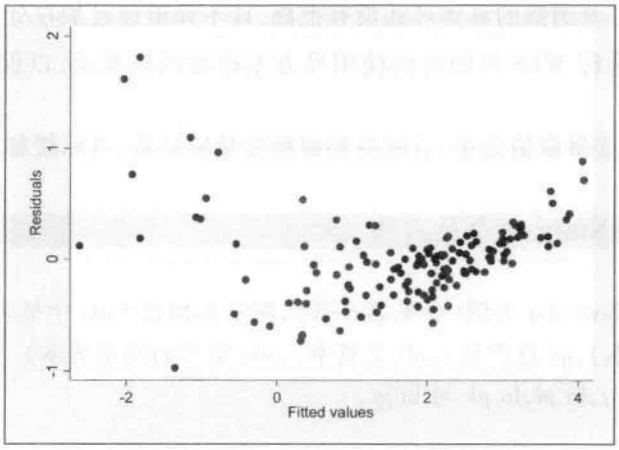



图 7.2 残差与拟合值的散点图

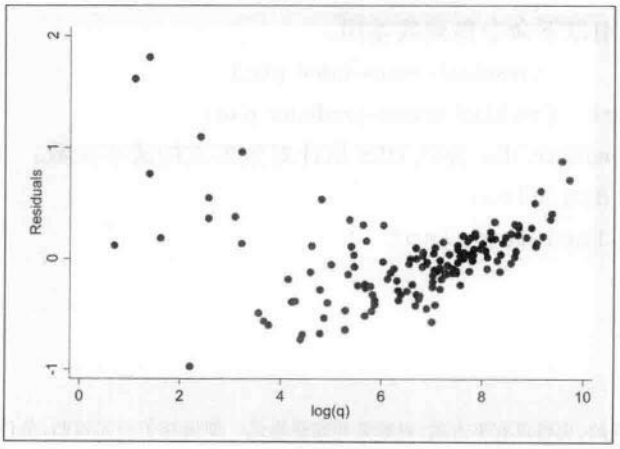



图 7.3 残差与解释变量 ln *q* 的散点图

从图 7.3 大致可知，产量（ln $q$）越小，则扰动项的方差越大。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. BP 检验**

在 Stata 中完成回归后，可使用以下命令进行 BP 检验：

```
estat hettest,iid rhs
```

其中，“estat”指 post-estimation statistics（估计后统计量），即在完成估计后所计算的后续统计量；“hettest”表示 heteroskedasticity test。选择项“iid”表示仅假定数据为 iid，而无需正态假定。选择项“rhs”表示，使用方程右边的全部解释变量进行辅助回归，默认使用拟合值 $\hat{y}$ 进行辅助回归。

如果想指定使用某些解释变量进行辅助回归，可使用如下命令：

```
estat hettest[varlist],iid
```

其中，“[varlist]”为指定的变量清单；而“[]”表示其中的内容可出现在命令中，也可不出现。回到 Nerlove（1963）的例子：

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

其中，前缀（prefix）“quietly”表示执行此命令，但不在 Stata 的结果窗口显示运行结果。首先，使用拟合值 $\hat{y}$ 进行 BP 检验。

```
. 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
```

其次，使用所有解释变量进行 BP 检验。

```
. estat hettest,iid rhs
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lnq lnpl lnpk lnpf

         chi2(4)      =    36.16
         Prob > chi2  =   0.0000
```

最后，使用变量 ln $q$ 进行 BP 检验。

```
. estat hettest lnq,iid
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lnq

         chi2(1)      =    32.10
         Prob > chi2  =   0.0000
```

以上各种形式 BP 检验的 $p$ 值都等于 0.000 0，故强烈拒绝同方差的原假设，认为存在异方差。此检验结果证实了根据残差图所做的大致判断。

### 3. 怀特检验

在 Stata 完成回归后，可使用如下命令进行怀特检验：

```
estat imtest,white
```

其中，“imtest”指 information matrix test（信息矩阵检验）。继续以 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（Prob > chi2）等于 0.000 0，故强烈拒绝同方差的原假设，认为存在异方差。此结果与 BP 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，可作为权重进行 WLS 估计。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。

继续以 Nerlove(1963)为例。首先计算残差，并记为 e1：

```
. quietly reg lntc lnq lnpl lnpk lnpf
. predict e1,residual
```

其次，生成残差的平方，并记为 e2：

```
. gen e2 = e1^2
```

将残差平方取对数：

```
. gen lne2 = log(e2)
```

假设 $\ln \hat{\sigma}_i^2$ 为变量 $\ln q$ 的线性函数，进行以下辅助回归：

```
. reg lne2 lnq
```

| Source | SS | df | MS | Number of obs = | 145 |
|---|---|---|---|---|---|
| | | | | F( 1, 143) = | 21.54 |
| Model | 105.722127 | 1 | 105.722127 | Prob > F = | 0.0000 |
| Residual | 701.999749 | 143 | 4.90908916 | R-squared = | 0.1309 |
| | | | | Adj R-squared = | 0.1248 |
| Total | 807.721876 | 144 | 5.60917969 | Root MSE = | 2.2156 |

| lne2 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | -.4479545 | .0965276 | -4.64 | 0.000 | -.6387597 | -.2571492 |
| _cons | -.7452062 | .6591018 | -1.13 | 0.260 | -2.048048 | .5576351 |

从上表可知，尽管变量 ln $q$ 在 1% 水平上显著，但 $R^2$ 仅为 0.130 9，而且常数项不显著（$p$ 值为 0.26）。因此，下面去掉常数项，重新进行辅助回归。

```
. reg lne2 lnq,noc
```

| Source | SS | df | MS | Number of obs = | 145 |
|---|---|---|---|---|---|
| | | | | F( 1, 144) = | 419.95 |
| Model | 2065.53636 | 1 | 2065.53636 | Prob > F = | 0.0000 |
| Residual | 708.275258 | 144 | 4.91857818 | R-squared = | 0.7447 |
| | | | | Adj R-squared = | 0.7429 |
| Total | 2773.81162 | 145 | 19.1297353 | Root MSE = | 2.2178 |

| lne2 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | -.5527533 | .0269733 | -20.49 | 0.000 | -.6060681 | -.4994384 |

上表显示，$R^2$ 上升为 0.744 7（尽管无常数项的 $R^2$ 与有常数项的 $R^2$ 不具有可比性，因为二者的定义不同），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。接着计算以上辅助回归的拟合值，并记为 lne2f：

```
. predict lne2f
(option xb assumed; fitted values)
```

去掉对数后，即得到方差的估计值，并记为 e2f：

```
. gen e2f = exp(lne2f)
```

最后，使用方差估计值的倒数作为权重，进行 WLS 回归：

```
. reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f]
```

| Source | SS | df | MS | Number of obs = 145 |
|---|---|---|---|---|
| | | | | F( 4, 140) = 895.03 |
| Model | 173.069988 | 4 | 43.2674971 | Prob > F = 0.0000 |
| Residual | 6.76790874 | 140 | .048342205 | R-squared = 0.9624 |
| | | | | Adj R-squared = 0.9613 |
| Total | 179.837897 | 144 | 1.24887428 | Root MSE = .21987 |

| lntc | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | .8759035 | .0153841 | 56.94 | 0.000 | .8454883 | .9063187 |
| lnpl | .5603879 | .1734141 | 3.23 | 0.002 | .2175389 | .9032369 |
| lnpk | -.0929807 | .1960402 | -0.47 | 0.636 | -.4805627 | .2946014 |
| lnpf | .4672438 | .0616476 | 7.58 | 0.000 | .3453632 | .5891243 |
| _cons | -5.522088 | .9928472 | -5.56 | 0.000 | -7.485 | -3.559176 |

WLS 回归的结果显示，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）。另外，使用 OLS 时，变量 ln *pl* 的 *p* 值为 0.13，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，使用 WLS 后提高了估计效率。

如果担心对条件方差函数的设定不准确，导致加权变换后的新扰动项仍有一定的异方差，可使用稳健标准误进行 WLS 估计：

```
. reg lntc lnq lnpl lnpk lnpf [aw=1/e2f],r
```

| Linear regression | | | | Number of obs = 145 |
|---|---|---|---|---|
| | | | | F( 4, 140) = 534.50 |
| | | | | Prob > F = 0.0000 |
| | | | | R-squared = 0.9624 |
| | | | | Root MSE = .21987 |

| lntc | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | .8759035 | .020787 | 42.14 | 0.000 | .8348064 | .9170006 |
| lnpl | .5603879 | .2090099 | 2.68 | 0.008 | .147164 | .9736118 |
| lnpk | -.0929807 | .3016444 | -0.31 | 0.758 | -.6893478 | .5033864 |
| lnpf | .4672438 | .0439915 | 10.62 | 0.000 | .3802702 | .5542173 |
| _cons | -5.522088 | 1.671596 | -3.30 | 0.001 | -8.826924 | -2.217252 |

从上表可知，无论是否使用稳健标准误，WLS 的回归系数都相同，但标准误有所不同。在此例中，多数解释变量（ln *q*，ln *pl*，ln *pk*）的稳健标准误大于普通标准误；但变量 ln *pf* 的稳健标准误反而小于普通标准误。

## 7.6 Stata 命令的批处理

在进行计量分析时，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 快捷键（参见图 7.4），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。

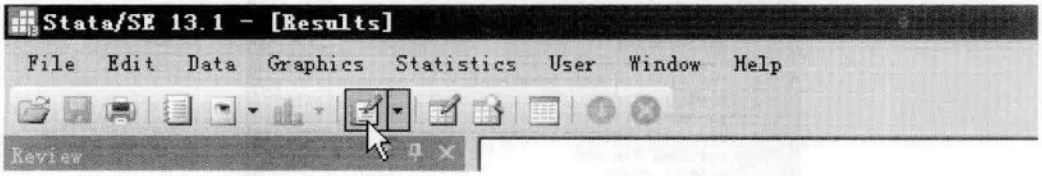



图 7.4 Stata 的 New Do-file Editor 快捷键

以上文的加权最小二乘法为例，假设数据文件 nerlove.dta 在当前路径，则可在“do 文件编辑器”中输入如下命令：

```
* WLS for Nerlove(1963)
capture log close
log using wls_nerlove,replace
set more off
use nerlove.dta,clear
reg lntc lnq lnpl lnpk lnpf
predict e1,r
gen e2 = e1^2
gen lne2 = log(e2)
reg lne2 lnq,noc
predict lne2f
gen e2f = exp(lne2f)
* Weighted least square regression
reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f]
reg lntc lnq lnpl lnpk lnpf [aw = 1/e2f],r
log close
exit
```

其中，“*”表示不执行其后的命令，常用来作为注释（便于自己或他人理解程序）。“capture log close”表示如有已打开的日志文件，先将其关闭（如有打开的日志文件，无法定义新的日志文件）。命令“log using wls_nerlove,replace”表示在当前路径创建名为“wls_nerlove.smcl”的日志文件（选择项 replace 表示可覆盖此文件的原有内容），并将 Stata 运行结果记录于此日志文件。命令“set more off”使得 Stata 输出结果可自动连贯显示，而无需点击

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。

“more”翻屏。

输入以上命令后，点击 Do-file Editor 的 Execute(do)快捷键即可运行此程序，参见图 7.5。

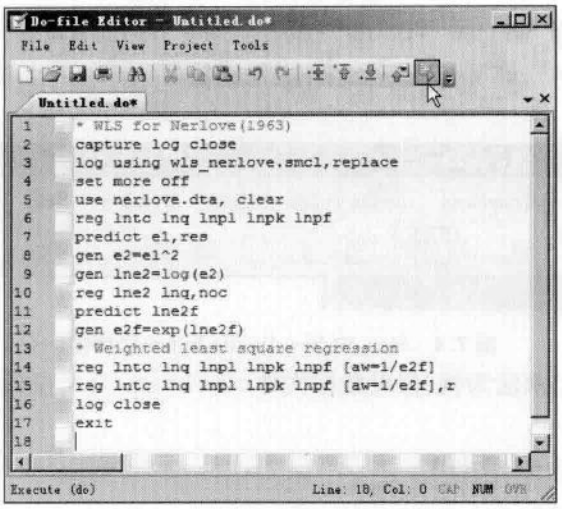



图 7.5 Do-file Editor 的 Execute(do)快捷键

如果要存储此程序文件，可点击 Do-file Editor 窗口的菜单 File→Save(或 Save As)，比如，将此程序文件存为“wls_nerlove. do”，参见图 7.6。


图 7.6 存储 do 文件

存储 Stata 的 do 文件后，可在 Stata 中点击菜单 File→Do，寻找“wls_nerlove. do”文件，然后执行此文件，参见图 7.7。

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，参见图 7.8。

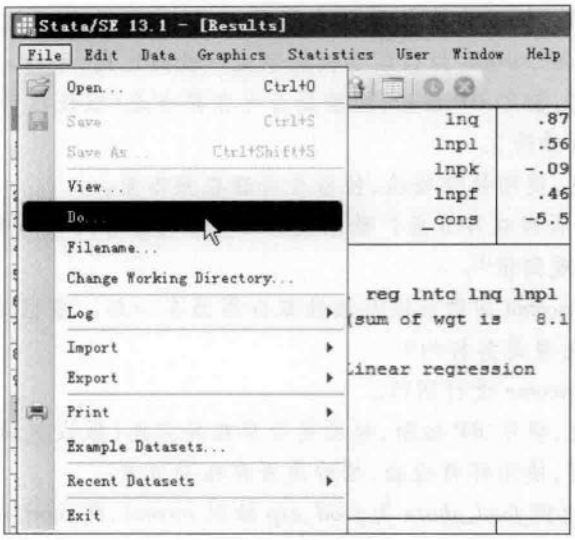



图 7.7 执行 do 文件

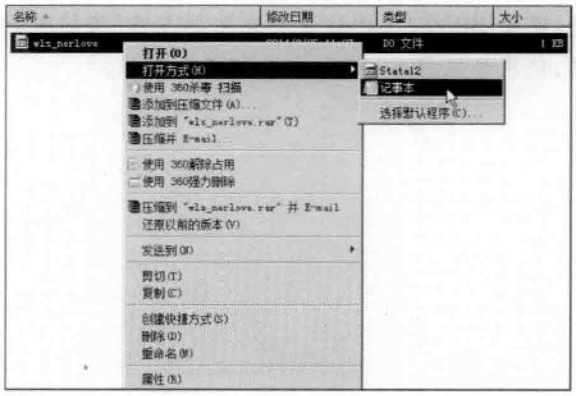



图 7.8 编辑 do 文件

## 习题

7.1[①] 考虑有关啤酒月消费量的线性模型：

$$beer = \beta_1 + \beta_2 inc + \beta_3 price + \beta_4 educ + \beta_5 female + \varepsilon \tag{7.21}$$

其中，$E(\varepsilon \mid inc, price, educ, female) = 0$，$Var(\varepsilon \mid inc, price, educ, female) = \sigma^2 inc^2$。对此模型进行变换，使得变换后的扰动项为同方差。

7.2 房价回归是否存在异方差？数据集 hprice2a. dta 包含美国波士顿 506 个社区的房屋中位数价格的横截面数据(参见第 6 章)[②]。考虑以下特征价格回归：

$$lprice_i = \beta_1 + \beta_2 lnox_i + \beta_3 ldist_i + \beta_4 rooms_i + \beta_5 stratio_i + \varepsilon_i \tag{7.22}$$

① 此题来自 Wooldridge(2009)。

② 此数据集来自 Baum(2006)。

其中，*lprice* 为房价的对数，*lnox* 为空气污染程度的对数，*ldist* 为社区到就业中心距离的对数，*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 $i$ 表示社区 $i$。

（1）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 iid，分别以拟合值 $\hat{y}$ 以及所有解释变量进行检验）。

（2）以5%的置信度，使用怀特检验，检验是否存在异方差。

7.3 恩格尔曲线是否存在异方差？数据集 food. dta 包含有关每周食物开支（*food_exp*）与每周收入（*income*）的 40 个观测值[①]。

（1）将 *food_exp* 与 *income* 的散点图与线性拟合图画在一起。根据此图，是否可能存在异方差？此异方差与收入的关系是怎样的？

（2）将 *food_exp* 对 *income* 进行回归。

（3）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 iid）。

（4）以5%的置信度，使用怀特检验，检验是否存在异方差。

（5）定义食物开支比例 *food_share* 为 *food_exp* 除以 *income*，将 *food_share* 与 *income* 的散点图与线性拟合图画在一起。从图上看，是否还存在异方差？

（6）将 *food_share* 对 *income* 进行回归。

（7）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 iid）。

（8）以5%的置信度，使用怀特检验，检验是否存在异方差。

① 此数据集来自 Adkins and Hill(2011)。

# 8. 自 相 关

## 8.1 自相关的后果

除异方差外，违反球形扰动项的另一情形是扰动项存在自相关。对于扰动项$\{\varepsilon_1,\cdots,\varepsilon_n\}$，如果存在$i\neq j$，使得$E(\varepsilon_i\varepsilon_j\mid \boldsymbol{X})\neq 0$，即协方差矩阵$Var(\boldsymbol{\varepsilon}\mid \boldsymbol{X})$的非主对角线元素不全为0，则存在“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、一致且渐近正态的，因为在证明这些性质时，并未用到“无自相关”的假定。

(2) OLS估计量方差$Var(\hat{\boldsymbol{\beta}}\mid \boldsymbol{X})$的表达式不再是$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，因为$Var(\boldsymbol{\varepsilon}\mid \boldsymbol{X})\neq\sigma^2\boldsymbol{I}$。因此，使用普通标准误的$t$检验、$F$检验失效。

(3) 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在自相关的情况下，OLS不再是BLUE，假设扰动项存在正自相关，即$E(\varepsilon_i\varepsilon_j\mid \boldsymbol{X})>0$，参见图8.1。

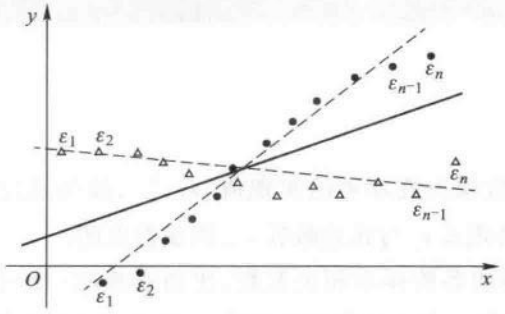



图8.1 自相关的后果

在图8.1中，实线表示真实的总体回归线。如果$\varepsilon_1>0$(图中左边小三角形)，由于扰动项存在正自相关，则$\varepsilon_2>0$的可能性也很大。而如果$\varepsilon_{n-1}<0$(图中右边小三角形)，则$\varepsilon_n<0$的可能性也就很大。此时，样本回归线(虚线)很可能左侧翘起、右侧下垂，使得对回归线斜率的估计过小。

反之，如果$\varepsilon_1<0$(图中左边小圆点)，由于扰动项存在正自相关，故$\varepsilon_2<0$的可能性也很大。而如果$\varepsilon_{n-1}>0$(图中右边小圆点)，则$\varepsilon_n>0$的可能性也就很大。此时，样本回归线(虚线)很可能左侧下垂、右侧翘起，使得对回归线斜率的估计过大。

总之，由于自相关的存在，使得样本回归线上下摆动幅度增大，导致参数估计变得不准确。从信息的角度来看，由于OLS估计忽略了扰动项自相关所包含的信息，故不是最有效率的估计方法。

## 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，某意外事件或新政策的效应需要随时间逐步释放出来。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) 横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 或 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。

(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(参见第 9 章)或季节调整(参见第 13 章)时，则可从理论上判断存在自相关。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，即便在横截面数据中也可能存在。

## 8.3 自相关的检验

### 1. 画图

由于残差 $\{e_t\}_{t=1}^n$ 可大致视为扰动项的实现值 $\{\varepsilon_t\}_{t=1}^n$，故可通过残差来考察扰动项的自相关[①]。一个直观的方法是将残差 $e_t$ 与残差滞后 $e_{t-1}$ 画成散点图。

进一步，可以计算残差的各阶样本相关系数，比如残差的一阶相关系数 $\hat{\rho}_1$，二阶相关系数 $\hat{\rho}_2$ 乃至 $k$ 阶相关系数 $\hat{\rho}_k$，等等。由于这些相关系数 $\hat{\rho}_k$ 是滞后阶数 $k$ 的函数，将 $(k,\hat{\rho}_k)$ 画图，即可得到残差的"自相关图"(correlogram)，参见图 8.6。

### 2. BG 检验(Breusch,1978;Godfrey,1978)

考虑以下多元线性模型：

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t \tag{8.1}$$

假设扰动项 $\varepsilon_t$ 存在一阶自相关，即

$$\varepsilon_t = \gamma \varepsilon_{t-1} + u_t \tag{8.2}$$

① 由于自相关多出现在时间序列中，故将下标以 $t$ 而非 $i$ 来表示。

其中，$u_t$ 为白噪声。方程(8.2)没有常数项，因为 $E(\varepsilon_t)=0$。为了检验是否存在一阶自相关，只要在方程(8.2)中检验 $H_0:\gamma=0$ 即可。更一般地，由于可能存在高阶自相关，可考虑扰动项的 $p$ 阶自回归：

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \cdots + \gamma_p\varepsilon_{t-p} + u_t \tag{8.3}$$

并检验原假设 $H_0:\gamma_1=\cdots=\gamma_p=0$。由于 $\varepsilon_t$ 不可观测，故用 $e_t$ 替代，并引入解释变量 $(x_{t2},\cdots,x_{tK})$，进行如下辅助回归：

$$e_t = \gamma_1 e_{t-1} + \cdots + \gamma_p e_{t-p} + \delta_2 x_{t2} + \cdots + \delta_K x_{tK} + v_t \quad (t=p+1,\cdots,n) \tag{8.4}$$

其中，由于残差 $e_t$ 是解释变量 $(x_{t2},\cdots,x_{tK})$ 的函数，故如果遗漏 $(x_{t2},\cdots,x_{tK})$，可能导致扰动项与 $(e_{t-1},\cdots,e_{t-p})$ 相关，使得估计不一致。在辅助回归(8.4)中，“无自相关”的原假设相当于检验 $H_0:\gamma_1=\cdots=\gamma_p=0$，通常使用 $nR^2$ 形式的 $LM$ 统计量进行检验：

$$LM = (n-p)R^2 \xrightarrow{d} \chi^2(p) \tag{8.5}$$

其中，由于辅助回归(8.4)使用了 $e_{t-p}$，损失 $p$ 个样本观测值[①]，故样本容量仅为 $(n-p)$。如果 $LM$ 统计量超过了 $\chi^2(p)$ 的临界值，则拒绝无自相关的原假设。此检验称为“Breusch-Godfrey 检验”(Breusch，1978；Godfrey，1978，简记 BG)。Davidson and MacKinnon(1993)建议，把残差中因滞后而缺失的项用其期望值 0 来代替[②]，以保持样本容量仍为 $n$，然后使用 $LM$ 统计量 $nR^2 \xrightarrow{d} \chi^2(p)$。Davidson-MacKinnon 方法为 Stata 的默认设置。

### 3. $Q$ 检验

记 $\rho_1,\cdots,\rho_p$ 分别为扰动项的 1 至 $p$ 阶自相关系数。检验自相关的另一思路是，检验各阶自相关系数均为 0，即 $H_0:\rho_1=\cdots=\rho_p=0$。为此，定义残差的各阶样本自相关系数为

$$\hat{\rho}_j \equiv \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2} \quad (j=1,\cdots,p) \tag{8.6}$$

如果 $H_0:\rho_1=\cdots=\rho_p=0$ 成立，则 $\hat{\rho}_j$ 应离 0 不远。事实上，根据大数定律，$\hat{\rho}_j$ 依概率收敛至 0。进一步，根据中心极限定理，$\sqrt{n}\hat{\rho}_j$ 服从渐近正态分布。因此，$\sqrt{n}\hat{\rho}_j$ 的平方和(对 $j$ 求和)为渐近卡方分布，这就是“Box-Pierce $Q$ 统计量”(Box and Pierce，1970)：

$$Q_{BP} \equiv n\sum_{j=1}^{p}\hat{\rho}_j^2 \xrightarrow{d} \chi^2(p) \tag{8.7}$$

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

$$Q_{LB} \equiv n(n+2)\sum_{j=1}^{p}\frac{\hat{\rho}_j^2}{n-j} \xrightarrow{d} \chi^2(p) \tag{8.8}$$

---

① 如果用 $e_1$ 作被解释变量，则需要知道 $e_0,e_{-1},\cdots,e_{-p+1}$，但我们并没有这些数据，故会损失 $p$ 个观测值。

② 即令 $e_0=e_{-1}=\cdots=e_{-p+1}=0$。

这两种 $Q$ 统计量在大样本下是等价的，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 $\min\{\text{floor}(n/2)-2, 40\}$，其中 floor$(n/2)$ 为不超过 $n/2$ 的最大整数，并在[floor$(n/2)-2$]与 40 之间取其小者。

#### 4. DW 检验

DW 检验（Durbin and Watson，1950）是较早出现的自相关检验，现已不常用。它的主要缺点是只能检验一阶自相关，且必须在解释变量满足严格外生性的情况下才成立（BG 检验无此限制）[①]。DW 检验的统计量为

$$\text{DW} \equiv d \equiv \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2}$$

$$\approx 2 - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \equiv 2(1-\hat{\rho}_1) \tag{8.9}$$

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，$\hat{\rho}_1 \approx 1$，存在一阶正自相关；当 $d=4$ 时，$\hat{\rho}_1 \approx -1$，存在一阶负自相关。

DW 检验的另一缺点是，其 $d$ 统计量还依赖于数据矩阵 $\boldsymbol{X}$，无法制成统计表，而必须使用其上限分布 $d_U$ 与下限分布 $d_L$（$d_L < d < d_U$）来间接地检验。但即便如此，也仍然存在“无结论区域”。从 DW 统计量的表达式(8.9)来看，其本质就是残差的一阶自相关系数，故不能指望它提供太多的信息。

DW 检验的具体检验方法，根据 $d_U$ 与 $d_L$ 的临界值，可做如下判断（参见图 8.2）：

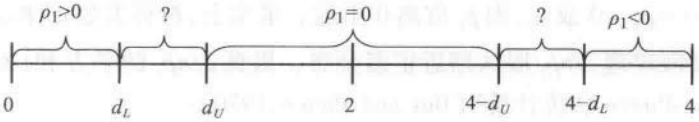



图 8.2 DW 检验的无结论区域

（1）如果 $0 < d \leqslant d_L$，则存在正自相关；

（2）如果 $d_L < d < d_U$，则无法确定；

（3）如果 $d_U \leqslant d \leqslant 4-d_U$，则无自相关；

（4）如果 $4-d_U < d < 4-d_L$，则无法确定；

（5）如果 $4-d_L \leqslant d$，则存在负自相关。

① 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。

## 8.4 自相关的处理

如果经过检验，发现存在自相关，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”

在自相关的情况下，OLS 估计量依然无偏且一致，故仍可使用 OLS 来估计回归系数。然而，为了正确地进行统计推断，需使用“异方差自相关稳健的标准误”（Heteroskedasticity and Autocorrelation Consistent Standard Error，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。

根据第 6 章，异方差稳健的协方差矩阵为夹心估计量：

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X})} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})}\,\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{8.10}$$

类似地，异方差自相关稳健的协方差矩阵也是夹心估计量，但考虑到自相关的存在，“三明治”中间的“菜”$\widehat{\mathrm{Var}(\varepsilon \mid \boldsymbol{X})}$更为复杂[①]。在计算 HAC 标准误时，一方面，如果仅考虑前几阶自相关系数（比如只考虑一阶自相关系数 $\rho_1$）将导致此标准误不一致，因为忽略了高阶自相关。另一方面，如果同时考虑所有各阶相关系数，即 $(\rho_1, \cdots, \rho_{n-1})$，则待估参数多达 $(n-1)$，且随样本容量 $n$ 同步增长，也将导致估计量不一致。另外，对 $\rho_{n-1}$ 的估计将很不准确，因为只有一对数据 $(e_1, e_n)$ 可用于此估计；类似地，对 $\rho_{n-2}$ 的估计也不准确，因为只有两对数据 $(e_1, e_{n-1})$、$(e_2, e_n)$ 可用于估计；以此类推。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 $n$ 而增长。一般建议取 $p = n^{1/4}$ 或 $p = 0.75n^{1/3}$，称为“截断参数”（truncation parameter），即比 $p$ 更高阶的自相关系数将被截断而不考虑。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 标准误是否对于截断参数敏感。

### 2. 准差分法

在自相关的情况下，由于 OLS 未充分利用此信息，故不是最有效率的 BLUE。根据加权最小二乘法的思路，如果能够变换原模型，使得转换后的扰动项变为球形扰动项（不再有自相关），则可得到更有效率的估计。假设原模型为

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_K x_{tK} + \varepsilon_t \quad (t = 1, \cdots, n) \tag{8.11}$$

其中，扰动项 $\varepsilon_t$ 存在自相关，且为一阶自回归形式：

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{8.12}$$

其中，自回归系数 $|\rho| < 1$，且 $u_t$ 为白噪声。将原模型（8.11）滞后一期，然后在方程两边同时乘以 $\rho$ 可得

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{t-1,2} + \cdots + \rho\beta_K x_{t-1,K} + \rho\varepsilon_{t-1} \tag{8.13}$$

---

① 参见陈强（2014，p. 104）。

将原方程(8.11)减去方程(8.13)可得

$$y_t - \rho y_{t-1} = (1-\rho)\beta_1 + \beta_2(x_{t2} - \rho x_{t-1,2}) + \cdots + \beta_K(x_{tK} - \rho x_{t-1,K}) + \underbrace{(\varepsilon_t - \rho\varepsilon_{t-1})}_{u_t} \tag{8.14}$$

其中，$t=2,\cdots,n$，故损失一个样本观测值。显然，在方程(8.14)中，新扰动项$(\varepsilon_t - \rho\varepsilon_{t-1}) = u_t$(白噪声)，故满足球形扰动项的假定。因此，对方程(8.14)进行 OLS 估计，可提高估计效率。这种方法称为“Cochrane-Orcutt 估计法”(Cochrane and Orcutt, 1949，简记 CO)。此法也称为“准差分法”(quasi differences)，因为在做变换时，只是减去滞后值的一部分(比如 $y_t - \rho y_{t-1}$)，而非全部(比如 $y_t - y_{t-1}$)。由于使用准差分法将损失一个样本容量，故仍然不是最有效率的 BLUE。

为了得到 BLUE 估计量，考虑补上损失的第一个方程：

$$y_1 = \beta_1 + \beta_2 x_{12} + \cdots + \beta_K x_{1K} + \varepsilon_1 \tag{8.15}$$

由于$\{u_t\}_{t=1}^n$为白噪声，故 $\varepsilon_1$ 与准差分后的新扰动项 $u_t = (\varepsilon_t - \rho\varepsilon_{t-1})$均不相关。因此，加入第一个方程(8.15)不会导致自相关，但却会导致异方差。这是因为，第一个方程(8.15)的扰动项方差为 $\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_t)$，而准差分方程(8.14)的扰动项方差为 $\sigma_u^2 \equiv \mathrm{Var}(u_t)$，二者并不相等。对方程(8.12)两边求方差，可得 $\sigma_\varepsilon^2$ 与 $\sigma_u^2$ 之间的关系：

$$\mathrm{Var}(\varepsilon_t) = \rho^2 \mathrm{Var}(\varepsilon_{t-1}) + \mathrm{Var}(u_t) \tag{8.16}$$

将上式移项整理可得

$$\sigma_u^2 = (1-\rho^2)\sigma_\varepsilon^2 \tag{8.17}$$

由此可知，$\sigma_u^2$ 是 $\sigma_\varepsilon^2$ 的$(1-\rho^2)$倍；除非 $\rho=0$(无自相关)，否则二者不会相等。因此，只要将第一个方程(8.15)两边同时乘以 $\sqrt{1-\rho^2}$，即可保证同方差：

$$\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}\beta_1 + \beta_2\sqrt{1-\rho^2}x_{12} + \cdots + \beta_K\sqrt{1-\rho^2}x_{1K} + \sqrt{1-\rho^2}\varepsilon_1 \tag{8.18}$$

此时，方程(8.18)的扰动项方差为

$$\mathrm{Var}(\sqrt{1-\rho^2}\varepsilon_1) = (1-\rho^2)\sigma_\varepsilon^2 = \sigma_u^2 \tag{8.19}$$

故这 $n$ 个方程满足同方差与无自相关的假定，为球形扰动项。因此，使用 OLS 估计这些方程，即可得到 BLUE。这种方法称为“Prais-Winsten 估计法”(Prais and Winsten, 1954，简记 PW)。

在某种意义上，无论 CO 法还是 PW 法均不可行(infeasible)，因为它们都假设知道一阶自回归系数 $\rho$。在实践中，必须用数据估计一阶自回归系数 $\hat{\rho}$。Stata 默认的估计方法为使用 OLS 残差进行辅助回归：

$$e_t = \hat{\rho}e_{t-1} + error_t \tag{8.20}$$

另外，也可使用残差的一阶自相关系数来估计 $\hat{\rho}$：

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \tag{8.21}$$

或通过 DW 统计量来估计 $\hat{\rho}$：

$$\hat{\rho} = 1 - \frac{\mathrm{DW}}{2} \tag{8.22}$$

在实践中，常使用迭代法，即首先用 OLS 估计原模型，使用 OLS 残差作辅助回归(8.20)，得到 $\hat{\rho}^{(1)}$(对 $\rho$ 的第一轮估计)，再用 $\hat{\rho}^{(1)}$ 进行 CO 或 PW 估计；然后，使用 CO 或 PW 法的新残差估计 $\hat{\rho}^{(2)}$(对 $\rho$ 的第二轮估计)，再用 $\hat{\rho}^{(2)}$ 进行 CO 或 PW 估计，以此类推，直至收敛(即相邻两轮的 $\rho$ 与系数估计值之差足够小)。

**3. 广义最小二乘法(GLS)**

第 7 章考虑了处理异方差的加权最小二乘法，而上文介绍了处理自相关的准差分法。更一般地，可能同时存在异方差与自相关。此时，应使用“广义最小二乘法”(Generalized Least Square，GLS)，同时处理异方差与自相关。

假设扰动项的协方差矩阵 $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{V}(\boldsymbol{X}) \neq \sigma^2 \boldsymbol{I}_n$，其中 $\boldsymbol{V}(\boldsymbol{X})$ 为对称正定矩阵且已知，但可能依赖于 $\boldsymbol{X}$。GLS 的基本思想是，通过变量转换，使得转换后的模型满足球形扰动项的假定。为了进行这种变换，首先介绍一个命题。

**命题** 对于对称正定矩阵 $\boldsymbol{V}_{n\times n}$，存在非退化矩阵 $\boldsymbol{C}_{n\times n}$，使得 $\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$。

直观来看，在一维情况下，“$\boldsymbol{V}$ 正定”要求 $\boldsymbol{V}$ 为正数，故 $\frac{1}{\boldsymbol{V}}$ 也是正数，可分解为 $\frac{1}{\sqrt{\boldsymbol{V}}} \cdot \frac{1}{\sqrt{\boldsymbol{V}}}$；反之，如果 $\boldsymbol{V}$ 为 0 或负数，则无法进行此分解。推广到多维情形，就是此命题。需要指出的是，此命题中的矩阵 $\boldsymbol{C}$ 并不唯一，但这不影响 GLS 的最终结果(详见下文)。

根据此命题，对于协方差矩阵 $\mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{V}(\boldsymbol{X})$，首先找到非退化矩阵 $\boldsymbol{C}$，使得 $\boldsymbol{V}^{-1} = \boldsymbol{C}'\boldsymbol{C}$。其次，将原回归模型 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 两边同时左乘矩阵 $\boldsymbol{C}$：

$$\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon} \tag{8.23}$$

定义以下变量转换：

$$\widetilde{\boldsymbol{y}} \equiv \boldsymbol{C}\boldsymbol{y}, \quad \widetilde{\boldsymbol{X}} \equiv \boldsymbol{C}\boldsymbol{X}, \quad \widetilde{\boldsymbol{\varepsilon}} \equiv \boldsymbol{C}\boldsymbol{\varepsilon} \tag{8.24}$$

则可将模型写为

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}} \tag{8.25}$$

容易验证，变换后的回归模型仍然满足严格外生性，因为

$$\mathrm{E}(\widetilde{\boldsymbol{\varepsilon}} \mid \widetilde{\boldsymbol{X}}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{C}\boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{C}\mathrm{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{0} \tag{8.26}$$

其中，由于 $\boldsymbol{C}$ 非退化，故 $\mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{C}\boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon} \mid \boldsymbol{X})$。而且，球形扰动项的假定也得到满足，因为

$$\begin{aligned}\mathrm{Var}(\widetilde{\boldsymbol{\varepsilon}} \mid \widetilde{\boldsymbol{X}}) &= \mathrm{E}(\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}' \mid \boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}' \mid \boldsymbol{X}) = \boldsymbol{C}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X})\boldsymbol{C}' = \sigma^2\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{V}^{-1})^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{C}')^{-1}\boldsymbol{C}' = \sigma^2\boldsymbol{I}_n\end{aligned} \tag{8.27}$$

因此，高斯－马尔可夫定理成立。对变换后的方程(8.25)使用 OLS 即得到 GLS 估计量：

$$\hat{\boldsymbol{\beta}}_{GLS} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}} = [(\boldsymbol{CX})'(\boldsymbol{CX})]^{-1}(\boldsymbol{CX})'\boldsymbol{Cy}$$
$$= (\boldsymbol{X}'\underbrace{\boldsymbol{C}'\boldsymbol{C}}_{\boldsymbol{V}^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\boldsymbol{C}'\boldsymbol{C}}_{\boldsymbol{V}^{-1}}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{8.28}$$

从上式可知，虽然 $\boldsymbol{C}$ 不唯一，但 $\hat{\boldsymbol{\beta}}_{GLS}$ 唯一，因为 $\hat{\boldsymbol{\beta}}_{GLS}$ 不依赖于 $\boldsymbol{C}$。由于高斯－马尔可夫定理成立，故 $\hat{\boldsymbol{\beta}}_{GLS}$ 是 BLUE，比 OLS 更有效率。使用 GLS 的前提是，必须知道协方差矩阵 $\boldsymbol{V}$。由于 $\boldsymbol{V}$ 通常未知，故在某种意义上，GLS 是不可行的。在实践中，需通过数据估计 $\hat{\boldsymbol{V}}$，再进行 GLS 估计，称为“可行广义最小二乘法”（Feasible GLS，FGLS）。显然，WLS 与 PW 都是 GLS 的特例，而 FWLS 与可行的 PW 法都是 FGLS 的特例。

何时使用 FGLS 处理自相关？在使用 FGLS 处理自相关时，如果对自相关系数 $\rho$ 的估计比较准确，而且满足严格外生性，则 FGLS 比 OLS 更有效率。但如果不满足严格外生性，而仅满足前定解释变量（同期外生）的假定，则 FGLS 可能不一致，尽管 OLS 依然一致。具体来说，在使用准差分法时，变换后的新扰动项为 $(\varepsilon_t - \rho\varepsilon_{t-1})$，而新解释变量为 $(x_t - \rho x_{t-1})$；故在同期外生的假定下，二者仍可能存在相关性，比如 $\mathrm{Cov}(\varepsilon_t, x_{t-1}) \neq 0$，导致不一致的估计。总之，FGLS 的适用条件比 OLS 更苛刻，不如 OLS 稳健。

#### 4. 修改模型设定

在有些情况下，自相关的深层原因可能是模型设定有误，比如，遗漏了自相关的解释变量；或将动态模型（解释变量中包含被解释变量的滞后值）误设为静态模型，而后者也可视为遗漏了解释变量。

具体来说，假设真实模型为

$$y_t = \alpha + \beta x_t + \rho y_{t-1} + \varepsilon_t \tag{8.29}$$

由于 $y_t$ 是 $y_{t-1}$ 的函数，故 $\{y_t\}$ 存在自相关。假设此模型被错误地设定为

$$y_t = \alpha + \beta x_t + \underbrace{(\rho y_{t-1} + \varepsilon_t)}_{v_t} \tag{8.30}$$

其中，$\rho y_{t-1}$ 被纳入到扰动项 $v_t$ 中，导致扰动项 $\{v_t\}$ 出现自相关，因为 $\{y_{t-1}\}$ 存在自相关。此例说明，对于时间序列存在的自相关，有时可通过引入被解释变量的滞后来消除。总之，对于模型设定误差所导致的自相关，最好从改进模型设定着手解决，而不是机械地使用 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（必须是时间序列或面板数据，才能定义时间变量）。假设时间变量为 `year`，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,且时间变量为 year。

常用的时间序列算子包括滞后(lag)与差分(difference),分别以“L.”与“D.”来表示(可以小写)。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,以下命令

```
. reg y L.x L2.x L3.x L4.x
```

可以简写为

```
. 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 D 2. $x_t = \Delta(\Delta x_t) = \Delta(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$(二阶差分为一阶差分的差分)。

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. $x_t$。有关时间序列算子的更多说明,参见“help tsvarlist”。

**2. 画残差图**

假设在作完回归后,将残差记为 e1,可输入如下命令画残差与其滞后的散点图:

```
scatter e1 L.e1
```

如果想看残差自相关图(即各阶自相关系数),可使用命令

```
ac e1
```

其中,“ac”表示 autocorrelation(自相关)。

**3. BG 检验**

作完 OLS 回归后,可使用如下命令进行 BG 检验:

```
estat bgodfrey,lags(p) nomiss0
```

其中,选择项“lags(p)”用来指定 BG 检验的滞后阶数 $p$,默认“lags(1)”,即 $p = 1$;选择项“nomiss0”表示进行不添加 0 的 BG 检验,默认以 0 代替缺失值,即 Davidson-MacKinnon 的方法。

如何确定滞后阶数 $p$? 一个简单的方法是,看自相关图。具体来说,在使用 Stata 命令 ac 画自相关图时,所有落在 95% 的置信区域(以阴影表示)以外的自相关系数均显著地不等于 0。

确定滞后阶数 $p$ 的另一方法是,设定一个较大的 $p$ 值,作回归

$$e_t = \gamma_1 e_{t-1} + \cdots + \gamma_p e_{t-p} + \delta_2 x_{t2} + \cdots + \delta_K x_{tK} + v_t \quad (t = p+1, \cdots, n) \tag{8.31}$$

然后看最后一个系数 $\gamma_p$ 的显著性;如果 $\gamma_p$ 不显著,则考虑滞后 $(p-1)$ 期,以此类推,直至显著为止。

#### 4. Q 检验

假设将 OLS 残差记为 e1,则可使用如下命令进行 $Q$ 检验:

```
wntestq e1,lags(p)
```

其中,"wntestq"指 white noise test $Q$,因为白噪声没有自相关。选择项"lags(p)"用来指定滞后阶数,默认滞后阶数为 min{floor($n/2$) -2,40}。

进行 $Q$ 检验的另一命令是

```
corrgram e1,lags(p)
```

其中,"corrgram"表示 correlogram,即画自相关图。选择项"lags(p)"用来指定滞后阶数,而默认滞后阶数也是 min{floor($n/2$) -2,40}。

#### 5. DW 检验

作完 OLS 回归后可使用命令"estat dwatson"显示 DW 统计量。由于 DW 检验的局限性,Stata 并不提供其临界值。

#### 6. HAC 稳健标准误

在 Stata 中进行 OLS 估计,但提供 Newey-West 标准误,可输入命令

```
newey y x1 x2 x3,lag(p)
```

其中,必选项"lag(p)"用来指定截断参数 $p$,即用于计算 HAC 标准误的最高滞后阶数。

#### 7. 处理一阶自相关的 FGLS

在 Stata 中使用准差分法处理自相关,可使用命令

```
prais y x1 x2 x3,corc
```

其中,选择项"corc "表示使用 CO 估计法,默认使用 PW 估计法。

下面以 Hildreth and Lu(1960)对冰淇淋需求函数的经典研究作为实例。数据集 icecream. dta 包含下列变量的 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)。

首先,打开数据集,并将其设为时间序列数据。

```
. use icecream.dta,clear
. tsset time
```

其次,为了看冰淇淋的消费量与气温的时间趋势图,输入命令:

```
. twoway connect consumption time,msymbol(circle) yaxis(1) || connect temp time,msymbol(triangle) yaxis(2)
```

其中,"connect"表示将观测点用线连接起来,选择项"msymbol(circle)"与"msymbol(triangle)"分别表示点的"图标"(marker symbol)分别为圆圈与三角形;选择项"yaxis(1)"与"yaxis(2)"指定使用不同的纵坐标,因为冰淇淋消费量与气温的取值范围很不相同,结果参见图 8.3。

从图 8.3 可知,冰淇淋消费量与温度明显地正相关。考虑以下线性回归模型:

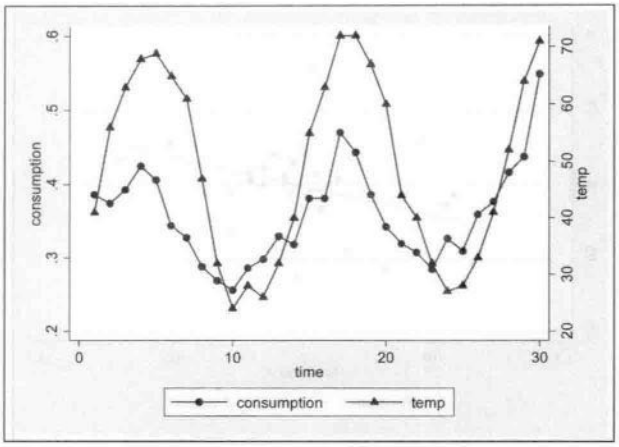



图 8.3 冰淇淋消费与气温的时间趋势

$$consumption_t = \beta_0 + \beta_1 temp_t + \beta_2 price_t + \beta_3 income_t + \varepsilon_t \tag{8.32}$$

首先进行 OLS 回归：

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

| Source | SS | df | MS | Number of obs = | 30 |
|---|---|---|---|---|---|
| | | | | F( 3, 26) = | 22.17 |
| Model | .090250523 | 3 | .030083508 | Prob > F = | 0.0000 |
| Residual | .035272835 | 26 | .001356647 | R-squared = | 0.7190 |
| | | | | Adj R-squared = | 0.6866 |
| Total | .125523358 | 29 | .004328392 | Root MSE = | .03683 |

| consumption | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| temp | .0034584 | .0004455 | 7.76 | 0.000 | .0025426 | .0043743 |
| price | -1.044413 | .834357 | -1.25 | 0.222 | -2.759458 | .6706322 |
| income | .0033078 | .0011714 | 2.82 | 0.009 | .0008999 | .0057156 |
| _cons | .1973149 | .2702161 | 0.73 | 0.472 | -.3581223 | .752752 |

上表显示，气温（*temp*）与收入（*income*）均在 1% 的水平上显著为正，表示气温越高、收入越高，则冰淇淋的消费量越大；价格（*price*）的系数为负，表明价格越高，则消费量越低，但并不显著（$p$ 值为 0.222）。由于这是时间序列，故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（l.e1），然后画残差与残差滞后的散点图。

```
. predict e1,r
. twoway scatter e1 l.e1 || lfit e1 l.e1
```

其中，“lfit”表示 linear fit（线性拟合），即画出 e1 与 l.e1 的拟合回归线，结果参见图 8.4。

图 8.4 显示，扰动项很可能存在一阶正自相关。作为对比，下面画残差与其二阶滞后的散点图，结果参见图 8.5。

```
. twoway scatter e1 l2.e1 || lfit e1 l2.e1
```

图 8.5 显示，残差似乎不存在二阶自相关（散点分布没有规律，且样本回归线的斜率接近于 0）。为了看各阶自相关系数及其显著性，下面画残差的自相关图，结果参见图 8.6。

```
. ac e1
```

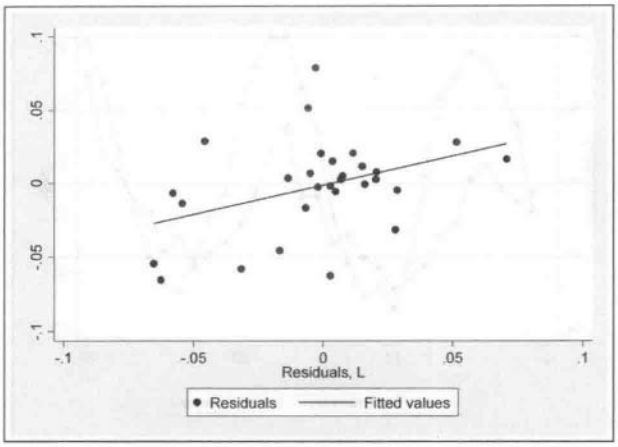



图 8.4 残差与残差滞后的散点图

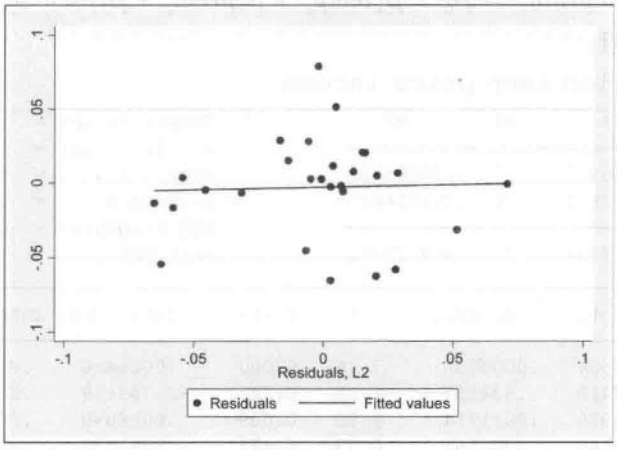



图 8.5 残差与二阶残差滞后的散点图

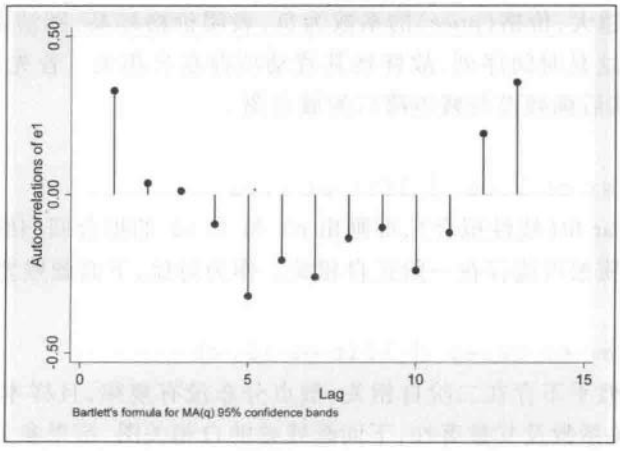



图 8.6 自相关图

图 8.6 的横轴为滞后阶数，纵轴为残差的自相关系数，而阴影部分为置信度为 95% 的置信区间(区域)。图 8.6 显示，各阶自相关系数的取值均在 95% 的置信区间之内，故可接受各阶自相关系数为 0 的原假设。然而，一阶自相关系数已很接近置信区间的边界，故仍怀疑存在一阶自相关，而更高阶自相关则可大致忽略。

下面进行正式的 BG 检验，考察是否存在一阶自相关：

```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation

| lags(p) | chi2 | df | Prob > chi2 |
|---|---|---|---|
| 1 | 4.237 | 1 | 0.0396 |

H0: no serial correlation

上表显示，BG 检验的 $p$ 值为 0.039 6，故可在 5% 的显著性水平上拒绝“无自相关”的原假设，而认为存在自相关。如果不以 0 取代缺失值，可输入命令

```
. estat bgodfrey,nomiss0
```

Breusch-Godfrey LM test for autocorrelation

| lags(p) | chi2 | df | Prob > chi2 |
|---|---|---|---|
| 1 | 4.704 | 1 | 0.0301 |

H0: no serial correlation

结果依然可在 5% 水平上拒绝“无自相关”的原假设。下面进行 $Q$ 检验。

```
. wntestq e1
```

Portmanteau test for white noise

```
Portmanteau (Q) statistic =    26.1974
Prob > chi2(13)           =     0.0160
```

其中，“Prob > chi2(13) = 0.016”表明默认的滞后阶数为 13 阶，且可在 5% 水平上拒绝“无自相关”的原假设。下面使用命令 corrgram 进行 $Q$ 检验。

```
. corrgram e1
```

| LAG | AC | PAC | Q | Prob>Q | [Autocorrelation] | [Partial Autocor] |
|---|---|---|---|---|---|---|
| 1 | 0.3298 | 0.3969 | 3.6 | 0.0578 | | |
| 2 | 0.0362 | -0.1681 | 3.645 | 0.1616 | | |
| 3 | 0.0111 | 0.0767 | 3.6494 | 0.3019 | | |
| 4 | -0.0934 | -0.1483 | 3.9715 | 0.4099 | | |
| 5 | -0.3186 | -0.3565 | 7.8703 | 0.1635 | | |
| 6 | -0.2058 | 0.0011 | 9.5645 | 0.1442 | | |
| 7 | -0.2582 | -0.4237 | 12.346 | 0.0897 | | |
| 8 | -0.1373 | -0.0721 | 13.169 | 0.1062 | | |
| 9 | -0.1035 | -0.3300 | 13.658 | 0.1350 | | |
| 10 | -0.2378 | -0.8928 | 16.372 | 0.0895 | | |
| 11 | -0.1193 | -0.5017 | 17.091 | 0.1052 | | |
| 12 | 0.1923 | -0.4590 | 19.064 | 0.0870 | | |
| 13 | 0.3554 | 0.0493 | 26.197 | 0.0160 | | |

上表汇报了从1-13阶的自相关系数(AC),$Q$统计量(Q)及其相应$p$值(Prob>Q)。其中,第13阶$Q$统计量及其$p$值与命令wntestq的结果完全相同。使用命令corrgram的好处在于,它同时计算了各阶$Q$统计量。

作为最后一个自相关检验,下面计算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较远而靠近0,故可大致判断存在正自相关。

由于扰动项存在自相关,故OLS估计所提供的普通标准误不准确,应使用异方差自相关稳健的HAC标准误。由于$n^{1/4}=30^{1/4}\approx 2.34$,故取Newey-West估计量的滞后阶数为$p=3$:

```
. newey consumption temp price income,lag(3)
```

```
Regression with Newey-West standard errors                Number of obs   =        30
maximum lag: 3                                            F(  3,    26)   =     27.63
                                                          Prob > F        =    0.0000
------------------------------------------------------------------------------
             |             Newey-West
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0034584   .0004002     8.64   0.000     .0026357    .0042811
       price |  -1.044413   .9772494    -1.07   0.295    -3.053178    .9643518
      income |   .0033078   .0013278     2.49   0.019     .0005783    .0060372
       _cons |   .1973149   .3378109     0.58   0.564    -.4970655    .8916952
------------------------------------------------------------------------------
```

Newey-West标准误与OLS标准误相差无几(但略大)。为考察Newey-West标准误是否对于截断参数敏感,下面将滞后阶数增大一倍,变为6,再重新估计。

```
. newey consumption temp price income,lag(6)
```

```
Regression with Newey-West standard errors                Number of obs   =        30
maximum lag: 6                                            F(  3,    26)   =     52.97
                                                          Prob > F        =    0.0000
------------------------------------------------------------------------------
             |             Newey-West
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0034584   .0003504     9.87   0.000     .0027382    .0041787
       price |  -1.044413   .9821798    -1.06   0.297    -3.063313    .9744864
      income |   .0033078    .00132     2.51   0.019     .0005945     .006021
       _cons |   .1973149   .3299533     0.60   0.555    -.4809139    .8755437
------------------------------------------------------------------------------
```

无论截断参数为3还是6,Newey-West标准误变化不大。由于存在自相关,OLS不再是BLUE,故可考虑使用FGLS,对模型进行更有效率的估计。首先使用CO估计法:

```
. prais consumption temp price income,corc
```

```
Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.4006
Iteration 2:  rho = 0.4008
Iteration 3:  rho = 0.4009
Iteration 4:  rho = 0.4009
Iteration 5:  rho = 0.4009
Iteration 6:  rho = 0.4009
Iteration 7:  rho = 0.4009

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  3,    25) =   15.40
       Model |  .047040596     3  .015680199           Prob > F      =  0.0000
    Residual |  .025451894    25  .001018076           R-squared     =  0.6489
-------------+------------------------------           Adj R-squared =  0.6068
       Total |  .072492491    28  .002589018           Root MSE      =  .03191

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0035584   .0005547     6.42   0.000      .002416    .0047008
       price |  -.8923963   .8108501    -1.10   0.282    -2.562373    .7775807
      income |   .0032027   .0015461     2.07   0.049      .0000186   .0063869
       _cons |   .1571479   .2896292     0.54   0.592    -.4393546    .7536504
-------------+----------------------------------------------------------------
         rho |   .4009256
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.021169
Durbin-Watson statistic (transformed) 1.548837
```

使用 CO 估计法得到的系数估计值与 OLS 比较接近，但样本容量降为 29（损失一个样本观测值）。上表最后一行显示，经过模型转换后 DW 值改进为 1.55。然后使用 PW 估计法：

```
. prais consumption temp price income,nolog
```

其中，选择项“nolog”表示不显示迭代过程。

```
Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   14.35
       Model |  .04494596     3  .014981987           Prob > F      =  0.0000
    Residual |  .027154354    26  .001044398           R-squared     =  0.6234
-------------+------------------------------           Adj R-squared =  0.5799
       Total |  .072100315    29  .002486218           Root MSE      =  .03232

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        temp |   .0029541   .0007109     4.16   0.000      .0014929   .0044152
       price |  -1.048854    .759751    -1.38   0.179    -2.610545    .5128361
      income |  -.0008022   .0020458    -0.39   0.698    -.0050074    .0034029
       _cons |   .5870049   .2952699     1.99   0.057    -.0199311    1.193941
-------------+----------------------------------------------------------------
         rho |   .8002264
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.021169
Durbin-Watson statistic (transformed) 1.846795
```

虽然使用 PW 估计法使 DW 统计量进一步改进为 1.85，但收入（*income*）的系数估计值却变为负数（-0.000 8）。尽管它只是绝对值很小的负数，且在统计上不显著，但由于 PW 估计法使得收入的系数估计值与理论预期相反，似乎 PW 估计法反而不如 OLS 稳健。

自相关的存在可能是由于模型设定不正确。为此，考虑在解释变量中加入气温（*temp*）的滞后值，然后进行 OLS 回归：

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

| Source | SS | df | MS | Number of obs = | 29 |
|---|---|---|---|---|---|
| | | | | F( 4, 24) = | 28.98 |
| Model | .103387183 | 4 | .025846796 | Prob > F = | 0.0000 |
| Residual | .021406049 | 24 | .000891919 | R-squared = | 0.8285 |
| | | | | Adj R-squared = | 0.7999 |
| Total | .124793232 | 28 | .004456901 | Root MSE = | .02987 |

| consumption | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| temp | | | | | | |
| --. | .0053321 | .0006704 | 7.95 | 0.000 | .0039484 | .0067158 |
| L1. | -.0022039 | .0007307 | -3.02 | 0.006 | -.0037119 | -.0006959 |
| price | -.8383021 | .6880205 | -1.22 | 0.235 | -2.258307 | .5817025 |
| income | .0028673 | .0010533 | 2.72 | 0.012 | .0006934 | .0050413 |
| _cons | .1894822 | .2323169 | 0.82 | 0.423 | -.2899963 | .6689607 |

上表显示，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（系数为 -0.002 2）；而当期气温仍然显著为正（系数为 0.005 3）。这可能意味着，当气温上升时，对冰淇淋的需求上升，但不会在当月全部消费完，而增加冰箱中的冰淇淋库存，导致下期对冰淇淋的开支下降。

使用 BG 检验判断重新设定的模型是否存在自相关：

```
. 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，故可放心接受“无自相关”的原假设。下面计算 DW 统计量。

```
. estat dwatson
```

```
Durbin-Watson d-statistic(  5,    29) =  1.582166
```

DW 值也改进为 1.58。因此，通过修改模型设定，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，在研究报告中同时列出各种模型的结果，以此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），从而给读者自己判断的机会。

## 习题

8.1 PW 估计法比 CO 估计法更有效率吗？为什么？

8.2 假设扰动项存在二阶自相关，即 $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t$，其中 $u_t$ 为白噪声。此时，还可以使用 CO 估计法吗？若可以，如何进行？

8.3[①] 使用数据集 gasoline. dta 估计美国 1953—2004 年的汽油需求函数。考虑如下回归：

$$lgasq_t = \beta_1 + \beta_2 lincome_t + \beta_3 lgasp_t + \beta_4 lpnc_t + \beta_5 lpuc_t + \varepsilon_t \tag{8.33}$$

其中，被解释变量 *lgasq* 为人均汽油消费量的对数，解释变量 *lincome* 为人均收入的对数，*lgasp* 为汽油价格指数的对数，*lpnc* 为新车价格指数的对数，*lpuc* 为二手车价格指数的对数。

(1) 使用 OLS 估计方程(8.33)。评论各变量系数的符号、显著性与经济意义。

(2) 计算残差，并记为 e1。将残差与其一阶滞后的散点图与线性拟合图画在一起。根据此图，是否可能存在自相关？

(3) 画残差的自相关图。

(4) 使用 BG 检验，检验扰动项是否存在自相关。

(5) 使用 $Q$ 检验，检验扰动项是否存在自相关。

(6) 计算 DW 统计量。

(7) 使用 HAC 标准误进行回归，将截断参数设为 $n^{1/4}$。

(8) 使用迭代式 CO 估计法进行 FGLS 估计。

(9) 使用迭代式 PW 估计法进行 FGLS 估计。

(10) 考虑到消费可能存在惯性，将被解释变量 *lgasq* 的一阶滞后作为解释变量，加入回归方程(8.33)。此滞后项是否显著？

(11) 对于修改后的模型，再次进行 BG 检验与 $Q$ 检验，是否还存在自相关？

① 此题改编自 Greene(2011)。

*If you torture the data long enough, Nature will confess.* ——Ronald Coase

# 9. 模型设定与数据问题

如果模型设定(model specification)不当,比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,比如多重共线性、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(true model)为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \tag{9.1}$$

其中,解释变量 $x_1, x_2$ 与扰动项 $\varepsilon$ 不相关,并省略了表示个体的下标 $i$(比如以 $y$ 表示 $y_i$)。而实际估计的模型(estimated model)为

$$y = \alpha + \beta x_1 + u \tag{9.2}$$

对比以上两个方程可知,遗漏变量(omitted variable)$x_2$ 被归入新扰动项 $u = \gamma x_2 + \varepsilon$ 中了。遗漏变量是否一定导致不一致的估计?为此,考虑以下两种情形:

(1) 遗漏变量 $x_2$ 与解释变量 $x_1$ 不相关,即 $\mathrm{Cov}(x_1, x_2) = 0$。在这种情况下,扰动项 $u = \gamma x_2 + \varepsilon$ 与解释变量 $x_1$ 不相关,因为

$$\mathrm{Cov}(x_1, u) = \mathrm{Cov}(x_1, \gamma x_2 + \varepsilon) = \gamma \mathrm{Cov}(x_1, x_2) + \mathrm{Cov}(x_1, \varepsilon) = 0 + 0 = 0 \tag{9.3}$$

此时,虽然存在遗漏变量,但 OLS 依然可一致地估计回归系数。然而,由于遗漏变量 $x_2$ 被归入扰动项 $u$ 中,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 $x_2$ 与解释变量 $x_1$ 相关,即 $\mathrm{Cov}(x_1, x_2) \neq 0$。在这种情况下,根据大样本理论,OLS 估计不一致,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,个人的先天能力因无法观测而遗漏,但能力与教育年限可能正相关。

总之,存在遗漏变量本身并不要紧,你甚至可以说"只对 $\mathrm{E}(y \mid x_1)$ 感兴趣,故不把 $x_2$ 放入解释变量中"。问题的关键是,遗漏变量不能与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) 随机实验与自然实验;

(iii) 工具变量法(第 10 章);

(iv) 使用面板数据(第 12 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,则需从理论上说明,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。

解决遗漏变量偏差的第(ii)种方法为随机实验或自然实验。物理学常使用"控制实验"(controlled experiment)的方法来研究 $x$ 对 $y$ 的因果关系,即给定影响 $y$ 的所有其他因素,单独让 $x$ 变化,然后观察 $y$ 如何变化。但这种控制实验的方法在其他学科未必可行,比如,医学上对新药 $x$ 疗效的实验。由于参加实验者的体质与生活方式不同,不可能完全控制所有其他因素(即使用老鼠做实验,老鼠之间仍然有差异),故无法进行严格的控制实验。

为此,当代统计学之父费舍尔(Ronald Fischer)提出了随机(控制)实验(randomized controlled experiment)的概念。考虑以下回归模型:

$$y = \alpha + \beta x + \varepsilon \tag{9.4}$$

其中,$x$ 是完全随机地决定的(比如,通过抛硬币或计算机随机数),故 $x$ 与世界上的任何其他变量都相互独立。由于 $x$ 独立于 $\varepsilon$,故 $\mathrm{Cov}(x,\varepsilon)=0$,因此无论遗漏了多少解释变量,OLS 都是一致的。

如果 $x$ 是取值为 0 或 1 的虚拟变量,则可将样本数据分为两组。通常称"$x=1$"的那一组为"实验组"或"处理组"(treatment group),比如医学实验中吃药的那组;而称"$x=0$"的那一组为"控制组"(control group)或"对照组",比如医学实验中不吃药的那组[①]。随机实验的核心思想是,将实验人群(或个体)随机分为两组,即实验组与控制组,则这两组除了 $x$ 不同外,在所有其他方面都没有系统性差别。

**例** 在农学中将地块随机分成三组(因为很难找到土壤条件完全一样的地块),分别给予不同的施肥量,然后考察施肥的效果。

**例** 班级规模是否影响学习成绩?班级规模过大(师生比过低)可能影响任课老师对每位学生的关注,从而影响学习效果。由于遗漏变量的存在,观测数据很难回答此问题。比如,规模较小的班级可能位于好学区,师资好,家庭也富有。为此,美国田纳西州进行了为期四年的随机实验(称为 Project STAR,即 Student-Teacher Achievement Ratio),将幼儿园至小学三年级的学生随机分为三组。第一组为普通班,每班 22 ~ 25 名学生;第二组为小班,每班 13 ~ 17 名学生;第三组也为小班,但配备一名教学助理(teacher's aide)。教师也随机分到这三类班级。实验结果发现,班级规模对学习成绩的影响在统计上显著,但在经济上并不显著(即此效应本身比较小,普通班与小班的成绩差距类似于男生与女生的成绩差距),详见 Stock and Watson(2012)第 13 章。

随机实验虽然说服力强,但通常成本较高。另外一种实验方法为"自然实验"(natural exper-

① 为避免心理作用的干扰,即使是"不吃药"的控制组,通常也服用"安慰药"(placebo),即没有任何医疗作用的替代品;并且不让实验参与者知道自己分在哪一组,到底吃的是真药还是假药。

iment)或"准实验"(quasi experiment),即由于某些并非为了实验目的而发生的外部突发事件,使得当事人仿佛被随机分在了实验组或控制组。

**例** 最低工资对就业的影响。提高法定最低工资(minimum wage)在多大程度上会影响对低技能工人的需求?为避免内生性(即解释变量与扰动项相关),Card and Krueger(1994)考虑了一个自然实验。在1992年,美国新泽西州通过法律将最低工资从每小时4.25美元提高到5.05美元,但在相邻的宾夕法尼亚州最低工资却保持不变。因此,这两个州的雇主仿佛被随机分配到实验组(新泽西州)与控制组(宾夕法尼亚州)。Card and Kruger(1994)收集了两个州的快餐店在实施新法前后雇佣人数的数据,结果发现,提高最低工资对低技能工人的就业几乎没有影响。

**例**(一个失败的例子) 京杭大运河流经省份的人均GDP平均而言高于其他省份。这是否可以归功于京杭大运河对区域经济增长的促进作用?问题在于,当隋炀帝确定京杭大运河的位置时,他是在地图上随机画了一条线吗?

解决遗漏变量偏差的第(iii)种方法为工具变量法,而第(iv)种方法为使用面板数据,将分别在第10章与第12章介绍。

总之,由于影响被解释变量的因素往往很多,而局限于数据的可得性(availability),故在任何实证研究中几乎总存在遗漏变量。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。假设真实的模型为

$$y = \alpha + \beta x_1 + \varepsilon \tag{9.5}$$

其中,$\mathrm{Cov}(x_1, \varepsilon) = 0$。而实际估计的模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + (\varepsilon - \gamma x_2) \tag{9.6}$$

其中,加入了与被解释变量$y$无关的解释变量$x_2$。由于真实参数$\gamma = 0$($x_2$对$y$无影响),故可将模型写为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \tag{9.7}$$

其中,扰动项仍是原来的$\varepsilon$。在方程(9.7)中,由于$x_2$与$y$无关,根据"无关变量"的定义,$x_2$也与$y$的扰动项$\varepsilon$无关,即$\mathrm{Cov}(x_2, \varepsilon) = 0$。因此,扰动项$\varepsilon$与所有解释变量均无关,故OLS仍然一致,即$\operatorname{plim}_{n\to\infty} \hat{\beta} = \beta$,$\operatorname{plim}_{n\to\infty} \hat{\gamma} = \gamma = 0$。

然而,引入无关变量后,由于受到无关变量的干扰,估计量$\hat{\beta}$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。比如,先将被解释变量 $y$ 对关键解释变量 $x$ 回归,然后再加入其他控制变量 $z$。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,导致系数估计量不一致,$t$ 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(可依次剔除最不显著,即 $p$ 值最大的变量)。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,一般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可以提高模型的解释力(比如增大拟合优度 $R^2$),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 $\overline{R}^2$:选择解释变量的个数 $K$ 以最大化 $\overline{R}^2$。

(2) "赤池信息准则"(Akaike Information Criterion,AIC):选择解释变量的个数 $K$,使得以下目标函数最小化:

$$\min_K \text{AIC} \equiv \ln\left(\frac{\text{SSR}}{n}\right) + \frac{2}{n}K \tag{9.8}$$

其中,SSR 为残差平方和 $\sum_{i=1}^{n} e_i^2$。上式右边的第一项为对模型拟合度的奖励(减少残差平方和 SSR),而第二项为对解释变量过多的惩罚(为解释变量个数 $K$ 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,BIC)或"施瓦茨信息准则"(Schwarz Information Criterion,SIC 或 SBIC):选择解释变量的个数 $K$,使得以下目标函数最小化:

$$\min_K \text{BIC} \equiv \ln\left(\frac{\text{SSR}}{n}\right) + \frac{\ln n}{n}K \tag{9.9}$$

BIC 准则与 AIC 准则仅第二项有差别。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。

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(AR($p$),详见第 13 章),

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t \tag{9.10}$$

其中,滞后阶数 $p$ 可通过信息准则来确定。可以证明,根据 BIC 准则计算的 $\hat{p}$ 是真实滞后阶数 $p$

的一致估计;但根据 AIC 计算的 $\hat{p}$ 却不一致,即使在大样本中也可能高估 $p$[①]。尽管如此,由于现实样本通常有限,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。

在 Stata 中作完回归后,计算信息准则的命令为

```
estat ic
```

其中,"ic"表示 information criterion(信息准则)。

(4)"由大到小的序贯 $t$ 规则"(general-to-specific sequential $t$ rule)。这种方法常用于时间序列模型,比如 AR($p$)。首先,设一个最大滞后期 $p_{max}$,令 $\hat{p}=p_{max}$ 进行估计,并对最后一阶系数的显著性进行 $t$ 检验。如果接受该系数为 0,则令 $\hat{p}=p_{max}-1$,重新进行估计,再对(新的)最后一阶系数的显著性进行 $t$ 检验,如果显著,则停止;否则,令 $\hat{p}=p_{max}-2$;以此类推。

下面以数据集 icecream. dta 为例(参见第 8 章),考虑应该引入气温(*temp*)的几阶滞后项。首先,使用信息准则。

```
. 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
```

| Model | Obs | ll(null) | ll(model) | df | AIC | BIC |
|---|---|---|---|---|---|---|
| . | 29 | 37.85248 | 63.41576 | 5 | -116.8315 | -109.995 |

从以上结果可知,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,仅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。

① 证明参见 Stock and Watson(2012,p. 623)或陈强(2014,p. 133)。

其次，使用序贯 $t$ 规则，并假设 $p_{max}=2$，估计以下模型：

```
. reg consumption temp L.temp L2.temp price income
```

| Source | SS | df | MS | Number of obs = | 28 |
|---|---|---|---|---|---|
| | | | | F( 5, 22) = | 21.92 |
| Model | .103722201 | 5 | .02074444 | Prob > F = | 0.0000 |
| Residual | .020822754 | 22 | .000946489 | R-squared = | 0.8328 |
| | | | | Adj R-squared = | 0.7948 |
| Total | .124544954 | 27 | .004612776 | Root MSE = | .03077 |

| consumption | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| temp | | | | | | |
| --. | .0047858 | .0013502 | 3.54 | 0.002 | .0019856 | .007586 |
| L1. | -.0010836 | .0022905 | -0.47 | 0.641 | -.0058338 | .0036666 |
| L2. | -.0008022 | .0013414 | -0.60 | 0.556 | -.0035841 | .0019797 |
| price | -.7326035 | .7214324 | -1.02 | 0.321 | -2.228763 | .7635558 |
| income | .0026704 | .0011308 | 2.36 | 0.027 | .0003252 | .0050156 |
| _cons | .1883478 | .23949 | 0.79 | 0.440 | -.3083241 | .6850196 |

从上表可知，L2.temp 的系数高度不显著（$p$ 值为 0.556）。因此，令 $\hat{p}=p_{max}-1=1$（即去掉 L2.temp），重新进行估计。

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

| Source | SS | df | MS | Number of obs = | 29 |
|---|---|---|---|---|---|
| | | | | F( 4, 24) = | 28.98 |
| Model | .103387183 | 4 | .025846796 | Prob > F = | 0.0000 |
| Residual | .021406049 | 24 | .000891919 | R-squared = | 0.8285 |
| | | | | Adj R-squared = | 0.7999 |
| Total | .124793232 | 28 | .004456901 | Root MSE = | .02987 |

| consumption | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| temp | | | | | | |
| --. | .0053321 | .0006704 | 7.95 | 0.000 | .0039484 | .0067158 |
| L1. | -.0022039 | .0007307 | -3.02 | 0.006 | -.0037119 | -.0006959 |
| price | -.8383021 | .6880205 | -1.22 | 0.235 | -2.258307 | .5817025 |
| income | .0028673 | .0010533 | 2.72 | 0.012 | .0006934 | .0050413 |
| _cons | .1894822 | .2323169 | 0.82 | 0.423 | -.2899963 | .6689607 |

从上表可知，L.temp 的系数在 1% 水平上显著（$p$ 值为 0.006），故最终选择 $\hat{p}=1$。此结果与信息准则的结果相同。

## 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，多元线性回归只能看作是非线性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？如果存在非线性项，但被遗漏了，也会导致遗漏变量偏差，这是模型设定误差（specification error）的一种形式。比如，假设真实模型为

$$y = \alpha + \beta x + (\gamma x^2 + \varepsilon) \tag{9.11}$$

其中，$\mathrm{Cov}(x,\varepsilon)=0$，而平方项 $\gamma x^2$ 被遗漏。容易证明，上式的解释变量与扰动项相关：

$$\mathrm{Cov}(x,\gamma x^2+\varepsilon) = \gamma\mathrm{Cov}(x,x^2) + \mathrm{Cov}(x,\varepsilon) = \gamma\mathrm{Cov}(x,x^2) \neq 0 \tag{9.12}$$

因此，遗漏高次项也会导致遗漏变量偏差。“Ramsey's RESET 检验”(Regression Equation Specification Error Test)(Ramsey,1969)的基本思想是，如果怀疑非线性项被遗漏，那么就把非线性项引入方程，并检验其系数是否显著。不失一般性，假设线性回归模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \tag{9.13}$$

记此回归的拟合值为

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_1 + \hat{\gamma} x_2 \tag{9.14}$$

既然 $\hat{y}$ 是解释变量的线性组合，$\hat{y}^2$ 就包含了解释变量二次项(含平方项与交叉项)的信息，$\hat{y}^3$ 就包含了解释变量三次项的信息，以此类推。考虑以下辅助回归：

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \delta_4 \hat{y}^4 + \varepsilon \tag{9.15}$$

然后对 $H_0:\delta_2=\delta_3=\delta_4=0$ 作 $F$ 检验，即检验拟合值 $\hat{y}$ 的高次项系数是否联合为 0。如果拒绝 $H_0$，说明模型中应有高次项；反之，如果接受 $H_0$，则可使用线性模型。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。

当然，也可以直接将解释变量 $x_1$ 与 $x_2$ 的高次项放入辅助回归中，比如

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 x_1^2 + \delta_3 x_2^2 + \delta_4 x_1 x_2 + \varepsilon \tag{9.16}$$

然后检验 $H_0:\delta_2=\delta_3=\delta_4=0$。关于如何确定回归方程的函数形式，最好从经济理论出发。在缺乏理论指导的情况下，可先从线性模型出发，然后进行 RESET 检验，看是否应加入非线性项。

在 Stata 中作完回归，进行 RESET 检验的命令为

```
estat ovtest,rhs
```

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。选择项“rhs”表示使用解释变量的幂为非线性项，即方程(9.16)；默认使用 $\hat{y}^2,\hat{y}^3,\hat{y}^4$ 为非线性项，即方程(9.15)。

下面以数据集 grilic.dta 为例进行演示。首先，打开数据集，并进行线性 OLS 回归。

```
. use grilic.dta,clear
. qui reg lnw s expr tenure smsa rns
```

然后，使用拟合值的高次项进行 RESET 检验。

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of lnw
       Ho:  model has no omitted variables
                 F(3, 749) =      1.51
                  Prob > F =      0.2114
```

上表显示，$p$ 值为 0.211 4，可接受原假设，未发现遗漏高次项。下面，直接使用解释变量的高次项进行 RESET 检验。

```
. estat ovtest,rhs

Ramsey RESET test using powers of the independent variables
      Ho:  model has no omitted variables
                 F(9, 743) =      2.03
                  Prob > F =      0.0336
```

上表显示,可在5%水平上拒绝原假设,即认为遗漏了高阶非线性项。根据劳动经济学的知识,工资对数与工龄(*expr*)的关系可能存在非线性。为此,引入工龄的平方项,记为 *expr2*,再进行回归。

```
. gen expr2 = expr^2
. reg lnw s expr expr2 tenure smsa rns
```

| Source | SS | df | MS | Number of obs = 758 |
|---|---|---|---|---|
| | | | | F( 6, 751) = 69.65 |
| Model | 49.7985818 | 6 | 8.29976364 | Prob > F = 0.0000 |
| Residual | 89.487568 | 751 | .11915788 | R-squared = 0.3575 |
| | | | | Adj R-squared = 0.3524 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = .34519 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .1029839 | .0058299 | 17.66 | 0.000 | .0915391 | .1144287 |
| expr | .0018351 | .0157708 | 0.12 | 0.907 | -.0291249 | .0327951 |
| expr2 | .0051819 | .0020645 | 2.51 | 0.012 | .001129 | .0092348 |
| tenure | .037085 | .0077374 | 4.79 | 0.000 | .0218954 | .0522746 |
| smsa | .1419045 | .0279979 | 5.07 | 0.000 | .086941 | .1968679 |
| rns | -.0843448 | .0286965 | -2.94 | 0.003 | -.1406797 | -.0280098 |
| _cons | 4.119327 | .0850277 | 48.45 | 0.000 | 3.952407 | 4.286247 |

上表显示,工龄平方(*expr2*)在1%显著为正,但工龄本身(*expr*)却变得很不显著;这是因为二者存在多重共线性(参见下一节)。再次使用解释变量的高次项进行 RESET 检验:

```
. estat ovtest,rhs

(note:  expr2 dropped because of collinearity)
(note:  expr2^2 dropped because of collinearity)

Ramsey RESET test using powers of the independent variables
      Ho:  model has no omitted variables
                 F(11, 741) =      1.73
                  Prob > F =      0.0626
```

从上表可知,可以在5%水平上,接受"无遗漏变量"的原假设。事实上,在本例中,最重要的模型设定误差乃是遗漏了对个人能力的度量,将在第10章进一步讨论。

## 9.6 多重共线性

如果在解释变量中,有某一解释变量可由其他解释变量线性表出,则存在"严格多重共线性"(strict multicollinearity)。在数学上,此时数据矩阵 $\boldsymbol{X}$ 不满列秩,$(\boldsymbol{X}'\boldsymbol{X})^{-1}$不存在,故无法定义OLS估计量 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$。比如,解释变量 $x_2$ 正好是解释变量 $x_3$ 的两倍,则无法区分 $x_2$ 与 $x_3$

对被解释变量 $y$ 的影响。

无严格多重共线性是对数据的最低要求，在现实中较少违背；即使出现，Stata 也会自动去掉多余的变量。在实践中更为常见的是近似（非严格）的多重共线性，简称“多重共线性”（multicollinearity）或“共线性”。多重共线性的主要表现是，如果将第 $k$ 个解释变量 $x_k$ 对其余的解释变量 $\{x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_K\}$ 进行回归，所得可决系数（记为 $R_k^2$）较高。

在多重共线性的情况下，OLS 仍是 BLUE，在所有线性无偏估计中方差最小，因为高斯 - 马尔可夫定理并未排除多重共线性的情形。然而，BLUE 只是保证 OLS 估计量在所有线性无偏估计量中相对而言方差最小，并不意味着 OLS 估计量的方差在绝对意义上小。

如果存在严格多重共线性，则矩阵 $(\boldsymbol{X}'\boldsymbol{X})$ 不可逆。类似地，在多重共线性的情况下，矩阵 $(\boldsymbol{X}'\boldsymbol{X})$ 变得“几乎不可逆”，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 变得很“大”，致使方差 $\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid X)=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 增大，系数估计变得不准确。此时，$\boldsymbol{X}$ 中元素轻微变化就会引起 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 很大变化，导致 OLS 估计值 $\hat{\boldsymbol{\beta}}$ 发生很大变化。

多重共线性的通常症状是，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。另一症状是，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。在严格多重共线性的极端情况下，一个变量刚好是其他变量的倍数，则完全无法区分。

总之，$R_k^2$ 越高，解释变量 $x_k$ 与其他解释变量的共线性越严重，则 $x_k$ 的系数估计量 $\hat{\beta}_k$ 的方差越大。具体来说，可以证明

$$\mathrm{Var}(\hat{\beta}_k\mid\boldsymbol{X})=\frac{\sigma^2}{(1-R_k^2)S_k} \tag{9.17}$$

其中，$\sigma^2\equiv\mathrm{Var}(\varepsilon)$ 为扰动项的方差；而 $S_k\equiv\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2$ 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（为常数），$S_k=0$，则完全无法估计 $\hat{\beta}_k$（此时 $x_k$ 与常数项构成严格多重共线性）。

从表达式(9.17)可知，方差 $\mathrm{Var}(\hat{\beta}_k\mid\boldsymbol{X})$ 与 $(1-R_k^2)$ 成反比。为此，定义解释变量 $x_k$ 的“方差膨胀因子”（Variance Inflation Factor，VIF）为

$$\mathrm{VIF}_k\equiv\frac{1}{1-R_k^2} \tag{9.18}$$

根据表达式(9.17)与(9.18)，可将方差写为

$$\mathrm{Var}(\hat{\beta}_k\mid\boldsymbol{X})=\mathrm{VIF}_k\cdot\frac{\sigma^2}{S_k} \tag{9.19}$$

其中，$\sigma^2$ 是扰动项本身的方差，$S_k$ 衡量变量 $x_k$ 自己的波动，而 $\mathrm{VIF}_k$ 才真正度量由于变量 $x_k$ 与其他解释变量的多重共线性所导致其方差 $\mathrm{Var}(\hat{\beta}_k\mid\boldsymbol{X})$ 的膨胀程度。方差膨胀因子 $\mathrm{VIF}_k$ 越大，则说明 $x_k$ 的多重共线性问题越严重，其方差 $\mathrm{Var}(\hat{\beta}_k\mid\boldsymbol{X})$ 将变得越大。

对于 $K$ 个解释变量$\{x_1,\cdots,x_K\}$，可计算相应的方差膨胀因子$\{\text{VIF}_1,\cdots,\text{VIF}_K\}$。判断是否存在多重共线性的一个经验规则是，$\{\text{VIF}_1,\cdots,\text{VIF}_K\}$ 的最大值不应超过 10。求解 $10=\dfrac{1}{1-R_k^2}$ 可知，相应的 $R_k^2$ 不应超过 0.9。显然，解释变量 $x_k$ 与其他变量的多重共线性越严重，$R_k^2$ 越接近于 1，则方差膨胀因子 $\text{VIF}_k$ 将急剧上升。更直观地，可在 Stata 中通过画出函数(9.18)来考察 $\text{VIF}_k$ 对 $R_k^2$ 的依赖性：

```
. twoway function VIF =1/(1 -x),xtitle(R2) xline(0.9,lp(dash)) yline
(10,lp(dash)) xlabel(0.1(0.1)1) ylabel(10 100 200 300)
```

其中，选择项“xtitle(R2)”指示横轴的标题为“R2”；选择项“xline(0.9,lp(dash))”与“yline(10,lp(dash))”分别表示在横轴 0.9 与纵轴 10 的位置画一条虚线；选择项“xlabel(0.1(0.1)1)”表示在横轴上，从 0.1 至 1，每隔 0.1 的位置给出标签；而选择项“ylabel(10 100 200 300)”则表示在纵轴上 10、100、200 与 300 的位置给出标签，结果参见图 9.1。

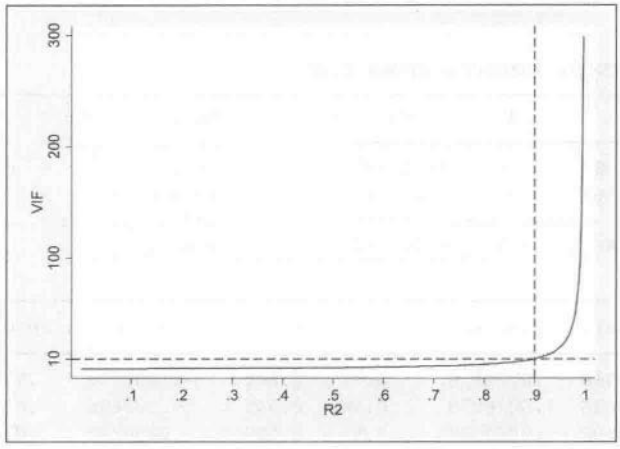



图 9.1 $\text{VIF}_k$ 与 $R_k^2$ 的关系

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，则通常可不必理会多重共线性(假设整个方程是显著的)。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，则也可不必理会。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，则应设法进行处理。比如，增大样本容量，剔除导致严重共线性的变量，将变量标准化(详见下文)，或对模型设定进行修改。

事实上，解释变量之间的相关性是普遍存在的，在一定程度上也是允许的。因此，处理多重共线性最常见的方法就是“无为而治”(do nothing)。

在 Stata 中作完回归后，可使用如下命令计算各变量的 VIF。

```
estat vif
```

仍以数据集 grilic.dta 为例。首先，考察线性回归的 VIF。

```
. use grilic.dta,clear
. qui reg lnw s expr tenure iq smsa rns
. estat vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| expr | 1.12 | 0.893267 |
| s | 1.07 | 0.930295 |
| tenure | 1.06 | 0.944083 |
| smsa | 1.04 | 0.964256 |
| rns | 1.03 | 0.970508 |
| Mean VIF | 1.06 | |

从上表可知，最大的 VIF 为 1.12，远小于 10，故不必担心存在多重共线性。

如果在模型中引入解释变量的平方项，则容易引起多重共线性，因为 $x$ 与 $x^2$ 通常较相关。考虑在上述回归中加入教育年限($s$)的平方项，记为 $s2$，再进行多重共线性检验。

```
. gen s2 = s^2
. reg lnw s s2 expr tenure smsa rns
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 6, 751) = | 68.26 |
| Model | 49.1549871 | 6 | 8.19249785 | Prob > F = | 0.0000 |
| Residual | 90.1311627 | 751 | .120014864 | R-squared = | 0.3529 |
| | | | | Adj R-squared = | 0.3477 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .34643 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .0339768 | .0729216 | 0.47 | 0.641 | -.1091776 | .1771312 |
| s2 | .0024636 | .0026079 | 0.94 | 0.345 | -.002656 | .0075832 |
| expr | .0375502 | .0063558 | 5.91 | 0.000 | .0250729 | .0500275 |
| tenure | .0356461 | .007743 | 4.60 | 0.000 | .0204456 | .0508466 |
| smsa | .1390585 | .0280915 | 4.95 | 0.000 | .0839113 | .1942058 |
| rns | -.0864204 | .0289057 | -2.99 | 0.003 | -.143166 | -.0296747 |
| _cons | 4.57118 | .5021415 | 9.10 | 0.000 | 3.585412 | 5.556948 |

从上表可知，教育年限($s$)与其平方项 $s2$ 都很不显著。显然，二者之间可能存在多重共线性。为此，计算各变量的 VIF 值。

```
. estat vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| s | 167.07 | 0.005986 |
| s2 | 166.30 | 0.006013 |
| expr | 1.13 | 0.885254 |
| tenure | 1.06 | 0.944065 |
| rns | 1.04 | 0.963378 |
| smsa | 1.04 | 0.963750 |
| Mean VIF | 56.27 | |

从上表可知，变量 $s$ 与 $s2$ 的 VIF 分别达到 167.07 与 166.30，远大于 10，故存在多重共线性。进一步，将 $s2$ 对 $s$ 进行回归。

```
. reg s2 s
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 1, 756) = | . |
| Model | 2916802.81 | 1 | 2916802.81 | Prob > F = | 0.0000 |
| Residual | 17941.0733 | 756 | 23.7315785 | R-squared = | 0.9939 |
| | | | | Adj R-squared = | 0.9939 |
| Total | 2934743.89 | 757 | 3876.8083 | Root MSE = | 4.8715 |

| s2 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | 27.81281 | .0793331 | 350.58 | 0.000 | 27.65707 | 27.96854 |
| _cons | -188.1622 | 1.078081 | -174.53 | 0.000 | -190.2785 | -186.0458 |

上表显示，此回归的 $R^2$ 高达 0.993 9，即变量 $s$ 可以解释其平方项 $s2$ 99% 的变动。这说明，$s$ 与 $s2$ 所包含的信息基本相同，故会导致严重的多重共线性。

一般来说，如果回归方程中包含解释变量的多项式（比如，$\beta x + \gamma x^2$），则通常会导致多重共线性。一个可能的解决方法是将变量标准化，即减去均值，除以标准差：

$$\tilde{x} \equiv \frac{x - \bar{x}}{s_x} \tag{9.20}$$

其中，$\bar{x}$ 为变量 $x$ 的样本均值，$s_x$ 为样本标准差，而 $\tilde{x}$ 为标准化之后的变量；然后，以 $\tilde{x}$ 及其平方 $\tilde{x}^2$ 作为解释变量。

继续上面的例子，先计算变量 $s$ 的均值与标准差，将其标准化，并记标准化变量为 $sd$：

```
. sum s
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| s | 758 | 13.40501 | 2.231828 | 9 | 18 |

```
. gen sd = (s - r(mean))/r(sd)
. gen sd2 = sd^2
. reg lnw sd sd2 expr tenure smsa rns
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 6, 751) = | 68.26 |
| Model | 49.154987 | 6 | 8.19249783 | Prob > F = | 0.0000 |
| Residual | 90.1311629 | 751 | .120014864 | R-squared = | 0.3529 |
| | | | | Adj R-squared = | 0.3477 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .34643 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sd | .2232421 | .014444 | 15.46 | 0.000 | .1948866 | .2515975 |
| sd2 | .0122714 | .0129899 | 0.94 | 0.345 | -.0132294 | .0377723 |
| expr | .0375502 | .0063558 | 5.91 | 0.000 | .0250729 | .0500275 |
| tenure | .0356461 | .007743 | 4.60 | 0.000 | .0204456 | .0508466 |
| smsa | .1390585 | .0280915 | 4.95 | 0.000 | .0839113 | .1942058 |
| rns | -.0864204 | .0289057 | -2.99 | 0.003 | -.143166 | -.0296747 |
| _cons | 5.469338 | .0319719 | 171.07 | 0.000 | 5.406574 | 5.532103 |

从上表可知，标准化的线性项 *sd* 在 1% 水平上显著为正，而标准化的平方项 *sd2* 不显著；多重共线性似乎有所缓解。下面，计算各变量的方差膨胀因子。

```
. estat vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| sd | 1.32 | 0.759911 |
| sd2 | 1.23 | 0.811990 |
| expr | 1.13 | 0.885254 |
| tenure | 1.06 | 0.944065 |
| rns | 1.04 | 0.963378 |
| smsa | 1.04 | 0.963750 |
| Mean VIF | 1.14 | |

上表显示，VIF 的最大值仅为 1.32，故认为基本不存在多重共线性。为了验证这一点，将 *sd2* 对 *sd* 进行回归：

```
. reg sd2 sd
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 1, 756) = | 159.77 |
| Model | 152.821515 | 1 | 152.821515 | Prob > F = | 0.0000 |
| Residual | 723.111439 | 756 | .956496612 | R-squared = | 0.1745 |
| | | | | Adj R-squared = | 0.1734 |
| Total | 875.932954 | 757 | 1.1571109 | Root MSE = | .97801 |

| sd2 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sd | .4493082 | .0355462 | 12.64 | 0.000 | .3795271 | .5190892 |
| _cons | .9986808 | .0355228 | 28.11 | 0.000 | .9289457 | 1.068416 |

此回归的 $R^2$ 仅为 0.174 5，相对于 *s2* 对 *s* 回归的 $R^2$（0.993 9）而言，大大下降。由于 *sd2* 在上面的回归中不显著，下面去掉 *sd2*，但保留 *sd*，再次进行回归：

```
. reg lnw sd expr tenure smsa rns
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 5, 752) = | 81.75 |
| Model | 49.0478812 | 5 | 9.80957624 | Prob > F = | 0.0000 |
| Residual | 90.2382686 | 752 | .119997698 | R-squared = | 0.3521 |
| | | | | Adj R-squared = | 0.3478 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .34641 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sd | .2290816 | .0130535 | 17.55 | 0.000 | .2034559 | .2547073 |
| expr | .0381189 | .0063268 | 6.02 | 0.000 | .0256986 | .0505392 |
| tenure | .0356146 | .0077424 | 4.60 | 0.000 | .0204153 | .0508138 |
| smsa | .1396666 | .0280821 | 4.97 | 0.000 | .0845379 | .1947954 |
| rns | -.0840797 | .0287973 | -2.92 | 0.004 | -.1406124 | -.0275471 |
| _cons | 5.479606 | .0300656 | 182.26 | 0.000 | 5.420584 | 5.538629 |

注意到 *sd* 的回归系数为 0.229 1，似乎偏高。但由于 *sd* 为标准化的变量，故 *sd* 变化一个单位，等价于 *s* 变化一个标准差，即 2.231 828 年。以此推算 *s* 的系数，即教育投资的年回报率应为

```
. dis .2290816/2.231828
.10264304
```

再次对比未将变量 *s* 标准化的回归：

```
. reg lnw s expr tenure smsa rns
```

| Source | SS | df | MS | Number of obs = | 758 |
|---|---|---|---|---|---|
| | | | | F( 5, 752) = | 81.75 |
| Model | 49.0478814 | 5 | 9.80957628 | Prob > F = | 0.0000 |
| Residual | 90.2382684 | 752 | .119997697 | R-squared = | 0.3521 |
| | | | | Adj R-squared = | 0.3478 |
| Total | 139.28615 | 757 | .183997556 | Root MSE = | .34641 |

| lnw | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| s | .102643 | .0058488 | 17.55 | 0.000 | .0911611 | .114125 |
| expr | .0381189 | .0063268 | 6.02 | 0.000 | .0256986 | .0505392 |
| tenure | .0356146 | .0077424 | 4.60 | 0.000 | .0204153 | .0508138 |
| smsa | .1396666 | .0280821 | 4.97 | 0.000 | .0845379 | .1947954 |
| rns | -.0840797 | .0287973 | -2.92 | 0.004 | -.1406124 | -.0275471 |
| _cons | 4.103675 | .085097 | 48.22 | 0.000 | 3.936619 | 4.270731 |

对比以上两表可知，是否将变量 *s* 标准化，对于回归结果（回归系数、标准误）没有任何实质性影响。

## 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，它们可能对 OLS 的回归系数产生很大影响。这些数据称为“极端观测值”（outliers）或“高影响力数据”（influential data），参见图 9.2。

以数据集 nerlove.dta 为例（参见第 7 章）。首先，打开数据集，进行回归；然后人为地构造一

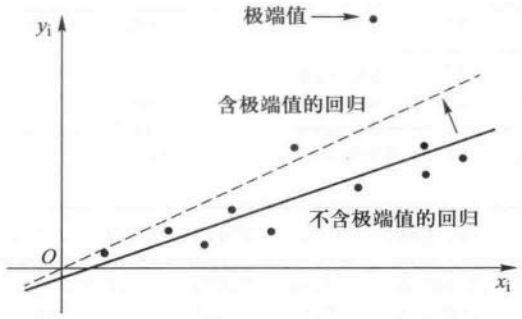



图 9.2 极端观测值对回归系数的影响

个极端值,再进行回归,并比较回归结果。

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

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

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

下面,将第一个观测值的产量对数($\ln q$)乘以 100,然后再次进行回归。

```
. replace lnq = lnq * 100 if _n == 1
(1 real change made)
```

其中,"_n"表示第 $n$ 个观测值,故"_n == 1"表示第 1 个观测值。

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

| Source | SS | df | MS | Number of obs = | 145 |
|---|---|---|---|---|---|
| | | | | F( 4, 140) = | 0.92 |
| Model | 7.4424142 | 4 | 1.86060355 | Prob > F = | 0.4551 |
| Residual | 283.624409 | 140 | 2.02588864 | R-squared = | 0.0256 |
| | | | | Adj R-squared = | -0.0023 |
| Total | 291.066823 | 144 | 2.02129738 | Root MSE = | 1.4233 |

| lntc | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | .0156026 | .0218326 | 0.71 | 0.476 | -.0275616 | .0587668 |
| lnpl | 1.214014 | 1.093264 | 1.11 | 0.269 | -.9474289 | 3.375456 |
| lnpk | -1.096614 | 1.233079 | -0.89 | 0.375 | -3.534477 | 1.341248 |
| lnpf | -.2427032 | .3641193 | -0.67 | 0.506 | -.9625867 | .4771803 |
| _cons | 7.230075 | 6.388544 | 1.13 | 0.260 | -5.400419 | 19.86057 |

对比以上两表可知，人为制造极端值后，回归系数的估计值变化很大，而且所有系数都变得不显著，$R^2$ 也从 0.926 降为 0.025 6（$\overline{R}^2$ 则变为负数）。所有这些变化都仅仅是因为在 145 个观测值中有一个观测值发生了变化，这正是“高影响力数据”（influential data）的含义。下面，将此人造极端值去掉，再对比回归结果。

```
. reg lntc lnq lnpl lnpk lnpf if _n >1
```

| Source | SS | df | MS | Number of obs = | 144 |
|---|---|---|---|---|---|
| | | | | F( 4, 139) = | 406.10 |
| Model | 251.560166 | 4 | 62.8900416 | Prob > F = | 0.0000 |
| Residual | 21.5261194 | 139 | .154864168 | R-squared = | 0.9212 |
| | | | | Adj R-squared = | 0.9189 |
| Total | 273.086286 | 143 | 1.90969431 | Root MSE = | .39353 |

| lntc | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnq | .7225462 | .0182135 | 39.67 | 0.000 | .6865348 | .7585576 |
| lnpl | .4445846 | .3028466 | 1.47 | 0.144 | -.1541969 | 1.043366 |
| lnpk | -.2219834 | .3415869 | -0.65 | 0.517 | -.8973614 | .4533947 |
| lnpf | .4311295 | .1019964 | 4.23 | 0.000 | .2294645 | .6327944 |
| _cons | -3.55227 | 1.78566 | -1.99 | 0.049 | -7.082837 | -.0217022 |

从上表可知，去掉极端值后的回归结果又“恢复正常”了，无论回归系数与显著性水平都类似于没有极端值的原始全样本。

如何发现极端数据？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行不通。从上例可知，某个观测值的影响力可通过去掉此观测值对回归系数的影响来衡量。记 $\hat{\boldsymbol{\beta}}$ 为全样本的 OLS 估计值，而 $\hat{\boldsymbol{\beta}}^{(i)}$ 为去掉第 $i$ 个观测值后的 OLS 估计值。我们关心 $(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^{(i)})$ 的变化幅度以及如何决定。

为此，定义第 $i$ 个观测数据对回归系数的“影响力”或“杠杆作用”（leverage）为

$$\text{lev}_i \equiv \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{9.21}$$

其中，$\boldsymbol{x}_i \equiv (1\ x_{i2}\cdots x_{iK})'$ 包含个体 $i$ 的全部解释变量，而 $\boldsymbol{X}=(\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\boldsymbol{x}_n)'$ 为数据矩阵。之所以这样定义，是因为 $\text{lev}_i$ 与 $(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^{(i)})$ 存在如下关系：

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)} = \left(\frac{1}{1 - \text{lev}_i}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i \tag{9.22}$$

由上式可知，$\text{lev}_i$越大，则$(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})$的变化越大。另外，可以证明，所有观测数据的影响力$\text{lev}_i$满足：(i) $0 \leqslant \text{lev}_i \leqslant 1, (i = 1, \cdots, n)$；(ii) $\sum_{i=1}^{n} \text{lev}_i = K$(解释变量个数)。因此，影响力$\text{lev}_i$的平均值为$K/n$。如果某些数据的$\text{lev}_i$比平均值$K/n$高很多，则可能对回归系数有很大影响。

在 Stata 中作完回归后，计算影响力$\text{lev}_i$的命令为

```
predict lev,leverage
```

此命令将计算所有观测数据的影响力，并记为变量 *lev*(可自行命名)。回到数据集 nerlove. dta 的例子。

```
. use nerlove.dta,clear
. qui reg lntc lnq lnpl lnpk lnpf
. predict lev,leverage
. sum 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
```

此命令将观测值按变量 *lev* 的降序排列。如果使用命令“sort lev”，则只能按升序排列。下面看 lev 取值最大的三个数据。

```
. list lev in 1/3
```

| | lev |
|---|---|
| 1. | .1177335 |
| 2. | .1001472 |
| 3. | .0983759 |

为了演示目的，再次人为制造极端数据，将第一个观测值的产量对数($\ln q$)乘以 100，然后计算 lev。

```
. replace lnq = lnq * 100 if _n == 1
. qui reg lntc lnq lnpl lnpk lnpf
. predict lev1,lev
. sum lev1
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lev1 | 145 | .0344828 | .0807897 | .0083048 | .9801415 |

```
. dis r(max)/r(mean)
28.424102
```

结果显示，lev 的最大值是其平均值的 28.42 倍，故存在高影响力的极端观测值。

如果发现存在极端数据，应如何处理呢？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，考察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。

## 9.8 虚拟变量

*Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex.* ——Fritz Machlup

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

$$D = \begin{cases} 1, & \text{男} \\ 0, & \text{女} \end{cases} \tag{9.23}$$

类似地，对于全球的五大洲[②]，则需要四个虚拟变量，即

$$D_1 = \begin{cases} 1, & \text{亚洲} \\ 0, & \text{其他} \end{cases}, \quad D_2 = \begin{cases} 1, & \text{美洲} \\ 0, & \text{其他} \end{cases},$$
$$D_3 = \begin{cases} 1, & \text{欧洲} \\ 0, & \text{其他} \end{cases}, \quad D_4 = \begin{cases} 1, & \text{非洲} \\ 0, & \text{其他} \end{cases} \tag{9.24}$$

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，则最多只能在回归方程中放入 $(M-1)$ 个虚拟变量。如果在回归方程中包含了 $M$ 个虚拟变量，则会产生严格多重共线性，因为如果将这 $M$ 个虚拟变量在数据矩阵 $\boldsymbol{X}$ 中对应的列向量相加，就会得到与常数项完全相同的向量，即 $(1\cdots1)'$（因为 $M$ 类中必居其一）。这种情况称为“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，则可以放入 $M$ 个虚拟变量。

**例** 假设样本中只有四位个体，分别属于三类。其中，前两位个体属于第一类，第三位个体属于第二类，而第四位个体属于第三类。相应地，定义三个虚拟变量 $D_1, D_2, D_3$，则其取值分别为：

① 比如，多输入了一个 0，或漏了一位数。

② 此处忽略南极洲，因为南极洲一般没有经济活动。

$$D_1 = \begin{pmatrix}1\\1\\0\\0\end{pmatrix}, \quad D_2 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \quad D_3 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{9.25}$$

考虑将这三个虚拟变量同时放入回归方程，并包含常数项：

$$y = \alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \tag{9.26}$$

显然，这三个虚拟变量之和正好就是常数项，因为

$$D_1 + D_2 + D_3 = \begin{pmatrix}1\\1\\0\\0\end{pmatrix} + \begin{pmatrix}0\\0\\1\\0\end{pmatrix} + \begin{pmatrix}0\\0\\0\\1\end{pmatrix} = \begin{pmatrix}1\\1\\1\\1\end{pmatrix} \tag{9.27}$$

因此，方程(9.26)存在严格多重共线性，即虚拟变量陷阱。解决方法之一是去掉一个虚拟变量，比如进行以下回归：

$$y = \alpha + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \tag{9.28}$$

解决方法之二是去掉常数项，即进行如下回归：

$$y = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \tag{9.29}$$

在模型中引入虚拟变量，会带来什么影响呢？考虑一个有关中国经济的时间序列模型：

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad (t = 1950, \cdots, 2000) \tag{9.30}$$

由于经济结构可能在1978年改革开放后有变化，故引入虚拟变量：

$$D_t = \begin{cases} 1, & \text{若 } t \geqslant 1978 \\ 0, & \text{其他} \end{cases} \tag{9.31}$$

根据引入虚拟变量的方式，考虑以下两种情况。

(1) 仅仅引入虚拟变量 $D_t$ 本身，则回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t \tag{9.32}$$

根据虚拟变量 $D_t$ 在不同时期的不同取值，该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + \beta x_t + \varepsilon_t, & \text{若 } t \geqslant 1978 \end{cases} \tag{9.33}$$

因此，仅仅引入虚拟变量相当于在不同时期给予不同的截距项，参见图9.3。

(2) 引入虚拟变量 $D_t$，以及虚拟变量与解释变量的“互动项”(interaction term) $D_t x_t$，则回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \tag{9.34}$$

显然，该模型等价于

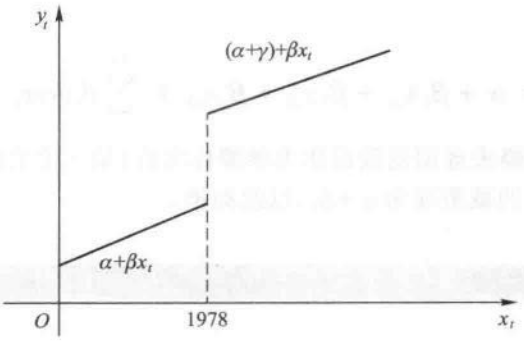



图 9.3 仅引入虚拟变量的效果

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + (\beta + \delta) x_t + \varepsilon_t, & \text{若 } t \geqslant 1978 \end{cases} \tag{9.35}$$

因此,引入虚拟变量及其互动项,相当于在不同时期使用不同的截距项与斜率,参见图 9.4。如果仅仅引入互动项,则仅改变斜率(这种情形比较少见)。

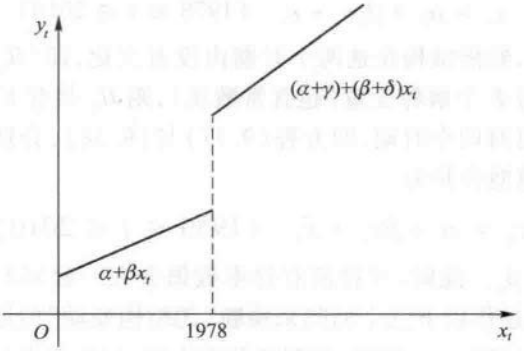



图 9.4 引入虚拟变量及其互动项的效果

在 Stata 中,假设时间变量为 *year*,可使用如下命令生成上文的虚拟变量:

```
gen d = (year > = 1978)
```

其中,"( )"表示对括弧内的表达式"year > =1978"进行逻辑判断。如果此表达式为真,则取值为 1;反之,取值为 0。

假设有 30 个省的名字储存于变量 *province*,希望为每个省设立一个虚拟变量,分别记为"*prov*1,*prov*2,…,*prov*30",则可使用如下 Stata 命令:

```
tabulate province,generate(prov)
```

其中,"tabulate"表示将变量按其取值列表;选择项"generate(prov)"表示根据此变量的不同取值生成以"*prov*"开头的虚拟变量。由此生成的这些虚拟变量,将按照变量 *province* 的字母顺序而排序。

在进行回归时可使用变量的简略写法,比如:

```
reg x1 x2 x3 prov2 - prov30
```

其中,为了避免虚拟变量陷阱而略去了第一个省的虚拟变量 prov1。与此相应的回归模型可

写为

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sum_{i=2}^{30} \delta_i prov_i + \varepsilon_i \tag{9.36}$$

根据上文的分析，(被略去虚拟变量而作为参照标准的)第一个省的截距项为 $\alpha$，第二个省的截距项为 $\alpha+\delta_2$，第三个省的截距项为 $\alpha+\delta_3$，以此类推。

## 9.9 经济结构变动的检验

对于时间序列而言，模型系数的稳定性(stability)是很重要的问题。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。在此，仅考虑结构变动的日期为已知的情形①。

继续上文的例子，假设要检验中国经济是否在 1978 年发生结构变动。定义第 1 个时期为 $1950 \leqslant t < 1978$，第 2 个时期为 $1978 \leqslant t \leqslant 2010$，则两个时期对应的回归方程可分别记为

$$y_t = \alpha_1 + \beta_1 x_t + \varepsilon_t \quad (1950 \leqslant t < 1978) \tag{9.37}$$

$$y_t = \alpha_2 + \beta_2 x_t + \varepsilon_t \quad (1978 \leqslant t \leqslant 2010) \tag{9.38}$$

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“$H_0: \alpha_1 = \alpha_2, \beta_1 = \beta_2$”，共有两个约束。更一般地，如果有 $K$ 个解释变量(包含常数项)，则 $H_0$ 共有 $K$ 个约束。

在无约束的情况下，可对两个时期，即方程(9.37)与(9.38)，分别进行回归。在有约束(即 $H_0$ 成立)的情况下，可将模型合并为

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad (1950 \leqslant t \leqslant 2010) \tag{9.39}$$

其中，$\alpha = \alpha_1 = \alpha_2, \beta = \beta_1 = \beta_2$。此时，可将所有样本数据合在一起回归，即方程(9.39)。传统的“邹检验”(Chow，1960)通过作以下三个回归来检验“无结构变动”的原假设。

首先，回归整个样本，$1950 \leqslant t \leqslant 2010$，得到残差平方和，记为 $SSR^*$。

其次，回归第 1 部分子样本，$1950 \leqslant t < 1978$，得到残差平方和 $SSR_1$。

最后，回归第 2 部分子样本，$1978 \leqslant t \leqslant 2010$，得到残差平方和 $SSR_2$。

其中，将整个样本一起回归为“有约束 OLS”，其残差平方和为 $SSR^*$。而将样本一分为二，分别进行回归则为“无约束 OLS”，其残差平方和为

$$SSR = SSR_1 + SSR_2 \tag{9.40}$$

显然，$SSR^* \geqslant SSR = SSR_1 + SSR_2$，因为有约束 OLS 的拟合优度比无约束 OLS 更差。如果 $H_0$ 成立(无结构变动)，则($SSR^* - SSR_1 - SSR_2$)应该比较小，即施加约束后，不应使得残差平方和上升很多。反之，如果($SSR^* - SSR_1 - SSR_2$)很大，则倾向于认为 $H_0$ 不成立，即存在结构变动。

根据第 5 章，在对 $m$ 个线性约束进行联合检验时，似然比检验原理的 $F$ 统计量为

$$F = \frac{(SSR^* - SSR)/m}{SSR/(n-K)} \sim F(m, n-K) \tag{9.41}$$

① 对于结构变动日期未知的情形，参见陈强(2014，p. 128)。

其中，SSR 为无约束的残差平方和，$SSR^*$ 为有约束的残差平方和，$n$ 为样本容量，而 $K$ 为无约束回归的参数个数。回到结构变动检验的情形，如果有 $K$ 个解释变量（包含常数项），则共有 $K$ 个约束条件，而无约束回归的参数个数为 $2K$。套用上面的公式，可得检验结构变动的 $F$ 统计量：

$$F = \frac{(SSR^* - SSR_1 - SSR_2)/K}{(SSR_1 + SSR_2)/(n-2K)} \sim F(K, n-2K) \tag{9.42}$$

其中，$n$ 为样本容量，$K$ 为有约束回归的参数个数（含常数项）。对于此一元回归的例子，$K=2$。

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，对于 $K=2$ 的情形，可进行如下回归：

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \tag{9.43}$$

然后检验联合假设 $H_0: \gamma = \delta = 0$。此检验所得 $F$ 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“球形扰动项”（同方差、无自相关）的假设下得到的，并不适用于异方差或自相关的情形。在异方差或自相关的情况下，仍可使用虚拟变量法，只要在估计方程（9.43）时，使用异方差自相关稳健的 HAC 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），而虚拟变量法则可同时提供这些信息。

下面以数据集 consumption. dta 为例，考察中国的消费函数是否在 1992 年发生了结构变化。数据来自国家统计局网站[①]。先考察中国 1978—2013 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 9.5），以当年价格计。

```
. use consumption.dta,clear
. twoway connect c y year,msymbol(circle) msymbol(triangle)
```

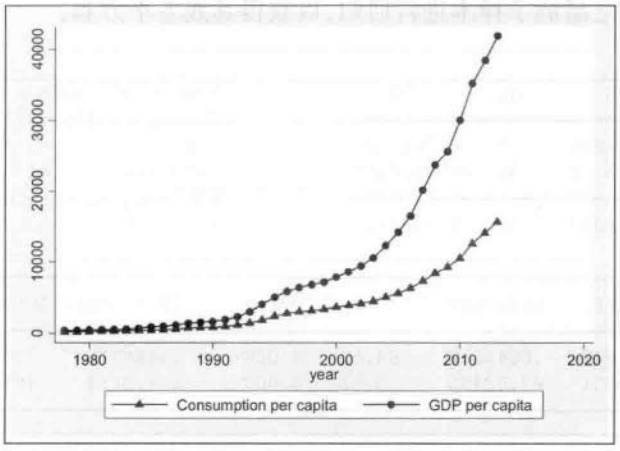



图 9.5 居民人均消费与人均国内总产值时间趋势

① 网址为 http://data.stats.gov.cn/workspace/index? m=hgnd。

从图 9.5 可知，二者的走势具有较强相关性。但图 9.5 的右边有些空白区域，不够美观。为此，将上述命令略加改进，并在 1992 年处画一条垂直线（结果见图 9.6）：

```
. twoway connect c y year,msymbol(circle) msymbol(triangle) xlabel
(1980(10)2010) xline(1992)
```

其中，选择项“xlabel(1980(10)2010)”指示在横轴（即 $X$ 轴）1980—2010 年之间，每隔 10 年做个标注（label）；选择项“xline(1992)”表示在横轴 1992 年的位置画一条直线。

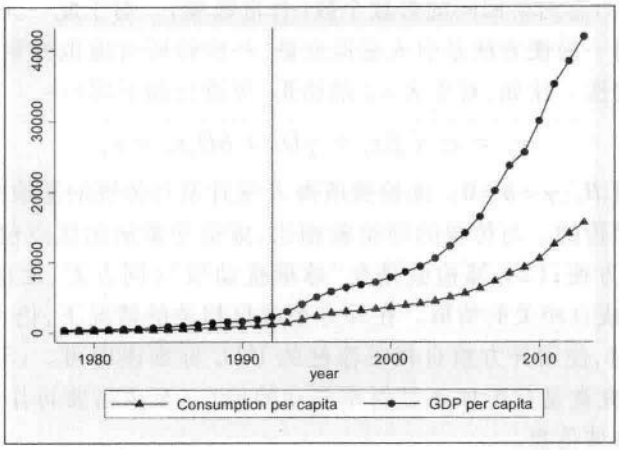



图 9.6 改进的趋势图

考察一个简单（粗糙）的消费函数：

$$c_t = \alpha + \beta y_t + \varepsilon_t$$

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 = | 36 |
|---|---|---|---|---|---|
| | | | | F( 1, 34) = | 7139.56 |
| Model | 617812224 | 1 | 617812224 | Prob > F = | 0.0000 |
| Residual | 2942143.12 | 34 | 86533.6213 | R-squared = | 0.9953 |
| | | | | Adj R-squared = | 0.9951 |
| Total | 620754367 | 35 | 17735839.1 | Root MSE = | 294.17 |

| c | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | .3572642 | .0042282 | 84.50 | 0.000 | .3486715 | .3658569 |
| _cons | 339.0701 | 63.83305 | 5.31 | 0.000 | 209.3458 | 468.7945 |

```
. scalar ssr = e(rss)
```

其中，“scalar”表示标量，即将此回归的残差平方和（e(rss)）记为标量 *ssr*。下面，对 1992 年之前的子样本进行回归。

```
. reg c y if year <1992
```

| Source | SS | df | MS | Number of obs = 14 |
|---|---|---|---|---|
| | | | | F( 1, 12) = 4344.64 |
| Model | 829125.648 | 1 | 829125.648 | Prob > F = 0.0000 |
| Residual | 2290.06599 | 12 | 190.838833 | R-squared = 0.9972 |
| | | | | Adj R-squared = 0.9970 |
| Total | 831415.714 | 13 | 63955.0549 | Root MSE = 13.814 |

| c | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | .4996452 | .0075803 | 65.91 | 0.000 | .4831292 | .5161612 |
| _cons | 12.89123 | 7.908212 | 1.63 | 0.129 | -4.339283 | 30.12174 |

```
. scalar ssr1 = e(rss)
```

此命令将 1992 年之前的子样本回归的残差平方和记为 *ssr*1。然后，对 1992 年及之后的子样本进行回归。

```
. reg c y if year >=1992
```

| Source | SS | df | MS | Number of obs = 22 |
|---|---|---|---|---|
| | | | | F( 1, 20) = 4889.80 |
| Model | 366038781 | 1 | 366038781 | Prob > F = 0.0000 |
| Residual | 1497151 | 20 | 74857.5501 | R-squared = 0.9959 |
| | | | | Adj R-squared = 0.9957 |
| Total | 367535932 | 21 | 17501711 | Root MSE = 273.6 |

| c | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | .3444589 | .004926 | 69.93 | 0.000 | .3341836 | .3547343 |
| _cons | 658.1088 | 95.04293 | 6.92 | 0.000 | 459.8527 | 856.3648 |

```
. scalar ssr2 = e(rss)
```

此命令将 1992 年之后的子样本回归的残差平方和记为 *ssr*2。由于 $n=36, K=2, n-2K=32$，故可计算 $F$ 统计量如下：

```
. di((ssr - ssr1 - ssr2)/2)/((ssr1 + ssr2)/32)
15.394558
```

故 $F$ 统计量等于 15.39。

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 $yd$，进行全样本 OLS 回归：

```
. reg c y d yd
```

| Source | SS | df | MS | Number of obs = | 36 |
|---|---|---|---|---|---|
| | | | | F( 3, 32) = | 4405.23 |
| Model | 619254926 | 3 | 206418309 | Prob > F = | 0.0000 |
| Residual | 1499441.07 | 32 | 46857.5333 | R-squared = | 0.9976 |
| | | | | Adj R-squared = | 0.9974 |
| Total | 620754367 | 35 | 17735839.1 | Root MSE = | 216.47 |

| c | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | .4996452 | .1187794 | 4.21 | 0.000 | .2576994 | .741591 |
| d | 645.2175 | 144.9484 | 4.45 | 0.000 | 349.9673 | 940.4678 |
| yd | -.1551863 | .1188434 | -1.31 | 0.201 | -.3972623 | .0868897 |
| _cons | 12.89123 | 123.9181 | 0.10 | 0.918 | -239.5216 | 265.3041 |

然后检验 $d$ 与 $yd$ 的联合显著性：

```
. test d yd
```

```
( 1)  d = 0
( 2)  yd = 0

       F(  2,    32) =   15.39
            Prob > F =    0.0000
```

上表显示，使用虚拟变量法所得 $F$ 统计量也为 15.39，与传统邹检验完全相同。该检验的 $p$ 值为 0.000 0，故可在 1% 水平上强烈拒绝“无结构变动”的原假设。然而，上述结构变化检验仅在球形扰动项（同方差、无自相关）的情况下才成立。为此，下面进行异方差与自相关的检验（参见第 7 章与第 8 章）。

```
. qui reg c y
. estat imtest,white
```

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(2)      =      6.31
         Prob > chi2  =    0.0427

Cameron & Trivedi's decomposition of IM-test
```

| Source | chi2 | df | p |
|---|---|---|---|
| Heteroskedasticity | 6.31 | 2 | 0.0427 |
| Skewness | 4.11 | 1 | 0.0425 |
| Kurtosis | 4.76 | 1 | 0.0291 |
| Total | 15.19 | 4 | 0.0043 |

上表显示，怀特检验的结果可在 5% 的水平上拒绝“同方差”的原假设。为了进行自相关的 BG 检验，首先设定变量 *year* 为时间变量。

```
. tsset year
```

```
        time variable:  year, 1978 to 2013
                delta:  1 unit
```

```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation

| lags(p) | chi2 | df | Prob > chi2 |
|---|---|---|---|
| 1 | 28.109 | 1 | 0.0000 |

H0: no serial correlation

上表显示,可在1%水平上强烈拒绝"无自相关"的原假设。总之,此模型的扰动项存在异方差与自相关。故应使用异方差自相关稳健的标准误,通过虚拟变量法检验结构变动。首先,计算HAC标准误的截断参数。

```
. dis 36^(1/4)
2.4494897
```

故应将截断参数设为3。下面进行Newey-West回归。

```
. newey c y d yd,lag(3)
```

Regression with Newey-West standard errors　　Number of obs = 36
maximum lag: 3　　F( 3, 32) = 2455.09
　　Prob > F = 0.0000

| c | Coef. | Newey-West Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| y | .4996452 | .0099228 | 50.35 | 0.000 | .4794332 | .5198573 |
| d | 645.2175 | 139.943 | 4.61 | 0.000 | 360.1629 | 930.2721 |
| yd | -.1551863 | .013774 | -11.27 | 0.000 | -.1832431 | -.1271295 |
| _cons | 12.89123 | 10.16563 | 1.27 | 0.214 | -7.815475 | 33.59793 |

然后,检验虚拟变量 $d$ 及其互动项 $yd$ 的联合显著性。

```
. test d yd
```

```
 (1)  d = 0
 (2)  yd = 0

       F(  2,    32) =   73.05
            Prob > F =    0.0000
```

该检验的 $p$ 值为0.000 0,故可在1%水平上强烈拒绝"无结构变动"的原假设,即认为中国的消费函数在1992年发生了结构变动。

## 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),会自动将缺失的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,为保持样本容量,

可采用“线性插值”(linear interpolation)的方法来补上缺失数据。

考虑最简单的情形。已知 $x_{t-1}$ 与 $x_{t+1}$,但缺失 $x_t$ 的数据,则 $x_t$ 对时间 $t$ 的线性插值为

$$\hat{x}_t = \frac{x_{t-1} + x_{t+1}}{2} \tag{9.44}$$

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 $\hat{y}$ 满足(参见图 9.7):

$$\frac{\hat{y} - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{9.45}$$

经整理可得

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0 \tag{9.46}$$

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 $\widehat{\ln y}$ 再取反对数(antilog),即计算 $\exp(\widehat{\ln y})$。

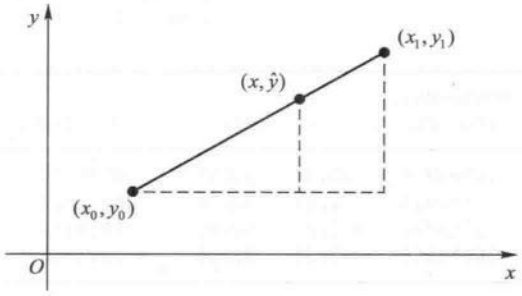



图 9.7 线性插值示意图

线性插值的 Stata 命令为

```
ipolate y x,gen(newvar)
```

其中,“ipolate”表示 interpolate,即将变量 $y$ 对变量 $x$ 进行线性插值,并将插值的结果记为新变量 *newvar*。

继续以数据集 consumption. dta 为例。

```
. use consumption.dta,clear
```

为了演示目的,假设 1980 年、1990 年、2000 年及 2010 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,并记为 $y1$。

```
. gen y1 =y
. replace y1 = . if year ==1980 | year ==1990 | year ==2000 | year ==2010
(4 real changes made,4 to missing)
```

下面,直接用 $y1$ 对 *year* 进行线性插值,并将结果记为 $y2$。

```
. ipolate y1 year,gen(y2)
```

由于人均 GDP 有指数增长趋势，故更好的做法是，先对 y1 取对数，进行线性插值，再取反对数，并将结果记为 y3。

```
. gen lny1 = log(y1)
(4 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 == 2010
```

| | year | y | y2 | y3 |
|---|---|---|---|---|
| 3. | 1980 | 463.25 | 455.705 | 454.2445 |
| 13. | 1990 | 1644 | 1705.88 | 1695.613 |
| 23. | 2000 | 7857.68 | 7890.105 | 7856.112 |
| 33. | 2010 | 30015.1 | 30402.659 | 30022.13 |

从上表可知，直接插值的结果 y2 倾向于高估真实值 y，而且整体估计效果不如先取对数再插值的结果 y3（1980 年的结果是个例外）。

## 9.11 变量单位的选择

在选择变量单位时，应尽量避免变量间的数量级差别过于悬殊，以免出现计算机运算的较大误差。比如，通货膨胀率通常小于 1，而如果模型中有 GDP 这个变量，则 GDP 应该使用亿或万亿作为单位。否则，变量 GDP 的取值将是通货膨胀率的很多倍，即数据矩阵 $\boldsymbol{X}$ 中某列的数值是另一列的很多倍，这可能使计算机在对 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 进行数值计算时出现较大误差。这是因为，计算机的存储空间有限，实际上只能作近似计算，即精确到小数点后若干位。

**习题**

9.1 在使用样本数据估计回归方程 $y=\alpha+\beta x+\varepsilon$ 时，如果怀疑 $x$ 对 $y$ 的作用还依赖于另一变量 $z$，应该如何检验此依赖性？

9.2 假设所估计的成本函数为 $\ln C=\alpha+\beta\ln Q+\gamma(\ln Q)^2+\varepsilon$，其中 $C$ 为成本，而 $Q$ 为产量。在 $\ln Q$ 的样本均值处，应该如何检验成本的产出弹性（elasticity of cost with respect to output）为 1 的原假设？

9.3 使用数据集 nerlove. dta，估计以下模型：

$$\ln tc_i = \beta_1 + \beta_2 \ln q_i + \beta_3 \ln pl_i + \beta_4 \ln pk_i + \beta_5 \ln pf_i + \varepsilon_i \tag{9.47}$$

其中，$\ln tc$，$\ln q$，$\ln pl$，$\ln pk$ 与 $\ln pf$ 分别为电力企业的总成本、总产量、小时工资率、资本使用成本、燃料价格的对数（参见第 6 章）。

（1）使用稳健标准误，对方程（9.47）进行 OLS 回归。

（2）计算 VIF。是否存在多重共线性？

（3）使用拟合值进行 RESET 检验。是否遗漏了非线性项？

(4) 在方程(9.47)中,加入 $\ln q$ 的平方项,重新进行回归。

(5) 再次使用拟合值进行 RESET 检验。是否还遗漏了非线性项?

(6) 再次计算 VIF。是否存在多重共线性?

(7) 从经济理论出发,以上两个回归结果,哪个更可信?

9.4[①] 使用数据集 Growth. dta 考察贸易与增长的关系。该数据集的被解释变量为 65 个国家 1960—1995 年的平均增长率(*growth*),而主要解释变量为 1960—1995 年的平均贸易开放度(*tradeshare*)。

(1) 将 *growth* 与 *tradeshare* 的散点图与线性拟合图画在一起。二者看上去是否有关系?

(2) 有一个国家马耳他(Malta),其贸易开放度比其他国家高很多。在散点图上找出马耳他。马耳他是否像极端值?

(3) 使用全样本,把 *growth* 对 *tradeshare* 进行回归。该回归的斜率与截距项估计值分别是多少?

(4) 计算每个观测值的影响力(*leverage*),以及此影响力的最大值与平均值之比。是否存在极端值?

(5) 去掉马耳他,重复上述回归,并再次回答(3)中的问题。(提示:可使用选择项"`if _n <65`"来去掉马耳他,其中"`_n`"表示第 $n$ 个观测值。)

(6) 马耳他在哪?马耳他的贸易开放度为什么这么高?是否应在本研究中去掉马耳他?

(7) 把 *growth* 对 *tradeshare*,*rgdp*60(1960 年的人均 GDP),*yearsschool*(1960 年的平均受教育年限),*rev-coups*(1960—1995 年的年平均政变次数),以及 *assassinations*(1960—1995 年的年平均政治暗杀次数)进行回归。评论各变量系数的符号、统计显著性与经济意义。

(8) 为什么将变量 *rgdp*60 与 *yearsschool* 的取值定为期初的 1960 年?

9.5[②] 美国的汽油需求函数是否稳定?使用数据集 gasoline. dta,估计美国 1953—2004 年的汽油需求函数(参见第 8 章):

$$lgasq_t = \beta_0 + \beta_1 lgasq_{t-1} + \beta_2 lincome_t + \beta_3 lgasp_t + \beta_4 lpnc_t + \beta_5 lpuc_t + \varepsilon_t \tag{9.48}$$

其中,被解释变量 *lgasq* 为人均汽油消费量的对数,解释变量 *lincome* 为人均收入对数,*lgasp* 为汽油价格指数的对数,*lpnc* 为新车价格指数的对数,*lpuc* 为二手车价格指数的对数。

(1) 将 *lgasq* 与 *lgasp* 的时间趋势图画在一起。根据此图,在 1953—2004 年期间,美国的汽油需求函数是否曾出现结构变动?

(2) 使用 OLS 估计方程(9.48)。

(3) 使用 BP 检验与怀特检验,检验是否存在异方差。

(4) 使用 BG 检验与 $Q$ 检验,检验是否存在自相关。

(5) 1973 年 10 月爆发石油危机,可能引起汽油需求的结构变动。使用虚拟变量法,检验美国的汽油需求函数是否在 1974 年发生结构变动。根据(3)与(4)的检验结果决定是否应使用稳健标准误。

---

① 此题改编自 Stock and Watson(2012)。

② 此题改编自 Greene(2011)。

*Give me a lever long enough and a place to stand, and I will move the world.* ——Archimedes

# 10. 工具变量法

OLS 能够成立的最重要条件是解释变量与扰动项不相关(即前定变量或同期外生的假设)。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,不一致的估计是无法接受的。然而,解释变量与扰动项相关(内生性)的例子却比比皆是。解决内生性的主要方法之一为工具变量法,它对于实证研究有着重要的价值。

内生性的主要来源包括遗漏变量偏差、联立方程偏差(双向因果关系),以及测量误差偏差(measurement error bias)。前者已在第 9 章讨论,下面首先介绍后二者。

## 10.1 联立方程偏差

**例** 考察如下农产品市场均衡模型:

$$\begin{cases} q_t^d = \alpha + \beta p_t + u_t & (\text{需求}) \\ q_t^s = \gamma + \delta p_t + v_t & (\text{供给}) \\ q_t^d = q_t^s & (\text{均衡}) \end{cases} \tag{10.1}$$

其中,$q_t^d$ 为农产品需求,$q_t^s$ 为农产品供给,而 $p_t$ 为农产品价格。市场出清(market clearing)的均衡条件要求 $q_t^d = q_t^s$。令 $q_t \equiv q_t^d = q_t^s$,可得

$$\begin{cases} q_t = \alpha + \beta p_t + u_t \\ q_t = \gamma + \delta p_t + v_t \end{cases} \tag{10.2}$$

显然,这两个方程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。

从数据生成过程(data generating process)的视角,可把线性方程组(10.2)中的$(p_t, q_t)$看成是未知数(内生变量),而把$(u_t, v_t)$看作已知,然后求解$(p_t, q_t)$为$(u_t, v_t)$的函数[①]。由此可知,解释变量 $p_t$ 与两个方程的扰动项$(u_t, v_t)$都相关,即 $\text{Cov}(p_t, u_t) \neq 0$,$\text{Cov}(p_t, v_t) \neq 0$。直观来看,一方面,对于需求函数的正冲击$(u_t > 0)$,将使得均衡价格 $p_t$ 上升,故二者正相关;另一方面,对于供给函数的正冲击$(v_t > 0)$,将使得均衡价格 $p_t$ 下降,故二者负相关。因此,OLS 估计值$(\hat{\beta}, \hat{\delta})$不是$(\beta, \delta)$的一致估计量,称这种偏差为"联立方程偏差"(simultaneity bias)或"内生性偏差"(endogeneity bias)。

---

① 即"上帝"先掷骰子,确定扰动项$(u_t, v_t)$,然后根据方程组(10.2)的均衡条件,决定市场出清的均衡价格与产量$(p_t, q_t)$。

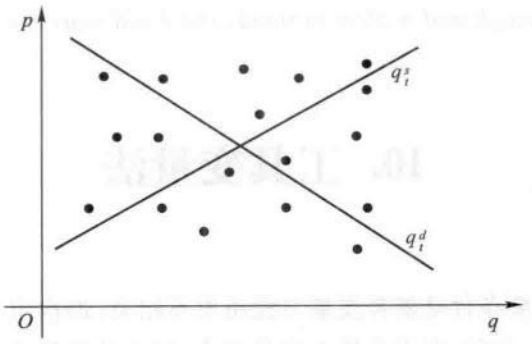



图 10.1 需求与供给决定市场均衡

**例** 考察宏观经济模型中的消费函数：

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases} \tag{10.3}$$

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。第一个方程为消费方程，而第二个方程为国民收入恒等式。显然，如果单独对消费方程进行 OLS 回归，将存在联立方程偏差，得不到一致估计。

## 10.2 测量误差偏差

导致内生性的另一来源是解释变量的测量误差（measurement error 或 errors-in-variables）。

**例** 假设真实模型为

$$y = \alpha + \beta x^* + \varepsilon \tag{10.4}$$

其中，$\beta \neq 0$，$\mathrm{Cov}(x^*, \varepsilon) = 0$。但 $x^*$ 无法精确观测，而只能观测到 $x$，二者满足如下关系：

$$x = x^* + u \tag{10.5}$$

其中，$\mathrm{Cov}(x^*, u) = 0$（测量误差 $u$ 与被测量变量 $x^*$ 不相关），$\mathrm{Cov}(u, \varepsilon) = 0$（测量误差与扰动项 $\varepsilon$ 不相关）。将表达式(10.5)代入方程(10.4)可得

$$y = \alpha + \beta x + (\varepsilon - \beta u) \tag{10.6}$$

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

$$\begin{aligned} \mathrm{Cov}(x, \varepsilon - \beta u) &= \mathrm{Cov}(x^* + u, \varepsilon - \beta u) \\ &= \underbrace{\mathrm{Cov}(x^*, \varepsilon)}_{=0} - \beta \underbrace{\mathrm{Cov}(x^*, u)}_{=0} + \underbrace{\mathrm{Cov}(u, \varepsilon)}_{=0} - \beta \mathrm{Cov}(u, u) \\ &= -\beta \mathrm{Var}(u) \neq 0 \end{aligned} \tag{10.7}$$

因此，OLS 估计不一致。由于解释变量测量误差所造成的 OLS 估计偏差，称为“测量误差偏差”（measurement error bias）。

如果被解释变量存在测量误差，后果却不严重，比如，只要被解释变量的测量误差与解释变量不相关，则 OLS 依然一致（参见习题）。

## 10.3 工具变量法

既然 OLS 的不一致性是由于内生变量与扰动项相关而引起，如果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[①]，而更常见的方法则借助另外一个“工具变量”来实现。

回到农产品市场均衡的例子。假设在图 10.1 中，存在某个因素（变量）使得供给曲线经常移动，而需求曲线基本不动。此时，就可以估计需求曲线，参见图 10.2。这个使得供给曲线移动的变量就是工具变量。

假设影响方程组（10.2）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 $z_t$ 与不可观测的其他因素：

$$q_t^s = \gamma + \delta p_t + \eta z_t + v_t \tag{10.8}$$

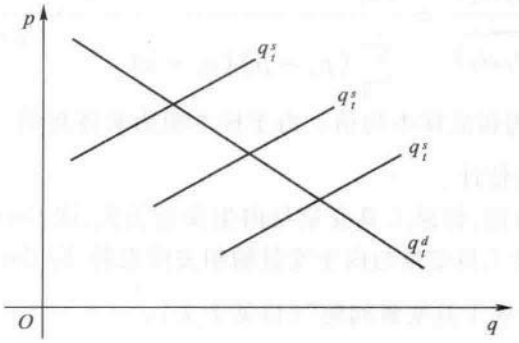



图 10.2 稳定的需求与变动的供给

假定气温 $z_t$ 是前定变量[②]，与需求方程的扰动项不相关，即 $\mathrm{Cov}(z_t, u_t) = 0$。由于气温 $z_t$ 的变化使得供给函数 $q_t^s$ 沿着需求函数 $q_t^d$ 移动，故可以估计出需求函数 $q_t^d$。在这种情况下，称 $z_t$ 为“工具变量”（Instrumental Variable，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（relevance）：工具变量与内生解释变量相关，即 $\mathrm{Cov}(z_t, p_t) \neq 0$。

（ii）外生性（exogeneity）：工具变量与扰动项不相关，即 $\mathrm{Cov}(z_t, u_t) = 0$。

显然，在本例中，气温 $z_t$ 满足这两个条件。

（i）相关性：从联立方程组可以解出 $p_t = p_t(z_t, u_t, v_t)$，故 $\mathrm{Cov}(z_t, p_t) \neq 0$。

（ii）外生性：假定气温 $z_t$ 是前定变量，故 $\mathrm{Cov}(z_t, u_t) = 0$。

利用工具变量的这两个性质，可得到对需求方程回归系数 $\beta$ 的一致估计。回顾需求方程为

---

① 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。

② 一般而言，人类活动对气温的影响可以忽略。

$$q_t = \alpha + \beta p_t + u_t \tag{10.9}$$

在此方程两边，同时求与 $z_t$ 的协方差：

$$\begin{aligned} \mathrm{Cov}(q_t, z_t) &= \mathrm{Cov}(\alpha + \beta p_t + u_t, z_t) \\ &= \beta \mathrm{Cov}(p_t, z_t) + \underbrace{\mathrm{Cov}(u_t, z_t)}_{=0} = \beta \mathrm{Cov}(p_t, z_t) \end{aligned} \tag{10.10}$$

其中，由于工具变量的外生性，故 $\mathrm{Cov}(u_t, z_t) = 0$。进一步，根据工具变量的相关性，$\mathrm{Cov}(p_t, z_t) \neq 0$。把上式两边同除以 $\mathrm{Cov}(p_t, z_t)$ 可得

$$\beta = \frac{\mathrm{Cov}(q_t, z_t)}{\mathrm{Cov}(p_t, z_t)} \tag{10.11}$$

上式相当于总体矩条件。以相应的样本矩取代上式的总体矩（即以样本协方差替代总体协方差），可得一致的“工具变量估计量”（instrumental variable estimator）：

$$\hat{\beta}_{\mathrm{IV}} = \frac{\widehat{\mathrm{Cov}(q_t, z_t)}}{\widehat{\mathrm{Cov}(p_t, z_t)}} = \frac{\sum_{t=1}^{n} (q_t - \bar{q})(z_t - \bar{z})}{\sum_{t=1}^{n} (p_t - \bar{p})(z_t - \bar{z})} \xrightarrow{p} \frac{\mathrm{Cov}(q_t, z_t)}{\mathrm{Cov}(p_t, z_t)} = \beta \tag{10.12}$$

其中，$\bar{q}, \bar{p}, \bar{z}$ 分别为 $q, p, z$ 的相应样本均值。由于样本矩为总体矩的一致估计，故工具变量估计量 $\hat{\beta}_{\mathrm{IV}}$ 是真实参数 $\beta$ 的一致估计。

从上式也可看出，一方面，如果工具变量与内生变量无关，即 $\mathrm{Cov}(z_t, p_t) = 0$，则无法定义工具变量法。另一方面，如果工具变量与内生变量的相关性很弱，即 $\mathrm{Cov}(z_t, p_t) \approx 0$，会导致估计量 $\hat{\beta}_{\mathrm{IV}}$ 的方差变得很大，称为“弱工具变量问题”（详见下文）。

## 10.4 二阶段最小二乘法

工具变量法一般通过“二阶段最小二乘法”（Two Stage Least Square，2SLS 或 TSLS）来实现（Theil，1953；Basmann，1957），顾名思义，即通过作两个回归来完成。

第一阶段回归：用内生解释变量对工具变量回归，即 $p_t \xrightarrow{\mathrm{OLS}} z_t$，得到拟合值 $\hat{p}_t$。

第二阶段回归：用被解释变量对第一阶段回归的拟合值进行回归，即 $q_t \xrightarrow{\mathrm{OLS}} \hat{p}_t$。

为什么这样做能得到好结果呢？首先，把需求方程 $q_t = \alpha + \beta p_t + u_t$ 分解为

$$q_t = \alpha_0 + \beta \hat{p}_t + \underbrace{[u_t + \beta(p_t - \hat{p}_t)]}_{=\varepsilon_t} \tag{10.13}$$

上式就是在第二阶段所作的回归，其扰动项为 $\varepsilon_t \equiv u_t + \beta(p_t - \hat{p}_t)$。

**命题** 在第二阶段回归中，$\hat{p}_t$ 与扰动项 $\varepsilon_t$ 不相关。

**证明：** 由于 $\varepsilon_t \equiv u_t + \beta(p_t - \hat{p}_t)$，故

$$\mathrm{Cov}(\hat{p}_t, \varepsilon_t) = \mathrm{Cov}(\hat{p}_t, u_t) + \beta \mathrm{Cov}(\hat{p}_t, p_t - \hat{p}_t) \tag{10.14}$$

首先，由于 $\hat{p}_t$ 是 $z_t$ 的线性函数（$\hat{p}_t$ 为第一阶段回归的拟合值），而 $\mathrm{Cov}(z_t, u_t) = 0$（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,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 $\text{Cov}(\hat{p}_t, p_t - \hat{p}_t) = 0$。因此,第二阶段回归的解释变量 $\hat{p}_t$ 与扰动项 $\varepsilon_t$ 不相关,故 2SLS 为一致估计。

从此例可看出,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 $z_t$ 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。

如果存在多个工具变量,也不难应用 2SLS 法。假设 $z_1$ 与 $z_2$ 为两个有效工具变量(都满足相关性与外生性),则第一阶段回归变为

$$p = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + u \tag{10.15}$$

由此可得拟合值 $\hat{p} = \hat{\alpha}_0 + \hat{\alpha}_1 z_1 + \hat{\alpha}_2 z_2$,而第二阶段回归不变。进一步,考虑存在多个内生变量的情形,比如

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{10.16}$$

其中,$x_1$ 与 $x_2$ 均为内生解释变量,都与 $\varepsilon$ 相关。由于有两个内生变量,则至少需要两个工具变量,才能进行 2SLS 估计,理由如下。如果只有一个工具变量 $z$,则由第一阶段回归可得,$\hat{x}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 z$,而 $\hat{x}_2 = \hat{\gamma}_0 + \hat{\gamma}_1 z$。将 $\hat{x}_1$ 与 $\hat{x}_2$ 代入原方程:

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + v_t \tag{10.17}$$

由于 $\hat{x}_1$ 与 $\hat{x}_2$ 都是 $z$ 的线性函数,故此方程存在严格多重共线性,即可以找到 $\hat{x}_1$ 与 $\hat{x}_2$ 的线性组合为常数(参见习题)。因此,如果存在两个内生变量,则至少需要两个工具变量,才能进行工具变量法的估计。更一般地,可将此结论推广为以下阶条件。

**阶条件**:进行 2SLS 估计的必要条件是工具变量个数不少于内生解释变量的个数,称为"阶条件"(order condition)。

根据阶条件是否满足可分为以下三种情况:

(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。

在恰好识别与过度识别的情况下,都可以使用 2SLS;但在不可识别的情况下,则无法使用 2SLS。更一般地,考虑多个内生变量,且包含外生解释变量的情形。比如,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 w + \varepsilon \tag{10.18}$$

其中,$x_1$ 与 $x_2$ 为内生变量,而 $w$ 为外生解释变量(与扰动项 $\varepsilon$ 不相关)。假设有三个有效工具变量 $z_1, z_2, z_3$。在 2SLS 的第一阶段回归中,应分别将两个内生解释变量($x_1, x_2$)对所有外生变量(包括工具变量 $z_1, z_2, z_3$ 及外生解释变量 $w$)进行回归:

$$x_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 w + u \tag{10.19}$$

$$x_2 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 w + v \tag{10.20}$$

其中,外生解释变量 $w$ 可视为自己的工具变量,因为满足工具变量的两个条件。首先,$w$ 显然与

$w$ 高度相关，故满足相关性。其次，$w$ 与扰动项 $\varepsilon$ 不相关（因为 $w$ 为外生解释变量）。因此，有时也称 $z_1, z_2, z_3$ 为方程外的工具变量。将方程(10.19)与(10.20)的拟合值分别记为 $\hat{x}_1$ 与 $\hat{x}_2$，并代入原方程(10.18)，进行第二阶段回归：

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \beta_3 w + \xi \tag{10.21}$$

其中，$\xi$ 为第二阶段回归的扰动项。记此估计量为 $\hat{\boldsymbol{\beta}}_{\text{IV}}$。可以证明，$\hat{\boldsymbol{\beta}}_{\text{IV}}$ 为真实参数 $\boldsymbol{\beta}$ 的一致估计，且服从渐近正态分布，可以照常进行大样本统计推断。由于 2SLS 的第二阶段回归就是 OLS，故 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的协方差矩阵 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}})$ 在形式上与 OLS 估计量相似。当然，考虑到可能存在异方差，建议使用异方差稳健的标准误。

需要注意的是，第二阶段回归所得残差为

$$\hat{\xi} \equiv y - (\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_1 + \hat{\beta}_2 \hat{x}_2 + \hat{\beta}_3 w) \tag{10.22}$$

而原方程真正的残差却是

$$e_{\text{IV}} \equiv y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 w) \tag{10.23}$$

一般来说，二者并不相等，即 $e_{\text{IV}} \neq \hat{\xi}$。因此，进行 2SLS 估计时，最好不要自己手工进行两次回归，而直接使用 Stata 命令（相信 Stata 会计算正确的残差！）。

2SLS 的 Stata 命令格式为

```
ivregress 2sls y x1 x2(x3 = z1 z2), robust first
```

其中，"y"为被解释变量，"x1 x2"为外生解释变量，"x3"为内生解释变量，而"z1 z2"为方程外的工具变量。选择项"robust"表示使用异方差稳健的标准误（默认为普通标准误）；选择项"first"表示显示第一阶段的回归结果。

可以证明，在球形扰动项的情况下，2SLS 是最有效率的工具变量法。然而，在异方差的情况下，则存在更有效率的工具变量法，即"广义矩估计"（Generalized Method of Moments，GMM），是数理统计中"矩估计"（Method of Moments，MM）的推广。直观上，GMM 之于 2SLS，正如 GLS 与 OLS 的关系。而且，在恰好识别的情况，GMM 就等价于 2SLS[①]。

## 10.5 弱工具变量

上文提及，如果工具变量与内生解释变量仅微弱地相关，则工具变量法估计量 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的方差将变得很大。直观上，由于工具变量仅包含极少与内生解释变量有关的信息，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，即 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的小样本真实分布离大样本的渐近正态分布相去甚远，致使基于大样本理论的统计推断失效（我们不知道 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的小样本真实分布）。

为了检验是否存在弱工具变量，可在第一阶段回归中，检验所有方程外的工具变量（不含外

① 有关 GMM 估计的介绍，参见陈强(2014，p.146)。

生解释变量)的系数是否联合为零。比如,假设原模型为

$$y = \beta_0 + \beta_1 x + \beta_2 w + \varepsilon \tag{10.24}$$

其中,$x$ 为内生变量,而 $w$ 为外生解释变量。假设有两个有效工具变量 $z_1, z_2$,则第一阶段回归为

$$x = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 w + u \tag{10.25}$$

然后检验 $H_0: \alpha_1 = \alpha_2 = 0$,即工具变量 $z_1, z_2$ 的系数联合为 0。由于工具变量的强弱程度可连续变化,故很难确定一个明确的标准。一个经验规则(rule of thumb)是,如果此检验的 $F$ 统计量大于 10[①],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 Stata 中作完 2SLS 回归后,可使用以下命令检验弱工具变量:

```
estat firststage
```

此命令将根据第一阶段回归计算一些统计量,包括上文的 $F$ 统计量。

如果发现存在弱工具变量,可能的解决方法包括:

(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, LIML);在大样本下,LIML 与 2SLS 渐近等价,但在弱工具变量的情况下,LIML 的小样本性质可能优于 2SLS。有关最大似然估计,参见第 11 章。

LIML 的 Stata 命令为

```
ivregress liml y x1 x2(x3 = z1 z2)
```

此命令在格式上与"`ivregress 2sls`"(二段最小二乘法)完全相同。

## 10.6 对工具变量外生性的过度识别检验

工具变量的外生性是保证 2SLS 一致性的重要条件。如果所使用的"工具变量"与扰动项相关,则可能导致严重的偏差(参见习题)。

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,则其影响被解释变量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,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

① 由于技术性原因,此处应使用普通标准误,而非异方差稳健的标准误(Stock and Watson, 2012, p. 507)。如果此 $F$ 统计量大于 10,则大致可保证 IV 估计量的偏差($\hat{\beta}_{IV} - \beta$)不大于 OLS 偏差($\hat{\beta}_{OLS} - \beta$)的 10%,即 IV 估计量约可减少 OLS 估计量 90% 的偏差。

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，则认为至少某个工具变量不是外生的，与扰动项相关。

不失一般性，假设共有 $K$ 个解释变量$\{x_1,\cdots,x_K\}$，其中前$(K-r)$个解释变量$\{x_1,\cdots,x_{K-r}\}$为外生解释变量，而后 $r$ 个解释变量$\{x_{K-r+1},\cdots,x_K\}$为内生解释变量：

$$y = \underbrace{\beta_1 x_1 + \cdots + \beta_{K-r}x_{K-r}}_{\text{外生}} + \underbrace{\beta_{K-r+1}x_{K-r+1} + \cdots \beta_K x_K}_{\text{内生}} + \varepsilon \tag{10.26}$$

同时假设共有 $m$ 个方程外的工具变量$\{z_1,\cdots,z_m\}$，其中 $m>r$；则过度识别的原假设为

$$H_0:\text{Cov}(z_1,\varepsilon) = 0, \quad \cdots, \quad \text{Cov}(z_m,\varepsilon) = 0 \tag{10.27}$$

由于扰动项 $\varepsilon$ 无法观测，故只能通过 2SLS 的残差 $e_{\text{IV}}$ 来考察工具变量与扰动项的相关性。为此，把 2SLS 的残差 $e_{\text{IV}}$ 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[①]：

$$e_{\text{IV}} = \gamma_1 x_1 + \cdots + \gamma_{K-r}x_{K-r} + \delta_1 z_1 + \cdots + \delta_m z_m + error \tag{10.28}$$

则原假设(10.27)可写为

$$H_0:\delta_1 = \cdots = \delta_m = 0 \tag{10.29}$$

记辅助回归(10.28)的可决系数为 $R^2$，则 Sargan 统计量(Sargan，1958)为

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

其中，Sargan 统计量的渐近分布为$\chi^2(m-r)$，其自由度$(m-r)$是过度识别约束的个数，即方程外工具变量个数$(m)$，减去内生解释变量个数$(r)$，也就是“多余”的工具变量个数。显然，如果恰好识别，则 $m-r=0$（自由度为 0），$\chi^2(0)$无定义，故无法使用此“过度识别检验”。

此检验背后的直观思想是，在过度识别的情况下，可以使用不同的工具变量组合来进行工具变量法估计；而如果所有工具变量都有效，则这些工具变量估计量 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 都将收敛到相同的真实参数 $\boldsymbol{\beta}$。为此，可以检验不同的工具变量估计量之间的差是否收敛于 $\boldsymbol{0}$；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。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，也并不能证明这些工具变量的外生性。这是因为，过度识别检验成立的大前提是，该模型至少是恰好识别的。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。直观上，即使不同的工具变量估计量 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的概率极限都相同，并不能保证它们都收敛到真实的参数 $\boldsymbol{\beta}$；因为也可能都收敛到其他值，比如 $\boldsymbol{\beta}^* \neq \boldsymbol{\beta}$。而恰好识别的大前提则保证了，在这些工具变量估计量中至少有一个估计量收敛到真实参数。此时，如果所有工具变量都外生，则所有工具变量估计量都会收敛到真实参数。

在 Stata 中作完 2SLS 估计后，可使用以下命令进行过度识别检验。

```
estat overid
```

① 其中，外生解释变量可视为自己的工具变量。

## 10.7 对解释变量内生性的豪斯曼检验：究竟该用 OLS 还是 IV

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。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 不一致，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 与工具变量法都一致，即在大样本下 $\hat{\boldsymbol{\beta}}_{IV}$ 与 $\hat{\boldsymbol{\beta}}_{OLS}$ 都收敛于真实的参数值 $\boldsymbol{\beta}$，因此 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$(称为“对比向量”，vector of contrast)依概率收敛于 $\boldsymbol{0}$。反之，如果 $H_0$ 不成立，则工具变量法一致而 OLS 不一致，故 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$ 不会收敛于 $\boldsymbol{0}$。豪斯曼检验正是基于这一思想进行的。如果 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$ 的距离很大，则倾向于拒绝原假设。根据沃尔德检验原理，以二次型来度量此距离可得

$$(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})'[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})]^{-1}(\hat{\boldsymbol{\beta}}_{IV}-\boldsymbol{\beta}_{OLS})\xrightarrow{d}\chi^2(r) \tag{10.31}$$

其中，$r$ 为内生解释变量的个数，$[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})]$ 为对比向量 $(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})$ 的协方差矩阵的样本估计值。如果豪斯曼统计量很大，超过了其渐近分布 $\chi^2(r)$ 的临界值，则可拒绝“所有解释变量均外生”的原假设，认为存在内生解释变量，故应使用 IV。

豪斯曼检验的 Stata 命令为

```
reg y x1 x2
estimates store ols                    (存储 OLS 的结果，记为 ols)
ivregress 2sls y x1(x2 = z1 z2)        (假设 x2 为内生变量，z1，z2 为 IV)
estimates store iv                     (存储 2SLS 的结果，记为 iv)
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率的估计量(即 OLS)所对应的残差来计算扰动项方差 $\hat{\sigma}^2$。这样有助于保证根据样本数据计算的 $[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})]$ 为正定矩阵，便于求其逆矩阵。选择项“constant”表示 $\hat{\boldsymbol{\beta}}_{IV}$ 与 $\hat{\boldsymbol{\beta}}_{OLS}$ 中都包括常数项(默认不含常数项)。

传统豪斯曼检验的缺点是，为了简化矩阵 $[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{IV}-\hat{\boldsymbol{\beta}}_{OLS})]$ 的计算，假设在 $H_0$ 成立的情况下，OLS 最有效率，故不适用异方差的情形(OLS 只在球形扰动项的情况下才最有效率)。而改进的“杜宾－吴－豪斯曼检验”(Durbin-Wu-Hausman Test，简记 DWH)即使在异方差的情况下也适用[①]。

① 有关 DWH 检验的详细说明，参见陈强(2014，p. 145)。

在 Stata 中作完 2SLS 估计后，可输入以下命令进行异方差稳健的 DWH 检验：

```
estat endogenous
```

## 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。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以下两步：

（i）列出与内生解释变量（$x$）相关的尽可能多的变量的清单（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）。

第（ii）步的操作有一定难度，因为扰动项不可观测。既然扰动项不可观测，那么如何判断某候选变量（$z$）是否与扰动项（$\varepsilon$）相关呢？由于扰动项是被解释变量（$y$）的扰动项，故可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，因为如果 $z$ 与 $\varepsilon$ 相关，则 $z$ 对 $y$ 的影响必然还有除 $x$ 以外的渠道，参见图 10.3。至于是否“$z$ 对 $y$ 的影响仅仅通过 $x$ 来起作用”，有时可以通过定性的讨论来确定。这就是上文提到的“排他性约束”（exclusion restriction）。

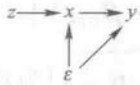



图 10.3 工具变量示意图

（箭头表示相关，无箭头表示不相关）

**例** 滞后变量。对于时间序列或面板数据，常使用内生解释变量的滞后作为工具变量。显然，内生解释变量与其滞后变量相关。另外，由于滞后变量已经发生，故为“前定”（从当期的角度看，其取值已经固定），可能与当期的扰动项不相关。比如，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，一方面，奖金占员工报酬比重越高，则越能促进生产率的提高；但另一方面，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。另外，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或许是外生的。

**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，就会出现内生性偏差。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（包括七种类型的犯罪）的工具变量。通常，在任市长在竞选连任时，为了拉选票，会增加警察人数以保证治安，故满足相关性。另外，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。

**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。从历史的角度,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十分不同。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。

**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,人们呆在室内的时间越长,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。

## 10.9 工具变量法的 Stata 实例

下面以数据集 grilic. dta 为例演示工具变量法,继续对教育投资回报率的探讨。此数据集的主要变量包括:ln *w*(工资对数),*s*(教育年限),*expr*(工龄),*tenure*(在现单位的工作年数),*iq*(智商),*med*(母亲的教育年限),*kww*(在"knowledge of the World of Work"测试中的成绩),*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)。

(1) 作为参照系,首先进行 OLS 回归,并使用稳健标准误。

```
. reg lnw s expr tenure rns smsa,r
```

其中,我们主要感兴趣的关键解释变量为 *s*(教育年限),而 *expr*,*tenure*,*rns*,*smsa* 为控制变量。

```
Linear regression                               Number of obs =     758
                                                F(  5,   752) =   84.05
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3521
                                                Root MSE      =  .34641
```

| lnw | Coef. | Robust 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 |

结果显示，教育投资的年回报率高达 10.26%，而且在 1% 的水平上显著不为 0。这意味着，多受一年教育，则未来工资将高出 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。

(2) 引入智商(*iq*)作为能力的代理变量(*proxy*)，再进行 OLS 回归。

```
. reg lnw s iq expr tenure rns smsa,r
```

```
Linear regression                               Number of obs =     758
                                                F(  6,   751) =   71.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3600
                                                Root MSE      =  .34454

------------------------------------------------------------------------------
             |               Robust
         lnw |      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%，更为合理些，但仍然显得过高。

(3) 由于用 *iq* 来度量能力存在测量误差，故 *iq* 是内生变量。为此，考虑使用变量(*med*, *kww*)作为 *iq* 的工具变量。显然，母亲的教育年限(*med*)与 KWW 测试成绩(*kww*)都与 *iq* 正相关；并假设 *med* 与 *kww* 为外生[①]。下面进行 2SLS 回归，使用稳健标准误，并显示第一阶段的回归结果。

```
. ivregress 2sls lnw s expr tenure rns smsa(iq = med kww),r first
```

① 比如，*med* 只通过 *iq* 影响子女工资 ln *w*，而不出现高学历母亲利用社会关系为子女找工作的情形。

```
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 | Coef. | Robust 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 |

| lnw | 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 |

```
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%，且在 1% 水平上显著，比较合理。

(4) 下面进行过度识别检验：

```
. estat overid
```

```
  Test of overidentifying restrictions:

  Score chi2(1)          =  .151451  (p = 0.6972)
```

由于 $p$ 值为 0.697，故接受原假设，认为(*med*，*kww*)外生，与扰动项不相关。

(5) 进一步考察有效工具变量的另一条件，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。

正式检验需计算第一阶段回归的普通(非稳健)$F$统计量,故首先使用普通标准误重新进行2SLS估计。

```
. quietly ivregress 2sls lnw s expr tenure rns smsa(iq = med kww)
. estat firststage
```

First-stage regression summary statistics

| Variable | R-sq. | Adjusted R-sq. | Partial R-sq. | F(2,750) | Prob > F |
|---|---|---|---|---|---|
| iq | 0.3066 | 0.3001 | 0.0382 | 14.9058 | 0.0000 |

Minimum eigenvalue statistic = 14.9058

Critical Values — # of endogenous regressors: 1
Ho: Instruments are weak — # of excluded instruments: 2

| | 5% | 10% | 20% | 30% |
|---|---|---|---|---|
| 2SLS relative bias | (not available) | | | |

| | 10% | 15% | 20% | 25% |
|---|---|---|---|---|
| 2SLS Size of nominal 5% Wald test | 19.93 | 11.59 | 8.75 | 7.25 |
| LIML Size of nominal 5% Wald test | 8.68 | 5.33 | 4.42 | 3.92 |

由于检验第一阶段回归的两个工具变量系数联合显著性的$F$统计量为14.91,超过10,故认为不存在弱工具变量。

(6) 为了稳健起见(也为了示范),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 lnw 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

| lnw | 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

LIML的系数估计值与2SLS非常接近,从侧面印证了"不存在弱工具变量"。

(7) 使用工具变量法的前提是存在内生解释变量。为此需进行豪斯曼检验,其原假设为"所有解释变量均为外生",即不存在内生变量。

```
. quietly reg lnw iq s expr tenure rns smsa
. estimates store ols
. quietly ivregress 2sls lnw s expr tenure rns smsa(iq = med kww)
. estimates store iv
. hausman iv ols,constant sigmamore
```

其中,由于传统的豪斯曼检验建立在同方差的前提下,故在上述回归中未使用稳健标准误(没有用选择项"r")。

```
Note: the rank of the differenced variance matrix (1) does not equal the number of
      coefficients being tested (7); be sure this is what you expect, or there may be
      problems computing the test.  Examine the output of your estimators for anything
      unexpected and possibly consider scaling your variables so that the coefficients
      are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       iv          ols         Difference          S.E.
-------------+------------------------------------------------------------
          iq |   .0139284     .0032792       .0106493        .0054318
           s |   .0607803     .0927874       -.032007        .0163254
        expr |   .0433237     .0393443       .0039794        .0020297
      tenure |   .0296442      .034209       -.0045648       .0023283
         rns |  -.0435271    -.0745325       .0310054        .0158145
        smsa |   .1272224     .1367369       -.0095145       .0048529
       _cons |   3.218043     3.895172       -.6771285       .3453751
------------------------------------------------------------------------------
                b = consistent under Ho and Ha; obtained from ivregress
     B = inconsistent under Ha, efficient under Ho; obtained from regress

  Test:  Ho:  difference in coefficients not systematic

                chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                        =         3.84
              Prob>chi2 =       0.0499
              (V_b-V_B is not positive definite)
```

上表显示,$p$ 值(Prob > chi2)为 0.049 9,故可在 5% 的显著性水平上拒绝"所有解释变量均为外生"的原假设,认为 $iq$ 为内生变量。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
```

```
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)
```

上表提供了一个 $F$ 统计量与一个 $\chi^2$ 统计量,二者在大样本下渐近等价。由于二者的 $p$ 值都小于 0.05,故认为 $iq$ 为内生解释变量。

(8) 汇报结果:如果希望将以上各种估计法的系数及标准误列在同一表格中(比如,用于论

文中),可使用以下命令:

```
. qui reg lnw s expr tenure rns smsa,r
. est sto ols_no_iq
. qui reg lnw iq s expr tenure rns smsa,r
. est sto ols_with_iq
. qui ivregress 2sls lnw s expr tenure rns smsa(iq=med kww),r
. est sto tsls
. qui ivregress liml lnw s expr tenure rns smsa(iq=med kww),r
. est sto liml
. estimates table ols_no_iq ols_with_iq tsls liml,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。

| Variable | ols_no_iq | ols_with~q | tsls | liml |
|---|---|---|---|---|
| s | .10264304 | .09278735 | .06078035 | .06063623 |
| | .00620988 | .00697626 | .01895051 | .01903397 |
| expr | .0381189 | .03934425 | .04332367 | .04334159 |
| | .00661439 | .00666033 | .00741179 | .0074185 |
| tenure | .03561456 | .03420896 | .02964421 | .02962365 |
| | .00799884 | .00789567 | .00831697 | .00832297 |
| rns | -.08407974 | -.07453249 | -.04352713 | -.04338751 |
| | .02953295 | .02997719 | .03447789 | .03452902 |
| smsa | .13966664 | .13673691 | .12722244 | .1271796 |
| | .02805598 | .02777116 | .02974144 | .02975994 |
| iq | | .00327916 | .01392844 | .01397639 |
| | | .00113212 | .00603931 | .00606812 |
| _cons | 4.103675 | 3.8951718 | 3.2180433 | 3.2149943 |
| | .08766646 | .11592863 | .39836829 | .40014925 |

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,而三颗星表示 1% 的显著性水平,可使用如下命令:

```
. estimates table ols_no_iq ols_with_iq tsls liml,star(0.1 0.05 0.01)
```

| Variable | ols_no_iq | ols_with_iq | tsls | liml |
|---|---|---|---|---|
| s | .10264304*** | .09278735*** | .06078035*** | .06063623*** |
| expr | .0381189*** | .03934425*** | .04332367*** | .04334159*** |
| tenure | .03561456*** | .03420896*** | .02964421*** | .02962365*** |
| rns | -.08407974*** | -.07453249** | -.04352713 | -.04338751 |
| smsa | .13966664*** | .13673691*** | .12722244*** | .1271796*** |
| iq | | .00327916*** | .01392844** | .01397639** |
| _cons | 4.103675*** | 3.8951718*** | 3.2180433*** | 3.2149943*** |

legend: * p<.1; ** p<.05; *** p<.01

但 Stata 官方命令"estimates table"无法同时显示回归系数、标准误与表示显著性的星号(在正式论文中通常需同时显示)。为此,下载非官方命令"estout"。

```
. ssc install estout
```

```
. esttab ols_no_iq ols_with_iq tsls liml,se r2 mtitle star( * 0.1 * *
0.05 * * * 0.01)
```

其中,选择项"se"表示在括弧中显示标准误(默认显示 $t$ 统计量,如果使用选择项"p"则显示 $p$ 值),选择项"r2"表示显示 $R^2$,选择项"mtitle"表示使用模型名称(model title)作为表中每列的标题(默认使用被解释变量作为标题)①,选择项"star( * 0.1 * * 0.05 * * * 0.01)"指定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 |
|---|---|---|---|---|
| s | 0.103*** | 0.0928*** | 0.0608*** | 0.0606*** |
| | (0.00621) | (0.00698) | (0.0190) | (0.0190) |
| expr | 0.0381*** | 0.0393*** | 0.0433*** | 0.0433*** |
| | (0.00661) | (0.00666) | (0.00741) | (0.00742) |
| tenure | 0.0356*** | 0.0342*** | 0.0296*** | 0.0296*** |
| | (0.00800) | (0.00790) | (0.00832) | (0.00832) |
| rns | -0.0841*** | -0.0745** | -0.0435 | -0.0434 |
| | (0.0295) | (0.0300) | (0.0345) | (0.0345) |
| smsa | 0.140*** | 0.137*** | 0.127*** | 0.127*** |
| | (0.0281) | (0.0278) | (0.0297) | (0.0298) |
| iq | | 0.00328*** | 0.0139** | 0.0140** |
| | | (0.00113) | (0.00604) | (0.00607) |
| _cons | 4.104*** | 3.895*** | 3.218*** | 3.215*** |
| | (0.0877) | (0.116) | (0.398) | (0.400) |
| N | 758 | 758 | 758 | 758 |
| R-sq | 0.352 | 0.360 | 0.278 | 0.277 |

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

如果要将上表输出到 Microsoft Word 文档,并以文件名 iv 来命名此文档,可运行如下命令:

```
. esttab ols_no_iq ols_with_iq tsls liml using iv.rtf,se r2 mtitle star
( * 0.1 * * 0.05 * * * 0.01)
```

```
(output written to iv.rtf)
```

其中,"iv.rtf"的扩展名"rtf"表示 rich text format。点击输出结果中的"iv.rtf"链接,即可打开此文件,然后在 Word 中继续编辑此文件。

**习题**

10.1 假设真实模型为 $y^* = \alpha + \beta x + \varepsilon$,其中 $\beta \neq 0$,而 $\mathrm{Cov}(x, \varepsilon) = 0$。$y^*$ 无法精确观测,但能观测到 $y$,二者满足 $y = y^* + v$,其中 $v$ 为测量误差。

① 此处由于被解释变量都是 $\ln w$,作为标题没有区别,故不使用此默认选项。

(1) 考虑回归模型 $y=\alpha+\beta x+u$,证明其扰动项 $u=\varepsilon+v$。

(2) 证明只要被解释变量的测量误差 $v$ 与解释变量 $x$ 不相关,则 OLS 为一致估计量。

(3) 被解释变量测量误差 $v$ 的存在,是否会增大扰动项 $u$ 的方差?

10.2 证明方程(10.17)存在严格多重共线性,即可以找到 $\hat{x}_1$ 与 $\hat{x}_2$ 的线性组合为常数。

10.3 如果"工具变量"与扰动项相关,$\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$是否为一致估计?(提示:根据 2SLS 的第二阶段回归进行说明。)

10.4 在方程(10.10)中,假设 $\mathrm{Cov}(u_t,z_t)\neq 0$(不满足外生性),而 $\mathrm{Cov}(p_t,z_t)\neq 0$(依然满足相关性)。

(1) 证明 $\hat{\beta}_{\mathrm{IV}}$不是 $\beta$ 的一致估计,即 $\operatorname{plim}_{n\to\infty}\hat{\beta}_{\mathrm{IV}}\neq\beta$。

(2) 计算大样本偏差 $\left(\operatorname{plim}_{n\to\infty}\hat{\beta}_{\mathrm{IV}}-\beta\right)$。在什么情况下,此偏差的绝对值会变大?

10.5 使用数据集 acemoglu.dta 复制 Acemoglu *et al.* (2001)的部分结果。该数据集包含 64 个曾为欧洲殖民地的国家,主要变量为 log *pgp*95(1995 年人均 GDP,购买力平价),*avexpr*(1985—1995 年间的平均产权保护程度,0 为最低,10 为最高),*lat_abst*(首都纬度的绝对值除以 90),以及 log *em*4(殖民者死亡率的对数)[①]。另外,变量 *shortnam* 以三个字母作为每个国家的简称。

(1) 为了直观地考察产权保护与经济发展的关系,将 log *pgp*95 与 *avexpr* 的散点图与线性拟合图画在一起,并为每个散点标注国家简称。

(2) 使用稳健标准误,把 log *pgp*95 对 *avexpr* 及 *lat_abst* 进行回归,评论变量系数的符号、统计显著性及经济意义。

(3) 由于 *avexpr* 可能为内生解释变量,使用 log *em*4 作为 *avexpr* 的工具变量,重新进行(2)的回归。工具变量回归的结果与 OLS 有何不同?

(4) log *em*4 是否为弱工具变量?

10.6[②] 生育行为如何影响劳动力供给?具体来说,如果妇女多生一位小孩,其劳动力供给将下降多少?本题使用来自美国 1980 年人口普查的数据集 fertility_small.dta 进行估计。此数据集包含美国 21~35 岁已婚且有两个或更多子女的妇女信息,主要变量为 *weeks*(1979 年的工作周数),*morekids*(是否有两个以上小孩),以及 *samesex*(头两个小孩是否性别相同)。

(1) 把 *weeks* 对虚拟变量 *morekids* 进行回归。有两个以上小孩的妇女是否比有两个小孩的妇女工作更少?少多少?此效应是否在统计上显著?

(2) 上面(1)的回归能否估计生育行为对劳动力供给的因果效应?为什么?

(3) 把 *morekids* 对 *samesex* 进行回归。如果头两个小孩性别相同,是否更可能生第三个小孩?此效应大吗?是否在统计上显著?

(4) 在 *weeks* 对 *morekids* 的回归中,能否将 *samesex* 作为有效工具变量?为什么?

(5) *samesex* 是否为弱工具变量?

(6) 以 *samesex* 为工具变量,把 *weeks* 对 *morekids* 进行回归。生育行为对劳动力供给的效应有多大?是否在统计上显著?

---

① 殖民者死亡率(settler mortality)为殖民者每年每千人的死亡人数。原文还有其他控制变量,在此从略。

② 此例来自 Stock and Watson(2012)。

(7) 数据集中还有其他控制变量，比如 *age*(年龄)，*black*(是否黑人)，*hispan*(是否拉丁裔)，以及 *othrace*(是否其他族裔)。在(6)的工具变量回归中加入这些控制变量(假定它们为外生)，回归结果是否有变化？为什么？

*To be, or not to be: that is the question.* ——*Hamlet* by William Shakespeare

# 11. 二值选择模型

## 11.1 二值选择模型

如果解释变量 $x$ 是离散的(比如,虚拟变量),这并不影响回归(参见第 9 章)。但有时被解释变量 $y$ 是离散的,而非连续的,称为"离散选择模型"(discrete choice model)或"定性反应模型"(qualitative response model)。

最常见的离散选择模型是二值选择行为(binary choices),因为人生充满了选择。比如: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。此时,由于被解释变量为虚拟变量,取值为 0 或 1,故通常不宜进行 OLS 回归。

假设个体只有两种选择,比如 $y=1$(考研)或 $y=0$(不考研)。最简单的建模方法为"线性概率模型"(Linear Probability Model, LPM):

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{11.1}$$

其中,解释变量 $\boldsymbol{x}_i \equiv (x_{i1}\ x_{i2} \cdots\ x_{iK})'$,而参数 $\boldsymbol{\beta} \equiv (\beta_1\ \beta_2 \cdots\ \beta_K)'$。线性概率模型的优点是,计算方便($y$ 为虚拟变量并不影响 OLS 估计),且容易得到边际效应(即回归系数)。其缺点是,虽然明知被解释变量 $y$ 的取值非 0 即 1,但根据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,故一般只将 LPM 作为粗略的参考。

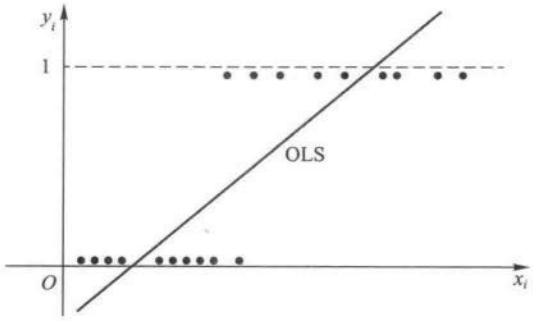



图 11.1 线性概率模型

为使 $y$ 的预测值总是介于[0,1]之间,在给定 $\boldsymbol{x}$ 的情况下,考虑 $y$ 的两点分布概率:

$$\begin{cases} P(y = 1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) \\ P(y = 0 \mid \boldsymbol{x}) = 1 - F(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases} \tag{11.2}$$

其中，函数 $F(\boldsymbol{x},\boldsymbol{\beta})$ 称为“连接函数”(link function)，因为它将解释变量 $\boldsymbol{x}$ 与被解释变量 $y$ 连接起来。由于 $y$ 的取值要么为 0，要么为 1，故 $y$ 肯定服从两点分布。连接函数的选择具有一定的灵活性。通过选择合适的连接函数 $F(\boldsymbol{x},\boldsymbol{\beta})$(比如，某随机变量的累积分布函数)，可以保证 $0\leqslant\hat{y}\leqslant1$，并将 $\hat{y}$ 理解为“$y=1$”发生的概率，因为

$$\mathrm{E}(y\mid\boldsymbol{x}) = 1\cdot P(y=1\mid\boldsymbol{x}) + 0\cdot P(y=0\mid\boldsymbol{x}) = P(y=1\mid\boldsymbol{x}) \tag{11.3}$$

上式之所以成立，正是由于被解释变量 $y$ 作为虚拟变量的两点分布特性。如果 $F(\boldsymbol{x},\boldsymbol{\beta})$ 为标准正态的累积分布函数(cdf)，则

$$P(y=1\mid\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}}\phi(t)\,\mathrm{d}t \tag{11.4}$$

其中，$\phi(\cdot)$ 与 $\Phi(\cdot)$ 分别为标准正态的密度函数与累积分布函数；此模型称为“Probit”。如果 $F(\boldsymbol{x},\boldsymbol{\beta})$ 为“逻辑分布”(logistic distribution)的累积分布函数，则

$$P(y=1\mid\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \tag{11.5}$$

其中，函数 $\Lambda(\cdot)$ 的定义为 $\Lambda(z)\equiv\dfrac{\exp(z)}{1+\exp(z)}$，此模型称为“Logit”。逻辑分布的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)，具有厚尾(fat tails)。这意味着，相对于标准正态的累积分布函数而言，逻辑分布的累积分布函数趋向于 0 或 1 的速度更慢，参见图 11.2。在实践中，Probit 与 Logit 都很常用，二者的估计结果(比如边际效应)也通常很接近。Logit 模型的优势在于，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 更为方便，而且 Logit 的回归系数更容易解释其经济意义。

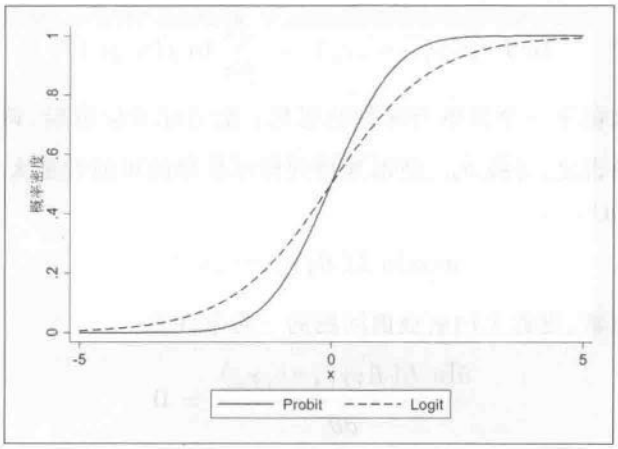



图 11.2 标准正态分布与逻辑分布的累积分布函数①

① 生成此图的 Stata 命令为“twoway function Probit = normal(x),range(-5 5) || function Logit = exp(x)/(1+exp(x)),range(-5 5) lpattern(dash) ytitle(概率密度)”。

## 11.2 最大似然估计的原理

Probit 与 Logit 模型在本质上都是非线性模型，无法通过变量转换而变为线性模型。对于非线性模型，常使用最大似然估计法（Maximum Likelihood Estimation，MLE 或 ML）。下面首先回顾概率统计中的最大似然估计法，然后再应用于 Probit 与 Logit 模型。

假设随机变量 $y$ 的概率密度函数为 $f(y;\theta)$，其中 $\theta$ 为未知参数。为了估计 $\theta$，从 $y$ 的总体中抽取样本容量为 $n$ 的随机样本 $\{y_1,\cdots,y_n\}$。假设 $\{y_1,\cdots,y_n\}$ 为 iid，则样本数据的联合密度函数为

$$f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta) \tag{11.6}$$

其中，$\prod_{i=1}^{n}$ 表示连乘。在抽样之前，$\{y_1,\cdots,y_n\}$ 为随机向量。抽样之后，$\{y_1,\cdots,y_n\}$ 就有了特定的样本值。因此，可将样本的联合密度函数视为在给定 $\{y_1,\cdots,y_n\}$ 的情况下，未知参数 $\theta$ 的函数。定义**似然函数**（likelihood function）为

$$L(\theta;y_1,\cdots,y_n) = \prod_{i=1}^{n} f(y_i;\theta) \tag{11.7}$$

由此可知，似然函数与联合密度函数完全相等，只是 $\theta$ 与 $\{y_1,\cdots,y_n\}$ 的角色互换，即把 $\theta$ 作为自变量，而视 $\{y_1,\cdots,y_n\}$ 为给定。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：

$$\ln L(\theta;y_1,\cdots,y_n) = \sum_{i=1}^{n} \ln f(y_i;\theta) \tag{11.8}$$

最大似然估计法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 $\theta$ 为何值的总体。换言之，寻找 $\hat{\theta}_{ML}$，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）：

$$\max_{\theta} \ln L(\theta;y_1,\cdots,y_n) \tag{11.9}$$

假设存在唯一内点解，则此无约束极值问题的一阶条件为

$$\frac{\partial \ln L(\theta;y_1,\cdots,y_n)}{\partial \theta} = 0 \tag{11.10}$$

求解此一阶条件，即可得到最大似然估计量 $\hat{\theta}_{ML}$。

**例** 假设 $y \sim N(\mu,\sigma^2)$，其中 $\sigma^2$ 已知，得到一个样本容量为 1 的样本 $y_1=2$，求对 $\mu$ 的最大似然估计。根据正态分布的密度函数可知，此样本的似然函数为

$$L(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(2-\mu)^2}{2\sigma^2}\right\} \tag{11.11}$$

显然，此似然函数在 $\hat{\mu}=2$ 处取最大值，参见图 11.3。

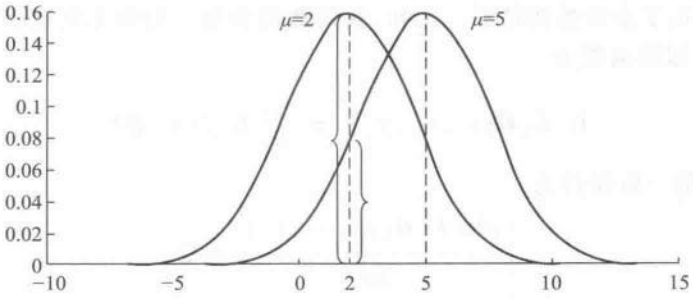



图 11.3 选择参数使观测到样本的可能性最大

**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。

可以证明，在一定的正则条件（regularity conditions）下，MLE 估计量具有以下良好的大样本性质，可照常进行大样本统计推断。

（1）$\hat{\theta}_{\mathrm{ML}}$ 为一致估计，即 $\operatorname{plim}_{n\to\infty}\hat{\theta}_{\mathrm{ML}}=\theta$。

（2）$\hat{\theta}_{\mathrm{ML}}$ 服从渐近正态分布。

（3）在大样本下，$\hat{\theta}_{\mathrm{ML}}$ 是最有效率的估计（渐近方差最小）。

由于模型存在非线性，故最大似然估计通常没有解析解，而只能寻找“数值解”（numerical solution）。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）。

MLE 的一阶条件可以归结为求非线性方程 $f(x)=0$ 的解。假设 $f(x)$ 的导数 $f'(x)$ 处处存在，参见图 11.4。记该方程的解为 $x^*$，满足 $f(x^*)=0$。首先猜一个初始值 $x_0$，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 $(x_1,f(x_1))$ 处再作一条切线，记此切线与横轴的交点为 $x_2$。以此类推，不断迭代，可得到序列 $\{x_0,x_1,x_2,x_3,\cdots\}$。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）。

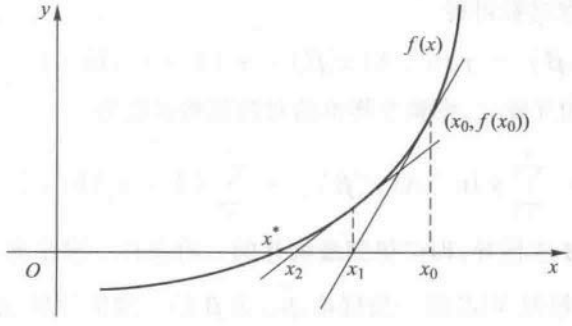



图 11.4 高斯 - 牛顿法

高斯 - 牛顿法之所以常用，原因之一是它的收敛速度很快，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 maxi-

mum)。

MLE 很容易应用于多参数的情形。比如,假设随机变量 $y$ 的概率密度函数为 $f(y;\boldsymbol{\theta})$,其中 $\boldsymbol{\theta}=(\theta_1\theta_2)'$,则对数似然函数为

$$\ln L(\boldsymbol{\theta};y_1,\cdots,y_n) = \sum_{i=1}^{n}\ln f(y_i;\boldsymbol{\theta}) \tag{11.12}$$

此最大化问题的一阶条件为

$$\begin{cases}\dfrac{\partial \ln L(\boldsymbol{\theta};y_1,\cdots,y_n)}{\partial\theta_1}=0\\[2ex] \dfrac{\partial \ln L(\boldsymbol{\theta};y_1,\cdots,y_n)}{\partial\theta_2}=0\end{cases} \tag{11.13}$$

求解联立方程组(11.13),即可得到最大似然估计量 $\hat{\theta}_{1,\mathrm{ML}}$ 与 $\hat{\theta}_{2,\mathrm{ML}}$。高斯 - 牛顿法也适用于多元函数 $f(\boldsymbol{x})=0$ 的情形,只要在上述迭代过程中,将切线替换为(超)切平面即可。

## 11.3 二值选择模型的 MLE 估计

下面以 Logit 模型为例,将 MLE 应用于二值选择模型。对于样本数据 $\{\boldsymbol{x}_i,y_i\}_{i=1}^{n}$,根据方程(11.5),第 $i$ 个观测数据的概率密度为

$$f(y_i\mid\boldsymbol{x}_i,\boldsymbol{\beta}) = \begin{cases}\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & 若\ y_i=1\\ 1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & 若\ y_i=0\end{cases} \tag{11.14}$$

其中,$\Lambda(z)\equiv\dfrac{\exp(z)}{1+\exp(z)}$为逻辑分布的累积分布函数。上式可紧凑地写为

$$f(y_i\mid\boldsymbol{x}_i,\boldsymbol{\beta}) = [\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]^{1-y_i} \tag{11.15}$$

显然,如果 $y_i=1$,则 $1-y_i=0$,故上式等于$[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]$;反之,如果 $y_i=0$,则 $1-y_i=1$,故上式等于$[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]$。将上式取对数可得

$$\ln f(y_i\mid\boldsymbol{x}_i,\boldsymbol{\beta}) = y_i\ln[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]+(1-y_i)\ln[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] \tag{11.16}$$

假设样本中的个体相互独立,则整个样本的对数似然函数为

$$\ln L(\boldsymbol{\beta}\mid\boldsymbol{y},\boldsymbol{x}) = \sum_{i=1}^{n}y_i\ln[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})]+\sum_{i=1}^{n}(1-y_i)\ln[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})] \tag{11.17}$$

把对数似然函数对 $\boldsymbol{\beta}$ 求偏导,即可得到最大化的一阶条件。满足此一阶条件的估计量即为 MLE 估计量,记为 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$。根据 MLE 的一般理论,$\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ 为 $\boldsymbol{\beta}$ 的一致估计量,服从渐近正态分布,且在大样本下具有最小渐近方差。

## 11.4 边际效应

对于线性模型,回归系数 $\beta_k$ 的经济意义十分明显,就是解释变量 $x_k$ 对被解释变量 $y$ 的边际

效应(marginal effects)。然而,在非线性模型中,估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ 一般并非边际效应。以 Probit 为例,计算解释变量 $x_k$ 的边际效应:

$$\frac{\partial P(y=1\mid \boldsymbol{x})}{\partial x_k}=\frac{\partial \Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k}=\frac{\partial \Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\cdot\frac{\partial(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k}=\phi(\boldsymbol{x}'\boldsymbol{\beta})\cdot\beta_k \tag{11.18}$$

其中,使用了微分的链锁法则(chain rule),而且假定 $x_k$ 为连续变量。需要注意的是,由于 Probit 与 Logit 所使用的分布函数不同,故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,由表达式(11.18)可知,对于非线性模型而言,边际效应通常不是常数,它随着解释向量 $\boldsymbol{x}$ 而变。

由于非线性模型的边际效应一般不是常数,故存在不同的边际效应概念。常用的边际效应概念包括:

(1) 平均边际效应(average marginal effect),即分别计算在每个样本观测值上的边际效应,然后进行简单算术平均。

(2) 样本均值处的边际效应(marginal effect at mean),即计算在 $\boldsymbol{x}=\bar{\boldsymbol{x}}$ 处的边际效应。

(3) 在某代表值处的边际效应(marginal effect at a representative value),即给定 $\boldsymbol{x}^*$,计算在 $\boldsymbol{x}=\boldsymbol{x}^*$ 处的边际效应。

以上三种边际效应的计算结果可能有较大差异。传统上,常计算样本均值处 $\boldsymbol{x}=\bar{\boldsymbol{x}}$ 的边际效应,因为计算方便。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。

## 11.5 回归系数的经济意义

既然 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ 并非边际效应,那么它究竟有什么含义?对于 Logit 模型,记事件发生的概率为 $p\equiv P(y=1\mid \boldsymbol{x})$,则事件不发生的概率为 $1-p=P(y=0\mid \boldsymbol{x})$。由于 $p=\dfrac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$,$1-p=\dfrac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$,故事件发生与不发生的几率比为

$$\frac{p}{1-p}=\exp(\boldsymbol{x}'\boldsymbol{\beta}) \tag{11.19}$$

其中,$\dfrac{p}{1-p}$ 称为"几率比"(odds ratio)或"相对风险"(relative risk)。例如,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(11.19)两边取对数可得

$$\ln\left(\frac{p}{1-p}\right)=\boldsymbol{x}'\boldsymbol{\beta}=\beta_1 x_1+\cdots+\beta_K x_K \tag{11.20}$$

其中，$\ln\left(\frac{p}{1-p}\right)$称为“对数几率比”(log-odds ratio)，而上式右边为线性函数。由此可知，回归系数$\hat{\beta}_j$表示解释变量$x_j$增加一个微小量引起对数几率比的边际变化。但“对数几率比”并不易直观理解。由于取对数意味着百分比的变化，故可把$\hat{\beta}_j$视为半弹性(semi-elasticity)，即$x_j$增加1单位引起几率比$\left(\frac{p}{1-p}\right)$的变化百分比。比如，$\hat{\beta}_j=0.12$，意味着$x_j$增加1单位引起几率比增加12%。

以上解释隐含地假设$x_j$为连续变量。如果$x_j$为离散变量(比如，性别、子女数)，则可使用另一解释方法。假设$x_j$增加1单位，从$x_j$变为$x_j+1$，记几率比$p$的新值为$p^*$，则新几率比与原几率比的比率可写为(此处无法使用微积分)

$$\frac{\frac{p^*}{1-p^*}}{\frac{p}{1-p}}=\frac{\exp[\beta_1+\beta_2x_2+\cdots+\beta_j(x_j+1)+\cdots+\beta_Kx_K]}{\exp(\beta_1+\beta_2x_2+\cdots+\beta_jx_j+\cdots+\beta_Kx_K)}=\exp(\beta_j) \tag{11.21}$$

为此，有些研究者偏好计算$\exp(\hat{\beta}_j)$，它表示解释变量$x_j$增加1单位引起几率比的变化倍数。正因为如此，Stata也称$\exp(\hat{\beta}_j)$为几率比(odds ratio)。

**例** $\hat{\beta}_j=0.12$，则$\exp(\hat{\beta}_j)=e^{0.12}=1.13$，故当$x_j$增加1单位时，新几率比是原几率比的1.13倍，或增加13%，因为$\exp(\hat{\beta}_j)-1=1.13-1=0.13$。

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。如果$x_j$至少必须变化1单位(比如性别、婚否等虚拟变量，以及年龄，子女个数等)，则应使用$\exp(\hat{\beta}_j)$。需要指出的是，对于Probit模型，无法对其系数$\hat{\boldsymbol{\beta}}_{ML}$进行类似的解释。这是Probit模型的劣势。

## 11.6 拟合优度

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$R^2$”或“伪$R^2$”(Pseudo $R^2$)，由McFadden(1974)所提出，其定义为

$$\text{准 } R^2\equiv\frac{\ln L_0-\ln L_1}{\ln L_0} \tag{11.22}$$

其中，$\ln L_1$为原模型的对数似然函数的最大值，而$\ln L_0$为以常数项为唯一解释变量的对数似然函数的最大值。

由于$y$为离散的两点分布，似然函数的最大可能值为1(即取值概率为1)，故对数似然函数的最大可能值为0，记为$\ln L_{max}$。显然，$0\geqslant\ln L_1\geqslant\ln L_0$，而$0\leqslant$准$R^2\leqslant1$，参见图11.5。由于$\ln L_{max}=0$，故可将“准$R^2$”写为

$$\text{准 } R^2=\frac{\ln L_1-\ln L_0}{\ln L_{max}-\ln L_0} \tag{11.23}$$

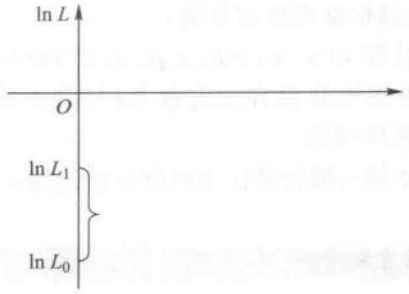



图 11.5 准 $R^2$ 的计算

其中，分子为加入解释变量后，对数似然函数的实际增加值（$\ln L_1 - \ln L_0$）；而分母为对数似然函数的最大可能增加值（$\ln L_{max} - \ln L_0$）。准 $R^2$ 即为前者占后者的比重。

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，即可计算正确预测的百分比。

## 11.7 准最大似然估计

使用 MLE 的前提是对总体的分布函数作具体的假定。比如，Probit 与 Logit 模型分别假设被解释变量 $y$ 的两点分布概率由标准正态或逻辑分布的累积分布函数所给出；但此分布函数的设定可能不正确，即存在“设定误差”（specification error）。

**定义** 使用不正确的分布函数所得到的最大似然估计量，称为“准最大似然估计”（Quasi MLE，QMLE）或“伪最大似然估计”（Pseudo MLE）。

准最大似然估计是否一定不一致？不一定！例如，假设线性模型的扰动项服从正态分布，则 MLE 估计量与 OLS 估计量完全相同[①]，而 OLS 估计量的一致性并不依赖于关于分布函数的具体假设。关于 QMLE 估计量的标准误可分为以下两种情况考虑。

（1）如果 QMLE 为一致估计量，考虑到可能存在对分布函数的设定误差，故应使用稳健标准误（robust standard errors），即相对于模型设定稳健的标准误。此稳健标准误与异方差稳健的标准误是一致的，因为扰动项方差是否相同也是一种模型设定。

（2）如果 QMLE 估计量不一致，则即使采用稳健标准误也无济于事。此时，QMLE 估计量 $\hat{\boldsymbol{\beta}}_{QML} \xrightarrow{p} \boldsymbol{\beta}^* \neq \boldsymbol{\beta}$，故首先应担心估计量的一致性。稳健标准误只是更精确估计了一个错误的“准真实参数”（pseudo true parameter）$\boldsymbol{\beta}^*$，而且通常不知道 $\boldsymbol{\beta}^*$ 的经济意义。换言之，在这种情况下，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）。

具体到二值选择模型（Probit 或 Logit 模型），可以证明，只要条件期望函数 $E(y \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$ 设定正确，则 MLE 估计就是一致的。由于两点分布的特殊性，在 iid 的情况下，只要 $E(y \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$ 成立，则稳健标准误就等于 MLE 的普通标准误。因此，如果认为模型设定正确，就没有

① 证明参见陈强（2014，p.68）。

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 $E(y \mid \boldsymbol{x}) \neq F(\boldsymbol{x},\boldsymbol{\beta})$),则 Probit 与 Logit 模型并不能得到对系数 $\boldsymbol{\beta}$ 的一致估计,使用稳健标准误也就没有太大意义(只是更精确地估计了错误参数的标准误),首先应解决参数估计的一致性问题。

因此,对于二值选择模型,使用普通标准误或稳健标准误都可以(文献中尚无定论)。

## 11.8 三类渐近等价的大样本检验

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。考虑以下线性回归模型:

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{11.24}$$

其中,解释变量 $\boldsymbol{x} \equiv (x_1\ x_2 \cdots\ x_K)'$,而参数 $\boldsymbol{\beta} \equiv (\beta_1\ \beta_2 \cdots\ \beta_K)'$。考虑检验以下原假设[①]:

$$H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0 \tag{11.25}$$

其中,$\boldsymbol{\beta}_0$ 已知,共有 $K$ 个约束。

(1) 沃尔德检验(Wald Test):沃尔德检验通过考察 $\boldsymbol{\beta}$ 的无约束估计量 $\hat{\boldsymbol{\beta}}$ 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 $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ 的绝对值不应该很大。由于 $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ 为多维向量,故使用以下二次型来衡量此距离:

$$W \equiv (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'[\mathrm{Var}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(K) \tag{11.26}$$

在大样本下,此 Wald 统计量服从渐近 $\chi^2(K)$ 分布,其中 $K$ 为约束条件的个数(在此为解释变量个数)。第 5~6 章所介绍的单一系数 $t$ 检验(在大样本下使用标准正态进行检验)、联合线性假设的 $F$ 检验(大样本下可使用 $\chi^2$ 分布进行检验)都是 Wald 检验。

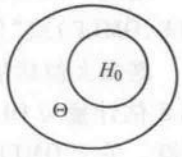


图 11.6 无约束与有约束的参数空间

(2) 似然比检验(Likelihood Ratio Test,$LR$):似然比检验通过比较无约束估计量 $\hat{\boldsymbol{\beta}}$ 与有约束估计量 $\hat{\boldsymbol{\beta}}^*$ 的差别来进行检验。一般来说,无约束的似然函数最大值 $\ln L(\hat{\boldsymbol{\beta}})$ 比有约束的似然函数最大值 $\ln L(\hat{\boldsymbol{\beta}}^*)$ 更大,因为在无约束条件下的参数空间 $\Theta$ 比有约束条件下(即 $H_0$ 成立时)参数的取值范围更大,参见图 11.6。

$LR$ 检验的基本思想是,如果 $H_0$ 正确,则 $[\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\hat{\boldsymbol{\beta}}^*)]$ 不应该很大。在这个简单例子中,有约束的估计量 $\hat{\boldsymbol{\beta}}^* = \boldsymbol{\beta}_0$。$LR$ 统计量为

$$LR \equiv -2\ln\left[\frac{L(\hat{\boldsymbol{\beta}}^*)}{L(\hat{\boldsymbol{\beta}})}\right] = 2[\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\hat{\boldsymbol{\beta}}^*)] \xrightarrow{d} \chi^2(K) \tag{11.27}$$

在大样本下,$LR$ 统计量也服从渐近 $\chi^2(K)$ 分布。第 5 章介绍的 $F$ 统计量的另一表达式 $F =$

① 对于一般的线性假设 $H_0: \boldsymbol{R\beta} = \boldsymbol{r}$,也有类似的结果。为了集中阐述这三类检验的思想,在此仅考虑最简单也最常见的情形 $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$。

$\frac{(\mathrm{SSR}^* - \mathrm{SSR})/(K-1)}{\mathrm{SSR}/(n-K)}$，就可看成是依据似然比原理而设计的[①]。在进行 Probit 或 Logit 回归时，Stata 会汇报一个似然比统计量，检验除常数项外所有参数的联合显著性，即考察原模型与只有常数项模型的似然函数最大值之比（准 $R^2$ 的计算也基于此）。

（3）拉格朗日乘子检验（Lagrange Multiplier Test，*LM*）：Wald 检验只考察无约束估计量 $\hat{\boldsymbol{\beta}}$，*LR* 检验同时考察无约束估计量 $\hat{\boldsymbol{\beta}}$ 与有约束估计量 $\hat{\boldsymbol{\beta}}^*$；而 *LM* 检验则只考察有约束估计量 $\hat{\boldsymbol{\beta}}^*$。考虑以下有约束条件的对数似然函数最大化问题：

$$\begin{aligned} &\max_{\tilde{\boldsymbol{\beta}}} \ln L(\tilde{\boldsymbol{\beta}}) \\ &s.t.\ \tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 \end{aligned} \tag{11.28}$$

其中，$\tilde{\boldsymbol{\beta}}$ 为在最大化过程中参数 $\boldsymbol{\beta}$ 的假想取值（hypothetical value）。对于约束极值问题，可引入以下拉格朗日乘子函数：

$$\max_{\tilde{\boldsymbol{\beta}},\boldsymbol{\lambda}} \ln L(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\lambda}'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{11.29}$$

其中，$\boldsymbol{\lambda}$ 为拉格朗日乘子向量（Lagrange multiplier），其经济含义为约束条件（比如资源约束）的影子价格（shadow price）。特别地，如果 $\hat{\boldsymbol{\lambda}} = \mathbf{0}$，则此约束条件完全不起作用（可以无偿获取任意数量的资源）。根据一阶条件（对 $\tilde{\boldsymbol{\beta}}$ 求导）可知：

$$\hat{\boldsymbol{\lambda}} = \frac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\boldsymbol{\beta}}} \equiv \begin{pmatrix} \dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\beta}_1} \\ \vdots \\ \dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\beta}_K} \end{pmatrix} \tag{11.30}$$

上式表明，最优的拉格朗日乘子向量 $\hat{\boldsymbol{\lambda}}$ 等于对数似然函数在约束估计量 $\hat{\boldsymbol{\beta}}^*$ 处的一阶偏导数（切线的斜率）。一方面，如果 $\hat{\boldsymbol{\lambda}} \approx \mathbf{0}$，则说明此约束条件不“紧”（tight）或不是“硬约束”（binding constraint），加上这个约束条件并不会使似然函数的最大值下降很多，即原假设 $H_0$ 很可能成立。另一方面，如果原假设 $H_0$ 成立，则（$\hat{\boldsymbol{\lambda}} - \mathbf{0}$）的绝对值不应很大。以二次型来度量此距离，可得 *LM* 统计量：

$$LM \equiv \hat{\boldsymbol{\lambda}}'[\mathrm{Var}(\hat{\boldsymbol{\lambda}})]^{-1}\hat{\boldsymbol{\lambda}} \xrightarrow{d} \chi^2(K) \tag{11.31}$$

其中，*LM* 统计量也服从渐近 $\chi^2(K)$ 分布，而 $\mathrm{Var}(\hat{\boldsymbol{\lambda}})$ 为 $\hat{\boldsymbol{\lambda}}$ 的协方差矩阵。由于似然函数的一阶导数 $\hat{\boldsymbol{\lambda}} = \frac{\partial \ln L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 称为“得分函数”（score function）或“得分向量”（score vector），故此检验也称为“得

① 可以证明，似然函数是残差平方和的单调减函数。

分检验"(score test)。另一直观理解是,由于在无约束估计量 $\hat{\boldsymbol{\beta}}$ 处,$\frac{\partial \ln L(\hat{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}}}=\boldsymbol{0}$(MLE 的一阶条件),故如果原假设 $H_0$ 成立,则在约束估计量 $\hat{\boldsymbol{\beta}}^*$ 处,此得分向量也应该接近于 $\boldsymbol{0}$,即 $\frac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \widetilde{\boldsymbol{\beta}}}\approx \boldsymbol{0}$,而 *LM* 统计量反映的就是此接近程度。在第 7 ~ 8 章,对异方差与自相关所进行的 $nR^2$ 形式的检验都来自于 *LM* 检验的推导。

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,它们只是从不同的侧面去考察同一事物。可以把这三类统计检验的思想画在同一张图上,参见图 11.7。

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"哪种更方便。如果无约束估计更方便,则常使用 Wald 检验(比如,对线性回归系数的显著性检验);如果有约束估计更方便,则常使用 *LM* 检验(比如,对异方差、自相关的检验);如果二者都方便,则可使用 *LR* 检验(比如,对非线性回归方程的显著性检验)。

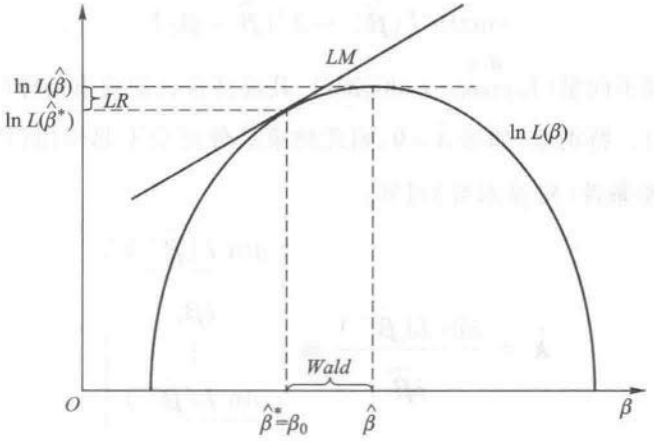



图 11.7 三类渐近等价的统计检验

## 11.9 二值选择模型的 Stata 命令与实例

二值选择模型的 Stata 命令为

```
probit y x1 x2 x3,r          (Probit 模型)
logit y x1 x2 x3,r or        (Logit 模型)
```

其中,选择项"r"表示使用稳健标准误(默认为普通标准误);选择项"or"表示显示几率比(odds ratio),而不显示回归系数。

完成 Probit 或 Logit 估计后,可进行预测,计算准确预测的百分比,或计算边际效应:

```
predict y1          (计算发生概率的预测值,记为 y1)
estat clas          (计算准确预测的百分比,clas 表示 classification)
```

```
margins,dydx( * )  (计算所有解释变量的平均边际效应;" * "代表所有解释变量)
margins,dydx( * ) atmeans     (计算所有解释变量在样本均值处的边际效应)
margins,dydx( * ) at(x1 =0)   (计算所有解释变量在 x1 =0 处的平均边际效应)
margins,dydx(x1)              (计算解释变量 x1 的平均边际效应)
margins,eyex( * )   (计算平均弹性,其中的两个"e"均指 elasticity)
margins,eydx( * )   (计算平均半弹性,x 变化一单位引起 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)
```

下面以数据集 titanic. dta 为例,演示二值选择模型的 Stata 估计。该数据集包括泰坦尼克号乘客的存活数据。泰坦尼克号邮轮是当时最大的客运轮船(参见图 11.8),在从英国南安普顿开往美国纽约的处女航途中于 1912 年 4 月 14 日撞冰山沉没。泰坦尼克号海难是和平时期死亡人数最多的海难之一(约死亡 1 500 人,专家意见不一),也是最广为人知的海难之一(1997 年同名好莱坞电影热映)。此数据集由 Dawson(1995)提供,原始数据来自英国贸易委员会(British Board of Trade)在沉船之后的调查。该数据集的被解释变量为 *survive*(存活 =1,死亡 =0);解释变量包括 *child*(儿童 =1,成年 =0),*female*(女性 =1,男性 =0),*class*1(头等舱 =1,其他 =0),*class*2(二等舱 =1,其他 =0),*class*3(三等舱 =1,其他 =0),*class*4(船员 =1,其他 =0)。在生死的紧要关头,各色人等的存活概率几何?

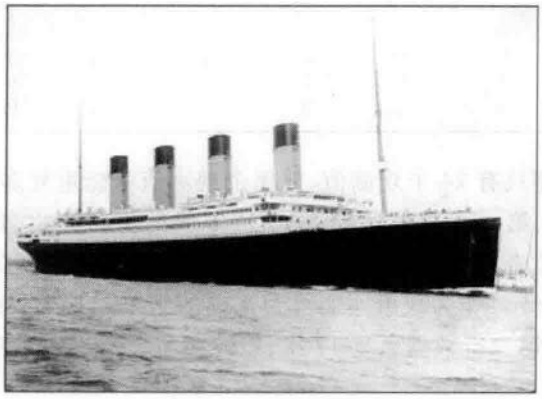


图 11.8 泰坦尼克号于 1914 年 4 月 10 日从英国南安普顿港出发

首先打开数据集,看一下原始数据。

```
. use titanic.dta,clear
. list
```

| | class1 | class2 | class3 | class4 | child | female | survive | freq |
|---|---|---|---|---|---|---|---|---|
| 1. | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 35 |
| 2. | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 17 |
| 3. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 118 |
| 4. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 154 |
| 5. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 387 |
| 6. | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 670 |
| 7. | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 4 |
| 8. | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 13 |
| 9. | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 89 |
| 10. | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 3 |
| 11. | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 5 |
| 12. | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 11 |
| 13. | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 13 |
| 14. | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 15. | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 13 |
| 16. | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 14 |
| 17. | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 57 |
| 18. | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 14 |
| 19. | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 75 |
| 20. | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 192 |
| 21. | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 140 |
| 22. | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 80 |
| 23. | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 76 |
| 24. | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 20 |

从上表可知，原始数据只有 24 个观测值，但每个观测值可能重复多次；其重复次数以最后一列变量 *freq* 来表示。比如，第一行数据显示，乘坐三等舱的男孩死亡者有 35 人；第二行数据显示，乘坐三等舱的女孩死亡者有 17 人；以此类推。对于这种观测值重复的数据，在进行计算与估计时，必须以重复次数（freq）作为权重才能得到正确的结果。其效果就相当于在数据文件中，将第一行数据重复 35 次，第二行数据重复 17 次，以此类推（完全不同于以方差倒数为权重的加权最小二乘法）。

假设观测值的重复次数记录于变量 *freq*，则在 Stata 中，可通过在命令的最后加上“[fweight = freq]”来实现此加权计算或估计；其中“fweight”指“frequency weight”（频数权重）。比如，首先考察各变量的统计特征。

```
. sum [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 2201 | .323035 | .4677422 | 0 | 1 |
| child | 2201 | .0495229 | .2170065 | 0 | 1 |
| female | 2201 | .2135393 | .4098983 | 0 | 1 |
| class1 | 2201 | .1476602 | .3548434 | 0 | 1 |
| class2 | 2201 | .1294866 | .335814 | 0 | 1 |
| class3 | 2201 | .3207633 | .466876 | 0 | 1 |

从上表可知，样本容量为 2201（旅客与船员总人数），而非 24。从变量 *survive* 的平均值可知，泰坦尼克号的平均存活率为 0.32。下面分别计算小孩、女士以及各等舱旅客的存活率。

```
. sum survive if child [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 109 | .5229358 | .5017807 | 0 | 1 |

```
. sum survive if female [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 470 | .7319149 | .4434342 | 0 | 1 |

```
. sum survive if class1 [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 325 | .6246154 | .4849687 | 0 | 1 |

```
. sum survive if class2 [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 285 | .4140351 | .493421 | 0 | 1 |

```
. sum survive if class3 [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 706 | .2521246 | .4345403 | 0 | 1 |

```
. sum survive if class4 [fweight = freq]
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| survive | 885 | .239548 | .427049 | 0 | 1 |

从以上结果可知，小孩、女士、一等舱、二等舱的存活率分别为 0.52、0.73、0.62、0.41，高于平均存活率；而三等舱、船员的存活率分别为 0.25、0.24，低于平均存活率。

下面进行更深入的回归分析。作为参照系，首先使用 OLS 估计线性概率模型。

```
. reg survive child female class1 class2 class3 [fweight = freq],r
```

Linear regression

| | |
|---|---|
| Number of obs | = 2201 |
| F( 5, 2195) | = 221.66 |
| Prob > F | = 0.0000 |
| R-squared | = 0.2529 |
| Root MSE | = .40474 |

| survive | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| child | .1812957 | .0479499 | 3.78 | 0.000 | .0872639 | .2753275 |
| female | .4906798 | .0239292 | 20.51 | 0.000 | .4437535 | .5376061 |
| class1 | .1755538 | .0291386 | 6.02 | 0.000 | .1184117 | .232696 |
| class2 | -.0105263 | .0258402 | -0.41 | 0.684 | -.0612 | .0401475 |
| class3 | -.1311806 | .0212996 | -6.16 | 0.000 | -.17295 | -.0894112 |
| _cons | .2267959 | .0139872 | 16.21 | 0.000 | .1993664 | .2542254 |

其中，由于所有乘客分为四类（*class*1—*class*4），故只能放入三个虚拟变量（*class*1—*class*3）；而将虚拟变量 *class*4（船员）作为参照类别，不放入回归方程。上表显示，儿童（*child*）、妇女（*female*）与头等舱旅客（*class*1）的存活几率比均显著地更高，三等舱旅客（*class*3）的存活几率比显著地更低，而二等舱旅客（*class*2）的存活几率比与船员无显著差异。

其次，使用 Logit 进行估计：

```
. logit survive child female class1 class2 class3 [fweight = freq],nolog
```

其中，选择项"nolog"表示不显示 MLE 数值计算的迭代过程。

```
Logistic regression                                     Number of obs   =       2201
                                                        LR chi2(5)      =     559.40
                                                        Prob > chi2     =     0.0000
Log likelihood = -1105.0306                             Pseudo R2       =     0.2020
```

| survive | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| child | 1.061542 | .2440257 | 4.35 | 0.000 | .5832608 | 1.539824 |
| female | 2.42006 | .1404101 | 17.24 | 0.000 | 2.144862 | 2.695259 |
| class1 | .8576762 | .1573389 | 5.45 | 0.000 | .5492976 | 1.166055 |
| class2 | -.1604188 | .1737865 | -0.92 | 0.356 | -.5010342 | .1801966 |
| class3 | -.9200861 | .1485865 | -6.19 | 0.000 | -1.21131 | -.6288619 |
| _cons | -1.233899 | .0804946 | -15.33 | 0.000 | -1.391666 | -1.076133 |

Logit 模型的估计结果在变量的显著性方面与 OLS 完全一致。上表显示，准 $R^2$ 为 0.20。检验整个方程显著性的 *LR* 统计量（LR chi2(5)）为 559.40，对应的 $p$ 值为 0.000，故整个方程的联合显著性很高。下面使用稳健标准误进行 Logit 估计。

```
. logit survive child female class1 class2 class3 [fweight = freq],nolog r
```

```
Logistic regression                                     Number of obs   =       2201
                                                        Wald chi2(5)    =     467.05
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1105.0306                       Pseudo R2       =     0.2020
```

| survive | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| child | 1.061542 | .2767452 | 3.84 | 0.000 | .5191318 | 1.603953 |
| female | 2.42006 | .1363096 | 17.75 | 0.000 | 2.152898 | 2.687222 |
| class1 | .8576762 | .1475218 | 5.81 | 0.000 | .5685387 | 1.146814 |
| class2 | -.1604188 | .1502193 | -1.07 | 0.286 | -.4548432 | .1340056 |
| class3 | -.9200861 | .1621035 | -5.68 | 0.000 | -1.237803 | -.602369 |
| _cons | -1.233899 | .0798876 | -15.45 | 0.000 | -1.390476 | -1.077322 |

对比以上两表可知，稳健标准误与普通标准误比较接近。由于此回归中的解释变量均为虚拟变量，只能变化一个单位（从 0 变为 1），为了便于解释回归结果，下面让 Stata 汇报几率比而非系数。

```
. logit survive child female class1 class2 class3 [fweight = freq],
or nolog
```

```
Logistic regression                                     Number of obs   =       2201
                                                        LR chi2(5)      =     559.40
                                                        Prob > chi2     =     0.0000
Log likelihood = -1105.0306                             Pseudo R2       =     0.2020

------------------------------------------------------------------------------
    survive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      child |   2.890826   .7054359     4.35   0.000     1.791872    4.663769
     female |   11.24654   1.579128    17.24   0.000     8.540859    14.80936
     class1 |   2.357675   .3709541     5.45   0.000     1.732036    3.209306
     class2 |    .851787   .1480291    -0.92   0.356     .6059037    1.197453
     class3 |   .3984847   .0592095    -6.19   0.000     .2978068    .5331983
      _cons |   .2911551   .0234364   -15.33   0.000     .2486608    .3409114
------------------------------------------------------------------------------
```

从上表可知，儿童的生存几率比是成年人的近 3 倍（几率比为 2.89），妇女的存活几率比是男人的 11 倍多（几率比为 11.25），头等舱旅客的存活几率比是船员的 2.36 倍，三等舱旅客的存活几率比只是船员的 39.8%，二等舱旅客的存活几率比也略低于船员（几率比为 0.85），但此差别在统计上不显著（$p$ 值为 0.356）。

为了与 OLS 估计的回归系数比较，下面计算 Logit 模型的平均边际效应：

```
. margins,dydx( * )
```

```
Average marginal effects                                Number of obs   =       2201
Model VCE    : OIM

Expression   : Pr(survive), predict()
dy/dx w.r.t. : child female class1 class2 class3

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       child |   .1732315   .0393799     4.40   0.000     .0960484    .2504147
      female |    .394926   .0171966    22.97   0.000     .3612214    .4286307
      class1 |   .1399629   .0250922     5.58   0.000     .0907831    .1891427
      class2 |  -.0261785   .0283616    -0.92   0.356    -.0817663    .0294093
      class3 |  -.1501475   .0238334    -6.30   0.000    -.1968602   -.1034348
------------------------------------------------------------------------------
```

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   =       2201
Model VCE    : OIM

Expression   : Pr(survive), predict()
dy/dx w.r.t. : child female class1 class2 class3
at           : child           =     .0495229 (mean)
               female          =     .2135393 (mean)
               class1          =     .1476602 (mean)
               class2          =     .1294866 (mean)
               class3          =     .3207633 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       child |   .2223422   .0510772     4.35   0.000     .1222328    .3224516
      female |   .5068865   .0303542    16.70   0.000     .4473934    .5663797
      class1 |    .179642   .0332374     5.40   0.000     .1144979    .2447861
      class2 |     -.0336   .0363774    -0.92   0.356    -.1048983    .0376983
      class3 |  -.1927139   .0308186    -6.25   0.000    -.2531173   -.1323105
------------------------------------------------------------------------------
```

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。下面计算 Logit 模型准确预测的比率：

```
. estat clas
```

```
Logistic model for survive

              -------- True --------
Classified |         D            ~D |      Total
-----------+--------------------------+-----------
         + |       349           126 |        475
         - |       362          1364 |       1726
-----------+--------------------------+-----------
     Total |       711          1490 |       2201

Classified + if predicted Pr(D) >= .5
True D defined as survive != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   49.09%
Specificity                     Pr( -|~D)   91.54%
Positive predictive value       Pr( D| +)   73.47%
Negative predictive value       Pr(~D| -)   79.03%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.46%
False - rate for true D         Pr( -| D)   50.91%
False + rate for classified +   Pr(~D| +)   26.53%
False - rate for classified -   Pr( D| -)   20.97%
--------------------------------------------------
Correctly classified                        77.83%
--------------------------------------------------
```

上表显示，正确预测的比率为(349 + 1364)/2201 = 77.83%。下面，根据 Logit 模型的回归结果，预测每位乘客的存活概率，并记为变量 *prob*。

```
. predict prob
```

```
(option pr assumed; Pr(survive))
```

由此,可以考察给定某种特征旅客的生存概率。比如,计算 Ms. Rose(头等舱、成年、女性)的存活概率:

```
. list prob survive freq if class1 ==1 & child ==0 & female ==1
```

| | prob | survive | freq |
|---|---|---|---|
| 7. | .8853235 | 0 | 4 |
| 21. | .8853235 | 1 | 140 |

从上表可知,Ms. Rose 的存活概率高达 88.5%。从频率上看,在所有头等舱的 144 位成年女性中,只有 4 位死亡。又比如,计算 Mr. Jack(三等舱、成年、男性)的存活概率:

```
. list prob survive freq if class3 ==1 & child ==0 & female ==0
```

| | prob | survive | freq |
|---|---|---|---|
| 5. | .1039594 | 0 | 387 |
| 19. | .1039594 | 1 | 75 |

从上表可知,Mr. Jack 的存活概率仅有 10.4%。从频率上看,在所有三等舱的 462 位成年男性中,只有 75 位生还。如此看来,Mr. Jack 与 Ms. Rose 生死相隔、阴阳两界实在是大概率事件。

类似地,可对此数据集进行 Probit 估计。

```
. probit survive child female class1 class2 class3 [fweight = freq],nolog
```

```
Probit regression                                 Number of obs   =       2201
                                                  LR chi2(5)      =     556.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -1106.3142                       Pseudo R2       =     0.2011
```

| survive | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| child | .5803382 | .1377535 | 4.21 | 0.000 | .3103463 | .85033 |
| female | 1.44973 | .0808635 | 17.93 | 0.000 | 1.29124 | 1.608219 |
| class1 | .5399101 | .0951552 | 5.67 | 0.000 | .3534092 | .7264109 |
| class2 | -.0898158 | .1028857 | -0.87 | 0.383 | -.2914681 | .1118364 |
| class3 | -.4875252 | .0800342 | -6.09 | 0.000 | -.6443893 | -.3306611 |
| _cons | -.7530486 | .0468804 | -16.06 | 0.000 | -.8449325 | -.6611648 |

由于 Probit 与 Logit 模型的回归系数并不直接可比,下面考察 Probit 模型的平均边际效应及预测准确度。

```
. margins,dydx( * )
```

```
Average marginal effects                        Number of obs   =       2201
Model VCE    : OIM

Expression   : Pr(survive), predict()
dy/dx w.r.t. : child female class1 class2 class3
```

| | dy/dx | Delta-method Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| child | .1640035 | .0386284 | 4.25 | 0.000 | .0882932 | .2397137 |
| female | .4096934 | .0177738 | 23.05 | 0.000 | .3748574 | .4445294 |
| class1 | .1525785 | .0262955 | 5.80 | 0.000 | .1010403 | .2041167 |
| class2 | -.0253819 | .0290666 | -0.87 | 0.383 | -.0823515 | .0315876 |
| class3 | -.1377745 | .0223131 | -6.17 | 0.000 | -.1815075 | -.0940416 |

```
. estat clas
```

```
Probit model for survive
```

| Classified | True D | ~D | Total |
|---|---|---|---|
| + | 349 | 126 | 475 |
| - | 362 | 1364 | 1726 |
| Total | 711 | 1490 | 2201 |

```
Classified + if predicted Pr(D) >= .5
True D defined as survive != 0
```

| | | |
|---|---|---|
| Sensitivity | Pr( +\| D) | 49.09% |
| Specificity | Pr( -\|~D) | 91.54% |
| Positive predictive value | Pr( D\| +) | 73.47% |
| Negative predictive value | Pr(~D\| -) | 79.03% |
| False + rate for true ~D | Pr( +\|~D) | 8.46% |
| False - rate for true D | Pr( -\| D) | 50.91% |
| False + rate for classified + | Pr(~D\| +) | 26.53% |
| False - rate for classified - | Pr( D\| -) | 20.97% |
| Correctly classified | | 77.83% |

从以上各表可知，Probit 模型的平均边际效应、准 $R^2$ 与正确预测比率与 Logit 模型十分接近，故可视为基本等价（二者的估计系数虽有差距，但估计系数没有可比性）。为了进一步验证这一点，下面使用 Probit 模型预测每位个体的存活概率，记为变量 *prob*1，并考察 *prob*1 与 *prob*（Logit 模型预测结果）的相关性。

```
. predict prob1
(option pr assumed; Pr(survive))
. corr prob prob1 [fweight = freq]
(obs =2201)
```

| | prob | prob1 |
|---|---|---|
| prob | 1.0000 | |
| prob1 | 0.9997 | 1.0000 |

从上表可知，Probit 与 Logit 模型对个体存活概率的预测结果相关系数高达 0.999 7，可以视为无差异。

## 11.10 其他离散选择模型

二值选择模型并非唯一的离散选择模型。其他离散选择模型还包括：

（1）多值选择（multiple choices）：比如，对交通方式的选择（步行、骑车、自驾车、乘出租车、地铁），对不同职业的选择，对手机品牌的选择。

（2）计数数据（count data）：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。

（3）排序数据（ordered data）：有些离散数据有着天然的排序。比如，公司债券的评级（AAA，AA，A，B，C 级），对“春节联欢晚会”的满意度（很满意、满意、不满意、很不满意）。

对于以上离散数据，一般也不宜直接进行 OLS 回归，主要估计方法仍为 MLE。由于离散选择模型主要用于微观经济学的实证研究中，故是“微观计量经济学”（Microeconometrics）的重要组成部分。除了离散数据外，微观计量经济学还关注的另一类数据类型为“受限被解释变量”（limited dependent variable），即被解释变量的取值范围受到限制（包括断尾回归、归并回归与样本选择模型等）。有关离散选择模型与受限被解释变量的具体介绍，参见 Wooldridge（2009）或陈强（2014）。

**习题**

11.1 假设离散型随机变量 $Y$ 服从如下概率分布：$P(Y=1)=p$，$P(Y=2)=q$，而 $P(Y=3)=1-p-q$。从此分布中抽取独立同分布的随机样本 $\{Y_1,\cdots,Y_n\}$。

（1）写出参数 $p$ 与 $q$ 的似然函数。

（2）推导 $p$ 与 $q$ 的最大似然估计量。

11.2 假设二值选择行为可通过不可观测的“潜变量”（latent variable）$y^*$ 来考察，其中 $y^*$ 是该行为的净收益（收益减成本）。如果净收益大于 0，则选择做，记 $y=1$；否则，选择不做，记 $y=0$。假设净收益的决定因素为

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \tag{11.32}$$

其中，$\boldsymbol{x}$ 为解释变量，而 $\varepsilon$ 为扰动项。证明：

（1）如果 $\varepsilon$ 服从逻辑分布，则 $y$ 为 Logit 模型。（提示：$P(y=1\mid\boldsymbol{x})=P(y^*>0\mid\boldsymbol{x})$。）

（2）如果 $\varepsilon\sim N(0,1)$，则 $y$ 为 Probit 模型。

（3）更一般地，如果 $\varepsilon\sim N(0,\sigma^2)$，其中 $\sigma\neq 1$，则 $y$ 为 Probit 模型。（提示：在方程（11.32）的两边同时除以 $\sigma$。）

11.3 泰坦尼克号在施救时奉行的政策是“妇女儿童优先”（women and children first）。此政策是否得到彻底执行？具体来说，三等舱的妇女或儿童的存活概率是否高于一等舱的男子？

根据数据集 titanic. dta 回答此问题。

11.4[①] 使用数据集 loanapp. dta 考察美国的按揭贷款(mortgage loan)是否存在种族歧视。如果申请人的按揭贷款获批,则被解释变量 *approve* 取值为 1;反之,*approve* 取值为 0。主要解释变量为 *white*(是否白人)。数据集中的其他种族为 *black*(是否黑人)与 *hispan*(是否拉丁裔)。本题统一使用稳健标准误。

(1) 把 *approve* 对 *white* 进行 OLS 回归。变量 *white* 的系数是否显著?此效应有多大?

(2) 把 *approve* 对 *white* 进行 Probit 回归。此回归结果与线性概率模型有何不同?

(3) 根据(1)与(2)的回归结果,是否可认为美国的按揭贷款市场对非白人(*nonwhite*)存在歧视?为什么?

(4) 加入控制变量 *hrat*(房供占总收入比例),*obrat*(其他债务支出占总收入比例),*loanprc*(贷款额占房价比例),*unem*(所在行业的失业率),*male*(是否男性),*married*(是否已婚),*dep*(家属人数),*sch*(是否受过 12 年及以上教育),*cosign*(是否有担保人),*chist*(1 = 怠账未及 60 天,0 = 怠账 60 天及以上),*pubrec*(是否曾申请破产),*mortlat*1(有 1 ~ 2 次逾期付款),*mortlat*2(有 2 次以上逾期付款),以及 *vr*(所在小区的空置率是否高于平均值),再次进行 *Probit* 回归。是否存在歧视非白人的统计证据?

(5) 使用 Logit,重复(4)的回归。Logit 模型的系数显著性是否与 Probit 模型相同?

(6) 使用 Logit,重复(4)的回归,但汇报几率比。在给定其他控制变量的情况下,比较白人与非白人成功申请贷款的几率比。

11.5 Chen(2015)研究中原王朝被游牧民族征服的概率,以每十年为观测单位建立公元前 221 年至 1911 年的时间序列。数据集 nomadic_conquest. dta 的被解释变量为 *conquered*(中原王朝是否被征服)。主要解释变量包括:*diff*(中原王朝早于游牧政权建立的年数),*age*(中原王朝的绝对年龄),*wall*(中原是否在长城的有效保护之下),以及 *drought*1(中国北方在十年中发生旱灾的年数比例的一阶滞后)[②]。另外,时间变量为 *decade*(十年)。

(1) 作为参照系,把 *conquered* 对 *diff*,*age*,*wall*,*drought*1 进行 OLS 回归,并使用稳健标准误。

(2) 使用 Logit,重复(1)的回归,并使用稳健标准误。评论变量系数的符号、统计显著性与经济意义。

(3) 计算所有变量的平均边际效应,并与线性概率模型的边际效应相比较。

(4) 通过几率比,说明 *diff* 与 *drought*1 对于游牧征服效应的大小。

(5) 计算 Logit 模型正确预测的百分比。

(6) 预测中原王朝被征服的概率,记为 *conquered*1。将预测征服概率(*conquered*1)与实际征服(*conquered*)的时间趋势画在一起进行对比。(提示:使用 Stata 命令"`tsline conquered1 conquered`"。)

(7) 使用 Probit,重复(1)的回归,并使用稳健标准误。

(8) 计算 Probit 模型的平均边际效应,并与 Logit 模型相比较。

(9) 计算 Probit 模型正确预测的百分比,并与 Logit 模型相比较。

---

① 此题改编自 Wooldridge(2009).

② 原文还有其他控制变量,在此从略。

*Panel data models have become increasingly popular among applied researchers due to their heightened capacity for capturing the complexity of human behavior as compared to cross-sectional or tim-series data models.*

——Cheng Hsiao

# 12. 面板数据

## 12.1 面板数据的特点

面板数据(panel data 或 longitudinal data),指的是在一段时间内跟踪同一组个体(individual)的数据。它既有横截面维度($n$ 位个体),又有时间维度($T$ 个时期)。一个 $T=3$ 的面板数据结构如表 12.1。

**表 12.1 面板数据的结构**

| | $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| 个体 1: $t=1$ | | | | |
| 个体 1: $t=2$ | | | | |
| 个体 1: $t=3$ | | | | |
| 个体 2: $t=1$ | | | | |
| 个体 2: $t=2$ | | | | |
| 个体 2: $t=3$ | | | | |
| …… | | | | |
| 个体 $n$: $t=1$ | | | | |
| 个体 $n$: $t=2$ | | | | |
| 个体 $n$: $t=3$ | | | | |

通常的面板数据 $T$ 较小,而 $n$ 较大,在使用大样本理论时让 $n$ 趋于无穷大。这种面板数据称为"短面板"(short panel)。反之,如果 $T$ 较大,而 $n$ 较小,则称为"长面板"(long panel)。在实践中,短面板较为常见。

在面板模型中,如果解释变量包含被解释变量的滞后值,则称为"动态面板"(dynamic panel);反之,则称为"静态面板"(static panel)。本书仅关注静态面板[①]。

如果在面板数据中,每个时期在样本中的个体完全一样,则称为"平衡面板"(balanced panel);反之,则称为"非平衡面板"(unbalanced panel)。我们主要关注平衡面板,但将在本章第 11 节讨论非平衡面板。

① 有关动态面板的介绍,参见陈强(2014)。

面板数据的主要优点如下。

(1) 可以解决遗漏变量问题:遗漏变量偏差是一个普遍存在的问题。虽然可以用工具变量法解决,但有效的工具变量通常很难找。遗漏变量常常是由于不可观测的个体差异或"异质性"(heterogeneity)造成的(比如个体能力),如果这种个体差异"不随时间而改变"(time invariant),则面板数据提供了解决遗漏变量问题的又一利器。

(2) 提供更多个体动态行为的信息:由于面板数据同时有横截面与时间两个维度,有时它可以解决单独的截面数据或时间序列所不能解决的问题。比如,考虑如何区分规模效应与技术进步对企业生产效率的影响。对于截面数据来说,由于没有时间维度,故无法观测到技术进步。而对于单个企业的时间序列数据来说,我们也无法区分其生产效率的提高究竟有多少是由于规模扩大,有多少是由于技术进步。又比如,对于失业问题,截面数据能告诉我们在某个时点上哪些人失业,而时间序列能告诉我们某个人就业与失业的历史,但这两种数据均无法告诉我们是否失业的总是同一批人(意味着低流转率,low turnover rate),还是失业的人群总在变动(意味着高流转率,high turnover rate)。如果有面板数据,就可能解决上述问题。

(3) 样本容量较大:由于同时有截面维度与时间维度,通常面板数据的样本容量更大,从而可以提高估计的精确度。

当然,面板数据也会带来一些问题,比如,样本数据通常不满足独立同分布的假定,因为同一个体在不同期的扰动项一般存在自相关。另外,面板数据的收集成本通常较高,不易获得。

## 12.2 面板数据的估计策略

估计面板数据的一个极端策略是将其看成是截面数据而进行混合回归(pooled regression),即要求样本中每位个体都拥有完全相同的回归方程。混合回归的缺点是,忽略了个体不可观测的异质性(heterogeneity),而该异质性可能与解释变量相关从而导致估计不一致。

另一极端策略则是,为每位个体估计一个单独的回归方程。分别回归的缺点是,忽略了个体的共性,而且可能没有足够大的样本容量(尤其对于短面板而言)。

实践中常采用折中的估计策略,即假定个体的回归方程拥有相同的斜率,但可有不同截距项,以捕捉异质性(参见图 12.1)。这种模型称为"个体效应模型"(individual-specific effects model),即

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\delta} + u_i + \varepsilon_{it} \quad (i = 1,\cdots,n;t = 1,\cdots,T) \tag{12.1}$$

其中,$z_i$ 为不随时间而变(time invariant)的个体特征(即 $z_{it} = z_i, \forall t$),比如性别;而 $\boldsymbol{x}_{it}$ 可以随个体及时间而变(time-varying)。扰动项由($u_i + \varepsilon_{it}$)两部分构成,称为"复合扰动项"(composite error term)。其中,不可观测的随机变量 $u_i$ 是代表个体异质性的截距项,即"个体效应"(individual effects)。在较早的文献中有时将 $u_i$ 视为常数(待估参数),但这也只是随机变量的特例,即退化的随机变量。$\varepsilon_{it}$ 为随个体与时间而改变的扰动项,称为"idiosyncratic error"。一般假设 $\{\varepsilon_{it}\}$ 为独立同分布,且与 $u_i$ 不相关。

如果 $u_i$ 与某个解释变量相关,则进一步称之为"固定效应模型"(Fixed Effects Model, FE)。在这种情况下,OLS 不一致。解决的方法是将模型转换,消去 $u_i$ 后获得一致估计量。

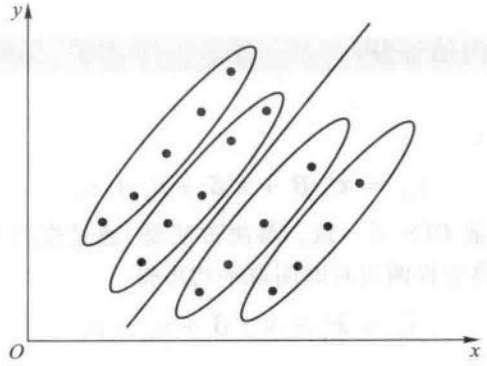



图 12.1 面板数据中不同个体的截距项可以不同

如果 $u_i$ 与所有解释变量($\boldsymbol{x}_{it}$,$\boldsymbol{z}_i$)均不相关,则进一步称之为"随机效应模型"(Random Effects Model,RE)。从经济理论的角度来看,随机效应模型比较少见[①],但仍需通过数据来检验究竟该用随机效应模型还是固定效应模型。

显然,与横截面数据相比,面板数据提供了更为丰富的模型与估计方法。

## 12.3 混合回归

如果所有个体都拥有完全一样的回归方程,则 $u_1 = u_2 = \cdots = u_n$。将这些相同的个体效应统一记为 $\alpha$,则方程(12.1)可写为:

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\delta} + \varepsilon_{it} \tag{12.2}$$

其中,$\boldsymbol{x}_{it}$不包括常数项。此时,可把所有数据放在一起,像对待横截面数据那样进行 OLS 回归,故称为"混合回归"(pooled regression)。

由于面板数据的特点,虽然通常可以假设不同个体之间的扰动项相互独立,但同一个体在不同时期的扰动项之间往往存在自相关。此时,每位个体不同时期的所有观测值即构成一个"聚类"(cluster)。这样,样本观测值可以分为不同的聚类,在同一聚类里的观测值互相相关,而不同聚类之间的观测值则不相关,称为"聚类样本"(cluster sample)。对于聚类样本,仍可进行 OLS 估计,但需使用"聚类稳健的标准误"(cluster-robust standard errors),在形式上也是一种夹心估计量,只是表达式更为复杂。

对于样本容量为 $nT$ 的平衡面板,共有 $n$ 个聚类,而每个聚类中包含 $T$ 期观测值。使用聚类稳健标准误的前提是,聚类中的观测值数目 $T$ 较小,而聚类数目 $n$ 较大($n\to\infty$);此时,聚类稳健标准误是真实标准误的一致估计。因此,聚类稳健标准误更适用于时间维度 $T$ 比截面维度 $n$ 小的短面板。另外,由于在其推导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。

混合回归的基本假设是不存在个体效应,对此假设需进行统计检验。由于个体效应以两种不同的形态存在(即固定效应与随机效应),故将在下文分别介绍其检验方法。

① 一般来说,不可观测的异质性通常会对解释变量有影响。

## 12.4 固定效应模型:组内估计量

考虑以下固定效应模型:

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\delta} + u_i + \varepsilon_{it} \tag{12.3}$$

其中,$u_i$ 与某解释变量相关,故 OLS 不一致。解决方法是,通过模型变换,消掉个体效应 $u_i$。在方程(12.3)中,给定个体 $i$,将方程两边对时间取平均可得

$$\bar{y}_i = \bar{\boldsymbol{x}}'_i\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i \tag{12.4}$$

其中,$\bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$,$\bar{\boldsymbol{x}}_i$ 与 $\bar{\varepsilon}_i$ 的定义类似;而 $\boldsymbol{z}_i$ 与 $u_i$ 由于不随时间而变,故对时间平均后仍不变。将原方程(12.3)减去平均方程(12.4),可得模型的离差形式:

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) \tag{12.5}$$

其中,$\boldsymbol{z}_i$ 与 $u_i$ 都被消去。定义 $\tilde{y}_{it} \equiv y_{it} - \bar{y}_i$,$\tilde{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$,$\tilde{\varepsilon}_{it} \equiv \varepsilon_{it} - \bar{\varepsilon}_i$,则

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}'_{it}\boldsymbol{\beta} + \tilde{\varepsilon}_{it} \tag{12.6}$$

由于上式中已将 $u_i$ 消去,故只要新扰动项 $\tilde{\varepsilon}_{it}$ 与新解释变量 $\tilde{\boldsymbol{x}}_{it}$ 不相关,则可用 OLS 一致地估计 $\boldsymbol{\beta}$,称为"固定效应估计量"(Fixed Effects Estimator),记为 $\hat{\boldsymbol{\beta}}_{FE}$。由于 $\hat{\boldsymbol{\beta}}_{FE}$ 主要使用了每位个体的组内离差信息,故也称为"组内估计量"(within estimator)。即使个体特征 $u_i$ 与解释变量 $\boldsymbol{x}_{it}$ 相关,只要使用组内估计量,即可得到一致估计,这是面板数据的一大优势。考虑到可能存在组内自相关,故应使用以每位个体为聚类的聚类稳健标准误。

然而,在作离差变换的过程中,$\boldsymbol{z}'_i\boldsymbol{\delta}$ 也被消掉,故无法估计 $\boldsymbol{\delta}$。也就是说,$\hat{\boldsymbol{\beta}}_{FE}$ 无法估计不随时间而变的变量的影响,这是 FE 的一大缺点。另外,为保证($\varepsilon_{it} - \bar{\varepsilon}_i$)与($\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$)不相关,需假定个体 $i$ 满足严格外生性(比前定变量或同期外生的假定更强),即 $E(\varepsilon_{it} \mid \boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iT}) = 0$,因为 $\bar{\boldsymbol{x}}_i$ 中包含了所有($\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iT}$)的信息。

## 12.5 固定效应模型:LSDV 法

对于方程(12.3)中的个体固定效应 $u_i$,传统上将其视为个体 $i$ 的待估参数;具体来说,可视 $u_i$ 为个体 $i$ 的截距项。根据第 9 章,对于 $n$ 位个体的 $n$ 个不同截距项,可通过在方程(12.3)中引入($n-1$)个个体虚拟变量来体现(如果没有截距项,则引入 $n$ 个虚拟变量),即估计以下模型:

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\delta} + \sum_{i=2}^{n} \gamma_i D_i + \varepsilon_{it} \tag{12.7}$$

其中,个体虚拟变量 $D_2 = 1$,如果为个体 2;否则,$D_2 = 0$。其他个体虚拟变量($D_3, \cdots, D_n$)的定义类似。常数项 $\alpha$ 表示被遗漏虚拟变量 $D_1$ 所对应的个体 1 的截距项,而个体 $i$($i > 1$)的截距项则为($\alpha + \gamma_i$)。

用 OLS 估计方程(12.7),称为"最小二乘虚拟变量法"(Least Square Dummy Variable,简记

LSDV)。可以证明,LSDV 法的估计结果与上述组内估计量 FE 完全相同。为什么二者的结果相同?直观来说,这正如线性回归与离差形式的回归在某种意义上是等价的(参见习题)。比如,

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \Leftrightarrow \quad y_i - \bar{y} = \beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \tag{12.8}$$

然而,如果做完 LSDV 后发现某些个体的虚拟变量不显著而删去,则 LSDV 的结果就不会与 FE 相同。使用 LSDV 的好处是可以得到对个体异质性 $u_i$ 的估计。LSDV 法的缺点是,如果 $n$ 很大,则需在回归方程中引入很多虚拟变量,可能超出 Stata 所允许的变量个数。

## 12.6 固定效应模型:一阶差分法

对于固定效应模型,还可对原方程(12.3)两边进行一阶差分,以消去个体效应 $u_i$(但同时也把 $\boldsymbol{z}_i'\boldsymbol{\delta}$ 消掉了):

$$y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) \tag{12.9}$$

对上式使用 OLS 即得到"一阶差分估计量"(First Differencing Estimator, FD),记为 $\hat{\boldsymbol{\beta}}_{FD}$。由于 $u_i$ 不再出现于差分方程中,只要扰动项的一阶差分($\varepsilon_{it} - \varepsilon_{i,t-1}$)与解释变量的一阶差分($\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}$)不相关,则 $\hat{\boldsymbol{\beta}}_{FD}$ 一致。此一致性条件比保证 $\hat{\boldsymbol{\beta}}_{FE}$ 一致的严格外生性假定更弱,这是 $\hat{\boldsymbol{\beta}}_{FD}$ 的主要优点。

可以证明(参见习题),如果 $T=2$,则 $\hat{\boldsymbol{\beta}}_{FD} = \hat{\boldsymbol{\beta}}_{FE}$。但对于 $T>2$,如果 $\{\varepsilon_{it}\}$ 为独立同分布的,则组内估计量 $\hat{\boldsymbol{\beta}}_{FE}$ 比一阶差分估计量 $\hat{\boldsymbol{\beta}}_{FD}$ 更有效率。因此,在实践上,主要使用 $\hat{\boldsymbol{\beta}}_{FE}$,而较少用 $\hat{\boldsymbol{\beta}}_{FD}$。

## 12.7 时间固定效应

个体固定效应模型解决了不随时间而变(time invariant)但随个体而异的遗漏变量问题。但还可能存在不随个体而变(individual invariant),但随时间而变(time varying)的遗漏变量问题;比如,企业经营的宏观经济环境。为此,在个体固定效应模型(12.3)中加入时间固定效应($\lambda_t$):

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + \lambda_t + u_i + \varepsilon_{it} \tag{12.10}$$

其中,$\lambda_t$ 随时间而变,但不随个体而变;故只有下标 $t$,而没有下标 $i$。在上式中,可视 $\lambda_t$ 为第 $t$ 期特有的截距项,并解释为"第 $t$ 期"对被解释变量 $y$ 的效应,故称 $\{\lambda_1, \cdots, \lambda_T\}$ 为"时间固定效应"(time fixed effects)。对于此方程,可使用 LSDV 法来估计,即对每个时期定义一个虚拟变量,然后把($T-1$)个时间虚拟变量包括在回归方程中(未包括的时间虚拟变量即为基期),比如

$$y_{it} = \alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + \sum_{t=2}^{T} \gamma_t D_t + u_i + \varepsilon_{it} \tag{12.11}$$

其中,时间虚拟变量 $D_2=1$,如果 $t=2$;否则,$D_2=0$。其他时间虚拟变量($D_3, \cdots, D_T$)的定义类似。常数项 $\alpha$ 表示被遗漏虚拟变量 $D_1$ 所对应的第 1 期截距项,而第 $t$ 期($t>1$)的截距项则为($\alpha+\gamma_t$)。

由于方程(12.11)既考虑了个体固定效应,又考虑了时间固定效应,故称为"双向固定效应"

(Two-way FE)。相应地,如果仅考虑个体固定效应,则称为"单向固定效应"(One-way FE)。有时为节省参数(比如,时间维度 $T$ 较大),可引入一个时间趋势项,以替代上述($T-1$)个时间虚拟变量:

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + z'_i\boldsymbol{\delta} + \gamma t + u_i + \varepsilon_{it} \tag{12.12}$$

上式隐含的假定是,每个时期的时间效应相等,即每期均增加 $\gamma$。如果此假定不太可能成立,则应在方程中加入时间虚拟变量。可通过检验这些时间虚拟变量的联合显著性来判断是否应使用双向固定效应模型。

## 12.8 随机效应模型

考虑以下随机效应模型:

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + z'_i\boldsymbol{\delta} + u_i + \varepsilon_{it} \tag{12.13}$$

其中,个体效应 $u_i$ 与解释变量 $\{\boldsymbol{x}_{it}, z_i\}$ 均不相关,故 OLS 一致。然而,由于扰动项由($u_i + \varepsilon_{it}$)组成,不是球型扰动项,故 OLS 不是最有效率的。

假设不同个体之间的扰动项互不相关。但即便如此,由于 $u_i$ 的存在,同一个体不同时期的扰动项之间仍存在自相关。对于 $t \neq s$,可以证明

$$\begin{aligned}\mathrm{Cov}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is}) &= \mathrm{Cov}(u_i, u_i) + \underbrace{\mathrm{Cov}(u_i, \varepsilon_{is})}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, u_i)}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is})}_{0} \\ &= \mathrm{Var}(u_i) \equiv \sigma_u^2 \neq 0\end{aligned} \tag{12.14}$$

其中,$\sigma_u^2 \equiv \mathrm{Var}(u_i)$ 为个体效应 $u_i$ 的方差(不随 $i$ 变化)。在上式中,如果 $t = s$,则

$$\mathrm{Var}(u_i + \varepsilon_{it}) = \sigma_u^2 + \sigma_\varepsilon^2 \tag{12.15}$$

其中,$\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_{it})$ 为 $\varepsilon_{it}$ 的方差(不随 $i, t$ 变化)。当 $t \neq s$ 时,个体 $i$ 扰动项的自相关系数为

$$\rho \equiv \mathrm{Corr}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is}) \equiv \frac{\mathrm{Cov}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is})}{\mathrm{Var}(u_i + \varepsilon_{it})} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \tag{12.16}$$

显然,自相关系数 $\rho$ 越大,则复合扰动项($u_i + \varepsilon_{it}$)中个体效应的部分($u_i$)越重要。Stata 记 $\rho$ 为"rho"。由于方程(12.13)的扰动项($u_i + \varepsilon_{it}$)存在组内自相关,故 OLS 不是最有效率的。可使用广义最小二乘法(GLS)对原模型进行转换,使得变换后的扰动项不再有自相关。

具体来说,首先定义

$$\theta \equiv 1 - \frac{\sigma_\varepsilon}{(T\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}} \tag{12.17}$$

其中,$T$ 为面板数据的时间维度。显然,$0 \leqslant \theta \leqslant 1$。给定个体 $i$,将方程(12.13)两边对时间进行平均,然后同乘以 $\theta$ 可得

$$\theta\bar{y}_i = \theta\bar{\boldsymbol{x}}'_i\boldsymbol{\beta} + \theta z'_i\boldsymbol{\delta} + \theta u_i + \theta\bar{\varepsilon}_i \tag{12.18}$$

将原方程(12.13)减去方程(12.18)可得“广义离差”(quasi-demeaned)模型:

$$y_{it} - \theta\bar{y}_i = (\boldsymbol{x}_{it} - \theta\bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (1-\theta)\boldsymbol{z}_i'\boldsymbol{\delta} + \underbrace{[(1-\theta)u_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i)]}_{\text{扰动项}} \tag{12.19}$$

由于$0 \leqslant \theta \leqslant 1$,故$(y_{it} - \theta\bar{y}_i)$只是减去平均值$\bar{y}_i$的一部分,故名“广义离差”。可以证明,广义离差方程(12.19)的扰动项$[(1-\theta)u_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i)]$不再有自相关(尽管它仍包含$u_i$),对此方程进行OLS估计即为GLS估计量。然而,$\theta$通常未知(取决于$u_i$与$\varepsilon_{it}$的方差),故需先估计$\hat{\theta}$,再进行FGLS估计。显然,可用下式来估计$\hat{\theta}$:

$$\hat{\theta} \equiv 1 - \frac{\hat{\sigma}_\varepsilon}{(T\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2)^{1/2}} \tag{12.20}$$

其中,$\hat{\sigma}_u$与$\hat{\sigma}_\varepsilon$分别为$\sigma_u$与$\sigma_\varepsilon$的样本估计值。Stata分别记$\hat{\sigma}_u$、$\hat{\sigma}_\varepsilon$与$\hat{\theta}$为`sigma_u`、`sigma_e`与`theta`。对于随机效应模型,由于OLS是一致的,且其扰动项为$(u_i + \varepsilon_{it})$,故可用OLS的残差来估计$(\sigma_u^2 + \sigma_\varepsilon^2)$。另外,FE也是一致的,且其扰动项为$(\varepsilon_{it} - \bar{\varepsilon}_i)$,故可用FE的残差来估计$\sigma_\varepsilon^2$。由此得到$\hat{\theta}$,再使用FGLS估计原模型,即可得到“随机效应估计量”(random effects estimator),记为$\hat{\boldsymbol{\beta}}_{RE}$。

对于随机效应模型,如果假设扰动项服从正态分布,则可写出样本的似然函数,然后进行最大似然估计(MLE)。

## 12.9 组间估计量

对于随机效应模型,还可以使用“组间估计量”。如果每位个体的时间序列数据较不准确或噪音较大,可对每位个体取时间平均值,然后用平均值来作横截面回归:

$$\bar{y}_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i \quad (i = 1, \cdots, n) \tag{12.21}$$

对上式使用OLS,即为“组间估计量”(Between Estimator,BE),记为$\hat{\boldsymbol{\beta}}_{BE}$。由于$\{\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{z}}_i\}$中包含了$\{\boldsymbol{x}_{it}, \boldsymbol{z}_i\}$的信息,如果$u_i$与解释变量$\{\boldsymbol{x}_{it}, \boldsymbol{z}_i\}$相关,则$\hat{\boldsymbol{\beta}}_{BE}$不一致。因此,不能在固定效应模型下使用组间估计法。即使在随机效应模型下,由于面板数据被压缩为截面数据,损失了较多信息量,故组间估计法也不常用。

## 12.10 拟合优度的度量

对于面板模型,如果进行混合回归,则可直接用混合回归的$R^2$衡量拟合优度。但如果使用固定效应或随机效应模型,拟合优度的度量略为复杂。对于有常数项的线性回归模型,其拟合优度$R^2$等于被解释变量$y$与预测值$\hat{y}$之间相关系数的平方,即$R^2 = [\text{Corr}(y, \hat{y})]^2$。有鉴于此,给定估计量$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}})$,Stata提供了以下三种$R^2$(不一定具有线性回归$R^2$的全部性质)。

(1) 对应于原模型(12.1),称$[\text{Corr}(y_{it}, \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i'\hat{\boldsymbol{\delta}})]^2$为“整体$R^2$”($R^2$overall),衡量估计量$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}})$对原模型的拟合优度。

(2) 对应于组内模型(12.6),称$[\mathrm{Corr}(\tilde{y}_{it},\tilde{\boldsymbol{x}}'_{it}\hat{\boldsymbol{\beta}})]^2$为"组内 $R^2$"($R^2$within),衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组内模型的拟合优度。

(3) 对应于组间模型(12.21),称$[\mathrm{Corr}(\bar{y}_i,\bar{\boldsymbol{x}}'_i\hat{\boldsymbol{\beta}}+\boldsymbol{z}'_i\hat{\boldsymbol{\delta}})]^2$为"组间 $R^2$"($R^2$between),衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组间模型的拟合优度。

无论固定效应、随机效应还是组间回归,都可以计算以上三种 $R^2$。对于固定效应模型,建议使用组内 $R^2$;对于组间回归模型,建议使用组间 $R^2$。对于随机效应模型,这三种 $R^2$ 都只是相应的相关系数平方而已(并非随机效应模型的 OLS $R^2$)。

## 12.11 非平衡面板

在面板数据中,如果每个时期在样本中的个体完全一样,则称为"平衡面板数据"(balanced panel)。但有时某些个体的数据可能缺失(比如,个体死亡、企业倒闭或被兼并、个体不再参与调查),或者新个体在后来才加入到调查中来。在这种情况下,每个时期观测到的个体不完全相同,称为"非平衡面板"(unbalanced panel)或"不完全面板"(incomplete panel)。

显然,非平衡面板数据并不影响计算离差形式的组内估计量(within estimator),因此,固定效应模型的估计可以照样进行。对于随机效应模型而言,非平衡面板数据也没有实质性影响。假设个体 $i$ 的时间维度为 $T_i$,则只要在做广义离差变换时,为每位个体定义

$$\hat{\theta}_i \equiv 1 - \frac{\hat{\sigma}_\varepsilon}{(T_i\hat{\sigma}_u^2+\hat{\sigma}_\varepsilon^2)^{1/2}} \tag{12.22}$$

即可照常进行 FGLS 估计。当然,非平衡面板数据使得估计量及其协方差矩阵的数学表达式更加复杂,但这些都由 Stata 在幕后进行。

非平衡面板可能出现的最大问题是,那些原来在样本中但后来丢掉的个体,如果其"丢掉"的原因是内生的(即与扰动项相关),则会导致样本不具有代表性(不再是随机样本),从而导致估计量不一致。比如,低收入的人群更容易从面板数据中丢掉。

如果从非平衡面板数据中提取一个平衡的面板数据子集[①],然后进行数据处理,则必然会损失样本容量,降低估计效率。更进一步,如果人为"丢掉"的个体并非完全随机,则同样会破坏样本的随机性。

## 12.12 究竟该用固定效应还是随机效应模型

在处理面板数据时,究竟应该使用固定效应还是随机效应模型是一个根本问题。为此,希望检验原假设"$H_0:u_i$ 与 $\boldsymbol{x}_{it},\boldsymbol{z}_i$ 不相关"(即随机效应模型为正确模型)。无论原假设成立与否,FE 都是一致的。如果原假设成立,则 RE 一致且比 FE 更有效率。但如果原假设不成立,则 RE 不

① 可通过连玉君老师提供的 Stata 命令"xtbalance"来实现,下载方法为"ssc install xtbalance",使用方法详见"help xtbalance"。

一致。因此，如果 $H_0$ 成立，则 FE 与 RE 估计量将共同收敛于真实的参数值，二者的差距将在大样本下消失，故 $(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})\xrightarrow{p}\boldsymbol{0}$。反之，如果二者的差距过大，则倾向于拒绝原假设。

以二次型度量此距离，豪斯曼检验（Hausman，1978）的统计量为

$$(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})'[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})}]^{-1}(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})\xrightarrow{d}\chi^2(K) \tag{12.23}$$

其中，$K$ 为 $\hat{\boldsymbol{\beta}}_{FE}$ 的维度，即 $\boldsymbol{x}_{it}$ 中所包含的随时间而变的解释变量个数（因为 $\hat{\boldsymbol{\beta}}_{FE}$ 无法估计不随时间而变的解释变量系数）。如果该统计量大于临界值，则拒绝 $H_0$。此检验的缺点是，为了计算 $\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})}$，它假设在 $H_0$ 成立的情况下，$\hat{\boldsymbol{\beta}}_{RE}$ 是最有效率的（fully efficient）。然而，如果扰动项存在异方差，则 $\hat{\boldsymbol{\beta}}_{RE}$ 并非最有效率的估计量。因此，传统的豪斯曼检验并不适用于异方差的情形，需使用异方差稳健的豪斯曼检验（参见下文）。

## 12.13 面板模型的 Stata 命令及实例

### 1. 面板数据的设定

设定面板数据的 Stata 命令为

```
xtset panelvar timevar
```

命令“xtset”告诉 Stata 你的数据为面板数据，其中面板（个体）变量“panelvar”的取值必须为整数且不重复，相当于将样本中每位个体进行编号；而“timevar”为时间变量。假如“panelvar”本来是字符串（比如，国家名字 *country*），则可使用以下命令将其转换为数字型变量：

```
encode country,gen(cntry)
```

其中，选择项“gen(cntry)”表示将新生成的数字型变量记为 *cntry*。这样，变量 *cntry* 就以“1，2，3，…”来指代不同的国家。

显示面板数据统计特征的 Stata 命令包括

```
xtdes        （显示面板数据的结构，是否为平衡面板）
xtsum        （显示组内、组间与整体的统计指标）
xtline varname    （对每位个体分别显示该变量的时间序列图；如果希望将所有个体的时间序列图叠放在一起，可加上选择项 overlay）
```

下面以数据集 lin_1992. dta 为例，取自 Lin（1992）对家庭联产承包责任制（household responsibility system）与中国农业增长的经典研究。该省际面板包含中国 28 个省 1970—1987 年有关种植业的数据。被解释变量为“种植业产值对数”（*ltvfo*，1980 年不变价格）。解释变量包括：耕地面积对数（*ltlan*，千亩），种植业劳动力（*ltwlab*），机械动力与畜力对数（*ltpow*，千马力），化肥使用量对数（*ltfer*，千吨），截止年底采用家庭联产承包制的生产队比重（*hrs*）[①]农村消费者价格与农村

① 严格来说，应使用 *hrs* 的一阶滞后来作为解释变量，以解释当年的农业产量（Xu，2012；Sun and Chen，2014）。但这里仍遵照 Lin（1992）的模型设定。

工业投入品价格之比的一阶滞后(*mipric*1,1950 年 =100),超额收购价格与农村工业投入品价格之比(*giprice*,1950 年 =100),复种指数(*mci*,播种面积除以耕地面积),非粮食作物占播种面积比重(*ngca*),时间趋势(*t*),*province*(省),*year*(年)。其中,为解决异方差问题,Lin(1992)将种植业产量、耕地面积、种植业劳动力、机械动力与畜力、化肥使用量这些传统的投入与产出变量都除以每省的生产队数目(*team*)。另外,两个价格变量 *mipric*1 与 *giprice* 为全国性指标,在各省都一样,只随时间变化。

首先,设定 *province* 与 *year* 为面板(个体)变量及时间变量:

```
. use lin_1992.dta,clear
. xtset province year
```

```
      panel variable:  province (strongly balanced)
       time variable:  year, 70 to 87
               delta:  1 unit
```

上表显示,这是一个平衡的面板数据(strongly balanced)。其次,显示数据集的结构:

```
. xtdes
```

```
province:  1, 2, ..., 28                                    n =         28
    year:  70, 71, ..., 87                                  T =         18
           Delta(year) = 1 unit
           Span(year)  = 18 periods
           (province*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        18      18      18        18        18      18      18

     Freq.  Percent     Cum. | Pattern
------------------------------+------------------
        28   100.00  100.00 | 111111111111111111
------------------------------+------------------
        28   100.00         | XXXXXXXXXXXXXXXXXX
```

上表清晰地显示,$n=28$,而 $T=18$。由于 $n$ 大而 $T$ 小,故这是一个短面板。再次,显示数据集中以上变量的统计特征:

```
. xtsum ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca
```

| Variable | | Mean | Std. Dev. | Min | Max | Observations | |
|---|---|---|---|---|---|---|---|
| ltvfo | overall | 7.647758 | .5331999 | 5.51 | 9.33 | N = | 504 |
| | between | | .4611992 | 6.982222 | 8.977222 | n = | 28 |
| | within | | .2806888 | 5.61498 | 8.471647 | T = | 18 |
| ltlan | overall | 5.837877 | .8084866 | 4.57 | 7.76 | N = | 504 |
| | between | | .8143036 | 4.617222 | 7.697778 | n = | 28 |
| | within | | .1138892 | 4.758988 | 6.163988 | T = | 18 |
| ltwlab | overall | 3.19752 | .4193496 | .98 | 3.86 | N = | 504 |
| | between | | .3195715 | 2.303889 | 3.646111 | n = | 28 |
| | within | | .2778123 | 1.618631 | 4.053631 | T = | 18 |
| ltpow | overall | 2.692778 | .9463811 | .2 | 5.04 | N = | 504 |
| | between | | .7702036 | 1.475 | 4.180556 | n = | 28 |
| | within | | .5678668 | .31 | 3.909444 | T = | 18 |
| ltfer | overall | 2.15119 | .7903761 | -.23 | 3.98 | N = | 504 |
| | between | | .5624935 | 1.081111 | 3.649444 | n = | 28 |
| | within | | .564791 | .4173016 | 3.510079 | T = | 18 |
| hrs | overall | .3497479 | .4526283 | 0 | 1 | N = | 476 |
| | between | | .0453814 | .2123529 | .4094118 | n = | 28 |
| | within | | .4504245 | -.0596639 | 1.053866 | T = | 17 |
| mipric1 | overall | 2.248889 | .2431379 | 1.76 | 2.73 | N = | 504 |
| | between | | 0 | 2.248889 | 2.248889 | n = | 28 |
| | within | | .2431379 | 1.76 | 2.73 | T = | 18 |
| giprice | overall | 2.858889 | .4537578 | 2.39 | 3.56 | N = | 504 |
| | between | | 0 | 2.858889 | 2.858889 | n = | 28 |
| | within | | .4537578 | 2.39 | 3.56 | T = | 18 |
| mci | overall | 1.538452 | .4931854 | .85 | 2.55 | N = | 504 |
| | between | | .4972044 | .8666667 | 2.487222 | n = | 28 |
| | within | | .0661412 | 1.323452 | 1.880119 | T = | 18 |
| ngca | overall | .199623 | .076145 | .06 | .91 | N = | 504 |
| | between | | .0631671 | .1144444 | .3466667 | n = | 28 |
| | within | | .0440777 | .1151786 | .8951786 | T = | 18 |

上表显示，除 *hrs* 外，所有变量的观测样本均为 $28 \times 18 = 504$；而关键变量 *hrs* 的样本容量仅为 $28 \times 17 = 476$，因为缺失 1980 年的 *hrs* 观测数据。

下面，看一下被解释变量 *ltvfo* 在 28 个省的时间趋势图，结果如图 12.2。

```
. xtline ltvfo
```

从图 12.2 可知，虽然不同省的种植业产值均随时间而增长，但变化的趋势与时机不尽相同。种植业产值的这些省际差异有助于估计决定种植业产值的因素。

### 2. 混合回归

作为参照系，首先进行混合回归。其 Stata 命令的基本格式为

```
reg y x1 x2 x3,vce(cluster id)
```

其中，“id”指用来确定每位个体的变量，而选择项“vce(cluster id)”表示以变量 id 作

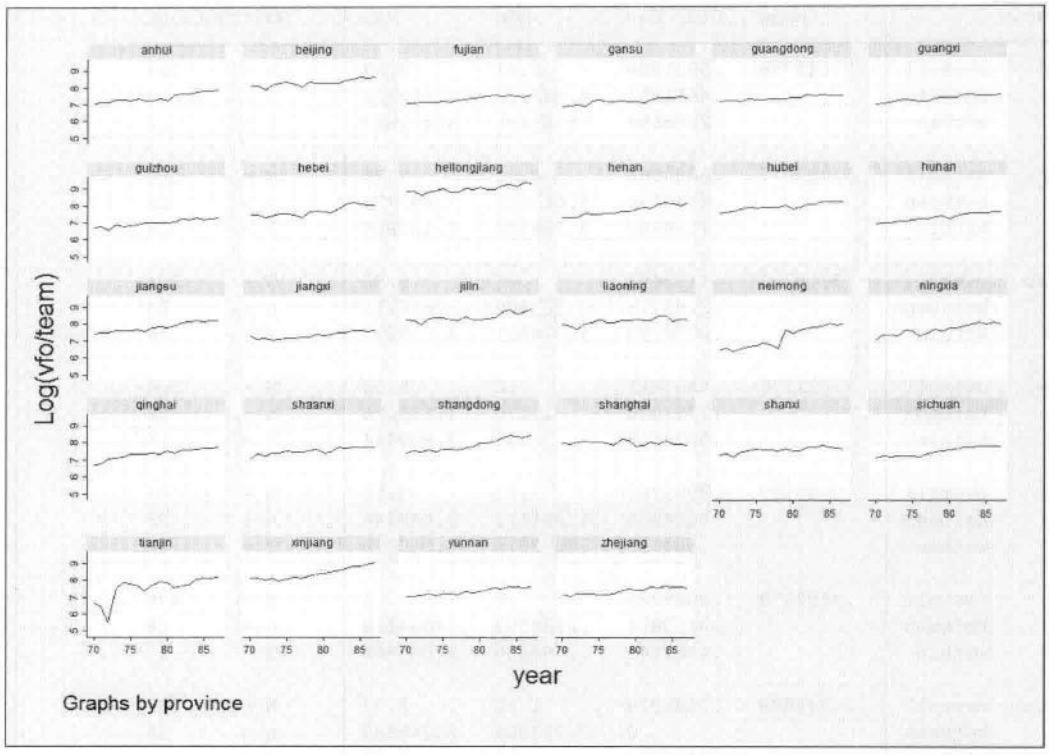



图 12.2　28 省种植业产值的时间趋势图

为聚类变量来计算聚类稳健的标准误。

```
.reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,vce
(cluster province)
```

其中,选择项“vce(cluster province)”表示,使用以 province 为聚类变量的聚类稳健标准误。将此结果储存,并记为“OLS”。

```
. estimates store OLS
```

```
Linear regression                               Number of obs =       476
                                                F(  9,    27) =     81.39
                                                Prob > F      =    0.0000
                                                R-squared     =    0.8685
                                                Root MSE      =    .19689
                                (Std. Err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |    .693795   .115024     6.03   0.000     .4577853    .9298048
      ltwlab |   .2650224   .0566294     4.68   0.000     .1488285    .3812164
       ltpow |  -.0291884   .0670385    -0.44   0.667    -.1667401    .1083633
       ltfer |   .3110617   .0531318     5.85   0.000     .2020443    .4200792
         hrs |   .2286926   .0489458     4.67   0.000     .1282642     .329121
     mipric1 |   .0122048   .0547799     0.22   0.825    -.1001943    .1246039
     giprice |  -.0538892   .0274468    -1.96   0.060    -.1102054     .002427
         mci |   .6949202   .1689692     4.11   0.000     .3482241    1.041616
        ngca |   .3053056   .5222639     0.58   0.564    -.7662914    1.376903
       _cons |   1.080587   .8269888     1.31   0.202    -.6162544    2.777427
------------------------------------------------------------------------------
```

上表显示,关键变量 *hrs* 在 1% 水平上显著为正。如果使用普通标准误,则可输入命令:

```
. reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca
```

```
      Source |       SS       df       MS              Number of obs =     476
-------------+------------------------------           F(  9,   466) =  342.09
       Model |  119.355964     9  13.2617737           Prob > F      =  0.0000
    Residual |  18.0652415   466  .038766613           R-squared     =  0.8685
-------------+------------------------------           Adj R-squared =  0.8660
       Total |  137.421205   475    .2893078           Root MSE      =  .19689

------------------------------------------------------------------------------
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |    .693795   .0368914    18.81   0.000     .6213008    .7662892
      ltwlab |   .2650224   .0238406    11.12   0.000     .2181741    .3118708
       ltpow |  -.0291884   .0331891    -0.88   0.380    -.0944073    .0360305
       ltfer |   .3110617   .0206459    15.07   0.000     .2704911    .3516324
         hrs |   .2286926   .0307121     7.45   0.000     .1683412     .289044
     mipric1 |   .0122048   .0533863     0.23   0.819    -.0927028    .1171125
     giprice |  -.0538892   .0271999    -1.98   0.048    -.1073389   -.0004395
         mci |   .6949202   .0522416    13.30   0.000      .592262    .7975784
        ngca |   .3053056   .1952732     1.56   0.119    -.0784195    .6890306
       _cons |   1.080587   .2832576     3.81   0.000     .5239661    1.637207
------------------------------------------------------------------------------
```

对比聚类稳健标准误与普通标准误可知,前者均大于后者。由于同一省不同年之间的扰动项一般存在自相关,而默认的普通标准误计算方法假设扰动项为独立同分布的,故普通标准误的估计并不准确①。

① Lin(1992)使用的是普通标准误,因为当时聚类稳健标准误刚被发明出来(Arellano,1987)。另外,如果使用聚类稳健标准误,就没有必要把传统投入与产出都除以每省的生产队数目以消除异方差,因为聚类稳健标准误也是异方差稳健的。事实上,我们对于异方差的具体形式并无把握。

### 3. 固定效应

由于每个省的“省情”不同，可能存在不随时间而变的遗漏变量，故考虑使用固定效应模型（FE）。固定效应模型（组内估计量）的 Stata 命令格式为

```
xtreg y x1 x2 x3,fe r
```

其中，选择项“fe”表示“fixed effects”（固定效应估计量），默认为“re”表示“random effects”（随机效应估计量）。其中，选择项“r”表示使用聚类稳健标准误；如果使用选择项“vce(cluster id)”也能达到完全相同的效果①。

LSDV 法的 Stata 命令为

```
reg y x1 x2 x3 i.id,vce(cluster id)
```

其中，“id”表示用来确定个体的变量，“i.id”则表示根据变量 id 而生成的虚拟变量。选择项“vce(cluster id)”表示使用聚类稳健的标准误。

首先使用组内估计量，并记其估计结果为“FE_robust”：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,fe r
. estimates store FE_robust
```

```
Fixed-effects (within) regression               Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8746                         Obs per group: min =        17
       between = 0.6483                                        avg =      17.0
       overall = 0.6993                                        max =        17

                                                F(9,27)            =    274.25
corr(u_i, Xb)  = -0.3877                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .6370234   .1681335     3.79   0.001     .2920421    .9820048
      ltwlab |   .1387786   .0624585     2.22   0.035     .0106242    .2669329
       ltpow |   .0577152   .0755568     0.76   0.452    -.0973146    .2127451
       ltfer |   .1826281    .043592     4.19   0.000     .0931846    .2720716
         hrs |   .2134022   .0391104     5.46   0.000     .1331542    .2936501
     mipric1 |   .0543577   .0590331     0.92   0.365    -.0667682    .1754837
     giprice |  -.0151451   .0245968    -0.62   0.543    -.0656135    .0353233
         mci |   .1943697   .0770515     2.52   0.018     .0362731    .3524663
        ngca |   .7562031   .3821261     1.98   0.058    -.0278549    1.540261
       _cons |   2.337895   .8552224     2.73   0.011      .583124    4.092667
-------------+----------------------------------------------------------------
     sigma_u |  .30549743
     sigma_e |  .10589274
         rho |  .89273901   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

上表的输出结果包括一个常数项（_cons），这是所有个体效应 $u_i$ 的平均值。上表最后一行

① 在使用命令 xtreg 时，Stata 已经知道这是面板数据，故使用选择项“r”或“vce(cluster id)”，都能得到完全相同的聚类稳健标准误。

显示，“rho = 0.89”，故复合扰动项（$u_i + \varepsilon_{it}$）的方差主要来自个体效应 $u_i$ 的变动。

究竟应该使用混合回归还是个体固定效应模型呢？在使用命令“xtreg，fe”时，如果不加选择项“r”（将估计结果记为“FE”），则输出结果还包含一个 $F$ 检验，其原假设为“$H_0$：所有 $u_i = 0$”，即混合回归是可以接受的：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,fe
. estimates store FE
```

```
Fixed-effects (within) regression               Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8746                         Obs per group: min =        17
       between = 0.6483                                        avg =      17.0
       overall = 0.6993                                        max =        17

                                                F(9,439)           =    340.20
corr(u_i, Xb)  = -0.3877                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .6370234   .0673191     9.46   0.000     .5047156    .7693312
      ltwlab |   .1387786   .0261554     5.31   0.000     .0873732     .190184
       ltpow |   .0577152   .0332508     1.74   0.083    -.0076352    .1230657
       ltfer |   .1826281   .0219921     8.30   0.000     .1394053     .225851
         hrs |   .2134022   .0223886     9.53   0.000        .1694    .2574043
     mipric1 |   .0543577   .0421659     1.29   0.198    -.0285145    .1372299
     giprice |  -.0151451   .0187457    -0.81   0.420    -.0519876    .0216975
         mci |   .1943697   .0876884     2.22   0.027     .0220285     .366711
        ngca |   .7562031   .2168141     3.49   0.001     .3300804    1.182326
       _cons |   2.337895    .385253     6.07   0.000     1.580726    3.095065
-------------+----------------------------------------------------------------
     sigma_u |  .30549743
     sigma_e |  .10589274
         rho |  .89273901   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(27, 439) =    43.41                 Prob > F = 0.0000
```

对于原假设“$H_0$：所有 $u_i = 0$”，由于上表最后一行 $F$ 检验的 $p$ 值为 0.000 0，故强烈拒绝原假设，即认为 FE 明显优于混合回归，应该允许每位个体拥有自己的截距项。然而，由于未使用聚类稳健标准误，故此 $F$ 检验并不有效，因为普通标准误均小于聚类稳健标准误。

为此，进一步通过 LSDV 法来考察（将估计结果记为“LSDV”）：

```
. reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca
i.province,vce(cluster province)
. estimates store LSDV
```

Linear regression

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | Number of obs = | 476 |
| | | | | | F( 8, 27) = | . |
| | | | | | Prob > F = | . |
| | | | | | R-squared = | 0.9642 |
| | | | | | Root MSE = | .10589 |

(Std. Err. adjusted for 28 clusters in province)

| ltvfo | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ltlan | .6370234 | .1732267 | 3.68 | 0.001 | .2815916 | .9924553 |
| ltwlab | .1387786 | .0643506 | 2.16 | 0.040 | .0067421 | .2708151 |
| ltpow | .0577152 | .0778457 | 0.74 | 0.465 | -.1020109 | .2174414 |
| ltfer | .1826281 | .0449126 | 4.07 | 0.000 | .0904751 | .2747811 |
| hrs | .2134022 | .0402952 | 5.30 | 0.000 | .1307233 | .296081 |
| mipric1 | .0543577 | .0608214 | 0.89 | 0.379 | -.0704375 | .1791529 |
| giprice | -.0151451 | .0253419 | -0.60 | 0.555 | -.0671423 | .0368522 |
| mci | .1943697 | .0793856 | 2.45 | 0.021 | .0314839 | .3572555 |
| ngca | .7562031 | .3937018 | 1.92 | 0.065 | -.0516063 | 1.564013 |
| province | | | | | | |
| beijing | -.1816498 | .1247829 | -1.46 | 0.157 | -.4376832 | .0743835 |
| fujian | .051657 | .0501916 | 1.03 | 0.313 | -.0513277 | .1546418 |
| gansu | -.8165674 | .1380808 | -5.91 | 0.000 | -1.099886 | -.5332489 |
| guangdong | -.010488 | .055811 | -0.19 | 0.852 | -.1250028 | .1040268 |
| guangxi | -.2304637 | .0570853 | -4.04 | 0.000 | -.3475932 | -.1133342 |
| guizhou | -.2350768 | .0615353 | -3.82 | 0.001 | -.3613369 | -.1088167 |
| hebei | -.2923854 | .0997217 | -2.93 | 0.007 | -.4969974 | -.0877733 |
| heilongjiang | -.1410195 | .2892268 | -0.49 | 0.630 | -.7344638 | .4524249 |
| henan | -.0904581 | .0435714 | -2.08 | 0.048 | -.1798593 | -.0010569 |
| hubei | .1118905 | .0340584 | 3.29 | 0.003 | .0420085 | .1817725 |
| hunan | -.0373775 | .0607647 | -0.62 | 0.544 | -.1620563 | .0873014 |
| jiangsu | .1150954 | .0342058 | 3.36 | 0.002 | .0449109 | .18528 |
| jiangxi | -.1352577 | .0579781 | -2.33 | 0.027 | -.2542188 | -.0162965 |
| jilin | -.2220282 | .2253552 | -0.99 | 0.333 | -.6844189 | .2403624 |
| liaoning | -.2789811 | .172656 | -1.62 | 0.118 | -.6332419 | .0752797 |
| neimong | -.9288069 | .2561317 | -3.63 | 0.001 | -1.454346 | -.403268 |
| ningxia | -.8813594 | .1975659 | -4.46 | 0.000 | -1.286731 | -.4759877 |
| qinghai | -.7062497 | .1521719 | -4.64 | 0.000 | -1.018481 | -.3940187 |
| shaanxi | -.3342067 | .0925991 | -3.61 | 0.001 | -.5242045 | -.144209 |
| shangdong | -.0049215 | .0581511 | -0.08 | 0.933 | -.1242377 | .1143947 |
| shanghai | .113901 | .0648627 | 1.76 | 0.090 | -.0191862 | .2469882 |
| shanxi | -.5312338 | .1514863 | -3.51 | 0.002 | -.8420581 | -.2204095 |
| sichuan | .0251618 | .0320732 | 0.78 | 0.440 | -.040647 | .0909707 |
| tianjin | -.3047612 | .1190042 | -2.56 | 0.016 | -.5489376 | -.0605848 |
| xinjiang | -.4007117 | .2397817 | -1.67 | 0.106 | -.8927032 | .0912797 |
| yunnan | -.2774542 | .0632713 | -4.39 | 0.000 | -.4072763 | -.1476321 |
| zhejiang | .1764445 | .0822195 | 2.15 | 0.041 | .007744 | .345145 |
| _cons | 2.568156 | .8279625 | 3.10 | 0.004 | .8693177 | 4.266995 |

从上表可知，不少个体虚拟变量在5%水平上显著，故可放心地拒绝“所有个体虚拟变量的系数都为0”的原假设，即认为存在个体固定效应，不应使用混合回归。另外，LSDV法的回归系数与组内估计量完全相同，但聚类稳健的标准误略有差别。

对于固定效应模型,也可使用一阶差分法(FD)。Stata 没有专门执行一阶差分法的命令,但在使用命令"xtserial,output"对组内自相关进行检验时①,可附带提供一阶差分法的估计结果(将此结果记为"FD"):

```
. xtserial ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,output
. estimates store FD
```

Linear regression — Number of obs = 420; F( 9, 27) = 902.61; Prob > F = 0.0000; R-squared = 0.5797; Root MSE = .11179

(Std. Err. adjusted for 28 clusters in province)

| D.ltvfo | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ltlan D1. | .9807158 | .0926143 | 10.59 | 0.000 | .790687 | 1.170745 |
| ltwlab D1. | .2420082 | .0734117 | 3.30 | 0.003 | .0913798 | .3926366 |
| ltpow D1. | -.0171023 | .0747984 | -0.23 | 0.821 | -.170576 | .1363714 |
| ltfer D1. | .2768317 | .0589799 | 4.69 | 0.000 | .155815 | .3978485 |
| hrs D1. | .2427773 | .0372382 | 6.52 | 0.000 | .1663709 | .3191837 |
| mipric1 D1. | .0250908 | .0357935 | 0.70 | 0.489 | -.0483513 | .0985329 |
| giprice D1. | -.0157708 | .021774 | -0.72 | 0.475 | -.0604473 | .0289057 |
| mci D1. | .1314675 | .1260309 | 1.04 | 0.306 | -.1271266 | .3900616 |
| ngca D1. | -.0260777 | .4846049 | -0.05 | 0.957 | -1.020405 | .9682494 |

```
Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,      27) =     12.511
          Prob > F =      0.0015
```

① 此处的组内自相关检验指的是检验 $H_0$:$\mathrm{Cov}(\varepsilon_{it},\varepsilon_{it})=0$,参见陈强(2014,p. 280),在此从略。

从上表可知，一阶差分估计量(FD)的估计系数与组内估计量(FE)有一定差别。一般认为，FE 比 FD 更有效率，故较少使用 FD。

也可以在固定效应模型中考虑时间效应，即双向固定效应(Two-way FE)，以捕捉技术进步等效应。为节省待估参数，首先考虑加入时间趋势项(将估计结果记为“FE_trend”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca t, fe r
. estimates store FE_trend
```

```
Fixed-effects (within) regression               Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8749                         Obs per group: min =        17
       between = 0.6490                                        avg =      17.0
       overall = 0.7006                                        max =        17

                                                F(10,27)           =    247.93
corr(u_i, Xb)  = -0.3767                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .6517195   .1843858     3.53   0.001     .2733911    1.030048
      ltwlab |   .1431791   .0589267     2.43   0.022     .0222716    .2640866
       ltpow |   .0366317   .0991178     0.37   0.715    -.1667413    .2400047
       ltfer |    .180359   .0428995     4.20   0.000     .0923365    .2683816
         hrs |   .1916276   .0295596     6.48   0.000     .1309763    .2522789
     mipric1 |   .0198772   .0515121     0.39   0.703    -.0858168    .1255713
     giprice |   -.026268   .0226875    -1.16   0.257    -.0728189    .0202829
         mci |   .2014685    .078794     2.56   0.016     .0397965    .3631404
        ngca |   .6761116    .421738     1.60   0.121    -.1892234    1.541447
           t |   .0063068   .0106492     0.59   0.559    -.0155436    .0281572
       _cons |    2.36174   .8262751     2.86   0.008     .6663633    4.057116
-------------+----------------------------------------------------------------
     sigma_u |  .30327958
     sigma_e |  .10589784
         rho |  .89132628   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

其中，时间趋势项 $t$ 并不显著($p$ 值为 0.559)，而主要变量的显著性不变。

其次，考虑加入年度虚拟变量。为了演示目的，定义年度虚拟变量：

```
. tab year,gen(year)
```

| year | Freq. | Percent | Cum. |
|---|---|---|---|
| 70 | 28 | 5.56 | 5.56 |
| 71 | 28 | 5.56 | 11.11 |
| 72 | 28 | 5.56 | 16.67 |
| 73 | 28 | 5.56 | 22.22 |
| 74 | 28 | 5.56 | 27.78 |
| 75 | 28 | 5.56 | 33.33 |
| 76 | 28 | 5.56 | 38.89 |
| 77 | 28 | 5.56 | 44.44 |
| 78 | 28 | 5.56 | 50.00 |
| 79 | 28 | 5.56 | 55.56 |
| 80 | 28 | 5.56 | 61.11 |
| 81 | 28 | 5.56 | 66.67 |
| 82 | 28 | 5.56 | 72.22 |
| 83 | 28 | 5.56 | 77.78 |
| 84 | 28 | 5.56 | 83.33 |
| 85 | 28 | 5.56 | 88.89 |
| 86 | 28 | 5.56 | 94.44 |
| 87 | 28 | 5.56 | 100.00 |
| Total | 504 | 100.00 | |

此命令将生成时间虚拟变量 *year*1，*year*2，…，*year*18。加入年度虚拟变量后，由于两个价格变量 *mipric*1 与 *giprice* 在各省都一样，故无法包括在回归方程中，以避免严格多重共线性。下面，进行含时间虚拟变量的双向固定效应估计（将结果记为“FE_TW”）：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca year2 -year18,fe r
. estimates store FE_TW
```

```
note: year11 omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8932                         Obs per group: min =        17
       between = 0.6596                                        avg =      17.0
       overall = 0.7156                                        max =        17

                                                F(23,27)           =    949.82
corr(u_i, Xb)  = -0.3425                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .5833594   .1745834     3.34   0.002     .2251439    .9415749
      ltwlab |   .1514909   .0585107     2.59   0.015     .0314368     .271545
       ltpow |   .0971114    .090911     1.07   0.295    -.0894225    .2836453
       ltfer |   .1693346   .0438098     3.87   0.001     .0794444    .2592248
         hrs |   .1503752   .0587581     2.56   0.016     .0298136    .2709368
         mci |   .1978373   .0810587     2.44   0.022     .0315186     .364156
        ngca |   .7784081   .4016301     1.94   0.063    -.0456688    1.602485
       year2 |  -.0240404    .023366    -1.03   0.313    -.0719836    .0239027
       year3 |  -.1323624   .0404832    -3.27   0.003    -.2154272   -.0492977
       year4 |  -.0377336   .0357883    -1.05   0.301    -.111165     .0356979
       year5 |   .0058554   .0500774     0.12   0.908    -.096895     .1086058
       year6 |   .0096731   .0566898     0.17   0.866    -.1066448    .1259911
       year7 |  -.0476465    .061423    -0.78   0.445    -.1736761    .0783832
       year8 |  -.0869336   .0680579    -1.28   0.212    -.2265767    .0527096
       year9 |  -.0325205   .0766428    -0.42   0.675    -.1897785    .1247376
      year10 |  -.0076332   .0833462    -0.09   0.928    -.1786454     .163379
      year11 |           0  (omitted)
      year12 |   -.093479   .1093614    -0.85   0.400    -.3178701    .1309121
      year13 |  -.0447862   .1207405    -0.37   0.714    -.2925251    .2029528
      year14 |  -.0309435   .1377207    -0.22   0.824     -.313523    .2516361
      year15 |   .0442535   .1428764     0.31   0.759    -.2489048    .3374117
      year16 |  -.0033372   .1561209    -0.02   0.983    -.3236709    .3169965
      year17 |     .00484    .157992     0.03   0.976    -.3193329    .3290129
      year18 |   .0386475   .1639608     0.24   0.815    -.2977723    .3750674
       _cons |   2.651286   .7738994     3.43   0.002     1.063376    4.239196
-------------+----------------------------------------------------------------
     sigma_u |  .29344594
     sigma_e |  .09930555
         rho |  .89724523   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

其中，*year*1（即 1970 年）被作为基期（对应于常数项_cons），而不包括在上述回归命令中（否则，将出现虚拟变量陷阱，导致完全多重共线性）。另外，由于 1980 年的 *hrs* 数据缺失，故 *year*11（即 1980 年）也被去掉。即使在双向固定效应模型中，*hrs* 也依然在 5% 水平上显著为正。另外，大多数的年度虚拟变量均不显著（但 *year*3 在 1% 水平上显著）。下面检验所有年度虚拟变量的联合显著性：

```
. test year2 year3 year4 year5 year6 year7 year8 year9 year10 year12
year13 year14 year15 year16 year17 year18
```

```
( 1)  year2 = 0
( 2)  year3 = 0
( 3)  year4 = 0
( 4)  year5 = 0
( 5)  year6 = 0
( 6)  year7 = 0
( 7)  year8 = 0
( 8)  year9 = 0
( 9)  year10 = 0
(10)  year12 = 0
(11)  year13 = 0
(12)  year14 = 0
(13)  year15 = 0
(14)  year16 = 0
(15)  year17 = 0
(16)  year18 = 0

      F( 16,    27) =   14.82
           Prob > F =    0.0000
```

结果强烈拒绝“无时间固定效应”的原假设,认为应在模型中包括时间固定效应。在 Stata 13 中,还可直接用以下命令来估计双向固定效应模型(不必先生成时间虚拟变量):

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca i.year,fe r
```

其中,“i.year”表示根据变量 year 的不同取值来生成年度虚拟变量。

```
Fixed-effects (within) regression               Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8932                         Obs per group: min =        17
       between = 0.6596                                        avg =      17.0
       overall = 0.7156                                        max =        17

                                                F(23,27)           =    949.82
corr(u_i, Xb)  = -0.3425                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 28 clusters in province)
```

| ltvfo | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ltlan | .5833594 | .1745834 | 3.34 | 0.002 | .2251439 | .9415749 |
| ltwlab | .1514909 | .0585107 | 2.59 | 0.015 | .0314368 | .271545 |
| ltpow | .0971114 | .090911 | 1.07 | 0.295 | -.0894225 | .2836453 |
| ltfer | .1693346 | .0438098 | 3.87 | 0.001 | .0794444 | .2592248 |
| hrs | .1503752 | .0587581 | 2.56 | 0.016 | .0298136 | .2709368 |
| mci | .1978373 | .0810587 | 2.44 | 0.022 | .0315186 | .364156 |
| ngca | .7784081 | .4016301 | 1.94 | 0.063 | -.0456688 | 1.602485 |
| year | | | | | | |
| 71 | -.0240404 | .023366 | -1.03 | 0.313 | -.0719836 | .0239027 |
| 72 | -.1323624 | .0404832 | -3.27 | 0.003 | -.2154272 | -.0492977 |
| 73 | -.0377336 | .0357883 | -1.05 | 0.301 | -.111165 | .0356979 |
| 74 | .0058554 | .0500774 | 0.12 | 0.908 | -.096895 | .1086058 |
| 75 | .0096731 | .0566898 | 0.17 | 0.866 | -.1066448 | .1259911 |
| 76 | -.0476465 | .061423 | -0.78 | 0.445 | -.1736761 | .0783832 |
| 77 | -.0869336 | .0680579 | -1.28 | 0.212 | -.2265767 | .0527096 |
| 78 | -.0325205 | .0766428 | -0.42 | 0.675 | -.1897785 | .1247376 |
| 79 | -.0076332 | .0833462 | -0.09 | 0.928 | -.1786454 | .163379 |
| 81 | -.093479 | .1093614 | -0.85 | 0.400 | -.3178701 | .1309121 |
| 82 | -.0447862 | .1207405 | -0.37 | 0.714 | -.2925251 | .2029528 |
| 83 | -.0309435 | .1377207 | -0.22 | 0.824 | -.313523 | .2516361 |
| 84 | .0442535 | .1428764 | 0.31 | 0.759 | -.2489048 | .3374117 |
| 85 | -.0033372 | .1561209 | -0.02 | 0.983 | -.3236709 | .3169965 |
| 86 | .00484 | .157992 | 0.03 | 0.976 | -.3193329 | .3290129 |
| 87 | .0386475 | .1639608 | 0.24 | 0.815 | -.2977723 | .3750674 |
| _cons | 2.651286 | .7738994 | 3.43 | 0.002 | 1.063376 | 4.239196 |
| sigma_u | .29344594 | | | | | |
| sigma_e | .09930555 | | | | | |
| rho | .89724523 | (fraction of variance due to u_i) | | | | |

#### 4. 随机效应

以上结果已基本确认了个体效应的存在，但个体效应仍可能以随机效应(RE)的形式存在。随机效应估计的 Stata 命令为

```
xtreg y x1 x2 x3,re r theta
```

其中，选择项“re”为默认选项(可省略)；选择项“r”表示使用聚类稳健标准误，如果使用选择项“vce(cluster id)”也能达到完全相同的效果。选择项“theta”表示显示用于进行广义离

差变换的 $\theta$ 值。

对于随机效应模型，也可以进行 MLE 估计，其 Stata 命令为

```
xtreg y x1 x2 x3,mle
```

下面，进行随机效应(RE)的估计(将结果记为“RE_robust”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,re r theta
. estimates store RE_robust
```

```
Random-effects GLS regression                   Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8700                         Obs per group: min =        17
       between = 0.8135                                        avg =      17.0
       overall = 0.8263                                        max =        17

                                                Wald chi2(7)       =   2452.50
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
theta          = .81012778

                                (Std. Err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .5655915   .1089863     5.19   0.000     .3519823    .7792007
      ltwlab |   .1441844   .0462225     3.12   0.002     .0535899     .234779
       ltpow |    .060477   .0508828     1.19   0.235    -.0392515    .1602055
       ltfer |   .1882741   .0386418     4.87   0.000     .1125376    .2640107
         hrs |   .2186096   .0377121     5.80   0.000     .1446952    .2925241
         mci |   .4702368   .0836862     5.62   0.000      .306215    .6342587
        ngca |   .6745175   .3663329     1.84   0.066    -.0434818    1.392517
       _cons |   2.387878   .5672669     4.21   0.000     1.276055    3.499701
-------------+----------------------------------------------------------------
     sigma_u |  .13324845
     sigma_e |  .10624809
         rho |  .6113231   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

上表最后三行显示，$\sigma_u = 0.133\,248\,45$，$\sigma_\varepsilon = 0.106\,248\,09$，而 $\rho \equiv \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = 0.611\,323\,1$。究竟应使用混合回归，还是个体随机效应模型？Breusch and Pagan(1980)提供了一个检验个体随机效应的 *LM* 检验，其原假设为 $H_0: \sigma_u^2 = 0$，而备择假设为 $H_1: \sigma_u^2 \neq 0$。如果拒绝 $H_0$，则说明原模型中应包括反映个体特性的随机扰动项 $u_i$，而不应该使用混合回归。该 *LM* 检验的 Stata 命令为“xttest0”(在执行命令“xtreg,re”之后才能进行)。

```
. xttest0
```

```
Breusch and Pagan Lagrangian multiplier test for random effects

        ltvfo[province,t] = Xb + u[province] + e[province,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                 ltvfo   |    .2893078        .5378734
                     e   |    .0112887        .1062481
                     u   |    .0177551        .1332484

        Test:   Var(u) = 0
                                 chibar2(01) =  1235.75
                              Prob > chibar2 =   0.0000
```

上表显示，*LM* 检验强烈拒绝“不存在个体随机效应”的原假设（$p$ 值为 0.000 0），即认为在随机效应与混合回归二者之间，应该选择随机效应。

下面，看一下使用普通标准误的随机效应估计结果（将结果记为“RE”）。

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,re
. estimates store RE
```

```
Random-effects GLS regression                   Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.8700                         Obs per group: min =        17
       between = 0.8135                                        avg =      17.0
       overall = 0.8263                                        max =        17

                                                Wald chi2(7)       =   2981.73
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       ltvfo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .5655915   .0478214    11.83   0.000     .4718633    .6593196
      ltwlab |   .1441844   .0233398     6.18   0.000     .0984394    .1899295
       ltpow |    .060477   .0252917     2.39   0.017     .0109062    .1100478
       ltfer |   .1882741   .0208337     9.04   0.000     .1474408    .2291075
         hrs |   .2186096   .0216932    10.08   0.000     .1760918    .2611275
         mci |   .4702368    .064681     7.27   0.000     .3434643    .5970093
        ngca |   .6745175   .2121571     3.18   0.001     .2586973    1.090338
       _cons |   2.387878   .2895274     8.25   0.000     1.820414    2.955341
-------------+----------------------------------------------------------------
     sigma_u |  .13324845
     sigma_e |  .10624809
         rho |  .6113231   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

作为对照，也可以对随机效应模型进行 MLE 估计（将结果记为“MLE”）：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,mle nolog
```

```
. estimates store MLE
```

```
Random-effects ML regression                    Number of obs      =       476
Group variable: province                        Number of groups   =        28

Random effects u_i ~ Gaussian                   Obs per group: min =        17
                                                               avg =      17.0
                                                               max =        17

                                                LR chi2(7)         =    961.00
Log likelihood  =  332.89739                    Prob > chi2        =    0.0000
```

| ltvfo | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ltlan | .5643288 | .0518396 | 10.89 | 0.000 | .462725 | .6659326 |
| ltwlab | .1408974 | .0235161 | 5.99 | 0.000 | .0948067 | .1869882 |
| ltpow | .0717013 | .0248585 | 2.88 | 0.004 | .0229794 | .1204231 |
| ltfer | .1772027 | .0206075 | 8.60 | 0.000 | .1368127 | .2175927 |
| hrs | .2153264 | .020987 | 10.26 | 0.000 | .1741926 | .2564602 |
| mci | .4064832 | .0712859 | 5.70 | 0.000 | .2667654 | .546201 |
| ngca | .7149811 | .2105466 | 3.40 | 0.001 | .3023172 | 1.127645 |
| _cons | 2.490512 | .3039688 | 8.19 | 0.000 | 1.894744 | 3.08628 |
| /sigma_u | .2141094 | .0317827 | | | .1600597 | .2864107 |
| /sigma_e | .1061021 | .0035665 | | | .0993371 | .1133278 |
| rho | .8028448 | .0486812 | | | .6942946 | .8840722 |

```
Likelihood-ratio test of sigma_u=0: chibar2(01)=  464.22 Prob>=chibar2 = 0.000
```

上表显示，随机效应 MLE 的系数估计值与随机效应 FGLS 有所不同，但在性质上依然类似。另外，上表最后一行的 *LR* 检验强烈拒绝原假设 $H_0:\sigma_u=0$，即认为存在个体随机效应，不应进行混合回归。

**5. 固定效应还是随机效应：豪斯曼检验**

在处理面板数据时，究竟使用固定效应还是随机效应模型，这是一个基本问题。为此，需进行豪斯曼检验。豪斯曼检验的 Stata 命令为

```
xtreg y x1 x2 x3,fe          (固定效应估计)
estimates store FE           (存储结果)
xtreg y x1 x2 x3,re          (随机效应估计)
estimates store RE           (存储结果)
hausman FE RE,constant sigmamore  (豪斯曼检验)
```

其中，选择项“constant”表示在比较系数估计值时包括常数项（默认不含常数项）；选择项“sigmamore”表示统一使用更有效率的那个估计量（即随机效应估计量）的方差估计。由于传统的豪斯曼检验假设球形扰动项，故在进行固定效应与随机效应的估计时，均不使用异方差或聚类稳健的标准误。

由于前面已存储了相应的估计结果，故可直接进行豪斯曼检验。

```
. hausman FE RE,constant sigmamore
```

| | Coefficients (b) FE | (B) RE | (b-B) Difference | sqrt(diag(V_b-V_B)) S.E. |
|---|---|---|---|---|
| ltlan | .6399658 | .5655915 | .0743743 | .0476709 |
| ltwlab | .1239927 | .1441844 | -.0201917 | .0125808 |
| ltpow | .0771604 | .060477 | .0166834 | .0081232 |
| ltfer | .1762775 | .1882741 | -.0119966 | .0078425 |
| hrs | .2075817 | .2186096 | -.0110279 | .0052769 |
| mci | .2580359 | .4702368 | -.2122009 | .0583709 |
| ngca | .7722795 | .6745175 | .097762 | .0828671 |
| _cons | 2.310114 | 2.387878 | -.0777638 | .2078242 |

```
            b = consistent under Ho and Ha; obtained from xtreg
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

              chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =          48.49
            Prob>chi2 =         0.0000
            (V_b-V_B is not positive definite)
```

由于 $p$ 值为 0.000 0，故强烈拒绝原假设“$H_0$：$u_i$ 与解释变量不相关”，认为应该使用固定效应模型，而非随机效应模型。

传统的豪斯曼检验假定，在 $H_0$ 成立的情况下，随机效应模型最有效率。这意味着，扰动项必须是同方差的，在异方差的情况下并不适用。事实上，在 Stata 中进行以上豪斯曼检验时，如果使用聚类稳健标准误，比如“xtreg y x1 x2 x3,fe vce(cluster id)”，则 Stata 将无法执行“hausman FE RE”命令。

为此，下载非官方命令 xtoverid 进行稳健的豪斯曼检验。此处“overid”指“overidentification test”（过度识别检验），因为随机效应模型与固定效应模型相比，前者多了“个体异质性 $u_i$ 与解释变量不相关”的约束条件，也可视为过度识别条件。

```
. ssc install xtoverid        (下载安装命令 xtoverid)
```

在使用命令 xtoverid 之前，需先以稳健标准误来执行命令“xtreg,re”。

```
. quietly xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,r
. xtoverid
```

```
Test of overidentifying restrictions: fixed vs random effects
Cross-section time-series model: xtreg re  robust cluster(province)
Sargan-Hansen statistic 221.225  Chi-sq(7)     P-value = 0.0000
```

上表显示，$\chi^2(7)$ 统计量为 221.225，$p$ 值为 0.000 0，故仍然强烈拒绝随机效应的原假设。

### 6. 组间估计量

纯粹为了演示目的，下面进行组间估计。

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,be
```

```
Between regression (regression on group means)  Number of obs      =       476
Group variable: province                        Number of groups   =        28

R-sq:  within  = 0.4673                         Obs per group: min =        17
       between = 0.9362                                        avg =      17.0
       overall = 0.0232                                        max =        17

                                                F(7,20)            =     41.93
sd(u_i + avg(e_i.))=  .1357173                  Prob > F           =    0.0000
```

| ltvfo | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ltlan | .7021718 | .1177241 | 5.96 | 0.000 | .4566037 | .9477398 |
| ltwlab | .5020747 | .0940029 | 5.34 | 0.000 | .3059881 | .6981612 |
| ltpow | -.1511518 | .1188815 | -1.27 | 0.218 | -.3991344 | .0968307 |
| ltfer | .1132485 | .0976941 | 1.16 | 0.260 | -.0905377 | .3170347 |
| hrs | -3.757418 | 1.242961 | -3.02 | 0.007 | -6.350189 | -1.164647 |
| mci | .586029 | .1838616 | 3.19 | 0.005 | .2025005 | .9695575 |
| ngca | .4443814 | .6565795 | 0.68 | 0.506 | -.9252194 | 1.813982 |
| _cons | 2.432839 | 1.005985 | 2.42 | 0.025 | .3343904 | 4.531288 |

由于豪斯曼检验选择了固定效应，而组间估计量仅在随机效应成立的情况下才一致，故组间估计的结果并不可信。例如，根据组间估计量，*hrs* 在 1% 水平上显著为负，即家庭联产承包责任制反而对种植业产值有负作用。

下面，将以上各主要方法的回归系数及标准误列表（为节省空间，不汇报包含年度虚拟变量的双向固定效应）：

```
. esttab OLS FE_robust FE_trend FE RE,b se mtitle
```

| | (1) OLS | (2) FE_robust | (3) FE_trend | (4) FE | (5) RE |
|---|---|---|---|---|---|
| ltlan | 0.694*** | 0.637*** | 0.652** | 0.640*** | 0.566*** |
| | (0.115) | (0.168) | (0.184) | (0.0644) | (0.0478) |
| ltwlab | 0.265*** | 0.139* | 0.143* | 0.124*** | 0.144*** |
| | (0.0566) | (0.0625) | (0.0589) | (0.0253) | (0.0233) |
| ltpow | -0.0292 | 0.0577 | 0.0366 | 0.0772** | 0.0605* |
| | (0.0670) | (0.0756) | (0.0991) | (0.0253) | (0.0253) |
| ltfer | 0.311*** | 0.183*** | 0.180*** | 0.176*** | 0.188*** |
| | (0.0531) | (0.0436) | (0.0429) | (0.0212) | (0.0208) |
| hrs | 0.229*** | 0.213*** | 0.192*** | 0.208*** | 0.219*** |
| | (0.0489) | (0.0391) | (0.0296) | (0.0213) | (0.0217) |
| mipric1 | 0.0122 | 0.0544 | 0.0199 | | |
| | (0.0548) | (0.0590) | (0.0515) | | |
| giprice | -0.0539 | -0.0151 | -0.0263 | | |
| | (0.0274) | (0.0246) | (0.0227) | | |
| mci | 0.695*** | 0.194* | 0.201* | 0.258** | 0.470*** |
| | (0.169) | (0.0771) | (0.0788) | (0.0831) | (0.0647) |
| ngca | 0.305 | 0.756 | 0.676 | 0.772*** | 0.675** |
| | (0.522) | (0.382) | (0.422) | (0.217) | (0.212) |
| t | | | 0.00631 | | |
| | | | (0.0106) | | |
| _cons | 1.081 | 2.338* | 2.362** | 2.310*** | 2.388*** |
| | (0.827) | (0.855) | (0.826) | (0.340) | (0.290) |
| N | 476 | 476 | 476 | 476 | 476 |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

从上表可知，无论使用何种方法，家庭联产承包责任制（*hrs*）均在1%的水平上对中国种植业产值有显著的正效应。

**习题**

12.1 考虑横截面数据的一元线性回归：

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{12.24}$$

证明OLS估计量$\hat{\beta}$等价以下离差模型的OLS估计量：

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \quad (i = 1, \cdots, n) \tag{12.25}$$

其中，$\bar{y}, \bar{x}, \bar{\varepsilon}$分别为$y, x, \varepsilon$的样本均值，比如$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$。（提示：使用无常数项的OLS公式。）

12.2 证明当$T=2$时（两期面板），差分估计量等价于组内估计量，即$\hat{\boldsymbol{\beta}}_{FD} = \hat{\boldsymbol{\beta}}_{FE}$。

12.3[①] 数据集 munnell.dta 包含美国 48 个州、1970—1986 年的年度数据。为了估计公共资本对经济增长的贡献，使用此数据集进行以下回归：

$$\ln y_{it} = \beta_0 + \beta_1 \ln k_{1,it} + \beta_2 \ln k_{2,it} + \beta_3 \ln labor_{it} + \beta_4 unemp_{it} + u_i + \varepsilon_{it} \quad (12.26)$$

其中，$y$ 为州产值(gross state product)，$k_1$ 为公共资本(包括高速公路、街道、供水、下水道及其他公共建筑)，$k_2$ 为私人资本存量(private capital stock)，*labor* 为非农劳动力，*unemp* 为州失业率(反映影响产出的经济周期因素)。面板变量为 *state*(州)，而时间变量为 *year*(年份)。

(1) 进行混合回归，评论 $\ln k_1$ 的系数符号、显著性与经济意义。

(2) 对随机效应模型进行 FGLS 估计。$\ln k_1$ 的系数符号与显著性是否有变化？检验是否存在个体随机效应。

(3) 对随机效应模型进行 MLE 估计。

(4) 对固定效应模型进行组内估计。$\ln k_1$ 的系数符号与显著性是否有变化？

(5) 对固定效应模型进行 LSDV 估计。检验是否存在个体固定效应。

(6) 进行传统的豪斯曼检验。

(7) 进行稳健的豪斯曼检验。

(8) 在组内估计中，加入时间趋势项。时间趋势项是否显著？

(9) 在组内估计中，加入时间虚拟变量，估计双向固定效应模型。时间效应是否显著？

(10) 计算一阶差分估计量。$\ln k_1$ 的系数符号与显著性是否有变化？

(11) 计算组间估计量。此估计量是否可信？

① 此数据集来自 Munnell(1990)，也为 Baltagi and Pinnoi(1995)与 Baltagi(2005)所使用。