Mc Graw Hill Education

“十一五”国家重点图书出版规划项目

·经/济/科/学/译/丛·

Basic Econometrics

(Fifth Edition)

计量经济学基础 上册

（第五版）

达摩达尔·N·古扎拉蒂（Damodar N. Gujarati）
唐·C·波特（Dawn C. Porter） 著

中国人民大学出版社

著作权合同登记号
图字：01－2009－2556号

本书配套图书有：

1.《计量经济学基础》（第五版）原版英文精编影印版本；

2.《计量经济学基础》（第五版）学生习题解答手册，提供了本书所有习题和问题的答案。

欲了解以上图书信息，请登录：www.crup.com.cn/jingji

本书英文原版图书还配有内容丰富的网络资源，使用本书作教材的教师可填写书后的《教师反馈表》申请获取以上资源。

http://www.mheducation.com

“十一五”国家重点图书出版规划项目

·经／济／科／学／译／丛·

Basic Econometrics

(Fifth Edition)

计量经济学基础 上册

（第五版）

达摩达尔·N·古扎拉蒂（Damodar N. Gujarati）
唐·C·波特（Dawn C. Porter） 著
费剑平 译

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

计量经济学基础：第5版/古扎拉蒂，波特著；费剑平译. —北京：中国人民大学出版社，2011

（经济科学译丛）

ISBN 978-7-300-13693-6

Ⅰ.①计… Ⅱ.①古…②波…③费… Ⅲ.①计量经济学 Ⅳ.①F224.0

中国版本图书馆CIP数据核字（2011）第080232号

“十一五”国家重点图书出版规划项目

经济科学译丛

计量经济学基础（第五版）

达摩达尔·N·古扎拉蒂 著

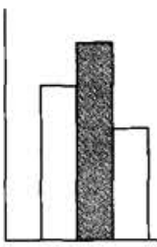
唐·C·波特

费剑平 译

Jiliang Jingjixue Jichu

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	涿州市星河印刷有限公司		
规　　格	185 mm×260 mm　16开本	**版　　次**	2011年6月第1版
印　　张	58.75插页6	**印　　次**	2011年6月第1次印刷
字　　数	1 234 000	**定　　价**	99.00元（上下册）

《经济科学译丛》编辑委员会



《经济科学译丛》总序

中国是一个文明古国，有着几千年的辉煌历史。近百年来，中国由盛而衰，一度成为世界上最贫穷、落后的国家之一。1949 年中国共产党领导的革命，把中国从饥饿、贫困、被欺侮、被奴役的境地中解放出来。1978 年以来的改革开放，使中国真正走上了通向繁荣富强的道路。

中国改革开放的目标是建立一个有效的社会主义市场经济体制，加速发展经济，提高人民生活水平。但是，要完成这一历史使命绝非易事，我们不仅需要从自己的实践中总结教训，也要从别人的实践中获取经验，还要用理论来指导我们的改革。市场经济虽然对我们这个共和国来说是全新的，但市场经济的运行在发达国家已有几百年的历史，市场经济的理论亦在不断发展完善，并形成了一个现代经济学理论体系。虽然许多经济学名著出自西方学者之手，研究的是西方国家的经济问题，但他们归纳出来的许多经济学理论反映的是人类社会的普遍行为，这些理论是全人类的共同财富。要想迅速稳定地改革和发展我国的经济，我们必须学习和借鉴世界各国包括西方国家在内的先进经济学的理论与知识。

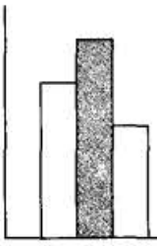
本着这一目的，我们组织翻译了这套经济学教科书系列。这套译丛的特点是：第一，全面系统。除了经济学、宏观经济学、微观经济学等基本原理之外，这套译丛还包括了产业组织理论、国际经济学、发展经济学、货币金融学、公共财政、劳动经济学、计量经济学等重要领域。第二，简明通俗。与经济学的经典名著不同，这套丛书都是国外大学通用的经济学教科书，大部分都已发行了几版或十几版。作者尽可能地用简明通俗的语言来阐述深奥的经济学原理，并附有案例与习题，对于初学者来说，更容

易理解与掌握。

经济学是一门社会科学，许多基本原理的应用受各种不同的社会、政治或经济体制的影响，许多经济学理论是建立在一定的假设条件上的，假设条件不同，结论也就不一定成立。因此，正确理解掌握经济分析的方法而不是生搬硬套某些不同条件下产生的结论，才是我们学习当代经济学的正确方法。

本套译丛于 1995 年春由中国人民大学出版社发起筹备并成立了由许多经济学专家学者组织的编辑委员会。中国留美经济学会的许多学者参与了原著的推荐工作。中国人民大学出版社向所有原著的出版社购买了翻译版权。北京大学、中国人民大学、复旦大学以及中国社会科学院的许多专家教授参与了翻译工作。前任策划编辑梁晶女士为本套译丛的出版做出了重要贡献，在此表示衷心的感谢。在中国经济体制转轨的历史时期，我们把这套译丛献给读者，希望为中国经济的深入改革与发展做出贡献。

《经济科学译丛》编辑委员会



前 言

□ 本书的编写目的

三十年前出版了《计量经济学基础》的第一版。这么多年过去了，计量经济理论与实践又取得了一些重要进展。在本书随后的各个版本中，我都试图把该领域的主要进展涵盖进来。第五版也继承了这一传统。

不过，这些年来，一直没变的是我坚定的信念：无需使用初级以上的线性代数、微积分和统计学知识，就能够向一个初学者讲授计量经济学。有些专题内容本身就有些技术性，在这种情况下，我就把相应的内容放在适当的附录中，或者让读者参考适当的资料。即便如此，我仍尽可能简化这些技术性材料，以便读者能对这种材料形成直觉上的理解。

令我分外惊喜的是，本书不仅生命力极强，而且除了被经济学与金融学专业的学生广泛使用外，还被政治学、国际关系学、农学和健康科学领域的研究者所钟爱。所有学生都将发现，这个新版本增加了一些非常有用的专题及其具体应用。我在本版中还特别注意书中所用现实数据的适用性和及时性。事实上，我增加了大约 15 个说明性例子和 30 多个章末习题。此外，我还更新了上一版中约 20 个例子和 20 多个习题的数据。

尽管我已经年过八旬，但我仍没有丧失对计量经济学的喜爱，而且我还竭力跟上该领域的主要进展。为了帮助我实现这一目的，很高兴洛杉矶南加州大学马歇尔商学院的统计学助教唐·C·波特博士愿意成为我的合作者。我们为《计量经济学基础》第五版的完成付出了艰辛的努力。

□ 第五版的主要特色

在讨论各章的具体改动之前，有必要指出这个新版本的如下主要特色：

1. 说明性例子中所用的所有数据实际上都被更新了。

2. 新增加了几个例子。

3. 我们还在几章的末尾增加了总结性的例子，以说明书中的各种观点。

4. 本书还包含了几个例子的具体的计算机输出结果。大多数结果都是基于EViews（第6版）、STATA（第10版）和MINITAB（第15版）而得到的。

5. 在许多章节包含了一些新的图表。

6. 许多章节还包含了一些以新数据为基础的习题。

7. 虽然书中包含了一些小型数据，但大型样本数据都张贴在本书的网站上，因而尽可能减少本书的分量。该网站还给出了本书所用到的所有数据，并定期更新。

8. 在少数章节我们还给出了一些课堂练习，鼓励学生搜集自己的数据并练习使用书中的各种方法。本书还包含了一些蒙特卡罗模拟。

□ 第五版的具体改动

各章的具体改动如下：

1. 第3章介绍的经典线性回归模型（CLRM）背后的假定现在清楚地区分了固定回归元（解释变量）和随机回归元。我们还讨论了这种区分的重要性。

2. 第6章的附录讨论了对数、博克斯-考克斯变换和各种增长表达式的性质。

3. 现在，第7章不仅讨论了单个回归元对因变量的边际影响，还讨论了所有解释变量同时变化对因变量的影响。本章还按照第3章讨论模型同样的结构重新组织了内容。

4. 第11章对各种异方差检验进行了比较。

5. 第12章就结构变化对自相关的影响进行了全新的讨论。

6. 第13章新增的专题包括数据缺失、非正态误差项和随机回归元。

7. 第14章讨论的非线性回归模型增加了博克斯-考克斯变换的一个具体应用。

8. 第15章增加了几个新的例子，以说明logit和probit模型在各个领域的应用。

9. 有关面板数据回归模型的第16章进行了全面修订，并用几个具体应用加以解释。

10. 现在的第17章增加了对希姆斯和格兰杰因果检验的深入讨论。

11. 第21章现在全面地讨论了平稳和非平稳时间序列以及与各种平稳性检验有关的几个问题。

12. 第22章增加了对如下问题的讨论：为什么在有些情况下为了使一个时间序列变得平稳而对它取一阶差分可能是不太适当的方法。

除了这些具体的改动之外，以前版本中的疏漏和打印错误也得以订正，而且还使得某些章节的专题讨论变得更加合理。

□ 课程安排和内容选择

此版更广泛的覆盖面使得教师在选择适合其教学对象的专题方面有充分的灵活性。这里给出如何使用本书的一些建议。

非专业人员一学期的课程：附录 A、第 1～9 章及对第 10～12 章作一简单了解（略去全部证明）。

经济学专业一学期的课程：附录 A 和第 1～13 章。

经济学专业两学期的课程：附录 A、B、C 和第 1～22 章。第 14 和 16 章可以有选择性地学习。某些技术性附录可以略去。

硕士生、博士生和研究者：将本书作为计量经济学主题方面必备的参考书。

□ 补充材料

一个综合性的网址提供了如下辅助材料：

——书中的数据以及书中提到的一些大型数据集；作者会定期更新这些数据。

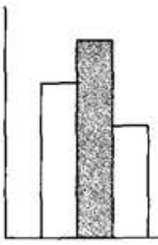
——唐·C·波特撰写的《习题解答手册》提供了全书所有习题和问题的答案。

——包含书中所有图表的数字图像库。

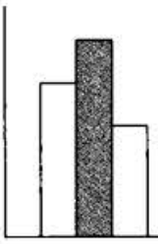
欲了解更多信息，请登录 www. mhhe. com/ gujarati5e。

达摩达尔·N·古扎拉蒂

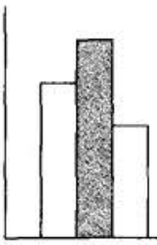
唐·C·波特



简要目录



目 录



引　言

1.1　什么是计量经济学?

从字面上解释，**计量经济学**（econometrics）意谓“经济测量”。虽然测量是计量经济学的一个重要部分，但计量经济学涉及的范围要广泛得多，这可以从下面的一些文献摘录看出：

> 计量经济学，是对经济学的作用存在某种期待的结果，它把数理统计学应用于经济数据，以使数理经济学构造出来的模型得到经验上的支持，并获得数值结果。①
>
> ……计量经济学可定义为实际经济现象的数量分析。这种分析基于理论与观测的并行发展，而理论与观测又通过适当的推断方法得以联系。②
>
> 计量经济学可定义为这样的社会科学：它把经济理论、数学和统计推断作为工具，应用于经济现象的分析。③

① Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics*, The University of Chicago Press, Chicago, 1968, p. 74.

② P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, “Report of the Evaluative Committee for Econometrica,” *Econometrica*, vol. 22, no. 2, April 1954, pp. 141-146.

③ Arthur S. Goldberger, *Econometric Theory*, John Wiley & Sons, New York, 1964, p. 1.

计量经济学研究经济定律的经验判定。①

计量经济学家的艺术，就在于找出一组足够具体且足够现实的假定，使他尽可能最好地利用他所获得的数据。②

计量经济学有助于在积极意义上驱散公众对经济学科（数量的或非数量的）的如下不良印象：这门学科犹如一个空箱子，即使有打开它的钥匙，对其空洞的内容，任何十位经济学家都会作出十一种解释。③

本质上，计量经济学的研究方法是，利用统计推断的理论和技术作为桥头堡，以达到经济理论和实际测算相衔接的目的。④

1.2 为什么是一门单独的学科？

上述各种定义表明，计量经济学是经济理论、数理经济、经济统计与数理统计的混合物。然而，这门学科值得作为一门独立的学科来研究，理由如下。

经济理论所作的陈述或假说大多数是定性的。例如，微观经济理论声称，在其他条件不变的情况下，一种商品的价格下降可望增加对该商品的需求量，即经济理论设想商品价格与其需求量之间存在一负向或逆向关系。但此理论并没有对两者的关系提供任何数值度量，也就是说，它没有说出随着商品价格的某一变化，需求量将会增加或减少多少。计量经济学家的工作就是要提供这一数值估计。换言之，计量经济学对大多数的经济理论赋予经验内容。

数理经济学的主要问题，是要用数学形式（方程式）来表述经济理论，而不管该理论是否可以量化或是否能够得到实证支持。如前所示，计量经济学的主要兴趣在于经济理论的经验论证。我们将看到，计量经济学家常常使用数理经济学家所提供的数学方程式，但要把这些方程式改造成适合于经验检验的形式。这种从数学方程到计量经济方程的转换需要有许多的创造性和实际技巧。

经济统计学的问题，主要是收集、加工并通过图表的形式来展现经济数据。这正是经济统计学家的工作。他们是收集国民生产总值（GNP）、就业、失业、价格等数据的主要负责人。这些数据从此构成了计量经济工作的原始资料。但是，经济统计学家的工作到此为止。他们不考虑怎样利用所收集来的数据去检验经济理论。当然，如果他们考虑的话，他们就变成计量经济学家了。

---

① H. Theil, *Principles of Econometrics*, John Wiley & Sons, New York, 1971, p. 1.

② E. Malinvaud, *Statistical Methods of Econometrics*, Rand McNally, Chicago, 1966, p. 514.

③ Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics*, Edward Elgar Publishing, Hants, England, 1990, p. 54.

④ T. Haavelmo, "The Probability Approach in Econometrics," Supplement to *Econometrica*, vol. 12, 1944, preface p. iii.

虽然数理统计学提供了这一行业中使用的许多工具，但由于大多数经济数据的独特性，即数据并非受控下的实验结果，计量经济学家常常需要有特殊的方法。如同气象学家那样，计量经济学家通常依赖于不能由他们直接控制的数据。如斯班诺斯（Spanos）所正确地观察到的那样：

> 在计量经济学中，建模者通常面对的是**观测**（observational）数据而非**实验**（experimental）数据。这对计量经济学中的经验建模有两方面的重要含义。首先，要求建模者掌握与分析实验数据极为不同的技巧……其次，数据搜集者与分析者的分离要求建模者十分熟悉所用数据的性质和结构。①

I.3 计量经济学方法论

对一个经济问题，计量经济学家是怎样进行分析的呢？他们的方法论是什么？尽管关于计量经济学的思想方法有了若干学派，但我们这里讲述的主要是至今仍在经济学及其他社会和行为科学领域的经验研究中占统治地位的**传统**（traditional）或**经典**（classical）方法论。②

大致说来，传统的计量经济学方法论按如下路线进行：

1. 理论或假说的陈述；
2. 理论的数学模型设定；
3. 统计或计量经济模型设定；
4. 获取数据；
5. 计量经济模型的参数估计；
6. 假设检验；
7. 预报或预测；
8. 利用模型进行控制或制定政策。

为了说明以上步骤，让我们考虑如下著名的凯恩斯消费理论。

□ 1. 理论或假说的陈述

凯恩斯说：

> 基本的心理定律……是，通常或平均而言，人们倾向于随着他们收入的增

① Aris Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, United Kingdom, 1999, p. 21.

② 关于对计量经济学方法论更完善但可能属于高级的讨论，可参见 David F. Hendry, *Dynamic Econometrics*, Oxford University Press, New York, 1995。也可参见阿里斯·斯班诺斯（Aris Spanos）的前引文献。

加而增加其消费，但不如收入增加的那么多。[①]

简言之，凯恩斯设想，**边际消费倾向**（marginal propensity to consume，MPC），即收入每变化一个单位的消费变化率，大于零而小于1。

□ 2. 消费的数学模型设定

虽然凯恩斯假设消费与收入之间存在正向关系，但他并没有明确指出二者之间准确的函数关系。为简单起见，数理经济学家也许建议采用如下形式的凯恩斯消费函数：

$$Y = \beta_1 + \beta_2 X \quad 0 < \beta_2 < 1 \tag{I. 3. 1}$$

其中 Y=消费支出，X=收入，而被称为模型**参数**（parameter）的 β_1 和 β_2 分别代表**截距**（intercept）和**斜率**（slope）系数。

斜率系数 β_2 度量了边际消费倾向，为说明其几何意义，将方程（I. 3. 1）表示如图 I—1。该方程表明消费与收入有线性关系。这种关系仅是消费与收入关系即经济学中所称的**消费函数**（consumption function）数学模型的一个例子。所谓数学模型无非就是一组数学方程而已。如果模型只有一个方程，像上例那样，就称之为**单方程模型**（single-equation model）；如果模型有不止一个方程，就称之为**多方程模型**（multiple-equation model）（后者将在以后讨论）。

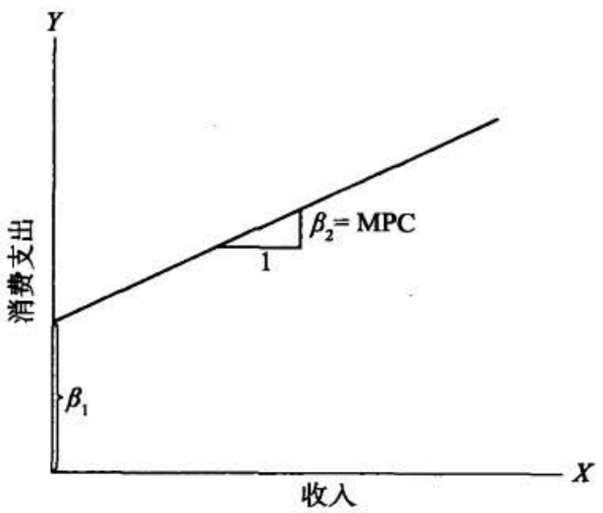


图 I—1　凯恩斯消费函数

出现在方程（I. 3. 1）等号左边的变量称为**因变量**（dependent variable），而出现在右边的变量（一个或多个）则称为**自变量**（independent）或**解释变量**（explanatory）。这样，在代表凯恩斯消费函数的方程（I. 3. 1）中，消费（支出）是因变量，而收入是解释变量。

① John Maynard Keynes, *The General Theory of Employment, Interest and Money*, Harcourt Brace Jovanovich, New York, 1936, p. 96.

□ 3. 消费的计量经济模型设定

由方程（I. 3. 1）给出的消费函数的纯数学模型，假定消费与收入之间有一个准确的或确定性的关系，因此它对计量经济学家的用处是有限的。一般地说，经济变量之间的关系是非准确的。例如，我们获得了（比如说）500 个美国家庭消费支出和可支配收入的一个样本数据，并把这些数据画在以消费支出为纵坐标，以可支配收入为横坐标的图纸上。我们不能指望所有的观测值都恰好落在方程（I. 3. 1）这条直线上，因为除了收入外，还有其他变量影响着消费支出。比方说，家庭规模、家庭成员的年龄、家庭的宗教信仰等，都会对消费有一定的影响。

考虑到经济变量之间的非准确关系，计量经济学家会把确定性的消费函数（I. 3. 1）修改如下：

$$Y = \beta_1 + \beta_2 X + u \tag{I. 3. 2}$$

其中 u 被称为**干扰项**（disturbance）或**误差项**（error term），是一个**随机变量**（random variable，stochastic variable），它有良好定义的概率性质。干扰项 u 可用来代表所有未经指明的对消费有所影响的那些因素。

方程（I. 3. 2）是**计量经济模型**（econometric model）之一例。更技术地讲，它是本书主要论述的**线性回归模型**（linear regression model）之一例。该计量经济消费函数假设了因变量 Y（消费）与解释变量 X（收入）之间存在线性关系。然而两者的关系不是准确的，它随着家庭的变化而有所变化。

可把消费函数的计量经济模型描绘成图 I—2 那样。

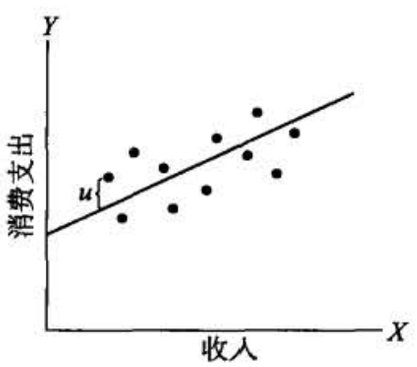


图 I—2　凯恩斯消费函数的计量经济模型

□ 4. 获取数据

为了估计（I. 3. 2）所给的计量经济模型，也就是为了得到 β_1 和 β_2 的数值，需要有数据。虽然我们在下一章将要更详细地讨论数据对经济分析的根本重要性，但现在不妨先看一下美国经济在 1960—2005 年间的数据，如表 I—1 所示。该表中的 Y 变量是（整个国家）个人消费支出（PCE）的加总，而 X 变量是国内生产总值（GDP），度量了美国的总收入，均以 2000 年不变美元价格计算，单位是十亿美元。因此，所列数据代表以 2000 年不变价格计算的“真实”消费和“真实”收入。现将

这些数据描绘在图 I—3 上（与图 I—2 相比较）。暂不考虑图中所画的直线。

表 I—1　　1960—2005 年间美国的 Y（个人消费支出）和 X（国内生产总值）数据

均以 2000 年十亿美元为单位

年份	PCE (Y)	GDP (X)	年份	PCE (Y)	GDP (X)
1960	1 597.4	2 501.8	1983	3 668.6	5 423.8
1961	1 630.3	2 560.0	1984	3 863.3	5 813.6
1962	1 711.1	2 715.2	1985	4 064.0	6 053.7
1963	1 781.6	2 834.0	1986	4 228.9	6 263.6
1964	1 888.4	2 998.6	1987	4 369.8	6 475.1
1965	2 007.7	3 191.1	1988	4 546.9	6 742.7
1966	2 121.8	3 399.1	1989	4 675.0	6 981.4
1967	2 185.0	3 484.6	1990	4 770.3	7 112.5
1968	2 310.5	3 652.7	1991	4 778.4	7 100.5
1969	2 396.4	3 765.4	1992	4 934.8	7 336.6
1970	2 451.9	3 771.9	1993	5 099.8	7 532.7
1971	2 545.5	3 898.6	1994	5 290.7	7 835.5
1972	2 701.3	4 105.0	1995	5 433.5	8 031.7
1973	2 833.8	4 341.5	1996	5 619.4	8 328.9
1974	2 812.3	4 319.6	1997	5 831.8	8 703.5
1975	2 876.9	4 311.2	1998	6 125.8	9 066.9
1976	3 035.5	4 540.9	1999	6 438.6	9 470.3
1977	3 164.1	4 750.5	2000	6 739.4	9 817.0
1978	3 303.1	5 015.0	2001	6 910.4	9 890.7
1979	3 383.4	5 173.4	2002	7 099.3	10 048.8
1980	3 374.1	5 161.7	2003	7 295.3	10 301.0
1981	3 422.2	5 291.7	2004	7 577.1	10 703.5
1982	3 470.3	5 189.3	2005	7 841.2	11 048.6

资料来源：*Economic Report of the President*，2007，Table B-2，p. 230.

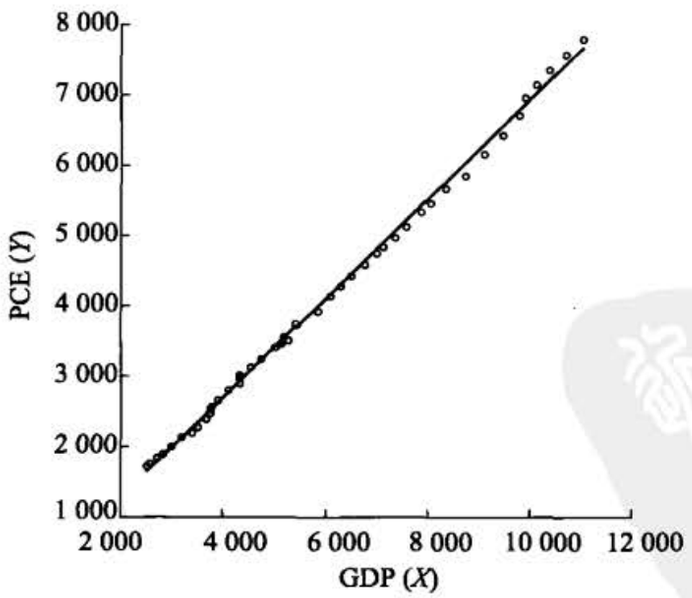


图 I—3　1960—2005 年间个人消费支出（Y）与国内生产总值（X）的关系
（均以 2000 年十亿美元为单位）

□ 5. 计量经济模型的参数估计

有了数据之后，下一步的任务就是估计消费函数中的参数。参数的数值估计将对消费函数赋予经验内容。估计参数的具体步骤将在第 3 章中说明。这里仅指出，**回归分析**（regression analysis）的统计学方法是获得估计值的主要手段。利用这种方法以及表 I—1 所给的数据，我们便得到 β_1 和 β_2 的估计值为 −299.591 3 和 0.721 8。于是所估计的消费函数是

$$\hat{Y}_t = -299.591\ 3 + 0.721\ 8X_t \qquad (I.3.3)$$

Y 顶上的帽符表示它是估计值。[①] 图 I—3 给出了估计的消费函数（即回归线）。

如图 I—3 所示，因为数据点很靠近回归线，所以回归线对数据拟合得相当好。从图 I—3 我们发现，在 1960—2005 年期间，斜率系数（即 **MPC**）约为 0.72，表明在此样本期间，真实收入每增加 1 美元，平均而言，真实消费支出将增加约 72 美分。[②] 我们说平均而言，是因为消费和收入之间没有准确的关系。这一点可以从图 I—3 看出：并非所有数据点都恰好位于回归线上。我们可以简单地说，根据我们的数据，真实收入每增加 1 美元，平均消费支出或消费支出均值会增加约 72 美分。

□ 6. 假设检验

假定所拟合的模型是现实的一个较好的近似，还必须制定适当的准则，借以判断如方程（I.3.3）中的估计值是否与待检验的理论预期值相一致。根据米尔顿·弗里德曼（Milton Friedman）这样的“实证”经济学家的意见，凡是不能通过经验证据来证实的理论或假设，都不可作为科学探索的一个部分。[③]

如前所述，凯恩斯曾预期 MPC 是正的，但小于 1。在我们的例子中，我们求得 MPC 约为 0.72。但在把这一发现看做是对凯恩斯消费理论的认可之前，还要追问这一估计值是否充分地低于 1，以使我们不再怀疑这个估计值仅是一次偶然的机会得来，或者怀疑我们用的数据太特殊了。换言之，0.72 是不是在统计意义上小于 1？如果是，就可用来支持凯恩斯理论。

以样本证据为依据去肯定或否定经济理论，是以所谓**统计推断**（statistical inference，即假设检验 hypothesis testing）这个统计理论分支为基础的。在本书中，我们会处处看到这种推断过程实际上是如何进行的。

□ 7. 预报或预测

如果所选的模型肯定了我们所考虑的假说或理论，就可以根据解释变量或**预测**

---

① 在一个变量或参数的上方画一个尖帽符号以表示它是一个估计值，已成为惯例。

② 第 3 章将表明，**最小二乘统计方法**（least squares）给出了这些估计值。这里暂且不去管它们是怎样得来的，而且也不必问为什么截距是负值。

③ 参看 Milton Friedman，“The Methodology of Positive Economics，” *Essays in Positive Economics*，University of Chicago Press，Chicago，1953。

变量（predictor variable）X 的已知或预期未来值，来预测因变量或**预报变量**（forecast variable）Y 的未来值。

为便于说明，假设我们想预测 2006 年的平均消费支出。2006 年 GDP 的值为 113 194 亿美元。[①] 将 GDP 的这个数字代入（I. 3. 3）的右边，我们得到：

$$\hat{Y}_{2006} = -299.591\ 3 + 0.721\ 8 \times 11\ 319.4 = 7\ 870.751\ 6 \qquad \text{(I. 3. 4)}^*$$

即约 78 700 亿美元。因此，给定 GDP 的值，预测的平均消费支出或消费支出的均值约为 78 700 亿美元。2006 年报告的实际消费支出值为 80 440 亿美元。于是估计模型（I. 3. 3）约**低估**（underpredicted）了实际消费支出 1 740 亿美元。我们可以说**预测误差**（forecast error）约为 1 740 亿美元，占 2006 年实际 GDP 值的 1.5%。当我们在以后章节详尽讨论了线性回归模型之后，我们会发现这样的误差是“小”还是“大”。但目前重要的是注意到，给定这种分析的统计性质，这种预测误差无法避免。

估计模型（I. 3. 3）还有另外一个用处。假设总统决定减少所得税。这种政策对收入及消费支出和最终就业会有什么影响呢？

假如政策改变的结果是投资有所下降，其对经济的影响将如何？宏观经济理论告诉我们，投资支出每改变 1 美元，收入的改变由**收入乘数**（income multiplier，M）：

$$M = \frac{1}{1 - \text{MPC}} \qquad \text{(I. 3. 5)}$$

给出。如利用由方程（I. 3. 3）得到的 MPC＝0. 72，此乘数就变成 M＝3. 57。也就是说，投资减少（增加）1 美元，将最终导致收入减少（增加）4 倍之多；注意，乘数的实现需要时间。

在这一计算中 MPC 是个关键值，因为乘数的大小取决于它。但 MPC 的估计来自诸如（I. 3. 3）的回归模型。所以，MPC 的数量估计为政策的制定提供了有价值的信息。一旦获知 MPC，即可跟踪政府财政政策的改变，预测收入和消费支出的未来变化过程。

□ 8. 利用模型进行控制或制定政策

若我们已估计出由方程（I. 3. 3）给出的凯恩斯消费函数，而且政府认为 87 500 亿美元（以 2000 年美元计）的（消费）支出水平即可维持当前约 4.2%的失业率（2006 年初），那么，什么样的收入水平将保证消费支出达到这个目标水平呢？

如果消费函数（I. 3. 3）是合理的，简单的数学运算就得到：

$$8\ 750 = -299.591\ 3 + 0.721\ 8(\text{GDP}_{2006}) \qquad \text{(I. 3. 6)}$$

解得 $X \approx 12\ 537$。也就是说，给定 MPC 约为 0. 72，125 370 亿美元的收入水平将导致约 87 500 亿美元的消费支出。

---

① 有 2006 年 PCE 和 GDP 的数据可用，但我们在说明本节所讨论的专题时故意不用。如同我们在以后章节中将讨论的那样，留下一部分数据用来检查拟合模型对样本外观测的预测力如何，是一个很好的主意。

* 公式中的单位均为十亿美元，后同。——译者注

上述计算提示我们，一个已估计出来的模型可服务于控制或政策的目的。通过适当的财政与货币政策的配合，政府可操纵**控制变量**（control variable）X 以实现**目标变量**（target variable）Y 的某个理想水平。

图 I—4 概括了经典计量经济学的建模方法。

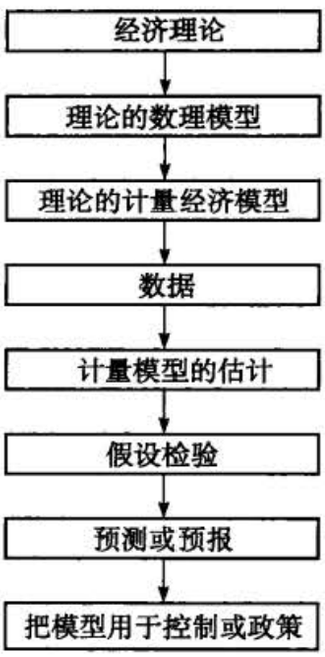


图 I—4　计量经济学建模剖视

□ 在竞争的模型之间进行选择

当一家政府机构（如美国商务部）搜集经济数据（如表 I—1 中所示）时，它们不一定有什么经济理论。那人们如何知道这些数据实际上是支持凯恩斯消费理论的呢？是因为图 I—3 所示的凯恩斯消费函数（即回归线）与实际数据点极为接近吗？还有其他的消费模型（理论）能同样好地拟合这些数据吗？比如，米尔顿·弗里德曼提出了一个被称为持久收入假说（permanent income hypothesis）[①] 的消费模型。罗伯特·霍尔（Robert Hall）也提出了一个被称为生命周期持久收入假说（life-cycle permanent income hypothesis）[②] 的消费模型。这些模型中有没有一两个也能拟合表 I—1 中的数据呢？

简言之，实践中的研究者面临的问题是，给定一种现象，如消费—收入关系，如何在几个竞争的假设或模型之间做出选择。如米勒（Miller）所据理力争的那样：

除非假设比某些自然的对手在处理数据上有更好的表现，否则，不针对具

① Milton Friedman, *A Theory of Consumption Function*, Princeton University Press, Princeton, N. J., 1957.

② R. Hall, "Stochastic Implications of the Life Cycle Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, vol. 86, 1978, pp. 971-987.

体数据就不可能真正证实原假设……这里，正是原假设的胜利，同时也是貌似可信的对立假设的失败强化了假设。[①]

那么，人们如何在竞争的模型或假设中进行选择呢？这里要记住克莱夫·格兰杰（Clive Granger）的建议[②]：

> 我想给的建议是，在你提出一种新理论或新的经验模型时，你要考虑这些问题：
>
> (i) 它的用途是什么？它有助于什么样的经济决策？
>
> (ii) 在已经提出的证据中，有没有某个证据让我能将这种新理论或新模型与其他理论或模型做比较？
>
> 我认为关注这样的问题将会加强经济研究或讨论。

通览全书，在解释各种经济现象时，我们将遇到几个竞争假设。比如，经济学的学生都熟悉生产函数的概念，它基本上就是指产出与投入（如资本和劳动）之间的关系。在文献中，最有名的两个就是柯布-道格拉斯生产函数和常替代弹性生产函数。给定投入和产出数据，如果可能的话，我们需要知道这两个生产函数中的哪一个能很好地拟合数据。

在能用于检验这些竞争假设的意义上，上述八步骤经典计量经济方法论是中性的。

有没有可能提出一种足以包含这些竞争假设的综合方法论呢？这是一个复杂而又有争议的问题。我们将在掌握了必要的计量经济理论之后，在第 13 章来讨论这个问题。

I.4 计量经济学的类型

如图 I—5 中的分类框架所示，计量经济学可划分为两大类：**理论计量经济学**（theoretical econometrics）和**应用计量经济学**（applied econometrics）。在每一大类中均可按**经典方法**（classical）或**贝叶斯方法**（Bayesian）进行研究。本书的重点在于经典方法。至于贝叶斯方法，读者可参阅本章末所附参考文献。

理论计量经济学是要找出适当的方法，去测度由计量经济模型设定的经济关系。为此，计量经济学家非常依赖于数理统计。例如，本书中广泛使用的方法之一是**最小二乘法**（least squares）。理论计量经济学家必须明确这一方法所涉及的假定、这一方法的性质，以及当某些假定不成立时，这些性质将会受到什么样的影响。

在应用计量经济学中，我们利用理论计量经济学工具去研究经济学或管理学中

---

① R. W. Miller, *Fact and Method: Explanation, Confirmation, and Reality in the Natural and Social Sciences*, Princeton University Press, Princeton, N. J., 1978, p. 176.

② Clive W. J. Granger, *Empirical Modeling in Economics*, Cambridge University Press, U. K., 1999, p. 58.

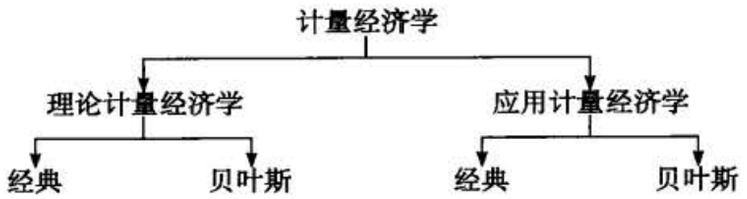


图 I—5　计量经济学分类

的某些特殊领域，诸如生产函数、投资函数、供求函数和证券组合理论等。

本书主要讨论计量经济学方法的发展、假定、用途及其局限性，并引用经济学和管理学各个领域的例子来加以说明。然而，这不是一本应用计量经济学的书，它并不深入研究任何一个特殊的经济应用领域。这种工作最好留给那些致力于专门研究的著作，本书末尾的文献提供了一些这样的著作。

I.5　数学与统计学预备知识

虽然本书是在一个初等水平上写作的，但作者仍假定读者熟悉一些统计估计和假设检验的基本概念。不过，为了便于读者重新复习有关内容，对于本书中所使用的基本统计概念，附录 A 提供了一个宽泛而又不太深入的概述。至于数学方面，希望读者对微积分的概念还不陌生，虽然这也不是必要的。在许多研究生用的计量经济学书中大量使用矩阵代数，作者在这里声明，阅读本书并无此必要。作者坚信，讲授计量经济学的基本思想并不需要使用矩阵代数。然而，为了顾及喜欢数学的学生，附录 C 仍然摘要地给出了基本回归理论的矩阵表述。对于这些学生，附录 B 还简明扼要地总结了矩阵代数的一些主要结论。

I.6　计算机的作用

当今，回归分析已是计量经济学的家常便饭，没有计算机和某些统计软件是难以想象的。（但请相信我，我本人却是在计算尺时代成长起来的!）幸亏，现在已有很多优秀的商用回归软件，对于大型计算机和个人微型计算机都适用，并且与日俱增。诸如 ET、LIMDEP、SHAZAM、MICROTSP、MINITAB、EVIEWS、SAS、SPSS、STATA、Microfit、PcGive 和 BMD 等回归软件，都包含了本书所讨论的大多数计量经济学方法和检验。

本书有时要求读者用一种或多种统计软件包做**蒙特卡罗**（Monte Carlo）模拟实验。蒙特卡罗实验是一些“有趣”的练习，它能使读者很好地体会本书所讲的几种

统计方法的性质。在适当的地方，我们还将详细讨论蒙特卡罗实验。

1.7 进一步阅读建议

计量经济学方法论是一个非常广泛且富有争议的论题。对有兴趣的读者，我建议阅读以下几本书：

Neil de Marchi and Christopher Gilbert, eds., *History and Methodology of Econometrics*, Oxford University Press, New York, 1989. 本书收集了早期的一些关于计量经济方法论的读物，对关于时间序列数据（即对同一个研究对象在不同时期收集的数据）的英式计量经济学方法有广泛的讨论。

Wojciech W. Charemza and Derek F. Deadman, *New Directions in Econometric Practice: General to Specific Modelling, Cointegration and Vector Autogression*, 2d ed., Edward Elgar Publishing Ltd., Hants, England, 1997. 作者们批判了传统计量经济学方法，并详细阐释了计量经济学方法论的新动向。

Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics*, Edward Elgar Publishing Ltd., Hants, England, 1990. 该书对计量经济学的各种方法论作了一个较为不偏不倚的论述，并表示要重新加盟到传统的计量经济学方法论中去。

Mary S. Morgan, *The History of Econometric Ideas*, Cambridge University Press, New York, 1990. 作者对计量经济学的理论和实践作了出色的历史剖析，还对哈维尔莫（Haavelmo，1990 年诺贝尔经济学奖得主）对计量经济学的早期贡献作了深入的讨论。出于同样的想法，David F. Hendry and Mary S. Morgan, *The Foundation of Econometric Analysis*, Cambridge University Press, U. K., 1995. 也搜集了计量经济学研讨会上的作品，以说明计量经济思想随着时间的演进过程。

David Colander and Reuven Brenner, eds., *Educating Economists*, University of Michigan Press, Ann Arbor, Michigan, 1992. 该书对经济学教学和实践提出了尖锐的，有时是不可知论的看法。

关于贝叶斯统计学和贝叶斯计量经济学，如下书目很有用处：John H. Dey: *Data in Doubt*, Basil Blackwell Ltd., Oxford University Press, England, 1985; Peter M. Lee, *Bayesian Statistics: An Introduction*, Oxford University Press, England, 1989; Dale J. Porier, *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge, Massachusetts, 1995. 一本高深的参考书是 Arnold Zellner, *An Introduction to Bayesian Inference in Econometric*, John Wiley & Sons, New York, 1971. 另一本比较高深的参考书是 *Palgrave Handbook of Econometrics*: Volume 1: *Econometric Theory*, edited by Terence C. Mills and Kerry Patterson, Palgrave Macmillan, New York, 2007.

第1篇

单方程回归模型

本书的第1篇介绍各种单方程回归模型。在这些模型中，一个因变量被表达为一个或多个所谓解释变量的线性函数。在这样的模型中有如下隐含假定：如果在因变量与解释变量之间存在某种因果关系的话，这个关系只有一个流向，就是从解释变量到因变量。

在第1章中，我们讨论回归的历史和现代含义。并用几个取自经济学和其他学科的例子来说明这两种含义的区别。

在第2章中，我们借助于双变量线性回归模型来介绍回归分析的一些基本概念。双变量线性模型，是指其中的因变量被表达成仅仅一个解释变量的线性函数。

在第3章中，我们继续同双变量模型打交道，并引进以经典线性回归模型为名的一种涉及若干简化假定的模型。有了这些假定，我们随即介绍普通最小二乘（ordinary least squares，OLS）方法，用以估计双变量回归模型中的参数。OLS易于应用，且有一些非常良好的统计性质。

在第4章中，我们介绍（双变量）经典正态线性回归模型。它假定随机因变量遵循正态概率分布。有了这一假定，在第3章中得到的OLS估计量，就有了一些比非正态经典线性回归模型更强的统计性质，使我们能从事于统计推断即假设检验。

第5章专门讨论假设检验。这里，我们试图分辨回归系数的估计值是否与它们的假设值相符（无矛盾），假设值是指由理论或先前的经验工作提出来的值。

第6章考虑由双变量回归模型延伸出来的一些细节问题。具体地说，这一章讨论如下问题：(1) 过原点回归，(2) 尺度与测量单位，以及 (3) 回归模型的函数形式，如双对数、半对数和倒数等模型。

在第7章中，我们考虑含有不止一个解释变量的多元回归或多变量回归模型。并说明怎样能把OLS推广应用到这种模型的参数估计上。

第8章把第5章介绍的概念推广到多元回归模型上，并指出由多个解释变量的引入而诱发的若干复杂性。

第1篇以讨论虚拟或定性解释变量的第9章作为结束。这一章强调，并不是所有的解释变量都必须是定量的（即比率尺度）。虽然诸如性别、种族、宗教、国籍和居住地等变量都不容易量化，但它们在解释许多经济现象时却起到重要作用。

第 1 章 回归分析的性质

如导言中所说，回归是计量经济学的主要工具。本章中，我们将扼要地考虑这一工具的性质。

1.1 “回归”一词的历史渊源

回归一词最先由弗朗西斯·高尔顿（Francis Galton）引人。在一篇著名的论文中，高尔顿发现，虽然有一个趋势，父母高，儿女也高；父母矮，儿女也矮，但给定父母的身高，儿女辈的平均身高却趋向于或者“回归”到全体人口的平均身高。[①] 换言之，尽管父母双亲都异常高或异常矮，而儿女的身高则有走向人口总体平均身高的趋势。高尔顿的普遍回归定律（law of universal regression）还被他的朋友卡尔·皮尔逊（Karl Pearson）证实。皮尔逊曾收集过一些家庭群体的一千多名成员的身高记录。[②] 他发现，对于一个父亲高的群体，儿辈的平均身高低于他们父辈的身高，而对于一个父亲矮的群体，儿辈的平均身高则高于其父辈的身高。这样就把高的和矮的儿辈一同“回归”到所有男子的平均身高。用高尔顿的话说，这是“回归到中等”（regression to mediocrity）。

① Francis Galton, “Family Likeness in Stature,” *Proceedings of Royal Society*, London, vol. 40, 1886, pp. 42-72.

② K. Pearson and A. Lee, “On the Laws of Inheritance,” *Biometrika*, vol. 2, 1903, pp. 357-462.

1.2 回归的现代含义

然而，对回归的现代解释却非常之不同，大致上，我们可以这样说：

回归分析是关于研究一个所谓的因变量对另一个或多个所谓解释变量的依赖关系，其用意在于通过后者（在重复抽样中）的已知或设定值，去估计和（或）预测前者的（总体）均值。

对回归分析的这种看法的全部含义，将随着本书的进程而渐明。但是用少数几个简单的例子，能把回归的基本概念弄得一清二楚。

□ 例子

1. 再次考虑高尔顿的普遍回归定律。高尔顿的兴趣在于发现为什么人口的身高分布有一种稳定性。但从现代观点考虑，我们并不关心这种解释。我们关心的是，给定父辈身高的情形下找出儿辈平均身高的变化。换言之，我们关心一旦知道了父辈的身高，怎样预测儿辈的平均身高。为了看清楚怎样能做到这一点，考虑图 1—1 这个**散点图**（scatter diagram，scattergram）。该图展示了对应于设定的父亲身高，儿子在一个假想人口总体中的身高分布。注意，对应于任一给定的父亲身高，都有着儿子身高的一个（分布）范围。然而，值得注意的是，随着父亲身高的增加，儿子的平均身高尽管有所波动，但总体上是增加的。为了看得明白无误，图中加圈的叉号表示对应于给定父辈身高情况下儿辈的平均身高。连接这些平均身高，我们就得到图中所示的直线。我们以后会知道，这条线就叫做**回归线**（regression line）。它表明了儿子的平均身高是怎样随父亲身高的增加而增加的。①

2. 考虑图 1—2 中的散点图。这是在不同的固定年龄处测度的一个假想的男孩身高总体的分布。注意，对应于任一给定年龄，都有一个身高的范围。显然，同一个给定年龄的男孩不会完全一样高。但身高随年龄的增加而增加（当然，只到一定的年龄为止），如果我们通过表示给定年龄下平均身高的加圈圆点画一条线（回归线），就可以清楚地看出这一点。于是，知道了年龄，就能预测相当于这个年龄的平均身高。

3. 转到经济学中的例子。经济学家也许想研究个人消费支出对税后或可支配的个人真实收入的依赖关系。这种分析会有助于估计边际消费倾向，即真实收入改变 1 美元导致消费支出的平均变化。

---

① 在此主题讨论的现阶段，我们干脆把这条回归线称为：对应于给定解释变量（父亲的身高）值，因变量（儿子的身高）的均值或平均值连线。注意，此线有一正的斜率；但此斜率小于 1，这和高尔顿的“回归到中等”相一致。（为什么？）

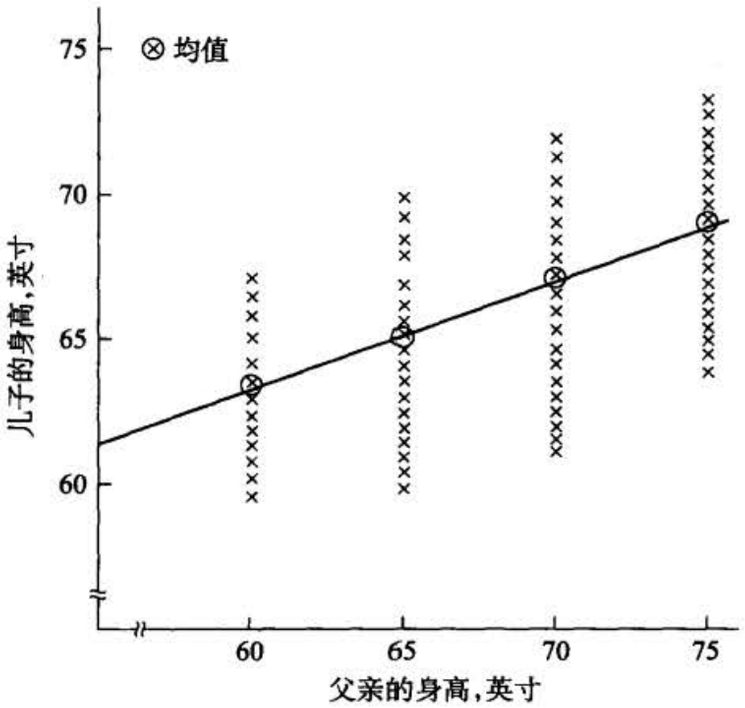


图 1—1　给定父亲身高时儿子身高的假想分布

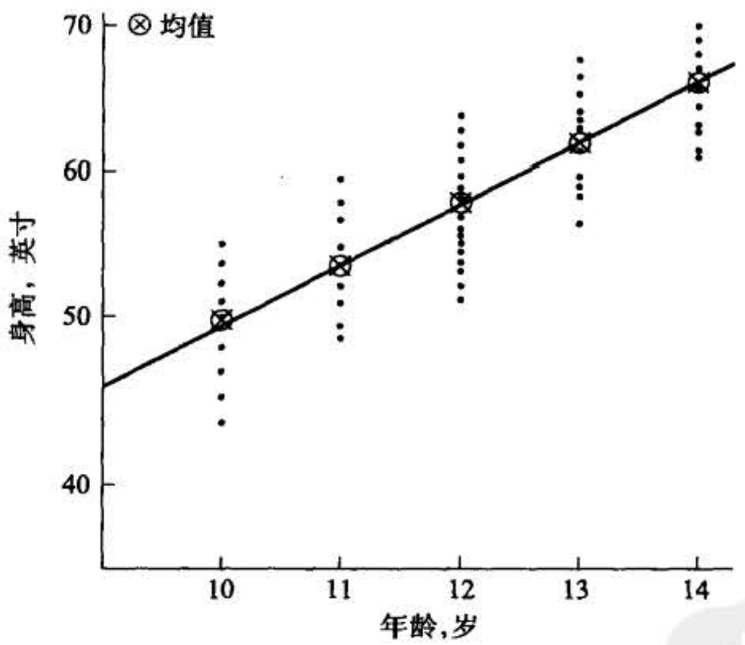


图 1—2　对应于选定年龄的假想身高分布

4. 一位能设定价格或产出（但不能同时设定两者）的垄断商，也许想知道产品需求对价格变化的反应，通过这种定价实验，也许能估计出产品需求的**价格弹性**（price elasticity，即对价格做出反应的敏感程度），从而有助于确定最有利可图的价格。

5. 一位劳动经济学家也许要研究货币工资变化率与失业率的关系。图 1—3 给出

了历史数据所表现的散点图。图中的曲线是把货币工资变化同失业率联系起来的著名菲利普斯曲线（Phillips curve）之一例。这样的散点图能使劳动经济学家在给定失业率下预测货币工资的平均变化。这种知识还有助于认识一个国家范围内的通货膨胀过程，因为货币工资的增长很可能在上涨的物价中得到反映。

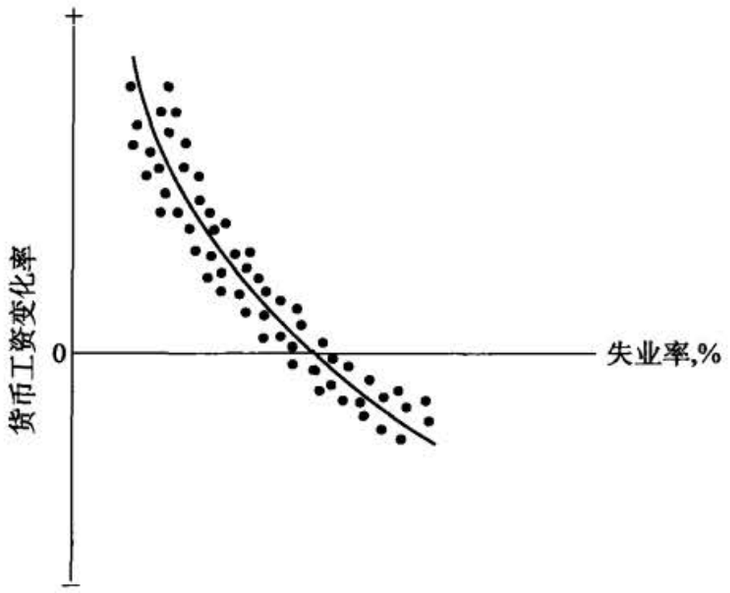


图 1—3　假想的菲利普斯曲线

6. 根据货币经济学，其他条件不变，通货膨胀率 π 越高，人们愿意以货币形式持有的收入比例 k 越低，如图 1—4 所示。这条线的斜率就表示给定通货膨胀率变化的情况下 k 的变化率。对这种关系作一数量分析，将使货币经济学家能够在不同通货膨胀率下预测人们愿意以货币形式持有的收入比例。

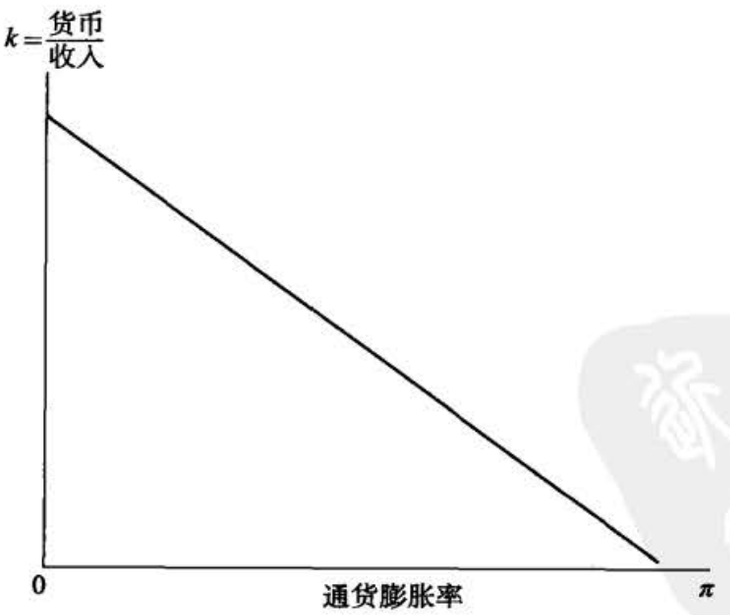


图 1—4　货币持有与通货膨胀率 π 的关系

7. 公司的销售部经理想知道人们对公司产品的需求与（比方说）广告支出的关系。这种研究在很大程度上有助于算出相对于广告支出的**需求弹性**（elasticity of demand），即广告费预算每变化百分之一导致需求变化的百分数。这种知识有助于制定“最优”的广告费预算。

8. 最后，也许农业经济学家想研究作物（比方说小麦）收成对气温、降雨量、阳光和施肥量的依赖关系。这种依赖关系使他能在给定解释变量信息的情况下预测或预报作物的平均收成。

相信读者能提供关于一个变量依赖于另一个或多个变量的大量事例。本书讨论的回归分析方法，就是要用来研究变量之间的这种依赖关系。

1.3 统计关系与确定性关系

读者能从 1.2 节所举的例子看到，不像经典物理学中考虑的那种变量之间的函数或确定性依赖关系，在回归分析中，我们考虑的是一种所谓统计依赖关系。在变量之间的统计关系式中，我们主要处理的是**随机**（random 或 stochastic）变量[①]，也就是有着概率分布的变量。但是在函数或确定性依赖关系中，我们要处理的变量不是随机的。

例如，作物收成对气温、降雨量、阳光以及施肥量的依赖关系是统计性质的。这个性质的意义在于：这些解释变量固然重要，但并不能使农业经济学家准确地预测作物的收成。一则对这些变量的测量有误差，二则还有一大堆整体地影响着收成的因素（变量），却难于一一辨认出来。因此，无论我们考虑了多少个解释变量，却无法完全地解释作物收成这个因变量。它的一些“内在”的或随机的变异是注定存在的。

另一方面，在确定性现象中，我们处理这样一类关系式，比如说，牛顿的万有引力定律所表示的关系式：宇宙间的每个粒子都相互吸引，其引力与它们的质量乘积成正比，而与它们之间距离的平方成反比，用符号表示就是 $F=k(m_1m_2/r^2)$，其中 F=引力，m_1 和 m_2 为两个粒子的质量，r=距离，而 k=比例常数。另一个例子是欧姆定律：对于金属导体在有限的温度范围内，电流 C 正比于电压 V，即 $C=V/k$，其中 $1/k$ 是比例常数。这类确定性现象的其他例子包括波以耳（Boyle）气体定律，基尔霍夫（Kirchhoff）的电流定律和牛顿的运动定律等。

在本书中，我们不去研究这类确定性现象。当然，如果，比方说，在牛顿的万有引力定律中，k 的测量有误差，则原来的确定性关系就变成了一个统计关系式。这

① “stochastic”（随机）一词源自希腊字 stokhos，意谓“公牛的眼睛”。把矛抛向一块盾牌的结果是一个随机过程，即一个带有击不中可能性的过程。

时，引力只能按给定的 k 值（还有 m_1，m_2 和 r）近似地加以预测。于是，变量 F 变成了一个随机变量。

1.4 回归与因果关系

虽然回归分析研究一个变量对另一（些）变量的依赖关系，但它并不一定意味着因果关系。用肯德尔（Kendall）和斯图亚特（Stuart）的话说："一个统计关系式，不管多强也不管多么有启发性，永远不能确立因果方面的联系：对因果关系的理念，必须来自统计学以外，最终来自这种或那种理论。"①

在前面所引的农作物收成一例中，没有任何统计上的理由可以认为降雨量不依赖于作物收成。我们把作物收成看作依赖于降雨量等的因变量，并非出于统计上的考虑。普通常识提示了我们不能把这种关系倒转过来，因为我们不能用改变作物收成的方法来控制降雨。

所有 1.2 节引用的例子都指出一个要点：**从逻辑上说，统计关系式本身不可能意味着任何因果关系。**要谈因果律，必须诉诸先验的或理论上的思考。例如，在上面所引的第三个例子中，我们说消费支出依赖于实际收入，是引用了经济理论的。②

1.5 回归与相关

与回归分析密切相关而在概念上明显不同的，是以测度两个变量之间的线性关联程度为其主要目的的**相关分析**（correlation analysis）。第 3 章中我们将要详细讨论的**相关系数**（correlation coefficient）就是用来测度这种（线性）关联强度的。例如，我们也许有兴趣去求吸烟与肺癌之间、统计学考分与数学考分之间、中学成绩与大学成绩之间的相关（系数）等。而在回归分析中，如前所述，我们并不主要对这种度量感兴趣。感兴趣的却是试图根据其他变量的设定值来估计或预测某一变量的平均值。例如，我们也许想知道能否从一个学生的已知数学考分，去预测他的统计学平均考分。

回归和相关有一些值得提出的基本分歧。在回归分析中，对因变量和解释变量的处理方法存在着不对称性。因变量被当作是统计的，随机的，也就是它有一个概

① M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Charles Griffin Publishers, New York, vol. 2, 1961, chap. 26, p. 279.

② 但在第 3 章中我们将看到，经典回归分析中假定了用于分析的模型是正确的。因此，在假设模型中，隐含着因果方向。

率分布。而解释变量则被看作是（在重复抽样中）取固定值的。[①] 这点在 1.2 节所给的回归定义中已说明白。因此，在图 1—2 中，我们假定年龄变量被固定在给定的水平上，而身高则是在这些水平上度量的。但在相关分析中，我们对称地对待任何（两个）变量；因变量和解释变量之间不加区别。毕竟，数学考分与统计学考分之间的相关就是统计学考分与数学考分之间的相关。此外，两个变量都被看作是随机的，如我们将会看到的，大部分相关理论都建立在变量的随机性假定之上。但是，本书要阐述的回归理论的大部分均以下述假定为条件：因变量是随机的，而解释变量是固定的或非随机的。[②]

1.6 术语与符号

在我们进入正式的回归理论分析之前，先来斟酌一下有关术语与符号的问题。因变量和解释变量两个名词在文献中都有过种种其他描述。一个有代表性的清单如下：

因变量（Dependent variable）	**解释变量**（Explanatory variable）
⇕	⇕
被解释变量（Explained variable）	*自变量（Independent variable）*
⇕	⇕
预测子（Predictand）	预测元（Predictor）
⇕	⇕
回归子（Regressand）	**回归元**（Regressor）
⇕	⇕
响应（Response）变量	刺激（Stimulus）变量
⇕	⇕
内生（Endogenous）变量	外生（Exogenous）变量
⇕	⇕
结果（Outcome）变量	协变量（Covariate）
⇕	⇕
被控变量（Controlled variable）	控制变量（Control variable）

虽然采用什么名词术语是一个个人爱好和传统习惯问题，但本书中我们采用的术语是因变量/解释变量或更中性的回归子和回归元。

如果我们在研究一个变量对仅仅一个解释变量的依赖关系，如消费支出对实际

① 须知，解释变量可能本来就是随机的。但出于回归分析的目的，我们假定它们的值在重复抽样中固定不变（即 X 在不同的多个样本中取同样的一组值），从而把它们转化成实质上非随机的。对此，第 3 章 3.2 节有更多的讨论。

② 在高级计量经济学教材中，解释变量非随机的这一假定可以去掉（见第 2 篇引言）。

收入的依赖，则称这种研究为简单或**双变量回归分析**（two-variable regression analysis）。但是，如果我们在研究一个变量对多于一个解释变量的依赖关系，有如农作物收成依赖于气温、降雨量、阳光和施肥量一例，则称之为**多元回归分析**（multiple regression analysis）。换句话说，在双变量回归中只有一个解释变量，而在多元回归中则有不止一个解释变量。

"Random"和"Stochastic"这两个单词是同义语，都是随机的意思。如前所述，一个随机变量是指这样的一个变量：它以给定的概率取任一特定数值，可正可负。[①]

除非另作声明，否则字母 Y 一律指因变量，而 X（X_1，X_2，…，X_k）一律指解释变量。其中 X_k 代表第 k 个解释变量。下标 i 或 t 则指第 i 次或第 t 次观测。这样，X_{ki}（或 X_{kt}）就指对变量 X_k 的第 i（或 t）次观测。N（或 T）指总体中的观测总个（次）数，而 n（或 t）则指样本中的观测值总个数。作为一种惯例，观测值下标 i 将用于**横截面数据**（cross-sectional data）（即在一个时间点上收集的数据），而下标 t 将用于**时间序列数据**（time series data）（即对同一个研究对象在不同时期收集的数据）。关于横截面数据和时间序列数据的性质以及经验分析所用数据的性质与来源等重要议题，将在下节讨论。

1.7 经济分析所用数据的性质与来源[②]

任何计量经济分析的成功最终都有赖于适当数据的获得。因此，我们有必要花点时间，来讨论一下经验分析中所遇到的数据的性质、来源及其局限性。

□ 数据类型

用于经济分析的数据有三类：**时间序列**（time series）、**横截面**（cross-section）以及**混合**（pooled，时间序列与横截面合并）数据。

时间序列数据。引言的表 I—1 展示的数据就是时间序列数据之一例。一个时间序列是对一个变量在不同时间取值的一组观测结果。这些数据可以是在有规则的时间间隔收集的。譬如**每日**（daily，如股票价格和天气预报），**每周**（weekly，如货币供给数字），**每月**（monthly，如失业率和消费者价格指数），**每季度**（quarterly，如 GDP），**每年**（annually，如政府预算），**每 5 年**（quinquennially，如制造业普查资料），**每 10 年**（decennially，如人口普查资料），有些数据每季度和每年都有公布，如 GDP 和消费者支出数据。随着高速计算机的出现，可以搜集极短的时间区间内的数据，如股票价格数据，几乎可以得到其连续数据（所谓的实时牌价）。

① 正式定义和更多的细节见附录 A。

② 一个富于信息的叙述，见 Michael D. Intriligator, *Econometric Models, Techniques, and Applications*, Prentice Hall, Englewood Cliffs, N. J., 1978, Chap. 3。

虽然许多计量经济研究都使用时间序列数据，但它们的使用给计量经济学家提出了特殊的问题。如我们在以**时间序列计量经济学**（time series econometrics）为名的篇章中所论，基于时间序列数据的经验研究，大多假定所依据的时间序列是**平稳的**（stationary）。虽然要介绍平稳性准确的技术含义为时尚早，但粗略地说，如果一个时间序列的均值和方差不随时间而系统地变化，那它就是平稳的。为了看出其含义，考虑图 1—5，它描绘的是美国从 1951 年 1 月 1 日至 1999 年 9 月期间 M1 货币供给的行为。（实际数据在习题 1.4 中给出。）如你从图中所见，随着时间的推移，M1 货币供给表现出稳定上升的**趋势**（trend）和逐年波动的特征，这就表明 M1 不是时间序列平稳的。[①] 我们在第 21 章将详尽讨论这一点。

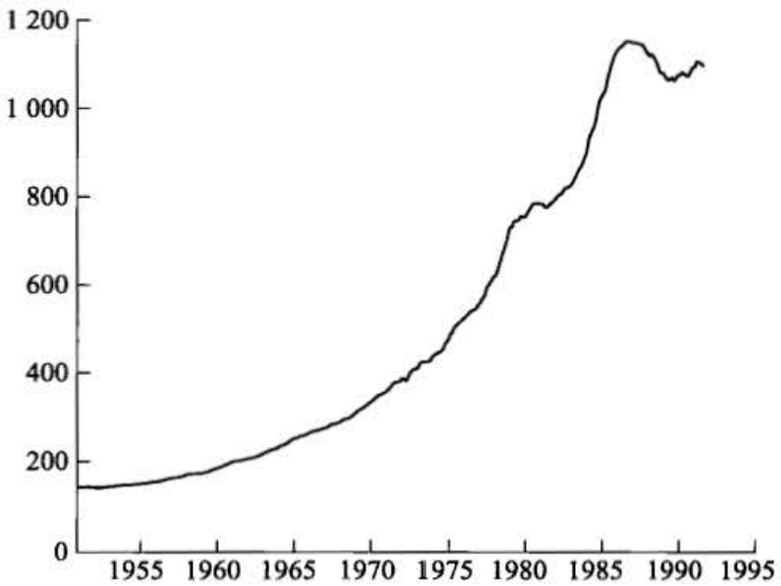


图 1—5　1951 年 1 月—1999 年 9 月美国的 M1 货币供给

横截面数据。横截面数据指对一个或多个变量在同一时间点上收集的数据。诸如人口普查局每 10 年进行一次的人口普查（最近一次是在 2000 年），密歇根大学举办的消费者支出普查。当然，由盖洛普（Gallup）和其他机构主导的一类民意调查更是属于这类数据。表 1—1 给出了横截面数据的一个实例。该表给出 1990 年和 1991 年美国 50 个州的鸡蛋产量和鸡蛋价格。每一年份 50 个州的数据就构成一个横截面数据样本。这样，表 1—1 中就有两个横截面样本。

表 1—1　　美国蛋类生产

州	Y_1	Y_2	X_1	X_2	州	Y_1	Y_2	X_1	X_2
AL	2 206	2 186	92.7	91.4	MT	172	164	68.0	66.0
AK	0.7	0.7	151.0	149.0	NE	1 202	1 400	50.3	48.9

① 为了更清楚地看出这一点，我们把数据分为四个时期：1951 年 1 月至 1962 年 12 月；1963 年 1 月至 1974 年 12 月；1975 年 1 月至 1986 年 12 月和 1987 年 1 月至 1999 年 9 月。对这些子期间，货币供给的均值（括号中是相应的标准差）分别是 165.88（23.27）、323.20（72.66）、788.12（195.43）和 1 099（27.84），所有数字都以十亿美元为单位。这大致表明了货币供给在整个期间非平稳的事实。

续前表

州	Y_1	Y_2	X_1	X_2	州	Y_1	Y_2	X_1	X_2
AZ	73	74	61.0	56.0	NV	2.2	1.8	53.9	52.7
AR	3 620	3 737	86.3	91.8	NH	43	49	109.0	104.0
CA	7 472	7 444	63.4	58.4	NJ	442	491	85.0	83.0
CO	788	873	77.8	73.0	NM	283	302	74.0	70.0
CT	1 029	948	106.0	104.0	NY	975	987	68.1	64.0
DE	168	164	117.0	113.0	NC	3 033	3 045	82.8	78.7
FL	2 586	2 537	62.0	57.2	ND	51	45	55.2	48.0
GA	4 302	4 301	80.6	80.8	OH	4 667	4 637	59.1	54.7
HI	227.5	224.5	85.0	85.5	OK	869	830	101.0	100.0
ID	187	203	79.1	72.9	OR	652	686	77.0	74.6
IL	793	809	65.0	70.5	PA	4 976	5 130	61.0	52.0
IN	5 445	5 290	62.7	60.1	RI	53	50	102.0	99.0
IA	2 151	2 247	56.5	53.0	SC	1 422	1 420	70.1	65.9
KS	404	389	54.5	47.8	SD	435	602	48.0	45.8
KY	412	483	67.7	73.5	TN	277	279	71.0	80.7
LA	273	254	115.0	115.0	TX	3 317	3 356	76.7	72.6
ME	1 069	1 070	101.0	97.0	UT	456	486	64.0	59.0
MD	885	898	76.6	75.4	VT	31	30	106.0	102.0
MA	235	237	105.0	102.0	VA	943	988	86.3	81.2
MI	1 406	1 396	58.0	53.8	WA	1 287	1 313	74.1	71.5
MN	2 499	2 697	57.7	54.0	WV	136	174	104.0	109.0
MS	1 434	1 468	87.8	86.7	WI	910	873	60.1	54.0
MO	1 580	1 622	55.4	51.5	WY	1.7	1.7	83.0	83.0

注：Y_1=1990 年鸡蛋产量（百万个）；Y_2=1991 年鸡蛋产量（百万个）；

X_1=1990 年每打鸡蛋的价格（美分）；X_2=1991 年每打鸡蛋的价格（美分）。

资料来源：*World Almanac*，1993，p. 119. 数据来自美国农业部经济研究服务部门。

正如时间序列数据由于平稳性问题而带来了它独有的问题，横截面数据也有其自身的问题，特别是异质性（heterogeneity）问题。我们从表 1—1 中给出的数据可以看出，有些州生产鸡蛋的数量巨大（如宾夕法尼亚州），而有些州则生产甚少（如阿拉斯加州）。当我们的统计分析包含有异质的单位时，我们必须考虑**尺度**（size）或**规模效应**（scale effect）以避免造成混乱。为了清楚地看出这一点，我们将美国 1990 年 50 个州的鸡蛋产量及其价格数据描绘在图 1—6 上。这些数字表明观测值散布得多么宽。在第 11 章中，我们将看到，在评价经济变量之间的关系时，规模效应怎样成为一个重要的因素。

混合数据。在混合或组合数据中兼有时间序列和横截面数据的成分。表 1—1 中的数据即混合数据之一例。对每一个年份我们有 50 个横截面观测，而对每一个州我们有蛋价和蛋产量的两个时期的观测序列，总共 100 个混合（或组合）观测。类似地，习题 1.1 给出的数据也是混合数据。因为 1980—2005 年间每个国家的 CPI 就构成一个时间序列。而对某一年来说，7 个国家的 CPI 又构成一个横截面。在此混合

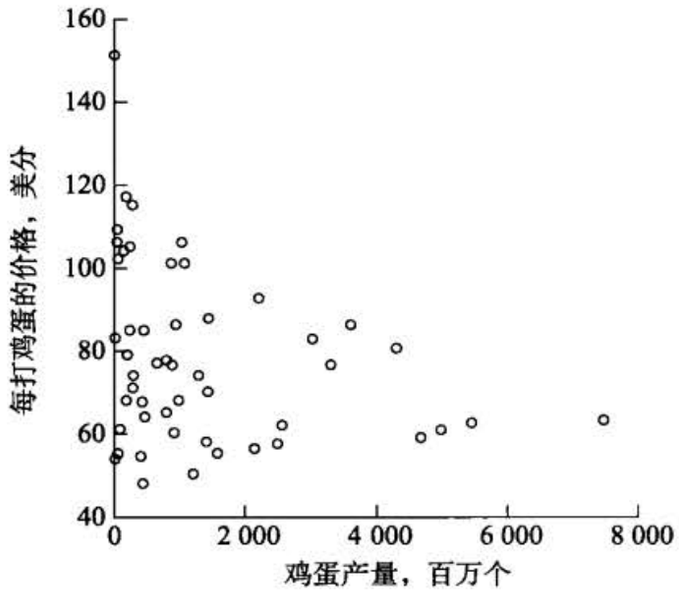


图 1—6　1990 年蛋产量与价格的关系

数据中，我们有 182 个观测——对 7 个国家中的每一个，都有 26 个年观测值。

面板、纵列或微观面板数据。这是混合数据的一种特殊类型。指对相同的横截面单位（比如家庭或厂家）在时间轴上进行跟踪调查的数据。例如，美国商务部定期举行的一种住房普查，每一次普查都对同样的住户（或对住在同样地址的人们）进行采访，以便发现自从上次普查以来该户的住房和财务状况是否有所变化。通过对相同住户的定期采访，面板数据对住户行为的动态特性提供了非常有用的信息，我们将在第 16 章看到这一点。

作为一个具体的例子，考虑表 1—2 中给出的数据。原本由格伦费尔德（Y. Grunfeld）搜集的表中数据指的是通用电气（GE）、美国钢铁（US）、通用汽车（GM）和西屋电气（WEST）美国四大公司的真实投资、企业真实价值和真实资本存量，数据搜集期为 1935—1954 年。[①] 既然对几家公司搜集了数年的数据，所以这是面板数据的一个典型例子。在这个表中，每个公司的观测次数都相同，但也并非总是这样。如果所有公司都具有相同的观测次数，我们便得到所谓的**平衡面板**（balanced panel）。如果每个公司的观测次数不尽相同，我们便得到一个**非平衡面板**（unbalanced panel）。我们在第 16 章面板数据回归模型中将分析这种数据并说明如何估计这种模型。

表 1—2　　1935—1954 年间美国四大公司的投资数据

年份	I	F_{-1}	C_{-1}	年份	I	F_{-1}	C_{-1}
	GE				US		
1935	33.1	1 170.6	97.8	1935	209.9	1 362.4	53.8
1936	45.0	2 015.8	104.4	1936	355.3	1 807.1	50.5

① Y. Grunfeld, "The Determinants of Corporate Investment," unpublished PhD thesis, Department of Economics, University of Chicago, 1958. 这些数据已成为说明面板数据回归模型的驱动力量。

续前表

年份	I	F_{-1}	C_{-1}	年份	I	F_{-1}	C_{-1}
	GE				US		
1937	77.2	2 803.3	118.0	1937	469.9	2 673.3	118.1
1938	44.6	2 039.7	156.2	1938	262.3	1 801.9	260.2
1939	48.1	2 256.2	172.6	1939	230.4	1 957.3	312.7
1940	74.4	2 132.2	186.6	1940	361.6	2 202.9	254.2
1941	113.0	1 834.1	220.9	1941	472.8	2 380.5	261.4
1942	91.9	1 588.0	287.8	1942	445.6	2 168.6	298.7
1943	61.3	1 749.4	319.9	1943	361.6	1 985.1	301.8
1944	56.8	1 687.2	321.3	1944	288.2	1 813.9	279.1
1945	93.6	2 007.7	319.6	1945	258.7	1 850.2	213.8
1946	159.9	2 208.3	346.0	1946	420.3	2 067.7	232.6
1947	147.2	1 656.7	456.4	1947	420.5	1 796.7	264.8
1948	146.3	1 604.4	543.4	1948	494.5	1 625.8	306.9
1949	98.3	1 431.8	618.3	1949	405.1	1 667.0	351.1
1950	93.5	1 610.5	647.4	1950	418.8	1 677.4	357.8
1951	135.2	1 819.4	671.3	1951	588.2	2 289.5	341.1
1952	157.3	2 079.7	726.1	1952	645.2	2 159.4	444.2
1953	179.5	2 371.6	800.3	1953	641.0	2 031.3	623.6
1954	189.6	2 759.9	888.9	1954	459.3	2 115.5	669.7
	GM				WEST		
1935	317.6	3 078.5	2.8	1935	12.93	191.5	1.8
1936	391.8	4 661.7	52.6	1936	25.90	516.0	0.8
1937	410.6	5 387.1	156.9	1937	35.05	729.0	7.4
1938	257.7	2 792.2	209.2	1938	22.89	560.4	18.1
1939	330.8	4 313.2	203.4	1939	18.84	519.9	23.5
1940	461.2	4 643.9	207.2	1940	28.57	628.5	26.5
1941	512.0	4 551.2	255.2	1941	48.51	537.1	36.2
1942	448.0	3 244.1	303.7	1942	43.34	561.2	60.8
1943	499.6	4 053.7	264.1	1943	37.02	617.2	84.4
1944	547.5	4 379.3	201.6	1944	37.81	626.7	91.2
1945	561.2	4 840.9	265.0	1945	39.27	737.2	92.4
1946	688.1	4 900.0	402.2	1946	53.46	760.5	86.0
1947	568.9	3 526.5	761.5	1947	55.56	581.4	111.1
1948	529.2	3 245.7	922.4	1948	49.56	662.3	130.6
1949	555.1	3 700.2	1 020.1	1949	32.04	583.8	141.8
1950	642.9	3 755.6	1 099.0	1950	32.24	635.2	136.7
1951	755.9	4 833.0	1 207.7	1951	54.38	732.8	129.7
1952	891.2	4 924.9	1 430.5	1952	71.78	864.1	145.5
1953	1 304.4	6 241.7	1 777.3	1953	90.08	1 193.5	174.8
1954	1 486.7	5 593.6	2 226.3	1954	68.60	1 188.9	213.5

注：$Y=I$=总投资=厂房与设备的增加以及维修和修理支出，经 P_1 折算并以百万美元为单位。

$X_2=F$=企业价值= 12 月 31 日普通股和优先股的价格（或本年 12 月 31 日与次年 1 月 31 日股价的平均值）乘以流通中普通股和优先股数量，再加上 12 月 31 日总债务的账面价值，经 P_2 折算并以百万美元为单位。

$X_3=C$=厂房与设备存量=净增加厂房与设备经 P_1 折算后的累计总和减去经 P_3 折算后的折旧金。

P_1=生产者耐用设备的暗含价格折算指数（1947=100）。

P_2=暗含的 GNP 折算指数（1947=100）。

P_3=折旧费的折算指数=金属及金属产品零售价格指数的 10 年移动平均（1947=100）。

资料来源：Reproduced from H. D. Vinod and Aman Ullah, *Recent Advances in Regression Methods*, Marcel Dekker, New York, 1981, pp. 259-261.

格伦费尔德搜集这些数据的目的是为了弄清楚真实总投资（I）与一年前的企业真实价值（F）和一年前的真实资本存量（C）之间的关系。由于样本中所包含的公司都在同一个资本市场上运作，所以通过把它们放在一起来研究，格伦费尔德希望弄明白它们是否具有相似的投资函数。

□ 数据来源[①]

用于经验分析的数据可以由一个政府机构（如商务部）、一个国际机构［如国际货币基金组织（IMF）、世界银行（World Bank）］、一个私人组织（如标准普尔公司）或某一个人来搜集。从表面上看，有成千上万的这种机构在搜集着具有各种用途的数据。

互联网。互联网简直使数据搜集发生了革命性的变化。如果只是用键盘“在网上冲浪”（比如汇率），那你将淹没在各种各样的数据来源之中。我们在附录 E 中提供了一些被频繁访问的网址，它们能提供各类经济和金融数据。许多数据都无须太多费用就能下载。你可能想把那些能为你提供有用的经济数据的各种网址制成书签。

这些机构所收集的数据可以是**实验**（experimental）或**非实验**（nonexperimental）性质的。在自然科学中，经常收集的是实验数据。这时，研究者希望在保持一些因素不变的情况下收集数据，以便评价另一些因素对某一现象的影响。例如，在评价肥胖对血压的影响时，研究者要在人们饮食、烟酒习惯都不变的情况下收集数据，以便尽可能减少这些变量对血压的影响。

在社会科学中，人们通常获得的数据是非实验性质的，就是说，这些数据不受研究者的控制。[②] 例如，GNP、失业、股票价格等数据并不受研究者的直接控制。如我们将要看到的那样，对数据缺乏控制常常给研究者在寻觅某种事态的准确原因时造成特别的困难。例如，究竟是货币供给决定（名义）GNP 呢，还是反过来 GNP 决定货币供给呢？

□ 数据的准确性[③]

虽然有大量的数据可供经济研究之用，但是数据的质量常常不那么好。对此有几点理由。

1. 如上所说，大部分社会科学数据是非实验性质的，有观测误差的可能；也可能出于疏漏（omission），或出于委托（commission）。

2. 即使是实验得来的数据，测量误差可由近似计算或进位而产生。

3. 在问卷调查中，无应答（nonresponse）的问题也可能相当严重；有问卷的

① 为看到一种明确的叙述，可参考 Albert T. Somers, *The U. S. Economy Demystified: What the Major Economic Statistics Mean and their Significance for Business*, D. C. Heath, Lexington, Mass., 1985.

② 人们在社会科学中有时也能进行控制试验，习题 1.6 就给出了一个这样的例子。

③ O. Morgenstern, *The Accuracy of Economic Observations*, 2d ed., Princeton University Press, Princeton, N. J., 1963，该书提有尖锐的意见。

40%应答者就算幸运。根据这样的部分答卷作的分析未必真正反映60%无应答者的行为。由此导致所谓的（样本）选择偏误（selectivity bias）。不但如此，回答问卷的人不一定回答所有的问题，特别是那些财务上敏感的问题，从而导致更多的选择偏误。

4. 获取数据的抽样方法可能变化很大，要比较不同样本得来的结果常常非常困难。

5. 通常获得的经济数据都是高度加总的。例如，大多数宏观数据（如GNP、就业、通货膨胀、失业）都是对整个国家或者至少是对一些很大的地（理）区（域）给出的。这种高度加总数据未必能告诉我们多少有关个人或微观单位的情况，而后者才是研究的最终目标。

6. 由于保密性质，某些数据只能以高度加总的形式公布。例如，法律不允许IRS公开个人税收回执数据；它只能透露一些高度概括性的数据。因此，尽管你想知道某一收入水平的个人在卫生保健方面花了多少钱，你是无法进行这种分析的，除非是在高度加总的水平上。然而，这样的宏观分析往往揭示不了微观单位的行为动态。类似地，商务部每5年进行一次企业普查，但法律却不允许它公布关于任何厂家的生产、人员雇佣、能源消耗、研究与开发费用等方面的信息。因此要在企业层次上研究在这些项目上的厂际差异是很困难的。

因为有这些和许多其他问题，研究者应时刻记住：**研究结果不可能比数据的质量更好。**所以，如果在一定情况下，研究者发现研究的结果“不能令人满意”的话，原因不一定是误用模型，而是数据的质量不好。不幸的是，由于大多数社会科学研究所用的数据都是非实验性质的，所以研究者常常别无选择，唯有依赖其所能获得的数据。但他还应时刻记住：所用的数据未必是最好的。因此不要过于教条地对待研究结果，尤其当数据的质量受到怀疑时。

□ 对变量测量尺度的注解[①]

我们通常遇到的变量分为如下四大类：比率尺度（ratio scale）、区间尺度（interval scale）、序数尺度（ordinal scale）和名义尺度（nominal scale）。理解其中的每一类对我们都很重要。

比率尺度。对于一个变量X，取其两个值X_1和X_2，比率X_1/X_2和距离（X_2-X_1）都是有意义的量。此外，这些值在这种尺度下存在着一种自然顺序（上升或下降）。因此，诸如$X_2 \leqslant X_1$或$X_2 \geqslant X_1$之类的比较也是有意义的。大多数经济变量都属于这一类。于是，问今年的GDP与去年的GDP相比有多大是有意义的。以美元为单位度量的个人收入是一个比率变量；挣10万美元的人，其收入就是挣5万美元的人的2倍（当然是税前）。

① 以下讨论很大程度上依据Aris Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, New York, 1999, p. 24。

区间尺度。一个区间尺度变量满足比率尺度变量的后面两个性质，但不满足第一个性质。因此，两个时期之内的距离（如 2000－1995）是有意义的，但两个时期的比率（2000/1995）就没有什么意义。2007 年 8 月 11 日上午 11 点天气预报说俄勒冈州波特兰市的温度是华氏 60 度，而佛罗里达州塔拉哈西市达到华氏 90 度。说塔拉哈西比波特兰暖和 50%没有意义，所以，温度不是比例尺度。这主要是因为华氏温标不是以 0 度作为起点所致。

序列尺度。只要一个变量满足比率尺度的第三个性质（即自然顺序），那它就属于这一类变量。例子有考试分数体系（A、B、C）或收入阶层（高、中、低）等。对于这些变量，自然顺序存在，但不同类别之间的差别不能量化。经济学的学生可能会想起两种商品之间的无差异曲线（indifference curves），虽然每条更高的无差异曲线标志着更高的效用水平，但不能量化一条无差异曲线比另一条无差异曲线到底高多少。

名义尺度。此类变量不具备比率尺度变量的任何一个特征。诸如性别（男、女）和婚姻状况（已婚、未婚、离婚、分居）之类的变量只表示了不同的类别。问题：这种变量不能用比率尺度、区间尺度或序列尺度表示的原因是什么？

以后将会看到，适合于比率尺度变量的计量经济方法可能不适合于名义尺度变量。因此，记住上面讨论的四类测量尺度之间的区别就很重要。

要点与结论

1. 回归分析的主要用意，是分析一个所谓因变量对另一个或多个所谓解释变量的统计依赖关系。

2. 这种分析的目的，是要在解释变量的已知或固定值的基础上，估计和（或）预测因变量的均值。

3. 实际上，回归分析的成功有赖于适用数据的获得。本章讨论了研究者通常能获得的（特别是在社会科学中）数据的性质、来源及其局限性。

4. 在任何一项研究中，研究者都应清楚地说明分析中所用数据的来源、定义及其搜集方法，数据中的任何差错或疏漏，以及对数据进行的任何改动。须知，政府公布的宏观经济数据是常有修改的。

5. 因为读者未必有这种时间、精力和资源去跟踪数据，所以他有权利假定研究者所用的数据是适当地采集的，并且计算和分析也都是正确的。

习　　题

1.1　表 1—3 给出了 7 个工业化国家的消费者价格指数（CPI）数据，以 1982—1984 年为该

指数的基期并令 1982—1984=100。

表 1—3　　1980—2005 年间 7 个工业化国家的 CPI（1982—1984=100）

年份	美国	加拿大	日本	法国	德国	意大利	英国
1980	82.4	76.1	91.0	72.2	86.7	63.9	78.5
1981	90.9	85.6	95.3	81.8	92.2	75.5	87.9
1982	96.5	94.9	98.1	91.7	97.0	87.8	95.4
1983	99.6	100.4	99.8	100.3	100.3	100.8	99.8
1984	103.9	104.7	102.1	108.0	102.7	111.4	104.8
1985	107.6	109.0	104.2	114.3	104.8	121.7	111.1
1986	109.6	113.5	104.9	117.2	104.6	128.9	114.9
1987	113.6	118.4	104.9	121.1	104.9	135.1	119.7
1988	118.3	123.2	105.6	124.3	106.3	141.9	125.6
1989	124.0	129.3	108.0	128.7	109.2	150.7	135.4
1990	130.7	135.5	111.4	132.9	112.2	160.4	148.2
1991	136.2	143.1	115.0	137.2	116.3	170.5	156.9
1992	140.3	145.3	117.0	140.4	122.2	179.5	162.7
1993	144.5	147.9	118.5	143.4	127.6	187.7	165.3
1994	148.2	148.2	119.3	145.8	131.1	195.3	169.3
1995	152.4	151.4	119.2	148.4	133.3	205.6	175.2
1996	156.9	153.8	119.3	151.4	135.3	213.8	179.4
1997	160.5	156.3	121.5	153.2	137.8	218.2	185.1
1998	163.0	157.8	122.2	154.2	139.1	222.5	191.4
1999	166.6	160.5	121.8	155.0	140.0	226.2	194.3
2000	172.2	164.9	121.0	157.6	142.0	231.9	200.1
2001	177.1	169.1	120.1	160.2	144.8	238.3	203.6
2002	179.9	172.9	119.0	163.3	146.7	244.3	207.0
2003	184.0	177.7	118.7	166.7	148.3	250.8	213.0
2004	188.9	181.0	118.7	170.3	150.8	256.3	219.4
2005	195.3	184.9	118.3	173.2	153.7	261.3	225.6

资料来源：*Economic Report of the President*，2007，Table 108，p. 354.

a. 利用所给数据计算每个国家的通货膨胀率。①

b. 绘制每个国家的通货膨胀率相对时间的描点图（即以时间为横轴，并以通货膨胀率为纵轴）。

c. 你从这 7 个国家的通货膨胀经历中能得出什么宽泛的结论？

d. 哪个国家的通货膨胀率波动最大？你能给出什么样的解释呢？

1.2　a. 利用表 1—3，绘制加拿大、法国、德国、意大利、日本和英国的通货膨胀率相对美

① 将当年的 CPI 减去上一年度的 CPI 后，再除以上一年度的 CPI，然后乘以 100 即可得到通货膨胀率。因此，加拿大 1981 年的通货膨胀率就是[(85.6—76.1)/76.1]×100=12.48%（近似）。

国通货膨胀率的散点图。

b. 一般性地评论这 6 个国家的通货膨胀率相对美国通货膨胀率的表现。

c. 如果你发现这 6 个国家的通货膨胀率与美国的通货膨胀率同向变化，那是否表明美国的通货膨胀导致了其他国家的通货膨胀？为什么？

1.3 表 1—4 给出了 9 个工业化国家 1985—2006 年间的外汇汇率数据。除英国外，汇率都定义为一美元兑换外币的数量；而英国的汇率定义为一英镑兑换美元的数量。

a. 画出这些汇率相对时间的散点图，并评论汇率在给定期间内的一般表现。

b. 如果一美元能买到更多单位的外币，则称之为美元升值（appreciate）。相反，如果一美元购买更少的外币，则称之为美元贬值（depreciate）。在 1985—2006 年间，美元的一般表现如何？顺便查阅一本宏观经济学或国际经济学教科书，以探明是哪些因素决定了货币的升值或贬值。

表 1—4　　九国汇率：1985—2006

年份	澳大利亚	加拿大	中国	日本	墨西哥	韩国	瑞典	瑞士	英国
1985	0.700 3	1.365 9	2.943 4	238.47	0.257	872.45	8.603 2	2.455 2	1.297 4
1986	0.670 9	1.389 6	3.461 6	168.35	0.612	884.60	7.127 3	1.797 9	1.467 7
1987	0.701 4	1.325 9	3.731 4	144.60	1.378	826.16	6.346 9	1.491 8	1.639 8
1988	0.784 1	1.230 6	3.731 4	128.17	2.273	734.52	6.137 0	1.464 3	1.781 3
1989	0.791 9	1.184 2	3.767 3	138.07	2.461	674.13	6.455 9	1.636 9	1.638 2
1990	0.780 7	1.166 8	4.792 1	145.00	2.813	710.64	5.923 1	1.390 1	1.784 1
1991	0.778 7	1.146 0	5.333 7	134.59	3.018	736.73	6.052 1	1.435 6	1.767 4
1992	0.735 2	1.208 5	5.520 6	126.78	3.095	784.66	5.825 8	1.406 4	1.766 3
1993	0.679 9	1.290 2	5.779 5	111.08	3.116	805.75	7.795 6	1.478 1	1.501 6
1994	0.731 6	1.366 4	8.639 7	102.18	3.385	806.93	7.716 1	1.366 7	1.531 9
1995	0.740 7	1.372 5	8.370 0	93.96	6.447	772.69	7.140 6	1.181 2	1.578 5
1996	0.782 8	1.363 8	8.338 9	108.78	7.600	805.00	6.708 2	1.236 1	1.560 7
1997	0.743 7	1.384 9	8.319 3	121.06	7.918	953.19	7.644 6	1.451 4	1.637 6
1998	0.629 1	1.483 6	8.300 8	130.99	9.152	1 400.40	7.952 2	1.450 6	1.657 3
1999	0.645 4	1.485 8	8.278 3	113.73	9.553	1 189.84	8.274 0	1.504 5	1.617 2
2000	0.581 5	1.485 5	8.278 4	107.80	9.459	1 130.90	9.173 5	1.690 4	1.515 6
2001	0.516 9	1.548 7	8.277 0	121.57	9.337	1 292.02	10.342 5	1.689 1	1.439 6
2002	0.543 7	1.570 4	8.277 1	125.22	9.663	1 250.31	9.723 3	1.556 7	1.502 5
2003	0.652 4	1.400 8	8.277 2	115.94	10.793	1 192.08	8.078 7	1.345 0	1.634 7
2004	0.736 5	1.301 7	8.276 8	108.15	11.290	1 145.24	7.348 0	1.242 8	1.833 0
2005	0.762 7	1.211 5	8.193 6	110.11	10.894	1 023.75	7.471 0	1.245 9	1.820 4
2006	0.753 5	1.134 0	7.972 3	116.31	10.906	954.32	7.371 8	1.253 2	1.843 4

资料来源：*Economic Report of the President*，2007，Table B-110，p. 356.

1.4 图 1—5 背后的 M1 货币供给数据由表 1—5 给出。你能给出货币供给在表中所示时期上升的原因吗？

表 1—5　　经季节调整的 M1 供给：1959 年 1 月—1999 年 7 月　　（单位：十亿美元）

1959 年 1 月	138.890 0	139.390 0	139.740 0	139.690 0	140.680 0	141.170 0
1959 年 7 月	141.700 0	141.900 0	141.010 0	140.470 0	140.380 0	139.950 0
1960 年 1 月	139.980 0	139.870 0	139.750 0	139.560 0	139.610 0	139.580 0
1960 年 7 月	140.180 0	141.310 0	141.180 0	140.920 0	140.860 0	140.690 0
1961 年 1 月	141.060 0	141.600 0	141.870 0	142.130 0	142.660 0	142.880 0
1961 年 7 月	142.920 0	143.490 0	143.780 0	144.140 0	144.760 0	145.200 0
1962 年 1 月	145.240 0	145.660 0	145.960 0	146.400 0	146.840 0	146.580 0
1962 年 7 月	146.460 0	146.570 0	146.300 0	146.710 0	147.290 0	147.820 0
1963 年 1 月	148.260 0	148.900 0	149.170 0	149.700 0	150.390 0	150.430 0
1963 年 7 月	151.340 0	151.780 0	151.980 0	152.550 0	153.650 0	153.290 0
1964 年 1 月	153.740 0	154.310 0	154.480 0	154.770 0	155.330 0	155.620 0
1964 年 7 月	156.800 0	157.820 0	158.750 0	159.240 0	159.960 0	160.300 0
1965 年 1 月	160.710 0	160.940 0	161.470 0	162.030 0	161.700 0	162.190 0
1965 年 7 月	163.050 0	163.680 0	164.850 0	165.970 0	166.710 0	167.850 0
1966 年 1 月	169.080 0	169.620 0	170.510 0	171.810 0	171.330 0	171.570 0
1966 年 7 月	170.310 0	170.810 0	171.970 0	171.160 0	171.380 0	172.030 0
1967 年 1 月	171.860 0	172.990 0	174.810 0	174.170 0	175.680 0	177.020 0
1967 年 7 月	178.130 0	179.710 0	180.680 0	181.640 0	182.380 0	183.260 0
1968 年 1 月	184.330 0	184.710 0	185.470 0	186.600 0	187.990 0	189.420 0
1968 年 7 月	190.490 0	191.840 0	192.740 0	194.020 0	196.020 0	197.410 0
1969 年 1 月	198.690 0	199.350 0	200.020 0	200.710 0	200.810 0	201.270 0
1969 年 7 月	201.660 0	201.730 0	202.100 0	202.900 0	203.570 0	203.880 0
1970 年 1 月	206.220 0	205.000 0	205.750 0	206.720 0	207.220 0	207.540 0
1970 年 7 月	207.980 0	209.930 0	211.800 0	212.880 0	213.660 0	214.410 0
1971 年 1 月	215.540 0	217.420 0	218.770 0	220.000 0	222.020 0	223.450 0
1971 年 7 月	224.850 0	225.580 0	226.470 0	227.160 0	227.760 0	228.320 0
1972 年 1 月	230.090 0	232.320 0	234.300 0	235.580 0	235.890 0	236.620 0
1972 年 7 月	238.790 0	240.930 0	243.180 0	245.020 0	246.410 0	249.250 0
1973 年 1 月	251.470 0	252.150 0	251.670 0	252.740 0	254.890 0	256.690 0
1973 年 7 月	257.540 0	257.760 0	257.860 0	259.040 0	260.980 0	262.880 0
1974 年 1 月	263.760 0	265.310 0	266.680 0	267.200 0	267.560 0	268.440 0
1974 年 7 月	269.270 0	270.120 0	271.050 0	272.350 0	273.710 0	274.200 0
1975 年 1 月	273.900 0	275.000 0	276.420 0	276.170 0	279.200 0	282.430 0
1975 年 7 月	283.680 0	284.150 0	285.690 0	285.390 0	286.830 0	287.070 0
1976 年 1 月	288.420 0	290.760 0	292.700 0	294.660 0	295.930 0	296.160 0
1976 年 7 月	297.200 0	299.050 0	299.670 0	302.040 0	303.590 0	306.250 0
1977 年 1 月	308.260 0	311.540 0	313.940 0	316.020 0	317.190 0	318.710 0
1977 年 7 月	320.190 0	322.270 0	324.480 0	326.400 0	328.640 0	330.870 0
1978 年 1 月	334.400 0	335.300 0	336.960 0	339.920 0	344.860 0	346.800 0
1978 年 7 月	347.630 0	349.660 0	352.260 0	353.350 0	355.410 0	357.280 0

续前表

1979 年 1 月	358.600 0	359.910 0	362.450 0	368.050 0	369.590 0	373.340 0
1979 年 7 月	377.210 0	378.820 0	379.280 0	380.870 0	380.810 0	381.770 0
1980 年 1 月	385.850 0	389.700 0	388.130 0	383.440 0	384.600 0	389.460 0
1980 年 7 月	394.910 0	400.060 0	405.360 0	409.060 0	410.370 0	408.060 0
1981 年 1 月	410.830 0	414.380 0	418.690 0	427.060 0	424.430 0	425.500 0
1981 年 7 月	427.900 0	427.850 0	427.460 0	428.450 0	430.880 0	436.170 0
1982 年 1 月	442.130 0	441.490 0	442.370 0	446.780 0	446.530 0	447.890 0
1982 年 7 月	449.090 0	452.490 0	457.500 0	464.570 0	471.120 0	474.300 0
1983 年 1 月	476.680 0	483.850 0	490.180 0	492.770 0	499.780 0	504.350 0
1983 年 7 月	508.960 0	511.600 0	513.410 0	517.210 0	518.530 0	520.790 0
1984 年 1 月	524.400 0	526.990 0	530.780 0	534.030 0	536.590 0	540.540 0
1984 年 7 月	542.130 0	542.390 0	543.860 0	543.870 0	547.320 0	551.190 0
1985 年 1 月	555.660 0	562.480 0	565.740 0	569.550 0	575.070 0	583.170 0
1985 年 7 月	590.820 0	598.060 0	604.470 0	607.910 0	611.830 0	619.360 0
1986 年 1 月	620.400 0	624.140 0	632.810 0	640.350 0	652.010 0	661.520 0
1986 年 7 月	672.200 0	680.770 0	688.510 0	695.260 0	705.240 0	724.280 0
1987 年 1 月	729.340 0	729.840 0	733.010 0	743.390 0	746.000 0	743.720 0
1987 年 7 月	744.960 0	746.960 0	748.660 0	756.500 0	752.830 0	749.680 0
1988 年 1 月	755.550 0	757.700 0	761.180 0	767.570 0	771.680 0	779.100 0
1988 年 7 月	783.400 0	785.080 0	784.820 0	783.630 0	784.460 0	786.260 0
1989 年 1 月	784.920 0	783.400 0	782.740 0	778.820 0	774.790 0	774.220 0
1989 年 7 月	779.710 0	781.140 0	782.200 0	787.050 0	787.950 0	792.570 0
1990 年 1 月	794.930 0	797.650 0	801.250 0	806.240 0	804.360 0	810.330 0
1990 年 7 月	811.800 0	817.850 0	821.830 0	820.300 0	822.060 0	824.560 0
1991 年 1 月	826.730 0	832.400 0	838.620 0	842.730 0	848.960 0	858.330 0
1991 年 7 月	862.950 0	868.650 0	871.560 0	878.400 0	887.950 0	896.700 0
1992 年 1 月	910.490 0	925.130 0	936.000 0	943.890 0	950.780 0	954.710 0
1992 年 7 月	964.600 0	975.710 0	988.840 0	1 004.340	1 016.040	1 024.450
1993 年 1 月	1 030.900	1 033.150	1 037.990	1 047.470	1 066.220	1 075.610
1993 年 7 月	1 085.880	1 095.560	1 105.430	1 113.800	1 123.900	1 129.310
1994 年 1 月	1 132.200	1 136.130	1 139.910	1 141.420	1 142.850	1 145.650
1994 年 7 月	1 151.490	1 151.390	1 152.440	1 150.410	1 150.440	1 149.750
1995 年 1 月	1 150.640	1 146.740	1 146.520	1 149.480	1 144.650	1 144.240
1995 年 7 月	1 146.500	1 146.100	1 142.270	1 136.430	1 133.550	1 126.730
1996 年 1 月	1 122.580	1 117.530	1 122.590	1 124.520	1 116.300	1 115.470
1996 年 7 月	1 112.340	1 102.180	1 095.610	1 082.560	1 080.490	1 081.340
1997 年 1 月	1 080.520	1 076.200	1 072.420	1 067.450	1 063.370	1 065.990
1997 年 7 月	1 067.570	1 072.080	1 064.820	1 062.060	1 067.530	1 074.870
1998 年 1 月	1 073.810	1 076.020	1 080.650	1 082.090	1 078.170	1 077.780
1998 年 7 月	1 075.370	1 072.210	1 074.650	1 080.040	1 088.960	1 093.350
1999 年 1 月	1 091.000	1 092.650	1 102.010	1 108.400	1 104.750	1 101.110
1999 年 7 月	1 099.530	1 102.400	1 093.460			

资料来源：Board of Governors，Federal Reserve Bank，USA.

1.5 假设你要做一个犯罪行为的经济学模型，比方说花在犯罪活动（如非法贩卖毒品）的小时数。在做这样的一个模型时，你要考虑哪些变量？看一下你的模型能否与诺贝尔奖得主加里·贝克尔（Gary Becker）的模型相媲美。①

1.6 **经济学中的控制试验：**2000年4月7日，克林顿（Clinton）总统签署了一项参众两院通过的法案，取消对社会保障金领取者的收入限制。此前，年龄介于65岁和69岁之间的受济者，年收入超过1.7万美元者，超出部分的每3美元减少1美元的社会保障救济金。你如何设计一个研究方案来分析这种法律修订的影响？注：原有法律对70岁以上的受济者没有设定收入限制。

1.7 表1—6中的数据发表在1984年3月1日的《华尔街日报》（*The Wall Street Journal*）上。它将1983年21家企业的广告预算（以百万美元计）与看报者每周对这些企业产品保留的印象次数（以百万次计）相联系。这些数据基于对4 000个成人的调查，在调查中要求产品使用者列出一条在过去的一周里见过的该类产品的商业广告。

a. 以印象数为纵轴、以广告支出为横轴画散点图。

b. 你认为这两个变量之间的关系具有什么样的性质？

c. 看一下你的图，你认为值得做广告吗？想想那些出现在星期天的超级碗杯赛（Super Bowl Sunday）上和世界职业棒球锦标赛期间的商业广告。

注：我们在以后的章节中将进一步探讨表1—6中给出的数据。

表1—6　　广告支出的影响

企业	印象（百万次）	支出（以1983年的百万美元计）
1. 米乐	32.1	50.1
2. 百事	99.6	74.1
3. 金鹰	11.7	19.3
4. 联邦快递	21.9	22.9
5. 汉堡王	60.8	82.4
6. 可口可乐	78.6	40.1
7. 麦当劳	92.4	185.9
8. 前世通公司	50.7	26.9
9. 健怡可乐	21.4	20.4
10. 福特	40.1	166.2
11. 李维斯	40.8	27.0
12. 百威	10.4	45.6
13. 美国电话电报公司/贝尔	88.9	154.9
14. 卡尔文·克莱恩（CK）	12.0	5.0
15. 温迪快餐	29.2	49.7
16. 宝丽来	38.0	26.9
17. 舍斯塔房车公司	10.0	5.7
18. 咪咪乐猫粮公司	12.3	7.6
19. 卡夫食品	23.4	9.2
20. 佳洁士	71.1	32.4
21. 基波斯狗粮公司	4.4	6.1

资料来源：http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html.

① G. S. Becker, "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, vol. 76, 1968, pp. 169-217.

第 2 章 双变量回归分析：一些基本思想

我们在第 1 章中概括地讨论了回归的概念。本章中我们将比较正式地继续探讨这一主题。具体地说，本章和随后的三章，我们向读者介绍最为简单的两个或双变量回归分析所依据的理论，其中因变量（回归子）仅与唯一的解释变量（回归元）相关。我们首先考虑双变量情形，不是因为它在实践中足够用了，而是因为它能使回归分析的基本概念表述得尽可能简单，而且，某些概念还能借助于二维图形进行说明。不仅如此，我们还将看到，更为一般的多元回归分析在许多方面都是双变量情形的逻辑推广。

2.1 一个假设的例子[①]

如 1.2 节所指出的，回归分析大体上说，是要根据解释变量的已知或给定值，去估计和（或）预测因变量的（总体）均值。[②] 为了理解如何能做到这一点，考虑表 2—1 中的数据。表中数据指的是，在一个假想的经济社会中，构成**总体**（population）的 60 个家庭及其周收入（X）和周消费支出（Y）的美元数量。这 60 个家庭被分成 10 个收入组（从 80 美元到 260 美元），各组中每个家庭的周消费支出都列在

① 在阅读本章之前，统计学知识有些生疏的读者可能愿意先读统计学附录即附录 A，以达到温故而知新的效果。

② 一个随机变量 Y 的期望值或期望或总体平均可记为 $E(Y)$。另一方面，从 Y 的总体的一个样本值中计算出来的均值记为 $\bar{Y}$，读作“Y 横”。

表中。因此，我们就有 10 个固定的 X 值及与每个 X 对应的 Y 值；可以说，有 10 个 Y 的子总体。

表 2—1 **周家庭收入 X** （单位：美元）

Y↓ \ X→	80	100	120	140	160	180	200	220	240	260
周家庭消费支出 Y（美元）	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	—	88	—	113	125	140	—	160	189	185
	—	—	—	115	—	—	—	162	—	191
共计	325	462	445	707	678	750	685	1 043	966	1 211
Y 的条件均值 $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

从图 2—1 可以清楚地看出，每个收入组的周消费支出都有可观的变化。但人们一般得到的图是，尽管每个收入组中的周消费支出可以变化，但平均来讲，周消费支出随着收入的上升而增加。为了清楚地看出这一点，我们在表 2—1 中已经给出了与 10 个收入水平分别对应的平均周消费支出或周消费支出均值。于是，对应于 80 美元的周收入水平，平均消费支出是 65 美元，而对应于 200 美元的收入水平，平均消费支出则是 137 美元。对 Y 的 10 个子总体，我们共有 10 个均值。我们称这些均值为**条件期望值**（conditional expected values），或称为条件均值，因为它们取决于（条件）变量 X 的给定值。我们用符号表示为 $E(Y \mid X)$，读作“给定 X 值下 Y 的期望值”（也可参见表 2—2）。

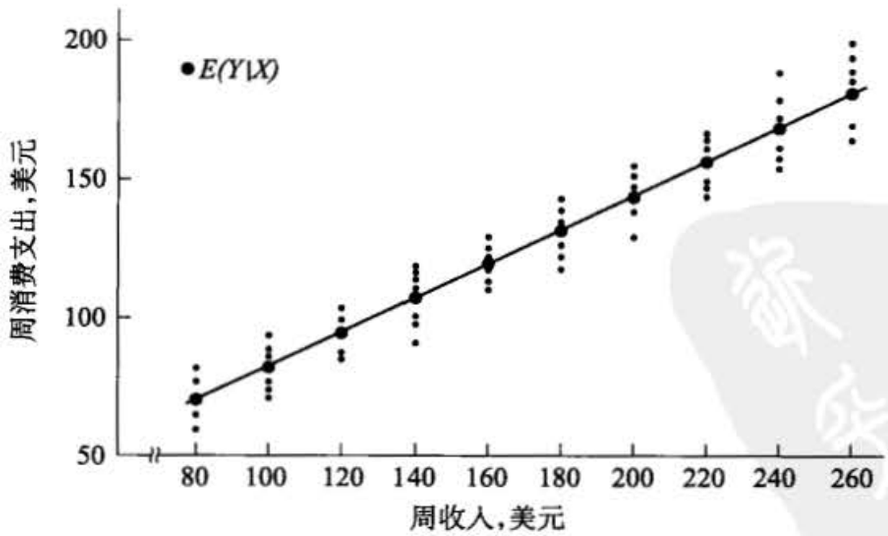


图 2—1 不同收入水平下支出的条件分布（表 2—1 的数据）

表 2—2　　与表 2—1 的数据相对应的条件概率 $p(Y \mid X_i)$

$X \rightarrow$ / $p(Y \mid X_i) \downarrow$	80	100	120	140	160	180	200	220	240	260
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
条件	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
概率	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
$p(Y \mid X_i)$	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	—	1/6	—	1/7	1/6	1/6	—	1/7	1/6	1/7
	—	—	—	1/7	—	—	—	1/7	—	1/7
Y 的条件均值	65	77	89	101	113	125	137	149	161	173

将周消费支出的这些条件期望值与**无条件期望值**（unconditional expected value）$E(Y)$ 区别开来至关重要。如果我们将总体中所有 60 个家庭的周消费支出都加起来，再将这个和除以 60，则得到的数字 121.20（=7 272/60）美元就是周消费支出的无条件均值或期望值 $E(Y)$；我们在得到这个数字时无视各个家庭的收入水平，从这个意义上讲，它是无条件的。[①] 显然，表 2—1 中给出的 Y 的各个条件期望值都不同于 Y 的无条件期望值 121.20 美元。当我们问“一个家庭周消费支出的期望值是多少”时，我们得到的回答是 121.20 美元（无条件均值）。但如果我们问“一个月收入为 140 美元的家庭的周消费支出的期望值是多少”时，我们得到的回答是 101 美元（条件均值）。换言之，如果我们问“对一个周收入 140 美元家庭的周消费的最佳（均值）预测是多少”，回答将是 101 美元。因此，对收入水平的了解使我们能比在不了解这些时更好地预测消费支出的均值。[②] 如我们将在整本书中讨论的那样，这可能正是回归分析的本质。

图 2—1 中加圈的黑点表示了不同 X 值下 Y 的条件均值。将这些条件均值连起来，就得到所谓的**总体回归线**（population regression line，PRL）或更一般地称为**总体回归曲线**（population regression curve）。[③] 更简单地说，就是**Y 对 X 的回归**（regression of Y on X）。形容词“总体”源于如下事实：我们一直在用 60 个家庭的整个总体来讨论这个例子。当然，在现实中，一个总体可能包含许多个家庭。

于是，在几何意义上，总体回归曲线就是（当）解释变量取给定值时因变量的条件均值或期望值的轨迹。更简单地说，对应于回归元 X 的每个给定值都有 Y 的一个子总体，连接这些子总体的均值就得到总体回归曲线。它可以画成图 2—2 的形状。

① 如附录 A 所示，条件均值和无条件均值通常是不同的。

② 十分感激戴维森（James Davidson）的这一见解。参见 James Davidson, *Econometric Theory*, Blackwell Publishers, Oxford, U. K., 2000, p. 11。

③ 虽然在本例中 PRL 是一条直线，但它也可以是一条曲线（见图 2—3）。

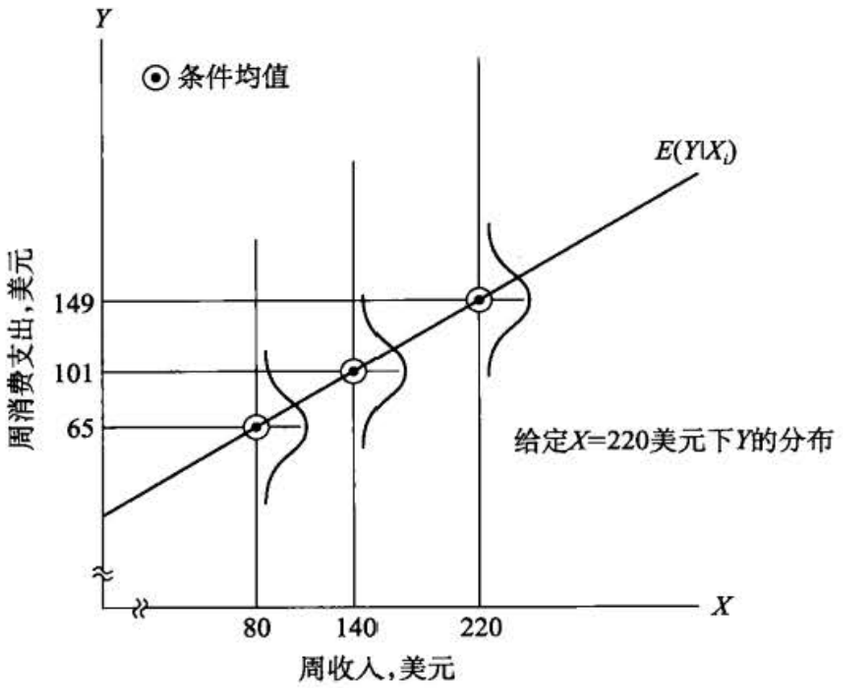


图 2—2　总体回归线（表 2—1 的数据）

此图表明，对于每个 X（即收入水平），都有 Y 值（周消费支出）的一个总体，这些 Y 值分散在其（条件）均值的左右。为简便起见，我们假定这些 Y 值对称地分布在其相应（条件）均值周围。而回归线（或曲线）穿过这些（条件）均值。

以此为背景，读者可能发现重温 1.2 节中对回归的定义有所裨益。

2.2　总体回归函数的概念

根据上述讨论和图 2—1 与图 2—2，我们清楚地看到，每一条件均值 $E(Y \mid X_i)$ 都是 X_i 的一个函数，其中 X_i 是 X 的某个给定值，用符号表示：

$$E(Y \mid X_i) = f(X_i) \tag{2.2.1}$$

其中 $f(X_i)$ 表示解释变量 X 的某个函数。在我们假设的例子中，$E(Y \mid X_i)$ 是 X_i 的一个线性函数。方程（2.2.1）被称为**条件期望函数**（conditional expectation function，CEF）或**总体回归函数**（population regression function，PRF）或简称为**总体回归**（population regression，PR）。它仅仅表明在给定 X_i 下 Y 的分布的（总体）均值与 X_i 有函数关系。简言之，它说出了 Y 的均值或平均响应是如何随 X 的变化而变化的。

函数 $f(X_i)$ 采取什么形式？这是一个重要的问题，因为在实际情况中我们不会对整个总体进行分析。因此，PRF 的函数形式是一个经验方面的问题，尽管对于一些特殊情形，理论也许能告诉我们一点什么。例如，一位经济学家可能提出消费支

出与收入存在线性关系。作为一个初次尝试或暂行假设，比如说，我们假定总体回归函数 $E(Y \mid X_i)$ 是 X_i 的线性函数，其形式是：

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \tag{2.2.2}$$

其中 β_1 和 β_2 为未知但却固定的参数，称为**回归系数**（regression coefficients）；β_1 和 β_2 也分别称为**截距**（intercept）和**斜率系数**（slope coefficients）。方程（2.2.2）本身则称为**线性总体回归函数**（linear population regression function）或简称**线性总体回归**。一些文献中曾用过其他术语，例如线性总体回归模型或线性总体回归方程等。在本书中，名词**回归**（regression）、**回归方程**（regression equation）和**回归模型**（regression model）将不加区别地当作同义词使用。

在回归分析中，我们的兴趣在于估计像方程（2.2.2）那样的 PRF。就是说，根据对 Y 和 X 的观测估计未知数 β_1 和 β_2 的值。这个问题将在第 3 章中详细研究。

2.3 “线性”一词的含义

由于本书主要讨论像方程（2.2.2）那样的线性模型，所以我们必须知道线性一词的真正含义，因为对它可作两种解释。

□ 对变量为线性

对线性的第一种并且也许是更“自然”的一种解释是，Y 的条件期望值是 X_i 的线性函数，比如说，方程（2.2.2）。[①] 从几何意义上说，这时回归曲线是一条直线。按照这种解释，诸如 $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ 的回归函数，由于变量 X 以幂或指数 2 出现，就不是线性的。

□ 对参数为线性

对线性的第二种解释是，Y 的条件期望 $E(Y \mid X_i)$ 是参数 β 的一个线性函数；它可以是或不是变量 X 的线性函数。[②] 对于这种解释，$E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ 就是一个线性（于参数）回归模型。为了看出这一点，让我们假设 X 取值为 3。因此，$E(Y \mid X = 3) = \beta_1 + 9\beta_2$，显然它是 β_1 和 β_2 的线性函数。图 2—3 中所示的所有模型因此也都是线性回

① 如果 X 仅以幂或指数 1 出现（即不包括 X^2 或 $\sqrt{X}$ 等项），并且它与其他变量也没有相乘或相除关系（比如 $X \cdot Z$ 或 X/Z，其中 Z 为另一变量），那么我们就说函数 $Y = f(X)$ 是 X 的线性函数。如果 Y 仅取决于 X，那么，Y 与 X 有线性关系的另一说法是，Y 对 X 的变化率（即 Y 对 X 的斜率）或导数（dY/dX）与 X 值无关。例如，若 $Y = 4X$，则 $dY/dX = 4$，这就与 X 值无关。但若 $Y = 4X^2$，则 $dY/dX = 8X$，这就不是与 X 值无关的了。从而它就不是 X 的线性函数。

② 如果在一个函数中，β_1 仅以一次方出现，而且不乘以或除以任何其他参数（例如，$\beta_1\beta_2$ 和 β_2/β_1 等），那么我们就说这个函数是参数 β_1 的线性函数。

归模型，即线性于参数的模型。

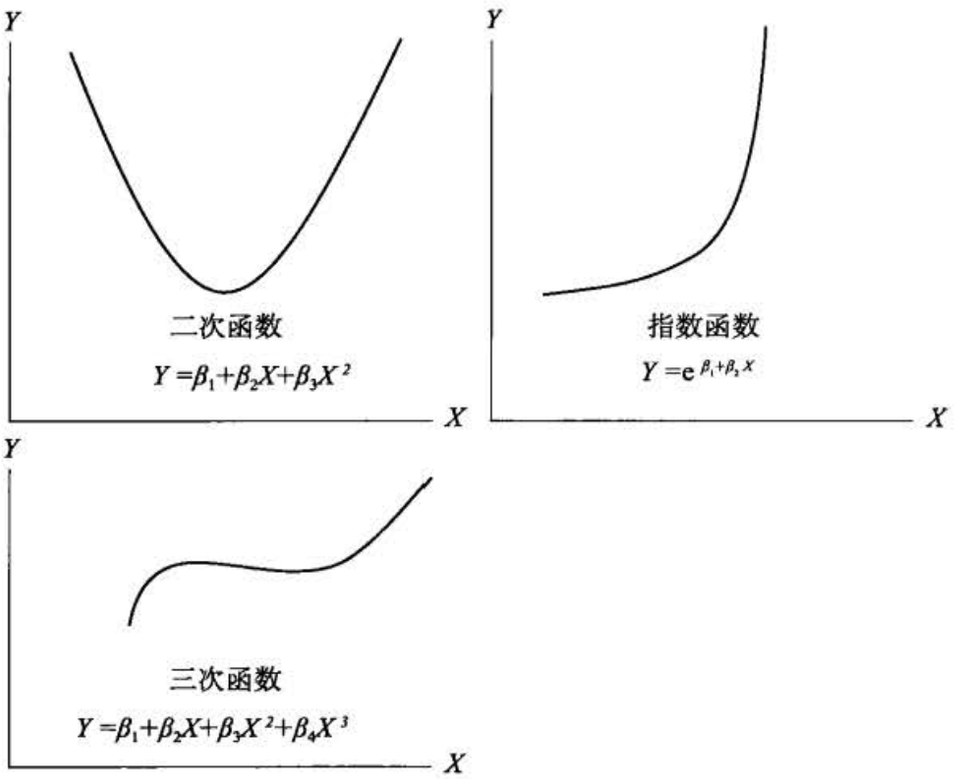


图 2—3　线性于参数的函数

现在考虑模型 $E(Y \mid X_i)=\beta_1+\beta_2^2X_i$。现在假设 $X=3$，则我们得到 $E(Y \mid X_i)=\beta_1+3\beta_2^2$，它显然不是 β_2 的线性函数。上述模型就是**非线性（于参数）回归模型**[nonlinear (in the parameter) regression model]。我们将在第 14 章讨论这种模型。

在对线性的两种解释中，对于下面即将展开讨论的回归理论来说，主要考虑的是线性于参数的情形。因此，从现在起，“线性回归”一词总是指对参数 β 为线性的一种回归（即参数只以它的一次方出现）；解释变量 X 则可以是或不是线性的。把上述讨论排成表格的形式，我们得到表 2—3。这样，$E(Y \mid X_i)=\beta_1+\beta_2X_i$ 兼对参数和变量为线性，是一个线性回归模型，而对参数为线性但对变量 X 为非线性的 $E(Y \mid X_i)=\beta_1+\beta_2X_i^2$，也是一个线性回归模型。

表 2—3　　线性回归模型

模型对参数为线性？	模型对变量为线性？	
	是	不是
是	LRM	LRM
不是	NLRM	NLRM

注：LRM＝线性回归模型；
NLRM＝非线性回归模型。

2.4 PRF 的随机设定

从图 2—1 清楚地看到，随着家庭收入的增加，家庭消费支出平均地说也增加。但是，对某一特定家庭来说，消费支出与其（固定的）收入水平的关系怎样？从表 2—1 和图 2—1 明显看出，某一特定家庭的消费支出不一定随收入水平增加而增加。例如，从表 2—1 我们观察到，对应于每周 100 美元的收入水平，有一个家庭的消费支出是 65 美元，少于每周收入仅为 80 美元的两个家庭的消费支出（70 美元和 75 美元）。但应看到，家庭每周收入为 100 美元的平均消费支出比家庭每周收入为 80 美元的平均消费支出大（77 美元对 65 美元）。

那么，在特定家庭的消费支出与给定收入水平之间能有什么关系呢？我们从图 2—1 看到，给定收入水平 X_i 的特定家庭的消费支出聚集在收入为 X_i 的所有家庭的平均消费支出的周围，也就是围绕着它的条件均值而分布。因此，我们可以把个别的 Y_i 围绕它的期望值的离差（deviation）表述如下：

$$u_i = Y_i - E(Y \mid X_i)$$

或者

$$Y_i = E(Y \mid X_i) + u_i \tag{2.4.1}$$

其中离差 u_i 是一个不可观测的可正可负的随机变量，在专业术语中，把 u_i 称为**随机干扰项**（stochastic disturbance term）或**随机误差项**（stochastic error term）。

我们该怎样解释方程（2.4.1）呢？我们可以说，给定 X 水平，特定家庭的支出可表示为两个成分之和：（1）$E(Y \mid X_i)$ 代表相同收入水平的所有家庭的平均消费支出。这一成分被称为**系统性**（systematic）或**确定性**（deterministic）成分，以及（2）u_i 为随机或**非系统性**（nonsystematic）成分。我们很快就要分析这个随机误差项的性质。但现在假定它是所有可能影响 Y 但又未能包括到回归模型中或被忽略的变量的替代（surrogate）或代理（proxy）变量。

假定 $E(Y \mid X_i)$ 对 X_i 为线性的，好比方程（2.2.2）那样，方程（2.4.1）可写为：

$$Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

方程（2.4.2）假定，一个家庭的消费支出线性地依赖于它的收入另加干扰项。例如，给定 X=80 美元（见表 2—1），各家庭的消费支出可表达为：

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1 \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2 \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3 \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4 \\ Y_5 &= 75 = \beta_1 + \beta_2(80) + u_5 \end{aligned} \tag{2.4.3}$$

现在，如果在方程（2.4.1）的两边取期望，就得到

$$E(Y_i \mid X_i) = E[E(Y \mid X_i)] + E(u_i \mid X_i) = E(Y \mid X_i) + E(u_i \mid X_i) \tag{2.4.4}$$

这里我们利用了常数的期望值就是它本身这一事实。[①] 要仔细看清，在方程（2.4.4）中我们取的是以给定的 X 值为条件的条件期望。

因为 $E(Y_i \mid X_i)$ 就是 $E(Y \mid X_i)$，故可从方程（2.4.4）推出：

$$E(u_i \mid X_i) = 0 \tag{2.4.5}$$

因此，假定回归线通过 Y 的条件均值（见图 2—2），意味着 u_i 的条件均值（以给定的 X_i 为条件）就是零。

从前面的讨论显见，如果 $E(u_i \mid X_i) = 0$，则方程（2.2.2）和方程（2.4.2）是等价的。[②] 但是方程（2.4.2）有它的优点，因为它清楚地表明，除收入外，还有影响消费支出的其他变量，并因此不能单凭回归模型中包含的（一个或多个）变量完全解释特定家庭的消费支出。

2.5 随机干扰项的意义

如 2.4 节所指出，干扰项是模型遗漏的而又一起影响着 Y 的全部变量的替代物。明显的问题是：为什么不把这些变量清晰地引进到模型中来？换句话说，为什么不构造一个含有尽可能多个变量的多元回归模型？理由是多方面的。

1. 理论的含糊性。即使有决定 Y 的行为理论，常常也是不完备的。我们可以肯定每周收入 X 影响每周消费支出 Y。但还有什么影响 Y 的其他变量呢，我们不是一无所知，就是不太确定。因此不妨用 u_i 作为模型所排除或忽略的全部变量的替代变量。

2. 数据的欠缺。即使我们明知被忽略变量中的一些变量，并因而考虑用一个多元回归而不是一个简单回归，我们却不一定能得到关于这些变量的数量信息。在经验研究中，人们得不到他们最想要的数据是司空见惯的事。例如，在原理上，除收入外，我们还可引进财富作为家庭消费支出的解释变量。但不幸的是，一般是得不到关于家庭财富的信息的。因此，我们不得不把财富变量从我们的模型中割舍掉。哪怕它在解释消费支出方面有很强的理论重要性。

3. 核心变量与周边变量。假定在我们的消费—收入例子中，除了收入 X_1 外，家庭的子女数 X_2、性别 X_3、宗教 X_4、教育 X_5 和地区 X_6 也影响消费支出。但很可能这些变量的全部或其中的一些，合起来的影响是如此之小，充其量是一种非系统的或随机的影响。从实际考虑以及从成本上计算，把它们一一引入模型是划不来的。

① 关于期望运算子 E 的性质的一个简要讨论，参见附录 A。请注意，一旦 X_i 值被固定，则 $E(Y \mid X_i)$ 就是一个常数。

② 事实上，在第 3 章所讲的最小二乘法中，明显地假定了 $E(u_i \mid X_i) = 0$。见 3.2 节。

所以人们希望把它们的共同影响当作一个随机变量u_i来看待。[①]

4. 人类行为的内在随机性。即使我们成功地把所有有关的变量都引进到模型中来，在个别的Y中仍不免有一些“内在”的随机性，无论我们花了多少力气都解释不了的。干扰项u_i也许能很好地反映这种随机性。

5. 糟糕的替代变量。虽然经典回归模型（将在第3章中讨论）假定变量Y和X能准确地观测，但实际上数据会受到测量误差的干扰。试看弗里德曼的著名的消费函数理论。[②] 他把持久消费（Y^p）看作持久收入（X^p）的函数。但由于这些变量不可直接观测，故实际上我们利用替代变量，诸如可观测的当前消费（Y）和当前收入（X）。而由于所观测的Y和X未必等于Y^p和X^p，这里就有一个测量误差的问题。这时干扰项u又可用来代表测量误差。我们在后面的一章中将会看到，如果有这种误差，回归系数β的估计会受到严重的影响。

6. 节省原则。仿效简单性原则[③]，我们想保持一个尽可能简单的回归模型。如果我们能用两个或三个变量就“基本上”解释了Y的行为，并且如果我们的理论完善或扎实的程度还没有达到足以提出可包含进来的其他变量，那么为什么要引进更多的变量呢？让u_i代表所有的其他变量好了。当然，我们不应该只为了保持回归模型简单而排除有关的和重要的变量。

7. 错误的函数形式。即使我们有了在理论上解释某种现象的正确变量，并且我们能获得这些变量的数据，我们却常常不知道回归子和回归元之间的函数关系式是什么形式。消费支出是收入的线性（对变量而言）函数抑或非线性（对变量而言）函数？如果属于前者，$Y_i=\beta_1+\beta_2 X_i+u_i$就是$Y$和$X$之间的适当函数关系式；但如果属于后者，$Y_i=\beta_1+\beta_2 X_i+\beta_3 X_i^2+u_i$也许才是正确的函数形式。在双变量模型中，人们往往能通过散点图来判断二者关系的函数形式。而在多变量回归模型中，由于无法从图形上想象一个多维的散点图，要决定适当的函数形式就更不容易。

由于所有这些理由，我们随后将看到，随机干扰项在回归分析中扮演着极为重要的角色。

2.6 样本回归函数

我们有意把至今的讨论局限于与固定X值相对应的Y值总体，以避免考虑抽样

① 还有一个困难，即诸如性别、教育、宗教等变量难以量化。

② Milton Friedman, *A Theory of the Consumption Function*, Princeton University Press, Princeton, N. J., 1957.

③ “That descriptions be kept as simple as possible until proved inadequate,” *The World of Mathematics*, vol. 2, J. R. Newman (ed.), Simon & Schuster, New York, 1956, p. 1247, or “Entities should not be multiplied beyond necessity,” Donald F. Morrison, *Applied Linear Statistical Methods*, Prentice Hall, Englewood Cliffs, N. J., 1983, p. 58.

的问题。（注意表 2—1 的数据代表总体，而不是一个样本。）但在大多数实际情况中，我们仅有对应于某些固定 X 的 Y 值的一个样本。所以现在是面对抽样问题的时候了。我们现在的任务是要在样本信息的基础上估计 PRF。

作为一个说明，假装我们不知道表 2—1 的总体数据，我们仅有的信息是表 2—4 给出的对应于固定 X 值的 Y 值的一个随机（抽取的）样本。它和表 2—1 不同，对应于给定的每个 X_i 只有一个 Y 值。表 2—4 中（给定 X_i）的每个 Y 都是从表 2—1 的总体中对应于同一 X_i 的同组 Y 值中随机抽取的。

问题是：我们能从表 2—4 的样本预测整个总体中对应于选定 X 的平均每周消费支出 Y 吗？换句话说，我们能从这些样本数据估计 PRF 吗？读者一定怀疑，由于抽样波动，我们未必能“准确”估计 PRF。为说明这点，设想我们从表 2—1 的总体中抽取另一个随机样本，如表 2—5 所示。

表 2—4　表 2—1 中总体的一个随机样本

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

表 2—5　表 2—1 中总体的另一个随机样本

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

将表 2—4 和表 2—5 的数据描点，得到图 2—4 中的散点图。在这个散点图中画两条样本回归线以尽可能好地拟合这些散点：SRF_1 是根据第一个样本画的；而 SRF_2 是根据第二个样本画的。那么，两条回归线中的哪一条代表“真实”的总体回归线呢？如果我们避免偷看表现出 PR 的图 2—1 的诱惑，我们就不可能有绝对的把握知道图 2—4 中的哪一条回归线代表真实的总体回归线（或曲线）。图 2—4 中的回归线称为**样本回归线**（sample regression lines）。姑且假定它们都代表总体回归线，但因抽样波动它们最多也不过是真实 PR 的一个近似而已。一般地说，从 N 个不同的样本会得到 N 个不同的 SRF，并且这些 SRF 不太可能是一样的。

类比于总体回归线有一个 PRF 作为其基础，现在我们能够写出一个代表样本回归线的**样本回归函数**（sample regression function，SRF）概念。对应于方程（2.2.2）的样本关系式可写为：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \tag{2.6.1}$$

其中 $\hat{Y}_i$ 读作“Y 帽”；

$\hat{Y}_i = E(Y \mid X_i)$ 的估计量；

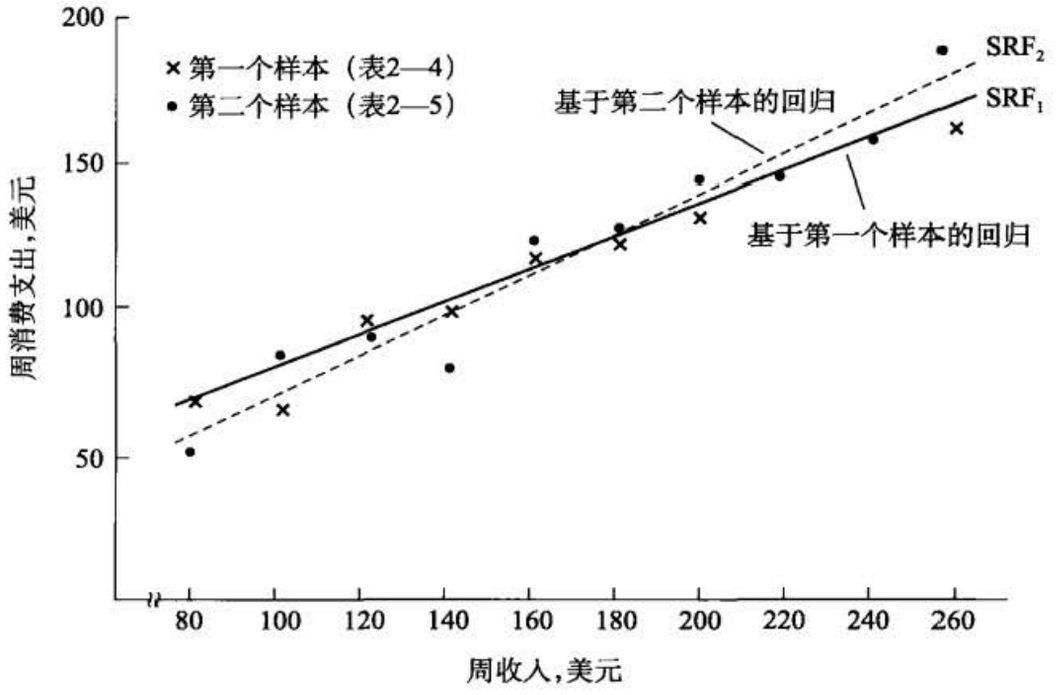


图 2—4　基于两个不同样本的回归线

$\hat{\beta}_1$ = β_1 的估计量；

$\hat{\beta}_2$ = β_2 的估计量。

注意，一个**估计量**（estimator），又称（样本）**统计量**（statistic），是指一个规则或公式或方法，它告诉人们怎样用手中样本所提供的信息去估计总体参数。在一项应用中，由估计量算出的一个具体的数值，称为**估计值**（estimate）。[①] 应该指出，估计量是随机的，而估计量算出的一个具体数值则是非随机的。（为什么?）

正如我们把 PRF 表达成方程（2.2.2）和方程（2.4.2）两种等价形式，我们也能把 SRF 的方程（2.6.1）表达成它的随机形式：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

其中，除已定义过的记号外，$\hat{u}_i$ 表示（样本）**残差**（residual）项。概念上，$\hat{u}_i$ 类似于 u_i，并可把它当作 u_i 的估计量，把它引进到 SRF 中来和把 u_i 引进到 PRF 中来，是出于同一理由。

至此，总的说来，由于我们的分析仅仅依据某总体单个样本的时候要比不是这样的时候多，我们看到，我们在回归分析中的主要目的是根据 SRF 的方程：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i \tag{2.4.2}$$

来估计 PRF 的方程：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.6.2}$$

然而，由于抽样的波动，我们根据 SRF 估计出来的 PRF 充其量也不过是一个近似结

① 在引言中曾经指出，在一个变量的上方加一个帽形，表示有关总体值的一个估计量。

果。图 2—5 对这种近似作了解析。

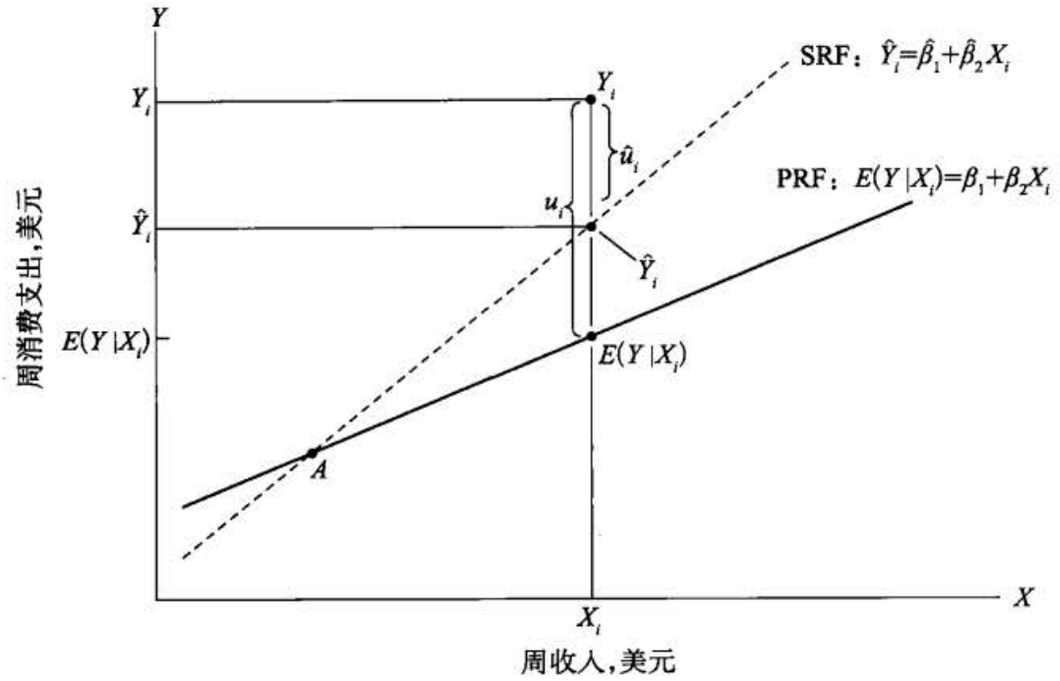


图 2—5　样本与总体回归线

对 $X=X_i$，我们有一个观测值 $Y=Y_i$。利用 SRF 的方程，可将所观测的 Y_i 表达为：

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

而通过 PRF，又可把它表达为：

$$Y_i = E(Y \mid X_i) + u_i \tag{2.6.4}$$

现在，对于图 2—5 中所示的 X_i，$\hat{Y}_i$ 明显过高地估计了真实的 $E(Y \mid X_i)$。类似的图形分析，对 A 点以左的任何 X_i，SRF 过低估计了真实的 PRF。但读者能容易看到，由于抽样的波动，这种过高或过低的估计是不可避免的。

现在，重要的问题是：既然认识到 SRF 只不过是 PRF 的一个近似，能不能设计一种规则或方法，使得这种近似是一种尽可能“接近”的近似？换一种说法，怎样构造 SRF 使得 $\hat{\beta}_1$ 尽可能“接近”真实的 β_1，$\hat{\beta}_2$ 尽可能地“接近”真实的 β_2？尽管真实的 β_1 和 β_2 永远都不知道。

对这些问题的回答，需要我们在第 3 章倾注大量的注意力。这里仅仅指出，我们能够给出这样一种程序，它告诉我们怎样构造 SRF，以尽可能忠实地反映 PRF。试想一想，虽然我们从来没有真正地确定过 PRF，做到这一点该是多么的不可思议。

2.7　说明性例子

我们以两个例子来结束本章。

例 2.1 不同受教育程度的平均小时工资

表 2—6 给出了受教育程度（以读书年数来度量）、每个受教育程度中人们的平均小时工资及各种受教育程度的人数三类数据。伯恩特（Ernst Berndt）最早获得表中列出的数据，他是从 1985 年 5 月的人口普查中推导出这些数据的。[①] 我们在下一章还将解释这些数据（及其他的解释变量）。

表 2—6　　不同受教育程度的平均小时工资

读书年数	平均小时工资，美元	人数
6	4.456 7	3
7	5.770 0	5
8	5.978 7	15
9	7.331 7	12
10	7.318 2	17
11	6.584 4	27
12	7.818 2	218
13	7.835 1	37
14	11.022 3	56
15	10.673 8	13
16	10.836 1	70
17	13.615 0	24
18	13.531 0	31
		总计 528

资料来源：Arthur S. Goldberger, *Introductory Econometrics*, Harvard University Press, Cambridge, Mass., 1998, Table 1.1, p. 5 (adapted).

以受教育程度为横轴、以（条件）平均小时工资为纵轴画图，可得到图 2—6。图中的回归曲线表明了平均小时工资如何随受教育程度而变化；它们通常随着受教育程度的提高而增加，这是一个不足为奇的结论。我们在下一章将研究，其他变量何以影响平均小时工资。

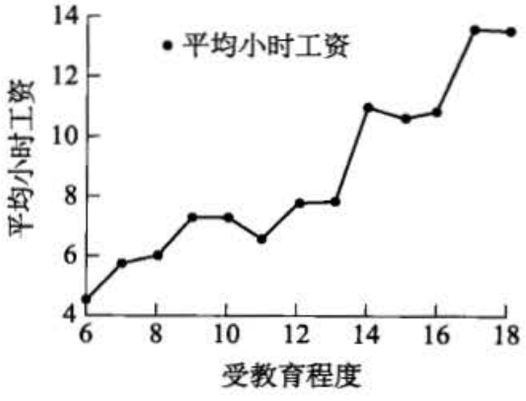


图 2—6　平均小时工资与受教育程度之间的关系

① Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary*, Addison Wesley, Reading, Mass., 1991. 顺便提一句，这是一本读者能从中发现计量经济学如何用于做研究的优秀教材。

例 2.2 SAT 数学成绩与家庭收入

习题 2.17 中的表 2—10，根据 2007 年参加 SAT 考试的 947 347 名考生，给出了他们在阅读、数学和写作方面的 SAT（学术能力倾向测试）平均成绩数据。将数学平均成绩对家庭平均收入进行描点，我们得到图 2—7 所示的图形。

注意，由于表 2—10 中收入的第一个类别和最后一个类别没有明确的区间限制，所以我们假定最低的家庭平均收入为 5 000 美元，而最高的家庭平均收入为 150 000 美元。

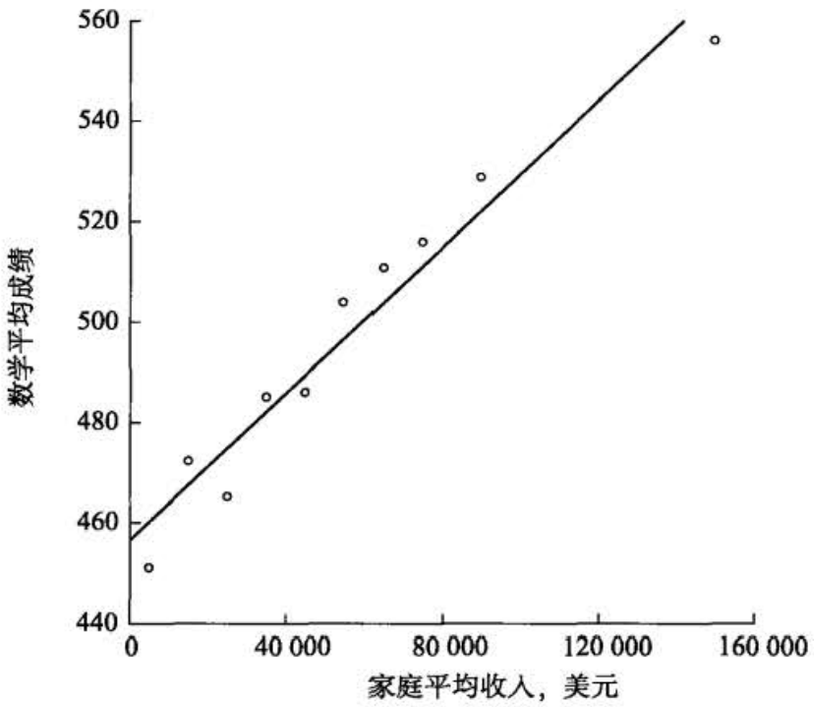


图 2—7 SAT 数学平均成绩与家庭平均收入之间的关系

如图 2—7 所示，数学平均成绩随着家庭平均收入的提高而提高。由于参加 SAT 考试的学生人数众多，所以它可能代表了参加 SAT 考试的考生总体。因此，图 2—7 中勾勒的回归线就可能表示了总体回归线。

观察到这两个变量之间的正相关关系，可能是有一些原因的。比如，有人认为家庭收入越高的学生越有能力支付 SAT 考试的辅导费用。此外，家庭收入越高的学生，其父母受教育程度也可能越高。还有一种可能，数学成绩越高的学生来自更好的学校。对于这两个变量之间观察到的正相关关系，读者也可以给出其他解释。

要点与结论

1. 作为回归分析基础的主要概念是**条件期望函数**或**总体回归函数**。我们做回归分析的目标就是要发现，因变量（回归子）的均值如何随着给定解释变量（回归元）的变化而变化。

2. 本书研究**线性 PRF**，也就是对未知参数为线性的回归。这些回归对因变量或回归子以及自变量或回归元（一个或多个）来说，可以是线性的，也可以不是线性的。

3. 出于经验研究的目的，重要的是**随机的 PRF**。在 PRF 的估计中，**随机干扰项** u_i 起着关键性作用。

4. PRF 是一个理想化的概念。实际上，人们很少得知他们所研究的整个总体。通常他们只拥有对这个总体的一个观测样本，因此要用**随机样本回归函数**去估计 PRF。第 3 章考虑怎样实现这一点。

习　　题

问答题

2.1　什么是条件期望函数或总体回归函数？

2.2　总体回归函数和样本回归函数之间的差别是什么？这是不是人为的区别？

2.3　回归分析中的随机误差项 u_i 有什么作用？它与残差 a_i 有何区别？

2.4　我们为什么需要回归分析？我们为什么不简单地用回归子的均值作为最优值？

2.5　线性回归模型的含义是什么？

2.6　判别如下模型是线性于参数、线性于变量还是同时线性于参数和变量，哪些模型是线性回归模型？

模型	描述性名称
a. $Y_i = \beta_1 + \beta_2\left(\frac{1}{X_i}\right) + u_i$	倒数
b. $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	半对数
c. $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	反半对数
d. $\ln Y_i = \ln\beta_1 + \beta_2 \ln X_i + u_i$	对数或双对数
e. $\ln Y_i = \beta_1 - \beta_2\left(\frac{1}{X_i}\right) + u_i$	对数倒数

注：ln 表示自然对数（即对数的底为 e）；u_i 为随机干扰项。我们将在第 6 章研究这些模型。

2.7　如下模型是线性回归模型吗？为什么？

a. $Y_i = e^{\beta_1+\beta_2 X_i+u_i}$　　b. $Y_i = \dfrac{1}{1+e^{\beta_1+\beta_2 X_i+u_i}}$

c. $\ln Y_i = \beta_1 + \beta_2\left(\dfrac{1}{X_i}\right) + u_i$　　d. $Y_i = \beta_1 + (0.75-\beta_1)e^{-\beta_2(X_i-2)} + u_i$

e. $Y_i = \beta_1 + \beta_2^3 X_i + u_i$

2.8　内在线性回归模型的含义是什么？如果习题 2.7d 中的 β_2 为 0.8，那它是一个线性回归模型，还是非线性回归模型？

2.9　考虑如下非随机模型（即不含随机误差项的模型）。它们是线性回归模型吗？若不是，可以通过适当的代数变换使之转化为线性模型吗？

a. $Y_i = \dfrac{1}{\beta_1 + \beta_2 X_i}$

b. $Y_i = \dfrac{X_i}{\beta_1 + \beta_2 X_i}$

c. $Y_i = \dfrac{1}{1 + \exp(-\beta_1 - \beta_2 X_i)}$

2.10 图 2—8 是一个散点图及其回归线。你从此图能得出什么一般性结论？图中勾画出的回归线是总体回归线还是样本回归线？

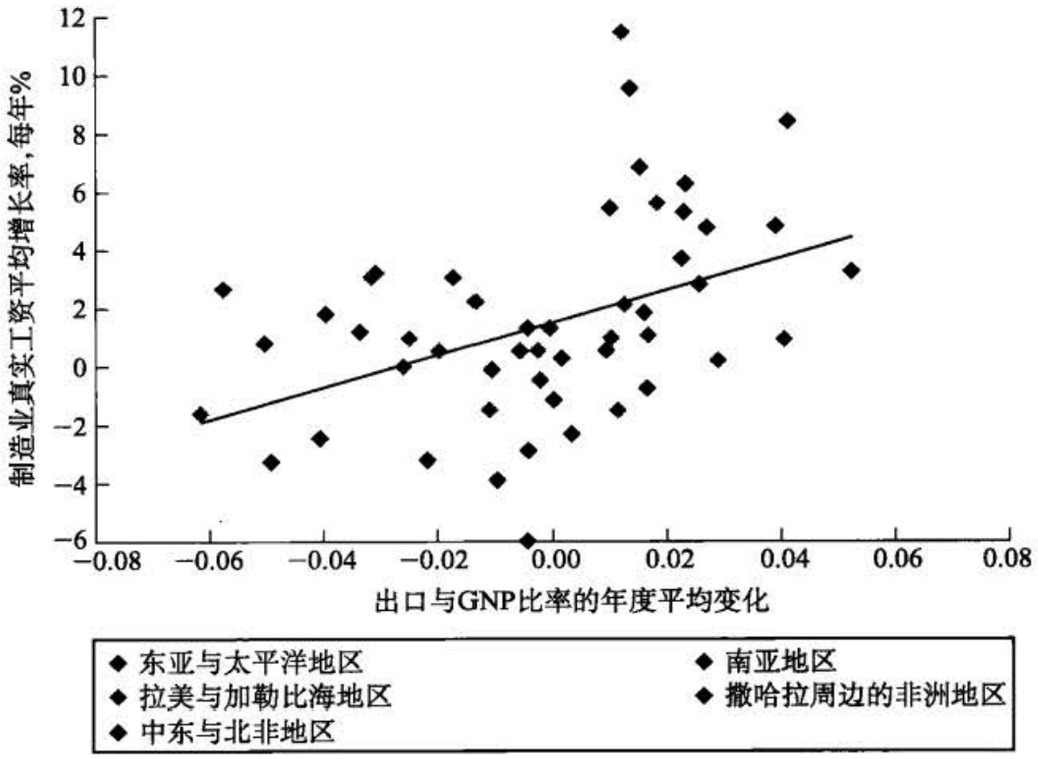


图 2—8 1970—1990 年间 50 个发展中国家的制造业真实工资增长率与出口数据

资料来源：The World Bank, *World Development Report 1995*, p. 55. The original source is UNIDO data, World Bank data.

2.11 你能从图 2—9 的散点图中得出什么一般性结论？其背后有什么经济理论？[提示：查一本国际经济学教科书并阅读贸易的赫克歇尔-俄林（Heckscher-Ohlin）模型。]

2.12 图 2—10 中的散点图揭示了什么关系？基于此图，你认为最低工资法有利于经济福利吗？

2.13 引言中图 I—3 所示的回归线是 PRF 还是 SRF？为什么？你如何解释回归线周围的散点？除 GDP 外，还会有什么其他的因素或变量决定着个人消费支出？

实证分析题

2.14 表 2—7 给出了美国 1980—2006 年间的数据。

a. 将城市男性劳动力劳动参与率相对城市男性失业率描点。目测一条穿过散点的回归线。推测二者之间的关系，其背后的经济理论是什么？这个散点图支持该理论吗？

b. 对女性重做（a）部分的练习。

c. 现在同时将男性和女性的劳动参与率相对平均小时工资（以 1982 年美元度量）描点。（你可以分开画图）现在又有何发现？你又将如何解释你的发现？

d. 你可以将劳动参与率同时对失业率和平均小时工资描点吗？若不能，你如何口头说明这三个变量之间的关系。

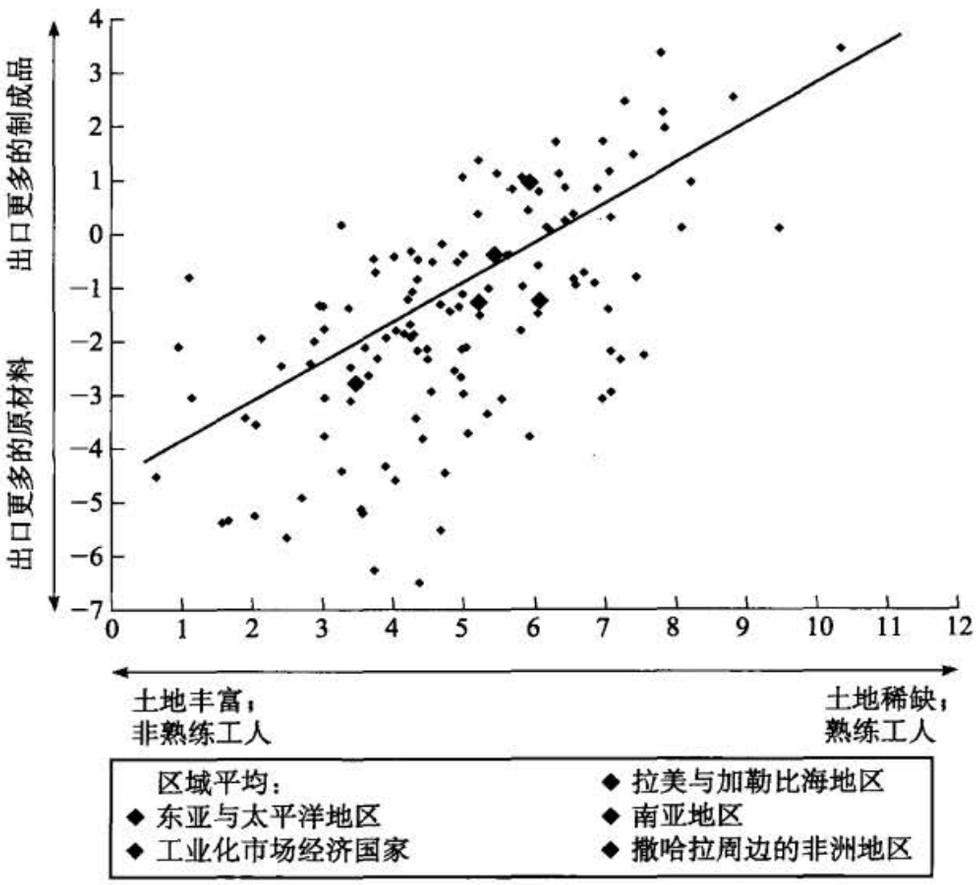


图 2—9　1985 年 126 个工业化和发展中国家的出口与技术密集性和人力资本数据

注：横轴数据为一个国家平均受教育年数与其土地面积之比的对数；纵轴为制成品出口与初级产品出口之比的对数。

资料来源：World Bank，*World Development Report* 1995，p. 59. 原始资料：出口数据摘自美国统计局 COMTRADE 数据库；受教育数据摘自 UNDP 1990；土地数据摘自世界银行。

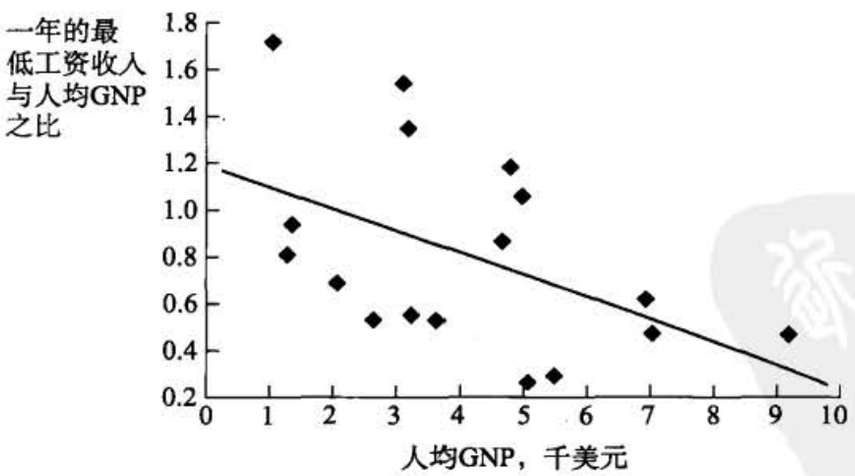


图 2—10　最低工资与人均 GNP

注：样本由 17 个发展中国家构成。各国数据采集的年份从 1988 年到 1992 年不等。数据以国际价格度量。

资料来源：World Bank，*World Development Report 1995*，p. 75.

表 2—7　　1980—2006 年间美国劳动参与率数据

年份	CLFPRM[1]	CLFPRF[2]	UNRM[3]	UNRF[4]	AHE82[5]	AHE[6]
1980	77.400 00	51.500 00	6.900 000	7.400 000	7.990 000	6.840 000
1981	77.000 00	52.100 00	7.400 000	7.900 000	7.880 000	7.430 000
1982	76.600 00	52.600 00	9.900 000	9.400 000	7.860 000	7.860 000
1983	76.400 00	52.900 00	9.900 000	9.200 000	7.950 000	8.190 000
1984	76.400 00	53.600 00	7.400 000	7.600 000	7.950 000	8.480 000
1985	76.300 00	54.500 00	7.000 000	7.400 000	7.910 000	8.730 000
1986	76.300 00	55.300 00	6.900 000	7.100 000	7.960 000	8.920 000
1987	76.200 00	56.000 00	6.200 000	6.200 000	7.860 000	9.130 000
1988	76.200 00	56.600 00	5.500 000	5.600 000	7.810 000	9.430 000
1989	76.400 00	57.400 00	5.200 000	5.400 000	7.750 000	9.800 000
1990	76.400 00	57.500 00	5.700 000	5.500 000	7.660 000	10.190 000
1991	75.800 00	57.400 00	7.200 000	6.400 000	7.580 000	10.500 000
1992	75.800 00	57.800 00	7.900 000	7.000 000	7.550 000	10.760 000
1993	75.400 00	57.900 00	7.200 000	6.600 000	7.520 000	11.030 000
1994	75.100 00	58.800 00	6.200 000	6.000 000	7.530 000	11.320 000
1995	75.000 00	58.900 00	5.600 000	5.600 000	7.530 000	11.640 000
1996	74.900 00	59.300 00	5.400 000	5.400 000	7.570 000	12.030 000
1997	75.000 00	59.800 00	4.900 000	5.000 000	7.680 000	12.490 000
1998	74.900 00	59.800 00	4.400 000	4.600 000	7.890 000	13.000 000
1999	74.700 00	60.000 00	4.100 000	4.300 000	8.000 000	13.470 000
2000	74.800 00	59.900 00	3.900 000	4.100 000	8.030 000	14.000 000
2001	74.400 00	59.800 00	4.800 000	4.700 000	8.110 000	14.530 000
2002	74.100 00	59.600 00	5.900 000	5.600 000	8.240 000	14.950 000
2003	73.500 00	59.500 00	6.300 000	5.700 000	8.270 000	15.350 000
2004	73.300 00	59.200 00	5.600 000	5.400 000	8.230 000	15.670 000
2005	73.300 00	59.300 00	5.100 000	5.100 000	8.170 000	16.110 000
2006	73.500 00	59.400 00	4.600 000	4.600 000	8.230 000	16.730 000

注：表中变量的定义及其在文献中的位置如下：

[1]CLFPRM，城市男性劳动力劳动参与率（%），Table B-39，p. 277。

[2]CLFPRF，城市女性劳动力劳动参与率（%），Table B-39，p. 277。

[3]UNRM，城市男性失业率（%），Table B-42，p. 280。

[4]UNRF，城市女性失业率（%），Table B-42，p. 280。

[5]AHE82，平均小时工资（以 1982 年美元度量），Table B-47，p. 286。

[6]AHE，平均小时工资（以当前价格度量），Table B-47，p. 286。

资料来源：*Economic Report of the President*，*2007*.

2.15　表 2—8 给出的是以卢比度量的食物支出和总支出数据，样本是印度的 55 个农户。（在 2000 年初，1 美元约兑换 40 卢比。）

表 2—8　　食物支出与总支出　　（单位：卢比）

观测	食物支出	总支出	观测	食物支出	总支出
1	217.000 0	382.000 0	29	390.000 0	655.000 0
2	196.000 0	388.000 0	30	385.000 0	662.000 0
3	303.000 0	391.000 0	31	470.000 0	663.000 0
4	270.000 0	415.000 0	32	322.000 0	677.000 0
5	325.000 0	456.000 0	33	540.000 0	680.000 0
6	260.000 0	460.000 0	34	433.000 0	690.000 0
7	300.000 0	472.000 0	35	295.000 0	695.000 0
8	325.000 0	478.000 0	36	340.000 0	695.000 0
9	336.000 0	494.000 0	37	500.000 0	695.000 0
10	345.000 0	516.000 0	38	450.000 0	720.000 0
11	325.000 0	525.000 0	39	415.000 0	721.000 0
12	362.000 0	554.000 0	40	540.000 0	730.000 0
13	315.000 0	575.000 0	41	360.000 0	731.000 0
14	355.000 0	579.000 0	42	450.000 0	733.000 0
15	325.000 0	585.000 0	43	395.000 0	745.000 0
16	370.000 0	586.000 0	44	430.000 0	751.000 0
17	390.000 0	590.000 0	45	332.000 0	752.000 0
18	420.000 0	608.000 0	46	397.000 0	752.000 0
19	410.000 0	610.000 0	47	446.000 0	769.000 0
20	383.000 0	616.000 0	48	480.000 0	773.000 0
21	315.000 0	618.000 0	49	352.000 0	773.000 0
22	267.000 0	623.000 0	50	410.000 0	775.000 0
23	420.000 0	627.000 0	51	380.000 0	785.000 0
24	300.000 0	630.000 0	52	610.000 0	788.000 0
25	410.000 0	635.000 0	53	530.000 0	790.000 0
26	220.000 0	640.000 0	54	360.000 0	795.000 0
27	403.000 0	648.000 0	55	305.000 0	801.000 0
28	350.000 0	650.000 0			

资料来源：Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries*, Routledge, New York, 1998, p. 457.

a. 以总支出为横轴，食物支出为纵轴将数据描点，并勾勒出一条穿过散点的回归线。

b. 你从此例中能得出什么一般性的结论？

c. 据经验，你会预测无论总支出水平如何食物支出总是随总支出线性地增加吗？为什么？你可以用总支出作为总收入的一个代理变量。

2.16　表 2—9 给出了 1972—2007 年间应届高中毕业生在 SAT 中的平均成绩数据。这些数据包括男生和女生在阅读和数学方面的成绩。2006 年才开始要求写作，所以这些数据中不包含写作成绩。

表 2—9　　1972—2007 年应届高中毕业生在 SAT 逻辑部分的分组平均成绩

年份	阅读			数学		
	男生	女生	总平均	男生	女生	总平均
1972	531	529	530	527	489	509
1973	523	521	523	525	489	506
1974	524	520	521	524	488	505
1975	515	509	512	518	479	498
1976	511	508	509	520	475	497
1977	509	505	507	520	474	496
1978	511	503	507	517	474	494
1979	509	501	505	516	473	493
1980	506	498	502	515	473	492
1981	508	496	502	516	473	492
1982	509	499	504	516	473	493
1983	508	498	503	516	474	494
1984	511	498	504	518	478	497
1985	514	503	509	522	480	500
1986	515	504	509	523	479	500
1987	512	502	507	523	481	501
1988	512	499	505	521	483	501
1989	510	498	504	523	482	502
1990	505	496	500	521	483	501
1991	503	495	499	520	482	500
1992	504	496	500	521	484	501
1993	504	497	500	524	484	503
1994	501	497	499	523	487	504
1995	505	502	504	525	490	506
1996	507	503	505	527	492	508
1997	507	503	505	530	494	511
1998	509	502	505	531	496	512
1999	509	502	505	531	495	511
2000	507	504	505	533	498	514
2001	509	502	506	533	498	514
2002	507	502	504	534	500	516
2003	512	503	507	537	503	519
2004	512	504	508	537	501	518
2005	513	505	508	538	504	520
2006	505	502	503	536	502	518
2007	504	502	502	533	499	515

注：1972—1986 年间用一个公式把原始均值及其标准差转化成标准分的均值。1987—1995 年间，把各个学生的成绩转化成标准分再计算均值。1996—1999 年间，几乎所有学生的成绩都是标准分。2000—2007 年间，所有成绩都是用标准分报告。

资料来源：College Board，2007.

a. 用横轴代表年度，纵轴代表 SAT 成绩，分别对男生和女生描绘阅读和数学成绩。

b. 你从这些图形中能得出什么一般性的结论？

c. 知道了男（女）生的阅读成绩，你会怎样预测相应的数学成绩？

d. 将女生总的数学成绩对男生的数学成绩描点。你看到了什么？

2.17　表 2—10 根据家庭收入给出了阅读、数学和写作三项 SAT 逻辑部分的平均成绩。我们在例 2.2 中给出了图 2—7，即将数学平均成绩对家庭平均收入的描点图。

表 2—10　　根据家庭收入分组的 SAT 逻辑部分的平均成绩

家庭收入（美元）	应试人数	阅读		数学		写作	
		均值	标准差	均值	标准差	均值	标准差
＜10 000	40 610	427	107	451	122	423	104
10 000～20 000	72 745	453	106	472	113	446	102
20 000～30 000	61 244	454	102	465	107	444	97
30 000～40 000	83 685	476	103	485	106	466	98
40 000～50 000	75 836	489	103	486	105	477	99
50 000～60 000	80 060	497	102	504	104	486	98
60 000～70 000	75 763	504	102	511	103	493	98
70 000～80 000	81 627	508	101	516	103	498	98
80 000～100 000	130 752	520	102	529	104	510	100
＞100 000	245 025	544	105	556	107	537	103

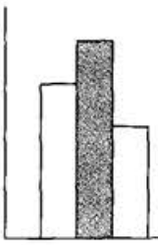
资料来源：College Board，2007，College-Bound Seniors，Table 11.

a. 参照图 2—7 画出阅读平均成绩与家庭平均收入之间的类似图形。将你的结论与图 2—7 中所示的图形进行比较。

b. 再做出写作平均成绩与家庭平均收入之间的图形，并与上述两个图形进行比较。

c. 根据这三个图形，你能得出什么一般性的结论？

第 3 章 双变量回归模型：估计问题



第 2 章中曾经指出，我们的首要任务是根据样本回归函数尽可能准确地估计总体回归函数。在附录 A 中，我们讨论了一般常用的两种估计方法：(1) **普通最小二乘法** (ordinary least squares，OLS) 和 (2) **极大似然法** (maximum likelihood，ML)。一般来说，普通最小二乘法在回归分析中应用得很广泛，主要是因为它颇具直觉吸引力，并且在数学上也比极大似然法简单得多。除此之外，就像我们将要说明的那样，在线性回归的背景中，这两种方法通常都会得到相同的结果。

3.1 普通最小二乘法

普通最小二乘法由德国数学家高斯 (Carl Friedrich Gauss) 提出。在一定的假定条件下 (见 3.2 节)，最小二乘法有一些非常有吸引力的统计性质，从而使之成为回归分析中最有功效和最为流行的方法之一。为了说明这个方法，我们先解释最小二乘原理。

回顾双变量 PRF：

$$Y_i=\beta_1+\beta_2 X_i+u_i \tag{2.4.2}$$

然而，如在第 2 章所提到的那样，这个 PRF 是无法直接观测的。我们通过 SRF 去估计它：

$$Y_i=\hat{\beta}_1+\hat{\beta}_2 X_i+\hat{u}_i \tag{2.6.2}$$

$$=\hat{Y}_i+\hat{u}_i \tag{2.6.3}$$

其中 $\hat{Y}_i$ 是 Y_i 的估计值（条件均值）。

但 SRF 又是怎样决定的呢？为了看清楚这个问题，让我们一步步解说如下。首先把方程（2.6.3）写成：

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{aligned} \tag{3.1.1}$$

这表明 $\hat{u}_i$（即残差）不过是 Y 的实际值与估计值之差。

对于给定的 Y 和 X 的 n 对观测值，我们希望这样决定 SRF，使得它尽可能靠近实际的 Y。为此，我们可以采用如下准则：选择这样的 SRF，使得残差和 $\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$ 尽可能小。这看来尽管有直观上的说服力，却不是一个很好的准则。这可以从图 3—1 中一个假想的散点图看出。

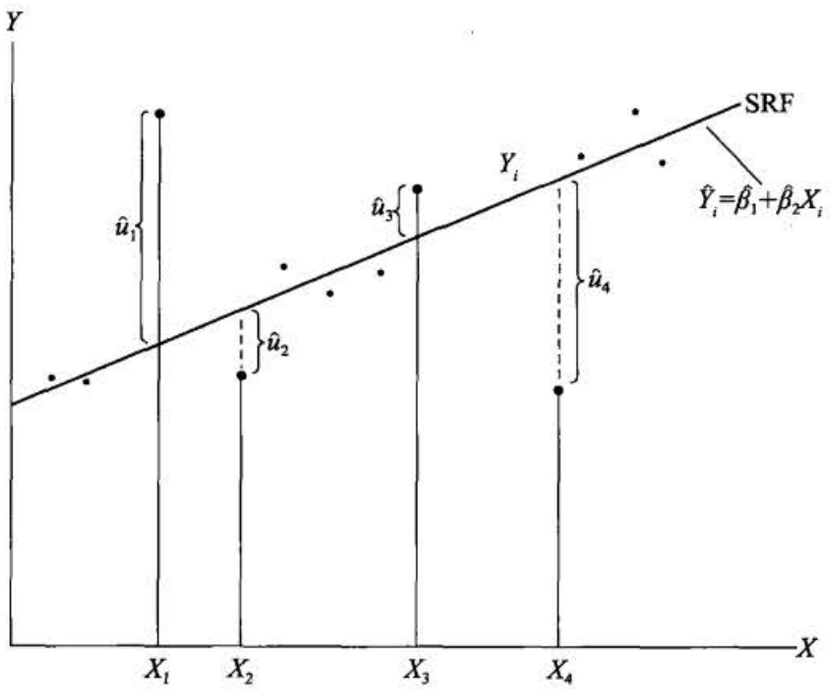


图 3—1　最小二乘准则

如果采用 $\sum \hat{u}_i$ 最小化的准则，那么在总和（$\hat{u}_1+\hat{u}_2+\hat{u}_3+\hat{u}_4$）中残差 $\hat{u}_2$ 及 $\hat{u}_3$ 得到的权重和 $\hat{u}_1$ 及 $\hat{u}_4$ 得到的权重一样多，尽管前两个残差比后两个残差更加靠近 SRF。换言之，不管各个观测点离 SRF 有多远，所有残差都受到同样的重视。因此，很可能 $\hat{u}_i$ 偏离 SRF 而散布得很远，但 $\hat{u}_i$ 的代数和却很小（甚至是零）。为了看清楚这一点，假定图 3—1 中的 $\hat{u}_1$、$\hat{u}_2$、$\hat{u}_3$ 和 $\hat{u}_4$ 分别取值 10、−2、+2、−10，虽然 $\hat{u}_1$ 和 $\hat{u}_4$ 比 $\hat{u}_2$ 和 $\hat{u}_3$ 偏离 SRF 远得多，但这些残差的代数和却是零。如果我们采用最小二乘准则，就可避免这种问题。最小二乘准则是要定出 SRF 使得下式尽可能地小：

$$\begin{aligned} \sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \end{aligned} \tag{3.1.2}$$

其中 a_i^2 是残差的平方。该方法通过对 a_i 平方而赋予诸如图 3—1 中的 a_1 和 a_4 比 a_2 和 a_3 更大的权重。如前所述，在 $\sum a_i$ 最小化的准则下，虽然 a_i 在 SRF 周围散布得很宽，但其总和可能很小。而在最小二乘法中，这是不可能的，因为 a_i（在绝对值上）越大，$\sum a_i^2$ 也越大。采用最小二乘法的理由如我们即将看到的那样，还在于由它得出的估计量有一些很好的统计性质。

由方程（3.1.2）明显地看到：

$$\sum a_i^2 = f(\hat{\beta}_1, \hat{\beta}_2) \tag{3.1.3}$$

也就是说，残差平方和是估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的某个函数。对任意给定的一组数据，选择不同的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 值将得到不同的 a，从而 $\sum a_i^2$ 有不同的值。为了清楚地看到这一点，考虑由表 3—1 前两列给出的 Y 和 X 的一些假想数据。现在做两个实验。在实验 1 中，取 $\hat{\beta}_1 = 1.572$ 和 $\hat{\beta}_2 = 1.357$（暂不必问这两个数值是怎样得来的，就算是一种猜测好了）。[①] 利用这些 $\hat{\beta}$ 值和表 3—1 第（2）列的 X 值，便容易算得该表第（3）列对 Y_i 的估计值，记作 $\hat{Y}_{1i}$（下标 1 表示第 1 个实验）。然后再做一个实验，但这回利用 $\hat{\beta}_1 = 3$ 和 $\hat{\beta}_2 = 1$ 两个值。把 Y_i 的估计值记为 $\hat{Y}_{2i}$，列在表 3—1 的第（6）列。如表所示，由于两个实验的 $\hat{\beta}$ 值有所不同，所估计的残差也有所不同；a_{1i} 是得自第 1 个实验的残差，而 a_{2i} 是得自第 2 个实验的残差，其平方列于第（5）列和第（8）列。显然，可从方程（3.1.3）预料到，由于所依据的 $\hat{\beta}$ 值不同，所以这些残差的平方和也不相同。

表 3—1　　通过实验确定 SRF

Y_i	X_i	$\hat{Y}_{1i}$	a_{1i}	a_{1i}^2	$\hat{Y}_{2i}$	a_{2i}	a_{2i}^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
4	1	2.929	1.071	1.147	4	0	0
5	4	7.000	−2.000	4.000	7	−2	4
7	5	8.357	−1.357	1.841	8	−1	1
12	6	9.714	2.286	5.226	9	3	9
总和：28	16		0.0	12.214		0	14

注：$\hat{Y}_{1i} = 1.572 + 1.357X_i$（即 $\hat{\beta}_1 = 1.572$ 和 $\hat{\beta}_2 = 1.357$）；

$\hat{Y}_{2i} = 3.0 + 1.0X_i$（即 $\hat{\beta}_1 = 3$ 和 $\hat{\beta}_2 = 1.0$）；

$a_{1i} = (Y_i - \hat{Y}_{1i})$；

$a_{2i} = (Y_i - \hat{Y}_{2i})$。

那么，我们应选取哪一组 $\hat{\beta}$ 值呢？因为第 1 个实验的 $\hat{\beta}$ 值比第 2 个实验的 $\hat{\beta}$ 值给出一个更低的 $\sum a_i^2$（12.214 小于 14）。所以说第 1 个实验的 $\hat{\beta}$ 值是“最优”值。但

① 对好奇的读者补充一句，这些值是由下面即将讨论的最小二乘法得到的。见方程（3.1.6）和（3.1.7）。

怎样知道是最优呢？如果我们的时间和耐心都是无限的，我们就能做许多类似的实验，每次选择不同的$\hat{\beta}$值，然后比较所得到的$\sum a_i^2$，并从中选择给出可能最小的$\sum a_i^2$值的那组$\hat{\beta}$值。当然，这里假定我们已经考虑过所有可想象到的β_1和β_2值。但由于时间和耐心无疑是有限的，所以我们必须考虑这种试错法的某种捷径。幸运的是，最小二乘法为我们提供了这一捷径。由最小二乘法原理或方法选出的$\hat{\beta}_1$和$\hat{\beta}_2$，将使得对于给定样本或一组数据，$\sum a_i^2$尽可能最小。换言之，对于一个给定样本，最小二乘法为我们提供使得$\sum a_i^2$达到最小可能值的β_1和β_2估计值。怎样做到这一点呢？这是微积分学的一个直接运算，如附录3A的3A.1节所示，通过微分法将得到用于估计β_1和β_2的下列方程：

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \tag{3.1.4}$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \tag{3.1.5}$$

其中n是样本容量。这组联立方程被称为**正规方程**（normal equations）。

解此联立方程得：

$$\begin{aligned}\hat{\beta}_2 &= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum x_i y_i}{\sum x_i^2}\end{aligned} \tag{3.1.6}$$

其中$\bar{X}$和$\bar{Y}$是X和Y的样本均值，并且定义$x_i = (X_i - \bar{X})$和$y_i = (Y_i - \bar{Y})$。从此以后，我们将遵循一个惯例：用小写字母表示一个变量与其均值的离差。

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned} \tag{3.1.7}$$

方程（3.1.7）中的最后一步是通过简单的代数运算而直接从方程（3.1.4）得到的。

顺便指出，利用简单的代数恒等式，用于估计β_2的公式（3.1.6）可另表述为：

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \tag{3.1.8}$$ ①

前面得到的估计量是从最小二乘原理推导出来的，所以叫做**最小二乘估计量**（least-squares estimators）。注意由普通最小二乘法得到的估计量的如下**数值性质**（numerical property）："数值性质是指由于运用普通最小二乘法而得以成立的那些性质，而不管数据是怎样生成的。"②稍后，我们还将考虑OLS估计量的**统计性质**（statistical property），也就是"仅在数据生成的方式满足一定假设条件下才得以成立"的性质。③（参看3.2节中的经典线性回归模型。）

I. OLS估计量是纯粹由可观测的（即样本）量（指X和Y）表达的，因此它们很容易计算。

II. 它们是**点估计量**（point estimators），即对于给定样本，每个估计量仅提供有关总体参数的一个（点）值。[在第5章，我们将考虑所谓**区间估计量**（interval estimators），后者对未知总体参数的可能值提供一个区间。]

III. 一旦从样本数据得到OLS估计值，便容易画出样本回归线（图3—1）。这样得到的回归线具有如下性质：

1. 如图3—2所示，它穿过Y和X的样本均值点。这是从方程（3.1.7）得出的显而易见的事实，因为该方程可写为$\bar{Y}=\hat{\beta}_1+\hat{\beta}_2\bar{X}$。

2. Y的估计值（$=\hat{Y}_i$）的均值等于实际Y值的均值。因为：

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ &= (\bar{Y} - \hat{\beta}_2\bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} + \hat{\beta}_2(X_i - \bar{X})\end{aligned} \tag{3.1.9}$$

将最后一个等式两边对样本值求和并同时除以样本容量n，即得：

---

① 注1：由于$\bar{X}$为一常数，所以$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2$。而且注意到$\sum X_i = n\bar{X}$和$\sum \bar{X}^2 = n\bar{X}^2$，最后便得到$\sum x_i^2 = \sum X_i^2 - n\bar{X}^2$。

注2：由于$\bar{Y}$为一常数，并且一个变量与其均值的离差的总和恒为零[例如，$\sum (X_i - \bar{X}) = 0$]，故有$\sum x_i y_i = \sum x_i(Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y}\sum x_i = \sum x_i Y_i - \bar{Y}\sum (X_i - \bar{X}) = \sum x_i Y_i$，类似地，有$\sum y_i = \sum (Y_i - \bar{Y}) = 0$。

② Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993, p. 3.

③ 同上。

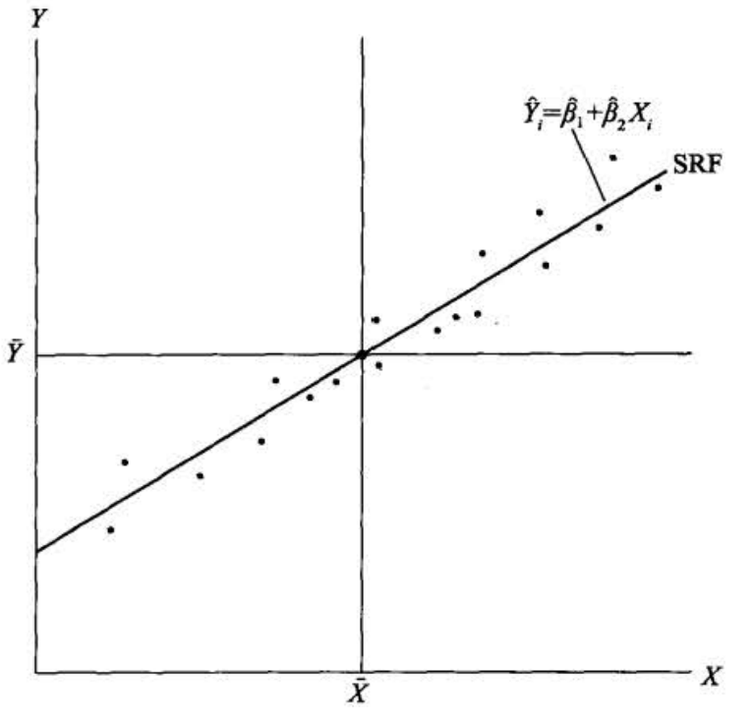


图 3—2　样本回归线穿过 Y 和 X 的样本均值点的图解

$$\bar{\hat{Y}} = \bar{Y} \tag{3.1.10}$$[①]

这里利用了等式 $\sum (X_i - \bar{X}) = 0$。(为什么?)

3. 残差 $\hat{u}_i$ 的均值为零。由附录 3A 的 3A.1 节，第一个方程是：

$$-2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

但由于 $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$，故上述方程可转化为 $-2\sum \hat{u}_i = 0$，从而 $\bar{u} = 0$。[②]

作为上述性质的一个结果，样本回归

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

可表达为另一种形式，其中 Y 和 X 都表示为对其均值的离差。为了看清楚这一点，由于 $\sum \hat{u}_i = 0$，所以对方程 (2.6.2) 两边求和，得到：

$$\begin{aligned}\sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \text{由于} \sum \hat{u}_i = 0\end{aligned} \tag{3.1.11}$$

方程 (3.1.11) 两边同时除以 n 得：

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \tag{3.1.12}$$

这无异于方程 (3.1.7)。从方程 (2.6.2) 减去方程 (3.1.12) 得到：

---

① 注意，这个结果仅当回归模型含有截距 β_1 时才是正确的。如附录 6A 中 6A.1 节所示，当模型不含有 β_1 时，这个结果不一定成立。

② 这一结果也要求截距项 β_1 必须在模型中出现（参看附录 6A 中 6A.1 节）。

$$Y_i - \bar{Y} = \hat{\beta}_2(X_i - \bar{X}) + a_i$$

或者

$$y_i = \hat{\beta}_2 x_i + a_i \tag{3.1.13}$$

其中，按照惯例，y_i 和 x_i 分别是 y 和 x 与其（样本）均值的离差。

方程（3.1.13）被称为**离差形式**（deviation form）。注意，在此形式中，截距项 $\hat{\beta}_1$ 不再出现。然而截距项总可以从方程（3.1.7）估计出来，也就是从样本回归线通过 Y 和 X 的样本均值这一事实估计出来。离差形式的好处是常常能够简化计算。

顺便指出，按照离差形式，SRF 可被写成：

$$\hat{y}_i = \hat{\beta}_2 x_i \tag{3.1.14}$$

而按照原来的度量单位则是 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$，如方程（2.6.1）所示。

4. 残差 a_i 和 Y_i 的预测值不相关。这一命题可验证如下，利用离差形式，可推出：

$$\begin{aligned}
\sum \hat{y}_i a_i &= \hat{\beta}_2 \sum x_i a_i \\
&= \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i) \\
&= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\
&= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \\
&= 0
\end{aligned} \tag{3.1.15}$$

其中我们利用了 $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$ 这一事实。

5. 残差 a_i 和 X_i 不相关；即 $\sum a_i X_i = 0$。这一事实可从附录 3A 第 3A.1 节的方程（2）推知。

3.2 经典线性回归模型：最小二乘法的基本假定

如果我们的目的仅是估计 β_1 和 β_2，那么上一节所讨论的 OLS 法就足够用了。但回顾一下第 2 章，在回归分析中我们的目的不仅仅是获得 $\hat{\beta}_1$ 和 $\hat{\beta}_2$，而且要对真实的 β_1 和 β_2 作出推断。例如，我们想知道 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 与它们相应的总体值有多接近，或者 $\hat{Y}_i$ 与其真实的 $E(Y \mid X_i)$ 有多接近。为达到这一目的，我们不仅要如同方程（2.4.2）那样设定模型的函数形式，还要对 Y_i 的生成方式做出一些假定。为了看清楚为什么需要做出这种要求，让我们看一下 PRF：$Y_i = \beta_1 + \beta_2 X_i + u_i$，它表明 Y_i 依赖于 X_i 和

u_i。因此，除非我们明确 X_i 和 u_i 是怎样生成的，否则我们将无法对 Y_i 进行任何统计推断，而且我们将会看到，也无法对 β_1 和 β_2 作出任何统计推断。也就是说，要对回归估计值做出可靠的解释，对 X_i 变量（一个或多个）和误差项做出假定是极其重要的。

经典（又称高斯或标准）线性回归模型（classical linear regression model，CLRM）。这一模型已成为大部分计量经济学理论的基石，它有 7 个假定。[①] 我们先从双变量回归模型的框架来讨论这些假定；等到第 7 章我们再把这些假定推广到多变量回归模型，即含有不止一个回归元的多元回归模型。

假定 1：线性回归模型。回归模型尽管对变量而言不一定是线性的，但它对于参数而言是线性的，也就是说，回归模型如方程（2.4.2）所示

$$Y_i=\beta_1+\beta_2 X_i+u_i \tag{2.4.2}$$

就像在第 7 章将要讨论的那样，这个模型可以扩展到不止一个解释变量的情形。

我们已在第 2 章中讨论过模型（2.4.2）。因为对于参数而言线性的回归模型是 CLRM 的出发点，所以我们在本书中将始终维持这一假定。[②] 如第 2 章所讨论的那样，回归子 Y 和回归元 X 本身可以是非线性的。

假定 2：X 值是固定的或独立于误差项。在重复样本中，回归元 X 所取的值被认为是固定的（固定回归元情形），或者与因变量 Y 同时抽取（随机回归元情形）。在后面的情形中，假定 X 变量与误差项是独立的，即 $\text{cov}(X_i, u_i)=0$。

这一点可用表 2—1 中给出的例子来解释。考虑与表中所示收入水平相对应的各个 Y 总体，比如说，把收入值 X 固定在 80 美元的水平上，我们随机地抽取一个家庭，并观测到它的每周家庭消费支出 Y，比方说为 60 美元；仍然把 X 固定在 80 美元，而随机地抽取一个家庭并观测到它的 Y 值是 75 美元。在每次抽取（即重复抽样）中，X 值都固定在 80 美元。我们可以对表中的全部 X 值重复这一过程。事实上，表 2—4 和表 2—5 中的样本数据就是这样抽取得来的。

我们为什么假定 X 值是非随机的呢？既然在大多数社会科学中，数据通常都是对 Y 和 X 变量同时随机抽取的，如此看来，做相反的假定是很自然的——X 变量与 Y 变量一样也是随机的，但出于以下原因，我们最初还是假定 X 变量是非随机的。

第一，最初这样做的目的是为了简化分析，并引导读者逐步深入地理解回归分析的复杂内容。第二，在实验环境中，X 值固定不变的假定可能是不太现实的。比

① 经典的意义在于，它于 1821 年由高斯首创，并且此后它都被作为范式或标准，与那些不满足高斯假定的回归模型进行比较。

② 第 14 章将对参数为非线性的回归模型做简要讨论。

如，一位农民可能把他的土地分成几块，并对这些实验田施加不同数量的肥料，然后看肥料对作物收成的影响。类似地，商店可以用不同的折扣率来分析它对消费者的影响。有时为了某种特殊的目的，我们也能固定 X 值。假设我们正试图弄清楚工人平均月薪（Y）与各种受教育程度（X）之间的关系，如表 2—6 所给出的数据那样。在这种情形中，X 变量可视为固定的或非随机的。第三，就像我们在第 13 章中将看到的那样，在一定的条件下，即使 X 是随机的，基于固定回归元的线性回归的统计结果也是可靠的。条件之一是回归元 X 与误差项 u_i 是独立的。正如戴维森（James Davidson）所指出的那样："……这个模型（即随机回归元）'模拟'了固定回归元模型，而且最小二乘法在固定回归元模型中具有的许多统计性质仍然成立。"[①]

出于这些原因，我们首先尽可能详尽地讨论（固定回归元）CLRM。不过，我们在第 13 章将较详尽地讨论随机回归元情形，并指出我们偶尔需要考虑随机回归元模型的情形。顺便指出，如果 X 变量是随机的，由此得到的模型被称为**新古典线性回归模型**（neo-classical linear regression model，NLRM）[②]，与 X 被处理成固定回归元或非随机回归元的 CLRM 形成对照。为便于讨论，我们称前者为**随机回归元模型**（stochastic regressor model），称后者为**固定回归元模型**（fixed regressor model）。

假定 3：干扰项 u_i 的均值为零。对给定的 X_i 值，随机干扰项 u_i 的均值或期望值为零，记为：

$$E(u_i \mid X_i) = 0 \tag{3.2.1}$$

或者在 X 是非随机的情形下记为

$$E(u_i) = 0$$

假定 3 是说，以给定的 X_i 为条件，u_i 的均值为零，其几何意义可由图 3—3 描绘出来。图中显示了变量 X 的几个值以及与每个 X 值相对应的一个 Y 总体。如图所示，对应于给定的 X，每一个 Y 总体都是围绕其均值（由 PRF 上打上圆圈的点来表示）而分布的；一些 Y 值位于均值之上，另一些则位于均值之下。偏离均值上方和下方的距离不是别的，正是 u_i。方程（3.2.1）要求，对于任一给定的 X，这些离差的均值应等于零。

鉴于 2.4 节［参看方程（2.4.5）］的讨论，这一假定应是不难理解的。这一假定无非是说，凡是模型中没有明显包含并因而归于 u_i 之中的因素，对 Y 的均值都没有系统的影响；换句话说，正的 u_i 值就抵消了负的 u_i 值，以致它们对 Y 的平均影响为零。[③]

① James Davidson, *Econometric Theory*, Blackwell Publishers, U. K., 2000, p. 10.

② 这个术语来自于 Arthur S. Goldberger, *A Course in Econometrics*, Harvard University Press, Cambridge, MA, 1991, p. 264。

③ 关于假定 3 为什么是必要的，还有一个更为技术性的理由，可参看 E. Malinvaud, *Statistical Methods of Econometrics*, Rand McNally, Chicago, 1966, p. 75。还可参看本书习题 3.3。

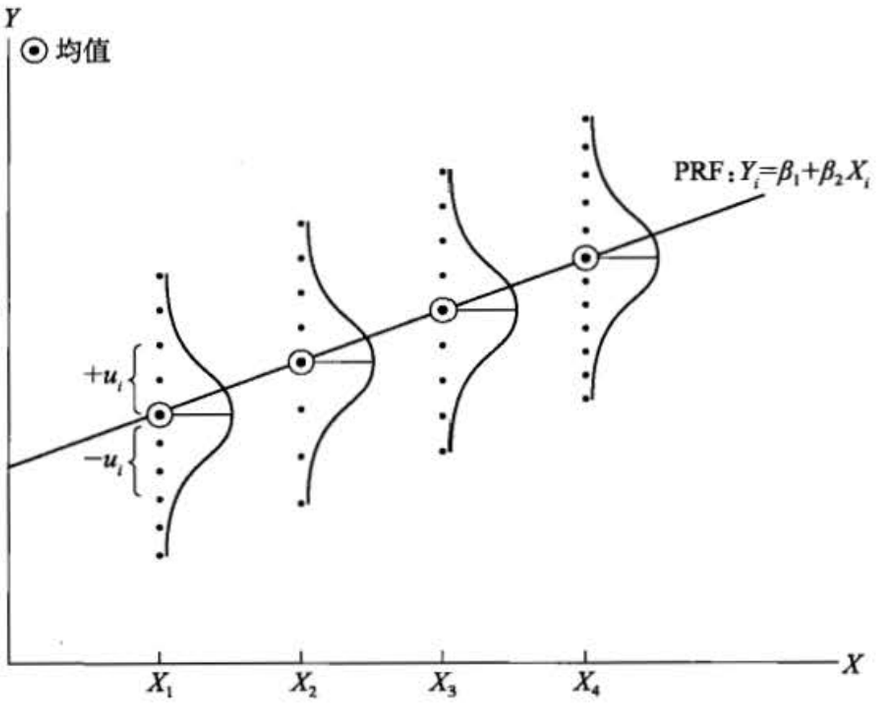


图 3—3　干扰项 u_i 的条件分布

顺便指出，假定 $E(u_i \mid X_i) = 0$ 意味着假定 $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$。(为什么?) 因此，这两个假定是等价的。

假定 3 意味着，经验分析所用的模型中不存在**设定偏误**（specification bias）或**设定误差**（specification error），指出这一点很重要。换句话说，回归模型设定是正确的。遗漏重要解释变量，包含一些不必要的变量，或对 Y 和 X 变量之间的关系选择了错误的函数形式，这些都是设定误差的例子。我们在第 13 章将详尽地讨论这个专题。

还要注意到，如果在给定一个随机变量的情况下另一个随机变量的条件均值为 0，那么这两个变量之间的协方差就是 0，因而这两个变量是无关的。因此，假定 3 意味着 X_i 和 u_i 是不相关的。[①]

假定干扰项 u 和解释变量 X 不相关的原因很简单，当我们像方程（2.4.2）那样表述 PRF 时，我们假定 X 和 u（代表所有被遗漏变量的影响）对 Y 具有独立（和加式）的影响。因而，如果 X 和 u 相关，就无法确定它们对 Y 的各自影响。因此，如果 X 和 u 正相关，那么 X 便随着 u 的增减而增减。类似地，如果 X 和 u 负相关，X 便随着 u 的增加而减小。在这类情形中，误差项很可能实际上包含了已经包含在模型中的一些变量。这就是假定 3 为什么是所选回归模型中不存在设定误差的另一种

① 不过，反之并不成立，因为相关系数只是度量了线性关联度。也就是说，即使 X_i 和 u_i 无关，给定 X_i，u_i 的条件均值也可能不等于 0。然而，如果 X_i 和 u_i 相关，$E(u_i \mid X_i)$就一定不等于 0，这就违背了假定 3。我们是从斯托克和沃森那里得到这种观点的，参见 James H. Stock and Mark W. Watson, *Introduction to Econometrics*, Addison-Wesley, Boston, 2003, pp. 104-105。

表述的原因。

假定 4：同方差性或 u_i 的方差相等。给定 X 值，对所有的观测，u_i 的方差都是相同的。也就是说 u_i 的条件方差是恒定的。用符号表示为：

$$\begin{aligned}\text{var}(u_i) &= E[u_i - E(u_i \mid X_i)]^2 \\ &= E(u_i^2 \mid X_i) \quad \text{由于假定 3} \\ &= E(u_i^2) \quad \text{如果 } X_i \text{ 是非随机的} \\ &= \sigma^2\end{aligned} \tag{3.2.2}$$

其中 var 表示方差。

方程（3.2.2）表明，对每个 X_i，u_i 的方差（即 u_i 的条件方差）都是某个等于 σ^2 的正常数。用专业术语说，方程（3.2.2）代表**同方差性**（homoscedasticity）或者说相同的散布或相等的方差。这一术语来自希腊词汇“skedanime”，指分散或散点。换言之，方程（3.2.2）意味着，对应于不同 X 值的 Y 总体均有同样的方差。简单地说，无论 X 的值如何变动，围绕着回归线即 X 与 Y 之间的平均关系线波动的方差是一样。图 3—4 用图形解释了这种情形。

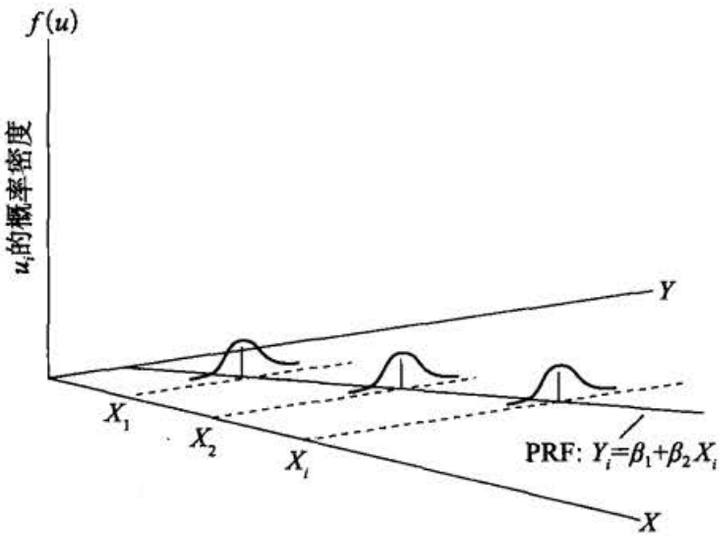


图 3—4　同方差性

与此相比，图 3—5 表示 Y 总体的条件方差随 X 而变化。这种情形的相应名称是**异方差性**（heteroscedasticity）或者说非相同的散布或非相等的方差。用符号表示，这时方程（3.2.2）可写为

$$\text{var}(u_i \mid X_i) = \sigma_i^2 \tag{3.2.3}$$

注意方程（3.2.3）中 σ^2 的下标，它表示 Y 总体的方差不再恒定不变。

为了清楚地区分这两种情形，令 Y 代表每周消费支出而 X 代表每周收入。图 3—4 和图 3—5 都表示随着收入的增加，平均消费支出也增加。但在图 3—4 中，消费支出的方差在所有的收入水平上都保持不变，而在图 3—5 中，这个方差随收入的

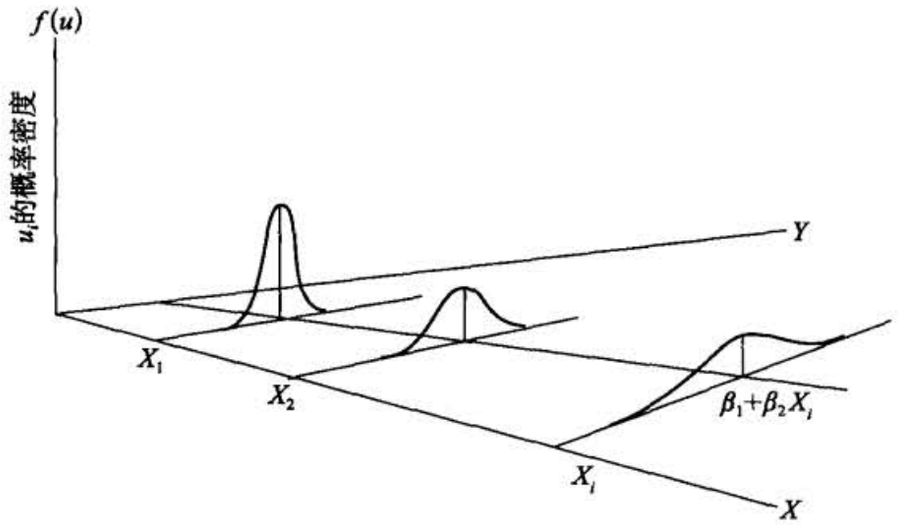


图 3—5　异方差性

增加而增加，换句话说，富有的家庭平均比贫穷的家庭消费更多，但前者的消费支出也有更大的差异。

为了了解其中的道理，请看图 3—5。如该图所示，$\text{var}(u \mid X_1) < \text{var}(u \mid X_2)$，…，$< \text{var}(u \mid X_i)$。因此，很有可能来自 $X=X_1$ 的 Y 总体与来自 $X=X_2$，$X=X_3$ 等的 Y 总体相比，其观测值更靠近 PRF。简言之，并不是对应于不同 X 的所有 Y 值都是同样可靠的。要根据 Y 值的分布与其均值（也就是 PRF 上的点）的分散程度来判断其可靠程度。如果这种想法符合实际，难道我们不认为那些离均值较近的 Y 总体的样本比远为分散的 Y 总体的样本更为可取吗？但这样做会限制 X 值的变动。

现在，借助假定 4，我们说，对应于不同 X 值的全部 Y 值都具有同样的重要性。在第 11 章中，我们将看到，如果情况不是这样，即存在异方差性，又会发生什么情况。

顺便提请注意，假定 4 意味着 Y_i 的条件方差也是同方差的，也就是说：

$$\text{var}(Y_i \mid X_i) = \sigma^2 \tag{3.2.4}$$

当然，Y 的无条件方差为 σ_Y^2。以后我们将会看到区分 Y 的条件方差和无条件方差的重要性。（关于条件方差和无条件方差的详细讨论，可参看附录 A。）

假定 5：各个干扰项之间无自相关。给定任意两个 X 值：X_i 和 X_j（$i \neq j$），u_i 和 u_j 之间的相关系数为零。简单地说，观测是相互独立的。用符号表示：

$$\text{cov}(u_i, u_j \mid X_i, X_j) = 0 \tag{3.2.5}$$

$$\text{cov}(u_i, u_j) = 0 \quad \text{若 } X \text{ 是非随机的}$$

其中 i 和 j 为两次不同的观测，而 cov 表示协方差。

字面上说，方程（3.2.5）设定干扰项 u_i 和 u_j 不相关。用专业术语来说，这是**无序列相关**（no serial correlation）或**无自相关**（no autocorrelation）假定。这就意

味着，给定 X_i，任意两个 Y 值与其均值的离差都不会表现出如图 3—6（a）和图 3—6（b）那样的模式。在图 3—6（a）中的 u 值是正相关的，即正的 u 伴随着正的 u，或负的 u 伴随着负的 u。图 3—6（b）的 u 值则是负相关的，即正（负）的 u 伴随着负（正）的 u。

如果这些干扰（离差）展现出某种系统性模式，如图 3—6（a）和图 3—6（b）那样，就表明它们有自相关或序列相关。假定 5 要求不要有这种相关。图 3—6（c）表示 u 之间没有这种系统性的模式，即零相关。

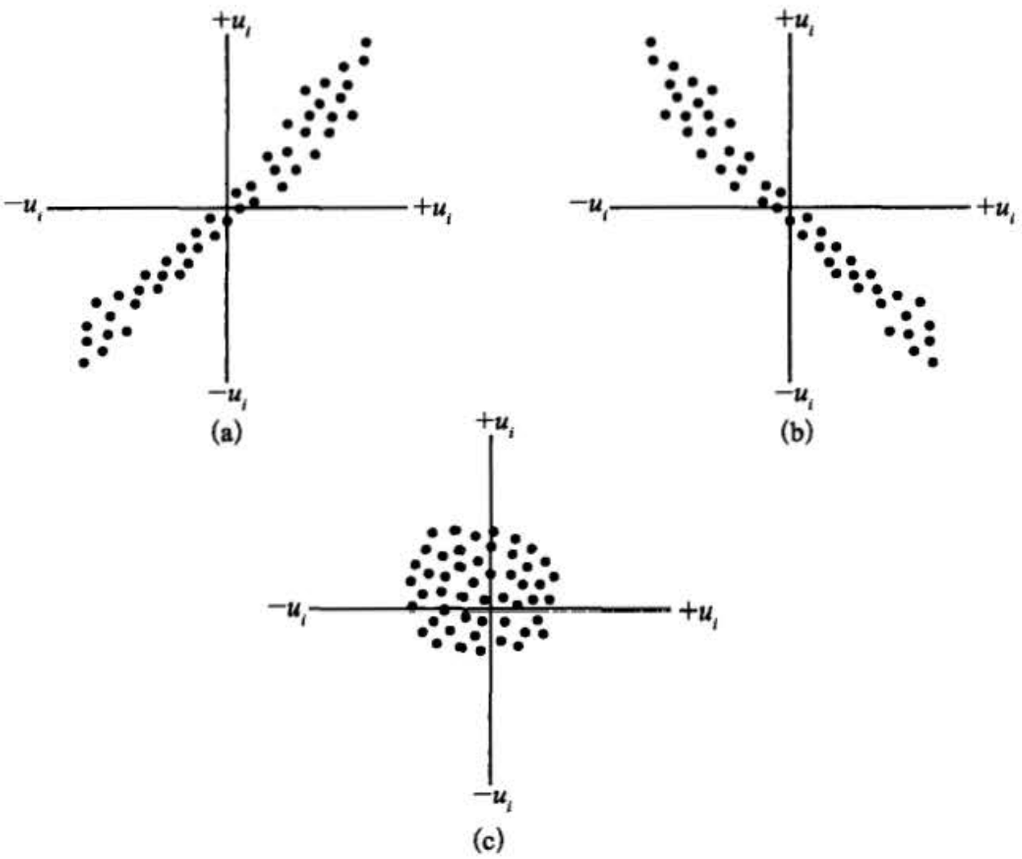


图 3—6　干扰之间的相关模式：
（a）正序列相关；（b）负序列相关；（c）零相关

在第 12 章里，我们将透彻地解释这一假定的全部含义。但直观上，我们可以对此假定作如下解释：设想在我们的 PRF（$Y_t=\beta_1+\beta_2 X_t+u_t$）中，$u_t$ 和 u_{t-1} 正相关，那么，Y_t 不仅依赖于 X_t，而且依赖于 u_{t-1}，因为 u_{t-1} 在一定程度上决定了 u_t。在讨论这一主题的现阶段，我们要利用假定 5，也就是说，我们将只考虑 X_t 对 Y_t 的系统性影响和是否有影响，而不去担心由于 u 之间可能的相关而造成的其他可能作用于 Y 的影响。但是，如同在第 12 章中指出的那样，我们能对这些干扰项之间的相关性作出分析并察看其后果。

但这里应该补充说明，此假定的合理性取决于分析中所用数据的类型，如果是横截面数据，又是取自一个相关总体的随机样本，那么这个假定通常是合理的。不过，如果是时间序列数据，那么由于 GDP 这类时间序列，其连续观测之间高度相

关，所以独立性假定难以成立，但在本书后面讲时间序列计量经济学时，我们会处理这种情形。

假定 6：观测次数 n 必须大于待估计的参数个数。另一种说法是，观测次数 n 必须大于解释变量的个数。

设立这一假定，并非是可有可无的。在表 3—1 假设的例子中，不妨设想我们只有对 Y 和 X 的第一对观测值（4 和 1），这样就无法由单一的观测去估计两个未知数 β_1 和 β_2。我们至少需要两对观测值来估计两个未知数。在下一章里，我们将会认识到这一假定的关键作用。

假定 7：X 变量的性质。在一个给定的样本中，X 值不可以全部相同。用专门术语来说，var(X) 是有限的正数。而且 X 变量的取值没有**异常**（outlier），即没有一个 X 值相对其余观测而言过大或过小。

这一关于 X 变量的假定也不是那么可有可无。且看方程（3.1.6）。如果全部 X 值都相同，则 $X_i=\overline{X}$。（为什么?）该方程的分母就变为零，从而无法估计 β_2，也就无法估计 β_1。凭直觉我们就能看出此假定为什么重要。看看我们第 2 章中的家庭消费支出例子，如果家庭收入很少变动，我们就不怎么能解释消费支出的变化。读者应该记住，要把回归分析作为一种研究工具来使用，Y 和 X 均有变化是根本重要的。简言之，变量必须在变！

X 变量的取值没有异常的要求是为了避免回归结果受到这种异常观测的支配。如果有少数 X 值是 X 均值的 20 倍，使用这些观测与否，所得到的回归线可能极为不同。这种异常常常是计算错误或混淆不同总体的样本所致。我们在第 13 章会进一步讨论这个问题。

现在我们对经典线性回归模型背后的假定已经讨论得足够充分。这里强调指出，所有这些假定都是针对 PRF 而非 SRF。有意思的是，我们前面讨论过的最小二乘法，其某些性质类似于我们对 PRF 所做的假定。比如，$\sum \hat{u}_i = 0$，因而 $\bar{\hat{u}} = 0$ 就类似于假定 $E(u_i \mid X_i) = 0$，同样，$\sum \hat{u}_i X_i = 0$ 的结论也类似于 $\text{cov}(u_i, X_i)=0$ 的假定。注意到最小二乘法试图“复制”我们对 PRF 所做的一些假定，令人欣喜不已。

当然，SRF 没有复制 CLRM 的所有假定，我们稍后将会看到，尽管 $\text{cov}(u_i, u_j)=0 (i\neq j)$，但 $\text{cov}(\hat{u}_i, \hat{u}_j)=0 (i\neq j)$并不成立。事实上，我们以后会说明，残差不仅自相关，而且还是异方差的（见第 12 章）。

□ 对这些假定的讨论

一个非常有价值的问题是：所有这些假定有多真实？这个“假定现实性”的问

题是科学哲学中的一个古老问题。有些人称假定是否真实无关紧要。重要的是基于这些假设的预测。以“假定无关紧要论”著称的有弗里德曼，对他来说，假定的非真实性有着积极的意义。“为了有意义，……一个假设在其假定中从描述上看必定是错误的。”①

我们可以不完全赞同这一观点，但回想一下在任何科学研究中我们做某些假定，都是因为它们便于逐步开展我们的主题研究，并不因为它们在准确地复制了现实的意义上必然是真实的。正如一位作者所说，“……如果简单性是好的理论所盼望的一个准则，那么所有好的理论都将毫无禁忌地理想化和简单化。”②

我们的计划是先透彻地研究 CLRM 的性质，然后在以后的篇章里深入分析如果 CLRM 的一个或多个假定不成立时会出现什么情况。在本章末我们在表 3—4 中给出一个指南，告诉读者到哪里去寻找当某一特定的假定不被满足时所发生的情况。

正如一位同僚向我指出的那样，当我们评阅他人的研究工作时，我们需要考虑研究者所作的假定是否切合他的数据和问题。经常出现的情况是，已发表的研究论文有赖于对问题和数据所做的隐含假定。这些假定可能不正确，而所做出的估计却以它们为依据。显然，有见识的读者看到这种问题，应对研究工作持有怀疑态度。所以表 3—4 中所列的假定，对于指导我们自己的研究和评价别人的研究都是一份检查目录。

有了这些背景，我们现在就可以开始研究 CLRM 了。具体而言，我们想比较 OLS 的**统计性质**（statistical properties）和先前讲的纯**数值性质**（numerical properties）。OLS 的统计性质以 CLRM 的假定为依据，并且在著名的**高斯-马尔可夫定理**（Gauss-Markov theorem）中被奉若神明。但在我们转到这个定理（为 OLS 广为应用提供理论解释的定理）之前，我们需要先考虑最小二乘估计的**精度**（precision）或**标准误**（standard error）。

3.3 最小二乘估计的精度或标准误

根据方程（3.1.6）和（3.1.7），显然最小二乘估计值是样本数据的函数。但因数据会随样本的变化而变化，所以估计值也必定随之而改变。因此需要有关于估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的“可靠性”或精度的某种度量。在统计学中，一个估计量的精度由它的标准误（se）来衡量。③ 附录 3A 第 3A.3 节中证明，在高斯的假定下，OLS 估计量

① Milton Friedman, *Essays in Positive Economics*, University of Chicago Press, Chicago, 1953, p. 14.

② Mark Blaug, *The Methodology of Economics: Or How Economists Explain*, 2d ed., Cambridge University Press, New York, 1992, p. 92.

③ 标准误无非是估计量的抽样分布的标准差，而一个估计量的抽样分布无非就是该估计量的概率或频率分布，也就是得自给定总体的容量相同的所有可能样本的估计值的一个分布。抽样分布的使用是为了能根据从一个或多个样本计算出来的估计值去推断总体的参数值。（详见附录 A。）

的标准误可求得如下：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{3.3.1}$$

$$\text{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \tag{3.3.2}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \tag{3.3.3}$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma \tag{3.3.4}$$

其中 var 是方差而 se 是标准误，并且 σ^2 为常数或假定 4 中 u_i 的共同方差。

除 σ^2 以外，上述方程中的一切变量均可从数据中估计出来。如附录 3A 第 3A.5 节所推导的那样，σ^2 由如下公式来估计：

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{3.3.5}$$

其中 $\hat{\sigma}^2$ 是真正的但未知的 σ^2 的 OLS 估计量，而表达式 $n-2$ 被称为**自由度**（number of degrees of freedom，df），$\sum \hat{u}_i^2$ 则表示**残差平方和或剩（残）余平方和**（residual sum of squares，RSS）。[①]

一旦获知 $\sum \hat{u}_i^2$，σ^2 就很容易计算。$\sum \hat{u}_i^2$ 既可由方程（3.1.2）算出，也可由下面的表达式（证明见 3.5 节）计算：

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \tag{3.3.6}$$

和方程（3.1.2）相比，方程（3.3.6）易于使用，后者并不要求计算每次观测的 $\hat{u}_i$，尽管这种计算本身有它的用处（在第 11 和 12 章中我们将看到这一点）。

因为

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

① 自由度一词指样本中观测值的总数（$=n$）减去相互独立的（线性）约束或限制的个数。换句话说，它是观测值的总个数中独立的观测值个数。例如，在计算出 RSS（3.1.2）之前必须先算出 $\hat{\beta}_1$ 和 $\hat{\beta}_2$。这两个估计值就是附加给 RSS 的两个约束。因此，在计算 RSS 时，就只有 $n-2$ 而不是 n 个独立观测值。按照这一逻辑，在三变量回归中 RSS 将有 $n-3$ 个自由度。至于 k 变量模型，它就有 $n-k$ 个自由度。**一般的规律是**：自由度个数 df=（n－待估参数的个数）。

所以计算 $\sum a_i^2$ 的另一表达式是：

$$\sum a_i^2 = \sum y_i^2 - \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2} \tag{3.3.7}$$

顺便提一下，$\hat{\sigma}^2$ 的正平方根

$$\hat{\sigma} = \sqrt{\frac{\sum a_i^2}{n-2}} \tag{3.3.8}$$

被称为**估计值的标准误**（standard error of estimate）或**回归标准误**（standard error of the regression，se）。它无非就是 Y 值围绕估计回归线波动的标准差，并常用于衡量所估计的回归线的**“拟合优度”**（goodness of fit），这是 3.5 节要讨论的专题。

我们在前面曾经指出，对于给定的 X_i，σ^2 同时代表 u_i 和 Y_i 的（条件）方差。因此，估计的标准误也可叫做 u_i 和 Y_i 的（条件）标准差。当然，和平常一样，σ_Y^2 和 σ_Y 分别代表 Y 的无条件方差和无条件标准差。

注意 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差（并因而它们的标准误）有如下特点：

1. $\hat{\beta}_2$ 的方差与 σ^2 成正比，而与 $\sum x_i^2$ 成反比。也就是说，对于给定的 σ^2，X 值的变化越大，$\hat{\beta}_2$ 的方差越小，从而 β_2 得以更大的精度加以估计。简言之，给定 σ^2，X 值有大的变化时比没有大的变化时，β_2 的测算更为准确。而且，给定 $\sum x_i^2$，方差 σ^2 越大，$\hat{\beta}_2$ 的方差也越大。注意，随着样本含量 n 的增加，总和 $\sum x_i^2$ 中的项数将增加。β_2 的估计精度可随 n 的增加而增加。（为什么？）

2. $\hat{\beta}_1$ 的方差与 σ^2 和 $\sum X_i^2$ 成正比，而与 $\sum x_i^2$ 和样本容量 n 成反比。

3. 由于 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是估计量，所以它们不仅随着样本的变化而变化，而且对于一个给定样本，它们还可能是互相依赖的。这种依赖性将由它们之间的协方差来衡量。在附录 3A 第 3A.4 节中，我们证明：

$$\begin{aligned}\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X}\operatorname{var}(\hat{\beta}_2) \\ &= -\bar{X}\left[\frac{\sigma^2}{\sum x_i^2}\right]\end{aligned} \tag{3.3.9}$$

既然和任何变量的方差一样，$\operatorname{var}(\hat{\beta}_2)$ 总是正的，所以 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 之间的协方差是正是负与 $\bar{X}$ 的符号有关。如果 $\bar{X}$ 是正的，那么从公式可以看出，协方差将是负的。这时，如果斜率系数 β_2 被过高估计（即斜率被估计得太陡），则截距系数 β_1 将被过低估计（即截距被估计得太小）。以后（特别是在讨论多重共线性的一章即第 10 章里）我们将看到，研究所估计的回归系数之间的协方差是有用的。

回归系数估计量（$\hat{\beta}_1$ 和 $\hat{\beta}_2$）的方差又怎样能够用来判断这些估计的可靠性呢？

这是一个统计推断问题，将在第 4 章和第 5 章里继续讨论。

3.4 最小二乘估计量的性质：高斯-马尔可夫定理[①]

前面曾指出，在给定经典线性回归模型的假定条件下，最小二乘估计具有一些理想的或最优的性质。这些性质包含在著名的高斯-马尔可夫定理之中。为阐明此定理，需要考虑一个估计量的**最优线性无偏性质**（best linear unbiasedness property，BLUE）。[②] 如在附录 A 中所解释的那样，一个估计量，比方说，OLS 估计量 $\hat{\beta}_2$，要成为 β_2 的最优线性无偏估计量，就必须满足下列条件：

1. 它是**线性的**（linear），即它是诸如回归模型中的因变量 Y 这种随机变量的线性函数。

2. 它是**无偏的**（unbiased），即它的均值或期望值 $E(\hat{\beta}_2)$ 等于真值 β_2。

3. 它在所有这样的线性无偏估计量中有最小方差；有最小方差的无偏估计量叫做**有效估计量**（efficient estimator）。

在回归的背景中，可以证明，OLS 估计量是 BLUE。这就是著名的高斯-马尔可夫定理的精髓，可叙述为：

> **高斯-马尔可夫定理**
>
> 在给定经典线性回归模型的假定下，最小二乘估计量在所有线性无偏估计量中具有最小方差，也就是说，它们是最优线性无偏估计量。

在附录 3A 第 3A.6 节中勾勒出这一定理的证明。随着本书的进展，高斯-马尔可夫定理的全部含义渐见分明。这里只需指出，该定理在理论上和实践上都是重要的。[③]

所有这些含义，都可以通过图 3—7 加以解释。

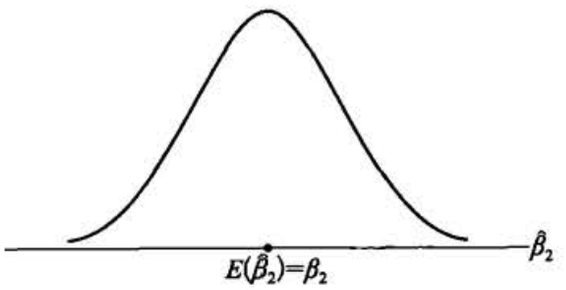
在图 3—7（a）中我们给出了 OLS 估计量 $\hat{\beta}_2$ 的**抽样分布**（sampling distribution），即在重复试验中 $\hat{\beta}_2$ 取值的分布（回忆表 3—1）。为方便起见，我们假定 $\hat{\beta}_2$ 是

---

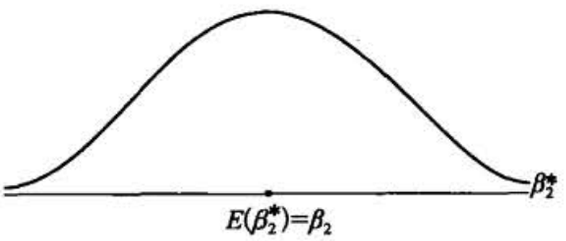
① 虽然以高斯-马尔可夫定理命名，但高斯的最小二乘法（1821）要早于马尔可夫的最小方差方法（1900）。

② 关于线性估计量的重要意义以及关于统计估计量的优良性质的一般讨论，读者可参考附录 A。

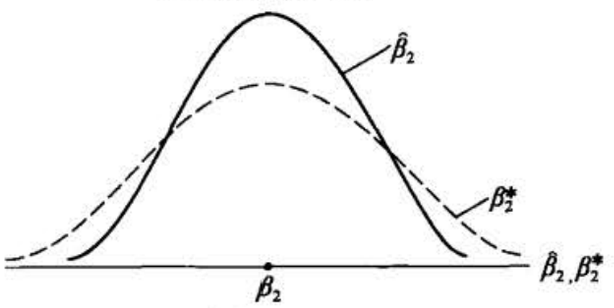
③ 例如，可以证明 β 的任意线性组合，如 $(\beta_1 - 2\beta_2)$ 可由 $(\hat{\beta}_1 - 2\hat{\beta}_2)$ 来估计，并且这一估计是 BLUE。详见 Henri Theil, *Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, N. J., 1978, pp. 401-402。注意对高斯-马尔可夫定理的一个技术性观点，即它只为 OLS 的有效性提供了一个充分（但不必要）条件。感谢西澳大利亚大学的迈克尔·麦卡利尔（Michael McAleer）让我注意到这个观点。



(a) $\hat{\beta}_2$的抽样分布



(b) β_2^*的抽样分布



(c) $\hat{\beta}_2$和β_2^*的抽样分布

图 3—7　OLS 估计量 $\hat{\beta}_2$ 和另一个估计量 β_2^* 的抽样分布

对称分布的（进一步的讨论见第 4 章）。如图所示，$\hat{\beta}_2$ 的均值 $E(\hat{\beta}_2)$ 等于 β_2。这时我们说 $\hat{\beta}_2$ 是 β_2 的一个无偏估计量。在图 3—7（b）中我们展示了使用另一种方法（不是 OLS 法）得到 β_2 的一个估计量 β_2^* 的抽样分布。为方便起见，仍假定 β_2^* 和 $\hat{\beta}_2$ 一样是无偏的，即其均值或期望值等于 β_2。再假定 $\hat{\beta}_2$ 和 β_2^* 都是线性估计量，即它们都是 Y 的线性函数。试问你会选取哪一个估计量，$\hat{\beta}_2$ 还是 β_2^*？

要回答这一问题，把两个图形重叠起来，如图 3—7（c）那样。显然，尽管 $\hat{\beta}_2$ 和 β_2^* 两者都是无偏的，但 β_2^* 的分布比 $\hat{\beta}_2$ 的分布围绕均值扩散得更广。换句话说，β_2^* 的方差比 $\hat{\beta}_2$ 的方差要大。既然两个估计量都是线性和无偏的，人们就会选择有较小方差的估计量，因为它比另一个估计量更可能接近 β_2。简单地说，人们会选择最优线性无偏估计量。

关于高斯-马尔可夫定理，值得一提的是，它并没有假设随机干扰项 u_i 的概率分布，对 Y_i 也一样（我们在下一章中将考虑这一点）。只要 CLRM 假设满足，定理结论就成立。这样，我们就无需寻找其他的线性无偏估计量了，因为我们将不会找到比 OLS 估计量的方差更小的估计量。当然，如果一个或多个假设条件不满足，定理

结论则不成立。例如，如果考虑关于参数非线性的回归模型（我们将在第 14 章中讨论），我们能够得到比 OLS 估计量更好的估计量。同样，如同我们将在异方差性一章中所讲的那样，如果同方差性的假设不满足，此时 OLS 估计量仍然是无偏的和一致的；但是，即使在线性估计量中，OLS 估计量也不是最小方差的。

刚才讨论的统计性质被称为**有限样本性质**（finite sample properties）：这些性质的成立与估计量所依据的样本容量无关。以后我们还有机会考虑**渐近性质**（asymptotic properties），即仅当样本非常大（从技术上讲，无穷大）时才会成立的性质。附录 A 对估计量的有限样本性质和大样本的渐近性质有一个一般性的讨论。

3.5 判定系数 r^2："拟合优度"的一个度量

直到现在为止，我们考虑的是估计回归系数的问题、它们的标准误以及它们的一些性质。现在我们来考虑对一组数据所拟合的回归线的**拟合优度**。也就是说，我们要说出这条样本回归线对数据拟合得有多好。由图 3—1 看出，如果全部观测点都落在样本回归线上，我们就得到一个"完美"的拟合。但是这种情形很少发生。一般的情形是，总有一些正的 a_i 和一些负的 a_i。我们所能希望的仅是这些围绕着回归线的残差尽可能小。**判定系数** r^2（coefficient of determination，双变量情形）或 R^2（多变量情形）就是告诉人们这条样本回归线对数据拟合效果的一个总度量。

在我们说明怎样计算 r^2 之前，先通过图形对 r^2 作一个直观的解释。这一图形称为**维恩图**（Venn diagram）或**巴伦坦图**（Ballentine），如图 3—8 所示。[①]

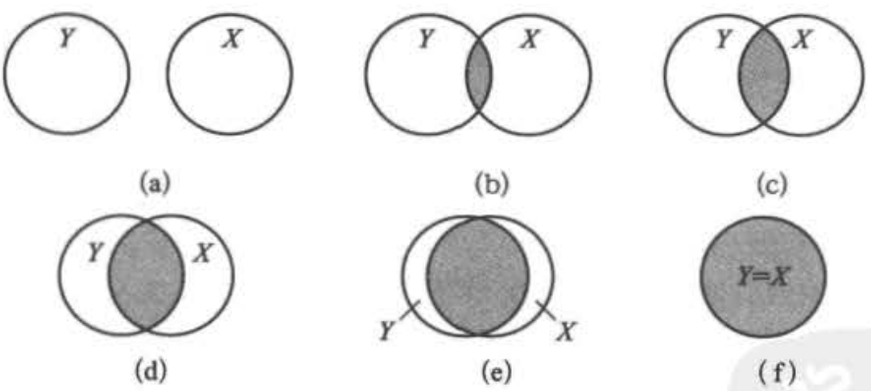


图 3—8　通过巴伦坦图看 r^2：(a) $r^2=0$；(f) $r^2=1$

在此图中，圆圈 Y 代表因变量 Y 的变异，圆圈 X 代表解释变量 X 的变异。[②] 两

① 参阅 Peter Kennedy，"Ballentine：A Graphical Aid for Econometrics，" *Australian Economics Papers*，vol. 20，1981，pp. 414-416。名字巴伦坦来自著名的 Ballantine 啤酒的圆圈徽标。

② 变异一词和方差有所不同。变异指一个变量对其均值的离差平方和，而方差指此平方和除以适当的自由度。简言之，方差＝变异/自由度。

圆的重叠部分代表 Y 的变异可由 X 的变异（比如说，通过 OLS 回归）解释的程度。重叠程度越大，Y 的变异被 X 的变异解释得越多。r^2 无非是这一重叠部分的一个数值度量。在此图中，从左到右，重叠部分渐增，即 Y 的变异被 X 的变异解释的部分依次变大，也就是说，越来越多的 Y 的变异被 X 的变异所解释，简单地说，r^2 在增加。在无重叠时，r^2 显然为零，但若全部重叠则 r^2 为 1，此时 Y 的变异百分之百地由 X 的变异所解释。下面我们很快看到 r^2 落在 0 和 1 之间。

计算 r^2 的步骤如下：回顾

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

或利用方程（3.1.13）和（3.1.14）写成离差形式：

$$y_i = \hat{y}_i + \hat{u}_i \tag{3.5.1}$$

两边平方并对样本求和，因为 $\sum \hat{y}_i \hat{u}_i = 0$（为什么?）且 $\hat{y}_i = \hat{\beta}_2 x_i$，便得到：

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2 \end{aligned} \tag{3.5.2}$$

出现在方程（3.5.2）中的各项平方和可描述为 $\sum y_i^2 = \sum (Y_i - \bar{Y})^2 =$ 实际 Y 值围绕其均值的总变异，称为**总平方和**（total sum of squares，TSS）。$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2 =$ 估计的 Y 值围绕其均值（$\bar{\hat{Y}} = \bar{Y}$）的变异，可适当地称为回归（即来自解释变量的）平方和，或者说，由回归解释的平方和，或简称**解释平方和**（explained sum of squares，ESS）。$\sum \hat{u}_i^2 =$ 估计的 Y 值围绕回归线的变异**未被解释**的（unexplained）或残差（或剩余）部分，或简称**残差平方和**（residual sum of squares，RSS）。因此，方程（3.5.2）就是：

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{3.5.3}$$

这说明 Y 的观测值围绕其均值的总变异可分解为两部分，一部分来自回归线，而另一部分则来自随机因素，因为 Y 的所有实际观测值并非都落在拟合线上。从几何意义上看，可画出图 3—9。

现用 TSS 除方程（3.5.3）的两边得：

$$\begin{aligned} 1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \\ &= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} \end{aligned} \tag{3.5.4}$$

我们定义 r^2 为：

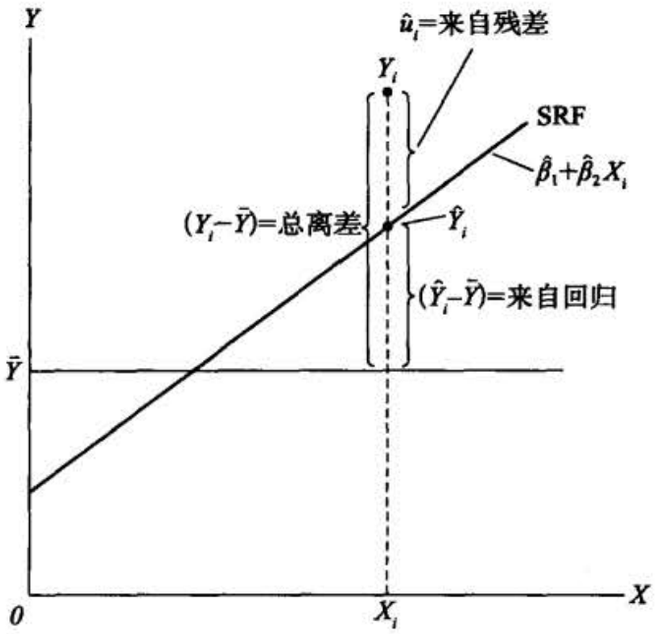


图 3—9　Y_i 的变异分解为两部分

$$r^2=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}=\frac{\text{ESS}}{\text{TSS}} \tag{3.5.5}$$

或者写成另一种形式：

$$r^2=1-\frac{\sum a_i^2}{\sum(Y_i-\bar{Y})^2}=1-\frac{\text{RSS}}{\text{TSS}} \tag{3.5.5a}$$

如上定义的数量 r^2 被称为（样本）**判定系数**，它是对回归线拟合优度的最为常用的度量，字面上讲，r^2 测度了在 Y 的总变异中由回归模型解释的那个部分所占的比例或百分比。

注意 r^2 的两个性质：

1. 它是一个非负量。（为什么？）

2. 它的界限为 $0\leqslant r^2\leqslant 1$。等于 1 的 r^2 意味着一个完美的拟合，即对每个 i 都有 $\hat{Y}_i=Y_i$。另一方面，等于 0 的 r^2 意味着回归值与回归元之间无任何关系（即 $\hat{\beta}_2=0$）。这时，如方程（3.1.9）所示，$\hat{Y}_i=\hat{\beta}_1=\bar{Y}$，也就是说，对任一 Y 值的最优预测值都是它的均值，从而回归线平行于 X 轴。

虽然 r^2 可按方程（3.5.5）所给的定义直接计算，但利用下面的公式能更快捷地求得：

$$
\begin{aligned}
r^2 &= \frac{\text{ESS}}{\text{TSS}} \\
&= \frac{\sum \hat{y}_i^2}{\sum y_i^2} \\
&= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} \\
&= \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right)
\end{aligned}
\tag{3.5.6}
$$

若将方程（3.5.6）中的分子与分母同时除以样本容量 n（或者对小样本用 $n-1$），就有：

$$
r^2 = \hat{\beta}_2^2 \left(\frac{S_x^2}{S_y^2} \right) \tag{3.5.7}
$$

其中 S_y^2 和 S_x^2 分别是 Y 和 X 的样本方差。

由于 $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$，方程（3.5.6）还可表达成：

$$
r^2 = \frac{\left(\sum x_i y_i \right)^2}{\sum x_i^2 \sum y_i^2} \tag{3.5.8}
$$

这是一个容易计算的表达式。

给定 r^2 的定义，可将上面讨论过的 ESS 和 RSS 表达如下：

$$
\begin{aligned}
\text{ESS} &= r^2 \cdot \text{TSS} \\
&= r^2 \sum y_i^2
\end{aligned}
\tag{3.5.9}
$$

$$
\begin{aligned}
\text{RSS} &= \text{TSS} - \text{ESS} \\
&= \text{TSS}(1 - \text{ESS}/\text{TSS}) \\
&= \sum y_i^2 \cdot (1 - r^2)
\end{aligned}
\tag{3.5.10}
$$

因此可写成：

$$
\begin{aligned}
&\text{TSS} = \text{ESS} + \text{RSS} \\
&\sum y_i^2 = r^2 \sum y_i^2 + (1 - r^2) \sum y_i^2
\end{aligned}
\tag{3.5.11}
$$

我们以后会看到，这是非常有用的表达式。

与 r^2 关系紧密但概念上与 r^2 很不相同的一个数量是**相关系数**（coefficient of correlation），如第 1 章所述，它测出两个变量之间的关联度。它既可由下式算出：

$$r = \pm\sqrt{r^2} \tag{3.5.12}$$

也可从它的定义算出：

$$r = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}} \tag{3.5.13}$$

该定义被称为**样本相关系数**（sample correlation coefficient）。[①]

r 具有如下性质（见图 3—10）：

1. 它可正可负，其符号与方程（3.5.13）的分子即两变量的协变异（covariation）的符号相同。

2. 它落在极限−1 和+1 之间；即 $-1 \leqslant r \leqslant 1$。

3. 它具有对称性；即 X 与 Y 之间的相关系数（r_{XY}）和 Y 与 X 之间的相关系数（r_{YX}）相同。

4. 它与原点和尺度都无关；即如果定义 $X_i^* = aX_i + c$ 和 $Y_i^* = bY_i + d$，其中 $a > 0$，$b > 0$，且 c 和 d 都是常数，则 X^* 和 Y^* 之间的 r 无异于原始变量 X 与 Y 之间的 r。

5. 如果 X 与 Y 统计上独立（参看附录 A 中的定义），则它们之间的相关系数为零；但反过来 $r=0$ 不等于说两个变量是独立的。换句话说，零相关并不一定意味着独立性。[见图 3—10（h）。]

6. 它仅是线性关联或线性相依的一个度量；它不能用于描述非线性关系。例如，在图 3—10（h）中 $Y = X^2$ 是一个准确的关系式，然而 r 为零。（为什么?）

7. 如第 1 章中所指出的那样，虽然它是两个变量之间线性关联的一个度量，却不一定有因果关系的含义。

在回归分析中，r^2 是一个比 r 更有意义的度量，因为前者告诉我们在因变量的变异中由解释变量解释的部分所占比例，因而对一个变量的变异在多大程度上决定另一个变量的变异提供了一个总的度量。而后者则没有这种价值。[②] 此外，如下面我们将会看到的，在一个多元回归模型中，对 r（$=R$）做的解释有多大价值，是个疑问。不管怎样，在第 7 章，我们还要进一步讨论 r^2。

顺便指出，前面定义的 r^2 还可作为实际 Y_i 与 Y_i 的估计值（即 $\hat{Y}_i$）之间的相关系数的平方来计算。也就是可利用方程（3.5.13）把它写为：

① 总体相关系数记为 ρ，其定义见附录 A。

② 回归模型所依据的基础理论表明，Y 与 X 之间的因果方向在简单回归中一般地说是从 X 到 Y 的。

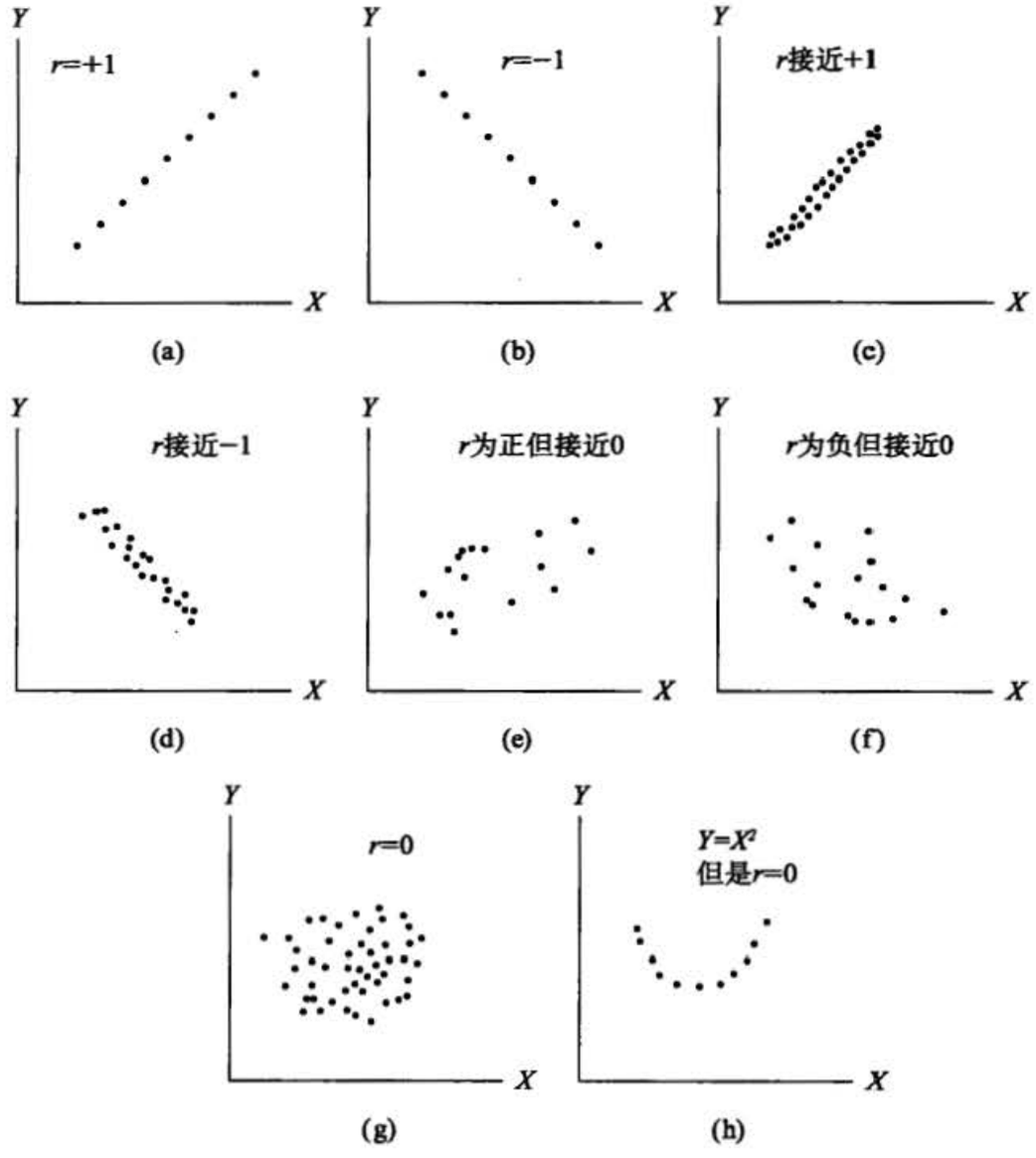


图 3—10　相关式样

资料来源：Adapted from Henri Theil, *Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, NJ, 1978, p. 86.

$$r^2 = \frac{\left[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}$$

也即：

$$r^2 = \frac{(\sum y_i \hat{y}_i)^2}{(\sum y_i^2)(\sum \hat{y}_i^2)} \tag{3.5.14}$$

其中 Y_i=Y 的实际值，$\hat{Y}_i$ = Y 的估计值，而 $\bar{Y} = \bar{\hat{Y}} = Y$ 的均值。有关证明见习题

3.15。表达式（3.5.14）解释了为什么把 r^2 描述成拟合优度的一个度量，这是因为它告诉我们 Y 的估计值和它的真实值靠得有多近。

3.6 一个数值例子

我们将通过考虑表 2—6 中给出的数据来解释到目前为止所介绍的计量经济理论，这些数据反映了平均小时工资（Y）与受教育程度（X）之间的关系。基本的劳动经济理论告诉我们，在各种变量中，受教育程度是工作的一个重要决定因素。

为了估计教育对工资的数量影响，我们在表 3—2 中给出了必要的原始数据。

表 3—2　　基于表 2—6 而计算的原始数据

观测	Y	X	x	y	x_i^2	$y_i x_i$
1	4.456 7	6	−6	−4.218	36	25.308
2	5.77	7	−5	−2.904 7	25	14.523 5
3	5.978 7	8	−4	−2.696	16	10.784
4	7.331 7	9	−3	−1.343	9	4.029
5	7.318 2	10	−2	−1.356 5	4	2.713
6	6.584 4	11	−1	−2.0903	1	2.090 3
7	7.818 2	12	0	−0.856 5	0	0
8	7.835 1	13	1	−0.839 6	1	−0.839 6
9	11.022 3	14	2	2.347 6	4	4.695 2
10	10.673 8	15	3	1.999 1	9	5.997 3
11	10.836 1	16	4	2.161 4	16	8.645 6
12	13.615	17	5	4.940 3	25	24.701 5
13	13.531 0	18	6	4.856 3	36	29.137 8
总计	112.771 2	156	0	0	182	131.785 6

观测	Y_i^2	X_i^2	$\hat{Y}_i$	$a_i = Y_i - \hat{Y}$	a_i^2
1	19.862 17	36	4.165 294	0.291 406	0.084 917
2	33.292 9	49	4.916 863	0.853 137	0.727 843
3	35.744 85	64	5.668 432	0.310 268	0.096 266
4	53.753 82	81	6.420 001	0.911 699	0.831 195
5	53.556 05	100	7.171 57	0.146 63	0.021 5

续前表

观测	Y_i^2	X_i^2	$\hat{Y}_i$	$a_i=Y_i-\hat{Y}$	a_i^2
6	43.354 32	121	7.923 139	−1.338 74	1.792 222
7	61.124 25	144	8.674 708	−0.856 51	0.733 606
8	61.388 79	169	9.426 277	−1.591 18	2.531 844
9	121.491 1	196	10.177 85	0.844 454	0.713 103
10	113.93	225	10.929 41	−0.255 62	0.065 339
11	117.421 1	256	11.680 98	−0.844 88	0.713 829
12	185.368 2	289	12.432 55	1.182 447	1.398 181
13	183.088	324	13.184 12	0.346 878	0.120 324
总计	1 083.376	2 054	112.771 2	≈0	9.830 17

注：$x_i=X_i-\overline{X}$；$y_i=Y_i-\overline{Y}$；

$$\hat{\beta}_2=\frac{\sum x_iy_i}{\sum x_i^2}=\frac{131.785\ 6}{182}=0.724\ 096\ 7;$$

$$\hat{\beta}_1=\overline{Y}-\hat{\beta}_2\overline{X}=8.674\ 708-0.724\ 096\ 7\times 12=-0.014\ 45;$$

$$\hat{\sigma}^2=\frac{\sum a_i^2}{n-2}=\frac{9.830\ 17}{11}=0.893\ 652;\ \hat{\sigma}=0.945\ 332;$$

$$\text{var}\ (\hat{\beta}_2)=\frac{\hat{\sigma}^2}{\sum x_i^2}=\frac{0.893\ 652}{182}=0.004\ 910;\ \text{se}(\hat{\beta}_2)=\sqrt{0.004\ 90}=0.070\ 072;$$

$$r^2=1-\frac{\sum a_i^2}{\sum (Y_i-\overline{Y})^2}=1-\frac{9.830\ 17}{105.118\ 8}=0.906\ 5;\ r=\sqrt{r^2}=0.952\ 1;$$

$$\text{var}\ (\hat{\beta}_1)=\frac{\sum X_i^2\hat{\sigma}^2}{n\sum x_i^2}=\frac{2\ 054\times 0.893\ 652}{13\times 182}=0.775\ 808;\ \text{se}(\hat{\beta}_1)=\sqrt{0.775\ 808}=0.880\ 800\ ^*。$$

根据表中数据，我们得到估计的回归线如下：

$$\hat{Y}_i=-0.014\ 4+0.724\ 0\ X_i \tag{3.6.1}$$

这条估计的回归线如图 3—11 所示。

正如我们所知道的那样，回归线上的每个点都给出在给定所选 X 值的情况下对 Y 均值的一个估计，也就是说，$\hat{Y}_i$ 是对 $E(Y\mid X_i)$ 的一个估计。$\hat{\beta}_2=0.724\ 0$ 度量了这条回归线的斜率，在 X 介于 6～18 年受教育程度的样本区间内，X 每增加 1 年，估计平均小时工资约提高 72 美分。也就是说，多接受 1 年教育，平均而言，每小时挣的工资将提高约 72 美分。

回归线的截距 $\hat{\beta}_1=-0.014\ 4$ 表示受教育程度为 0 时的平均工资水平。在目前的情况下，对截距做这种字面解释没有意义。工资怎么可能为负呢？正如我们在全书中将看到的那样，截距项经常没有可靠的实际意义。而且，受教育程度为 0 也不是我们的样本中观测的受教育水平。正如我们在第 5 章将看到的那样，截距的观测值

* 注意，最后两个公式的计算中原书有错误，此处已修改。——译者注

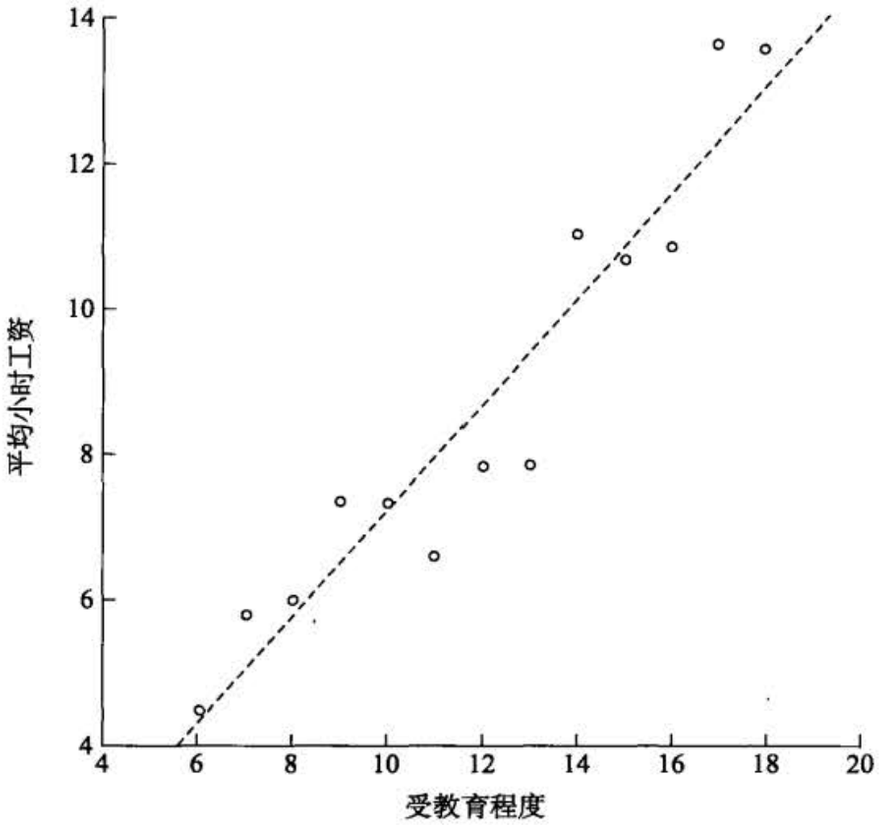


图 3—11　根据表 2—6 中工资—受教育程度数据而估计的回归线

在统计上与 0 没有差异。

约等于 0.90 的 r^2 值表明，受教育程度解释了平均小时工资变异的约 90%。考虑到 r^2 充其量也只能等于 1，所以我们的回归线对数据的拟合相当不错。相关系数 $r=0.952\ 1$ 表明，工资与受教育程度高度正相关。

在我们结束本例之前，注意我们的模型是极其简单的。劳动经济理论告诉我们，除了受教育程度之外，诸如性别、种族、居住地、是否加入工会以及语言也是决定小时工资的重要因素。在我们学习了第 7 章和第 8 章的多元回归之后，我们会考虑工资决定的一个更大的模型。

3.7　说明性例子

例 3.1　美国消费—收入关系：1960—2005 年

让我们回到引言的表 I—1 中给出的消费和收入数据。我们已经在图 1—3 及估计的回归线方程 (1.3.3) 中展示了这些数据。现在我们给出基本的 OLS 回归结果。(此结果得自统计软件 EViews 6。) 注：Y 表示个人消费支出 (PCE)，X 表示国内生产总值 (GDP)，均以 2000 年十亿美元计。

此例中的数据是时间序列数据。

$$\hat{Y}_t = -299.591\,3 + 0.721\,8X_t \tag{3.7.1}$$

$\text{var}(\hat{\beta}_1) = 827.419\,5$　　$\text{se}(\hat{\beta}_1) = 28.746\,9$

$\text{var}(\hat{\beta}_2) = 0.000\,019\,5$　　$\text{se}(\hat{\beta}_2) = 0.004\,423$

$r^2 = 0.998\,3$　　$\hat{\sigma}^2 = 73.566\,89$

方程（3.7.1）是总量（即对整个国家而言）凯恩斯消费函数。如该方程所示，**边际消费倾向**（MPC）约为 0.72，它表明如果收入增加 1 美元，平均个人消费支出（PCE）约上升 72 美分。根据凯恩斯理论，MPC 介于 0 与 1 之间。

本例中的截距值为负，也没有任何经济意义。从字面解释，它意味着，如果 GDP 的值为 0，则私人消费支出的平均水平约为 −2 990 亿美元。

r^2 的值为 0.998 3 意味着，PCE 变异的 99%都可由 GDP 的变异来解释。考虑到 r^2 最高等于 1，所以这个值相当高。如我们将在本书中所见到的那样，对时间序列数据的回归通常都能得到很高的 r^2 值。在自相关和有关时间序列的章节中，我们将看到这一现象背后的原因。

例 3.2　印度的食物支出

查阅习题 2.15 中表 2—8 所给出的数据。数据涉及印度 55 个农户的一个样本。本例中的回归子是食物支出，而回归元则是收入的代理变量——总支出，都以卢比为单位。因此本例中的数据为横截面数据。

基于给定数据，我们得到如下回归：

$$\widehat{\text{FoodExp}}_i = 94.208\,7 + 0.436\,8\text{TotalExp}_i \tag{3.7.2}$$

$\text{var}(\hat{\beta}_1) = 2\,560.940\,1$　　$\text{se}(\hat{\beta}_1) = 50.856\,3$

$\text{var}(\hat{\beta}_2) = 0.006\,1$　　$\text{se}(\hat{\beta}_2) = 0.078\,3$

$r^2 = 0.369\,8$　　$\hat{\sigma}^2 = 4\,469.691\,3$

我们从方程（3.7.2）可见，如果总支出增加 1 卢比，那么平均食物支出将增加 44 派沙（1 卢比＝100 派沙）。如果总支出为零，则平均的食物支出为 94 卢比。同样，对截距项的这种机械解释可能没有意义。但在本例中，人们可以认为，即使总支出为零（如因为失业），但人们仍可能通过借贷或动用储蓄将食物支出维持在某个最低水平。

约为 0.37 的 r^2 值表明，食物支出变动中的 37%由总支出来解释。看上去这是一个相当低的值，但如我们全书所见，在横截面数据中，通常获得低 r^2 值都可能是因为样本单位的分散性所致。我们将在异方差一章（第 11 章）中讨论这一专题。

例 3.3　手机和个人计算机需求与人均收入的关系

表 3—3 给出了 34 个样本国家的如下数据：每 100 个人中使用手机的人数，每 100 个人中拥有个人计算机的人数，以及经购买力平价调整后的人均收入（以美元计）。因而我们的数据是横截面数据。这些数据是 2003 年的数据，摘自《2006 年美国统计摘要》（*Statistical Abstract of the United States*，2006）。

表 3—3　2003 年样本国家每百人手机用户、每百人个人计算机用户与人均收入数据

	手机用户	个人计算机用户	人均收入（美元）
阿根廷	17.76	8.2	11 410
澳大利亚	71.95	60.18	28 780
比利时	79.28	31.81	28 920
巴西	26.36	7.48	7 510
保加利亚	46.64	5.19	75.4
加拿大	41.9	48.7	30 040
中国	21.48	2.76	4 980
哥伦比亚	14.13	4.93	6 410
捷克	96.46	17.74	15 600
厄瓜多尔	18.92	3.24	3 940
埃及	8.45	2.91	3 940
法国	69.59	34.71	27 640
德国	78.52	48.47	27 610
希腊	90.23	8.17	19 900
危地马拉	13.15	1.44	4 090
匈牙利	76.88	10.84	13 840
印度	2.47	0.72	2 880
印度尼西亚	8.74	1.19	3 210
意大利	101.76	23.07	26 830
日本	67.9	38.22	28 450
墨西哥	29.47	8.3	8 980
荷兰	76.76	46.66	28 560
巴基斯坦	1.75	0.42	2 040
波兰	45.09	14.2	11 210
俄罗斯	24.93	8.87	8 950
沙特阿拉伯	32.11	13.67	13 230
南非	36.36	7.26	10 130
西班牙	91.61	19.6	22 150
瑞典	98.05	62.13	26 710
瑞士	84.34	70.87	32 220
泰国	39.42	3.98	7 450
英国	91.17	40.57	27 690
美国	54.58	65.98	37 750
委内瑞拉	27.3	6.09	4 750

注：数据为每百人手机使用人数和每百人个人计算机使用人数。

资料来源：*Statistical Abstract of the United States*，2006，表 1364 为手机和计算机用户数据，表 1327 为经调整的人均收入购买力。

尽管手机和个人计算机在美国已经广泛使用，但在许多国家还做不到。为了看出人均收入是否影响手机和个人计算机使用数量，我们利用这 34 个样本国家将这些通信工具对人均收入回归。得到的结果如下：

手机需求。令 Y＝手机用户数，X＝经购买力平价调整后的人均收入，我们得到如下回归：

$$\hat{Y}_i = 14.477\ 3 + 0.002\ 2X_i \tag{3.7.3}$$

$$se(\hat{\beta}_1) = 6.152\ 3 \qquad se(\hat{\beta}_2) = 0.000\ 32$$

$$r^2 = 0.602\ 3$$

斜率系数表明，如果人均收入提高 1 000 美元，则每 100 个人中手机用户将平均增加 2.2 个。约等于 14.47 的截距值表明，就算人均收入为 0，每 100 个人中手机用户平均也能达到约 14 个。同样，这个截距没有多大意义，因为在我们的样本中没有一个国家的人均收入为 0。r^2 值比较高。但注意，我们的样本包含了收入水平千差万别的一系列国家。在如此分散的一个样本中，我们不能指望 r^2 值能高到哪里去。

在我们学习了第 5 章之后，我们将知道方程（3.7.3）中报告的估计标准误如何用于评价系数估计值的统计显著性。

个人计算机需求。尽管个人计算机价格近年来明显下降，但个人计算机的使用仍不是那么普遍。个人计算机需求的一个重要决定因素就是个人收入。另一个决定因素是价格，但对于我们的样本国家，我们没有可以比较的个人计算机价格数据。

令 Y 表示个人计算机用户数，X 表示人均收入，我们得到对个人计算机的如下“局部需求”（之所以说是局部需求，因为我们没有比较价格数据或其他可能影响个人计算机需求的变量数据）：

$$\hat{Y}_i = -6.583\ 3 + 0.001\ 8X_i \tag{3.7.4}$$

$$se(\hat{\beta}_1) = 2.743\ 7 \qquad se(\hat{\beta}_2) = 0.000\ 14$$

$$r^2 = 0.829\ 0$$

正如这些结果所示，人均收入与个人计算机需求有正相关关系。在我们学习了第 5 章之后将看到，从统计上看，人均收入是个人计算机需求的一个重要决定因素。在目前的情况下，截距为负没有实际含义。尽管我们的样本很分散，但估计的 r^2 值还是相当高。对斜率系数的解释是，如果人均收入提高 1 000 美元，则平均而言，每 100 个人中个人计算机需求会增加约 2 个单位。

尽管个人计算机的使用迅速普及，但仍有许多国家仍使用大型计算机。因此，在这样一些国家，计算机的总使用量可能远大于个人计算机销售额所显示的数量。

3.8 关于蒙特卡罗实验的一个注记

本章说过，在 CLRM 的假定下，最小二乘估计量有一些优良的统计性质，可概括为 BLUE 性质。在本章的附录中，我们更规范地给出了这一性质的证明。但实际上我们怎样才能知道这一 BLUE 性质是否成立？比如，怎样知道 OLS 估计量是否无偏？所谓的**蒙特卡罗**（Monte Carlo）实验能提供这一答案，它本质上就是一种计算机模拟或抽样实验。

为了介绍其基本思想，且考虑我们的双变量 PRF：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.8.1}$$

蒙特卡罗实验的程序如下：

1. 假定参数有如下的真实值：$\beta_1 = 20$ 和 $\beta_2 = 0.6$。

2. 选定样本容量，比方说 $n=25$。

3. 每次观测固定一个 X 值，因此共有 25 个 X 值。

4. 从一张随机数表中选出 25 个数值，且称它们为 u_i（如今的统计软件多数都包含随机数发生器）。[①]

5. 由于 β_1，β_2，X_i 和 u_i 已知，便可利用方程（3.8.1）得到 25 个 Y_i 值。

6. 现在利用如此生成的 25 个 Y_i 值，对在第 3 步中所选的 25 个 X 值做回归，求得最小二乘估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$。

7. 假使重复这一实验 99 次，每次都用相同的 β_1，β_2 和 X 值。当然，u_i 在每次实验中有所变化，因而在总共的 100 次实验中，就产生 β_1 和 β_2 的各 100 个值。（实际上，人们做过许多这样的实验，有时重复 1 000 次～2 000 次。）

8. 取这 100 个估计值的均值，并称它们为 $\bar{\beta}_1$ 和 $\bar{\beta}_2$ 。

9. 如果这些均值和在第 1 步中所假定的 β_1 和 β_2 的真实值（20，0.6）相差不多，那么蒙特卡罗法就“证实”了最小二乘估计量确实是无偏的。回想在 CLRM 假定下，$E(\hat{\beta}_1)=\beta_1$ 和 $E(\hat{\beta}_2)=\beta_2$ 。

以上步骤刻画了蒙特卡罗试验的一般性质。这种实验常被用来研究各种估计总体参数的方法的统计性质。它们在研究小样本或有限样本的估计量的性态时尤其有用。这些实验对于彻底掌握**重复抽样**（repeated sampling）的概念也是绝好的手段。重复抽样的概念，如在第 5 章中我们将会看到的那样，是大部分经典统计推断的基础。我们将通过课堂布置的习题提供蒙特卡罗实验的若干例子。（见习题 3.27。）

要点与结论

1. 回归分析的基本构架是 **CLRM**。

2. CLRM 是以一组假定为基础的。

3. 基于这些假定，最小二乘估计量便具有一些可概括为高斯-马尔可夫定理的性质。该定理认为，在所有线性无偏估计量中，最小二乘估计量有最小的方差。简单地说，这些估计量是 BLUE。

4. OLS 估计量的精度由其**标准误**来衡量，在第 4 章和第 5 章中，我们将看到这些标准误如何用来推断总体参数——系数 β。

5. 回归模型的总拟合优度由**判定系数** r^2 来衡量。它表明在因变量或回归子的变异中，由解释变量或回归元解释的部分所占的比例。r^2 在 0 与 1 之间；它越靠近 1，回归拟合得越好。

6. 与判定系数相关的一个概念是**相关系数** r。它是两个变量之间线性关联的一个度量，并位于 -1 与 $+1$ 之间。

7. CLRM 仅是一个理论上的构想或抽象，因为它是以一组严谨的或者说“不真实”的假定作

① 实践中，假定 u_i 服从给定参数（如均值和方差）的某一概率分布，如正态分布。一旦设定了参数，就容易用统计包生成 u_i。

为依据的。但是这种抽象不管在哪个研究领域中，在其初始阶段常常都是必需的。一旦掌握了CLRM，就能研究如果某一或某些假定不成立，将会出现什么情况。本书的第1篇专门讨论CLRM。其他几篇则考虑CLRM的引申。表3—4给出了一幅本书的学习图。

表3—4　　违反CLRM假定的种种后果

假定编号	违反类型	在何处研究
1	对参数非线性	第14章
2	随机回归元（一个或多个）	第13章
3	u_i 有非零均值	第2篇导论
4	异方差性	第11章
5	干扰项自相关	第12章
6	样本观测次数小于回归元个数	第10章
7	回归元缺乏变异	第10章
8	多重共线性*	第10章
9	设定偏误*	第13、14章
10**	干扰项的非正态性	第13章

注：* 这些假定将在第7章讨论多元回归模型时介绍。

** 干扰项 u_i 为正态分布的假定不属于CLRM。进一步的讨论见第4章。

习　　题

问答题

3.1　给定下表第（1）列中的假定，证明第（2）列中的假定与之等价。

关于经典模型的假定

(1)	(2)
$E(u_i \mid X_i) = 0$	$E(Y_i \mid X_i) = \beta_1 + \beta_2 X$
$\text{cov}(u_i, u_j) = 0 \quad i \neq j$	$\text{cov}(Y_i, Y_j) = 0 \quad i \neq j$
$\text{var}(u_i \mid X_i) = \sigma^2$	$\text{var}(Y_i \mid X_i) = \sigma^2$

3.2　证明表3—1的第1个实验所用的估计值 $\hat{\beta}_1 = 1.572$ 和 $\hat{\beta}_2 = 1.357$ 事实上是OLS估计量。

3.3　按照马林伍得（Malinvaud）（参看3.2节假设3中的注释）的意见，假设 $E(u_i \mid X_i) = 0$ 是相当重要的。为了看到这一点，考虑PRF：$Y = \beta_1 + \beta_2 X_i + u_i$。现区分两种情形：(i) $\beta_1 = 0, \beta_2 = 1$ 及 $E(u_i) = 0$；和 (ii) $\beta_1 = 1, \beta_2 = 0$ 及 $E(u_i) = (X_i - 1)$。然后在这两种情形中以 X 为条件求PRF的数学期望，并看你是否同意马林伍得的观点，即假定 $E(u_i \mid X_i) = 0$ 非常重要。

3.4　考虑样本回归：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + a_i$$

在如下约束条件下：(i) $\sum a_i = 0$ 和 (ii) $\sum a_i X_i = 0$，求估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$，并证明它们无异于方程（3.1.6）和方程（3.1.7）中所给出的最小二乘估计量。这种求估计量的方法叫做**类比原理**（analogy principle）。试述施加约束条件（i）和（ii）的直觉理由。（提示：回顾关于 u_i 的CLRM假

定。）顺便指出，估计未知参数的类比原理又叫做**矩法**（method of moments），即用样本矩（如样本均值）去估计总体矩（如总体均值）。如在附录 A 中所指出的那样，矩是概率分布的一个摘要统计量，比如期望和方差。

3.5 证明由方程（3.5.5）定义的 r^2 落在 0 与 1 之间。你可以利用柯西-施瓦茨（Cauchy-Schwarz）不等式，即对任意随机变量 X 和 Y，下列关系式总是成立的：

$$[E(XY)]^2 \leqslant E(X^2)E(Y^2)$$

3.6 令 $\hat{\beta}_{YX}$ 和 $\hat{\beta}_{XY}$ 分别为 Y 对 X 回归和 X 对 Y 回归中的斜率。证明：

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r^2$$

其中 r 为 X 与 Y 之间的相关系数。

3.7 假设在习题 3.6 中 $\hat{\beta}_{YX}\hat{\beta}_{XY}=1$。那么求 Y 对 X 的回归和求 X 对 Y 的回归有什么差别？请解释。

3.8 斯皮尔曼（Spearman）等级相关系数 r_s 的定义如下：

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

其中 d=编排给同一单元或现象的等级差，n=参与等级编排的单元或现象个数。试从方程（3.5.13）中定义的 r 推出 r_s。提示：从 1 到 n 将 X 和 Y 编排等级。注意 X 和 Y 的等级和为 $n(n+1)/2$，因而它们的均值都是 $(n+1)/2$。

3.9 考虑如下双变量 PRF 表达式：

模型 I：$Y_i = \beta_1 + \beta_2 X_i + u_i$

模型 II：$Y_i = \alpha_1 + \alpha_2(X_i - \bar{X}) + u_i$

a. 求 β_1 和 α_1 的估计量。它们是否相同？它们的方差是否相同？

b. 求 β_2 和 α_2 的估计量，它们是否相同？它们的方差是否相同？

c. 如果模型 II 比模型 I 好，好在哪里？

3.10 假设你做了如下回归：

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

其中 y_i 和 x_i 是它们与其各自均值的离差，问 $\hat{\beta}_1$ 将取何值？为什么？$\hat{\beta}_2$ 会不会和方程（3.1.6）的 $\hat{\beta}_2$ 一样？为什么？

3.11 令 r_1 为 n 对（X_i，Y_i）值的相关系数，而 r_2 为 n 对（aX_i+b，cY_i+d）值的相关系数，其中 a，b，c 和 d 为常数。证明 $r_1 = r_2$，从而证实相关系数对度量单位和原点的改变保持不变的性质。

提示：应用方程（3.5.13）中所给的 r 定义。

注：运算 aX_i，X_i+b 和 aX_i+b 分别叫做尺度变换、原点变换和尺度与原点同时变换。

3.12 如果 n 对（X_i，Y_i）值的相关系数是正的，试判断以下各个命题的对错：

a.（$-X_i$，$-Y_i$）之间的 r 可正可负。

b.（$-X_i$，Y_i）之间以及（X_i，$-Y_i$）之间的 r 可正可负。

c. 斜率系数 β_{yx} 和 β_{xy} 都是正的，其中 β_{yx} 为 Y 对 X 回归的斜率系数，而 β_{xy} 为 X 对 Y 回归的斜率系数。

3.13 如果 X_1，X_2 和 X_3 是有同样方差但互不相关的变量，试证明 X_1+X_2 和 X_2+X_3 之间的相关系数等于 1/2。为什么这个相关系数不是零？

3.14 假设在回归 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 中，我们将每个 X 值都乘以 2，这会不会改变 Y 的残差及拟合值？为什么？如果我们给每个 X 值都加上一个常数 2，又会怎样？

3.15 证明方程（3.5.14）事实上就是判定系数。

提示：应用方程（3.5.13）中所给 r 的定义，并回忆 $\sum y_i\hat{y}_i = \sum (\hat{y}_i + \hat{u}_i)\hat{y}_i = \sum \hat{y}_i^2$ 和方程（3.5.6）。

3.16 判断以下命题对错，并给出原因。

a. 由于两个变量 Y 和 X 之间的相关系数取值范围为 $[-1, 1]$，所以这意味着 cov(Y, X) 也在此范围内。

b. 如果两个变量之间的相关系数为零，那就意味着这两个变量之间不存在相关关系。

c. 如果你将 Y_i 对 $\hat{Y}_i$ 回归（即实际的 Y 对估计的 Y 回归），那么截距和斜率的值分别为 0 和 1。

3.17 **不含回归元的回归。**假如给你一个模型 $Y_i = \beta_1 + u_i$。利用 OLS 求出 β_1 的估计量。其方差和 RSS 是多少？估计的 β_1 有直觉上的意义吗？现在考虑双变量模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$。值得在此模型中增加 X_i 吗？否则，为什么要进行回归分析呢？

实证分析题

3.18 表 3—5 给出了 10 名学生在统计学期中和期末考试中的名次。计算斯皮尔曼等级相关系数并加以解释。

表 3—5

	学生									
等级	A	B	C	D	E	F	G	H	I	J
期中	1	3	7	10	9	5	4	8	2	6
期末	3	2	8	7	9	6	5	10	1	4

3.19 **名义汇率与相对价格之间的关系。**根据 1985—2005 年的年度观测，可得到如下回归结果：

$$\hat{Y}_t = -0.912 + 2.250X_t \qquad r^2 = 0.440$$

$$se = \qquad 0.096$$

其中 Y 表示加元与美元的汇率（CD/$），$X$ 表示美国 CPI 与加拿大 CPI 之比；即 X 代表了这两个国家的相对价格。

a. 解释这个回归。你如何解释 r^2？

b. X_t 的系数为正有经济意义吗？其背后的经济理论是什么？

c. 假如我们将 X 定义为加拿大 CPI 与美国 CPI 之比。X 的符号会改变吗？为什么？

3.20 表 3—6 给出了美国在 1960—2005 年间商业和非农商业部门的小时产出指数（X）和真实小时工资（Y）的数据。基年（1992）指数为 100，且指数经过了季节性调整。

表 3—6　　1960—2005 年商业部门的生产力及相关数据

（1992 年指数＝100；季度数据经过了季节性调整）

年份	所有人的每小时产出[1]		真实小时工资[2,3]	
	商业部门	非农商业部门	商业部门	非农商业部门
1960	48.9	51.9	60.8	63.3
1961	50.6	53.5	62.5	64.8
1962	52.9	55.9	64.6	66.7
1963	55.0	57.8	66.1	68.1
1964	56.8	59.6	67.7	69.3
1965	58.8	61.4	69.1	70.5

续前表

年份	所有人的每小时产出[1]		真实小时工资[2,3]	
	商业部门	非农商业部门	商业部门	非农商业部门
1966	61.2	63.6	71.7	72.6
1967	62.5	64.7	73.5	74.5
1968	64.7	66.9	76.2	77.1
1969	65.0	67.0	77.3	78.1
1970	66.3	68.0	78.8	79.2
1971	69.0	70.7	80.2	80.7
1972	71.2	73.1	82.6	83.2
1973	73.4	75.3	84.3	84.7
1974	72.3	74.2	83.3	83.8
1975	74.8	76.2	84.1	84.5
1976	77.1	78.7	86.4	86.6
1977	78.5	80.0	87.6	88.0
1978	79.3	81.0	89.1	89.6
1979	79.3	80.7	89.3	89.7
1980	79.2	80.6	89.1	89.6
1981	80.8	81.7	89.3	89.8
1982	80.1	80.8	90.4	90.8
1983	83.0	84.5	90.3	90.9
1984	85.2	86.1	90.7	91.1
1985	87.1	87.5	92.0	92.2
1986	89.7	90.2	94.9	95.2
1987	90.1	90.6	95.2	95.5
1988	91.5	92.1	96.5	96.7
1989	92.4	92.8	95.0	95.1
1990	94.4	94.5	96.2	96.1
1991	95.5	96.1	97.4	97.4
1992	100.0	100.0	100.0	100.0
1993	100.4	100.4	99.7	99.5
1994	101.3	101.5	99.0	99.1
1995	101.5	102.0	98.7	98.8
1996	104.5	104.7	99.4	99.4
1997	106.5	106.4	100.5	100.3
1998	109.5	109.4	105.2	104.9
1999	112.8	112.5	108.0	107.5
2000	116.1	115.7	112.0	111.5
2001	119.1	118.6	113.5	112.8
2002	124.0	123.5	115.7	115.1
2003	128.7	128.0	117.7	117.1
2004	132.7	131.8	119.0	118.2
2005	135.7	134.9	120.2	119.3

注：[1]产出指该部门真实 GDP。

[2]雇员的工资和薪水加上雇主对社会保障和私人福利方案的支付。

[3]对于城市消费者，近几个季度的小时工资除以消费者价格指数。

资料来源：*Economic Report of the President*，2007，Table 49.

a. 分别对两个部门将 Y 对 X 描点。

b. 这两个变量之间关系的背后有什么经济理论？散点图支持该理论吗？

c. 估计 Y 对 X 的 OLS 回归，在学完第 5 章后，再回头看一下你的结果。

3.21 根据一个包含 10 次观测的样本，得到如下结果：

$$\sum Y_i = 1\,110 \quad \sum X_i = 1\,700 \quad \sum X_i Y_i = 205\,500 \quad \sum X_i^2 = 322\,000 \quad \sum Y_i^2 = 132\,100$$

并且相关系数 r=0.975 8。但在重新核对这些计算时，发现有两组观测的记录是

Y	X
90	120
140	220

而不是

Y	X
80	110
150	210

问这一错误对 r 有何影响？求正确的 r。

3.22 表 3—7 给出 1974—2006 年间美国的黄金价格、消费者价格指数（CPI）和纽约证券交易所指数（NYSE Index）数据。NYSE 指数包括在 NYSE 上市的 1 500 多种股票中的大多数。

表 3—7　　1974—2006 年间美国的黄金价格、NYSE 指数和消费者价格指数数据

年份	黄金价格	NYSE 指数	CPI
1974	159.260 0	463.540 0	49.300 00
1975	161.020 0	483.550 0	53.800 00
1976	124.840 0	575.850 0	56.900 00
1977	157.710 0	567.660 0	60.600 00
1978	193.220 0	567.810 0	65.200 00
1979	306.680 0	616.680 0	72.600 00
1980	612.560 0	720.150 0	82.400 00
1981	460.030 0	782.620 0	90.900 00
1982	375.670 0	728.840 0	96.500 00
1983	424.350 0	979.520 0	99.600 00
1984	360.480 0	977.330 0	103.900 0
1985	317.260 0	1 142.970	107.600 0
1986	367.660 0	1 438.020	109.600 0
1987	446.460 0	1 709.790	113.600 0
1988	436.940 0	1 585.140	118.300 0
1989	381.440 0	1 903.360	124.000 0
1990	383.510 0	1 939.470	130.700 0
1991	362.110 0	2 181.720	136.200 0
1992	343.820 0	2 421.510	140.300 0
1993	359.770 0	2 638.960	144.500 0
1994	384.000 0	2 687.020	148.200 0
1995	384.170 0	3 078.560	152.400 0
1996	387.770 0	3 787.200	156.900 0
1997	331.020 0	4 827.350	160.500 0
1998	294.240 0	5 818.260	163.000 0
1999	278.880 0	6 546.810	166.600 0
2000	279.110 0	6 805.890	172.200 0

续前表

年份	黄金价格	NYSE 指数	CPI
2001	274.040 0	6 397.850	177.100 0
2002	309.730 0	5 578.890	179.900 0
2003	363.380 0	5 447.460	184.000 0
2004	409.720 0	6 612.620	188.900 0
2005	444.740 0	7 349.000	195.300 0
2006	603.460 0	8 357.990	201.600 0

a. 在同一散点图中描绘黄金价格、CPI 和 NYSE 指数。

b. 如果一项投资的价格和（或）回报率至少跟得上通货膨胀，就认为它是保值（能抵御通货膨胀）的。为检验这一假设，假定在（a）中的散点图表明适当的情况下，你拟合如下模型：

$$\text{Gold price}_t = \beta_1 + \beta_2 \text{CPI}_t + u_t$$

$$\text{NYSE Index}_t = \beta_1 + \beta_2 \text{CPI}_t + u_t$$

如果虚拟假设正确，你对 β_2 的值有什么样的期望？

3.23 表 3—8 给出 1959—2005 年间美国国内生产总值（GDP）数据。

表 3—8　　1959—2005 年间美国名义和真实 GDP

年份	NGDP	RGDP	年份	NGDP	RGDP
1959	506.6	2 441.3	1983	3 536.7	5 423.8
1960	526.4	2 501.8	1984	3 933.2	5 813.6
1961	544.7	2 560.0	1985	4 220.3	6 053.7
1962	585.6	2 715.2	1986	4 462.8	6 263.6
1963	617.7	2 834.0	1987	4 739.5	6 475.1
1964	663.6	2 998.6	1988	5 103.8	6 742.7
1965	719.1	3 191.1	1989	5 484.4	6 981.4
1966	787.8	3 399.1	1990	5 803.1	7 112.5
1967	832.6	3 484.6	1991	5 995.9	7 100.5
1968	910.0	3 652.7	1992	6 337.7	7 336.6
1969	984.6	3 765.4	1993	6 657.4	7 532.7
1970	1 038.5	3 771.9	1994	7 072.2	7 835.5
1971	1 127.1	3 898.6	1995	7 397.7	8 031.7
1972	1 238.3	4 105.0	1996	7 816.9	8 328.9
1973	1 382.7	4 341.5	1997	8 304.3	8 703.5
1974	1 500.0	4 319.6	1998	8 747.0	9 066.9
1975	1 638.3	4 311.2	1999	9 268.4	9 470.3
1976	1 825.3	4 540.9	2000	9 817.0	9 817.0
1977	2 030.9	4 750.5	2001	10 128.0	9 890.7
1978	2 294.7	5 015.0	2002	10 469.6	10 048.8
1979	2 563.3	5 173.4	2003	10 960.8	10 301.0
1980	2 789.5	5 161.7	2004	11 712.5	10 703.5
1981	3 128.4	5 291.7	2005	12 455.8	11 048.6
1982	3 255.0	5 189.3			

注：除非特别说明，均以十亿美元为单位；季度数据经季节性调整为年度数据；RGDP 以 2000 年不变价格十亿美元计算。

资料来源：*Economic Report of the President*，2007，Tables B-1 and B-2.

a. 将当年美元和不变（即 2000 年）美元数据对时间描图。

b. 用 Y 表示 GDP，X 表示时间（按年从 1 代表 1959 年，2 代表 1960 年开始，直至 47 代表 2005 年）。看以下模型是否适合 GDP 数据：

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

试用当年美元和不变美元两种数据分别估计此模型。

c. 你会怎样解释 β_2？

d. 如果用当年美元 GDP 估计的 β_2 和用不变美元 GDP 估计的 β_2 有所不同，怎样解释这个差异？

e. 从你的计算结果，你能对样本时期美国通货膨胀的性质得出什么结论？

3.24 利用引言中表 I—1 所给数据，验证方程（3.7.1）。

3.25 对习题 2.16 中 SAT 一例做以下练习：

a. 将女生阅读成绩相对于男生阅读成绩描点。

b. 如果散点图表明两者似有线性关系，试求女生阅读成绩对男生阅读成绩的回归。

c. 如果这两个阅读成绩之间有某种关系，它是不是因果关系？

3.26 用数学成绩代替阅读成绩，重做习题 3.25。

3.27 蒙特卡罗研究课堂作业：回到表 2—4 中所列的 10 个 X 值。令 $\beta_1 = 25$ 和 $\beta_2 = 0.5$。假定 $u_i \sim N(0, 9)$，即 u_i 服从均值为 0、方差为 9 的正态分布。用这两个参数值生成 100 个样本，求出 β_1 和 β_2 的 100 个估计值，然后对这些估计值描图。从这一蒙特卡罗研究中，你能得出什么结论？注：当今大多数统计包都能从一些最熟悉的概率分布中生成随机数。如果你在生成这些随机数时遇到困难，请向你的老师求助。

3.28 利用表 3—3 中给出的数据，将手机用户数对计算机用户数进行描点。二者之间有明显的关系吗？如果有，你如何对这种关系给出合理的解释。

附录 3A

□ 3A.1 最小二乘估计的推导

将方程（3.2.1）对 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 求偏导数我们得到：

$$\frac{\partial\left(\sum a_i^2\right)}{\partial \hat{\beta}_1} = -2\sum\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = -2\sum a_i \tag{1}$$

$$\frac{\partial\left(\sum a_i^2\right)}{\partial \hat{\beta}_2} = -2\sum\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)X_i = -2\sum a_i X_i \tag{2}$$

令这些导数为零。经过代数上的简化和运算，便得到方程（3.1.6）和（3.1.7）所给的估计量。

□ 3A.2 最小二乘估计量的线性和无偏性质

由方程（3.1.8）得：

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \tag{3}$$

其中

$$k_i = \frac{x_i}{\sum x_i^2}$$

这说明 $\hat{\beta}_2$ 是 Y 的一个**线性函数**；它是 Y_i 的一个加权平均，以 k_i 为权数，从而它是一个线性估计量。同理，$\hat{\beta}_1$ 也是一个线性估计量。

顺便指出权数 k_i 的一些性质：

1. 因 X_i 被假定为是非随机的，故 k_i 也是非随机的。

2. $\sum k_i = 0$ 。

3. $\sum k_i^2 = 1/\sum x_i^2$ 。

4. $\sum k_i x_i = \sum k_i X_i = 1$ 。这些性质均可直接从 k_i 的定义中验证。

例如，

$$\sum k_i = \sum \left[\frac{x_i}{\sum x_i^2}\right] = \frac{1}{\sum x_i^2}\sum x_i \qquad \text{因为对于一个给定样本，} \sum x_i^2 \text{ 是已知的}$$

$$= 0 \qquad \text{因为与均值的离差之和 } \sum x_i \text{ 恒为 0}$$

现将 PRF 即 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 代入方程（3）得：

$$\begin{aligned}\hat{\beta}_2 &= \sum k_i(\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i u_i \\ &= \beta_2 + \sum k_i u_i\end{aligned} \tag{4}$$

其中利用了前面提到的 k_i 的性质。

对方程（4）两边求数学期望并注意到 k_i 是非随机的，即可视同常数，于是得到：

$$E(\hat{\beta}_2) = \beta_2 + \sum k_i E(u_i) = \beta_2 \tag{5}$$

这是因为，根据假定，有 $E(u_i) = 0$ 。因此，$\hat{\beta}_2$ 是 β_2 的一个无偏估计量。同理可证，$\hat{\beta}_1$ 也是 β_1 的一个无偏估计量。

□ 3A.3 最小二乘估计量的方差和标准误

根据方差定义，可写：

$$\begin{aligned}\text{var}(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2 \\ &= E(\hat{\beta}_2 - \beta_2)^2 \qquad \text{因为 } E(\hat{\beta}_2) = \beta_2 \\ &= E(\sum k_i u_i)^2 \qquad \text{利用上面的方程（4）} \\ &= E(k_1^2 u_1^2 + k_2^2 u_2^2 + \cdots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \cdots + 2k_{n-1} k_n u_{n-1} u_n)\end{aligned} \tag{6}$$

既然根据假定，对每个 i 都有 $E(u_i^2) = \sigma^2$，而且 $E(u_i u_j) = 0$（$i \neq j$），所以：

$$\begin{aligned}\text{var}(\hat{\beta}_2) &= \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum x_i^2} \qquad \text{利用 } k_i^2 \text{ 的定义} \\ &= \text{方程（3.3.1）}\end{aligned} \tag{7}$$

按照同样的思路可求得 $\hat{\beta}_1$ 的方差。一旦得到 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差，取其正的平方根即是相应的标准误。

□ 3A.4 $\hat{\beta}_1$ 与 $\hat{\beta}_2$ 的协方差

根据定义有

$$\begin{aligned}\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2) &= E\{[\hat{\beta}_1 - E(\hat{\beta}_1)][\hat{\beta}_2 - E(\hat{\beta}_2)]\} \\ &= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \qquad \text{为什么?} \\ &= -\bar{X}E(\hat{\beta}_2 - \beta_2)^2 \\ &= -\bar{X}\operatorname{var}(\hat{\beta}_2) \\ &= \text{方程(3.3.9)}\end{aligned} \tag{8}$$

这里用到关系式 $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$ 及 $E(\hat{\beta}_1) = \bar{Y} - \beta_2\bar{X}$，并由此给出 $\hat{\beta}_1 - E(\hat{\beta}_1) = -\bar{X}(\hat{\beta}_2 - \beta_2)$。注：$\operatorname{var}(\hat{\beta}_2)$ 由方程（3.3.1）给出。

□ 3A.5 σ^2 的最小二乘估计量

回顾：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{9}$$

因此

$$\bar{Y} = \beta_1 + \beta_2\bar{X} + \bar{u} \tag{10}$$

从方程（9）减去方程（10）得：

$$y_i = \beta_2 x_i + (u_i - \bar{u}) \tag{11}$$

再想到

$$\hat{u}_i = y_i - \hat{\beta}_2 x_i \tag{12}$$

从而把方程（11）代入方程（12）便得到：

$$\hat{u}_i = \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \tag{13}$$

合并同类项、平方并对两边同时求和则得到：

$$\sum \hat{u}_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2)\sum x_i(u_i - \bar{u}) \tag{14}$$

两边取数学期望又得到：

$$\begin{aligned}E(\sum \hat{u}_i^2) &= \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 + E[\sum (u_i - \bar{u})^2] - 2E[(\hat{\beta}_2 - \beta_2)\sum x_i(u_i - \bar{u})] \\ &= \sum x_i^2 \operatorname{var}(\hat{\beta}_2) + (n-1)\operatorname{var}(u_i) - 2E[\sum k_i u_i(x_i u_i)] \\ &= \sigma^2 + (n-1)\sigma^2 - 2E[\sum k_i x_i u_i^2] \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 \\ &= (n-2)\sigma^2\end{aligned} \tag{15}$$

其中最后一步用到方程（3）对 k_i 的定义和方程（4）给出的关系。还注意到

$$\begin{aligned}E\sum (u_i - \bar{u}_i)^2 &= E[\sum u_i^2 - n\bar{u}^2] \\ &= E\left[\sum u_i^2 - n\left(\frac{\sum u_i}{n}\right)^2\right] \\ &= E\left[\sum u_i^2 - \frac{1}{n}\sum (u_i^2)\right] \\ &= n\sigma^2 - n\sigma^2/n = (n-1)\sigma^2\end{aligned}$$

其中用到 u_i 互不相关及每个 u_i 的方差都是 σ^2 的事实。

因此，我们得到：

$$E(\sum \hat{u}_i^2) = (n-2)\sigma^2 \tag{16}$$

于是，若定义：

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{17}$$

那么，利用方程（16），它的期望值就是：

$$E(\hat{\sigma}^2) = \frac{1}{n-2}E(\sum \hat{u}_i^2) = \sigma^2 \tag{18}$$

这就表明 $\hat{\sigma}^2$ 是真实 σ^2 的一个无偏估计量。

□ 3A.6 最小二乘估计量的最小方差性质

在附录 3A 第 3A.2 节中曾证明，最小二乘估计量 $\hat{\beta}_2$ 是线性和无偏的（而且 $\hat{\beta}_1$ 也如此）。为了证明这些估计量在所有线性无偏估计量中有最小方差，考虑最小二乘估计量 $\hat{\beta}_2$：

$$\hat{\beta}_2 = \sum k_i Y_i$$

其中

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} = \frac{x_i}{\sum x_i^2} \quad \text{见附录 3A.2} \tag{19}$$

这表明 $\hat{\beta}_2$ 是 Y_i 的加权平均，以 k_i 为权重。

让我们定义 β_2 的另一线性估计量 β_2^* 如下：

$$\beta_2^* = \sum w_i Y_i \tag{20}$$

其中权重 w_i 不一定等于 k_i，于是：

$$\begin{aligned} E(\beta_2^*) &= \sum w_i E(Y_i) \\ &= \sum w_i(\beta_1 + \beta_2 X_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i X_i \end{aligned} \tag{21}$$

欲使 β_2^* 无偏，必须有

$$\sum w_i = 0 \tag{22}$$

以及

$$\sum w_i X_i = 1 \tag{23}$$

而且，我们还可以将其方差写成：

$$\begin{aligned} \text{var}(\beta_2^*) &= \text{var}\sum w_i Y_i \\ &= \sum w_i^2 \text{var}(Y_i) \qquad \text{注：} \text{var}(Y_i) = \text{var}(u_i) = \sigma^2 \\ &= \sigma^2 \sum w_i^2 \qquad \text{注：} \text{cov}(Y_i, Y_j) = 0 (i \neq j) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} + \frac{x_i}{\sum x_i^2} \right)^2 \qquad \text{注意其中的数学技巧} \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right) \left(\frac{x_i}{\sum x_i^2} \right) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \left(\frac{1}{\sum x_i^2} \right) \end{aligned} \tag{24}$$

因为倒数第二步的最后一项消失了。(为什么?)

由于方程(24)中的最后一项是常数，所以 β_2^* 的方差只能通过对第一项的处理使之最小化。若令：

$$w_i = \frac{x_i}{\sum x_i^2}$$

则方程(24)简化为：

$$\text{var}(\beta_2^*) = \frac{\sigma^2}{\sum x_i^2} = \text{var}(\hat{\beta}_2) \tag{25}$$

一般来说，当权重 w_i = 最小二乘的权重 k_i 时，线性估计量 β_2^* 的方差等于最小二乘估计量 $\hat{\beta}_2$ 的方差；否则，$\text{var}(\beta_2^*) > \text{var}(\hat{\beta}_2)$。也就是说，如果存在 β_2 的一个最小方差线性无偏估计量，那么，它必定是最小二乘估计量。同理，可以证明 $\hat{\beta}_1$ 是 β_1 的最小方差线性无偏估计量。

□ 3A.7 最小二乘估计量的一致性

我们在经典线性回归模型的框架中已经证明，最小二乘估计量在无论大样本或小样本容量的情况下都是无偏的(和有效的)。但如附录A中所讲到的那样，一个估计量有时候可能不满足一个或多个优良的小样本统计性质。但随着样本容量无限扩大，这些估计量就具有一些优良的统计性质。这些性质被称为**大样本**(large sample)**或渐近性质**(asymptotic properties)。在此附录中，我们将讨论一个大样本性质即**一致性**(consistency)，在附录A中将更充分地对此展开讨论。对双变量模型，我们已经证明了OLS估计量 $\hat{\beta}_2$ 是真实 β_2 的一个无偏估计量。现在我们来证明 $\hat{\beta}_2$ 也是 β_2 的一个一致估计量。如附录A中证明的那样，一致性的一个充分条件是，$\hat{\beta}_2$ 是无偏的，且随着样本容量趋于无穷，其方差趋于零。

既然我们已经证明了无偏性，现在就只需证明 $\hat{\beta}_2$ 的方差在 n 无限增加时趋于零。我们知道

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2/n}{\sum x_i^2/n} \tag{26}$$

将分子和分母同时除以 n 不会改变这个等式的值。

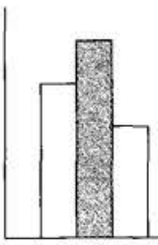
现在

$$\underbrace{\lim \text{var}(\hat{\beta}_2)}_{n \to \infty} = \underbrace{\lim \left(\frac{\sigma^2/n}{\sum x_i^2/n} \right)}_{n \to \infty} = 0 \tag{27}$$

其中用到如下事实：(1)比率的极限等于分子的极限与分母的极限之比(参考任何一本微积分方面的书)；(2)随着 n 趋于无穷，由于 σ^2 是一个有限的数，所以 σ^2/n 趋于零；而由CLRM的假定7，X 的方差是一个有限的正数，所以 $[(\sum x_i^2)/n] \neq 0$。

上述讨论的结果是，OLS估计量 $\hat{\beta}_2$ 是真实 β_2 的一个一致估计量。与此相仿，我们可以证明，$\hat{\beta}_1$ 也是一个一致估计量。因此，在重复(或小)样本下OLS估计量是无偏的，而随着样本容量无限增加，OLS估计量是一致的。如以后所见，即使CLRM的某些假定不满足，我们在某几种情况下也能得到回归系数的一致估计量。

第 4 章 经典正态线性回归模型



所谓**统计推断的经典理论**（classical theory of statistical inference）由两个分支构成，即**估计**（estimation）和**假设检验**（hypothesis testing）。我们到目前为止已讨论了（双变量）线性回归模型的参数估计问题。用 OLS 的方法，我们能估计参数 β_1、β_2 和 σ^2。在经典线性回归模型的假定下，我们可以证明，$\hat{\beta}_1$、$\hat{\beta}_2$、$\hat{\sigma}^2$ 和这些参数的估计量都满足一些理想的统计性质，如无偏性和最小方差等。（回顾 BLUE 性质。）注意，既然它们都是估计量，所以它们的值将随样本而变化。因此，这些估计量都是随机变量。

但估计是成功的一半。假设检验是另一半。回想我们在回归分析中的目标不仅仅是估计样本回归函数（SRF），而是像第 2 章所强调的那样，我们要用估计来对总体回归函数（PRF）进行推断。于是，我们想知道，$\hat{\beta}_1$ 和 $\hat{\sigma}^2$ 与真实的 β_1 和 σ^2 到底有多接近。比如，在例 3.2 中，我们在方程（3.7.2）中估计了 SRF。但这个回归只是基于一个由 55 个家庭构成的样本，我们怎么知道估计的 MPC 为 0.436 8 代表了整个总体中（真实）的 MPC 呢？

因此，由于 $\hat{\beta}_1$、$\hat{\beta}_2$ 和 $\hat{\sigma}^2$ 是随机变量，所以我们需要清楚它们的概率分布，若不知其概率分布，那我们就无法将它们与其真实值相联系。

4.1 干扰项 u_i 的概率分布

为得到 OLS 估计量的概率分布，我们将如下进行。具体地，考虑 $\hat{\beta}_2$。如我们在

附录 3A.2 中证明的那样，

$$\hat{\beta}_2 = \sum k_i Y_i \tag{4.1.1}$$

其中 $k_i = x_i / \sum x_i^2$。但由于假定 X 为固定或非随机的，所以我们的条件回归分析就以 X_i 的固定值为条件。方程（4.1.1）表明，$\hat{\beta}_2$ 是 Y_i 的一个线性函数，根据假定 Y_i 也是随机的。但由于 $Y_i = \beta_1 + \beta_2 X_i + u_i$，所以我们可以把方程（4.1.1）写成

$$\hat{\beta}_2 = \sum k_i (\beta_1 + \beta_2 X_i + u_i) \tag{4.1.2}$$

由于 k_i、β 系数和 X_i 都是固定的，所以 $\hat{\beta}_2$ 最终是 u_i 的一个线性函数，而根据假定，u_i 是随机变量。因此，$\hat{\beta}_2$（及 $\hat{\beta}_1$）的概率分布将取决于对 u_i 的概率分布所做的假定。由于对 OLS 估计量的概率分布知识足以对其总体值做出推断，所以 u_i 概率分布的性质在假设检验中就起到极为重要的作用。

由于在最小二乘法中我们并没有对干扰项 u_i 的概率性质做任何假定，所以尽管有了高斯-马尔可夫定理，但仍无助于从 SRF 去推断 PRF。如果我们愿意假定 u 服从某种概率分布的话，就可弥补这一缺憾。在回归分析中，人们常常假定 u 服从正态分布，其原因稍后即明。在第 3 章中讨论的经典线性回归模型的假定中增加 u_i 的正态性假定，我们就得到所谓**经典正态线性回归模型**（classical normal linear regression model，CNLRM）。

4.2 关于 u_i 的正态性假定

经典正态线性回归假定每个 u_i 都是正态分布的，且其

均值：$E(u_i) = 0$ (4.2.1)

方差：$E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$ (4.2.2)

协方差 $\text{cov}(u_i, u_j)$：$E\{[u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i u_j) = 0 \quad i \neq j$ (4.2.3)

这些假定可更简洁地叙述为：

$$u_i \sim N(0, \sigma^2) \tag{4.2.4}$$

其中～表示“其分布为”，而 N 代表“正态分布”，括号中的项代表正态分布的两个参数：均值与方值。

就像在附录 A 中提到的那样，**对两个正态分布变量来说，零协方差或零相关就意味着两个变量互相独立。**因此，在正态性假定下，方程（4.2.4）不仅意味着 u_i 与 u_j 不相关，而且意味着它们是独立分布的。

于是，可将方程（4.2.4）写为：

$$u_i \sim \text{NID}(0, \sigma^2) \tag{4.2.5}$$

其中 **NID** 表示“正态且独立分布”。

□ 为什么要做正态假定?

我们为什么采用正态性假定呢? 有如下几个理由:

1. 如 2.5 节所指出的, u_i 代表回归模型中未明显引进的许多自变量(对因变量)的总影响。如同已指出的那样,我们希望这些被忽略的变量所起的作用是微小的,而且充其量是随机的。于是,利用统计学中著名的**中心极限定理**(central limit theorem, CLT)就能证明(详见附录 A),如果存在大量独立且相同分布的随机变量,那么,除了少数例外情形,随着这些变量的个数无限地增加,它们的总和将趋向服从正态分布。[①] 正是这个中心极限定理为 u_i 的正态性假定提供了理论基础。

2. 中心极限定理的另一个说法是,即使变量个数并不很大或这些变量并不是严格独立的,但它们的总和仍可视为正态分布的。[②]

3. 如附录 A 中所言,**正态分布的一个性质是,正态分布变量的任何线性函数都是正态分布的。**因此,在正态性假定下,OLS 估计量的概率分布很容易推导。前面曾讨论过,OLS 估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是 u_i 的线性函数。因此,若 u_i 是正态分布的,则 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 也是正态分布的,这就使得我们的假设检验工作十分简单。

4. 正态分布是一个比较简单的、仅涉及两个参数(均值和方差)的分布;它为人们所熟知,它的理论性质在数理统计学中受到广泛的研究。而且看上去许多现象都是正态分布。

5. 如果我们在处理小样本或有限样本时,比方说数据少于 100 次观测,那么正态假定就起到关键作用。它不仅有助于我们推导出 OLS 估计量精确的概率分布,而且使我们能用 t 检验、F 检验和 χ^2 检验来对回归模型进行统计检验。附录 A 讨论了 t、F 和 χ^2 概率分布的统计性质。如我们稍后所见,如果样本容量大到合理的程度,我们或许能放宽正态性假定。

6. 最后,在大样本中,t 和 F 统计量近似服从 t 和 F 概率分布,所以基于误差项正态分布这一假定的 t 和 F 检验仍能可靠地使用。[③] 如今有许多包含大量观测的横截面数据和时间序列数据。因此,在大样本数据集中,正态性假定也许不是非常关键。

一句忠告:既然我们"施加"了正态性假定,那我们就有必要在一些涉及小样本容量数据的实际应用中分析正态性假定是否适当。稍后,我们将对此进行检验。此外,我们以后还会偶尔遇到一些正态性假定不适当的情况。但出于前面讨论的原因,我们仍继续做正态性假定,除非我们看到这个假定并不适当。

---

① 对此定理的一个简单直接的讨论,见 Sheldon M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, 2d ed, Harcourt Academic Press, New York, 2000, pp. 193-194。此定理的一个例外情形是柯西分布;见 M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Charles Griffin & Co., London, 1960, vol. 1, pp. 248-249。

② 至于中心极限定理的各种形式,可参见 Harald Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946, Chap. 17。

③ 对这一论点的技术性讨论,参见 Christiaan Heij et al., *Econometric Methods with Applications in Business and Economics*, Oxford University Press, Oxford, 2004, p. 197。

4.3 在正态性假定下 OLS 估计量的性质

在方程（4.2.5）中 u_i 为正态分布的假定下，OLS 估计量有如下统计性质（附录 A 中对估计量的统计性质做了一个一般性的讨论）：

1. 它们是无偏的。

2. 它们有最小方差。结合性质 1 就意味着它们是**最小方差无偏的**（minimum-variance unbiased）或者说它们是**有效估计量**（efficient estimators）。

3. **一致性**。也就是说，随着样本容量无限增大，估计量将收敛于它们的真值。

4. $\hat{\beta}_1$（u_i 的线性函数）是正态分布的，且

$$\text{均值：} E(\hat{\beta}_1) = \beta_1 \tag{4.3.1}$$

$$\text{方差 var}(\hat{\beta}_1)\text{：} \quad \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \quad = (3.3.3)\ (4.3.2)$$

或更简洁地写成：

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

然后，利用正态分布的性质，定义：

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \tag{4.3.3}$$

可知变量 Z 服从**标准正态分布**（standard normal distribution），即零均值和单位（=1）方差的正态分布，或写成：

$$Z \sim N(0, 1)$$

5. $\hat{\beta}_2$（u_i 的线性函数）是正态分布的，且

$$\text{均值：} E(\hat{\beta}_2) = \beta_2 \tag{4.3.4}$$

$$\text{方差 var}(\hat{\beta}_2)\text{：} \quad \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2} \quad = (3.3.1)\ (4.3.5)$$

或更简洁地写成：

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$$

然后，如同方程（4.3.3）那样，

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \tag{4.3.6}$$

也服从标准正态分布。

图 4—1 从几何图形上描绘了 $\hat{\beta}_1$ 的概率分布。

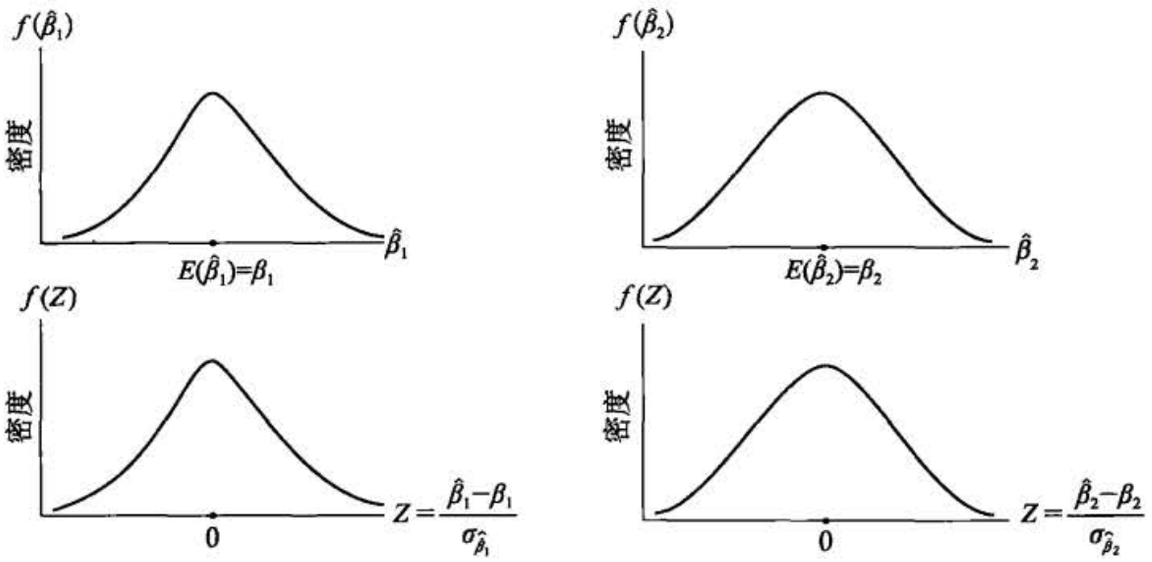


图 4—1 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的概率分布

6. $(n-2)\hat{\sigma}^2/\sigma^2$ 服从 $n-2$ 个自由度的 χ^2（即卡方）分布。[①] 如我们在第 5 章将看到的那样，这一点有助于我们从估计的 σ^2 中对真实的 σ^2 进行推断。（关于 χ^2 分布及其性质的讨论见附录 A。）

7. $(\hat{\beta}_1, \hat{\beta}_2)$ 的分布独立于 $\hat{\sigma}^2$。这一重要性质将在下一章进行解释。

8. $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 在所有无偏估计中，无论是线性的还是非线性的，都有最小方差。饶（C. R. Rao）给出的这一结论是非常强有力的。它与高斯-马尔可夫定理不同，它的成立，不仅限于线性估计量。[②] 因此，我们可以说最小二乘估计量是**最优无偏估计量**（best unbiased estimators，BUE）。即在所有无偏估计中，这些估计量具有最小方差。

总结：重要的是要看到，正态性假定使我们能够推导出 $\hat{\beta}_1$、$\hat{\beta}_2$（都是正态的）以及 $\hat{\sigma}^2$（与 χ^2 相关的）的概率或抽样分布。在下一章中我们将看到，这将简化构造置

① 对这个命题的证明略显复杂，一个比较容易理解的证明可参阅 Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics*, 2d ed., Macmillan, New York, 1965, p. 144。

② C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York, 1965, p. 258.

信区间和（统计）假设检验的工作。

顺便指出，如果假定 u_i 服从以 0 为均值 σ^2 为方差的正态分布，则 Y_i 作为 u_i 的线性函数，本身也服从正态分布，其均值和方差依次为：

$$E(Y_i) = \beta_1 + \beta_2 X_i \tag{4.3.7}$$

$$\text{var}(Y_i) = \sigma^2 \tag{4.3.8}$$

或更简洁地写为：

$$Y_i \sim N\ (\beta_1 + \beta_2 X_i,\ \sigma^2) \tag{4.3.9}$$

4.4 极大似然法

和 OLS 相比，**极大似然法**（maximum likelihood，ML）是一种具有更强的理论特征的点估计方法。由于此法较复杂，故放在本章的附录中讨论。对一般读者来说，只需知道：如果假定 u_i 是正态分布的（假定的理由已讨论在前），则回归系数 β 的 ML 估计量和 OLS 估计量是相同的，无论所考虑的是简单回归还是多元回归。但 σ^2 的 ML 估计量是 $\sum \hat{u}_i^2/n$，这是一个有偏误的估计量，而 σ^2 的 OLS 估计量 $\sum \hat{u}_i^2/(n-2)$ 则是无偏的。比较 σ^2 的这两种估计量，可知随着样本容量 n 的变大，两者趋于相等。因此，σ^2 的 ML 估计量是渐近（即随着 n 无限增大）无偏的。

既然补充了 u_i 的正态性假定，最小二乘法便为我们提供了对线性回归模型进行估计和假设检验的全部必备工具。即使读者由于极大似然法在数学上略为复杂而不愿意探讨它，也不致蒙受什么损失。

要点与结论

1. 本章讨论了经典正态线性回归模型。

2. 此模型与经典线性回归模型（CLRM）的差异，在于它特意假定了进入回归模型的干扰项 u_i 是正态分布的。CLRM 则不要求对 u_i 的概率分布作任何假定；仅要求 u_i 的均值为零以及方差为一个有限常数。

3. 正态性假定的理论依据是**中心极限定理**。

4. 在没有正态性假定的情况下，在第 3 章所讨论的其他假定下，高斯-马尔可夫定理表明，OLS 估计量是最优线性无偏估计量。

5. 由于正态性假定，OLS 估计量就不仅是**最优无偏估计量**，而且服从熟知的概率分布。截距和斜率的 OLS 估计量本身是正态分布的，并且 u_i 的方差的 OLS 估计量（$=\hat{\sigma}^2$）与 χ^2 分布有关。

6. 在第 5 章和第 8 章中，我们将说明怎样把这些知识用于推断总体参数的真值。

7. 取代最小二乘法的一个方法是**极大似然法**。然而，为了使用此法，必须对干扰项 u_i 的概率分布作一假定。在回归分析中，最常作的假定就是 u_i 服从正态分布。

8. 在正态性假定下，截距和斜率参数的 ML 估计量和 OLS 估计量是完全相同的。但是，u_i 的方差的 OLS 估计量和 ML 估计量却有差别。然而，在大样本中，这两个估计量趋于一致。

9. 通常称 ML 为大样本方法。ML 有更为广泛的应用。也就是说，它可应用于对参数为非线性的回归模型。对非线性情形，一般都不用 OLS。对此更多的讨论内容见第 14 章。

10. 在本教材中，我们基本上依靠的是 OLS，具体理由如下：(a) 相对于 ML 来说，OLS 易于应用；(b) β_1 和 β_2 的 ML 估计量和 OLS 估计量是相同的（对多元回归也是如此）；(c) 即使样本不很大，σ^2 的 OLS 估计量和 ML 估计量也相差无几。

然而，为了方便爱好数学的读者，本章附录以及附录 A 中附有 ML 的一个简要介绍。

附录 4A

□ 4A.1 双变量回归模型的极大似然估计

假定在双变量模型 $Y_i=\beta_1+\beta_2X_i+u_i$ 中，Y_i 是正态且独立分布的，其均值$=\beta_1+\beta_2X_i$，其方差$=\sigma^2$。[参看方程 (4.3.9)。] 从而，给定上述均值和方差，Y_1，Y_2，…，Y_n 的联合概率密度函数就可写为：

$$f(Y_1,Y_2,\cdots,Y_n \mid \beta_1+\beta_2X_i,\sigma^2)$$

但由于各个 Y 的独立性，此联合概率密度函数可写为 n 个单个密度函数之积：

$$\begin{aligned} &f(Y_1,Y_2,\cdots,Y_n \mid \beta_1+\beta_2X_i,\sigma^2) \\ &=f(Y_1 \mid \beta_1+\beta_2X_i,\sigma^2)f(Y_2 \mid \beta_1+\beta_2X_i,\sigma^2)\cdots f(Y_n \mid \beta_1+\beta_2X_i,\sigma^2) \end{aligned} \tag{1}$$

其中

$$f(Y_i)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{(Y_i-\beta_1-\beta_2X_i)^2}{\sigma^2}\right\} \tag{2}$$

这是给定均值和方差的一个正态分布变量的密度函数。(注：exp 指以 {} 中的表达式为 e 的幂。)

将方程 (2) 中的每个 Y_i 代入方程 (1) 便得到：

$$f(Y_1,Y_2,\cdots,Y_n \mid \beta_1+\beta_2X_i,\sigma^2)=\frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left\{-\frac{1}{2}\sum\frac{(Y_i-\beta_1-\beta_2X_i)^2}{\sigma^2}\right\} \tag{3}$$

若 Y_1，Y_2，…，Y_n 为已知或给定，而 β_1，β_2 和 σ^2 为未知，则称 (3) 为**似然函数** (likelihood function)，记为 LF(β_1，β_2，σ^2) 并写为[①]：

$$\mathrm{LF}(\beta_1,\beta_2,\sigma^2)=\frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left\{-\frac{1}{2}\sum\frac{(Y_i-\beta_1-\beta_2X_i)^2}{\sigma^2}\right\} \tag{4}$$

极大似然法，顾名思义，就是要在估计未知参数时使得观测到给定的这些 Y_i 的概率尽可能大。因此，有必要求方程 (4) 的最大值。这不过是微分运算中的一个简单练习。为了求微分，将

① 当然，若 β_1，β_2 和 σ^2 已知而 Y_i 未知，则方程 (4) 代表联合概率密度函数——指联合观测 Y_i 的概率。

方程（4）表达成如下对数形式更为容易。[①]（注：ln 表示自然对数。）

$$\ln \mathrm{LF} = -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i-\beta_1-\beta_2X_i)^2}{\sigma^2}$$

$$= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i-\beta_1-\beta_2X_i)^2}{\sigma^2} \tag{5}$$

将方程（5）对 β_1，β_2 和 σ^2 求偏导数得：

$$\frac{\partial \ln \mathrm{LF}}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum(Y_i-\beta_1-\beta_2X_i)(-1) \tag{6}$$

$$\frac{\partial \ln \mathrm{LF}}{\partial \beta_2} = -\frac{1}{\sigma^2}\sum(Y_i-\beta_1-\beta_2X_i)(-X_i) \tag{7}$$

$$\frac{\partial \ln \mathrm{LF}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i-\beta_1-\beta_2X_i)^2 \tag{8}$$

令这些方程为零（最优化的一阶条件），并记 ML 估计量为 $\tilde{\beta}_1$，$\tilde{\beta}_2$ 和 $\tilde{\sigma}^2$，便得到[②]：

$$\frac{1}{\tilde{\sigma}^2}\sum(Y_i-\tilde{\beta}_1-\tilde{\beta}_2X_i) = 0 \tag{9}$$

$$\frac{1}{\tilde{\sigma}^2}\sum(Y_i-\tilde{\beta}_1-\tilde{\beta}_2X_i)X_i = 0 \tag{10}$$

$$-\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum(Y_i-\tilde{\beta}_1-\tilde{\beta}_2X_i)^2 = 0 \tag{11}$$

经过简化，方程（9）和（10）给出：

$$\sum Y_i = n\tilde{\beta}_1 + \tilde{\beta}_2\sum X_i \tag{12}$$

$$\sum Y_iX_i = \tilde{\beta}_1\sum X_i + \tilde{\beta}_2\sum X_i^2 \tag{13}$$

这正是在方程（3.1.4）和（3.1.5）中得到的最小二乘理论的正规方程。由此可见，ML 估计量 $\tilde{\beta}$ 无异于由方程（3.1.6）和（3.1.7）给出的 OLS 估计量 $\hat{\beta}$，这个等同的结果并非偶然。分析一下似然函数（5），我们看到最后一项是带有负号的。因此，方程（5）的最大化就是这一项的最小化，而后者如同我们能从方程（3.1.2）中看到的那样，正是最小二乘法所采取的路线。

将 ML（=OLS）估计量代入方程（11）并加以简化，就得到 $\tilde{\sigma}^2$ 的 ML 估计量为：

$$\tilde{\sigma}^2 = \frac{1}{n}\sum(Y_i-\tilde{\beta}_1-\tilde{\beta}_2X_i)^2$$

$$= \frac{1}{n}\sum(Y_i-\hat{\beta}_1-\hat{\beta}_2X_i)^2$$

$$= \frac{1}{n}\sum\hat{u}_i^2 \tag{14}$$

从方程（14）明显看出，ML 估计量 $\tilde{\sigma}^2$ 不同于 OLS 估计量 $\hat{\sigma}^2 = [1/(n-2)]\sum\hat{u}_i^2$，后者在附录 3A 第 3A.5 节中已被证明是 σ^2 的一个无偏估计量，因此，σ^2 的 ML 估计量是有偏误的。偏误的程度也很容易确定如下。

对方程（14）两边取数学期望得：

$$E(\tilde{\sigma}^2) = \frac{1}{n}E(\sum\hat{u}_i^2)$$

$$= \left(\frac{n-2}{n}\right)\sigma^2 \qquad \text{利用附录 3A 第 3A.5 的方程（16）} \tag{15}$$

---

① 由于对数函数是单调函数，故 ln LF 和 LF 在同一点上达到最大。

② 我们用“~”（波浪号）表示 ML 估计量；用“^”（尖帽）表示 OLS 估计量。

$$= \sigma^2 - \frac{2}{n}\sigma^2$$

这表明，在小样本中，$\tilde{\sigma}^2$ 偏小（即低估了真实的 σ^2）。但应看到，随着样本容量 n 无限增大，方程（15）中的第二项即偏误因子将趋于零。因此 $\tilde{\sigma}^2$ 是渐近（即在很大的样本中）无偏的，也就是说，当 $n \to \infty$ 时，$\lim E(\tilde{\sigma}^2) = \sigma^2$。还可进一步证明，$\tilde{\sigma}^2$ 也是一致估计量。[①] 即当 n 无限增大时，$\tilde{\sigma}^2$ 收敛于真值 σ^2。

□ 4A.2 印度食物支出的极大似然估计

回到例 3.2 和方程（3.7.2），在那里，利用印度 55 个农户的数据将食物支出对总支出做回归。由于在正态性假定下，回归系数的 OLS 估计量和 ML 估计量相同，所以我们得到 ML 估计量为 $\tilde{\beta}_1 = \hat{\beta}_1 = 94.208\ 7$ 和 $\tilde{\beta}_2 = \hat{\beta}_2 = 0.436\ 8$。$\sigma^2$ 的 OLS 估计量为 $\hat{\sigma}^2 = 4\ 469.691\ 3$，但 ML 估计量 $\tilde{\sigma}^2 = 4\ 407.156\ 3$，小于 OLS 估计量。如前面指出的那样，ML 估计量在小样本情形下有向下的偏误；即它总体上低估了真实方差 σ^2。当然，如你所料，随着样本容量的增加，这两个估计量之间的差别将越来越小。将估计量的值放到似然函数中，我们得到似然函数值为 $-308.162\ 5$。若需要 LF 的极大值，只需取 $-308.162\ 5$ 的反对数即可。没有其他任何参数值能以更高的概率得到你用以分析的样本。

□ 附录 4A 习题

4.1 “若两个随机变量在统计上独立，则两者的相关系数为零。但反之未必成立。也就是说，零相关不意味着统计独立性。然而，如果两个变量都是正态分布的，则零相关必然意味着统计独立性。”试利用下面的两个正态分布变量 Y_1 和 Y_2 的联合概率密度函数（又称**双变量正态概率密度函数**，bivariate normal probability density function）来证明这一命题：

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \times \left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(Y_1-\mu_1)(Y_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

其中，$\mu_1 = Y_1$ 的均值；

$\mu_2 = Y_2$ 的均值；

$\sigma_1 = Y_1$ 的标准差；

$\sigma_2 = Y_2$ 的标准差；

$\rho = Y_1$ 与 Y_2 之间的相关系数。

4.2 试用取极值的二阶条件（即二阶导数检验），证明通过解方程（9）、（10）和（11）而得到的 β_1、β_2 和 σ^2 的 ML 估计量，确实可使方程（4）中的似然函数取极大值。

4.3 随机变量 X 服从**指数分布**（exponential distribution），如果它有如下的概率密度函数：

$$f(X) = (1/\theta)e^{-X/\theta} \quad \text{当 } X > 0$$
$$= 0 \quad \text{当 } X \text{ 取其他值}$$

① 关于极大似然估计量性质的一般讨论，以及渐近无偏性与一致性之间的差别所在，参看附录 A。粗略地说，对于渐近无偏性，我们要设法求出当 n 趋于无穷大时的 $\lim E(\tilde{\sigma}_n^2)$，其中 n 代表估计量所依据的样本容量，而对于一致性，我们要设法求出当 n 无限增大时 $\tilde{\sigma}_n^2$ 的变化过程。注意，无偏性是指在给定样本容量的情况下，一个估计量的重复抽样性质，而一致性是关于一个估计量在样本容量无限增大过程中所表现出来的性质。

其中 $\theta > 0$ 是此分布的参数。试用 ML 证明 θ 的 ML 估计量是 $\tilde{\theta} = \sum X_i/n$，其中 n 为样本容量。也就是证明 θ 的 ML 估计量是样本均值 $\overline{X}$。

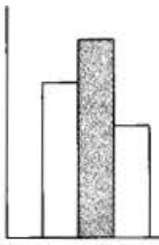
4.4 假设一项实验的结果要么成功要么失败。在实验结果成功时令 $X=1$，在实验结果失败时令 $X=0$，X 的概率密度函数如下

$$p(X=0) = 1-p$$

$$p(X=1) = p,\ 0 \leqslant p \leqslant 1$$

成功概率 p 的极大似然估计量是什么？

第 5 章 双变量回归：区间估计与假设检验



警惕过多地检验假设；你对数据越苛求，数据会越多地向你供认，但在威逼下得到的供词，在科学研究的法庭上是不容许的。[①]

如在第 4 章中指出的那样，估计与假设检验构成经典统计学的两个主要分支。估计理论由两部分组成：点估计与区间估计。在前面两章中我们介绍 OLS 和 ML 估计方法时已透彻地讨论过点估计。在本章中，我们先考虑区间估计，然后再讨论假设检验的问题。后面这个问题与区间估计有着紧密的关系。

5.1 统计学的预备知识

在讲解如何构造置信区间与检验统计假设的具体步骤之前，我们假定读者已熟悉概率与统计学的基本概念。附录 A 虽然不能代替一门统计学的基础课程，却具备了读者所必须掌握的统计学要义。一些基本概念如**概率**（probability）、**概率分布**（probability distributions）、**第Ⅰ类错误**（Type Ⅰ error）和**第Ⅱ类错误**（Type Ⅱ error）、**显著（性）水平**（level of significance）、**统计检验**（statistical test）的功效以及**置信区间**（confidence interval）等，对于理解本章和以后各章的内容都起着关键

① Stephen M. Stigler, "Testing Hypothesis of Fitting Models? Another Look at Mass Extinctions," in Matthew H. Nitecki and Antoni Hoffman, eds., *Neutral Models in Biology*, Oxford University Press, Oxford, 1987, p. 148.

作用。

5.2 区间估计：一些基本思想

为了集中注意力，不妨考虑第3章的工资—受教育程度的例子。方程（3.6.1）表明，受教育程度提高1年，估计平均小时工资（$\hat{\beta}_2$）将提高约0.724 0，这是对未知总体参数β_2的一个数字（点）估计值。这个估计值的可靠度如何呢？就像在第3章曾指出的那样，由于抽样波动，单个估计值很可能与真实值不同，尽管在重复抽样的过程中，预计它的均值会等于真实值。[注：$E(\hat{\beta}_2)=\beta_2$。]在统计学中，一个点估计量的可靠性由它的标准误来衡量。因此，我们不能完全信赖一个点估计值，而是要围绕点估计量构造一个区间。比方说，在点估计量的两旁各宽2个或3个标准误的一个区间，使得它有95%的概率包含着真实的参数值。这就是**区间估计**（interval estimation）的粗略概念。

说得更确切些，假定我们想知道究竟$\hat{\beta}_2$距离β_2有多“近”。为此，我们试求两个正数δ和α，α位于0与1之间，使得**随机区间**（random interval）$(\hat{\beta}_2-\delta,\ \hat{\beta}_2+\delta)$包含真实$\beta_2$的概率为$1-\alpha$。用符号表示：

$$\Pr(\hat{\beta}_2-\delta\leqslant\beta_2\leqslant\hat{\beta}_2+\delta)=1-\alpha \tag{5.2.1}$$

如果这样的区间存在，就被称为**置信区间**（confidence interval）；$1-\alpha$被称为**置信系数**（confidence coefficient）；而α（$0<\alpha<1$）则被称为**显著（性）水平**（level of significance）。[①] 置信区间的端点被称为**置信限**（confidence limits）或**临界值**（critical values）。$\hat{\beta}_2-\delta$被称为**置信下限**（lower confidence limit），而$\hat{\beta}_2+\delta$被称为**置信上限**（upper confidence limit）。顺便指出，在实践中，α和$1-\alpha$常用百分数表示成$100\alpha\%$和$100(1-\alpha)\%$的形式。

方程（5.2.1）表明，和点估计量相对应，**区间估计量**（interval estimator）是一个构造出来的区间，要使得它以一个特定的概率$1-\alpha$把参数的真实值包括在区间的界限内。比方说，$\alpha=0.05$或5%，那么方程（5.2.1）就可读为：式中的（随机）区间包含真实β_2的概率为0.95或95%。从而区间估计量给出了一个真实β_2可能会落入其中的数值范围。

理解区间估计的下列特征是非常重要的：

1. 方程（5.2.1）并没有说β_2落入给定界限内的概率是$1-\alpha$。因为β_2虽然未知，但被假定为某个定数，或者落在区间内，或者落在区间外。方程（5.2.1）所表述的是，使用本章所描述的方法构造出来的一个区间包含β_2的概率为$1-\alpha$。

---

① 又称**犯第Ⅰ类错误的概率**（probability of committing a Type Ⅰ error）。第Ⅰ类错误，是指拒绝一个正确假设的错误。而第Ⅱ类错误，则指接受一个错误假设的错误。（对这个问题，附录A中有充分的讨论。）符号α又被称为**（统计）检验的尺度**[size of the (statistical) test]。

2. 方程（5.2.1）中的区间是一个**随机区间**；它随样本的变化而变化，因为它是根据 $\hat{\beta}_2$ 来构造的，而 $\hat{\beta}_2$ 是随机的。（为什么？）

3. 既然置信区间是随机的，与之相关的概率命题就应从长远的意义上或从重复抽样的意义上加以理解。说得更具体些，方程（5.2.1）表明，如果在重复抽样中，像方程（5.2.1）那样基于概率 $1-\alpha$ 而构造许多置信区间，那么，从长期看，平均来说，这些区间中将有 $1-\alpha$ 的比例包含着参数的真实值。

4. 如在（2）中提到的那样，只要 $\hat{\beta}_2$ 尚不知道，方程（5.2.1）中的区间就是随机的。但是，一旦我们有了一个特定的样本并获得 $\hat{\beta}_2$ 的一个特定数值，方程（5.2.1）中的区间就不再是随机的，而是固定的了。这时我们不可做如同方程（5.2.1）那样的表述；也就是说，我们不能说一个给定了的固定区间包含真实 β_2 的概率是 $1-\alpha$。在这种情况下，β_2 或者落入这个固定区间内或者落在固定区间之外，从而概率只能是 1 或 0。因此，在我们的工资—受教育程度的例子中，如果我们求得的 95%置信区间是 $0.570\ 0 \leqslant \beta_2 \leqslant 0.878\ 0$，如同我们在方程（5.3.9）中很快就要看到的那样，我们就**不可以**说这个区间包含真实 β_2 的概率是 95%。这个概率不是 1 就是 0。

怎样构造置信区间？从上面的讨论，读者也许预想到，如果估计量的**抽样**（sampling）或**概率分布**（probability distribution）已知，就可以做出如同方程（5.2.1）那样的置信区间的表达式。在第 4 章中，我们曾看到，在干扰项 u_i 的正态性假定下，*OLS 估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 本身是正态分布的，而 OLS 估计量 $\hat{\sigma}^2$ 则与 χ^2 分布有关*。这样看来，构造置信区间是一桩简单的事情。确实如此！

5.3 回归系数 β_1 和 β_2 的置信区间

□ β_2 的置信区间

第 4 章 4.3 节已表明，在 u_i 的正态性假定下，OLS 估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 本身就是正态分布的，其均值和方差已随之列出。因此，如在方程（4.3.6）中所指出的那样，变量

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\mathrm{se}(\hat{\beta}_2)} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma} \tag{5.3.1}$$

是一个标准正态变量。因此，如果真实的总体方差 σ^2 已知，就可利用正态分布对 β_2 作概率判断。当 σ^2 已知时，以 μ 为均值和 σ^2 为方差的正态分布变量有一个重要性质，就是正态曲线下在 $\mu \pm \sigma$ 之间的面积约为 68%；在 $\mu \pm 2\sigma$ 之间的面积约为 95%；在 $\mu \pm 3\sigma$ 之间的面积约为 99.7%。

但是我们很少知道 σ^2，在实践中是用无偏估计量 $\hat{\sigma}^2$ 来测定的。如果我们用 $\hat{\sigma}$ 代

替σ，方程（5.3.1）就可写为：

$$t=\frac{\hat{\beta}_2-\beta_2}{\mathrm{se}(\hat{\beta}_2)}=\frac{\text{估计量}-\text{参数}}{\text{估计量的标准误的估计}}$$

$$=\frac{(\hat{\beta}_2-\beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}} \tag{5.3.2}$$

其中 $\mathrm{se}(\hat{\beta}_2)$ 在这里用来表示估计量的标准误。可以证明（见附录 5A 第 5A.2 节），这样定义的 t 变量服从自由度为 $n-2$ 的 t 分布。[注意方程（5.3.1）与（5.3.2）之间的区别。] 因此，我们不用正态分布，而是要用 t 分布来构造 β_2 的置信区间：

$$\Pr(-t_{\alpha/2}\leqslant t\leqslant t_{\alpha/2})=1-\alpha \tag{5.3.3}$$

其中位于两个不等号中间的 t 值就是由方程（5.3.2）给出的 t 值，而 $t_{\alpha/2}$ 是由显著水平为 $\alpha/2$ 和自由度为 $n-2$ 的 t 分布给出的 t 变量值，常常被称为在 $\alpha/2$ 显著水平上的**临界值**。将方程（5.3.2）代入方程（5.3.3）得：

$$\Pr\left[-t_{\alpha/2}\leqslant\frac{\hat{\beta}_2-\beta_2}{\mathrm{se}(\hat{\beta}_2)}\leqslant t_{\alpha/2}\right]=1-\alpha \tag{5.3.4}$$

重新整理方程（5.3.4）得：

$$\Pr\left[\hat{\beta}_2-t_{\alpha/2}\mathrm{se}(\hat{\beta}_2)\leqslant\beta_2\leqslant\hat{\beta}_2+t_{\alpha/2}\mathrm{se}(\hat{\beta}_2)\right]=1-\alpha \tag{5.3.5}$$ ①

方程（5.3.5）给出 β_2 的一个 $100(1-\alpha)\%$**置信区间**，可更简洁地把它写成：

β_2 的 $100(1-\alpha)\%$置信区间：

$$\hat{\beta}_2\pm t_{\alpha/2}\mathrm{se}(\hat{\beta}_2) \tag{5.3.6}$$

利用方程（4.3.1）和（4.3.2），类似地推理，就能写出：

$$\Pr\left[\hat{\beta}_1-t_{\alpha/2}\mathrm{se}(\hat{\beta}_1)\leqslant\beta_1\leqslant\hat{\beta}_1+t_{\alpha/2}\mathrm{se}(\hat{\beta}_1)\right]=1-\alpha \tag{5.3.7}$$

或更简洁地写为：

β_1 的 $100(1-\alpha)\%$置信区间：

$$\hat{\beta}_1\pm t_{\alpha/2}\mathrm{se}(\hat{\beta}_1) \tag{5.3.8}$$

注意，由方程（5.3.6）和（5.3.7）给出的置信区间有一重要的特点：在这两个方程中置信区间的宽度都与估计量的标准误成比例。也就是说，标准误越大，置信区间越宽。换句话说，估计量的标准误越大，对未知参数的真值进行估计的不确定性越大。因此，估计量的标准误常被喻为估计量的**精度**，即用估计量去测定真实的总体值有多精确。

---

① 一些作者喜欢在方程（5.3.5）中把自由度显示出来，从而把方程（5.3.5）写为：

$$\Pr\left[\hat{\beta}_2-t_{(n-2),\alpha/2}\mathrm{se}(\hat{\beta}_2)\leqslant\beta_2\leqslant\hat{\beta}_2+t_{(n-2),\alpha/2}\mathrm{se}(\hat{\beta}_2)\right]=1-\alpha$$

但为简单起见，我们将保持我们的记号；有关的自由度在行文中加以澄清。

回到在第3章中（3.6节）将平均小时工资（Y）对受教育程度（X）进行回归的例子。回想我们在表3—2中发现，$\hat{\beta}_2 = 0.724\ 0$；$se(\hat{\beta}_2) = 0.070\ 0$。由于只有13个观测，所以自由度就是11。如果我们假定$\alpha = 5\%$，即构造一个95%的置信区间，那么，t表告诉我们，自由度为11的**临界值** $t_{\alpha/2} = 2.201$。将这些值代入方程(5.3.5)，即可证实β_2的95%置信区间为[①]：

$$0.570\ 0 \leqslant \beta_2 \leqslant 0.878\ 0 \tag{5.3.9}$$

或者，按照方程（5.3.6）的形式把它写为：

$$0.724\ 0 \pm 2.201 \times 0.070\ 0$$

即：

$$0.724\ 0 \pm 0.154\ 0 \tag{5.3.10}$$

对这个置信区间的解释是：给定置信系数为95%，每100个像方程（5.3.9）这样的区间中，将有95个包含真实的β_2。但就像前面曾告诫的那样，我们不可以说方程(5.3.9)这个特定区间有95%的概率包含着真实的β_2，因为这个区间已经固定而不再是随机的了；因此，β_2要么落入其中，要么落在其外：这个特定的固定区间包含着真实β_2的概率不是1就是0。

根据方程（5.3.7）和表3—2中的数据，读者应很容易验证，本例中β_1的95%置信区间是

$$-1.887\ 1 \leqslant \beta_1 \leqslant 1.858\ 3 \tag{5.3.11}$$

同样，在解释这个置信区间时，你应该格外小心。在100个像方程（5.3.11）这样的区间中，将有95个包含真实的β_1；这个特定区间包含真实β_1的概率不是1就是0。

□ β_1 和 β_2 的联合置信区间

有时我们需要构造β_1和β_2的一个联合置信域（joint confidence interval），使得β_1和β_2同时落在其中的置信系数$(1-\alpha)$等于，比方说，95%。由于这个问题比较复杂，有兴趣的读者可参考有关文献。[②] 我们将在第8章和第10章简要地提到这个主题。

5.4 σ^2 的置信区间

如在第4章4.3节中所指出的那样，在正态性假定下，变量：

① 由于表3—2中的四舍五入误差，以下给出的答案与统计软件给出的答案可能不是完全一致。

② 一个简明的讨论，见 John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Regression Models*, Richard D. Irwin, Homewood, Ill., 1983, Chap. 5。

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{5.4.1}$$

服从自由度为 $n-2$ 的 χ^2 分布。[①] 故可利用 χ^2 分布来构造 σ^2 的置信区间：

$$\Pr(\chi^2_{1-\alpha/2} \leqslant \chi^2 \leqslant \chi^2_{\alpha/2}) = 1-\alpha \tag{5.4.2}$$

其中两个不等号中间的 χ^2 值由方程（5.4.1）给出，而 $\chi^2_{1-\alpha/2}$ 和 $\chi^2_{\alpha/2}$ 是得自 χ^2 表中自由度为 $n-2$ 的两个 χ^2 值（χ^2 **临界值**），使得它们各自切去 χ^2 分布尾部 $100(\alpha/2)\%$ 的面积，如图 5—1 所示。

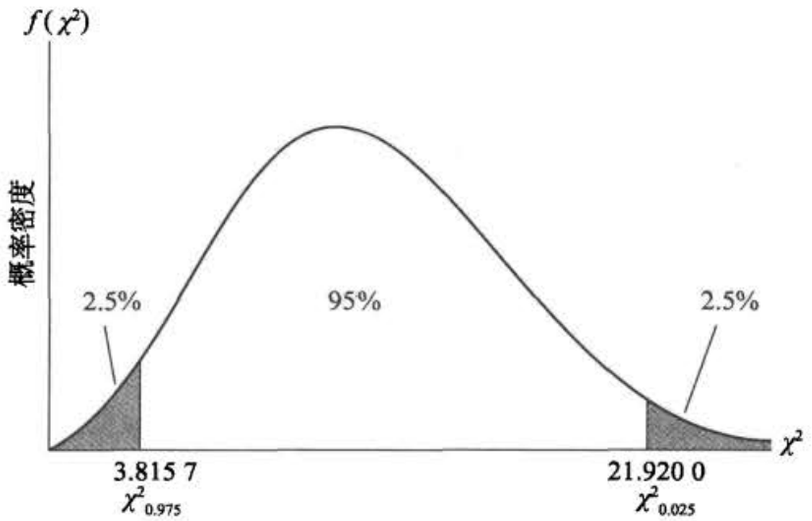


图 5—1　χ^2 的 95%置信区间（11 个自由度）

将方程（5.4.1）的 χ^2 代入方程（5.4.2），并加以整理便得到

$$\Pr\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha \tag{5.4.3}$$

这就给出 σ^2 的 $100(1-\alpha)\%$ 置信区间。

继续我们前面有关工资—受教育程度的例子，我们在表 3—2 中发现，对于我们的数据，有 $\hat{\sigma}^2 = 0.893\ 6$。如果取 α 为 5%，则对于自由度为 11 的 χ^2 表给出下列临界值：$\chi^2_{0.025} = 21.920\ 0$ 和 $\chi^2_{0.975} = 3.815\ 7$。这些值表示 χ^2 值超过 21.920 0 的概率是 2.5%；超过 3.815 7 的概率是 97.5%。因此，这两值之间的区间构成 χ^2 的一个 95%置信区间，如图 5—1 所示。（注意 χ^2 分布的偏态。）

读者将此例的数据代入方程（5.4.3），便能证实 σ^2 的 95%置信区间为：

$$0.448\ 4 \leqslant \sigma^2 \leqslant 2.576\ 0 \tag{5.4.4}$$

对这个区间的解释是：如果我们建立了 σ^2 的 95%置信（界）限，并且事先声称这些界限将包含真实的 σ^2，那么从长远看，我们将有 95%的概率是正确的。

① 证明见 Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics*, 2d ed., Macmillan, New York, 1965, p. 144。

5.5 假设检验：概述

我们已经讨论了点估计和区间估计的问题，现在考虑假设检验的问题。在本节中我们只对这个问题作一个简要的概述；更多的细节见附录 A。

统计假设检验的问题可简单地叙述如下：某一给定的观测或发现与某声称的假设是否相符？这里用“相符”(compatible) 一词表示与假设的值“足够接近”，因而我们不拒绝所声称的假设。例如，如果某种理论或先前经验使我们相信工资—受教育程度一例的真实斜率系数 β_2 等于 1，那么从表 3—2 的样本得到的观测值 $\hat{\beta}_2 = 0.724$ 是否与声称的假设值 1 相一致呢？如果是，我们不拒绝该假设；否则就可拒绝它。

用统计学的语言说，这个声称的假设叫做**虚拟假设** (null hypothesis) [通常代表一种信以为真或意在维护的所谓**维持假设** (maintained hypothesis)]，并用符号 H_0 来表示。通常在检验虚拟假设时要有一个**对立假设** (alternative hypothesis)，常用符号 H_1 来表示。比如，H_1 表示真实的 β_2 不等于 1。对立假设可以是**简单的**或**复合的**。[①] 例如，$H_1: \beta_2 = 1.5$ 是一个简单假设，但 $H_1: \beta_2 \neq 1.5$ 则是一个复合假设。

假设检验理论是要提出一个观测或程序，以便决定拒绝抑或不拒绝一个虚拟假设。为了设计这样的规则有两种互为补充的方法，就是**置信区间**和**显著性检验** (test of significance)。两种方法都宣称所考虑的变量（统计量或估计量）服从某个概率分布，并且做假设检验就在于对这个分布的参数值发表意见或作出判断。例如，我们知道在正态性假定下，$\hat{\beta}_2$ 是正态分布的，其均值等于 β_2；其方差由方程 (4.3.5) 给出。当我们假设 $\beta_2 = 1$ 时，我们就是在对这个正态分布的两参数之一即均值作出判断。本书中所遇到的统计假设大多数都属于这种类型，都是对某些假定的概率分布诸如正态分布、F 分布、t 分布或 χ^2 分布中的一个或多个参数的值作出判断。这些判断是怎样作出的将在下面两节讨论。

5.6 假设检验：置信区间方法

□ 双侧或双尾检验

为了说明置信区间方法，再次回到工资—受教育程度一例。我们从方程

① 一个统计假设如果规定了一个概率分布参数的准确值，就叫做**简单假设** (simple hypothesis)；否则就叫做**复合假设** (composite hypothesis)。例如，在正态概率密度函数 $[1/(\sigma\sqrt{2\pi})]\exp\left\{-\frac{1}{2}[(X-\mu)/\sigma]^2\right\}$ 中，如果我们断言 $H_1: \mu = 15$ 和 $\sigma = 2$，它就是一个简单假设；但如果 $H_1: \mu = 15$ 和 $\sigma > 15$，由于标准差没有一个准确值，它就是一个复合假设。

(3.6.1) 中所给的回归结果得知，斜率系数是 0.724 0。假使我们假设：

$$H_0: \beta_2 = 0.5$$

$$H_1: \beta_2 \neq 0.5$$

也就是说，真实斜率系数在虚拟假设下是 0.5，而在对立假设下不等于 0.5。虚拟假设是一个简单假设，而对立假设则是一个复合假设；实际上这就是所谓的**双侧假设**(two-sided hypothesis)。这样的双侧对立假设，常常反映着我们对于对立假设偏离虚拟假设的方向没有一个强有力的先验或理论的期望。

所观测的 $\hat{\beta}_2$ 是否与 H_0 相符呢？为了回答这个问题，不妨引用方程（5.3.9）中的置信区间。我们知道，从长期看，像（0.570 0，0.878 0）这样的许多区间将有 95%的概率包含真实的 β_2。因此，在长期（即重复抽样）的意义上，这样的区间以（比方说）95%的置信系数给出真实的 β_2 落入其中的一个范围或界限，从而置信区间给出了一个可信的虚拟假设集合。因此，如果虚拟假设 H_0 下的 β_2 落入这个 100 $(1-\alpha)$%置信区间，我们就不拒绝虚拟假设；如果它落在区间之外，我们就可拒绝它。[①] 在图 5—2 中，我们勾画了这一区间范围。

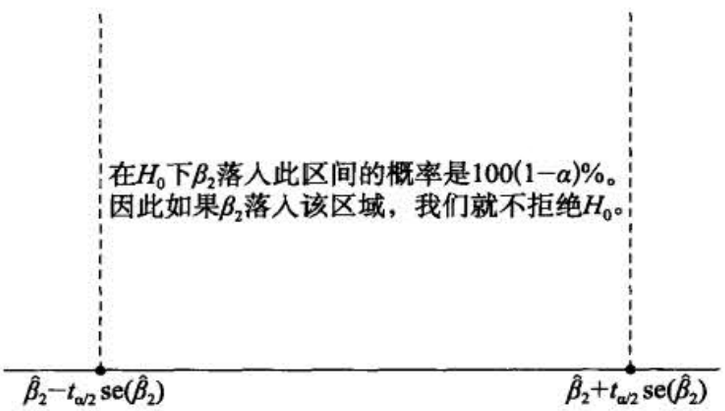


图 5—2　β_2 的一个 100$(1-\alpha)$%置信区间

决策规则

构造一个 β_2 的 100$(1-\alpha)$%置信区间。如果在假设 H_0 下 β_2 落入此区间，就不拒绝 H_0。但如果它落在此区间之外，就要拒绝 H_0。

遵照此规则，拿我们假设的例子来说，$H_0: \beta_2 = 0.5$ 显然落在由方程（5.3.9）给出的 95%置信区间之外，因此我们能以 95%的置信度拒绝真实斜率为 0.5 的假设。即令虚拟假设是真实的，我们得到一个大到 0.724 0 的斜率值，最多也只有 5%的机会，这是一个小概率。

① 请牢记，即使假设 H_0 正确，在 H_0 下仍有 100α%的机会，这个区间不包含 β_2。简单地说，有 100α%的机会犯第 I 类错误。比方说，如果 $\alpha=0.05$，就会有 5%的机会我们拒绝了一个正确的虚拟假设。

在统计学中，当我们拒绝虚拟假设时，我们说我们的发现是**统计上显著的**（statistically significant）。反之，当我们不拒绝虚拟假设时，我们说我们的发现**不是统计上显著的**（not statistically significant）。

一些作者使用“统计上高度显著的”一词，该词通常是指，当他们拒绝虚拟假设时，犯第Ⅰ类错误的概率（即 α）是一个小数，通常指1%。但在5.8节中，我们对 ***p*** **值**的讨论将表明，较好的做法是让研究者自己去决定一个统计上的发现究竟是“显著的”、“中度显著的”，还是“高度显著的”。

□ 单侧或单尾检验

有时，我们有一种强烈的先验预期或理论预期（或基于前人的一些经验研究），认定对立假设是单侧或单向的，而不是刚才讨论过的双侧假设。比如，就拿我们的工资—受教育程度的例子来说，我们可以假设：

$$H_0:\beta_2 \leqslant 0.5 \quad 和 \quad H_1:\beta_2 > 0.5$$

也许是经济理论或以前的经验研究提示了边际消费倾向大于0.5。尽管检验上述假设的程序容易从方程（5.3.5）推导出来，而实际的步骤最好通过下面即将讨论的显著性检验方法予以说明。[①]

5.7 假设检验：显著性检验方法

□ 检验回归系数的显著性：*t* 检验

作为对置信区间方法的一种补充，检验统计假设的另一种方法是分别由费希尔（R. A. Fisher）以及由内曼（Neyman）和皮尔逊（Pearson）提出的**显著性检验方法**（test-of-significance approach）。[②] **概括地说，显著性检验是一种利用样本结果来证实一个虚拟假设真伪的检验程序。**显著性检验背后的关键思想在于一个**检验统计量**（test statistic）及其在虚拟假设下的抽样分布。根据手头数据算出的检验统计量值决定是否接受 H_0。

作为一个说明，回忆在正态假定下，变量：

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}} \tag{5.3.2}$$

服从自由度为 $n-2$ 的 t 分布。如果在虚拟假设下 β_2 的真值被设定，则容易从现有样本计算出方程（5.3.2）中的 t 值。因此，这个 t 变量就可作为一个统计量。并且由

① 如果你想采用置信区间方法，就要构造 β_2 的一个 $100(1-\alpha)\%$ 单侧或单尾置信区间。为什么？

② 详见 E. L. Lehman, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959。

于这个统计量服从 t 分布，故可做出如下置信区间表述：

$$\Pr\left[-t_{\alpha/2} \leqslant \frac{\hat{\beta}_2 - \beta_2^*}{\mathrm{se}(\hat{\beta}_2)} \leqslant t_{\alpha/2}\right] = 1-\alpha \tag{5.7.1}$$

其中 β_2^* 是在 H_0 下的 β 值，而 $-t_{\alpha/2}$ 和 $t_{\alpha/2}$ 是得自 t 表中相对于 $\alpha/2$ 显著水平和 $n-2$ 个自由度的 t 值（t 临界值）[参看方程（5.3.4）]。t 表见于附录 D。

将方程（5.7.1）整理得：

$$\Pr\left[\beta_2^* - t_{\alpha/2}\mathrm{se}(\hat{\beta}_2) \leqslant \hat{\beta}_2 \leqslant \beta_2^* + t_{\alpha/2}\mathrm{se}(\hat{\beta}_2)\right] = 1-\alpha \tag{5.7.2}$$

此式给出在给定 $\beta_2 = \beta_2^*$ 时，$\hat{\beta}_2$ 以概率 $1-\alpha$ 落入其中的区间。用假设检验的语言说，方程（5.7.2）中建立的 $100(1-\alpha)\%$ 置信区间叫做（虚拟假设的）**接受域**（region of acceptance），而置信区间以外的（一个或多个）区域叫做（虚拟假设的）**拒绝域**（region of rejection）或**临界域**（critical region）。如前所说，置信限，即置信区间的端点，又叫做**临界值**。

现在，通过比较方程（5.3.5）和方程（5.7.2），就能看清假设检验的置信区间方法和显著性检验方法之间的密切联系。在置信区间程序中，我们试图建立一个以某种概率包含有真实但未知的 β_2 的一个范围或区间，而在显著性检验步骤中，我们假设 β_2 为某值，然后来看所计算的 $\hat{\beta}_2$ 是否位于该假设值周围的某个合理（置信）范围之内。

让我们再次回到工资—受教育程度一例。我们知道 $\hat{\beta}_2 = 0.724\ 0$，$\mathrm{se}(\hat{\beta}_2) = 0.070\ 0$ 和 df=11。若取 $\alpha = 5\%$，则 $t_{\alpha/2} = 2.201$。

若令 $H_0: \beta_2 = \beta_2^* = 0.5$ 和 $H_1: \beta_2 \neq 0.5$，方程（5.7.2）即为：

$$\Pr(0.346\ 0 \leqslant \hat{\beta}_2 \leqslant 0.654\ 0) \tag{5.7.3}$$[①]

如图 5—3 所示。因所测 $\hat{\beta}_2$ 的值落在临界域中，故拒绝真实 $\beta_2 = 0.5$ 的虚拟假设。

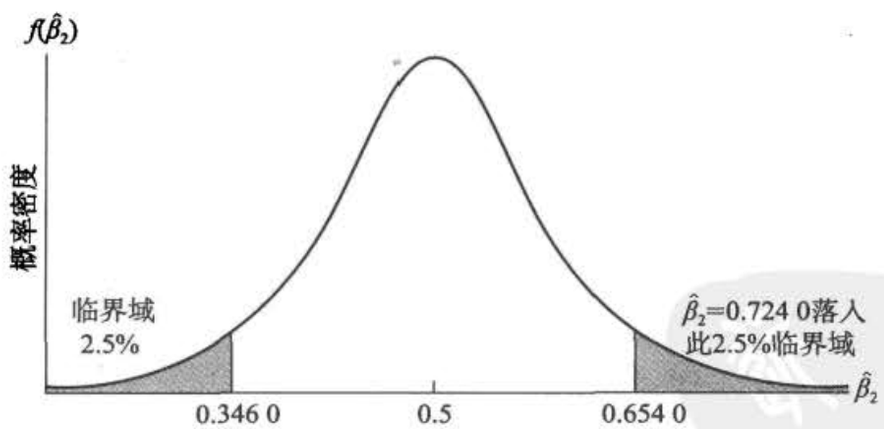


图 5—3　在假设 $\beta_2 = 0.5$ 下 $\hat{\beta}_2$ 的 95%置信区间

在实践中，并不需要明确地估计方程（5.7.2），而是按照方程（5.7.1）给出的

① 5.2 节第 4 点曾表明，我们不可以说固定了的区间（0.570 0，0.878 0）包含真实 β_2 的概率是 95%。但是作为估计量的 $\hat{\beta}_2$ 是一个随机变量，所以我们就可以给出方程（5.7.3）中的概率表述。

双重不等式计算居中的 t 值，然后看它是落在两个 t 临界值之间还是之外。对于我们的例子

$$t=\frac{0.724\ 0-0.5}{0.070\ 0}=3.2 \tag{5.7.4}$$

它明显落在图 5—4 的临界域内。结论仍然是一样的：我们拒绝 H_0。

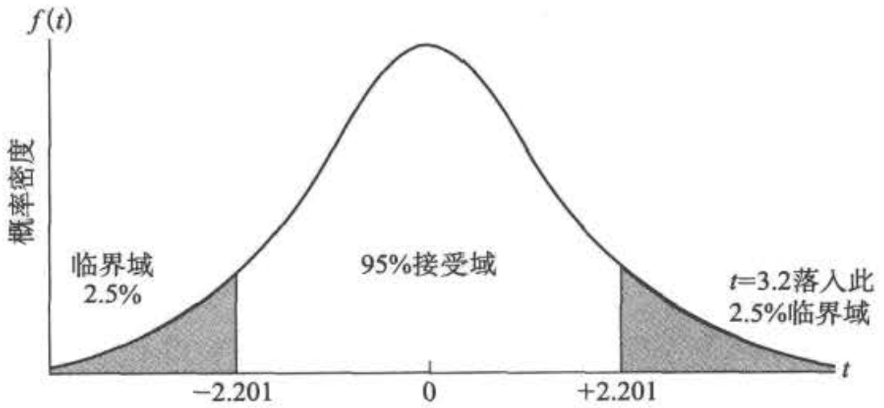


图 5—4　11 个自由度下 t 的 95%置信区间

注意，如果估计的 β_2（$=\hat{\beta}_2$）等于假设的 β_2，方程（5.7.4）中的 t 将为零。然而，随着估计的 β_2 值远离假设的 β_2 值，$|t|$（即 t 的绝对值；注意 t 可正可负）将越来越大。因此，一个“大”的 $|t|$ 值便是与虚拟假设相抵触的迹象。当然，我们总可以利用 t 表来决定一个特定的 t 值是大还是小；我们知道，这个答案依赖于自由度的个数和我们愿意接受的第 I 类错误的概率。如果你翻阅一下附录 D 中的 t 表，你将察觉，对给定的自由度，得到的 $|t|$ 值越大，其概率越小。比如，对 20 个自由度来说，得到一个 1.725 或更大的 $|t|$ 值的概率是 0.10 或 10%。但对于同样的自由度，得到一个 3.552 或更大的 $|t|$ 值的概率仅为 0.002 或 0.2%。

因为我们利用了 t 分布，所以上述检验程序适合称为 **t 检验**。**用显著性检验的语言说，如果一个统计量的值落在临界域内，那么这个统计量就是统计上显著的。这时我们拒绝虚拟假设。同理，如果一个检验统计量的值落在接受域中，那么它就是统计上不显著的，这时我们不拒绝虚拟假设。**在我们的例子中 t 是显著的，从而我们拒绝虚拟假设。

在结束我们对假设检验的讨论前，请注意我们刚才描述的检验程序是一种**双侧或双尾**显著性检验程序，因为我们把有关概率分布的两个尾端当作拒绝域；如果虚拟假设值落入任一尾端，就拒绝该假设，这样做的原因是我们的 H_1 是一个双侧复合假设；$\beta_2\neq0.5$ 表示 β_2 或者大于 0.5，或者小于 0.5。但是，假设先前的经验提示我们，预计斜率要比 0.5 大，这样，我们就有 $H_0:\beta_2\leqslant0.5$ 和 $H_1:\beta_2>0.5$。虽然 H_1 仍是一个复合假设，但它却是单侧的。为了检验此假设，我们利用**单尾检验**（one-tail test，右尾部），如图 5—5 所示。（并参考 5.6 节的讨论。）

除了置信上限或临界值现在是 $t_\alpha=t_{0.05}$ 即 5%的水平外，检验的程序同前。如图 5—5 所示，在此情形中，我们并不需要考虑 t 分布的左尾端。究竟使用双尾还是单

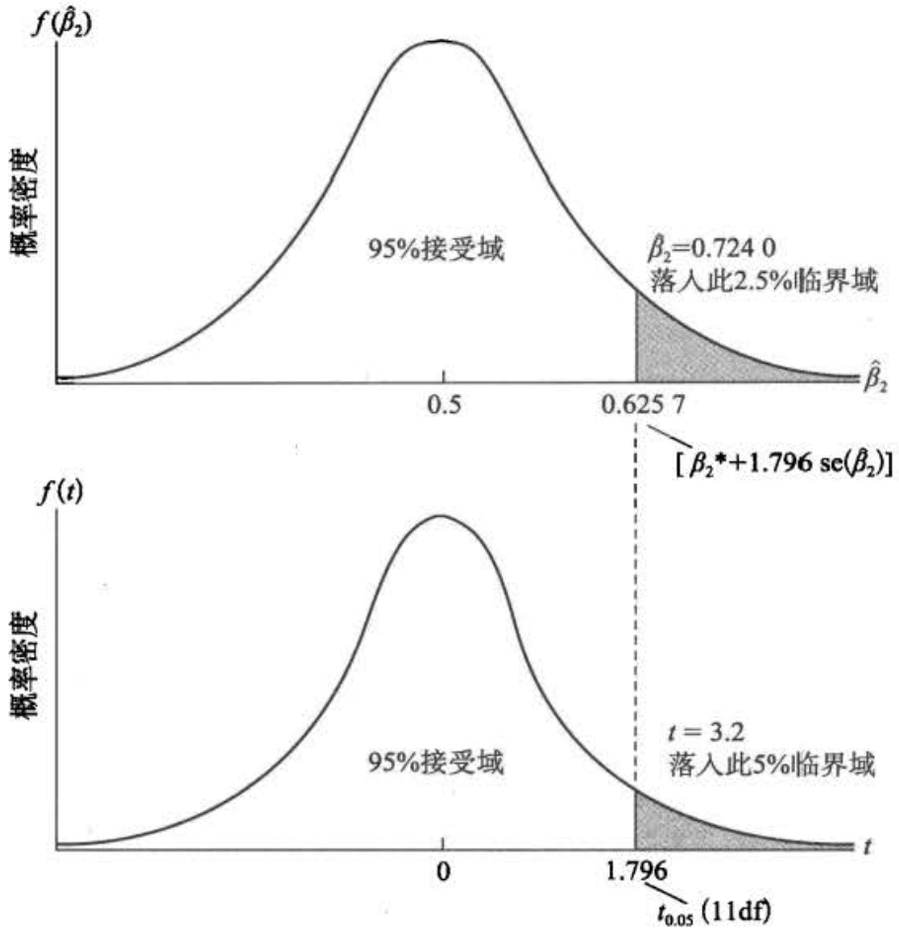


图 5—5　单尾显著性检验

尾显著性检验，要看对立假设是怎样构成的。而后者又有赖于某种先验思考或先前的实际经验。（进一步的讨论见 5.8 节。）

假设检验的显著性 t 检验方法可概括为表 5—1。

表 5—1　　显著性 t 检验：决策规则

假设类型	H_0 ：虚拟假设	H_1 ：对立假设	决策规则：拒绝 H_0 ，如果
双尾	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2,\mathrm{df}}$
右尾	$\beta_2 \leqslant \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,\mathrm{df}}$
左尾	$\beta_2 \geqslant \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,\mathrm{df}}$

注：β_2^* 是 β_2 的假设数值。

$\lvert t \rvert$ 指 t 的绝对值。

t_α 或 $t_{\alpha/2}$ 指在 α 或 $\alpha/2$ 显著水平上的 t 临界值。

df 是指自由度，对双变量模型是（$n-2$），对三变量模型是（$n-3$）。依此类推，同样的程序适用于 β_1 的假设检验。

□ 检验 σ^2 的显著性：χ^2 检验

作为显著性检验方法的另一说明，考虑以下变量：

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{5.4.1}$$

前面已指出，这个变量服从自由度为 $n-2$ 的 χ^2 分布。对于我们的例子，$\hat{\sigma}^2 = 0.893\ 7$ 并且 df=11。如果假设 H_0：$\sigma^2 = 0.6$ 和 H_1：$\sigma^2 \neq 0.6$，方程（5.4.1）便给出关于 H_0 的检验统计量。把相应的数值代入方程（5.4.1）就能求出在 H_0 下 $\chi^2 = 16.384\ 5$。如果我们取 $\alpha = 5\%$，χ^2 的两个临界值便是 3.815 75 和 21.920 0。由于计算出来的 χ^2 落在这两个界限之间，表明数据支持虚拟假设，因此我们不拒绝它。（见图 5—1。）这一检验程序叫做 **χ^2 显著性检验**（chi-square test of significance）。假设检验的 χ^2 显著性检验方法可概括为表 5—2。

表 5—2　　χ^2 检验概要

H_0：虚拟假设	H_1：对立假设	临界域：拒绝 H_0，如果
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\mathrm{df}(\hat{\sigma}^2)/\sigma_0^2 > \chi^2_{\alpha,\mathrm{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\mathrm{df}(\hat{\sigma}^2)/\sigma_0^2 < \chi^2_{(1-\alpha),\mathrm{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\mathrm{df}(\hat{\sigma}^2)/\sigma_0^2 > \chi^2_{\alpha/2,\mathrm{df}}$ 或 $< \chi^2_{(1-\alpha/2),\mathrm{df}}$

注：σ_0^2 是在虚拟假设下的 σ^2 值。最后列中 χ^2 的第一个下标指显著水平，而第二个下标指自由度。这些 χ^2 均是临界值。注意，对双变量回归模型，自由度为（$n-2$），对三变量回归模型自由度为（$n-3$），依此类推。

5.8 假设检验：一些实际操作问题

□“接受”或“拒绝”假设的含义

在显著性检验中，比如说在 t 检验的基础上，如果我们决定“接受”虚拟假设，我们不过是说，根据样本证据，我们还没有理由拒绝它；而不是说，虚拟假设毫无疑问是真的。为什么？为了回答此问题，让我们回到工资—受教育程度的例子中，并假定 H_0：$\beta_2 = 0.70$。既然斜率的估计值是 $\hat{\beta}_2 = 0.724\ 1$，并且 $\mathrm{se}(\hat{\beta}_2) = 0.070\ 1$，那么，根据 t 检验，我们求得 $t = (0.724\ 1 - 0.70)/0.070\ 1 = 0.343\ 8$，这在 $\alpha = 5\%$ 的水平上是不显著的。因此，我们“接受” H_0。但是，让我们再假定 H_0：$\beta_2 = 0.60$，应用 t 检验，我们又得到 $t = (0.724\ 1 - 0.6)/0.070\ 1 = 1.770\ 3$，这仍然是统计上不显著的。于是，再“接受”此 H_0。但这两个虚拟假设哪一个“真实”呢？我们不知道。所以，在“接受”一个虚拟假设时，应时刻警觉到另一个虚拟假设也可能会同样地与数据相符。有鉴于此，我们宁可说可以接受一个虚拟假设，而不能说我们（确实）接受它。更好的说法是：

> ……正如一个法庭宣告某一判决为“无罪”而不为“清白”，统计检验的结论也应为“不拒绝”而不为“接受”[①]。

① Jan Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971, p. 114.

□“零”虚拟假设与“2-t”经验法则

在经验工作中经常检验的一个虚拟假设是 $H_0:\beta_2 = 0$，即斜率系数是零。这个“零”虚拟假设像是一个草人，目的是要明确 Y 是否与解释变量 X 有某种关系。如果要从 Y 和 X 之间无任何关系开始，那么检验诸如 $\beta_2 = 0.3$ 或任何其他值的虚拟假设就没有意义。

可以容易地用前几节所讨论的置信区间方法或 t 检验方法来检验这个虚拟假设。但常常可采用“2-t”显著性法则将这类按部就班的检验方法简化如下：

> **“2-t”经验法则（“2-t” Rule of Thumb）**
>
> 如果自由度为 20 或更大且显著水平定在 0.05，那么，从方程（5.3.2）算得 t 值［$=\hat{\beta}_2/\mathrm{se}(\hat{\beta}_2)$］在绝对值上超过 2 时，就可拒绝虚拟假设 $\beta_2=0$。

此准则的合理性不难领会。由方程（5.7.1）知，对于适当的自由度，如果

$$t = \hat{\beta}_2/\mathrm{se}(\hat{\beta}_2) > t_{\alpha/2} \qquad \text{当 } \hat{\beta}_2 > 0 \text{ 时}$$

或者

$$t = \hat{\beta}_2/\mathrm{se}(\hat{\beta}_2) < -t_{\alpha/2} \qquad \text{当 } \hat{\beta}_2 < 0 \text{ 时}$$

我们将拒绝 $H_0:\beta_2 = 0$。也可以说，如果满足如下条件，我们将拒绝 H_0：

$$|t| = \left|\frac{\hat{\beta}_2}{\mathrm{se}(\hat{\beta}_2)}\right| > t_{\alpha/2} \tag{5.8.1}$$

检查一下附录 D 中的 t 表便看到，当自由度约为 20 或更大时，计算的 t 值在绝对值上超过 2，比方说 2.1，在 5%的水平上是统计上显著的，即意味着对虚拟假设的拒绝。因此，对于 20 或更大的自由度，如果计算的 t 值是 2.5 或 3，我们就不需要查阅 t 表以评定所估斜率系数的显著性。当然，为了得知准确的显著水平，我们可随时查阅 t 表，而当自由度小于 20 时，我们一定要查阅 t 表。

顺便指出，如果我们相对 $\beta_2 > 0$ 或 $\beta_2 < 0$ 检验单侧假设 $\beta_2 = 0$，则当下式成立时，应拒绝虚拟假设：

$$|t| = \left|\frac{\hat{\beta}_2}{\mathrm{se}(\hat{\beta}_2)}\right| > t_{\alpha} \tag{5.8.2}$$

如果把 α 定在 0.05，则从 t 表我们看到，对于 20 或更多的自由度，一个超过 1.73 的 t 值在 5%的显著水平上是（单尾）统计上显著的。因而，每当 t 值超过比方说 1.8（在绝对值上）且自由度为 20 或更大，就不需要查阅 t 表以评定所测系数的统计显著性。当然，如果 α 选定在 0.01 或任何其他水平上，则还必须决定适当的 t 值作为临界参考点。至此，读者该知道怎么去做了。

□ 构造虚拟假设和对立假设[①]

给定了虚拟假设和对立假设，如何检验它们的统计显著性就不再是什么神秘的事了。但是怎样构造这些假设并没有什么一成不变的规则。常常是我们所研究的现象会提示我们虚拟假设和对立假设的性质。例如，考虑证券组合理论中的资本市场线，该理论假设 $E_i = \beta_1 + \beta_2 \sigma_i$，其中 E=组合证券的期望回报，而 σ_i=回报的标准差。后者是风险的一种度量。因为人们预期回报与风险存在正相关关系——回报越高，风险越大，因此相对于虚拟假设 $\beta_2 = 0$，自然的对立假设便是 $\beta_2 > 0$。也就是说，人们不会作出考虑 β 为负值的选择。

然而，试看对货币的需求问题。以后我们将说明，货币需求的重要决定因素之一是收入。先前关于货币需求函数的研究曾表明，货币需求的收入弹性（指收入变化1%时货币需求变化的百分数）典型地位于0.7和1.3之间。因此，如果在一项新的货币需求研究中假设收入弹性系数 β_2 为1，则对立假设可取为 $\beta_2 \neq 1$，即一双侧对立假设。

可见，理论预期或经验工作或两者同时可作为构造假设的依据，但不管怎样构造这些假设，一件极为重要的事，是研究者要在进行经验调查研究之前先构造这些假设。否则，他或她就犯了迂回推理或自欺欺人的错误。也就是说，如果先分析经验结果再进行假设的话，就不免受到一种诱惑，要构造一种假设来袒护自己所得到的结果。要不惜一切代价去避免这种做法，至少为了科学事业要如此，请牢记本章一开头就引用的施蒂格勒（Stigler）的话！

□ 选择显著性水平 α

讨论至此，应该清楚，拒绝或不拒绝虚拟假设，关键在于 α 这个显著性水平或犯**第Ⅰ类错误的概率**——拒绝了真实假设的概率。在附录A中我们充分地讨论了第Ⅰ类错误的性质，它和**第Ⅱ类错误**（接受了错误的假设）的关系，以及为什么经典统计学通常都集中于讨论第Ⅰ类错误。但是，即使我们讨论了这些问题，人们仍会问，为什么 α 通常都固定在1%、5%，也许还有10%的水平上？其实，这些值并不是神圣不可侵犯的；任何其他值都是可以的。

在像本书这样的导论中，不可能深入地讨论为什么人们选择1%，5%或10%的显著性水平。这样做就会把我们引入到本身就是一个学科分支的统计决策领域。然而，这里可以作一个简短的概括。如同在附录A中所讨论的那样，对于给定的样本容量，如果我们要减少犯第Ⅰ类错误，第Ⅱ类错误就要增加；反之亦然。就是说，给定了样本容量，如果我们企图减少拒绝真实假设的概率，我们就同时增加了接受错误假设的概率，因此，对于给定的样本容量，这两种错误类型之间有一种替代关

① 关于如何建立假设的一个饶有趣味的讨论，见J. Bradford De Long and Kevin Lang, "Are All Economic Hypotheses False?" *Journal of Political Economy*, vol. 100, no. 6, 1992, pp. 1257-1272。

系。解决这一替代关系的唯一途径，就是找出两类错误的相对代价。于是，

> 如果错误地拒绝一个其实是真实的虚拟假设（第Ⅰ类错误）的代价比起错误地未拒绝一个其实是错误的虚拟假设（第Ⅱ类错误）的代价相对高昂，那么把第Ⅰ类错误的概率定得低些将是合理的。反之，如果犯第Ⅰ类错误的代价比犯第Ⅱ类错误的代价相对低廉，就值得把第Ⅰ类错误的概率定得高些（从而使犯第Ⅱ类错误的概率低些）。[①]

当然，困难在于我们很少知道犯这两类错误的代价。因此，应用计量经济学家一般看来都是把 α 定在1%、5%甚至10%的水平上。然后选择一个能使犯第Ⅱ类错误的概率尽可能小的检验统计量。由于1减去犯第Ⅱ类错误的概率被称为**检验功效**（power of the test），这一程序相当于求检验功效的最大化。（关于检验功效的讨论，参看附录A。）

但是，如果我们采用下节将讨论的检验统计量的 ***p* 值**，则所有有关选择适当 α 值的问题均可避免。

□ 精确的显著性水平：*p* 值

如方才指出的那样，经典假设检验方法的痛处在于选择 α 时的武断性。当我们对给定的样本算出一个检验统计量（如 t 统计量）的值时，为什么不干脆查阅适当的统计表，看看得到一个和从样本得到的检验统计量那样大或者更大的数值的确切概率？这个概率就叫做 ***p* 值**（p value），即**概率值**（probability value），也叫做**观测或精确显著水平**（observed or exact level of significance），或**犯第Ⅰ类错误的精确概率**（exact probability of committing a Type Ⅰ error）。用更专业化的语言说，p 值被定义为**一个虚拟假设可被拒绝的最低显著水平**。

作为说明，仍回到工资—受教育程度一例。给定虚拟假设：受教育程度的真实系数是0.5，我们得到方程（5.7.4）中的 t 值为3.2。得到一个大到3.2或更大的 t 值的 p 值是什么？查阅附录D中的 t 表，我们看到，对于11个自由度，得到这样的 t 值的概率一定小于0.005（单尾）或0.010（双尾）。

如果你利用Stata或EViews等统计软件，你将发现得到3.2或更大的 t 值的概率约为0.000 01，即非常小。这就是我们所测到这个 t 统计量的 p 值。这一精确的 t 统计量显著水平，比起常用并任意固定的任何一个显著水平如1%、5%或10%等，都要小得多。事实上，如果我们真的使用刚才算的 p 值来拒绝“受教育程度的真实系数是0.5”的虚拟假设，那么我们犯第Ⅰ类错误的概率只有十万分之一！

我们在前面曾指出，如果数据不支持虚拟假设，则在虚拟假设下得到的 $|t|$ 值将会很“大”，得到这样一个 $|t|$ 值的 p 值因而就很“小”。换言之，对于给定样本容量，随着 $|t|$ 的增加，p 值会不断下降，我们也就越来越有信心拒绝虚拟假设。

---

① Jan Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971, pp. 126-127.

p 值和显著性水平 α 是怎样一种关系呢？如果我们养成一种习惯，把 α 固定在一个检验统计量（如 t 统计量）的 p 值上，这两个值就没有任何矛盾。换句话说，**与其人为地把 α 固定在某一水平，不如干脆选取检验统计量的 p 值**。让读者自己决定是否在给定的 p 值水平上拒绝虚拟假设好了。如果在一项应用中，检验统计量的 p 值正好是 0.145 或 14.5%，并且如果读者想要在这一精确显著水平上拒绝虚拟假设，就让他这样做好了。采用一个 14.5%犯（第Ⅰ类）错误（犯拒绝了真实的虚拟假设的错误）的机会，并没有任何过错。同样，在我们的工资—受教育程度的例子中，研究者如要采用一个约为 0.02%的 p 值，而不愿接受比万分之二更大的犯错误机会，也没有什么过错。毕竟有一些研究者是风险偏好者，而另一些则是风险厌恶者！

在本书的其余部分，我们一般都标出给定检验统计量的 p 值。读者可以把 α 固定在某一水平上，并在 p 值小于 α 时拒绝虚拟假设。这是他们的选择自由。

□ 统计显著性与实际显著性

回到我们的例 3.1，看方程（3.7.1）中给出的回归结果。这个回归将美国 1960—2005 年间的个人消费支出与国内生产总值相联系，这两个变量都是以 2000 年的十亿美元为单位度量的。

我们从这个回归中看到，边际消费倾向（MPC）即收入（用 GDP 度量）每增加 1 美元导致消费增加的数量约为 0.72 美元或 72 美分。利用方程（3.7.1）中的数据，读者很容易验证 MPC 的 95%置信区间为（0.712 9，0.730 6）。（注意，由于在这个问题中自由度为 44，所以我们得不到这些自由度下的精确 t 值，因此你可以利用2-t 经验法则来计算这个 95%置信区间。）

假设有人坚持认为真正的 MPC 是 0.74，这个假设与 0.72 不同吗。如果我们严格地使用上述构造的置信区间，二者的确是不同的。

但是我们的这一发现有什么实际或实质显著性呢？也就是说，如果我们把 MPC 当作 0.74 而不是 0.72，会有什么差别呢？两个 MPC 之间 0.02 的差别有什么实际重要意义？

问题的回答有赖于我们要用这些估计值来做什么。例如，从宏观经济学中我们得知，收入乘数是 1/(1－MPC)。因此，如果 MPC 是 0.72，这个乘数就是 3.57；但如果 MPC 是 0.74，乘数就是 3.84。这就是说，如果政府打算增加 1 美元的开支以拯救经济萧条，如果 MPC 是 0.72，收入将最终增加 3.57 美元；而如果 MPC 是 0.74，则最终增加 3.84 美元。而这一差异对于经济的复苏也许很重要。

全部讨论的要点在于：不要把统计上的显著性和实际上或经济上的显著性混同起来，正如戈德伯格（Goldberger）所说：

> 当人们设定一个虚拟假设，比方说，H_0：$\beta_i=1$ 时，其用意很可能是说 β_i 接近 1，且接近到这样一个程度，以致为了一切实际目的，都可以看作它就是 1。然而，1.1 是否“实际上无异于”1.0？这是一个经济学问题，而不是统计学问题。我们不能依靠假设检验来解决这个问题。因为检验统计量 [$t=$]

$(b_j-1)/\hat{\sigma}_{bj}$ 用标准误做单位来衡量所估计的系数，而标准误并不是衡量经济参数 β_j-1 的一个有意义的单位。一个好的办法也许是把显著性一词当作统计学概念，而在经济学概念中使用“重要性”一词。①

戈德伯格的论点是重要的。当样本容量变得非常大时，统计显著性的问题会变得黯然失色，而经济显著性的问题会变得至关重要。的确，在样本非常大的情况下，几乎任何虚拟假设都一定会被拒绝，点估计就成为唯一可研究的问题。

□ 假设检验的置信区间方法和显著性检验方法的选择

在大多数应用性的经济分析中，虚拟假设的建立犹如一个稻草人的竖立。经验研究工作的目的，是要把它打倒，即拒绝这个虚拟假设。以我们的消费—收入关系为例，虚拟假设 MPC 即 $\beta_2=0$ 是显然荒谬的，可是我们却常用它来把经验研究的结果写得更引人入胜。现任著名期刊的编辑都发觉发表一篇不拒绝虚拟假设的经验性文章是不足以激动人心的。发现 MPC 在统计上异于零，多少要比发现它等于 0.7 更值得作为新闻报道！

因此，J·布拉德福康·德·朗（J. Bradford De Long）和凯文·兰（Kevin Lang）辩称，对经济学家而言，较好的做法是，

> ……集中讨论系数的大小并报告其置信水平，而不去提显著性检验。如果全部或几乎全部虚拟假设都是错误的，那么，讨论一个估计值是否无异于它在虚拟假设下的预测值，就是无意义的，相反，我们也许想探明什么模型是较好的近似，这就需要知道为经验估计所排除的参数值域。②

简言之，这些作者认为，置信区间方法优于显著性检验方法。读者不妨把这一忠告铭记在心。③

5.9 回归分析与方差分析

本节我们从方差分析的视角来研究回归分析，从而向读者介绍一种对待统计推断问题更有启发意义和补充作用的方法。

---

① 参见 Arthur S. Goldberger, *A Course in Econometrics*, Harvard University Press, Cambridge, Massachusetts, 1991, p. 240。注意 b_j 是 β_j 的 OLS 估计量，而 $\hat{\sigma}_{bj}$ 是它的标准误。赞同的观点还见于 D. N. McCloskey, “The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests,” *American Economic Review*, vol. 75, 1985, pp. 201-205。也可参见 D. N. McCloskey and S. T. Ziliak, “The Standard Error of Regression,” *Journal of Economic Literature*, vol. 37, 1996, pp. 97-114。

② 见第 123 页注释①所引的他们的论文，p. 1271。

③ 至于多少有些不同的观点，参见 Carter Hill, William Griffiths, and George Judge, *Undergraduate Econometrics*, Wiley & Sons, New York, 2001, p. 108。

在第 3 章 3.5 节中，我们曾导出如下恒等式：

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2 \tag{3.5.2}$$

即 TSS=ESS+RSS，它把总平方和（TSS）分解为两个构成部分：解释平方和（ESS）与剩余平方和（RSS）。对 TSS 的这些构成部分的研究从回归的观点就叫做**方差分析**（analysis of variance，ANOVA）。

同任一个平方和联系在一起的是它所依据的自由度，即独立观测值的个数。因为在计算样本均值 $\bar{Y}$ 时，我们失去 1 个自由度，故 TSS 有 $n-1$ 个自由度。RSS 有 $n-2$ 个自由度。（为什么?）（注：仅对有截距 β_1 的双变量回归模型才是对的。）ESS 有1 个自由度（也仅对双变量情形才正确），这是因为当 $\sum x_i^2$ 为已知时 ESS= $\hat{\beta}_2^2 \sum x_i^2$ 仅是 $\hat{\beta}_2$ 的函数。

把各项平方和及其相应的自由度在表 5—3 中列出，就成为 AOV 表的标准形式，有时又称 ANOVA 表。给定表 5—3 中的条目，现考虑以下变量：

$$\begin{aligned} F &= \frac{\text{MSS of ESS}}{\text{MSS of RSS}} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2} \end{aligned} \tag{5.9.1}$$

表 5—3　　双变量回归模型的 ANOVA 表

变异来源	SS*	df	MSS†
回归部分（ESS）	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
剩余部分（RSS）	$\sum \hat{u}_i^2$	$n-2$	$\frac{\sum \hat{u}_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n-1$	

注：*SS 指平方和。

†MSS 指均方和，得自 SS 除以其自由度。

如同我们在 CNLRM 所做的一样，假定干扰项 u_i 是正态分布的且虚拟假设 H_0：$\beta_2=0$，就可证明方程（5.9.1）中的 F 服从自由度为 1 和 $n-2$ 的 F 分布。（证明见附录 5A 第 5A.3 节。F 分布性质的讨论见附录 A。）

上述 F 有什么用处呢？可以证明[①]：

$$E(\hat{\beta}_2^2 \sum x_i^2) = \sigma^2 + \beta_2^2 \sum x_i^2 \tag{5.9.2}$$

以及

$$E\frac{\sum \hat{u}_i^2}{n-2} = E(\hat{\sigma}^2) = \sigma^2 \tag{5.9.3}$$

① 证明见 K. A Brownlee，*Statistical Theory and Methodology in Science and Engineering*，John Wiley & Sons，New York，1960，pp. 278-280。

（注意出现在这些方程右端的 β_2 和 σ^2 是真实的参数。）因此，若 β_2 确实是零，方程（5.9.2）和（5.9.3）两者都给出相同的真实 σ^2 的估计，这时解释变量 X 与 Y 没有任何线性影响，Y 的全部变异均由随机干扰项 u_i 来解释。而另一方面，若 β_2 不是零，则方程（5.9.2）和（5.9.3）将有所不同，从而 Y 的部分变异将归因于 X。于是，（5.9.1）的 F 提供了对虚拟假设 H_0：$\beta_2=0$ 的一个检验。由于此方程中的每一个量都可从已有的样本算得，所以这个 F 就为检验虚拟假设真实 $\beta_2=0$ 提供了一个检验统计量。我们所需做的无非就是算出 F，再用它同从 F 表在选定显著水平上读出的 F 临界值相比较，或者是查找所算 F 统计量的 p 值。

为便于说明，继续使用上述说明性例子。此例的 ANOVA 表见于表 5—4。我们看到计算的 F 值为 108.302 6。这个 F 值对应于 1 和 11 个自由度的 p 值不能从附录 D 给出的 F 表读出。但利用电子统计表可以显示 p 值是 0.000 000 1，确实是一个极小的概率。如果你决定选择假设检验的显著性水平方法，并把 α 固定在 0.01 或 1% 的水平上，你便能看到所算的 108.302 6 这一 F 值在此水平上明显是显著的。所以，如果我们拒绝虚拟假设 $\beta_2=0$，犯第Ⅰ类错误的概率就非常小。从一切实际意义来考虑，我们的样本都不可能来自一个 β_2 为零的总体，从而我们能够很有信心地做出受教育程度 X 对工资 Y 有影响的结论。

表 5—4　　工资—受教育程度一例的 ANOVA 表

变异来源	SS	df	MSS	
回归部分（ESS）	95.425 5	1	95.425 5	$F=\frac{95.425\ 5}{0.881\ 1}$
剩余部分（RSS）	9.692 8	11	0.881 1	$=108.302\ 6$
TSS	105.118 3	12		

翻到附录 5A.1 的定理 5.7，该定理指出，自由度为 k 的 t 值的平方是一个分子自由度为 1 和分母自由度为 k 的 F 值。对本例来说，如果假设 H_0：$\beta_2=0$，则由方程（5.3.2）容易验证，估计 t 值是 10.41，这个 t 值有 11 个自由度。在同样的虚拟假设下，F 值曾算出是 108.302 6 且有自由度 1 和 11。因此，在不计四舍五入误差的情况下，应该有 $(10.342\ 8)^2$ 等于这个 F 值。

因此，t 检验和 F 检验是检验虚拟假设 $\beta_2=0$ 的两个互为补充的备选方法。若果真如此，为什么不仅仅使用 t 检验就够了，还要麻烦到用 F 检验及伴随的方差分析呢？对于双变量模型，确实不需要用 F 检验。但当我们考虑多元回归问题时，我们将看到 F 的一些有趣的应用，使得它成为检验统计假设的非常有用和有效的方法。

5.10　回归分析的应用：预测问题

根据表 3—2 的样本数据，我们曾得到如下样本回归：

$$\hat{Y}_i = -0.014\ 4 + 0.724\ 0X_i \tag{3.6.1}$$

其中 $\hat{Y}_i$ 是给定 X 下真实 $E(Y_i)$ 的估计量。这一描述**历史回归**（historical regression）能有什么用处呢？一个用途是“预测”或“预报”给定受教育程度 X 下未来的平均工资 Y。现有两种预测：(1) 对应于选定的 X，比方说 X_0，预测 Y 的条件均值，也就是预测总体回归线本身的点（见图 2—2），以及 (2) 预测对应于 X_0 的 Y 的个别值。我们将把这两种预测分别称为**均值预测**（mean prediction）和**个值预测**（individual prediction）。

□ 均值预测[①]

为便于说明，假定 $X_0=20$，我们要预测 $E(Y \mid X_0 = 20)$。可以表明，历史回归方程（3.6.2）给出这个均值预测的点估计如下：

$$\begin{aligned}\hat{Y}_0 &= \hat{\beta}_1 + \hat{\beta}_2 X_0 \\ &= -0.014\ 4 + 0.724\ 0 \times 20 \\ &= 14.465\ 6\end{aligned} \tag{5.10.1}$$

其中 $\hat{Y}_0$ 为 $E(Y \mid X_0)$ 的估计量。可以证明，这个点预测量是一个最优线性无偏估计量。

既然 $\hat{Y}_0$ 是一个估计量，就可能不同于它的真值。二者之差将大致给出预测或预报的误差。为了评估这个误差，我们需要求出 $\hat{Y}_0$ 的抽样分布。附录 5A 第 5A.4 节表明，方程（5.10.1）中的 $\hat{Y}_0$ 是正态分布的，其均值为 $(\beta_1 + \beta_2 X_0)$，而方差由下式给出：

$$\operatorname{var}(\hat{Y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \tag{5.10.2}$$

将未知的 σ^2 代以它的无偏估计量 $\hat{\sigma}^2$，就可推知变量：

$$t = \frac{Y_0 - (\beta_1 + \beta_2 X_0)}{\operatorname{se}(\hat{Y}_0)} \tag{5.10.3}$$

服从 $n-2$ 个自由度的 t 分布。因而 t 分布可用来推导真实 $E(Y_0 \mid X_0)$ 的置信区间，并且用惯常的方式去检验关于它的假设，即：

$$\Pr\left[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2}\operatorname{se}(\hat{Y}_0) \leqslant \beta_1 + \beta_2 X_0 \leqslant \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2}\operatorname{se}(\hat{Y}_0)\right] = 1 - \alpha \tag{5.10.4}$$

其中 $\operatorname{se}(\hat{Y}_0)$ 由方程（5.10.2）求得。

对于我们的数据（见表 3—2），有

$$\begin{aligned}\operatorname{var}(\hat{Y}_0) &= 0.893\ 6 \times \left[\frac{1}{13} + \frac{(20-13)^2}{182}\right] \\ &= 0.382\ 6\end{aligned}$$

① 关于各个命题的证明，见附录 5A 第 5A.4 节。

以及

$$\text{se}(\hat{Y}_0) = 0.618\ 5$$

因此，真实 $E(Y \mid X_0) = \beta_1 + \beta_2 X_0$ 的95%置信区间由下式给出：

$$14.465\ 6 - 2.20\ 1 \times 0.618\ 5 \leqslant E(Y_0 \mid X = 20) \leqslant 14.465\ 6 + 2.20\ 1 \times 0.618\ 5$$

即

$$13.104\ 3 \leqslant E(Y_0 \mid X = 20) \leqslant 15.826\ 0 \qquad (5.10.5)$$

这就是说，给定 $X_0 = 20$，在重复抽样中，每100个类似于方程（5.10.5）的区间中，将有95个包含着真实的均值；真实均值的单个最优估计，当然是点估计值14.465 6。

如果我们对表3—2中的每一个 X 值，都求出类似方程（5.10.5）的95%置信区间，把这些区间的端点联结起来，我们就得到如图5—6所示的一个关于总体回归函数的所谓**置信带**（confidence interval）或**置信域**（confidence band）。

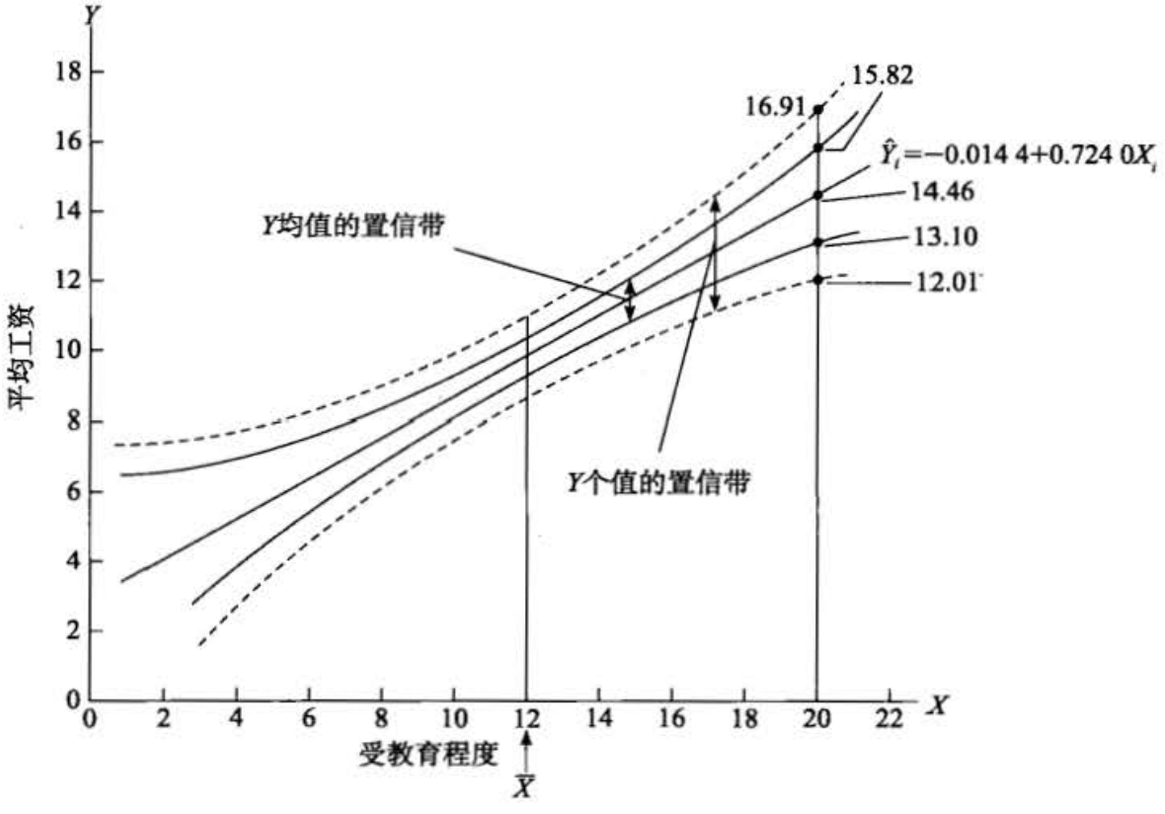


图5—6　Y 均值与 Y 个值的置信带（域）

□ 个值预测

如果我们的兴趣在于预测对应于给定 X 值（比方说 X_0）的单个 Y 值（Y_0），那么，如附录5A第5A.3节所示，Y_0 的一个最优线性无偏估计量仍由方程（5.10.1）给出，但是它的方差如下

$$\text{var}(Y_0 - \hat{Y}_0) = E[Y_0 - \hat{Y}_0]^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \qquad (5.10.6)$$

可以进一步证明 Y_0 也服从正态分布，其均值和方差分别由方程（5.10.1）和方程（5.10.6）给出。用 $\hat{\sigma}^2$ 代替未知的 σ^2，即推出下式也服从 t 分布：

$$t = \frac{Y_0 - \hat{Y}_0}{\text{se}(Y_0 - \hat{Y}_0)}$$

因此 t 分布可用来对真实 Y_0 进行推断。继续上述例子，我们看到，Y_0 的点预测和 $\hat{Y}_0$ 的点预测一样，都是 14.465 6。但它的方差是 1.235 7（读者可验证这一计算）。由此可见，对应于 $X_0=20$ 的 Y_0 的 95%置信区间是：

$$12.019\ 0 \leqslant E(Y_0 \mid X = 20) \leqslant 16.912\ 2 \tag{5.10.7}$$

拿此区间同方程（5.10.5）相比，即看出 Y_0 的置信区间比 Y_0 均值的置信区间要宽。（为什么?）以表 3—2 所给的 X 值为条件计算类似于方程（5.10.7）的诸多置信区间，联结起来就得到对应于这些 X 值的单个 Y 值的 95%置信带。图 5—6 同时展示了对应于同样 X 值的 Y 和 $\hat{Y}_0$ 的置信带。

注意图 5—6 展示的置信带的一个重要特点。当 $X_0=\overline{X}$ 时这些置信带的宽度达到最小。（为什么?）随着 X_0 远离 $\overline{X}$，这个宽度急剧地变大。（为什么?）这种变化说明历史样本回归线的预测能力随着 X_0 越来越远离 $\overline{X}$ 而显著下降。**因此，当 X_0 远离 $\overline{X}$ 时，人们凭借“外推”历史回归线来预测对应于给定 X_0 的 $E(Y \mid X_0)$ 或 Y_0 时，必须保持高度警觉。**

5.11 报告回归分析的结果

报告回归分析的结果有许多方式，本书将采用以下方式，这里仍用第 3 章的工资—受教育程度的例子进行说明：

$$\begin{aligned}
&\hat{Y}_i = -0.014\ 4 + 0.724\ 0X_i \\
&\text{se} = (0.931\ 7) \quad (0.070\ 0) \qquad r^2 = 0.906\ 5 \\
&t = (-0.015\ 4) \quad (10.342\ 8) \qquad \text{df} = 11 \\
&p = (0.987) \qquad (0.000) \qquad F_{1,11} = 108.30
\end{aligned} \tag{5.11.1}$$

方程（5.11.1）中第一组括号内的数字代表回归系数的估计标准误，第二组数字代表在每个回归系数的真实总体值为零的虚拟假设下由方程（5.3.2）计算出来的 t 估计值（例如 10.342 8=0.724 0/0.070 0），而第三组数字代表估计的 p 值。比如，当自由度为 11 时，得到一个等于 10.342 8 或更大的 t 值的概率是 0.000 09，它在实践中可视为 0。

把这些估计的 t 系数的 p 值显示出来，我们就能马上看到每一个 t 估计值的精确的显著性水平。例如，在真实总体截距值为零（即受教育程度对平均工资没有影响）的虚拟假设下，得到一个大到 10.342 8 或更大 t 值的精确概率（即 p 值）实践中几乎为 0。记得 p 值越小，我们犯拒绝正确虚拟假设的错误的概率就越低。

我们在前面曾指出 F 和 t 这两个统计量之间有一种内在联系，即 $F_{1,k} = t_k^2$。在真实 $\beta_2 = 0$ 的虚拟假设下，方程（5.11.1）表明，F 值为 108.30（对于 1 个分子自由度和 11 个分母自由度），而 t 值约为 10.34（有 11 个自由度）；如同预料，在不考虑进位误差的情况下，前一数值正好是后一数值的平方。这个问题的 ANOVA 表已经在前面讨论过。

5.12 评价回归分析的结果

在引言的表 I—4 中，我们对计量经济模型的建立作过一个简要的剖析。现在，我们又在方程（5.11.1）中给出了我们的工资—受教育程度一例的回归分析结果，我们不免要问，这个模型的拟合效果如何。这个拟合的模型有多“好”？为了回答这个问题，需要有一些准则。

第一，所估系数的符号是否与理论或事前预期相一致？先验地说，工资—受教育程度一例中的 β_2 理应为正。在本例中的确如此。第二，如果理论上认为这个关系式不仅是正的，而且是统计上显著的，那么在本例中是这样的吗？如 5.11 节中所讨论的那样，受教育程度的系数不仅是正的，而且统计上显著地异于零，t 估计值的 p 值极小，同样的评价也适用于截距系数。第三，回归模型在多大的程度上解释了平均工资的变异？可以用 r^2 来回答此问题。本例中 r^2 约为 0.90，考虑到 r^2 最多只能大到 1，这个值也算很高了。

如此看来，为了解释平均工资，我们选用的模型算是够好的了。但在我们结束讨论之前，我们还想看看我们的模型是否满足 CNLRM 的假定。因为模型明显如此简单，所以我们现在不去审查这些假定。但有一个假定是我们想要查对的，就是关于干扰项 u_i 的正态性。回想一下，前面所用的 t 检验和 F 检验都要求误差项服从正态分布。否则，在小样本或有限样本中，检验的程序将是无效的。

□ 正态性检验

虽然文献中有多种正态性检验，但我们只想讨论三种：（1）残差直方图；（2）正态概率图；（3）**雅克-贝拉检验**（Jarque-Bera test）。

残差直方图。残差直方图是用于了解随机变量概率密度函数（PDF）形状的一个简单图示。在水平轴上，我们将所关注变量的值（比如 OLS 残差）分成适当的区间，在每个区间里，我们做垂直的矩形，并让矩形的高等于落在该区间内的观测次数（即频数）。如果你从心里在直方图上估画一条钟形的正态分布曲线，那么你就对正态 PDF 近似是否适当有些认识。对于工资—受教育程度的回归，残差直方图如图 5—7 所示。

此图表明，残差不是完美的正态分布；对于一个正态分布变量，其偏态值（对

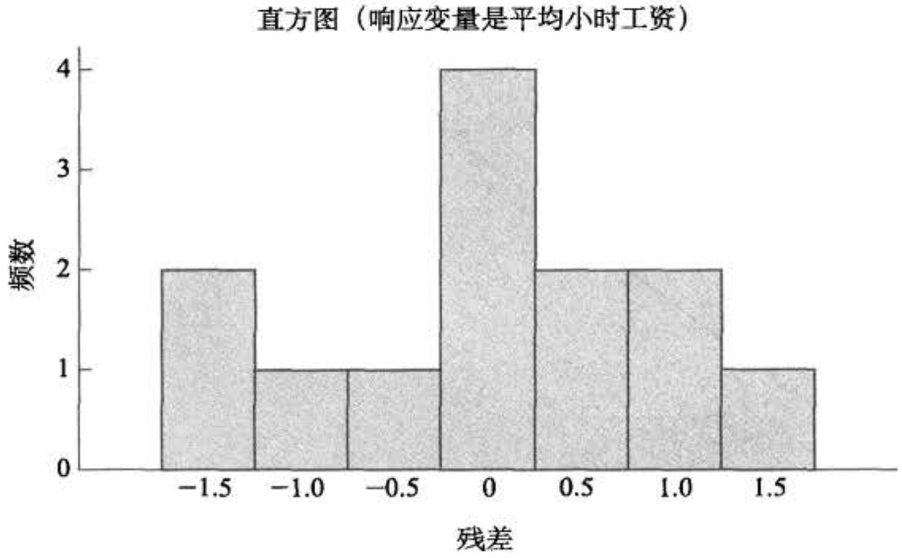


图 5—7　工资—受教育程度一例数据的残差直方图

称性的一个度量指标）应该为 0，而峰态值（它度量了一个正态分布的高矮）应该为 3。

但作为检验正态假定的一个简易方法，绘制回归残差的直方图总是一个好的做法。

正态概率图。要研究随机变量的 PDF 的形状，一个相对简单的图示法就是，利用正态概率纸（一种专门设计的坐标纸）做出**正态概率图**（normal probability plot，NPP）。在水平轴或 X 轴上，我们描绘所关注变量的值（如 OLS 残差 $\hat{u}_i$），而在纵轴或 Y 轴上，我们标出这个变量服从正态分布时的期望值。因此，如果这个变量实际上来自正态总体，那么 NPP 将近似为一条直线。从工资—受教育程度回归所得到残差的 NPP 如图 5—8 所示，它是从 MINITAB 软件（15 版）得到的。如前所述，如果 NPP 的拟合线近似为一条直线，那人们就可以断定这个变量是正态分布的。在图 5—8 中，由于一条直线相当好地拟合了这些数据，所以我们说明性例子中的残差是渐近正态分布的。

MINITAB 还给出**安德森-达林正态性检验**（Anderson-Darling normality test），称 A^2 **统计量**（A^2 statistic）。其背后的虚拟假设是，所考虑的变量是正态分布的。如图 5—8 所示，就我们的例子而言，计算出来的 A^2 统计量为 0.289。得到这样一个 A^2 值的 p 值为 0.558，这是相当高的。因此，我们不能拒绝此例中残差为正态分布的假设。顺便提一句，图 5—8 表明，此（正态）分布的两个参数均值近似为 0，标准差约为 0.898 7。

正态性的雅克-贝拉（JB）检验。[①] 正态性的 JB 检验是一项渐近或大样本检验。它仍以 OLS 残差为依据。此检验先计算 OLS 残差的**偏态**（skewness）和**峰态**（kur-

① 见 C. M. Jarque and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, vol. 55, 1987, pp. 163-172。

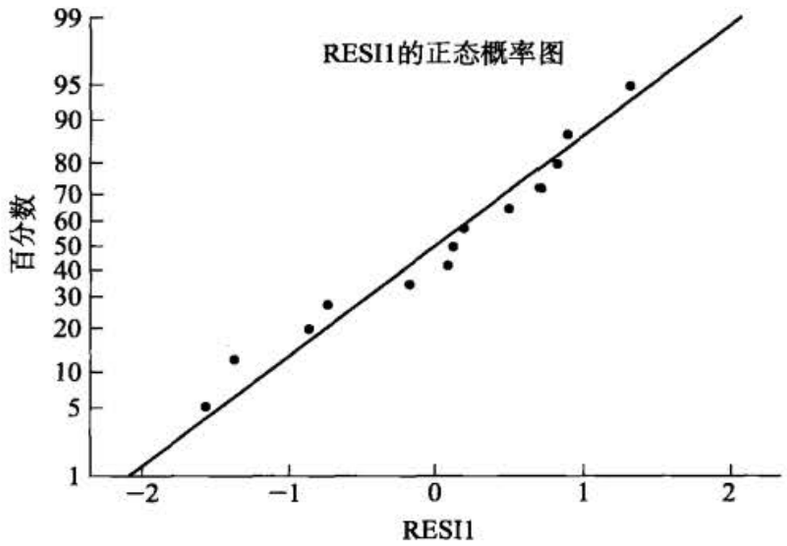


图 5—8　工资—受教育程度的残差

tosis)（附录 A 中有所描述），再使用下列检验统计量：

$$\text{JB}= n\left[\frac{S^2}{6}+\frac{(K-3)^2}{24}\right] \tag{5.12.1}$$

其中 n=样本容量，S=偏态系数，K=峰态系数。对于一个正态分布变量，$S=0$，而 $K=3$。因此，正态性的 JB 检验是对 $S=0$ 和 $K=3$ 这一联合假设进行检验。

在残差为正态分布的虚拟假设下，雅克和贝拉证明了方程（5.12.1）所给的 JB 统计量渐近地（即在大样本中）服从自由度为 2 的 χ^2 分布。如果在一项应用中算出来的 JB 统计量的 p 值充分地小，就可拒绝残差为正态分布的假设。但如果 p 值合理地高，就不要拒绝正态性假定。

再看我们的例子，工资—受教育程度回归中 JB 统计量的估计值为 0.828 6。得到 JB 统计量高达 0.828 6 的 p 值约为 0.66 或 66%，所以，我们不能拒绝本例中残差为正态分布的虚拟假设。这个概率相当高。注意，尽管我们的回归有 13 个观测，但这些观测得自于由 528 个观测构成的样本，这个样本容量已经足够大了。

□ 模型适宜性的其他检验

除了误差项的正态性假定之外，记得 CNLRM 还做了许多其他假定。随着我们进一步分析我们的计量经济理论，我们还将考虑模型适宜性的若干其他检验（见第 13 章）。在这之前，请记住，我们的回归模型是建立在一些不一定总是成立的简化假定之上的。

一个总结性的例子

让我们回到关于印度食物支出的例 3.2。利用方程（3.7.2）中给出的数据并采用方程（5.11.1）的格式，我们得到如下支出方程

$$\begin{aligned}
\widehat{\text{FoodExp}_i} &= 94.208\ 7 \quad + \quad 0.436\ 8\ \text{TotalExp}_i \\
\text{se} &= (50.856\ 3) \quad (0.078\ 3) \\
t &= (1.852\ 4) \quad (5.577\ 0) \\
p &= (0.069\ 5) \quad (0.000\ 0)^* \\
r^2 &= 0.369\ 8 \quad \text{df} = 53 \\
F_{1,53} &= 31.103\ 4 \quad (p\text{ 值} = 0.000\ 0)^*
\end{aligned} \tag{5.12.2}$$

其中 * 表示极小。

首先，让我们解释这个回归，食物支出与总支出之间存在着预期的正相关。如果总支出增加 1 卢比，食物支出平均增加约 44 派沙。如果总支出为零，食物支出约为 94 卢比。当然，对截距的这种机械解释可能没有多大经济意义。约为 0.37 的 r^2 值意味着，食物支出的变异中有 37%可由收入的代理变量总支出来解释。

假设我们想检验食物支出与总支出之间没有关系的虚拟假设，即真正的斜率系数 $\beta_2=0$。β_2 的估计值为 0.436 8。如果虚拟假设正确，那么得到 0.436 8 这样一个值的概率是多大？在这个虚拟假设下，我们从方程 (5.12.2) 中观测到，t 值为 5.577 0，得到这样一个 t 值的 p 值实际上为零。换句话说，我们完全可以拒绝这个虚拟假设。但若虚拟假设是 $\beta_2=0.5$，情况又怎样呢？利用 t 检验，我们得到

$$t = \frac{0.436\ 8 - 0.5}{0.078\ 3} = -0.807\ 1$$

得到 $|t| = 0.807\ 1$ 的概率大于 20%。因此我们不能拒绝真正的 β_2 为 0.5 的假设。

注意，在真正的斜率系数为零的虚拟假设下，F 值如方程 (5.12.2) 所示为 31.103 4。在同样的假设下，我们得到的 t 值为 5.577 0。它的平方就是 31.102 9，约等于 F 值，这又再次表明了 t 统计量与 F 统计量之间的密切关系。(注：F 统计量的分子自由度必须为 1，这里正是如此。)

利用回归的估计残差，我们对误差项的概率分布有何见解？图 5—9 给出了这方面的信息。如图 5—9 所示，食物支出回归的残差看起来是对称分布的。应用雅克-贝拉检验表明，JB 统计量约为 0.257 6，在正态性假定下，得到这样一个统计量的概率约为 88%。因此我们不能拒绝误差项为正态分布的假设。但要记住，55 个观测的样本容量可能不算太大。

至于构造回归系数的置信区间并得出正态概率图，以及进行均值和个值预测，则留给读者自己完成。

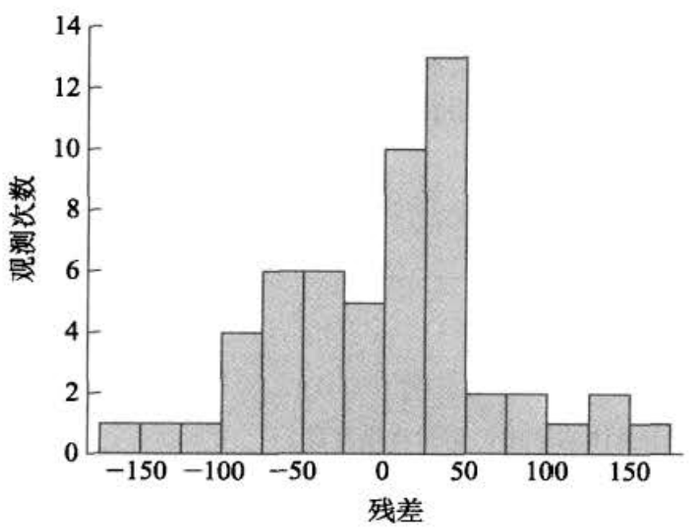


图 5—9　食物支出回归的残差

要点与结论

1. 估计与假设检验是经典统计学的两个主要分支。在第 3 章和第 4 章讨论了估计问题之后，本章讨论假设检验的问题。

2. 假设检验要回答这样的问题：一个给定的发现是否与声称的假设相符？

3. 为回答上述问题，有两个互为补充的方法：**置信区间方法**与**显著性检验方法**。

4. 置信区间方法建立在**区间估计**的概念之上。一个区间估计量是指一个区间或变化域的构造，要使得它以预定的概率把未知参数的真值包含在其界限内。如此构造的区间被称为**置信区间**，这个区间常用百分数比如说 90%或 95%的形式来表述。置信区间对未知参数的取值提供了一个可信假设集。如果虚拟假设值落入置信区间，就不拒绝假设；如果它落在此区间之外，就可拒绝虚拟假设。

5. 在**显著性检验**程序中，我们找出一个**检验统计量**，并研究它在虚拟假设下的抽样分布。通常这个检验统计量都服从一个明确定义的概率分布，比如正态、t、F 或 χ^2 分布。一旦从现有的数据算出某个检验统计量（如 t 统计量），就容易求出它的 p 值。这个 p 值给出在虚拟假设下得到所估算的检验统计量的精确概率。如果 p 值小，就可拒绝虚拟假设；但如果 p 值大，就不可拒绝。什么才算小的或大的 p 值应由研究者来决定。在选择 p 值时，研究者要切实考虑犯**第Ⅰ类错误**和**第Ⅱ类错误**的概率。

6. 在实践中，犯**第Ⅰ类错误**的概率 α 被任意选定为 1%、5%或 10%，要仔细考虑。较好的做法是引用检验统计量的 ***p* 值**。而且，不要把统计显著性和实际显著性相混淆。

7. 当然，假设检验事先就认定，选用来做经验分析的模型不违背经典正态线性回归模型中的任一个或多个假定，所以是适宜的。因此，应把模型适宜性的检验放在假设检验之前进行。本章介绍的一个模型适宜性检验就是**正态性检验**，借以发现误差项是否服从正态分布。因为在小的或有限的样本中，t、F 和 χ^2 检验都需要有正态性假定，所以对此假定正式地加以核实就至关重要。

8. 如果模型在实践中被认为是适宜的，就可用于预报。但在预报回归子的未来值时，切勿超出回归元的样本取值范围太远。否则，预测误差会急剧增大。

习　　题

问答题

5.1　判断下述命题的正误。并给出具体的理由，切勿含糊其辞。

a. 本章所讨论的显著性 t 检验要求估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的抽样分布是正态分布。

b. 即使 CLRM 中的干扰项不是正态分布的，OLS 估计量仍然是无偏的。

c. 如果回归模型中没有截距项，u_i 估计值（$=\hat{u}_i$）的总和将不为零。

d. p 值和检验统计量的尺度指的是一回事。

e. 在一个含有截距的回归模型中，残差的总和必定为零。

f. 如果一个虚拟假设不被拒绝，它就是真实的。

g. σ^2 的值越大，方程（3.3.1）所给的 $\hat{\beta}_2$ 的方差也越大。

h. 一个随机变量的条件均值和无条件均值是一样的。

i. 在双变量 PRF 中，如果斜率系数 β_2 是零，则截距 β_1 由样本均值 $\bar{Y}$ 来估计。

j. 如果 X 对 Y 无影响，条件方差 $\text{var}(Y_i \mid X_i) = \sigma^2$ 和 Y 的无条件方差 $\text{var}(Y) = \sigma_Y^2$ 将是一样的。

5.2 对方程（3.7.2）所给的回归模型建立像表 5—4 那样的 ANOVA 表，并检验印度的食物支出与总支出无关的假设。

5.3 参考方程（3.7.3）中所给出的手机需求回归。

a. 在 5%的显著水平上，截距系数估计值显著吗？你进行检验的虚拟假设是什么？

b. 在 5%的显著水平上，斜率系数估计值显著吗？其背后的虚拟假设是什么？

c. 构造真实斜率系数的 95%置信区间。

d. 如果人均收入是 9 000 美元，手机需求的平均预测值是多少？这个预测值的 95%置信区间是什么？

5.4 令 ρ^2 代表真实的总体判定系数，假定你想检验假设：$\rho^2 = 0$，用文字说明你会怎样检验此假设。提示：利用方程（3.5.11），也可参看习题 5.7。

5.5 现代投资分析中所谓的**特征线**（characteristic line）就是得自以下模型的回归线：

$$r_{it} = \alpha_i + \beta_i r_{mt} + u_i$$

其中 r_{it} = 第 i 种证券在时间 t 的回报率；

r_{mt} = 市场组合证券在时间 t 的回报率；

u_t = 随机干扰项。

在此模型中，β_i 被称为第 i 种证券的**β 系数**（beta coefficient），是对证券的市场（或系统）风险的一种度量。①

福格勒（Fogler）和加纳帕赛（Ganapathy）根据 1956—1976 年间的 240 个月回报率，算得 IBM 股票相对于芝加哥大学研制的市场组合证券指数的特征线如下②：

$$\hat{r}_{it} = 0.726\,4 + 1.059\,8\, r_{mt} \qquad r^2 = 0.471\,0$$

$$\text{se} = (0.300\,1)\ (0.072\,8) \qquad \text{df} = 238$$

$$F_{1,238} = 211.896$$

a. β 系数大于 1 的证券被称为易波动或进攻型证券。问在此研究期间 IBM 是易波动证券吗？

b. 问截距系数是否显著地异于零？如果是，它的实际意义何在？

5.6 方程（5.3.5）还可写为：

$$\Pr[\hat{\beta}_2 - t_{\alpha/2}\,\text{se}(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + t_{\alpha/2}\,\text{se}(\hat{\beta}_2)] = 1 - \alpha$$

也就是说，弱不等式（≤）可代之以强不等式（<）。为什么？

5.7 费希尔曾推导出由方程（3.5.13）定义的相关系数的抽样分布。如果假定变量 X 和 Y

---

① 参看 Haim Levy and Marshall Sarnat, *Portfolio and Investment Selection: Theory and Practice*, Prentice-Hall International, Englewood Cliffs, NJ, 1984, Chap. 12。

② H. Russell Fogler and Sundaram Ganapathy, *Financial Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1982, p. 13.

是联合正态分布的，即如果它们来自双变量正态分布（见附录4A，习题4.1），则在总体相关系数 ρ 为零的假定下，可以证明 $t=r\sqrt{n-2}/\sqrt{1-r^2}$ 服从自由度为 $n-2$ 的 t 分布。[①] 试说明这个 t 值等同于在虚拟假设 $\beta_2=0$ 下由方程（5.3.2）给出的 t 值，进而证明在相同的虚拟假设下 $F=t^2$。（见5.9节。）

5.8 考虑如下回归输出结果[②]：

$$\hat{Y}_i = 0.203\ 3 + 0.656\ 0X_i$$

$$\text{se} = (0.097\ 6) \quad (0.196\ 1)$$

$$r^2 = 0.397 \quad \text{RSS} = 0.054\ 4 \quad \text{ESS} = 0.035\ 8$$

其中 Y=1972年妇女的劳动参与率（LFPR），X=1968年妇女的劳动参与率。这个回归结果得自于美国19个城市构成的一个数据样本。

a. 你如何解释这个回归结果？

b. 在对立假设为 H_1：$\beta_2>1$ 的前提下，检验 H_0：$\beta_2=1$ 的虚拟假设。你使用什么检验？为什么？你使用的检验所依据的假定有哪些？

c. 假设1968年的LFPR为0.58（或58%）。基于上述回归结果，1972年LFPR的均值是多少？构造这一均值预测的一个95%置信区间。

d. 你如何检验总体回归中误差项服从正态分布的虚拟假设？给出必要的计算。

实证分析题

5.9 表5—5给出了1985年50个州和哥伦比亚特区公共教师的平均工资（以美元计的年薪）和对公立学校每个学生的支出（美元）方面的数据。

为了探明公立学校中教师工资与对每个学生的支出之间是否存在某种关系，有人提出如下模型：$\text{Pay}_i = \beta_1 + \beta_2 \text{Spend}_i + u_i$，其中Pay表示教师工资，Spend表示对每个学生的支出。

表5—5　　1985年的教师平均工资与学生支出　　（单位：美元）

观测	年薪	支出	观测	年薪	支出
1	19 583	3 346	13	30 168	3 782
2	20 263	3 114	14	26 525	4 247
3	20 325	3 554	15	27 360	3 982
4	26 800	4 642	16	21 690	3 568
5	29 470	4 669	17	21 974	3 155
6	26 610	4 888	18	20 816	3 059
7	30 678	5 710	19	18 095	2 967
8	27 170	5 536	20	20 939	3 285
9	25 853	4 168	21	22 644	3 914
10	24 500	3 547	22	24 624	4 517
11	24 274	3 159	23	27 186	4 349
12	27 170	3 621	24	33 990	5 020

① 若 ρ 确实是零，费希尔曾证明，只要 X 或 Y 服从正态分布，则 r 便服从同样的 t 分布。但若 ρ 不等于零，则必须两个变量都是正态分布的。参看 R. L. Anderson and T. A. Bancroft, *Statistical Theory in Research*, McGraw-Hill, New York, 1952, pp. 87-88。

② 节选自 Samprit Chatterjee, Ali S. Hadi, and Bertram Price, *Regression Analysis by Example*, 3d ed., Wiley Interscience, New York, 2000, pp. 46-47。

续前表

观测	年薪	支出	观测	年薪	支出
25	23 382	3 594	39	22 482	3 947
26	20 627	2 821	40	20 969	2 509
27	22 795	3 366	41	27 224	5 440
28	21 570	2 920	42	25 892	4 042
29	22 080	2 980	43	22 644	3 402
30	22 250	3 731	44	24 640	2 829
31	20 940	2 853	45	22 341	2 297
32	21 800	2 533	46	25 610	2 932
33	22 934	2 729	47	26 015	3 705
34	18 443	2 305	48	25 788	4 123
35	19 538	2 642	49	29 132	3 608
36	20 460	3 124	50	41 480	8 349
37	21 419	2 752	51	25 845	3 766
38	25 160	3 429			

资料来源：National Education Association，as reported by *Albuquerque Tribune*，Nov.，7，1986.

a. 描出这些数据点并目测一条回归线。

b. 假设你想根据（a）估计上述回归模型。求参数估计值及其标准误、r^2、RSS 和 ESS。

c. 解释这个回归。它有经济意义吗？

d. 构造 β_2 的一个 95%置信区间。你会拒绝真正的斜率系数为 3.0 的假设吗？

e. 若对每个学生的支出为 5 000 美元，求 Pay 的均值和个值预测值。同样分别构造它们的 95%置信区间。

f. 你如何检验误差项的正态性假定？说明你所用的检验。

5.10　根据习题 3.20 中的数据，构造 ANOVA 表，以检验生产率与真实工资报酬之间没有关系的虚拟假设。对商业部门和非农商业部门分别做这个检验。

5.11　根据习题 1.7 中的数据：

a. 以印象为纵轴和广告支出为横轴描点。你观察到哪种关系？

b. 对数据拟合一个双变量线性回归模型合适吗？为什么？若不合适，你将用哪种类型的回归模型来拟合数据？我们有拟合这种模型的必要工具吗？

c. 假设你不描点而简单地对数据拟合一个双变量回归模型。给出通常的回归结果。留存结果等以后再看这个问题。

5.12　根据习题 1.1 中的数据：

a. 将美国的 CPI 相对加拿大的 CPI 描图，这个图说明了什么？

b. 假设你想基于加拿大的 CPI 来预测美国的 CPI，给出一个适当的模型。

c. 检验这两个 CPI 之间没有关系的假设（$\alpha=5\%$）。如果拒绝了虚拟假设，是否意味着加拿大的 CPI“导致”了美国的 CPI？为什么？

5.13　根据习题 3.22 中的数据：

a. 估计那里的两个回归，以获得通常的输出（结果），如标准误等等。

b. 检验两个回归模型的干扰项都是正态分布的假设。

c. 在黄金价格回归中，检验假设 $\beta_2=1$，即黄金价格和 CPI 之间有 1∶1 的关系（也就是说，黄金是一种完美的保值工具）。所估计的检验统计量的 p 值是多少？

d. 对 NYSE 指数回归重做（c）题。投资于股票市场是防范通货膨胀的完美保值手段吗？你检验的虚拟假设是什么？它的 p 值是多少？

e. 在黄金与股票之间，你会选择哪一种投资？你的决策依据是什么？

5.14　表 5—6 给出 1970—1983 年美国的 GNP 和 4 种不同定义的货币存量。将 GNP 对各种定义的货币作回归，并将所得结果列入表 5—7。

表 5—6　　GNP 和货币存量的四种度量

年份	GNP，十亿美元	货币存量度量，十亿美元计			
		M1	M2	M3	L
1970	992.70	216.6	628.2	677.5	816.3
1971	1 077.6	230.8	712.8	776.2	903.1
1972	1 185.9	252.0	805.2	886.0	1 023.0
1973	1 326.4	265.9	861.0	985.0	1 141.7
1974	1 434.2	277.6	908.5	1 070.5	1 249.3
1975	1 549.2	291.2	1 023.3	1 174.2	1 367.9
1976	1 718.0	310.4	1 163.6	1 311.9	1 516.6
1977	1 918.3	335.4	1 286.7	1 472.9	1 704.7
1978	2 163.9	363.1	1 389.1	1 647.1	1 910.6
1979	2 417.8	389.1	1 498.5	1 804.8	2 117.1
1980	2 631.7	414.9	1 632.6	1 990.0	2 326.2
1981	2 957.8	441.9	1 796.6	2 238.2	2 599.8
1982	3 069.3	480.5	1 965.4	2 462.5	2 870.8
1983	3 304.8	525.4	2 196.3	2 710.4	3 183.1

注：M1＝现金＋活期存款＋旅行支票＋其他支票存款（OCDS）。

M2＝M1＋隔日回购及欧洲美元＋货币市场共同基金（MMMF）余额＋货币市场存款账户＋储蓄及小额存款。

M3＝M2＋大额定期存款＋定期回购＋机构 MMMF。

L＝M3＋其他流动资产。

资料来源：*Economic Report of the President*，1985，GNP data from Table B-1，p. 232；money stock data from Table B-61，p. 303.

货币主义者或货币数量理论家声称，名义收入（即名义 GNP）主要由货币存量的数量变化决定，虽然什么是货币的“合适”定义尚无一致意见。给定上表中的结果，试考虑如下问题：

表 5—7　　GNP—货币存量回归：1970—1983 年

1)	$\widehat{\text{GNP}}_t = -787.472\,3 + 8.086\,3M_{1t}$ (77.966 4)　(0.219 7)	$r^2 = 0.991\,2$
2)	$\widehat{\text{GNP}}_t = -44.062\,6 + 1.587\,5M_{2t}$ (61.013 4)　(0.044 8)	$r^2 = 0.990\,5$
3)	$\widehat{\text{GNP}}_t = 159.136\,6 + 1.203\,4M_{3t}$ (42.988 2)　(0.026 2)	$r^2 = 0.994\,3$
4)	$\widehat{\text{GNP}}_t = 164.207\,1 + 1.029\,0L_t$ (44.765 8)　(0.023 4)	$r^2 = 0.993\,8$

注：括号中的数字是估计的标准误（差）。

a. 哪一种货币定义看似与名义 GNP 有密切关系？

b. 既然 r^2 项都很高，这是否意味着无论怎样选择货币定义都关系不大？

c. 如果美联储想控制货币供给，那么，这些货币度量中的哪一种可作为较好的目标？你能从这些回归结果看出来吗？

5.15 假设两种物品的**无差异曲线**（indifference curve）方程是：

$$X_iY_i = \beta_1 + \beta_2 X_i$$

你会怎样估计此模型的参数？将此模型应用于表 5—8 中的数据并评述你所得到的结果：

表 5—8

消费品 X	1	2	3	4	5
消费品 Y	4	3.5	2.8	1.9	0.8

5.16 《经济学家》（*Economist*）杂志自从 1986 年开始发布巨无霸指数，作为一种粗糙而喧闹一时的指标，用以评价各国（地区）货币的汇率是否“正确”地确定在**购买力平价**（purchasing power parity，PPP）的理论水平上。PPP 认为，一单位货币能在所有国家（地区）购买到同样的商品组合。PPP 的支持者声称，从长远看，各国（地区）货币都向其 PPP 移动。《经济学家》把麦当劳餐厅的巨无霸当作代表性商品并给出表 5—9 中的信息。

表 5—9　　汉堡包标准（The hamburger standard）

	巨无霸价格		美元的隐含 PPP*	实际美元汇率，1 月 31 日	当地货币相对美元定价过低（−）/过高（+），%
	以当地货币计	以美元计			
美国†	3.22 美元	3.22	—	—	—
阿根廷	8.25 比索	2.65	2.56	3.11	−18
澳大利亚	3.45 澳大利亚元	2.67	1.07	1.29	−17
巴西	6.4 瑞亚尔	3.01	1.99	2.13	−6
英国	1.99 英镑	3.90	1.62††	1.96††	+21
加拿大	3.63 加拿大元	3.08	1.13	1.18	−4
智利	1 670 智利比索	3.07	519	544	−5
中国大陆	11.0 元	1.41	3.42	7.77	−56
哥伦比亚	6 900 比索	3.06	2 143	2 254	−5
哥斯达黎加	1 130 科朗	2.18	351	519	−32
捷克	52.1 克朗	2.41	16.2	21.6	−25
丹麦	27.75 丹麦克朗	4.84	8.62	5.74	+50
埃及	9.09 磅	1.60	2.82	5.70	−50
爱沙尼亚	30 克朗	2.49	9.32	12.0	−23
欧元区§	2.94 欧元	3.82	1.10**	1.30**	+19
中国香港	12.0 港币	1.54	3.73	7.81	−52
匈牙利	590 福林	3.00	183	197	−7
冰岛	509 克朗	7.44	158	68.4	+131
印度尼西亚	15 900 盾	1.75	4 938	9 100	−46
日本	280 日元	2.31	87.0	121	−28
拉脱维亚	1.35 拉脱	2.52	0.42	0.54	−22
立陶宛	6.50 立特	2.45	2.02	2.66	−24
马来西亚	5.50 林吉特	1.57	1.71	3.50	−51

续前表

	巨无霸价格		美元的隐含 PPP*	实际美元汇率，1月31日	当地货币相对美元定价过低（－）/过高（＋），%
	以当地货币计	以美元计			
墨西哥	29.0 墨西哥比索	2.66	9.01	10.9	－17
新西兰	4.60 新西兰元	3.16	1.43	1.45	－2
挪威	41.5 克朗	6.63	12.9	6.26	＋106
巴基斯坦	140 卢比	2.31	43.5	60.7	－28
巴拉圭	10 000 瓜拉尼	1.90	3 106	5 250	－41
秘鲁	9.50 新索尔	2.97	2.95	3.20	－8
菲律宾	85.0 菲律宾比索	1.74	26.4	48.9	－46
波兰	6.90 兹罗提	2.29	2.14	3.01	－29
俄罗斯	49.0 卢布	1.85	15.2	26.5	－43
沙特阿拉伯	9.00 里亚尔	2.40	2.80	3.75	－25
新加坡	3.60 新加坡元	2.34	1.12	1.54	－27
斯洛伐克	57.98 克朗	2.13	18.0	27.2	－34
南非	15.5 南非兰特	2.14	4.81	7.25	－34
韩国	2 900 圆	3.08	901	942	－4
斯里兰卡	190 卢比	1.75	59.0	109	－46
瑞典	32.0 瑞典克朗	4.59	9.94	6.97	＋43
瑞士	6.30 瑞士法郎	5.05	1.96	1.25	＋57
中国台湾	75.0 新台币	2.28	23.3	32.9	－29
泰国	62.0 泰铢	1.78	19.3	34.7	－45
土耳其	4.55 里拉	3.22	1.41	1.41	—
阿联酋	10.0 迪拉姆	2.72	3.11	3.67	－15
乌克兰	9.00 格里夫尼亚	1.71	2.80	5.27	－47
乌拉圭	55.0 比索	2.17	17.1	25.3	－33
委内瑞拉	6 800 玻利瓦尔	1.58	2 112	4 307	－51

注：* PPP：当地价格除以美国价格。

** 每欧元兑换美元数。

†取纽约、芝加哥、三藩市和亚特兰大的平均值。

††每英镑兑换美元数。

§欧元区的加权平均价格。

资料来源：McDonald's；*The Economist*，February 1，2007.

考虑如下回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

其中 Y＝实际汇率，X＝美元的隐含 PPP。

a. 若 PPP 成立，你会先验地预期 β_1 和 β_2 取什么值？

b. 回归的结果是否支持你的预期？你用什么形式的检验去检验你的假设？

c.《经济学家》是否应继续发布巨无霸指数？为什么？

5.17　参照习题 2.16 中所给的 SAT 数据。假设你想根据女生的数学成绩（X），通过做如下回归去预测男生的数学成绩（Y）：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

a. 估计上述模型。

b. 从所估计的残差看，正态性假定是否可以维系？

c. 检验假设：$\beta_2=1$，即男生和女生的数学成绩有一个 1∶1 的对应关系。

d. 建立对此问题的 ANOVA 表。

5.18 把上题重做一遍，但这回令 Y 和 X 分别代表男生和女生的阅读成绩。

5.19 表 5—10 给出了美国 1980—2006 年的消费者价格指数（CPI）和批发价格指数（WPI）或生产者价格指数（PPI）。

表 5—10　　1980—2006 年美国的 CPI 和 PPI

年份	CPI	PPI	年份	CPI	PPI
1980	82.4	88.0	1994	148.2	125.5
1981	90.9	96.1	1995	152.4	127.9
1982	96.5	100.0	1996	156.9	131.3
1983	99.6	101.6	1997	160.5	131.8
1984	103.9	103.7	1998	163.0	130.7
1985	107.6	104.7	1999	166.6	133.0
1986	109.6	103.2	2000	172.2	138.0
1987	113.6	105.4	2001	177.1	140.7
1988	118.3	108.0	2002	179.9	138.9
1989	124.0	113.6	2003	184.0	143.3
1990	130.7	119.2	2004	188.9	148.5
1991	136.2	121.7	2005	195.3	155.7
1992	140.3	123.2	2006	201.6	160.3
1993	144.5	124.7			

资料来源：*Economic Report of the President*，2007，Tables B-62 and B-65.

a. 以 PPI 为横轴、CPI 为纵轴描点。据经验，你预期这两个指数之间有何种关系？为什么？

b. 假设你想基于一个指数预测另一个指数。你会用哪个指数作为回归元，哪个指数作为回归子？为什么？

c. 做（b）部分你确定的回归。给出标准的结果，检验这两个指数有 1∶1 变化关系的假设。

d. 从（c）部分回归所得到的残差，你能接受真实误差项正态分布的假设吗？说明你所用的检验。

5.20 表 5—11 给出了 25 个职业人群的肺癌死亡指数（100＝平均水平）和抽烟指数（100＝平均水平）。

表 5—11　　抽烟与肺癌

职业	抽烟指数	肺癌死亡指数
农业、林业、渔业工人	77	84
挖掘和采石工人	137	116
天然气、焦炭和化工生产者	117	123
玻璃与陶器制造者	94	128
锻造锻压工人	116	155
电气电子工人	102	101
工程及相关行业工人	111	118
木工业工人	93	113
皮革业工人	88	104

续前表

职业	抽烟指数	肺癌死亡指数
纺织业工人	102	88
服装业工人	91	104
食物、饮料及烟草行业工人	104	129
造纸印刷业工人	107	86
其他产品制造者	112	96
建筑工人	113	144
油漆工和装潢工人	110	139
发动机、起重机等操作员	125	113
其他劳动力	113	146
交通运输业工人	115	128
库房仓库保管员等	105	115
文书办事员	87	79
销售员	91	85
服务业、体育和休闲场所工人	100	120
行政人员和经理人员	76	60
艺术家、科学家及技术工人	66	51

资料来源：http：//lib. stat. cmu. edu/DASL/datafiles/SmokingandCancer. html.

a. 将肺癌死亡指数相对抽烟指数做描点图，你能观察到什么形式的关系？

b. 令 Y＝肺癌死亡指数和 X＝抽烟指数，估计一个线性回归模型，并得到常用的回归统计量。

c. 在 $\alpha=5\%$的水平上，检验抽烟对肺癌没有影响的假设。

d. 从肺癌死亡率来看，哪个行业的风险较高？你能给出之所以如此的理由吗？

e. 有把职业分类明确引进回归分析的办法吗？

附录 5A

□ 5A.1 与正态分布有关的概率分布

在附录 A 中讨论了 t、χ^2 和 F 概率分布，它们与正态分布都有内在的联系。由于我们在后面的章节中要大量使用这些概率分布，所以我们在下面的一些定理中归纳出它们与正态分布之间的关系；至于定理的证明则超出本书的范围，可以在参考书目中查找。①

定理 5.1 若 Z_1，Z_2，…，Z_n 都是服从 $Z_i\sim N(\mu_i,\ \sigma_i^2)$ 的独立正态分布变量，那么它们的线性和 $Z=\sum k_iZ_i$ 也服从均值为 $\sum k_i\mu_i$ 和方差为 $\sum k_i^2\sigma_i^2$ 的正态分布，即 $Z\sim N(\sum k_i\mu_i,\sum k_i^2\sigma_i^2)$，其中 k_i 是不全为零的常数。注：μ 表示均值。

① 至于各个定理的证明，见 Alexander M. Mood，Franklin A. Graybill，and Duane C. Bose，*Introduction to the Theory of Statistics*，3d ed.，McGraw-Hill，New York，1974，pp. 239-249。

简言之，正态变量的线性组合本身还是正态分布的。比如 Z_1 和 Z_2 是独立分布的正态变量：$Z_1 \sim N(10, 2)$，$Z_2 \sim N(8, 1.5)$，那么线性组合 $Z = 0.8Z_1 + 0.2Z_2$ 也是正态分布的，均值为 $0.8\times10 + 0.2\times8 = 9.6$，方差为 $0.64\times2 + 0.04\times1.5 = 1.34$，即 $Z \sim (9.6, 1.34)$。

定理 5.2 若 $Z_1, Z_2, \cdots, Z_n$ 服从正态分布但不独立，则 $Z = \sum k_i Z_i$ 也服从均值为 $\sum k_i\mu_i$ 和方差为 $[\sum k_i^2\sigma_i^2 + 2\sum k_i k_j \operatorname{cov}(Z_i, Z_j), i \neq j]$ 的正态分布，其中 k_i 是不全为零的常数。

因此，若 $Z_1 \sim N(6, 2)$，$Z_2 \sim N(7, 3)$ 和 $\operatorname{cov}(Z_1, Z_2) = 0.8$，则线性组合 $0.6Z_1 + 0.4Z_2$ 也是正态分布的，均值为 $0.6\times6 + 0.4\times7 = 6.4$，方差为 $[0.36\times2 + 0.16\times3 + 2\times0.6\times0.4\times0.8] = 1.584$。

定理 5.3 若 $Z_1, Z_2, \cdots, Z_n$ 相互独立，且都服从 $Z_i \sim N(0, 1)$ 正态分布，即标准正态分布，则 $\sum Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ 服从自由度为 n 的 χ^2 分布。用符号表示即 $\sum Z_i^2 \sim \chi_n^2$，其中 n 表示自由度。

简言之，“独立标准正态变量的平方和服从自由度等于正态变量个数的 χ^2 分布。”①

定理 5.4 若 $Z_1, Z_2, \cdots, Z_n$ 为服从自由度为 k_i 的 χ^2 分布独立变量，则 $\sum Z_i = Z_1 + Z_2 + \cdots + Z_n$ 也服从自由度为 $\sum k_i$ 的 χ^2 分布。

于是，若 Z_1 和 Z_2 为自由度分别为 k_1 和 k_2 的 χ^2 分布独立变量，则 $Z = Z_1 + Z_2$ 是一个自由度为 $(k_1 + k_2)$ 的 χ^2 变量。此即 χ^2 分布的**再生性质**（reproductive property）。

定理 5.5 若 Z_1 是一个标准化的正态变量 $[Z_1 \sim N(0, 1)]$，另一变量 Z_2 服从自由度为 k 的 χ^2 分布并独立于 Z_1，则如下定义的变量

$$t = \frac{Z_1}{\sqrt{Z_2/k}} = \frac{Z_1\sqrt{k}}{\sqrt{Z_2}} = \frac{\text{标准正态变量}}{\sqrt{\text{独立 } \chi^2 \text{ 分布变量 /df}}} \sim t_k$$

服从 df $= k$ 的 t 分布。注：此分布在附录 A 中讨论，并在第 5 章加以说明。

顺便一提，注意随着自由度 k 无限增大（即 $k \to \infty$），t 分布趋近于标准正态分布。② 作为惯例，记号 t_k 表示自由度为 k 的 t 分布或变量。

---

① 至于各个定理的证明，见 Alexander M. Mood, Franklin A. Graybill, and Duane C. Bose, *Introduction to the Theory of Statistics*, 3d ed., McGraw-Hill, New York, 1974, p. 243。

② 证明参见 Henri Theil, *Introduction to Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1978, pp. 237-245。

定理 5.6 若 Z_1 和 Z_2 分别是自由度为 k_1 和 k_2 的 χ^2 分布独立变量，则变量

$$F=\frac{Z_1/k_1}{Z_2/k_2}\sim F_{k_1,k_2}$$

服从自由度为 k_1 和 k_2 的 F 分布，其中 k_1 表示**分子自由度**（numerator degrees of freedom），k_2 表示**分母自由度**（denominator degrees of freedom）。

同样，出于习惯，记号 $F_{k1,k2}$ 表示自由度为 k_1 和 k_2 的 F 变量，分子自由度放在前面。

换言之，定理 5.6 说明，F 变量无非就是两个 χ^2 分布独立变量分别除以其自由度后的比率。

定理 5.7 自由度为 k 的 t 变量的平方服从分子自由度 $k_1=1$ 和分母自由度 $k_2=k$ 的 F 分布。① 即

$$F_{1,k}=t_k^2$$

注意，欲使此等式成立，F 变量的分子自由度必须为 1。于是 $F_{1,4}=t_4^2$ 或 $F_{1,23}=t_{23}^2$ 等。

我们以后将逐渐看到前面这些定理的实际用处。

定理 5.8 对很大的分母自由度，分子自由度乘以 F 值就近似等于具有分子自由度的 χ^2 值。即

$$mF_{m,n}=\chi_m^2 \qquad \text{当 } n\to\infty \text{时}$$

定理 5.9 在自由度充分大时，χ^2 分布可由标准正态分布近似如下：

$$Z=\sqrt{2\chi^2}-\sqrt{2k-1}\sim N(0,1)$$

其中 k 表示自由度。

□ 5A.2 方程（5.3.2）的推导

假设：

$$Z_1=\frac{\hat{\beta}_2-\beta_2}{\text{se}(\hat{\beta}_2)}=\frac{(\hat{\beta}_2-\beta_2)\sqrt{x_i^2}}{\sigma} \tag{1}$$

以及

$$Z_2=(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{2}$$

如果 σ 已知，则 Z_1 服从标准正态分布；也就是说，$Z_1\sim N(0,1)$。（为什么？）Z_2 服从 $n-2$ 个自由度的 χ^2 分布。② 而且，可以证明 Z_2 的分布独立于 Z_1。③ 因此，借助定理 5.5，变量

① 证明参见方程（5.3.2）和（5.9.1）。

② 证明参见 Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics*, 2d ed., Macmillan, New York, 1965, p. 144。

③ 证明见 J. Johnston, *Econometric Methods*, 3d. ed., McGraw-Hill, New York, 1984, pp. 181-182（但要读懂它，需要具备矩阵代数的知识）。

$$t=\frac{Z_1\sqrt{n-2}}{\sqrt{Z_2}} \tag{3}$$

服从 $n-2$ 个自由度的 t 分布。将方程（1）和（2）代入方程（3）即得出方程（5.3.2）。

□ 5A.3 方程（5.9.1）的推导

方程（1）表明 $Z_1 \sim N(0, 1)$。因此，由定理 5.3，这个量的平方：

$$Z_1^2=\frac{(\hat{\beta}_2-\beta_2)^2\sum x_i^2}{\sigma^2}$$

服从 1 个自由度的 χ^2 分布。如 5A.1 节所指出：

$$Z_2=(n-2)\frac{\hat{\sigma}^2}{\sigma^2}=\frac{\sum \hat{u}_i^2}{\sigma^2}$$

服从 $n-2$ 个自由度的 χ^2 分布。此外，如 4.3 节所提到的，Z_2 的分布独立于 Z_1。然后应用定理 5.6 推出：

$$F=\frac{Z_1^2/1}{Z_2/(n-2)}=\frac{(\hat{\beta}_2-\beta_2)^2(\sum x_i^2)}{\sum \hat{u}_i^2/(n-2)}$$

服从自由度分别是 1 和 $n-2$ 的 F 分布。在虚拟假设 H_0：$\beta_2=0$ 下，上述 F 比率简化为方程（5.9.1）。

□ 5A.4 方程（5.10.2）和（5.10.6）的推导

均值预测的方差

给定 $X_i=X_0$，对真实均值的预测 $E(Y_0 \mid X_0)$ 由下式给出：

$$E(Y_0 \mid X_0)=\beta_1+\beta_2 X_0 \tag{1}$$

我们用

$$\hat{Y}_0=\hat{\beta}_1+\hat{\beta}_2 X_0 \tag{2}$$

来估计方程（1）。给定 X_0，因为 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 都是无偏估计量，取方程（2）的数学期望得：

$$\begin{aligned}E(\hat{Y}_0)&=E(\hat{\beta}_1)+E(\hat{\beta}_2)X_0\\&=\beta_1+\beta_2 X_0\end{aligned}$$

因此，

$$E(\hat{Y}_0)=E(Y_0 \mid X_0)=\beta_1+\beta_2 X_0 \tag{3}$$

也就是说，$\hat{Y}_0$ 是 $E(Y_0 \mid X_0)$ 的一个无偏误预测元。

现利用性质：

$$\text{var}(a+b)=\text{var}(a)+\text{var}(b)+2\text{cov}(a,b)$$

我们得到：

$$\text{var}(\hat{Y}_0)=\text{var}(\hat{\beta}_1)+\text{var}(\hat{\beta}_2)X_0^2+2\text{cov}(\hat{\beta}_1\hat{\beta}_2)X_0 \tag{4}$$

利用方程（3.3.1）、（3.3.3）和（3.3.9）中所给 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差与协方差公式，将各项合并整理即得：

$$\text{var}(\hat{Y}_0)=\sigma^2\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right] \tag{5.10.2}$$

个值预测的方差

我们要预测对应于 $X=X_0$ 的个值 Y，也就是，我们要得到：

$$Y_0=\beta_1+\beta_2 X_0+u_0 \tag{5}$$

我们把 Y_0 预测为：

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \tag{6}$$

预测误差 $Y_0 - \hat{Y}_0$ 是：

$$\begin{aligned} Y_0 - \hat{Y}_0 &= \beta_1 + \beta_2 X_0 + u_0 - (\hat{\beta}_1 + \hat{\beta}_2 X_0) \\ &= (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2) X_0 + u_0 \end{aligned} \tag{7}$$

因此，

$$\begin{aligned} E(Y_0 - \hat{Y}_0) &= E(\beta_1 - \hat{\beta}_1) + E(\beta_2 - \hat{\beta}_2) X_0 - E(u_0) \\ &= 0 \end{aligned}$$

这是因为 $\hat{\beta}_1$，$\hat{\beta}_2$ 是无偏的，X_0 是固定数，而 $E(u_0)$ 根据假定为零。

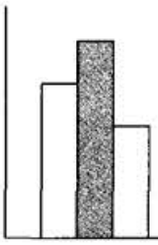
将方程（7）两边同时平方再取期望值，就得到：

$$\text{var}(Y_0 - \hat{Y}_0) = \text{var}(\hat{\beta}_1) + X_0^2 \text{var}(\hat{\beta}_2) + 2X_0 \text{cov}(\beta_1, \beta_2) + \text{var}(u_0)$$

利用先前给出的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差公式并注意到 $\text{var}(u_0) = \sigma^2$，得到：

$$\text{var}(Y_0 - \hat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2} \right] \quad = (5.10.6)$$

第 6 章 双变量线性回归模型的延伸



线性回归分析的某些特征，能够很容易地纳入我们至今已讨论过的双变量线性回归模型的框架之中。首先让我们考虑**过原点的回归**情形。也就是，模型中不存在截距项 β_1。然后，我们考虑**度量单位**（units of measurement）的问题，即 Y 和 X 变量用什么单位来度量，度量单位的变化会不会影响回归的结果。最后，我们考虑线性回归模型的**函数形式**（functional form）问题。至今我们所考虑的模型既是参数的线性函数，又是变量的线性函数。但请回顾前面各章所讲的回归理论，仅要求模型是参数的线性函数；进入模型的变量则可以是线性的，也可以不是线性的。本章中我们将表明，考虑参数为线性而变量不一定为线性的模型能处理一些有意思的实际问题。

我们一旦掌握好本章所介绍的思想，当我们在第 7 章和第 8 章中看到这些思想被推广到多元回归模型中时，我们就了如指掌了。

6.1 过原点回归

有时双变量 PRF 采取如下形式：

$$Y_i = \beta_2 X_i + u_i \tag{6.1.1}$$

在此模型中截距项不出现或者为零。因此取名为**过原点回归**（regression through the origin）。

作为一个说明性例子，考虑现代证券组合理论中的资本资产定价模型（capital asset pricing model，CAPM）。可用风险溢价或升水（risk-premium）的形式把它表述为[①]：

$$(ER_i - r_f) = \beta_i(ER_m - r_f) \tag{6.1.2}$$

其中 ER_i 为证券 i 的期望回报率；ER_m 为比方说，由标准普尔（S&P）500 综合股票指数所代表的市场证券组合的期望回报率；r_f 为无风险回报率，比方说，90 天国债回报率；β_i 为 Beta 系数，指不能通过分散投资而消除的系统风险（systematic risk）的一种度量，又指第 i 种证券回报率与市场互动程度的一种度量。一个大于 1 的 β_i 意味着证券 i 是一种易波动或进攻型证券，而一个小于 1 的 β_i 则意味着证券 i 是一种防御型证券。（注：不要把这个 β_i 和双变量回归的斜率系数 β_2 混同起来。）

如果资本市场有效运行，则 CAPM 要求证券 i 的期望风险溢价（$=ER_i - r_f$）等于期望市场风险溢价（$=ER_m - r_f$）乘以该证券的 β 系数。如果 CAPM 成立，我们就得到图 6—1 所描述的情形。图中所展示的直线叫做**证券市场线**（security market line，SML）。

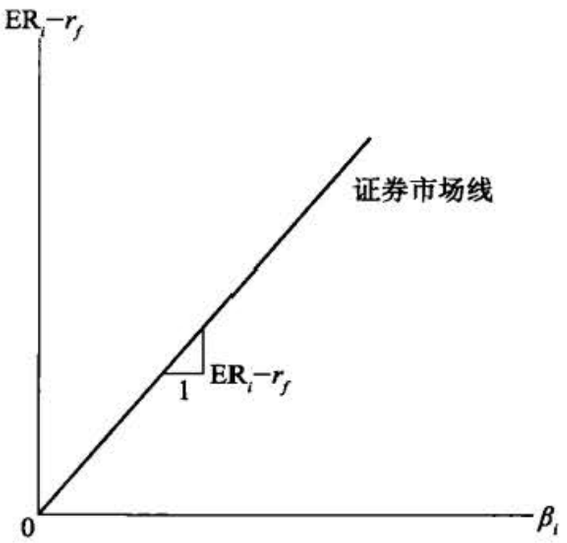


图 6—1　系统风险

为便于进行经验研究，方程（6.1.2）常被表达为：

$$R_i - r_f = \beta_i(R_m - r_f) + u_i \tag{6.1.3}$$

或者

$$R_i - r_f = \alpha_i + \beta_i(R_m - r_f) + u_i \tag{6.1.4}$$

后一个式子叫做**市场模型**（market model）。[②] 如果 CAPM 成立，则预期 α_i 为零。（见图 6—2。）

---

① 见 Haim Levy and Marshall Sarnat, *Portfolio and Investment Selection: Theory and Practice*, Prentice-Hall International, Englewood Cliffs, NJ, 1984, Chap. 14。

② 例如，参看 Diana R. Harrington, *Modern Portfolio Theory and the Capital Asset Pricing Model: A User's Guide*, Prentice Hall, Englewood Cliffs, NJ, 1983, p. 71。

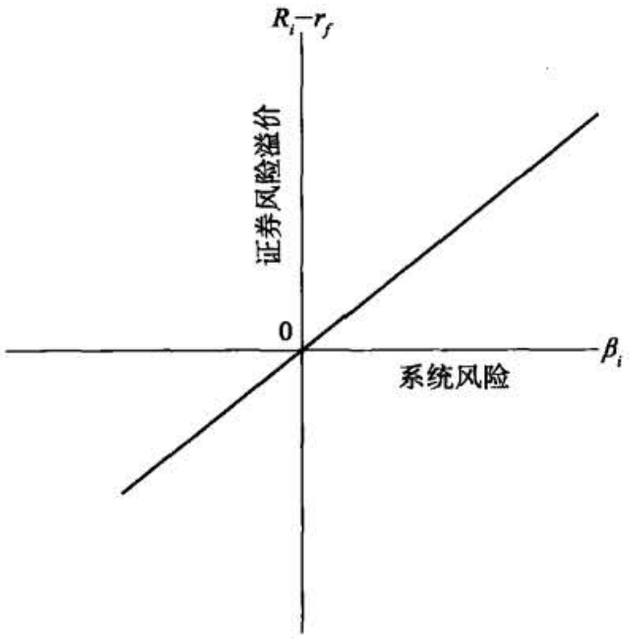


图 6—2　证券组合理论的市场模型（假定 $\alpha_i = 0$）

顺便指出，方程（6.1.4）中的因变量 Y 是（$R_i - r_f$），但解释变量 X 是波动性系数 β_i，而不是（$R_m - r_f$）。因此，为了做回归方程（6.1.4），必须先估计 β_i，如同习题 5.5 所描述的，β_i 通常要从**特征线**（characteristic line）导出。（更多细节，参见习题 8.28。）

如本例所示，有时基本理论能断定某个模型没有截距项。其他适合零截距模型的例子还有弗里德曼的永久收入假说（permanent income hypothesis）：永久消费正比于永久收入；成本分析理论：生产的可变成本正比于产出；以及货币主义理论的某些解说，如价格变化率（即通货膨胀率）正比于货币供给变化率。

如何估计类似于方程（6.1.1）这样的模型呢？这类模型提出了什么特殊问题？为了回答这些问题，可先把方程（6.1.1）的样本回归函数写成：

$$Y_i = \hat{\beta}_2 X_i + \hat{u}_i \tag{6.1.5}$$

现对方程（6.1.5）应用 OLS 法，得到 $\hat{\beta}_2$ 的如下公式及其方差（证明见附录 6A 第 6A.1 节）：

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{6.1.6}$$

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_i^2} \tag{6.1.7}$$

其中 σ^2 被估计为：

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-1} \tag{6.1.8}$$

将这些公式同含有截距项的模型的公式相比是有趣的，后者是：

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{3.1.6}$$

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{3.3.1}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{3.3.5}$$

两组公式之间的差异应是明显的：在没有截距项的模型中，我们使用**原始**（raw）变量的平方和及其交叉乘积和，而在有截距项的模型中，我们使用变量偏离其均值的离差平方和及其交叉乘积和。其次，在计算 $\hat{\sigma}^2$ 时，前者的自由度是（$n-1$），而后者的自由度是（$n-2$）。（为什么?）

虽然无截距或零截距模型在某些情况下是适宜的，但要注意这种模型的一些特点。第一，对有截距项的模型（惯用的模型）来说，$\sum \hat{u}_i = 0$ 总是成立的。但当截距项不出现时，$\sum \hat{u}_i = 0$ 就不一定成立。简言之，在过原点回归中，$\sum \hat{u}_i$ 不一定是零。第二，第 3 章所介绍的判定系数 r^2 对惯用的模型来说总是非负的。但对无截距模型来说，有时可能出现负值！这些异常结果的出现，是因为第 3 章中所介绍的 r^2 明确地假定模型包含有截距。因此，按习惯计算的 r^2 未必适合于过原点回归模型。①

□ 过原点回归模型的 r^2

如刚才所指出并在附录 6A 第 6A.1 节中进一步讨论的那样，第 3 章给出的惯用的 r^2 并不适合于不含截距的回归。但是我们可以对这类模型计算以 **raw r^2** 为名并定义如下的 raw r^2：

$$\text{raw } r^2 = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2} \tag{6.1.9}$$

注：这些是原始（而不是经过均值校正的）平方和及交叉乘积和。

虽然 raw r^2 满足关系 $0 < r^2 < 1$，却不能直接同惯用的 r^2 值相比。因此，一些作者并不对零截距回归模型报告 r^2 值。

由于此模型的这些异常特性，在使用零截距回归模型时须特别小心。除非有非常强的先验预期，否则以采取习惯含有截距的模型为好。这样做有两方面的好处：第一，尽管模型含有截距项，但若该项的出现是统计上不显著的（即统计上等于零），则从任何实际方面考虑，都可认为这个结果是一个过原点回归模型。② 第二，并且更为重要的是，如果在模型中确实有截距，而我们却执意拟合一个过原点回归，

---

① 更多的讨论见 Dennis J. Aigner, *Basic Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1971, pp. 85-88。

② 亨利·瑟尔（Henri Theil）指出，如果确实没有截距，那么斜率系数要比硬放进一个截距项估计的准确得多。见 *Introduction to Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1978, p. 76。还可参看他随之给出的数值例子。

我们就犯了**设定错误**（specification error）。我们在第 7 章将进一步深入讨论这个问题。

例 6.1

表 6—1 给出了 1980—1999 年间同周期消费品部门中 104 种股票构成的一个指数的超额回报率 Y_t（%）和英国总体股票指数的超额回报率 X_t（%）的月度数据，共 240 个观测。[①] 超额回报率指的是超过无风险资产回报率的部分（参见 CAPM 模型）。

表 6—1

OBS	Y	X	OBS	Y	X
1980 年 1 月	6.080 228 52	7.263 448 404	1982 年 2 月	−11.129 075 03	−4.033 607 075
1980 年 2 月	−0.924 185 461	6.339 895 504	1982 年 3 月	1.724 627 956	3.042 525 777
1980 年 3 月	−3.286 174 252	−9.285 216 834	1982 年 4 月	0.157 879 967	0.734 564 665
1980 年 4 月	5.211 976 571	0.793 290 771	1982 年 5 月	−1.875 202 616	2.779 732 288
1980 年 5 月	−16.164 211 11	−2.902 420 985	1982 年 6 月	−10.624 817 67	−5.900 116 576
1980 年 6 月	−1.054 703 649	8.613 150 875	1982 年 7 月	−5.761 135 416	3.005 344 385
1980 年 7 月	11.172 376 99	3.982 062 848	1982 年 8 月	5.481 432 596	3.954 990 619
1980 年 8 月	−11.063 275 51	−1.150 170 907	1982 年 9 月	−17.022 074 59	2.547 127 067
1980 年 9 月	−16.776 996 09	3.486 125 868	1982 年 10 月	7.625 420 708	4.329 008 106
1980 年 10 月	−7.021 834 032	4.329 850 278	1982 年 11 月	−6.575 721 646	0.191 940 594
1980 年 11 月	−9.716 846 68	0.936 875 279	1982 年 12 月	−2.372 829 861	−0.921 675 55
1980 年 12 月	5.215 705 717	−5.202 455 846	1983 年 1 月	17.523 749 36	3.394 682 577
1981 年 1 月	−6.612 000 956	−2.082 757 509	1983 年 2 月	1.354 655 809	0.758 714 353
1981 年 2 月	4.264 498 443	2.728 522 893	1983 年 3 月	16.268 610 49	1.862 073 664
1981 年 3 月	4.916 710 821	0.653 397 106	1983 年 4 月	−6.074 547 158	6.797 751 341
1981 年 4 月	22.204 959 46	6.436 071 962	1983 年 5 月	−0.826 650 702	−1.699 253 628
1981 年 5 月	−11.298 685 24	−4.259 197 932	1983 年 6 月	3.807 881 996	4.092 592 402
1981 年 6 月	−5.770 507 783	0.543 909 707	1983 年 7 月	0.575 700 91	−2.926 299 262
1981 年 7 月	−5.217 764 717	−0.486 845 933	1983 年 8 月	3.755 563 441	1.773 424 306
1981 年 8 月	16.196 201 75	2.843 999 508	1983 年 9 月	−5.365 927 271	−2.800 815 667
1981 年 9 月	−17.169 953 95	−16.457 214 2	1983 年 10 月	−3.750 302 815	−1.505 394 995
1981 年 10 月	1.105 334 728	4.468 938 171	1983 年 11 月	4.898 751 703	4.186 962 84
1981 年 11 月	11.685 336 7	5.885 519 658	1983 年 12 月	4.379 256 151	1.201 416 981
1981 年 12 月	−2.301 451 728	−0.390 698 164	1984 年 1 月	16.560 161 88	6.769 320 788
1982 年 1 月	8.643 728 679	2.499 567 896	1984 年 2 月	1.523 127 464	−1.686 027 417

① 这些数据源自 *DataStream* 数据库，这里直接复制于 Christiaan Heij et al.，*Econometrics Methods with Applications in Business and Economics*，Oxford University Press，Oxford，U. K.，2004。

续前表

OBS	Y	X	OBS	Y	X
1984年3月	1.020 607 8	5.245 806 105	1987年4月	3.662 578 335	4.295 976 077
1984年4月	−3.899 307 684	1.728 710 264	1987年5月	7.157 455 113	7.719 692 529
1984年5月	−14.325 016 15	−7.279 075 595	1987年6月	4.774 901 623	3.039 887 622
1984年6月	3.056 627 177	−0.779 470 67	1987年7月	4.237 701 66	2.510 223 804
1984年7月	−0.021 535 92	−2.439 634 487	1987年8月	−0.881 352 219	−3.039 443 563
1984年8月	3.355 102 212	8.445 977 813	1987年9月	11.496 884 16	3.787 092 018
1984年9月	0.100 006 778	1.221 080 129	1987年10月	−35.566 176 24	−27.869 693 11
1984年10月	1.691 250 318	2.733 386 772	1987年11月	−14.591 373 69	−9.956 367 094
1984年11月	8.200 753 01	5.127 533 29	1987年12月	14.872 716 64	7.975 865 948
1984年12月	3.527 866 16	3.191 554 763	1988年1月	1.748 599 294	3.936 938 398
1985年1月	4.554 587 707	3.907 838 688	1988年2月	−0.606 016 446	−0.327 970 64
1985年2月	5.365 478 677	−1.708 567 484	1988年3月	−6.078 095 523	−2.161 544 202
1985年3月	4.525 231 564	0.435 218 492	1988年4月	3.976 153 828	2.721 787 842
1985年4月	2.944 654 344	0.958 067 845	1988年5月	−1.050 910 058	−0.514 825 422
1985年5月	−0.268 599 528	1.095 477 375	1988年6月	3.317 856 956	3.128 796 482
1985年6月	−3.661 040 481	−6.816 108 909	1988年7月	0.407 100 105	0.181 502 075
1985年7月	−4.540 505 062	2.785 054 354	1988年8月	−11.879 325 24	−7.892 363 786
1985年8月	9.195 292 816	3.900 209 023	1988年9月	8.801 026 046	3.347 081 899
1985年9月	−1.894 817 019	−4.203 004 414	1988年10月	6.784 211 277	3.158 592 144
1985年10月	12.006 612 74	5.601 798 02	1988年11月	−10.205 781 19	−4.816 470 363
1985年11月	1.233 987 382	1.570 093 976	1988年12月	−6.738 053 81	−0.008 549 997
1985年12月	−1.446 329 607	−1.084 427 121	1989年1月	12.839 036 43	13.460 982 19
1986年1月	6.023 618 851	0.778 669 473	1989年2月	3.302 860 922	−0.764 474 692
1986年2月	10.512 357 56	6.470 651 262	1989年3月	−0.155 918 301	2.298 491 097
1986年3月	13.400 710 24	8.953 781 192	1989年4月	3.623 090 767	0.762 074 588
1986年4月	−7.796 262 998	−2.387 761 685	1989年5月	−1.167 680 873	−0.495 796 117
1986年5月	0.211 540 446	−2.873 838 588	1989年6月	−1.221 603 303	1.206 636 013
1986年6月	6.471 111 064	3.440 269 098	1989年7月	5.262 902 744	4.637 026 116
1986年7月	−9.037 475 168	−5.891 053 375	1989年8月	4.845 013 219	2.680 874 116
1986年8月	−5.478 380 91	6.375 582 004	1989年9月	−5.069 564 838	−5.303 858 035
1986年9月	−6.756 881 852	−5.734 839 396	1989年10月	−13.579 635 26	−7.210 655 599
1986年10月	−2.564 960 223	3.630 884 08	1989年11月	1.100 607 603	5.350 185 944
1986年11月	2.456 599 468	−1.316 066 87	1989年12月	4.925 083 189	4.106 245 855
1986年12月	1.476 421 303	3.521 601 216	1990年1月	−2.532 068 851	−3.629 547 374
1987年1月	17.069 400 4	8.673 412 896	1990年2月	−6.601 872 876	−5.205 804 299
1987年2月	7.565 726 727	6.914 361 923	1990年3月	−1.023 768 943	−2.183 244 863
1987年3月	−3.239 325 817	−0.460 660 854	1990年4月	−7.097 917 266	−5.408 563 794

续前表

OBS	Y	X	OBS	Y	X
1990年5月	6.376 626 925	10.575 991 69	1993年7月	4.374 024 535	1.943 061 568
1990年6月	1.861 974 711	−0.338 612 099	1993年8月	1.733 528 075	4.961 979 827
1990年7月	−5.591 527 585	−2.213 162 02	1993年9月	−3.659 808 969	−1.618 729 936
1990年8月	−15.317 589 75	−8.476 177 427	1993年10月	5.856 907 64	4.215 408 608
1990年9月	−10.172 273 58	−7.459 414 71	1993年11月	−1.365 550 294	1.880 360 165
1990年10月	−2.217 396 045	−0.085 887 763	1993年12月	−1.346 979 017	5.826 352 413
1990年11月	5.974 205 798	5.034 770 534	1994年1月	12.895 787 58	2.973 540 693
1990年12月	−0.857 289 036	−1.767 714 908	1994年2月	−5.346 700 561	−5.479 858 563
1991年1月	−3.780 184 589	0.189 108 456	1994年3月	−7.614 726 564	−5.784 547 088
1991年2月	20.647 214 37	10.387 415 04	1994年4月	10.220 429 23	1.157 083 438
1991年3月	10.940 680 18	2.921 913 827	1994年5月	−6.928 422 261	−6.356 199 493
1991年4月	−3.145 639 589	0.971 720 188	1994年6月	−5.065 919 037	−0.843 583 888
1991年5月	−3.142 887 645	−0.431 781 9	1994年7月	7.483 498 556	5.779 953 224
1991年6月	−1.960 866 141	−3.342 924 986	1994年8月	1.828 762 662	3.298 130 184
1991年7月	7.330 964 031	5.242 811 509	1994年9月	−5.692 932 79	−7.110 010 085
1991年8月	7.854 387 926	2.880 654 691	1994年10月	−2.426 962 489	2.968 005 597
1991年9月	2.539 177 843	−1.121 472 224	1994年11月	2.125 100 668	−1.531 245 158
1991年10月	−1.233 244 642	−3.969 577 956	1994年12月	−4.225 370 964	0.264 280 259
1991年11月	−11.746 040 4	−5.707 995 062	1995年1月	−6.302 392 617	−2.420 388 431
1991年12月	1.078 226 286	1.502 567 049	1995年2月	1.278 676 37	0.138 795 213
1992年1月	5.937 904 622	2.599 565 094	1995年3月	10.908 905 16	3.231 656 585
1992年2月	4.113 184 542	0.135 881 087	1995年4月	2.497 849 434	2.215 804 682
1992年3月	−0.655 199 392	−6.146 138 064	1995年5月	2.891 526 594	3.856 813 589
1992年4月	15.284 302 78	10.457 368 31	1995年6月	−3.773 000 069	−0.952 204 306
1992年5月	3.994 517 585	1.415 987 046	1995年7月	8.776 288 715	4.020 036 363
1992年6月	−11.944 509 98	−8.261 109 424	1995年8月	2.882 560 97	1.423 600 345
1992年7月	−2.530 701 327	−3.778 812 167	1995年9月	2.146 913 33	−0.037 912 571
1992年8月	−9.842 366 221	−5.386 818 488	1995年10月	−4.590 104 662	−1.176 553 29
1992年9月	18.115 737 24	11.194 363 72	1995年11月	−1.293 255 187	3.760 277 356
1992年10月	0.200 950 206	3.999 870 038	1995年12月	−4.244 101 531	0.434 626 357
1992年11月	1.125 853 097	3.620 674 752	1996年1月	6.647 088 904	1.906 345 103
1992年12月	7.639 180 786	2.887 222 251	1996年2月	1.635 900 742	0.301 898 961
1993年1月	2.919 569 408	1.336 746 091	1996年3月	7.858 189 9	−0.314 132 324
1993年2月	−1.062 404 105	1.240 273 846	1996年4月	0.789 544 896	3.034 331 741
1993年3月	1.292 641 409	0.407 144 312	1996年5月	−0.907 725 397	−1.497 346 299
1993年4月	0.420 241 384	−1.734 930 047	1996年6月	−0.392 246 948	−0.894 676 854
1993年5月	−2.514 080 553	1.111 533 687	1996年7月	−1.035 896 351	−0.532 816 274
1993年6月	0.419 362 276	1.354 127 742	1996年8月	2.556 816 005	3.863 737 088

续前表

OBS	Y	X	OBS	Y	X
1996年9月	3.131 830 038	2.118 254 897	1998年5月	3.293 686 179	−1.613 131 159
1996年10月	−0.020 947 358	−0.853 553 262	1998年6月	−13.254 618 02	−0.397 288 954
1996年11月	−5.312 287 782	1.770 340 939	1998年7月	−7.714 205 916	−2.237 365 283
1996年12月	−5.196 176 326	1.702 551 635	1998年8月	−15.263 404 83	−12.463 199 3
1997年1月	−0.753 247 124	3.465 753 348	1998年9月	−15.228 651 41	−5.170 734 985
1997年2月	−2.474 343 938	1.115 253 221	1998年10月	15.962 180 38	11.705 447 88
1997年3月	2.476 478 02	−2.057 818 461	1998年11月	−8.684 089 113	−0.380 200 223
1997年4月	−1.119 104 196	3.570 899 55	1998年12月	17.138 423 69	4.986 705 187
1997年5月	3.352 076 269	1.953 480 438	1999年1月	−1.468 448 611	2.493 727 994
1997年6月	−1.910 172 239	2.458 700 404	1999年2月	8.503 6	0.937 105 259
1997年7月	0.142 814 607	2.992 341 297	1999年3月	10.894 307 3	4.280 082 506
1997年8月	10.501 992 63	−0.457 968 038	1999年4月	13.034 973 94	3.960 824 402
1997年9月	12.985 019 43	8.111 278 967	1999年5月	−5.654 671 597	−4.499 198 079
1997年10月	−4.134 761 655	−6.967 124 504	1999年6月	8.321 969 316	3.656 745 699
1997年11月	−4.148 579 856	−0.155 924 791	1999年7月	0.507 652 273	−2.503 971 473
1997年12月	−1.752 478 236	3.853 283 433	1999年8月	−5.022 980 561	−0.121 901 923
1998年1月	−3.349 121 498	7.379 466 014	1999年9月	−2.305 448 839	−5.388 032 432
1998年2月	14.074 713 04	4.299 097 886	1999年10月	−1.876 879 466	4.010 989 716
1998年3月	7.791 650 968	3.410 780 517	1999年11月	1.348 824 769	6.265 312 975
1998年4月	5.154 679 109	−0.081 494 993	1999年12月	−2.641 649 38	4.045 658 427

我们首先对这些数据拟合模型（6.1.3）。利用 EViews 6，我们得到标准的 EViews 回归结果如下：

Dependent Variable: *Y*
Method: Least Squares
Sample: 1980M01 1999M12
Included observations: 240

	Coefficient	Std. Error	*t*-Statistic	Prob.
X	**1.155512**	**0.074396**	**15.53200**	**0.0000**
R-squared	0.500309	Mean dependent var.		0.499826
Adjusted *R*-squared†	0.500309	S.D. dependent var.		7.849594
S.E. of regression	5.548786	Durbin-Watson stat.*		1.972853
Sum squared resid.	7358.578			

注：* 我们将在第12章讨论这个统计量。

† 见第7章。

如这些结论所示，斜率系数即 β 系数是高度显著的，因为它的 p 值极小。这里所做的解释是，如果超额市场回报率提高一个百分点，那么消费品部门指数的超额回报率将提高约1.15个百分点。斜率系数不仅统计显著，而且明显大于1（你能验证这个结论吗?）。如果 β 系数大于1，那么这个证券（这里指104种股票的投资组合）就是易波动的；它的波动超过了整个股指的波动。但这个结论无足为奇，因为在本例中，我们所考虑的股票来自家庭耐用品、汽车、纺织和运动设备等同周期消费品部门。

如果我们拟合模型（6.1.4），我们便得到如下结论：

Dependent Variable: *Y*
Method: Least Squares
Sample: 1980M01 1999M12
Included observations: 240

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−0.447481	0.362943	−1.232924	0.2188
X	1.171128	0.075386	15.53500	0.0000

R-squared	0.503480	Mean dependent var.	0.499826
Adjusted *R*-squared	0.501394	S.D. dependent var.	7.849594
S.E. of regression	5.542759	Durbin-Watson stat.	1.984746
Sum squared resid.	7311.877	Prob. (*F*-statistic)	0.000000
F-statistic	241.3363		

我们从这些结论中看到，尽管斜率系数（β系数）是高度统计显著的，但截距并非显著异于0。这就表明，过原点回归对这些数据的拟合效果更好。此外，从统计上讲，这两个模型中的斜率系数值没有差别。注意，过原点回归模型中斜率系数的标准误略低于含截距模型中斜率系数的标准误，这就支持了第152页注释②中瑟尔所做的论断。即便如此，斜率系数在统计上仍大于1，这就再次肯定了同周期消费品部门的股票更具波动性。

同时还注意到，由于在过原点回归模型中，常用的 r^2 表达式不再适用，所以应该有所保留地看待这种模型给出的 r^2 值。不过，即便是这样的模型，EViews仍例行给出标准的 r^2 值。

6.2 尺度与测量单位

为领会本节所介绍的思想，考虑表6—2中的数据。表中数据是以2000年美元（按链式法则）计算的美国国内私人总投资（gross private domestic investment，GPDI）和国内生产总值（GDP），分别以十亿美元和百万美元为单位。

表6—2　　1990—2005年美国国内私人总投资与GDP

（除非特别指出，都是以2000年美元按链式法则计算；季度数据按季节调整的年率折算）

年份	GPDIBL	GPDIM	GDPB	GDPM
1990	886.6	886 600.0	7 112.5	7 112 500.0
1991	829.1	829 100.0	7 100.5	7 100 500.0
1992	878.3	878 300.0	7 336.6	7 336 600.0
1993	953.5	953 500.0	7 532.7	7 532 700.0
1994	1 042.3	1 042 300.0	7 835.5	7 835 500.0
1995	1 109.6	1 109 600.0	8 031.7	8 031 700.0
1996	1 209.2	1 209 200.0	8 328.9	8 328 900.0

续前表

年份	GPDIBL	GPDIM	GDPB	GDPM
1997	1 320.6	1 320 600.0	8 703.5	8 703 500.0
1998	1 455.0	1 455 000.0	9 066.9	9 066 900.0
1999	1 576.3	1 576 300.0	9 470.3	9 470 300.0
2000	1 679.0	1 679 000.0	9 817.0	9 817 000.0
2001	1 629.4	1 629 400.0	9 890.7	9 890 700.0
2002	1 544.6	1 544 600.0	10 048.8	10 048 800.0
2003	1 596.9	1 596 900.0	10 301.0	10 301 000.0
2004	1 713.9	1 713 900.0	10 703.5	10 703 500.0
2005	1 842.0	1 842 000.0	11 048.6	11 048 600.0

注：GPDIBL＝以 2000 年十亿美元计国内私人总投资。
GPDIM＝以 2000 年百万美元计国内私人总投资。
GDPB＝以 2000 年十亿美元计国内生产总值。
GDPM＝以 2000 年百万美元计国内生产总值。
资料来源：*Economic Report of the President*，2007，Table B-2，p. 328.

假设在 GPDI 对 GDP 的回归中某一研究者使用以十亿美元计的数据，而另一研究者使用以百万美元计的同样变量的数据。这两种情形的回归结果会不会是一样的？如果不一样，哪一种结果应被采用？简言之，Y 和 X 的测量单位会造成回归结果的差异吗？如果存在差异，在选择回归分析的测量单位时要采取哪些合理的途径？为了回答这些问题，让我们系统地进行分析。令：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{6.2.1}$$

其中 Y＝GPDI，X＝GDP。定义：

$$Y_i^* = w_1 Y_i \tag{6.2.2}$$

$$X_i^* = w_2 X_i \tag{6.2.3}$$

其中 w_1 和 w_2 为常数，称作**尺度因子**（scale factors）；w_1 和 w_2 可以相等或相异。

由方程（6.2.2）和（6.2.3）显见 Y_i^* 和 X_i^* 是重新度量（rescaled）的 Y_i 和 X_i。例如，Y_i 和 X_i 是以十亿美元计量的而某人想改用百万美元去表达它们，于是有 $Y_i^* = 1\,000 Y_i$，$X_i^* = 1\,000 X_i$；这里 $w_1 = w_2 = 1\,000$。

现考虑使用变量 Y_i^* 和 X_i^* 的回归：

$$Y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^* + \hat{u}_i^* \tag{6.2.4}$$

其中 $Y_i^* = w_1 Y_i$，$X_i^* = w_2 X_i$，并且 $\hat{u}_i^* = w_1 \hat{u}_i$。（为什么？）

我们要找出以下两两变量之间的关系式：

1. $\hat{\beta}_1$ 和 $\hat{\beta}_1^*$；
2. $\hat{\beta}_2$ 和 $\hat{\beta}_2^*$；
3. $\text{var}(\hat{\beta}_1)$ 和 $\text{var}(\hat{\beta}_1^*)$；
4. $\text{var}(\hat{\beta}_2)$ 和 $\text{var}(\hat{\beta}_2^*)$；
5. $\hat{\sigma}^2$ 和 $\hat{\sigma}^{*2}$；
6. r_{xy}^2 和 $r_{x^*y^*}^2$。

由最小二乘理论，我们知道（见第 3 章）：

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \tag{6.2.5}$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{6.2.6}$$

$$\text{var}\,(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2} \cdot \sigma^2 \tag{6.2.7}$$

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{6.2.8}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{6.2.9}$$

类似地，把 OLS 应用于方程（6.2.4），我们得到：

$$\hat{\beta}_1^* = \overline{Y}^* - \hat{\beta}_2^* \overline{X}^* \tag{6.2.10}$$

$$\hat{\beta}_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} \tag{6.2.11}$$

$$\text{var}\,(\hat{\beta}_1^*) = \frac{\sum X_i^{*2}}{n \sum x_i^{*2}} \cdot \sigma^{*2} \tag{6.2.12}$$

$$\text{var}\,(\hat{\beta}_2^*) = \frac{\sigma^{*2}}{\sum x_i^{*2}} \tag{6.2.13}$$

$$\hat{\sigma}^{*2} = \frac{\sum \hat{u}_i^{*2}}{n-2} \tag{6.2.14}$$

根据这些结果，就很容易证明这两组参数估计值之间的关系。所要做的仅是回忆如下定义的关系式：$Y_i^* = w_1 Y_i$（或 $y_i^* = w_1 y_i$）；$X_i^* = w_2 X_i$（或 $x_i^* = w_2 x_i$）；$\hat{u}_i^* = w_1 \hat{u}_i$；$\overline{Y}^* = w_1 \overline{Y}$ 和 $\overline{X}^* = w_2 \overline{X}$。读者利用这些定义很容易就能证明：

$$\hat{\beta}_2^* = \left(\frac{w_1}{w_2}\right)\hat{\beta}_2 \tag{6.2.15}$$

$$\hat{\beta}_1^* = w_1 \hat{\beta}_1 \tag{6.2.16}$$

$$\hat{\sigma}^{*2} = w_1^2 \hat{\sigma}^2 \tag{6.2.17}$$

$$\text{var}\,(\hat{\beta}_1^*) = w_1^2\,\text{var}\,(\hat{\beta}_1) \tag{6.2.18}$$

$$\text{var}\,(\hat{\beta}_2^*) = \left(\frac{w_1}{w_2}\right)^2 \text{var}\,(\hat{\beta}_2) \tag{6.2.19}$$

$$r_{xy}^2 = r_{x^*y^*}^2 \tag{6.2.20}$$

由上述结果显见，一旦尺度因子 w 已知，给定了一种测量尺度的回归结果，便可导出另一种测量尺度的回归结果。在实践中，虽然我们应该合理地选择测量单位，但要用许多零去把数字表达成百万或十亿是没有多少意义的。

通过方程（6.2.15）到（6.2.20）这些结果很容易推导出一些特例。例如，当 $w_1 = w_2$，即尺度因子相同时，斜率系数及其标准误不受尺度从（Y_i，X_i）变到（Y_i^*，X_i^*）的影响。这一点是显而易见的。然而，截距及其标准误却放大或缩小了

w_1 倍。但若 X 尺度不变（即 $w_2=1$），而 Y 尺度按因子 w_1 改变，则斜率和截距系数以及它们各自的标准误都要乘以相同的 w_1 因子。最后，如果 Y 尺度不变（即 $w_1=1$），而 X 尺度按因子 w_2 改变，则斜率系数及其标准误要乘以因子 $(1/w_2)$，但截距系数及其标准误不变。

然而，应该知道从 (Y, X) 到 (Y^*, X^*) 的尺度变换并不影响前面各章所讨论的 OLS 估计量的性质。

例 6.2　1990—2005 年美国 GPDI 与 GDP 之间的关系

为了证实上述理论结果，让我们回到表 6—2 给出的数据，并分析下述回归结果（括号中的数字为估计标准误。）

GPDI 和 GDP 都以十亿美元计算：

$$\widehat{\text{GPDI}}_t = -926.090 + 0.253\,5\ \text{GDP}_t$$

$$\text{se} = (116.358)\ (0.012\,9) \qquad r^2 = 0.964\,8 \tag{6.2.21}$$

GPDI 和 GDP 都以百万美元计算：

$$\widehat{\text{GPDI}}_t = -926\,090 + 0.253\,5\ \text{GDP}_t$$

$$\text{se} = (116\,358)\ (0.012\,9) \qquad r^2 = 0.964\,8 \tag{6.2.22}$$

注意，如理论所示，截距及其标准误都是回归（6.2.21）中相应值的 1 000 倍（即从十亿美元变到百万美元的 $w_1=1\,000$），但斜率系数及其标准误均不变。

GPDI 以十亿美元计算而 GDP 以百万美元计算：

$$\widehat{\text{GPDI}}_t = -926.090 + 0.000\,253\,5\ \text{GDP}_t$$

$$\text{se} = (116.358)\ (0.000\,012\,9) \qquad r^2 = 0.964\,8 \tag{6.2.23}$$

如同所料，因为仅仅 X 或 GDP 改变了尺度，所以斜率系数及其标准误都是它们在方程（6.2.21）中对应值的 1/1 000 倍。

GPDI 以百万美元计算而 GDP 以十亿美元计算：

$$\widehat{\text{GPDI}}_t = -926\,090 + 253.524\ \text{GDP}_t$$

$$\text{se} = (116\,358.7)\ (12.946\,5) \qquad r^2 = 0.964\,8 \tag{6.2.24}$$

再次看到如理论结果所示，截距和斜率系数以及它们各自的标准误都是它们在方程（6.2.21）中对应值的 1 000 倍。

注意，上面给出所有回归的 r^2 值都保持不变，由于 r^2 值是一个纯数字或没有维度，所以它不随度量单位而变化，上述 r^2 都相同也就无足为奇了。

□ 为结果的解释进一言

因为斜率系数 β_2 无非就是变化率，它的单位就是如下比率的单位：

$$\frac{\text{因变量 } Y \text{ 的单位}}{\text{解释变量 } X \text{ 的单位}}$$

例如，在回归（6.2.21）中，斜率系数 0.253 5 的意义是，GDP 每改变一个单位，即 10 亿美元，GPDI 平均改变 2.535 亿美元。在回归（6.2.23）中，GDP 的一个单位

即 100 万美元的变化，平均导致 GPDI 变化 0.002 535 亿美元。当然，这两个结果从它们的 GDP 对 GPDI 的影响看是完全相同的；只不过用不同的测量单位来表达而已。

6.3 标准化变量的回归

我们在上一节看到，回归子和回归元的单位会影响到回归系数的截距。如果我们愿意把回归子和回归元表示成标准化变量，这种影响就得以避免。如果将一个变量在减去其均值后再除以其标准差，我们就说把这个变量标准化了。

于是，在 Y 对 X 的回归中，如果我们把这些变量重新定义为

$$Y_i^* = \frac{Y_i - \overline{Y}}{S_Y} \tag{6.3.1}$$

$$X_i^* = \frac{X_i - \overline{X}}{S_X} \tag{6.3.2}$$

其中 $\overline{Y}$ 为 Y 的样本均值，S_Y 为 Y 的样本标准差，$\overline{X}$ 为 X 的样本均值，S_X 为 X 的样本标准差；变量 Y_i^* 和 X_i^* 则被称为**标准化变量**（standardized variables）。

标准化变量的一个有趣特征是，其均值总是 0 和标准差总是 1。（证明见附录 6A 第 6A.2 节。）

因此，回归子和回归元如何度量就无所谓了。于是，不再做标准（双变量）回归：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{6.3.3}$$

我们对标准化变量做回归

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + u_i^* \tag{6.3.4}$$

$$= \beta_2^* X_i^* + u_i^* \tag{6.3.5}$$

由于对标准化的回归子和回归元做回归，所以截距项总是零。[①] 标准化变量的回归系数（由 β_1^* 和 β_2^* 表示）在文献中被称为 **β 系数**（beta coefficients）。[②] 顺便指出，方程（6.3.5）是一个过原点的回归。

如何解释这些 β 系数呢？其解释是，如果（标准化）回归元增加一个单位的标准差，则（标准化）回归子平均增加 β_2^* 单位个标准差。于是，与方程（6.3.3）中的传统模型不同，我们度量的变量影响不再使用原来表示 Y 和 X 的单位，而是用其标准差作为单位。

为说明方程（6.3.3）与（6.3.5）之间的差别，让我们回到上一节中讨论的 GPDI 和 GDP 一例。为便于讨论，将上一节得到的结果再次给出如下：

$$\widehat{\text{GPDI}}_t = -926.090 + 0.253\,5\ \text{GDP}_t$$

$$\text{se} = (116.358)\ (0.012\,9) \qquad r^2 = 0.964\,8 \tag{6.3.6}$$

① 记得在方程（3.1.7）中，截距＝因变量的均值－斜率×回归元的均值。但对标准化变量而言，因变量和回归元的均值都是零，所以截距值为零。

② 不要将这些 β 系数与金融理论中的 β 系数相混淆。

其中 GPDI 和 GDP 均以十亿美元计。

对应于方程（6.3.5）的结果如下，其中带星号的变量为标准化变量：

$$\widehat{\text{GPDI}}_t^* = 0.982\ 2\ \text{GDP}_t^*$$
$$\text{se} = (0.048\ 5) \tag{6.3.7}$$

我们知道如何解释方程（6.3.6）：若 GDP 提高 1 美元，则 GPDI 平均提高 25 美分。方程（6.3.7）又该如何解释呢？这里的解释是，若（标准化）GDP 增加一个标准差，则（标准化）GPDI 平均增加约 0.98 个标准差。

标准化回归模型与传统模型相比有什么优势呢？若不止一个回归元，则优势更加明显，我们在第 7 章将讨论这个论题。通过将回归元标准化，我们就能将它们放到同等地位并直接进行比较。如果一个标准化回归元的系数比模型中另一个标准化回归元的系数大，那么前者就能比后者更多地解释回归子。换言之，我们可以用 β 系数作为各个回归元相对解释力的一种度量。在接下来的两章有更多的说明。

在结束本论题之前，须注意两点。第一，由于标准化回归（6.3.7）是一个过原点回归，而我们在 6.1 节已经指出通常的 r^2 不能使用，所以我们就没有给出其 r^2 值。第二，传统模型的 β 系数与这里的 β 系数之间存在一种有趣的关系。在双变量情形中，这种关系如下：

$$\hat{\beta}_2^* = \hat{\beta}_2 \left(\frac{S_x}{S_y}\right) \tag{6.3.8}$$

其中 S_x 为回归元 X 的样本标准差，S_y 为回归子 Y 的样本标准差。因此，若我们知道回归元和回归子的（样本）标准差，则可以将两个系数相互转换。我们在下一章中会看到，在多元回归中，这一关系仍然成立。针对我们的说明性例子验证方程（6.3.8）成立，则留给读者作为一个练习。

6.4 回归模型的函数形式

如在第 2 章中所指出的那样，本书主要考虑参数的线性模型；对变量则可以是或不是线性的。在下面的几节中，我们考虑一些常用的回归模型，它们也许对变量是非线性的，但对参数则是线性的。或者可通过适当的变量代换而变为参数的线性函数。具体地说，我们讨论如下回归模型：

1. 对数线性模型；
2. 半对数模型；
3. 倒数模型；
4. 对数倒数模型。

我们讨论每一种模型的特点，这些模型在什么场合适用，以及怎样估计它们。每一种模型都用适当的例子加以说明。

6.5 怎样度量弹性：对数线性模型

考虑以**指数回归模型**（exponential regression model）命名的如下模型：

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i} \tag{6.5.1}$$

它又可表达为[①]

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i \tag{6.5.2}$$

其中 ln 为自然对数（即以 e 为底的对数，e=2.718）。[②]

如果将方程（6.5.2）写成：

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i \tag{6.5.3}$$

其中 $\alpha = \ln\beta_1$，这个模型就是参数 α 和 β_2 的线性函数，并且是变量 Y 和 X 的对数的线性函数，从而可用 OLS 回归来估计。由于这种线性性质，该模型被称为**对数—对数**（log-log）、**双对数**（double-log）或**对数线性**（log-linear）模型。对数的性质参见附录 6A.3 节。

如果经典线性回归模型的假定均得到满足，则可用 OLS 估计方程（6.5.3）中的参数。令：

$$Y_i^* = \alpha + \beta_2 X_i^* + u_i \tag{6.5.4}$$

其中 $Y_i^* = \ln Y_i$，而 $X_i^* = \ln X_i$。所得的 OLS 估计量 $\hat{\alpha}$ 和 $\hat{\beta}_2$ 将分别是 α 和 β_2 的最优线性无偏估计量。

对数—对数模型的一个诱人且致使它获得普遍应用的特点，就是斜率系数 β_2 测度了 Y 对 X 的**弹性**（elasticity），也就是给定 X 变化的百分数引起 Y 变化的百分数。[③] 比

---

① 须知对数的这些性质：(1)$\ln(AB) = \ln A + \ln B$，(2)$\ln(A/B) = \ln A - \ln B$ 和(3)$\ln(A^k) = k \ln A$，这里假定 A 和 B 是正数而且 k 是某常数。

② 在实践中可以用常用对数，即以 10 为底的对数。自然对数与常用对数的关系是 $\ln_e X = 2.302\,6 \log_{10} X$。按惯例，ln 指自然对数，而 log 指以 10 为底的对数；从而没有必要写明下标 e 和 10。

③ 用微积分符号，弹性系数被定义为 $(dY/Y)/(dX/X) = [(dY/dX)(X/Y)]$。熟悉微分学的读者容易看出 β_2 确实是弹性系数。

一个技术性的注解：习惯于微积分表达的读者将看到 $d(\ln X)/dX = 1/X$ 或 $d(\ln X) = dX/X$，即 $\ln X$ 的无穷小变化（注意微分算子 d）等于 X 的相对或比例变化。在实践中，对于 X 的较小变化，可将此关系式写成 $\ln X$ 的变化≐X 的相对变化，这里≐表示近似等于。因此，对于小的变化，

$$(\ln X_t - \ln X_{t-1}) \doteq (X_t - X_{t-1})/X_{t-1} = X \text{ 的相对变化}$$

顺便指出，读者应留意这些常常出现的名词：(1) **绝对变化**（absolute change），(2) **相对或比例变化**（relative or proportional change），(3) **百分比变化**（percentage change）或**百分数增长率**（percent growth rate）。比如 $(X_t - X_{t-1})$ 表示绝对变化，$(X_t - X_{t-1})/X_{t-1} = X_t/X_{t-1} - 1$ 为相对或比例变化，而 $[(X_t - X_{t-1})/X_{t-1}]100$ 为百分比变化或增长率。X_t 和 X_{t-1} 分别是变量 X 的现期和前期值。

如说，Y 代表对某一商品的需求量，X 代表其单位价格，则 β_2 度量了需求的价格弹性，这是一个颇具经济含义的参数。如果需求量与价格的关系如图 6—3（a）所示，则由图 6—3（b）显示的双对数变换将给出价格弹性（$-\beta_2$）的估计值。

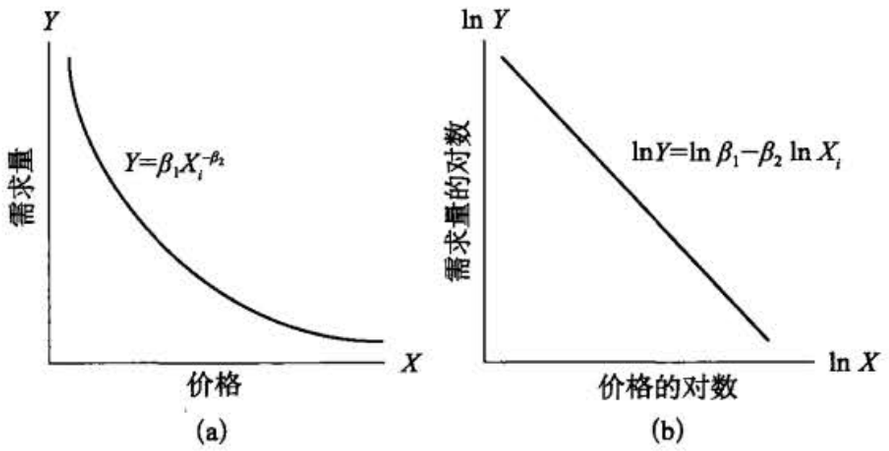


图 6—3　不变弹性模型

可以指出对数线性模型的两个特点：该模型假定 Y 与 X 之间的弹性系数 β_2 在整个研究范围内保持不变（为什么?），因此又名**不变弹性模型**（constant elasticity model）。[①] 换言之，如图 6—3（b）所示，不管在 $\ln X$ 哪一处测度弹性 β_2，$\ln X$ 每变化一个单位所引起的 $\ln Y$ 的变化都是一样的。模型的另一特点是，虽然 $\hat{\alpha}$ 和 $\hat{\beta}_2$ 是 α 和 β_2 的无偏估计量，（进入原始模型的参数）β_1 的估计值 $\hat{\beta}_1$（即 $\hat{\alpha}$ 的反对数）本身却是一个有偏误的估计量。然而在大多数实际问题中，截距项都居于次要地位，我们没有必要为得到一个无偏估计值而发愁。[②]

在双变量模型中，决定对数线性模型能否拟合好数据的最简单方法，是描绘出 $\ln Y_i$ 对 $\ln X_i$ 的散点图，看看这些散点是否差不多落在一条如同图 6—3（b）那样的直线上。

例 6.3　耐用品支出与个人消费总支出之间的关系

表 6—3 给出了个人消费总支出(PCEXP)、耐用品支出(EXPDUR)、非耐用品支出(EXPNONDUR)和劳务支出(EXPSERVICES) 方面的数据，均以 2000 年的十亿美元计。[③]

---

① 对于给定的价格百分比变化，不管价格的绝对水平是什么，一个不变弹性模型将给出一个不变的总收入变化。读者可将此结果同一个简单的线性需求函数 $Y_i=\beta_1+\beta_2 X_i+u_i$ 所蕴含的弹性情形相比较。然而，简单的线性函数却给出价格每单位变化导致数量需求量的恒定变化量。再将此情形同价格改变 1 美元时对数线性模型所给出的变化相比较。

② 关于偏误的性质以及怎样对付这种偏误，参看 Arthur S. Goldberger, *Topics in Regression Analysis*, Macmillan, New York, 1978, p. 120。

③ 耐用品包括机动车辆及其部件、家具和住房设施；非耐用品包括食物、衣物、汽油、石油、燃油和煤炭；劳务包括家务、电力和煤气、交通和医疗等。

表 6—3　　个人消费总支出及其分类

（均以 2000 年十亿美元按链式法则计算，季度数据按年增长率进行季节调整）

年份与季度	EXPSERVICES	EXPDUR	EXPNONDUR	PCEXP
2003 - I	4 143.3	971.4	2 072.5	7 184.9
2003 - II	4 161.3	1 009.8	2 084.2	7 249.3
2003 - III	4 190.7	1 049.6	2 123.0	7 352.9
2003 - IV	4 220.2	1 051.4	2 132.5	7 394.3
2004 - I	4 268.2	1 067.0	2 155.3	7 479.8
2004 - II	4 308.4	1 071.4	2 164.3	7 534.4
2004 - III	4 341.5	1 093.9	2 184.0	7 607.1
2004 - IV	4 377.4	1 110.3	2 213.1	7 687.1
2005 - I	4 395.3	1 116.8	2 241.5	7 739.4
2005 - II	4 420.0	1 150.8	2 268.4	7 819.8
2005 - III	4 454.5	1 175.9	2 287.6	7 895.3
2005 - IV	4 476.7	1 137.9	2 309.6	7 910.2
2006 - I	4 494.5	1 190.5	2 342.8	8 003.8
2006 - II	4 535.4	1 190.3	2 351.1	8 055.0
2006 - III	4 566.6	1 208.8	2 360.1	8 111.2

注：EXPSERVICES＝劳务支出，以 2000 年十亿美元为单位。
EXPDUR＝耐用品支出，以 2000 年十亿美元为单位。
EXPNONDUR＝非耐用品支出，以 2000 年十亿美元为单位。
PCEXP＝个人消费总支出，以 2000 年十亿美元为单位。

资料来源：Department of Commerce, Bureau of Economic Analysis. *Economic Report of the President*, 2007, Table B-17, p. 347.

假设我们想求出耐用品支出对个人消费总支出的弹性。将耐用品支出的对数相对个人消费总支出的对数描点，你将看到二者之间存在线性关系。因此，双对数模型适用。回归结果如下：

$$\widehat{\ln \text{EXDUR}_t} = -7.541\,7 + 1.626\,6 \ln \text{PCEX}_t$$
$$\text{se} = (0.716\,1) \quad (0.080\,0)$$
$$t = (-10.530\,9)^* \quad (20.315\,2)^* \quad r^2 = 0.969\,5 \qquad (6.5.5)$$

其中 * 表示 p 值极小。

这些结论表明，EXPDUR 对 PCEXP 的弹性约为 1.63，这就意味着，如果个人消费总支出提高 1%，则耐用品支出平均提高约 1.63%。因此，耐用品支出对个人消费总支出的变化非常敏感。这正是耐用品生产厂商为什么总是对个人收入和个人消费支出保持警觉的原因之一。习题 6.18 要求读者对非耐用品做一个类似的练习。

6.6　半对数模型：线性到对数与对数到线性模型

□ 怎样测量增长率：线性到对数模型

经济学家、企业人员与政府常常对于求出某些经济变量的增长率感兴趣，如人口、GNP、货币供给、就业、生产力、贸易赤字等。

假设我们想对表 6—3 中的数据求出个人劳务消费支出的增长率。令 Y_t 表示在 t 时期对劳务的真实支出，Y_0 表示劳务支出的初始值（即 2002 年第 4 季度末的值）。回忆你在经济学入门课程中学到的如下著名的复利公式：

$$Y_t = Y_0 (1+r)^t \tag{6.6.1}$$

其中 r 是 Y 的复合（即在时间上）增长率。取方程（6.6.1）的自然对数，则可写为：

$$\ln Y_t = \ln Y_0 + t\ln (1+r) \tag{6.6.2}$$

现假设：

$$\beta_1 = \ln Y_0 \tag{6.6.3}$$

$$\beta_2 = \ln (1+r) \tag{6.6.4}$$

就可把方程（6.6.2）写为：

$$\ln Y_t = \beta_1 + \beta_2 t \tag{6.6.5}$$

在方程（6.6.5）中加入一个干扰项便得到[①]：

$$\ln Y_t = \beta_1 + \beta_2 t + u_t \tag{6.6.6}$$

此模型和任何其他线性模型一样，也是参数 β_1 和 β_2 的线性函数。唯一的区别在于回归子是 Y 的对数，而回归元是“时间”，取值为 1、2、3 等等。

像方程（6.6.6）那样的模型叫做**半对数模型**（semilog models），因为只有一个变量（在本例中为回归子）以对数形式出现。为了便于叙述，只是回归子取对数的模型叫做**线性到对数模型**（log-lin model）。稍后我们将考虑回归子是线性而回归元是对数的另一种模型，并称之为**对数到线性模型**（lin-log model）。

在我们列出回归结果之前，先来检查模型（6.6.5）的一些性质。在此模型中，斜率系数度量了给定回归元（在本例中为时间变量 t）取值的绝对改变量时 Y 的恒定改变比例或相对改变量，也就是[②]：

$$\beta_2 = \frac{\text{回归子的相对改变量}}{\text{回归元的绝对改变量}} \tag{6.6.7}$$

如果将 Y 的相对改变量乘以 100，则方程（6.6.7）将给出相对于回归元 X 的绝对改变量 Y 的百分比变化或增长率（growth rate）。即 100 乘以 β_2 给出 Y 的增长率；100 乘以 β_2 在文献中被称为 Y 对 X 的**半弹性**（semielasticity）。（提问：要得到弹性，我们该怎么做。）[③]

例 6.4 劳务支出的增长率

为了说明增长模型（6.6.6），考虑表 6—3 中给出的劳务支出数据。对时间 t 的回归结果如下：

---

① 我们增加这个误差项是因为复利公式并不准确地成立。至于为什么要在作对数变换之后才加进这个误差项，将在 6.8 节中加以解释。

② 由微分学可以推出 $\beta_2 = \mathrm{d}(\ln Y)/\mathrm{d}X = (1/Y)(\mathrm{d}Y/\mathrm{d}X) = (\mathrm{d}Y/Y)/\mathrm{d}X$，而这就是方程(6.6.7)。对于 Y 和 X 的微小变化，这个关系式可近似地写为：

$$\frac{(Y_t - Y_{t-1})/Y_{t-1}}{X_t - X_{t-1}}$$

注：其中 $X=t$。

③ 至于各种增长表达式，参见附录 6A.4。

$$\widehat{\ln EXS_t} = 8.322\,6 + 0.007\,05t \qquad (6.6.8)$$
$$se = (0.001\,6) \quad (0.000\,18) \quad r^2 = 0.991\,9$$
$$t = (5\,201.625)^* \quad (39.166\,7)^*$$

注：EXS表示劳务支出，而*表示p值极小。

对方程（6.6.8）的解释是，在2003年第1季度到2006年第3季度期间，劳务支出以（每季度）0.705%的速度增加。粗略地讲，这等于2.82%的年增长率。由于8.322 6等于研究期初EXS的对数，所以取其反对数则得到EXS期初值（即2003年初）为41 159.6亿美元。图6—4勾勒了方程（6.6.8）中得到的回归线。

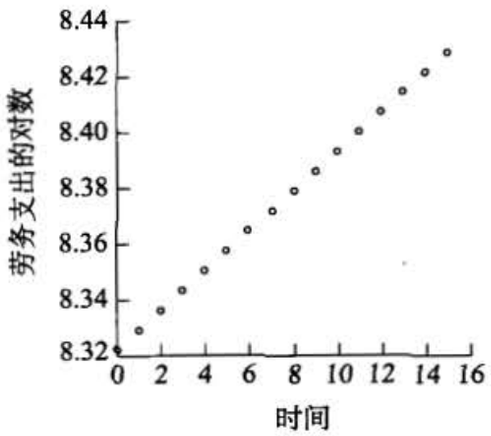


图6—4

瞬时与复合增长率。增长模型（6.6.6）中趋势变量的系数β_2给出**瞬时**（instantaneous）（指一个时点）增长率而不是**复合**（compound）（指一个时期）增长率。然而后者很容易由方程（6.6.4）求出：只需取β_2估计值的反对数，再从中减去1，然后用100乘以这个差值即可。于是，对我们的说明性例子，估计的斜率系数为0.007 05。因此［antilog(0.007 05) −1］=0.007 08或0.708%。因此，在说明性例子中，劳务支出的复合增长率约为每季度0.708%，略高于0.705%的瞬时增长率。这当然是由于复合效应所致。

线性趋势模型。有时研究者不去估计模型（6.6.6），而代之以如下的模型：

$$Y_t = \beta_1 + \beta_2 t + u_t \qquad (6.6.9)$$

即不做ln$Y$对时间的回归，而是做Y对时间的回归。这样的模型叫做**线性趋势模型**（linear trend model），并且把时间变量t取名为趋势变量。趋势的意思是指一个变量的行为中的一种持续上升或下降运动。如果方程（6.6.9）中的斜率系数是正的，则Y中存在**上升趋势**（upward trend）；反之，如果它是负的，则Y中存在**下降趋势**（downward trend）。

对于我们前面考虑的劳务支出数据，拟合线性趋势模型（6.6.9）的结果如下：

$$\widehat{EXS_t} = 4\,111.545 + 30.674t$$
$$t = (655.562\,8) \quad (44.467\,1) \quad r^2 = 0.993\,5 \qquad (6.6.10)$$

对照方程（6.6.8），对方程（6.6.10）的解释如下：在2003年第1季度至2006年第3季度期间，劳务支出以每季度约300亿美元的绝对速度（注意不是相对速度）增加。即劳务支出有上涨的趋势。

增长模型（6.6.8）与线性趋势模型（6.6.10）之间的取舍，有赖于人们对实际GDP的相对或绝对变化的兴趣。尽管对许多研究目的来说，相对变化是更为重要的。顺便提请注意，因为（6.6.8）和（6.6.10）两个模型的回归子不相同，所以不能比较它们的 r^2 值。我们在第7章将说明，该如何比较像模型（6.6.8）和（6.6.10）的 r^2。

□ 对数到线性模型

在刚才讨论的增长模型中，我们感兴趣的是，对 X 的一个单位的绝对变化，找出 Y 的百分比增长率。但现在我们感兴趣的是对 X 的一个百分比变化，找出 Y 的绝对变化量。能实现这一目的的模型可写为：

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i \tag{6.6.11}$$

为便于描述，我们把这个模型叫做**对数到线性模型**。

让我们来解释斜率系数 β_2。[①] 如同平常，

$$\beta_2 = \frac{Y\text{的变化}}{\ln X\text{的变化}} = \frac{Y\text{的变化}}{X\text{的相对变化}}$$

上面第二步是因为一个数字的对数变化就是它的相对变化。

用符号表示，我们有：

$$\beta_2 = \frac{\Delta Y}{\Delta X / X} \tag{6.6.12}$$

其中，Δ 照例表示一个微小的变化。方程（6.6.12）又可等价地写为：

$$\Delta Y = \beta_2 (\Delta X / X) \tag{6.6.13}$$

这个方程说明，Y 的绝对变化（ΔY）等于 β_2 乘以 X 的相对变化。如果后者乘以100，则方程（6.6.13）给出了 X 变化1%时 Y 的绝对变化量。例如，$\Delta X/X$ 改变0.01单位（或1%）时，Y 的绝对变化量是 $0.01\beta_2$。比如，人们在某一应用中求得 $\beta_2 = 500$，那么，Y 的绝对变化量就是 $0.01 \times 500 = 5.0$。因此，当人们用OLS来估计类似于方程（6.6.11）的回归时，要将斜率系数 β_2 的估计值乘以0.01，或者除以100也是一样的。如果不牢记这一点，你在应用中的解释将具有高度的误导性。

实际的问题是，像方程（6.6.11）这样的一个对数到线性模型什么时候有用？一个有趣的应用在于所谓的**恩格尔支出**（Engel expenditure）模型——以德国统计学家恩斯特·恩格尔（Ernst Engel，1821—1896）的名字命名。（见习题6.10。）恩格尔写道："用于食物的总支出以算术级数增加，而总支出以几何级数增加。"[②]

---

① 仍然利用微分学，我们得到 $dY/dX = \beta_2(1/X)$。因此，$\beta_2 = dY/(dX/X) =$ (6.6.12)。

② 见 Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries*, Routledge, London, 1998, p. 158。这里所引用部分归功于 H. Working, "Statistical Laws of Family Expenditure," *Journal of the American Statistical Association*, vol. 38, 1943, pp. 43-56。

例 6.5

作为对对数到线性模型的一个说明，让我们回顾有关印度食物支出的例 3.2。我们在那里拟合了一个线性于变量的模型作为初步近似。但若描点则得到图 6—5 中的散点图。如此图所示，食物支出比总支出增加得更缓慢，这可能印证了恩格尔法则。对数据拟合对数到线性模型的结果如下：

$$\widehat{\text{FoodExp}_i} = -1\,283.912 + 257.270\,0\ \ln\text{TotalExp}_i \qquad (6.6.14)$$

$$t = (-4.384\,8)^{*}\ (5.662\,5)^{*} \qquad r^2 = 0.376\,9$$

注：* 号表示 p 值极小。

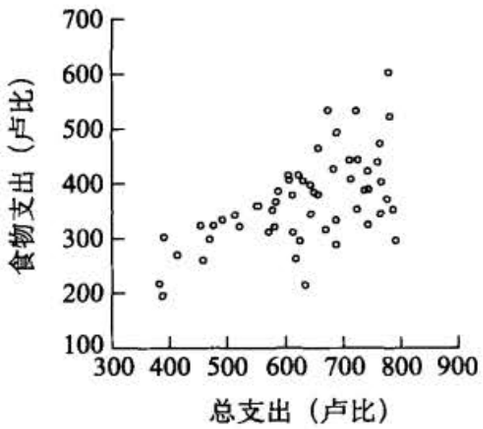


图 6—5

按前面描述的方法来解释，约等于 257 的斜率系数意味着总支出每提高 1%，导致样本中包含的 55 个家庭的食物支出平均增加约 2.57 卢比。（注：我们已将估计系数除以 100。）

在进一步说明之前，注意，如果你想计算线性到对数或对数到线性模型的弹性系数，你可以通过前面给出的弹性定义来计算，即

$$\text{弹性} = \frac{dY}{dX} \cdot \frac{X}{Y}$$

事实上，一旦一个模型的函数形式已知，就能利用上述定义来计算弹性。（表 6—6 对各种模型总结了弹性系数的计算方法。）

或许应该指出，对数变换有时被用来消除异方差性和偏态。（见第 11 章。）许多经济变量的一个共同特征就是它们是正偏的［比如企业的规模分布或收入（财富）的分布］。对这种变量进行对数变换就能够降低偏斜程度和异方差性。这正是劳动经济学家在工资对受教育程度（用受教育年数度量）的回归中将工资取对数的原因所在。

6.7 倒数模型

属于以下类型的模型均称**倒数**（reciprocal）模型：

$$Y_i = \beta_1 + \beta_2\left(\frac{1}{X_i}\right) + u_i \tag{6.7.1}$$

虽然此模型对变量 X 而言是非线性的（因为它以倒数形式进入模型），但模型对 β_1 和 β_2 而言却是线性的，因此它是一个线性回归模型。①

此模型有这样一些特点：随着 X 无限地增大，$\beta_2(1/X)$ 项趋于零（注意 β_2 是一常数）而 Y 趋于极限或渐近值 β_1。因此，方程（6.7.1）这类模型在结构上有一内在的**渐近线**（asymptote）或极限值。当变量 X 值无限增大时因变量将取此极限值。②图 6—6 给出了与方程（6.7.1）对应的几种可能的曲线形状。

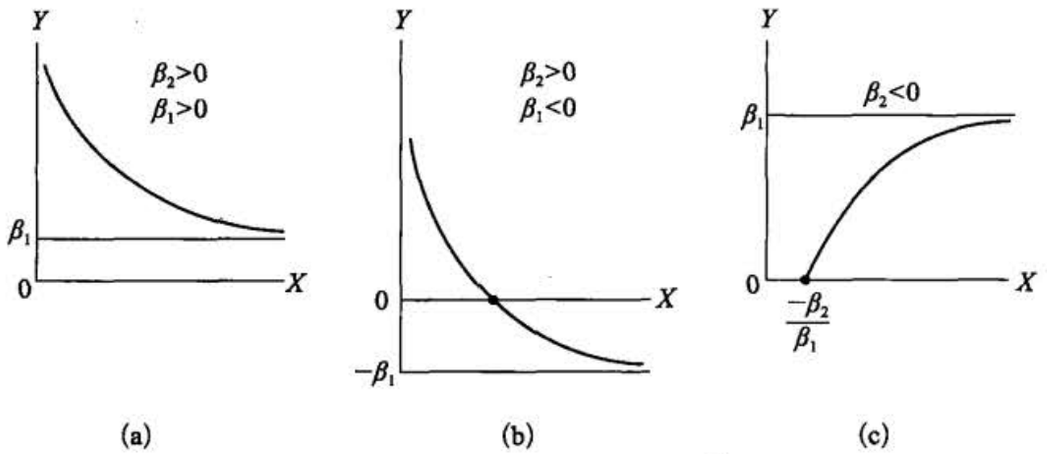


图 6—6　倒数模型：$Y=\beta_1+\beta_2\left(\frac{1}{X}\right)$

例 6.6

作为对图 6—6（a）的一个说明，考虑表 6—4 中给出的数据。这是 64 个国家的儿童死亡率及其他变量数据。目前主要考虑儿童死亡率（CM）和人均 GNP 这两个变量，并在图 6—7 中描出相应的点。

表 6—4　　64 个国家的生育率及其他数据

观测	CM	FLR	PGNP	TFR	观测	CM	FLR	PGNP	TFR
1	128	37	1 870	6.66	8	240	29	300	5.89
2	204	22	130	6.15	9	241	11	120	5.89
3	202	16	310	7.00	10	55	55	290	2.36
4	197	65	570	6.25	11	75	87	1 180	3.93
5	96	76	2 050	3.81	12	129	55	900	5.99
6	209	26	200	6.44	13	24	93	1 730	3.50
7	170	45	670	6.19	14	165	31	1 150	7.41

① 若令 $X_i^* = (1/X_i)$，则方程（6.7.1）既是参数又是变量 Y 和 X^* 的线性函数。

② 方程（6.7.1）的斜率是 $dY/dX = -\beta_2(1/X^2)$，其含义是，如果 β_2 是正的，斜率就一直是负的；而如果 β_2 是负的，斜率就总是正的，分别见图 6—6（a）和图 6—6（c）。

续前表

观测	CM	FLR	PGNP	TFR	观测	CM	FLR	PGNP	TFR
15	94	77	1 160	4.21	40	262	22	230	6.50
16	96	80	1 270	5.00	41	215	12	140	6.25
17	148	30	580	5.27	42	246	9	330	7.10
18	98	69	660	5.21	43	191	31	1 010	7.10
19	161	43	420	6.50	44	182	19	300	7.00
20	118	47	1 080	6.12	45	37	88	1 730	3.46
21	269	17	290	6.19	46	103	35	780	5.66
22	189	35	270	5.05	47	67	85	1 300	4.82
23	126	58	560	6.16	48	143	78	930	5.00
24	12	81	4 240	1.80	49	83	85	690	4.74
25	167	29	240	4.75	50	223	33	200	8.49
26	135	65	430	4.10	51	240	19	450	6.50
27	107	87	3 020	6.66	52	312	21	280	6.50
28	72	63	1 420	7.28	53	12	79	4 430	1.69
29	128	49	420	8.12	54	52	83	270	3.25
30	27	63	19 830	5.23	55	79	43	1 340	7.17
31	152	84	420	5.79	56	61	88	670	3.52
32	224	23	530	6.50	57	168	28	410	6.09
33	142	50	8 640	7.17	58	28	95	4 370	2.86
34	104	62	350	6.60	59	121	41	1 310	4.88
35	287	31	230	7.00	60	115	62	1 470	3.89
36	41	66	1 620	3.91	61	186	45	300	6.90
37	312	11	190	6.70	62	47	85	3 630	4.10
38	77	88	2 090	4.20	63	178	45	220	6.09
39	142	22	900	5.43	64	142	67	560	7.20

注：CM=儿童死亡率，每千名儿童中每年不足5岁便死亡的儿童人数。
FLR=妇女识字率，%。
PGNP=1980年的人均GNP。
TFR=1980—1985年的总生育率，即一位妇女生育的平均子女数，用给定年份按年龄划分的生育率表示。

资料来源：Chandan Mukherjee, Howard White, and Marc Whyte, *Econometrics and Data Analysis for Developing Countries*, Routledge, London, 1998, p. 456.

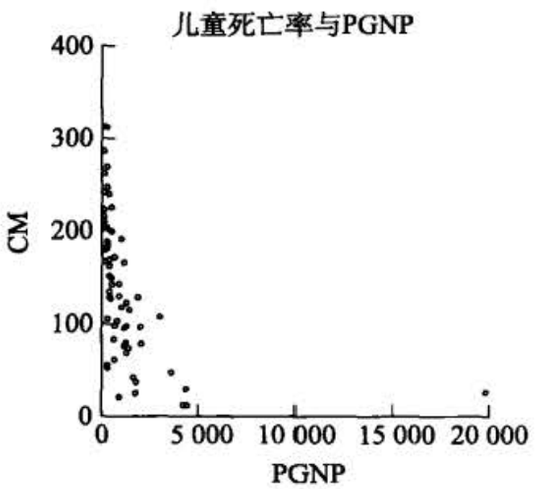


图 6—7　64个国家中儿童死亡率与人均GNP的关系

如你所见，此图与图 6—6（a）相似：假定所有其他变量保持不变，随着人均 GNP 的提高，预计儿童死亡率会因人们能承担更多的健康医疗费用而下降。但这种关系不是一条直线：随着人均 GNP 的增加，CM 首先有明显下降，但随着人均 GNP 继续增加，CM 的下降逐渐减弱。

如果我们试图拟合倒数模型（6.7.1），我们得到如下回归结果：

$$\widehat{CM}_i = 81.794\ 36 + 27\ 237.17\ (1/PGNP_i)$$
$$se = (10.832\ 1)\ (3\ 759.999) \qquad (6.7.2)$$
$$t = (7.551\ 1)\quad (7.253\ 5)\qquad r^2 = 0.459\ 0$$

随着人均 GNP 无限增加，儿童死亡率趋近其渐近值，每千人中死亡 82 人。如第 170 页注释②中解释的那样，$(1/PGNP_i)$ 的正系数意味着 CM 随着 PGNP 负向变化。

图 6—6（b）的重要应用之一，是宏观经济学中著名的菲利普斯曲线。根据 1861—1957 年间英国货币工资的百分比变化（Y）和失业的百分比（X）数据，菲利普斯得到一条在一般形状上类似于图 6—6（b）的曲线（图 6—8）。[①]

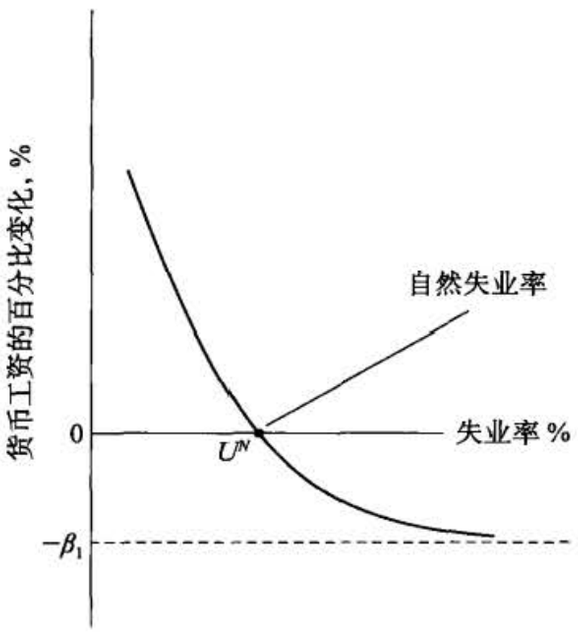


图 6—8　菲利普斯曲线

图 6—8 表明，在工资变化对失业水平的反应中，存在不对称性：当失业率低于经济学家所称的自然失业率［被定义为保持（工资）通货膨胀不变所需要的失业率］U^N 时，由失业的单位变化引起的工资上升要**快于**当失业率高于自然水平时由失业的同样变化引起的工资下降。而 β_1 表示工资变化的渐近底限。菲利普斯曲线的这一具体特征可能缘于工会的讨价还价能力、最低工资规定和失业补贴等制度因素。

---

① A. W. Phillips, "The Relationship between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861—1957," *Economica*, November 1958, vol. 15, pp. 283-299. 注意，原始曲线并没有穿过失业轴。而图 6—8 则代表菲利普斯曲线的后来版本。

自从菲利普斯的论文发表以来，从理论和经验上对菲利普斯曲线的研究十分广泛。篇幅所限，不容许我们详细介绍围绕着菲利普斯曲线而展开的争辩。菲利普斯曲线本身也几经演变。一个相对近期的表述由奥利维尔·布兰查德（Olivier Blanchard）提供。[①] 如果我们令 π_t 表示 t 时期的通货膨胀率，其定义是价格水平（由一个有代表性的价格来度量，如消费者价格指数）的百分比变化，令 UN_t 表示 t 时期的失业率，那么现代版本的菲利普斯曲线可表示为如下格式：

$$\pi_t - \pi_t^e = \beta_2(\mathrm{UN}_t - U^N) + u_t \tag{6.7.3}$$

其中π_t＝第 t 年的实际通货膨胀率；

π_t^e＝在第 $t-1$ 年对第 t 年通货膨胀率的预期；

UN_t＝第 t 年的实际失业率；

U^N＝第 t 年的自然失业率；

u_t＝随机误差项。[②]

由于 π_t^e 不能直接观测，所以可以从简化假定 $\pi_t^e = \pi_{t-1}$ 开始；即今年的预期通货膨胀率为去年的通货膨胀率；当然，在形成预期时也可以做更复杂的假定，我们在讨论分布滞后模型的第 17 章中讨论这个问题。

将这个假定代入方程（6.7.3），并将回归模型写成标准形式，我们就得到如下估计方程：

$$\pi_t - \pi_{t-1} = \beta_1 + \beta_2 \mathrm{UN}_t + u_t \tag{6.7.4}$$

其中 $\beta_1 = -\beta_2 U^N$。方程（6.7.4）说明，两个时期之间通货膨胀率的变化与当前失业率线性相关。据经验，预计 β_2 为负（为什么?），而 β_1 为正（因为 β_2 为负且 U^N 为正）。

顺便一提，方程（6.7.3）中的菲利普斯曲线在文献中被称为**修正的菲利普斯曲线**（modified Phillips curve），或**附加预期的菲利普斯曲线**（expectations-augmented Phillips curve）（意味着 π_{t-1} 表示预期通货膨胀率）或**加速主义者菲利普斯曲线**（accelerationist Phillips curve）（表明低失业率导致通货膨胀的上升，并因而成为价格水平的加速器）。

例 6.7

作为对修正的菲利普斯曲线的一个说明，我们在表 6—5 中给出 1960—2006 年间的通货膨胀和失业率数据，其中通货膨胀由消费者价格指数的年百分比变化（CPI 膨胀）来度量，失业率指城镇失业率。我们由这些数据可以得到通货膨胀率的变化（$\pi_t - \pi_{t-1}$），并相对城市失业率描点；我们用 CPI 作为对通货膨胀的一种度量。由此得到图 6—9。

恰如所料，通货膨胀率的变化和失业率之间存在负向关系——低失业率导致通货膨胀率的上升，并因此使价格水平加速上升，加速主义者菲利普斯曲线由此得名。

① 参见 Olivier Blanchard, *Macroeconomics*, Prentice Hall, Englewood Cliffs, NJ, 1997, Chap. 17。

② 经济学家相信这个误差项代表某种供给冲击，如 OPEC 1973 年和 1979 年的石油禁运。

表 6—5　　美国通货膨胀率与失业率：1960—2006 年

（对所有城市消费者；除非特别指出，否则令 1982—1984＝100）

观测	通货膨胀率	失业率	观测	通货膨胀率	失业率
1960	1.718	5.5	1984	4.317	7.5
1961	1.014	6.7	1985	3.561	7.2
1962	1.003	5.5	1986	1.859	7.0
1963	1.325	5.7	1987	3.650	6.2
1964	1.307	5.2	1988	4.137	5.5
1965	1.613	4.5	1989	4.818	5.3
1966	2.857	3.8	1990	5.403	5.6
1967	3.086	3.8	1991	4.208	6.8
1968	4.192	3.6	1992	3.010	7.5
1969	5.460	3.5	1993	2.994	6.9
1970	5.722	4.9	1994	2.561	6.1
1971	4.381	5.9	1995	2.834	5.6
1972	3.210	5.6	1996	2.953	5.4
1973	6.220	4.9	1997	2.294	4.9
1974	11.036	5.6	1998	1.558	4.5
1975	9.128	8.5	1999	2.209	4.2
1976	5.762	7.7	2000	3.361	4.0
1977	6.503	7.1	2001	2.846	4.7
1978	7.591	6.1	2002	1.581	5.8
1979	11.350	5.8	2003	2.279	6.0
1980	13.499	7.1	2004	2.663	5.5
1981	10.316	7.6	2005	3.388	5.1
1982	6.161	9.7	2006	3.226	4.6
1983	3.212	9.6			

注：通货膨胀率为 CPI 的年百分比变化。失业率为城市失业率。

资料来源：*Economic Report of the President*，2007，Table B-60，p. 399，for CPI changes and Table B-42，p. 376，for the unemployment rate.

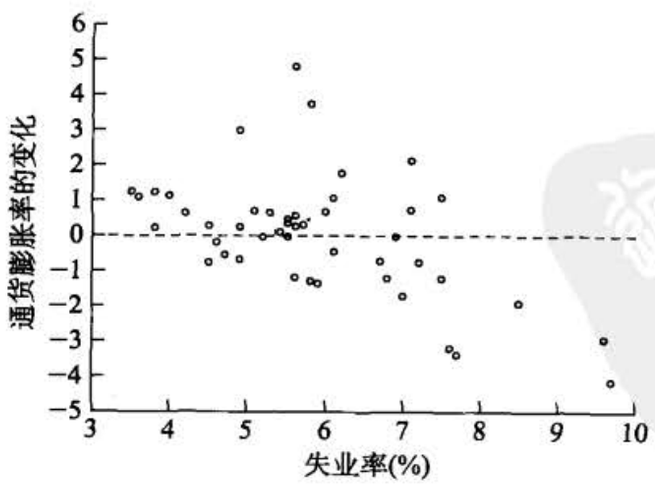


图 6—9　修正的菲利普斯曲线

从图6—9来看，是一个线性（直线）回归模型还是一个倒数模型拟合数据并不明显；这两个变量之间也可能存在曲线关系。我们基于这两个模型给出如下回归。但记住，如第172页注释①所指出的那样，倒数模型的截距项预计为负，斜率为正。

$$\text{线性模型：}\widehat{(\pi_t - \pi_{t-1})} = 3.784\,4 - 0.638\,5\mathrm{UN}_t$$

$$t = (4.191\,2)\ (-4.275\,6) \quad r^2 = 0.293\,5 \tag{6.7.5}$$

$$\text{倒数模型：}\widehat{(\pi_t - \pi_{t-1})} = -3.068\,4 + 17.207\,7\left(\frac{1}{\mathrm{UN}_t}\right)$$

$$t = (-3.163\,5)\ (3.288\,6) \quad r^2 = 0.197\,3 \tag{6.7.6}$$

这两个模型的所有估计系数都是个别统计显著的，所有的 p 值都低于0.005的水平。

模型（6.7.5）表明，若失业率下降1个百分点，则通货膨胀率平均上升约0.64个百分点，反之亦然。模型（6.7.6）表明，即便失业率无限增加，通货膨胀率的最大变化也就是下降约3.07个百分点。顺便提一句，我们从方程（6.7.5）可以计算出其背后的自然失业率为

$$U^N = \frac{\hat{\beta}_1}{-\hat{\beta}_2} = \frac{3.784\,4}{0.638\,5} = 5.927\,0 \tag{6.7.7}$$

即自然失业率约为5.93%。经济学家认为自然失业率介于5%与6%之间，尽管美国最近的实际失业率远低于这个数字。

□ 对数双曲线或对数倒数模型

通过考虑形如下式的对数倒数模型，

$$\ln Y_i = \beta_1 - \beta_2\left(\frac{1}{X_i}\right) + u_i \tag{6.7.8}$$

我们结束对倒数模型的讨论。其形状如图6—10所示。如此图所示，Y 首先以递增的速度增加（即曲线先是凸的），然后以递减的速度增加（即曲线变成凹的）。[①]这类模型可能因此适合于短期生产函数模型。回想在微观经济学中，如果劳动和资本是一个生产函数的投入，而且我们保持资本投入不变但增加劳动投入，那么产出与劳动之间的短期关系就类似图6—10。（见第7章例7.3。）

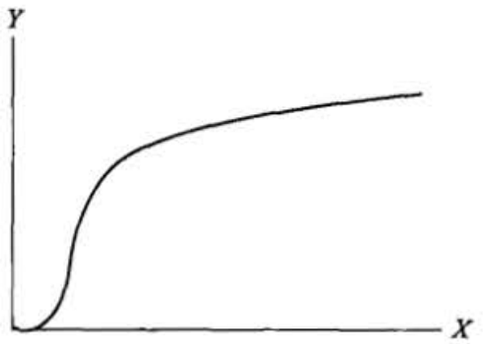


图6—10　对数倒数模型

① 由微分学可以得到 $\frac{\mathrm{d}}{\mathrm{d}X}(\ln Y) = -\beta_2\left(-\frac{1}{X^2}\right) = \beta_2\left(\frac{1}{X^2}\right)$，但是 $\frac{\mathrm{d}}{\mathrm{d}X}(\ln Y) = \frac{1}{Y}\frac{\mathrm{d}Y}{\mathrm{d}X}$，替换后得到 $\frac{\mathrm{d}Y}{\mathrm{d}X} = \beta_2\frac{Y}{X^2}$，这就是 Y 对 X 回归的斜率。

6.8 函数形式的选择

我们在本章讨论了经验模型可以利用的几种函数形式（它们都是参数的线性回归模型）。在双变量情形中，由于通过对变量描点就能基本知道哪个模型合适，所以特定函数形式的选择就相对容易。当我们考虑涉及不止一个回归元的多元回归模型时，这种选择将困难得多，我们在下面两章中讨论这个问题时将会认识到这一点。不可否认，在对经验估计选择适当模型时，需要大量的技巧和经验，但仍有一些指导原则可以参考：

1. 模型背后的理论（如菲利普斯曲线）可能给出了一个特定的函数形式。

2. 最好能求出回归子相对回归元的变化率（即斜率）和回归子对回归元的弹性。我们在表 6—6 中针对本章考虑的各种模型列出了其斜率和弹性系数的公式。了解这些公式将有助于我们比较各种不同的模型。

表 6—6

模型	方程	斜率 $\left(=\frac{dY}{dX}\right)$	弹性 $\left(=\frac{dY}{dX}\cdot\frac{X}{Y}\right)$
线性	$Y=\beta_1+\beta_2 X$	β_2	$\beta_2\left(\frac{X}{Y}\right)^*$
对数线性（对数－对数）	$\ln Y=\beta_1+\beta_2\ln X$	$\beta_2\left(\frac{Y}{X}\right)$	β_2
线性到对数	$\ln Y=\beta_1+\beta_2 X$	$\beta_2(Y)$	$\beta_2(X)^*$
对数到线性	$Y=\beta_1+\beta_2\ln X$	$\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{1}{Y}\right)^*$
倒数	$Y=\beta_1+\beta_2\left(\frac{1}{X}\right)$	$-\beta_2\left(\frac{1}{X^2}\right)$	$-\beta_2\left(\frac{1}{XY}\right)^*$
对数倒数	$\ln Y=\beta_1-\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{Y}{X^2}\right)$	$\beta_2\left(\frac{1}{X}\right)^*$

注：* 表示弹性系数是可变的，它依赖于 X 或 Y 或二者的取值。在 X 和 Y 未给定时，实践中常常在均值 $\bar{X}$ 和 $\bar{Y}$ 处测度这些弹性。

3. 所选模型的系数应该满足一定的先验预期。比如，如果我们考虑对汽车的需求是价格和其他变量的函数，那我们应该预期价格变量的系数为负。

4. 有时不止一个模型都能相当不错地拟合一个给定的数据集。在修正的菲利普斯曲线中，我们对同样的数据拟合了一个线性模型和一个倒数模型。在这两种情况下，系数都与先验预期相一致，也都是统计显著的。一个重要的区别在于，线性模型的 r^2 值比倒数模型的 r^2 值大。因此人们略微倾向于使用线性模型。但一定要注意，在比较两个 r^2 值时，两个模型的因变量或回归子必须相同；回归元则可采用任

何形式。我们在下一章将解释其原因。

5. 通常不应该过分强调 r^2 这个指标，也就是说，并非模型的 r^2 值越大越好。如我们在下一章中将讨论的那样，当我们在模型中添加更多的回归元时，r^2 不断地提高。更重要的地方在于所选模型的理论基础、估计系数的符号及其统计显著性。如果一个模型从这些准则来看不错，那么较低的 r^2 值也是完全可以接受的。我们将在第 13 章更深入地讨论这个重要问题。

6. 在有些情形中，确定一个特定的函数形式不是那么容易，此时，我们或许可以使用所谓的博克斯-考克斯变换（Box-Cox transformations）。由于这个专题相当技术化，所以我们在附录 6A.5 节中予以讨论。

*6.9 关于随机误差项性质的一个注记：加式与乘式随机误差项

考虑如下回归模型

$$Y_i = \beta_1 X_i^{\beta_2} \tag{6.9.1}$$

这是一个与方程（6.5.1）相同但没有误差项的模型。为便于估计，可把此模型表达成三种不同的形式：

$$Y_i = \beta_1 X_i^{\beta_2} u_i \tag{6.9.2}$$

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i} \tag{6.9.3}$$

$$Y_i = \beta_1 X_i^{\beta_2} + u_i \tag{6.9.4}$$

对这些方程两边取对数得：

$$\ln Y_i = \alpha + \beta_2 \ln X_i + \ln u_i \tag{6.9.2a}$$

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i \tag{6.9.3a}$$

$$\ln Y_i = \ln (\beta_1 X_i^{\beta_2} + u_i) \tag{6.9.4a}$$

其中 $\alpha = \ln\beta_1$。

像方程（6.9.2）这样的模型是本质上对参数而言线性的回归模型，因为通过适当的（对数）变换即可将该模型变成参数 α 和 β_2 的线性函数。（注：这些模型对 β_1 而言是非线性的。）但模型（6.9.4）是本质上对参数而言非线性的。因为 $\ln(A+B) \neq \ln A + \ln B$，所以没有对方程（6.9.4）取对数的简单方法。

虽然方程（6.9.2）和（6.9.3）同是线性回归模型，并且都可用 OLS 或 ML 加以估计，但我们必须注意进入模型的随机误差项的性质。记得 OLS 的 BLUE 性质要求 u_i 有零均值、恒定方差和零自相关。对于假设检验，我们还假定 u_i 是正态分布的，并有方才所说的均值和方差。简言之，我们假定 $u_i \sim N(0, \sigma^2)$。

* 选读内容，全书同。

现在来考虑模型（6.9.2）及其统计上的对应方程（6.9.2a）。为了利用经典正态线性回归模型，我们必须假定：

$$\ln u_i \sim N(0, \sigma^2) \tag{6.9.5}$$

因此，当我们做回归（6.9.2a）时，我们必须把第5章讨论的正态性检验应用到从这个回归得到的残差中。顺便指出，如果 $\ln u_i$ 服从零均值和恒定方差的正态分布，则统计学理论证明方程（6.9.2）中的 u_i 必然服从均值为 $e^{\sigma^2/2}$ 且方差为 $e^{\sigma^2} \cdot (e^{\sigma^2} - 1)$ 的**对数正态分布**（log-normal distribution）。

上述分析表明，为了进行回归分析，我们必须十分留意变换一个模型时的误差项。至于方程（6.9.4），这个本质上对参数而言非线性的回归模型，还必须用某种计算机迭代程序来求解。模型（6.9.3）理应没有什么估计上的问题。

总之，当你为回归分析而变换一个模型时，要对干扰项给予非常仔细的注意，否则，对变换后的模型盲目地应用OLS将不会得到一个统计性质优良的模型。

要点与结论

本章介绍了经典线性回归模型中的若干更为细致的问题。

1. 有时一个回归模型并不明显包含截距项。这样的模型被称为过原点回归。虽然估计这种模型的代数方法很简单，但应小心使用这些模型。对于这种模型，残差和 $\sum \hat{u}_i$ 是非零的；此外，通常计算的 r^2 不一定有意义。除非有很强的理论原因，否则还是在模型中明显地引入一个截距为好。

2. 因为单位和尺度是回归系数赖以解释的关键，所以用什么单位和尺度来表达回归子和回归元是很重要的。在经验研究中，研究者不仅要注明数据的来源，还要声明变量是怎样度量的。

3. 同样重要的是回归子与回归元之间的函数关系式。本章讨论的一些重要的函数形式是：（a）对数线性或恒定弹性模型；（b）半对数回归模型；（c）倒数模型。

4. 在对数线性模型中，回归子和回归元都用对数形式来表达。附着于对数回归元的回归系数被解释为回归子对回归元的弹性。

5. 在半对数模型中，或者是回归子或者是回归元以对数形式出现。在回归子为对数形式且回归元为时间的半对数模型中，所估计的斜率系数（乘以100）度量着回归子的（瞬时）增长率。这样的模型常被用来度量许多经济现象的增长率。在半对数模型中，如果回归元是对数形式，它的系数就测出回归元取值的给定百分比变化所引起的回归子的绝对变化率。

6. 在倒数模型中，或者将回归子或者将回归元表达为倒数或反比形式，以刻画经济变量之间的非线性关系，如同著名的菲利普斯曲线那样。

7. 在选择各种函数形式时，须对随机干扰项 u_i 给予高度关注。如在第5章中所指出的那样，CLRM明确地假定了干扰项有零均值和恒定方差（同方差性），并且它与回归元不相关。在这种假定下OLS估计量才是BLUE。此外，在CNLRM下，OLS估计量还是正态分布的。因此，在选择函数形式进行经验分析时，需要查对一下这些假定是否成立。在算完一个回归之后，还应做诊断性检验，如第5章中讨论的正态性检验。这一点再三强调也不为过，因为经典的假设检验，诸如

t、F 和 χ^2 检验都是以干扰项的正态分布为依据的。这对小样本情形尤为重要。

8. 虽然直到现在为止的讨论都限于双变量回归模型，随后各章将表明，在许多情况下，把我们的讨论推广到多元回归模型上也不过涉及更多的代数，而无须引进更多的基本概念。这说明读者牢牢掌握双变量回归模型是非常重要的。

习　题

问答题

6.1　考虑回归模型：

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

其中 $y_i = (Y_i - \overline{Y})$，$x_i = (X_i - \overline{X})$。这时回归线必定经过原点。该观点正确或错误？给出你的计算。

6.2　根据 1978 年 1 月至 1987 年 12 月每月数据获得以下回归结果：

$$\hat{Y}_t = 0.006\ 81 + 0.758\ 15X_t$$

$$\text{se} = (0.025\ 96)\ (0.270\ 09)$$

$$t = (0.262\ 29)\ (2.807\ 00)$$

$$p\text{值} = (0.798\ 4)\ (0.018\ 6) \quad r^2 = 0.440\ 6$$

$$\hat{Y}_t = 0.762\ 14X_t$$

$$\text{se} = (0.265\ 799)$$

$$t = (2.954\ 08)$$

$$p\text{值} = (0.013\ 1) \quad r^2 = 0.436\ 84$$

其中 Y=德士古（Texaco）普通股的月回报率，%，

X=市场回报率，%。①

a. 这两个回归模型有什么区别？

b. 给定上述结果，你会在第一个模型中保留截距项吗？为什么？

c. 你怎样解释这两个模型的斜率系数？

d. 两个模型所依据的理论是什么？

e. 你能不能比较两模型的 r^2 项？为什么？

f. 在此问题中第一个模型的雅克-贝拉正态性统计量是 1.116 7，而第二个模型的是 1.117 0。你能从这些统计量中得出什么结论？

g. 在零截距的模型中斜率系数的 t 值约为 2.95，而在有截距的模型中则约为 2.81。你能对这一结果做出合理的解释吗？

6.3　考虑如下回归模型：

$$\frac{1}{Y_i} = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$$

① 所用数据得自如下教材：Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley, Reading, Mass., 1991。

注：Y 和 X 都不为零。

a. 这是一个线性回归模型吗？

b. 你怎样估计这个模型？

c. 随着 X 趋于无穷大，Y 有怎样的行为？

d. 你能给出该模型可能适用的一个例子吗？

6.4　考虑对数线性模型

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

把 Y 画在纵轴上，并把 X 画在横轴上。分别描绘出 $\beta_2=1$，$\beta_2>1$ 和 $\beta_2<1$ 时表现 Y 与 X 之间关系的曲线。

6.5　考虑下列模型：

模型Ⅰ：$Y_i = \beta_1 + \beta_2 X_i + u_i$

模型Ⅱ：$Y_i^* = \alpha_1 + \alpha_2 X_i^* + u_i$

其中 Y^* 和 X^* 是习题 6.7 所定义的标准化变量。试说明 $\hat{\alpha}_2 = \hat{\beta}_2(S_x/S_y)$ 并证明如下命题：虽然回归的斜率系数与原点的变化无关，但与尺度的变化有关。

6.6　考虑下列模型：

$$\ln Y_i^* = \alpha_1 + \alpha_2 \ln X_i^* + u_i^*$$

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

其中 $Y_i^* = w_1 Y_i$，$X_i^* = w_2 X_i$，这里 w 是常数。

a. 构造这两组回归系数与其标准误之间的关系。

b. 两模型的 r^2 有所不同吗？

6.7　在回归（6.6.8）和（6.6.10）之间，你觉得哪个模型好？为什么？

6.8　对回归（6.6.8）检验假设：斜率系数与 0.005 无显著差异。

6.9　能否从方程（6.7.3）所给的菲利普斯曲线，估计出自然失业率？如何估计？

6.10　恩格尔支出曲线把一个消费者在某一商品上的支出同他或她的总收入联系起来。令$Y=$对某一商品的消费支出，$X=$消费者收入，考虑下列模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = \beta_1 + \beta_2 (1/X_i) + u_i$$

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$$

$$\ln Y_i = \ln \beta_1 + \beta_2 (1/X_i) + u_i$$

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

你会选择哪个（些）模型作为恩格尔支出曲线，为什么？（提示：解释各种斜率系数，求出支出的收入弹性，等等。）

6.11　考虑如下模型

$$Y_i = \frac{e^{\beta_1+\beta_2 X_i}}{1+e^{\beta_1+\beta_2 X_i}}$$

它表示一个线性回归模型吗？若否，你能用什么“技巧”使它成为一个线性回归模型？你如何解释由此得到的模型？在什么情况下，这种模型比较合适？

6.12　画出如下模型（为便于说明，我们省略了观测下标 i）：

a. $Y = \beta_1 X^{\beta_2}$，对 $\beta_2 > 1$，$\beta_2 = 1$，$0 < \beta_2 < 1$，…

b. $Y = \beta_1 e^{\beta_2 X}$，对 $\beta_2 > 0$ 和 $\beta_2 < 0$。

讨论何时这些模型比较适合。

6.13 考虑如下回归[①]：

$$SPI_i = -17.8 + 33.2Gini_i$$

$$se = (4.9)\ (11.8) \quad r^2 = 0.16$$

其中 SPI 表示 1960—1985 年间平均社会政治不稳定程度的一个指标，Gini 表示 1975 年或 1970—1980 年间最近年份的基尼系数。样本由 40 个国家组成。

基尼系数度量了收入不均等程度，并介于 0 与 1 之间。它越接近 0，收入就越均等，它越接近 1，收入就越不均等。

a. 你如何解释这个回归？

b. 假设基尼系数从 0.25 提高到 0.55。SPI 上升多少？这在实践中有何含义？

c. 在 5%的显著性水平上，斜率系数估计值是统计显著的吗？给出必要的计算。

d. 基于上述回归，你能认为收入越不平均的国家政治也越不稳定吗？

实证分析题

6.14 给定表 6—7 中的数据。[②]用这些数据拟合以下模型，求出通常的回归统计量并解释结果。

$$\frac{100}{100 - Y_i} = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right)$$

表 6—7

Y_i	86	79	76	69	65	62	52	51	51	48
X_i	3	7	12	17	25	35	45	55	70	120

6.15 为了研究投资率（投资占 GDP 的比例）与储蓄率（储蓄占 GDP 的比例）之间的关系，马丁·费尔德斯坦（Martin Feldstein）和查尔斯·堀冈（Charles Horioka）得到 21 个国家的样本数据。（见表 6—8。）每个国家的投资率是 1960—1974 年间的平均投资率，储蓄率是同期的平均储蓄率。变量 INVRATE 表示投资率，变量 SAVRATE 表示储蓄率。[③]

表 6—8

	SAVRATE	INVRATE
澳大利亚	0.250	0.270
奥地利	0.285	0.282
比利时	0.235	0.224
加拿大	0.219	0.231
丹麦	0.202	0.224
芬兰	0.288	0.305
法国	0.254	0.260
德国	0.271	0.264

① 参见 David N. Weil, *Economic Growth*, Addison Wesley, Boston, 2005, p. 392。

② 节选自 J. Johnston, *Econometric Methods*, 3d ed., McGraw-Hill, New York, 1984, p. 87, 实际上取自牛津大学 1975 年的计量经济学考试题目。

③ Martin Feldstein and Charles Horioka, "Domestic Saving and International Capital Flows," *Economic Journal*, vol., 90, June 1980, pp. 314-329. 数据复制于 Michael P. Murray, *Econometrics: A Modern Introduction*, Addison-Wesley, Boston, 2006。

续前表

	SAVRATE	INVRATE
希腊	0.219	0.248
爱尔兰	0.190	0.218
意大利	0.235	0.224
日本	0.372	0.368
卢森堡	0.313	0.277
荷兰	0.273	0.266
新西兰	0.232	0.249
挪威	0.278	0.299
西班牙	0.235	0.241
瑞典	0.241	0.242
瑞士	0.297	0.297
英国	0.184	0.192
美国	0.186	0.186

注：SAVRATE=储蓄占 GDP 的比例。INVRATE=投资占 GDP 的比例。

a. 将投资率对储蓄率描点。

b. 基于这个描点图，你认为如下模型对这些数据的拟合效果同样好吗？

$$\text{INVRATE}_i = \beta_1 + \beta_2 \text{SAVRATE}_i + u_i$$

$$\ln \text{INVRATE}_i = \alpha_1 + \alpha_2 \ln \text{SAVRATE}_i + u_i$$

c. 估计这两个模型并求出常用的统计量。

d. 你如何解释线性模型中的斜率系数？又如何解释对数线性模型中的斜率系数？对这些系数的解释有何差异？

e. 你如何解释这两个模型的截距？二者的解释有何差异？

f. 你会比较这两个模型中的 r^2 吗？为什么？

g. 假设你想计算投资率对储蓄率的弹性？你如何在线性模型中求这个弹性？又如何在对数线性模型中求这个弹性？注意这个弹性被定义为储蓄率改变 1%导致投资率改变的百分比。

h. 给定这两个回归模型的结果，你更喜欢哪个模型？为什么？

6.16 表 6—9① 给出了各种支出、总支出、收入、家长年龄和子女数的变量定义，样本取自 1980—1982 年间英国家庭支出调查中 1 519 个家庭。实际数据集可在本书网站上找到。数据只包括住在伦敦市区和市郊有 1～2 个子女的家庭。样本不包括自我雇佣和退休家庭。

表 6—9

变量列表
wfood=食物支出预算份额
wfuel=汽油支出预算份额
wcloth=服装支出预算份额
walc=酒类支出预算份额

① 数据来自 Richard Blundell and Krishna Pendakur, "Semiparametric Estimation and Consumer Demand," *Journal of Applied Econometrics*, vol. 13, no. 5, 1998, pp. 435-462。数据复制于 R. Carter Hill, William E. Griffiths, and George G. Judge, *Undergraduate Econometrics*, 2d ed., John Wiley & Sons, New York, 2001。

续前表

变量列表
wtrans＝交通支出预算份额
wother＝其他支出预算份额
totexp＝家庭总支出（近似到 10 英镑）
income＝家庭总的净收入（近似到 10 英镑）
age＝家长年龄
nk＝子女数
预算份额的定义（以食物为例）是
wfood＝食物支出/总支出

a. 利用食物支出与总支出数据，判断表 6—6 中概括的模型中的哪一个能够较好地拟合这些数据？

b. 基于（a）中得到的回归结果，哪个模型看来在这里比较适当？

注：保留这些数据，以便在下一章讨论多元回归时进一步分析使用。

6.17　参考表 6—3。求出耐用品支出增长率。半弹性估计值是多少？解释你的结果。以耐用品支出为回归子和时间为回归元做一个双对数回归说得过去吗？你如何解释此时的斜率系数？

6.18　根据表 6—3 给出的数据求非耐用品支出的增长率，并将得到的结论与习题 6.17 中的结论相比较。

6.19　表 6—10 给出了英国 29 类商品的总消费支出 CONEXP（百万英镑）和广告支出 ADEXP（百万英镑）数据。①

a. 考虑我们在本章所讨论的各种函数形式，哪个函数形式能够拟合表 6—10 中的数据？

b. 估计所选回归模型的参数，并解释你的结果。

c. 如果你取广告支出占总消费支出的比率 RATIO，你将有何发现？有哪种商品的这一比率看上去异常高吗？对于广告支出异常高的商品，有什么特别原因能够解释这些商品的广告支出相对较高？

表 6—10　　英国 29 类商品的广告支出与总消费支出　　（单位：百万英镑）

观测	ADEXP	CONEXP	RATIO
1	87 957.00	13 599.00	0.006 468
2	23 578.00	4 699.000	0.005 018
3	16 345.00	5 473.000	0.002 986
4	6 550.000	6 119.000	0.001 070
5	10 230.00	8 811.000	0.001 161
6	9 127.000	1 142.000	0.007 992
7	1 675.000	143.0000	0.011 713
8	1 110.000	138.0000	0.008 043
9	3 351.000	85.000 00	0.039 424
10	1 140.000	108.000 0	0.010 556

① 这些数据来自 *Advertising Statistics Year Book*，1996。复制于 http://www.economicswebinstitute.org/ecdata.htm。

续前表

观测	ADEXP	CONEXP	RATIO
11	6 376.000	307.000 0	0.020 769
12	4 500.000	1 545.000	0.002 913
13	1 899.000	943.000 0	0.002 014
14	10 101.00	369.000 0	0.027 374
15	3 831.000	285.000 0	0.013 442
16	99 528.00	1 052.000	0.094 608
17	15 855.00	862.000 0	0.018 393
18	8 827.000	84.000 00	0.105 083
19	54 517.00	1 174.000	0.046 437
20	49 593.00	2 531.000	0.019 594
21	39 664.00	408.000 0	0.097 216
22	327.000 0	295.000 0	0.001 108
23	22 549.00	488.000 0	0.046 207
24	416 422.0	19 200.00	0.021 689
25	14 212.00	94.000 00	0.151 191
26	54 174.00	5 320.000	0.010 183
27	20 218.00	357.000 0	0.056 633
28	11 041.00	159.000 0	0.069 440
29	22 542.00	244.000 0	0.092 385

注：ADEXP=广告支出（百万英镑）；CONEXP=总消费支出（百万英镑）。
资料来源：http：//www. economicswebinstitute. org/ecdata. htm.

6.20 参考第3章例3.3，回答如下问题。

a. 将手机需求相对经购买力调整的人均收入描图。

b. 将手机需求的对数相对经购买力调整的人均收入的对数描图。

c. 这两个图有何差异？

d. 根据这两个图，你认为双对数模型比线性模型能够更好地拟合这些数据吗？估计双对数模型。

e. 你如何解释双对数模型中的斜率系数？

f. 在5%的显著性水平上，双对数模型中的斜率系数估计值是统计显著的吗？

g. 在方程（3.7.3）给出的线性模型中，你如何估计手机需求对经购买力调整的收入的弹性？如果你还需要其他信息，是什么样的信息呢？如此估计的弹性被称为收入弹性。

h. 从双对数模型和线性模型中估计的收入弹性有差别吗？如果有，你将选择哪个模型？

6.21 参考方程（3.7.4）中给出的个人计算机需求，重新回答习题6.20中的问题。手机和个人计算机的估计收入弹性有差别吗？如果有，哪些因素可以解释这种差别？

6.22 参考表3—3中的数据。为了弄清楚拥有个人计算机的人是否也拥有手机，做如下回归：

$$\text{CellPhone}_i = \beta_1 + \beta_2 \text{PCs}_i + u_i$$

a. 估计此回归的参数。

b. 斜率系数估计值是统计显著的吗？

c. 如果你做如下回归也没有关系吗？

$$PCs_i = \alpha_1 + \alpha_2\ \text{CellPhone}_i + u_i$$

d. 估计上述回归并检验斜率系数估计值的统计显著性。

e. 你在这两个回归之间如何选择？

附录 6A

□ 6A.1 过原点回归的最小二乘估计量的推导

我们想选择 $\hat{\beta}_2$ 使得下式尽可能最小化：

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_2 X_i)^2 \tag{1}$$

求（1）对 $\hat{\beta}_2$ 的导数得：

$$\frac{d\sum \hat{u}_i^2}{d\hat{\beta}_2} = 2\sum (Y_i - \hat{\beta}_2 X_i)(-X_i) \tag{2}$$

令方程（2）等于零并化简得：

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2} \qquad (6.1.6) = (3)$$

将 PRF：$Y_i = \beta_2 X_i + u_i$ 代入此方程得：

$$\hat{\beta}_2 = \frac{\sum X_i(\beta_2 X_i + u_i)}{\sum X_i^2} = \beta_2 + \frac{\sum X_i u_i}{\sum X_i^2} \tag{4}$$

[注：$E(\hat{\beta}_2) = \beta_2$。] 因此，

$$E(\hat{\beta}_2 - \beta_2)^2 = E\left[\frac{\sum X_i u_i}{\sum X_i^2}\right]^2 \tag{5}$$

将方程（5）的右端展开，并注意到 X_i 是非随机的以及 u_i 具有同方差性且无自相关，故得：

$$\text{var}(\hat{\beta}_2) = E(\hat{\beta}_2 - \beta_2)^2 = \frac{\sigma^2}{\sum X_i^2} \qquad (6.1.7) = (6)$$

顺便指出，令方程（2）等于零即有：

$$\sum \hat{u}_i X_i = 0 \tag{7}$$

从附录 3A 第 3A.1 节我们看到，当截距项在模型中出现时，除了方程（7）以外，还可以得到条件 $\sum \hat{u}_i = 0$。根据刚才的数学推导，应该很容易明白为什么过原点回归模型的误差总和 $\sum \hat{u}_i$ 不一定为零。

假设我们要增加 $\sum \hat{u}_i = 0$ 这个条件，那么我们便得到：

$$\begin{aligned}\sum Y_i &= \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= \hat{\beta}_2 \sum X_i \quad \text{因为根据假定有} \sum \hat{u}_i = 0\end{aligned} \tag{8}$$

于是这一表达式给出：

$$\hat{\beta}_2 = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}} = \frac{Y\text{ 的均值}}{X\text{ 的均值}} \tag{9}$$

但这个估计量和方程（3）或（6.1.6）并不一样。而由于方程（3）中的 $\hat{\beta}_2$ 是无偏的（为什么?），所以方程（9）中的 $\tilde{\beta}_2$ 不可能是无偏的。

要点在于：在过原点回归中，我们不可能像惯用的模型那样，同时令 $\sum a_i X_i$ 和 $\sum a_i$ 都等于零。得到满足的唯一条件是 $\sum a_i X_i$ 等于零。

回忆：

$$Y_i = \hat{Y}_i + a \tag{2.6.3}$$

两边求和再除以样本容量 N 便得到：

$$\bar{Y} = \bar{\hat{Y}} + \bar{a} \tag{10}$$

因为对零截距模型来说 $\sum a_i$ 从而 $\bar{a}$ 不一定为零，所以：

$$\bar{Y} \neq \bar{\hat{Y}} \tag{11}$$

也就是说，实际的 Y 均值不一定等于估计的 Y 均值；但对有截距的模型而言，这两个均值是等同的，这可从方程（3.1.10）看出。

前面说过，对零截距模型来说，r^2 可能是负的，而对惯用的模型来说，它永远不会是负的。下面说明这一情况。

利用方程（3.5.5a），可写为：

$$r^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum a_i^2}{\sum y_i^2} \tag{12}$$

对于惯用的或含有截距的模型，方程（3.3.6）表明：除非 $\hat{\beta}_2$ 是零（即 X 对 Y 无任何影响），否则总有

$$\text{RSS} = \sum a_i^2 = \sum y_i^2 - \tilde{\beta}_2^2 \sum x_i^2 \leqslant \sum y_i^2 \tag{13}$$

也就是说，对于惯用的模型，RSS≤TSS，或者说，r^2 不可能是负的。

对于无截距模型，可以类似地证明：

$$\text{RSS} = \sum a_i^2 = \sum Y_i^2 - \tilde{\beta}_2^2 \sum X_i^2 \tag{14}$$

[注：Y 和 X 的平方和并未经过均值调整。] 现在并不能保证 RSS 一定小于 $\sum y_i^2 = \sum Y_i^2 - N\bar{Y}^2$（即 TSS）。这就表明，RSS 可能大于 TSS，也就意味着惯用的定义 r^2 可能是负的。其实不难看出，如果 $\tilde{\beta}_2^2 \sum X_i^2 < N\bar{Y}^2$，RSS 将大于 TSS。

□ 6A.2 证明标准化变量的均值为零和方差为 1

考虑随机变量 Y，其（样本）均值为 $\bar{Y}$，（样本）标准差为 S_y。定义

$$Y_i^* = \frac{Y_i - \bar{Y}}{S_y} \tag{15}$$

因此，Y_i^* 就是一个标准化变量。注意标准化涉及双重变化：（1）原点即方程（15）的分子发生变化；（2）分母所代表的尺度也发生了变化。因此，标准化同时涉及原点和尺度的变化。

现在，由于一个变量与其均值的离差之和恒等于零，所以

$$\bar{Y}_i^* = \frac{1}{S_y} \frac{\sum (Y_i - \bar{Y})}{n} = 0 \tag{16}$$

因此标准化变量的均值为零。（注：我们之所以能把 S_y 项放到求和符号之外，因为它的值是已知的。）

于是

$$S_{y^*}^2 = \sum \frac{(Y_i - \bar{Y})^2/(n-1)}{S_y^2}$$

$$= \frac{1}{(n-1)S_y^2} \sum (Y_i - \bar{Y})^2 \tag{17}$$

$$= \frac{(n-1)S_y^2}{(n-1)S_y^2} = 1$$

注意

$$S_y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

是 Y 的样本方差。

□ 6A.3 对数

考虑数字 5 和 25。我们知道

$$25 = 5^2 \tag{18}$$

我们说指数 2 是 25 以 5 为底的对数。更规范地，一个数字（如 25）以给定底的对数就是要得到这个数字（25）而必须给底（5）赋予的指数（2）。

更一般地，如果

$$Y = b^X \quad (b > 0) \tag{19}$$

那么

$$\log_b Y = X \tag{20}$$

在数学上，函数（19）被称为指数函数，而函数（20）被称为对数函数。从方程（19）和（20）显然可见，这两个函数互为反函数。

尽管任何一个（正）底都可以使用，但实践中两个常用的底是 10 和数字 e=2.718 28…。

以 10 为底的对数被称为常用对数。因此，

$$\log_{10} 100 = 2 \qquad \log_{10} 30 \approx 1.48$$

也就是说，在第一种情形中有 $100 = 10^2$，而在第二种情形中有 $30 \approx 10^{1.48}$。

以 e 为底的对数被称为自然对数。因此，

$$\log_e 100 \approx 4.605\ 1 \qquad \log_e 30 \approx 3.401\ 2$$

所有这些计算都可用一个计算器来完成。

根据惯例，以 10 为底的对数用字母 log 表示，而以 e 为底的对数用 ln 表示。因此，在上例中，我们可以把它们写成 log100、log30、ln100 或 ln30。

常用对数与自然对数之间有一个固定关系，即

$$\ln X = 2.302\ 6 \log X \tag{21}$$

也就是说，数字 X 的自然对数等于 2.302 6 乘以 X 以 10 为底的常用对数。于是，和前面一样

$$\ln 30 = 2.302\ 6 \log 30 = 2.302\ 6 \times 1.48 \approx 3.4012 \text{（近似）}$$

因此，使用常用对数还是自然对数没有实质性的影响。但数学上通常喜欢使用以 e 为底，即自然对数。因此，在本书中，除非明确说明，否则，所有对数都是自然对数。当然，利用方程（21），使用这两个底的对数可相互转换。

记住，负数的对数没有定义。因此，log(−5) 或 ln(−5) 都没有意义。

对数有如下性质：如果 A 和 B 是任意正数，那么可以证明：

1. $\ln(A \times B) = \ln A + \ln B$ (22)

也就是说，两个（正）数 A 和 B 乘积的对数等于它们的对数之和。

2. $\ln(A/B) = \ln A - \ln B$ (23)

也就是说，A 和 B 之比的对数等于它们的对数之差。

3. $\ln(A \pm B) \neq \ln A \pm \ln B$ (24)

也就是说，A 和 B 之和或差的对数不等于它们的对数之和或差。

4. $\ln(A^k) = k\ln A$ (25)

也就是说，A 的 k 次幂的对数等于 k 乘以 A 的对数。

5. $\ln e = 1$ (26)

也就是说，e 以它自己为底的对数等于 1（就像 10 的常用对数等于 1 一样）。

6. $\ln 1 = 0$ (27)

也就是说，1 的自然对数等于 0（就像 1 的常用对数等于 0 一样）。

7. 若 $Y = \ln X$，则

$$dY/dX = 1/X \quad (28)$$

也就是说，Y 相对 X 的变化率（即导数）就是 X 分之一。图 6A—1 给出了指数函数和（自然）对数函数的图示。

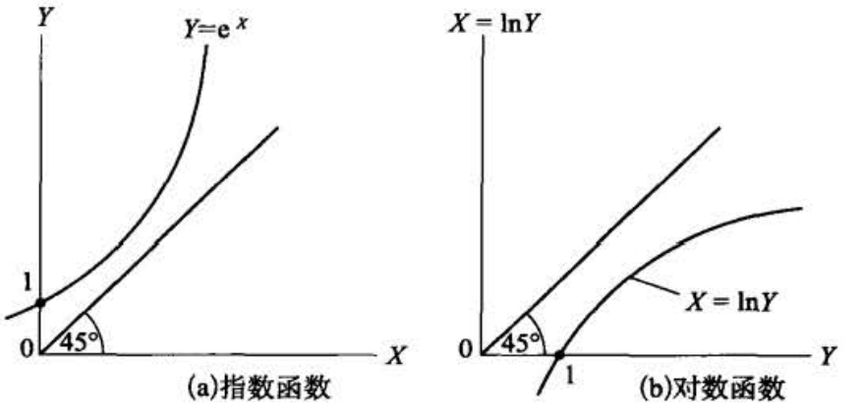


图 6A—1　指数函数与对数函数

尽管只有正数才能取对数，但对数值却可正可负。很容易证明

若 $0<Y<1$　则　$\ln Y<0$

若 $Y=1$　则　$\ln Y=0$

若 $Y>1$　则　$\ln Y>0$

还注意到，尽管图 6A—1（b）中所示的对数曲线斜率为正，即越大的数字，对数值也越大，但该曲线的增加速度越来越慢（从数学上讲，此函数的二阶导数为负）。于是，$\ln 10 = 2.302\,6$（近似），而 $\ln 20 = 2.995\,7$（近似）。也就是说，一个数字加倍，但其对数值不会加倍。

这就是对数变换为什么被称为非线性变换的原因所在。由方程（28）也能看出这一点，即若 $Y = \ln X$，则 $dY/dX = 1/X$。这就意味着对数函数的斜率取决于 X 值；当然就不是常数（回忆变量的线性定义）。

对数与百分比变化。既然 $d(\ln X)/dX = 1/X$，或 $d(\ln X) = dX/X$，所以对于 X 的一个很小变化，$\ln X$ 的变化就等于 X 的相对变化或比例变化。在实践中，如果 X 的变化足够小，上述关系式表明：$\ln X$ 的变化近似等于 X 的相对变化。

于是，对于 X 的一个较小变化，有

$\ln X_t - \ln X_{t-1} \approx (X_t - X_{t-1})/X_{t-1} = X$ 的相对变化

□ 6A.4 增长率表达式

令变量 Y 表示时间的一个函数，即 $Y=f(t)$，其中 t 表示时间。Y 的瞬时（即在一个时点上的）增长率 g_Y 被定义为

$$g_Y = \frac{\frac{dY}{dt}}{Y} = \frac{1}{Y}\frac{dY}{dt} \tag{29}$$

注意，如果我们把 g_Y 乘以 100，就得到增长百分数，其中 dY/dt 表示 Y 相对时间的变化率。

现在，如果我们令 $\ln Y=\ln f(t)$，其中 ln 表示自然对数，那么

$$\frac{d\ln Y}{dt} = \frac{1}{Y}\frac{dY}{dt} \tag{30}$$

它正好与方程（29）相同。

因此，对数变换在计算增长率时非常有用，特别是在 Y 是某些与时间有关的变量的函数时，下面的例子就是一个很好的说明。令

$$Y=X\cdot Z \tag{31}$$

其中 Y 表示名义 GDP，X 表示真实 GDP，而 Z 表示（GDP）价格缩减指数。用语言表述，就是名义 GDP 等于真实 GDP 乘以（GDP）价格缩减指数。所有这些变量都是时间的函数，并随着时间的变化而变化。

现在将方程（31）的两边取对数，我们得到

$$\ln Y=\ln X+\ln Z \tag{32}$$

将方程（32）的两边同时对时间求导，我们得到

$$\frac{1}{Y}\frac{dY}{dt} = \frac{1}{X}\frac{dX}{dt} + \frac{1}{Z}\frac{dZ}{dt} \tag{33}$$

即 $g_Y=g_X+g_Z$，其中 g 表示增长率。

用文字表述，即 Y 的瞬时增长率等于 X 的瞬时增长率和 Z 的瞬时增长率之和。在本例中，就是名义 GDP 的瞬时增长率等于真实 GDP 的瞬时增长率与 GDP 价格缩减指数的瞬时增长率之和。

更一般地，乘积的瞬时增长率等于各个部分的瞬时增长率之和。这个结论可推广到多于两个变量的乘积。

类似地，如果我们有

$$Y=X/Z \tag{34}$$

那么

$$\frac{1}{Y}\frac{dY}{dt} = \frac{1}{X}\frac{dX}{dt} - \frac{1}{Z}\frac{dZ}{dt} \tag{35}$$

即 $g_Y=g_X-g_Z$。换言之，Y 的瞬时增长率等于 X 的瞬时增长率和 Z 的瞬时增长率之差。于是，如果 Y 表示人均收入，X 表示 GDP，Z 表示人口，则人均收入的瞬时增长率就等于 GDP 的瞬时增长率与人口的瞬时增长率之差。

现在令 $Y=X+Z$。Y 的增长率是多少呢？如果 Y 表示总就业量，X 表示蓝领就业量，Z 表示白领就业量。由于

$$\ln(X+Z)\neq\ln X+\ln Z$$

所以不太容易计算 Y 的增长率，但在数学上可以证明

$$g_Y = \frac{X}{X+Z}g_X + \frac{Z}{X+Z}g_Z \tag{36}$$

即和的增长率等于其各个加数的增长率的加权平均。对于我们的例子而言，总就业量的增长率等于白领就业量增长率与蓝领就业量增长率的加权平均，权重是两种就业在总就业量中所占的比例。

□ 6A.5 博克斯-考克斯回归模型

考虑如下回归模型：

$$Y_i^{\lambda} = \beta_1 + \beta_2 X_i + u_i, \quad Y > 0 \tag{37}$$

其中 λ 是一个可正可负亦可为零的参数。由于 Y 的幂指数为 λ，所以根据 λ 值的不同，我们对 Y 有不同的变换。

方程（37）被称为博克斯-考克斯回归模型，因统计学家博克斯和考克斯而得名。[①]根据 λ 值的不同，我们得到下表所示的各种回归模型：

λ 值	模型
1	$Y_i = \beta_1 + \beta_2 X_i + u_i$
2	$Y_i^2 = \beta_1 + \beta_2 X_i + u_i$
0.5	$\sqrt{Y_i} = \beta_1 + \beta_2 X_i + u_i$
0	$\ln Y_i = \beta_1 + \beta_2 X_i + u_i$
−0.5	$1/\sqrt{Y_i} = \beta_1 + \beta_2 X_i + u_i$
−1	$1/Y_i = \beta_1 + \beta_2 X_i + u_i$

你可以看到，线性和对数线性模型是博克斯-考克斯变换模型中的两个特殊情形而已。

当然，我们也可以对 X 变量进行这种变换。注意到，有意思的是，当 λ 为零时，我们得到 Y 的对数变换。证明略显复杂，最好留给读者查阅相关文献。[擅长微积分的读者在证明时注意要援引洛必达法则（l'Hopital Rule）。]

但在一个给定情形中，我们实际上该如何确定适当的 λ 值呢？我们不能直接估计方程（37），因为这个方程不仅包含参数 β_1 和 β_2，还包含参数 λ，而且参数 λ 是以非线性形式进入方程的。不过可以证明，我们可以利用极大似然法估计所有参数。有回归软件专门进行这种估计。

由于估计程序多少有些复杂，我们在此不再深究。

然而，我们也可以利用试错法进行估计。选择几个 λ 值，并相应地对 Y 进行变换，做回归方程（37）便得到变换后每个回归的残差平方和。选择得到最小残差平方和的 λ 值即可。[②]

① G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society*, B26, 1964, pp. 211-243.

② 一个容易接受的讨论参见 John Neter, Michael Kutner, Christopher Nachtsheim, and William Wasserman, *Applied Linear Regression Models*, 3rd ed., Richard D. Irwin, Chicago, 1996。

第 7 章 多元回归分析：估计问题

前面各章所广泛讨论的双变量模型在实践中往往是不适宜的。比如，在我们的消费—收入一例（例 3.1）中，我们无形地假定只有收入 X 影响着消费 Y。但经济理论少见有这般简单的情形，因为除了收入，还有许多其他变量会影响消费支出。一个显然的变量是消费者的财富。作为另一个例子，对某商品的需求很可能不仅依赖于其价格本身，而且还依赖于其他替代品或互补品的价格、消费者的收入、社会地位，等等。因此，我们需要把这个简单的双变量模型推广到包含多于两个变量的模型。加入更多的变量，就把我们引到多元（多变量）回归模型的讨论中去。也就是说，要讨论因变量或回归子 Y，依赖于两个或更多个解释变量或回归元的模型。

最为简单的多元回归模型是含有一个因变量和两个解释变量的三变量回归模型。在本章和下一章中，我们将研究这种模型，而在第 9 章中，我们将把它推广到多于三个变量的情形。通观全书，我们考虑的是多元线性回归模型，即参数的线性模型；对变量而言，它们可以是线性的，也可以是非线性的。

7.1 三变量模型：符号与假定

将双变量的总体回归函数方程（2.4.2）进行推广，便可写出三变量 PRF 为：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{7.1.1}$$

其中 Y 是因变量，X_2 和 X_3 是解释变量（或回归元），u 是随机干扰项，而 i 指第 i

次观测。当数据为时间序列时，下标 t 指第 t 次观测。[①]

在方程（7.1.1）中 β_1 是截距项。虽然机械地解释，它代表 X_2 和 X_3 均为零时 Y 的均值，但像通常所描述的那样，它给出了所有未包含到模型中的变量对 Y 的平均影响。系数 β_2 和 β_3 被称为**偏回归系数**（partial regression coefficients），其含义稍后解释。

我们继续最初在第 3 章介绍的经典线性回归模型框架中讨论。具体地说，我们做如下假定：

1. 线性回归模型，或模型是参数的线性函数。 (7.1.2)

2. X 值固定或独立于误差项。这里，这个假定意味着 u_i 和每个 X 变量之间的协方差为 0：

$$\text{cov}(u_i, X_{2i}) = \text{cov}(u_i, X_{3i}) = 0 \quad (7.1.3)^{[②]}$$

3. 干扰项 u_i 均值为零，或对每一个 i，都有：

$$E(u_i \mid X_{2i}, X_{3i}) = 0 \quad (7.1.4)$$

4. 同方差性，或 u_i 的方差保持不变：

$$\text{var}(u_i) = \sigma^2 \quad (7.1.5)$$

5. 干扰项之间无自相关或序列相关：

$$\text{cov}(u_i, u_j) = 0, i \neq j \quad (7.1.6)$$

6. 观测次数 n 必须大于待估计参数个数，这里有 3 个参数。 (7.1.7)

7. X 变量的值必须存在变异。 (7.1.8)

此外，我们还提出另外两个要求。

8. X 变量之间不存在完全共线性：

X_2 与 X_3 之间无**精确的线性关系**（exact linear relationship） (7.1.9)

在 7.7 节我们会花更多的时间讨论最后一个假定。

9. 无设定偏误，或：

模型被正确地设定 (7.1.10)

从假定（7.1.2）到假定（7.1.10）的合理性都无异于 3.2 节的讨论。假定（7.1.9）是说，X_2 与 X_3 之间无精确的线性关系，或者涉及不止一个精确线性关系式，专业上称为无共线性或**无多重共线性**（no multicollinearity）。

非正式地说，无共线性是指没有一个解释变量可以被写成模型中其余解释变量的线性组合。正式地说，无共线性的含义是，不存在一组不全为零的数 λ_2 和 λ_3 使得：

$$\lambda_2 X_{2i} + \lambda_3 X_{3i} = 0 \quad (7.1.11)$$

如果这一关系式存在，则说 X_2 和 X_3 是**共线的**（collinear）或线性相关。另一方面，如

① 出于符号上的对称性，方程（7.1.1）也可写成：
$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

② 若 X_2 和 X_3 是非随机的，且方程（7.1.4）成立，则这个假定自动得到满足。

果假定（7.1.8）仅当 $\lambda_2 = \lambda_3 = 0$ 时才成立，则说 X_2 和 X_3 线性独立。

因此，如果：

$$X_{2i} = -4X_{3i} \quad 或 \quad X_{2i} + 4X_{3i} = 0 \tag{7.1.12}$$

这两个变量就是线性相关的。如果这样两个变量包含在一个回归模型中，我们就遇到一种完全共线性，或两回归元之间存在一个精确的线性关系。

虽然我们将在第 10 章中详细讨论多重共线性的问题，但不难从直观上去掌握无多重共线性假定的道理。假设方程（7.1.1）中的 Y、X_2 和 X_3 分别代表消费者的消费支出、收入和财富。在消费支出与收入和财富有线性关系的假设中，经济理论设想财富和收入也许对消费各有一些独立的影响，否则把收入和财富两个变量都包括到模型中来就是没有意义的。在极端的情形中，如果收入与财富之间存在准确的线性关系，我们就只有一个独立变量而不是两个，也就无从区分收入和财富对消费的各自影响了。为了看清楚这点，在消费—收入—财富的表达式中，令 $X_{3i} = 2X_{2i}$。于是回归（7.1.1）就变成：

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 (2X_{2i}) + u_i \\ &= \beta_1 + (\beta_2 + 2\beta_3) X_{2i} + u_i \\ &= \beta_1 + \alpha X_{2i} + u_i \end{aligned} \tag{7.1.13}$$

其中 $\alpha = (\beta_2 + 2\beta_3)$。也就是说，事实上我们有一个双变量而不是三变量的回归。而且，如果我们做回归（7.1.13）并得到 α，那么，α 给出的是 X_2 和 X_3 对 Y 的联合影响，并且没有什么方法能分别估计出 X_2 的单独影响（β_2）和 X_3 的单独影响（β_3）。[①]

总之，无多重共线性假定要求我们在 PRF 中仅仅把那些不是模型中其他变量的线性函数的变量包括进来。对此我们将在第 10 章中作广泛的探讨，但这里有几点值得注意：

首先，无多重共线性的假定是对我们的理论（即 PRF）模型而言。实际上，当我们为经验分析搜集数据时，不能保证回归元之间不存在相关。事实上，我们在本章稍后的说明性例子中将会发现，在多数应用研究中，几乎不可能找到两个或多个在某种程度上不相关的（经济）变量。我们只是要求不存在像方程（7.1.12）中那样精确的线性关系。

其次，记住我们只是在讨论两个或多个变量之间的完全线性关系。多重共线性并不排除变量之间的非线性关系。假设 $X_{3i} = X_{2i}^2$，这就不违背不完全共线性的假定，因为变量之间的关系不是线性的。

7.2 对多元回归方程的解释

给定经典回归模型的假定，那么，在方程（7.1.1）的两边对 Y 求条件期望便

① 从数学的角度说，$\alpha = \beta_2 + 2\beta_3$ 是包含两个未知数的一个方程，因此没有从 α 估计 β_2 和 β_3 的唯一方法。

得到：

$$E(Y_i \mid X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_{3i} X_{3i} \tag{7.2.1}$$

用文字表述，方程（7.2.1）给出**以变量 X_2 和 X_3 的固定值为条件的 Y 的条件均值或期望值**。因此，如同双变量情形那样，多元回归分析是以多个解释变量的固定值为条件的回归分析，并且我们所得到的，是给定回归元值时 Y 的平均值或 Y 的平均响应。

7.3 偏回归系数的含义

前面曾指出，回归系数 β_2 和 β_3 被称为**偏回归**（partial regression）或**偏斜率系数**（partial slope coefficients）。偏回归系数的含义如下：β_2 度量着在 X_3 保持不变的情况下，X_2 每变化 1 单位时，Y 的均值 $E(Y)$ 的变化。换句话说，它给出 X_2 的单位变化对 Y 的均值的“直接”或“净”影响，净在不存在 X_3 的影响。[①] 类似地，β_3 度量着在保持 X_2 不变的情况下 X_3 每变化 1 单位时 Y 的均值的变化。即它给出 X_3 的单位变化对 Y 的均值的“直接”或“净”影响，即在没有 X_2 的影响下 X_3 对 Y 的均值的影响。[②]

我们实际上应如何理解保持一个回归元的影响不变呢？为了解释这一点，让我们回到儿童死亡率的例 6.6。记得在那个例子中，Y＝儿童死亡率（CM），X_2＝人均 GNP（PGNP），X_3＝妇女识字率（FLR）。假设我们想保持 FLR 的影响不变。由于在任何一个给定的具体数据中，FLR 对 CM 和 PGNP 都有影响，我们所能做的是，通过分别做 CM 对 FLR 和 PGNP 对 FLR 的回归，看一下从这些回归中得到的残差，从而消除 CM 和 PGNP 中 FLR 的（线性）影响。利用表 6—4 中的数据，我们得到如下回归：

$$\widehat{\mathrm{CM}}_i = 263.863\,5 - 2.390\,5\mathrm{FLR}_i + \hat{u}_{1i}$$
$$\mathrm{se} = (12.224\,9)\ (0.213\,3) \qquad r^2 = 0.669\,5 \tag{7.3.1}$$

其中 $\hat{u}_{1i}$ 表示此回归的残差项。

$$\widehat{\mathrm{PGNP}}_i = -39.303\,3 + 28.142\,7\mathrm{FLR}_i + \hat{u}_{2i}$$
$$\mathrm{se} = (734.952\,6)\ (12.821\,1) \qquad r^2 = 0.072\,1 \tag{7.3.2}$$

其中 $\hat{u}_{2i}$ 表示此回归的残差项。

现在

$$\hat{u}_{1i} = \mathrm{CM}_i - 263.863\,5 + 2.390\,5\mathrm{FLR}_i \tag{7.3.3}$$

① 习惯于用微积分思考的读者会立即看出，β_2 和 β_3 是 $E(Y \mid X_2, X_3)$ 对 X_2 和 X_3 的偏导数。

② 顺便指出，控制、保持不变、考虑到或解释……的影响、修正……的影响和去掉……的影响等词句是同义语，并将在本书中交替使用。

表示 CM 中除去 FLR（线性）影响余下的部分。类似地，

$$\hat{a}_{2i} = \text{PGNP}_i + 39.303\,3 - 28.142\,7\text{FLR}_i \tag{7.3.4}$$

表示 PGNP 中除去 FLR（线性）影响余下的部分。

因此，如果我们现在做 $\hat{a}_{1i}$ 对 $\hat{a}_{2i}$（即去除 FLR 影响后净的 CM 和 PGNP）的如下回归，我们不就得到 PGNP 对 CM 的净影响了吗？实际上确实如此（见附录 7A 第 7A.2 节）。回归结果如下：

$$\hat{a}_{1i} = -0.005\,6\hat{a}_{2i}$$

$$\text{se} = (0.001\,9) \quad r^2 = 0.115\,2 \tag{7.3.5}$$

注：此回归没有截距项，因为 OLS 残差 $\hat{a}_{1i}$ 和 $\hat{a}_{2i}$ 的均值都为零。（为什么?）

现在的斜率系数−0.005 6 就给出了 PGNP 的单位变化对 CM 的“真实”影响或净影响，或者说 CM 对 PGNP 的真实斜率，即 CM 对 PGNP 的偏回归系数 β_2。

想得到 CM 对 FLR 的偏回归系数的读者，通过将 CM 对 PGNP 回归得到残差 $\hat{a}_{1i}$，再将 FLR 对 PGNP 回归得到残差 $\hat{a}_{2i}$，再将 $\hat{a}_{1i}$ 对 $\hat{a}_{2i}$ 回归即可。我确信读者明白了这一点。

我们每次要求出真实的偏回归系数都必须经过这个多步骤的程序吗？幸运的是，不必如此，通过下一节讨论的 OLS 程序，可以相当迅速而又例行地做到这一切。刚才概括的多步骤程序只是出于教学目的，让读者明白“偏”回归系数的含义。

7.4 偏回归系数的 OLS 与 ML 估计

为了估计三变量回归模型（7.1.1）的参数，我们先考虑第 3 章介绍的普通最小二乘法（OLS），然后再扼要地考虑第 4 章讨论的极大似然法（ML）。

□ OLS 估计量

为了求 OLS 估计量，让我们先写出与方程（7.1.1）的 PRF 相对应的样本回归函数：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{a}_i \tag{7.4.1}$$

其中 $\hat{a}_i$ 是残差项，为随机干扰项 u_i 的样本对应部分。

如第 3 章所看到的，OLS 方法是要选择未知参数的值，以使残差平方和（RSS）$\sum \hat{a}_i^2$ 尽可能小，用符号表示为：

$$\min \sum \hat{a}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2 \tag{7.4.2}$$

其中 RSS 的表达式得自方程（7.4.1）的简单代数运算。

为了求方程（7.4.2）最小化的估计量，最直接的方法是将它对未知数求微分，令所得的表达式为零，然后解联立方程。如同附录 7A 第 7A.1 节证明的那样，此方

法给出如下正规方程［比较方程（3.1.4）和（3.1.5）］：

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 \tag{7.4.3}$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} \tag{7.4.4}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 \tag{7.4.5}$$

由方程（7.4.3）我们立即得到：

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \tag{7.4.6}$$

这就是总体截距 β_1 的 OLS 估计量。

按照用小写字母表示对样本均值离差的惯例，我们从正规方程（7.4.3）至（7.4.5）导出以下的公式：

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{7.4.7}$$①

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{7.4.8}$$

分别给出总体偏回归系数 β_2 和 β_3 的 OLS 估计量。

顺便指出以下性质：（1）可以从方程（7.4.7）和（7.4.8）中的一个方程通过对调 x_2 和 x_3 的位置而得到另一个方程，所以它们本质上是对称的；（2）两个方程的分母完全相同；以及（3）三变量情形是双变量情形的自然推广。

□ OLS 估计量的方差和标准误

得到了偏回归系数的 OLS 估计量，就可按照附录 3A.3 节所指示的方法推出这些估计量的方差和标准误。如同双变量情形，我们计算标准误有两个主要目的：建立置信区间和检验统计假设。有关的公式如下②：

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}\right] \cdot \sigma^2 \tag{7.4.9}$$

$$\text{se}(\hat{\beta}_1) = +\sqrt{\text{var}(\hat{\beta}_1)} \tag{7.4.10}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \sigma^2 \tag{7.4.11}$$

① 这个估计量等于方程（7.3.5）中的估计量，参看附录 7A 第 7A.2 节。

② 这些公式的推导利用矩阵符号较为容易，有基础的读者可参见附录 C。

或等价地

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)} \tag{7.4.12}$$

其中 r_{23} 是在第 3 章中定义的 X_2 和 X_3 的样本相关系数。[①]

$$\text{se}(\hat{\beta}_2) = +\sqrt{\text{var}(\hat{\beta}_2)} \tag{7.4.13}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\sigma^2 \tag{7.4.14}$$

或等价地：

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)} \tag{7.4.15}$$

$$\text{se}(\hat{\beta}_3) = +\sqrt{\text{var}(\hat{\beta}_3)} \tag{7.4.16}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}} \tag{7.4.17}$$

在所有的这些公式中 σ^2 是总体干扰项 u_i 的（同方差性）方差。

参照附录 3A 第 3A.5 节的证明，读者能证明 σ^2 的一个无偏估计量是：

$$\hat{\sigma}^2 = \frac{\sum u_i^2}{n-3} \tag{7.4.18}$$

注意 σ^2 的这一估计量与双变量模型中的对应估计量 $[\hat{\sigma}^2 = (\sum u_i^2)/(n-2)]$ 之间的相似性。现在的自由度是 $n-3$，这是因为在估计 $\sum u_i^2$ 之前，我们必须先估计 β_1，β_2 和 β_3，从而消耗了 3 个自由度。（这种证明很具有一般性，例如，在四变量的情形中，自由度就是 $n-4$。）

一旦算出残差 u_i，就能从方程（7.4.18）算出估计量 $\hat{\sigma}^2$。但更容易算出 $\hat{\sigma}^2$ 的方法是利用下述关系式（证明见附录 7A 第 7A.3 节）：

$$\sum u_i^2 = \sum y_i^2 - \hat{\beta}_2\sum y_i x_{2i} - \hat{\beta}_3\sum y_i x_{3i} \tag{7.4.19}$$

这是方程（3.3.6）所给的关系式在三变量模型中的对应方程。

□ OLS 估计量的性质

多元回归模型的 OLS 估计量和双变量模型的 OLS 估计量有着平行的性质。具体地说：

---

① 由第 3 章给出的 r 的定义，我们有：$r_{23}^2 = \dfrac{(\sum x_{2i}x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2}$

1. 三变量回归线（面）通过均值 $\bar{Y}$、$\bar{X}_2$ 和 $\bar{X}_3$。方程（7.4.3）说明了这一点[与双变量模型的方程（3.1.7）进行比较]。这个性质可以推广到一般情形。例如，在 k 变量线性回归模型［一个回归子和（$k-1$）个回归元］中：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{7.4.20}$$

我们有：

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 - \cdots - \hat{\beta}_k \bar{X}_k \tag{7.4.21}$$

2. 估计的 Y_i（$= \hat{Y}_i$）的均值等于真实 Y_i 的均值，这是容易证明的：

$$\begin{aligned} \bar{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3) + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \quad \text{（为什么?）} \\ &= \bar{Y} + \hat{\beta}_2 (X_{2i} - \bar{X}_2) + \hat{\beta}_3 (X_{3i} - \bar{X}_3) \\ &= \bar{Y} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \end{aligned} \tag{7.4.22}$$

其中，和平常一样，小写字母表示有关变量对各自均值的离差。

将方程（7.4.22）两边对所有样本值求和并除以样本容量 n 即得 $\bar{\hat{Y}} = \bar{Y}$。（注：$\sum x_{2i} = \sum x_{3i} = 0$。为什么?）注意，利用方程（7.4.22），我们可写

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \tag{7.4.23}$$

其中 $y_i = \hat{Y}_i - \bar{Y}$。

因此，SRF 方程（7.4.1）可用离差形式表达为

$$y_i = \hat{y}_i + \hat{u}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \tag{7.4.24}$$

3. $\sum \hat{u}_i = \bar{\hat{u}} = 0$，这可从方程（7.4.24）得到证实。[提示：方程（7.4.24）两边对样本值求和。]

4. 残差 $\hat{u}_i$ 与 X_{2i} 和 X_{3i} 都不相关，即 $\sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = 0$。（其证明见附录 7A.1。）

5. 残差 $\hat{u}_i$ 与 $\hat{Y}_i$ 不相关，即 $\sum \hat{u}_i \hat{Y}_i = 0$。为什么?[提示：方程（7.4.23）两边同时乘以 $\hat{u}_i$，然后对样本值求和。]

6. 根据方程（7.4.12）和（7.4.15），显然随着 X_2 和 X_3 的相关系数 r_{23} 朝着 1 增大，对给定的 σ^2 和 $\sum x_{2i}^2$ 或 $\sum x_{3i}^2$ 来说，$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差也不断增大。在 $r_{23} = 1$ 即完全共线性的极限情形中，这些方差变成无穷大。这种情形的含义将在第 10 章中充分探讨，但直觉上读者能看出，随着 r_{23} 的增加，要知道 β_2 和 β_3 的真值所在将变得越来越困难。[参见方程（7.1.13），但下章有更多的讨论。]

7. 由方程（7.4.12）和（7.4.15）还明显看到，对给定的 r_{23} 和 $\sum x_{2i}^2$ 或 $\sum x_{3i}^2$ 值，OLS 估计值的方差正比于 σ^2，即这些方差随 σ^2 的增加而增加。类似地，对给定的 σ^2 和 r_{23} 值，$\hat{\beta}_2$ 的方差反比于 $\sum x_{2i}^2$，即 X_2 的样本值变化越大，$\hat{\beta}_2$ 的方差越小，并从而能更精确地估计 β_2。关于 $\hat{\beta}_3$ 的方差也可作类似的叙述。

8. 在 7.1 节详细说明的经典线性模型假定下，可以证明偏回归系数的 OLS 估计量不仅是线性的和无偏的，而且在所有线性无偏估计量中具有最小方差。简单地说，

它们是 BLUE：换一种方式说，它们满足高斯-马尔可夫定理。（其证明完全类似于附录 3A 第 3A.6 节所证明的双变量情形，而我们将在第 9 章中用矩阵符号更简洁地给出这个证明。）

□ ML 估计量

在第 4 章中我们曾指出，在总体干扰项 u_i 服从零均值和常数方差 σ^2 的正态分布的假定下，双变量模型回归系数的 ML 估计量和 OLS 估计量是相等的。这种关系可推广到包含任意多个变量的模型中去。（证明见附录 7A 第 7A.4 节。）然而，对 σ^2 的估计量而言则不然。可以证明，不管模型中有多少个变量，σ^2 的 ML 估计量都是 $\sum \hat{u}_i^2/n$，而 σ^2 的 OLS 估计量则对双变量情形为 $\sum \hat{u}_i^2/(n-2)$，对三变量情形为 $\sum \hat{u}_i^2/(n-3)$，对 k 变量模型（7.4.20）情形为 $\sum \hat{u}_i^2/(n-k)$。总之，σ^2 的 OLS 估计量考虑了自由度的个数，而 ML 估计量无此考虑。当然，如果 n 很大，σ^2 的 ML 和 OLS 估计量将趋于一致。（为什么?）

7.5 多元判定系数 R^2 与多元相关系数 R

在双变量的情形中我们曾看到，由方程（3.5.5）定义的 r^2 是回归方程拟合优度的一个指标；即它给出在因变量 Y 的总变异中由（单个）解释变量 X 解释的比例或百分比。容易把 r^2 这个符号推广应用到含有多于两个变量的回归模型中去。因此，在三变量模型中，我们也许想知道 Y 的变异由变量 X_2 和 X_3 联合解释的比例。提供这一信息的数量被称为**多元判定系数**（multiple coefficient of determination），记为 R^2；概念上 R^2 近似于 r^2。

可仿照 3.5 节推导 r^2 的方法来推导 R^2。回忆：

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \tag{7.5.1}$$

其中 $\hat{Y}_i$ 是从所拟合的回归线估计出来的 Y_i 值，它是真实 $E(Y_i \mid X_{2i}, X_{3i})$ 的一个估计量。把它换成小写字母，用以表示对均值的离差，方程（7.5.1）就可写为：

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \tag{7.5.2}$$

将方程（7.5.2）两边平方，再对样本值求和，便得到：

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \quad \text{（为什么?）} \end{aligned} \tag{7.5.3}$$

用文字表述，方程（7.5.3）是说，总平方和等于解释平方和加上残差平方和。用方程（7.4.19）代替 $\sum \hat{u}_i^2$ 则得到：

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$$

整理后得到：

$$\text{ESS} = \sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} \tag{7.5.4}$$

于是，按定义：

$$R^2 = \frac{\text{ESS}}{\text{TSS}}$$

$$= \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \tag{7.5.5}$$[①]

[比较方程（7.5.5）和（3.5.6）。]

因为方程（7.5.5）的各项通常已按既定程序算出，故 R^2 很容易计算。R^2 和 r^2 一样，介于 0 与 1 之间。如果是 1，则所拟合的回归线 100%地解释了 Y 的变异；如果是 0，则模型不解释 Y 的任何变异。典型的情形是 R^2 位于这两个极端值之间。R^2 越靠近 1，我们说模型“拟合”得越好。

回顾双变量情形，我们把 r 这个量定义为相关系数，它度量着两个变量之间的（线性）相关程度。类似于 r 的三变量或多变量指标是**多元相关**（multiple correlation）系数，记为 R。它度量着 Y 和所有解释变量的共同相关程度。虽然 r 可正可负，而 R 则永远取正值。实际上，R 没有多大重要性，更有意义的量是 R^2。

在继续讨论之前，让我们指出 R^2 与 k 变量多元回归模型（7.4.20）中的一个偏回归系数的方差 $\text{var}(\hat{\beta}_j)$ 之间的下述关系：

$$\text{var}\ (\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right) \tag{7.5.6}$$

其中 $\hat{\beta}_j$ 是回归元 X_j 的偏回归系数，而 R_j^2 是 X_j 对其余 $k-2$ 个回归元进行回归的 R^2。（注：k 变量回归模型中有 $k-1$ 个回归元。）虽然方程（7.5.6）的用途要等到第 10 章讲多重共线性时才变得明显，但应看到此方程不过是公式（7.4.12）或（7.4.15）的推广。后面这两个公式适用于三变量回归模型，即一个回归子和两个回归元的模型。

7.6 一个说明性例子

例 7.1 儿童死亡率与人均 GNP 和妇女识字率的关系

我们在第 6 章中考虑了儿童死亡率（CM）与人均 GNP（PGNP）之间的关系。在那里，我们发现 PGNP 如预期般对 CM 有负影响。现在让我们引入由妇女识字率（FLR）度量的妇女文化程度变量。据经验，预计 FLR 对 CM 也具有负影响。现在在模型中同时引入这两个变量之后，我们

① 注意，R^2 还可计算为 $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} = 1 - \frac{(n-3)\hat{\sigma}^2}{(n-1)S_Y^2}$。

就不需要专门净化每个回归元的净影响。即我们需要估计每个回归元的（偏）回归系数。于是我们的模型为：

$$CM_i = \beta_1 + \beta_2\, PGNP_i + \beta_3\, FLR_i + u_i \tag{7.6.1}$$

所需数据由表 6—4 给出。记住，CM 为每 1 000 名儿童中不足 5 岁便死亡的人数，PGNP 为 1980 年的人均 GNP，而 FLR 以百分比度量。我们的样本由 64 个国家构成。

使用 EViews 6 统计软件，我们得到如下结果：

$$\widehat{CM}_i = 263.641\,6 - 0.005\,6\,PGNP_i - 2.231\,FLR_i \tag{7.6.2}$$

$$se = (11.593\,2)\ (0.001\,9)\qquad (0.209\,9)\qquad R^2 = 0.707\,7$$

$$\bar{R}^2 = 0.698\,1$$ ①

括号中的数字为估计的标准误。在我们解释这个回归之前，先观察 PGNP 的偏回归系数－0.005 6。它与我们在上一节中用三步法所得到的结果［见方程（7.3.5）］不完全一样吗？你对此感到惊讶吗？不仅如此，两个标准误也完全相同，这些都无足为奇。但我们没有使用麻烦的三步法也得到了这些结果。

让我们现在来解释这些回归系数：－0.005 6 是 PGNP 的偏回归系数，它告诉我们，保持 FLR 的影响不变，PGNP 提高 1 美元，儿童死亡率平均下降 0.005 6 个单位。为了在经济上更容易解释，若人均 GNP 提高 1 000 美元，则每 1 000 名儿童中不足 5 岁便死亡的儿童数平均下降约 5.6 人。系数－2.231 6 告诉我们，保持 PGNP 的影响不变，妇女识字率每提高 1%，每 1 000 名儿童中不足 5 岁便死亡的儿童数平均减少约 2.23 人。约等于 263 的截距值，机械地解释就是，若 PGNP 和 FLR 固定为零，则每 1 000 个产婴中儿童死亡人数的均值为 263。当然，对这种解释应该有所保留。从实际情况可以推断出来的是，若这两个回归元都固定为零，儿童死亡率应该相当高。约为 0.71 的 R^2 值意味着，儿童死亡率变异中约有 70%可由 PGNP 和 FLR 来解释，考虑到 R^2 的最大值充其量为 1，这个值已相当高了。所有这些都说明回归结果讲得通。

估计系数的统计显著性如何？我们将在第 8 章讨论这个问题。我们在那里会看到，本章在许多方面都是对讨论双变量模型的第 5 章的一个推广。我们还将证明，双变量和多变量回归模型在统计推断（即假设检验）方面有一些重要的差别。

□ 标准化变量的回归

在上一章中，我们介绍了对标准化变量进行回归的问题，并说过这种分析可推广至多元回归的情况。记住，如果将一个变量用它与其均值的离差除以其标准差的形式来表示，就称之为标准化变量或以标准差为单位表示的变量。

对儿童死亡率一例，结果如下：

$$\widehat{CM}^* = -0.202\,6PGNP_i^* - 0.763\,9FLR_i^*$$

$$se = (0.071\,3)\qquad (0.071\,3)\qquad r^2 = 0.707\,7 \tag{7.6.3}$$

注：加星号的变量都是标准化变量。还要注意，出于上一章已经讨论过的原因，此模型中不包含截距项。

① 对 $\bar{R}^2$，可参见 7.8 节。

你从这个回归中可以看到，保持 FLR 不变，PGNP 提高一个标准差，导致 CM 平均下降 0.202 6 个标准差。类似地，保持 PGNP 不变，FLR 提高一个标准差，导致 CM 平均下降 0.763 9 个标准差。相对而言，妇女识字率比人均 GNP 对儿童死亡率的影响更大。这里你将看到使用标准化变量的好处，由于所有标准化变量的均值都是零，方差都是 1，所以标准化使所有变量都处在同一个标准之下。

□ 不止一个回归元的单位变化对因变量的影响

在进一步讨论之前，假设我们想知道：如果 PGNP 和 FLR 同时提高，儿童死亡率会如何变化呢？假设人均 GNP 提高 1 美元，同时妇女识字率也提高 1%。这种同时变化对儿童死亡率有何影响？为了弄清楚这一点，我们所要做的无非就是把 PGNP 和 FLR 的系数乘以各自的变化，再把二者加总起来即可。在本例中，我们得到

$$-0.005\ 6\times 1-2.231\ 6\times 1=-2.237\ 2$$

即 PGNP 和 FLR 这种同时改变的结果是，每 1 000 名儿童中不足 5 岁便死亡的儿童数约降低 2.24 人。

更一般地，如果我们想得到不止一个回归元的单位变化对因变量的影响，我们所要做的无非就是把对应变量的系数相加。注意截距项没有在这些计算中出现。（为什么?）

7.7 从多元回归的角度看简单回归：设定偏误初探

记得经典线性回归模型的假定（7.1.10）声称，分析中所用的回归模型是正确设定的，也就是说，不存在**设定偏误**或**误差**（对此第 3 章曾有过一些初步的评论）。虽然第 13 章将对设定误差问题作更透彻的讨论，但上一节的说明性例子给我们提供机会去深入认识假定（7.1.10）的重要性，不仅如此，还启发我们更好地去认识偏回归系数的含义，从而比较正式地把我们引导到设定偏误的问题上来。

假定（7.6.1）是解释儿童死亡率与人均 GNP 和妇女识字率（FLR）之间关系的“真实”模型。但假设我们去掉 FLR 而估计如下简单回归：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_{1i} \tag{7.7.1}$$

其中 Y=CM，X_2=PGNP。

既然方程（7.6.1）是真实模型，所以估计方程（7.7.1）将构成设定误差；这里的误差因省略妇女识字率变量 X_3 而形成。注意，我们在方程（7.7.1）中使用了不同的参数符号（α），以有别于真实模型（7.6.1）中的参数（β）。

现在 α_2 能给出模型（7.6.1）中 β_2 所表示的 PGNP 的真实影响的一个无偏估计吗？换言之，$E(\hat{\alpha}_2) = \beta_2$ 吗？其中 $\hat{\alpha}_2$ 是 α_2 的估计值。换句话说，知道我们从模型中

省略了变量 X_3（即 FLR）后，方程（7.7.1）中 PGNP 的系数 α_2 是 PGNP 对 CM 真实影响的无偏估计吗？恰如所料，α_2 通常不是 β_2 的无偏估计。为粗略地了解一下偏误，让我们做回归（7.7.1），并得到如下结论：

$$\widehat{CM}_i = 157.424\ 4 - 0.011\ 4\ PGNP_i \qquad (7.7.2)$$
$$se = (9.845\ 5)\ \ (0.003\ 2) \qquad r^2 = 0.166\ 2$$

与“真实”多元回归（7.6.1）相比，从此回归中可观察到如下几点：

1. 从绝对值看（即去掉符号），PGNP 系数从 0.005 6 增加到 0.011 4，几乎扩大一倍。

2. 标准误不同。

3. 截距值不同。

4. r^2 值明显不同，而随着模型中回归元个数的增加，r^2 值通常都会提高。

现在假设你将儿童死亡率对妇女识字率回归，而无视 PGNP 的影响，并得到如下结果：

$$\widehat{CM}_i = 263.863\ 5 - 2.390\ 5\ FLP_i \qquad (7.7.3)$$
$$se = (21.224\ 9)\ \ (0.213\ 3) \qquad r^2 = 0.669\ 6$$

同样，你如果将此（误设）回归的结果与“真实”多元回归的结果相比较，你会看出其差别，只是这里的差别没有回归（7.7.2）中的那么明显。

要指出的重要一点是，错误拟合一个模型会导致严重后果。我们在有关设定误差的第 13 章中将更全面地讨论这个专题。

7.8 R^2 及调整 R^2

R^2 的一个重要性质是，它是出现在模型中的解释变量或回归元的个数的非减函数；随着回归元个数的增加，R^2 几乎必然增加并永不减小。换一种说法，增加一个 X 变量必不会减少 R^2。比如，将回归（7.7.2）或（7.7.3）与（7.6.2）相比较。为了看清楚这点，回忆一下判定系数的定义：

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= 1 - \frac{ESS}{TSS} \\ &= 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \end{aligned} \qquad (7.8.1)$$

这里，$\sum y_i^2$ 就是 $\sum (Y_i - \bar{Y})^2$，与模型中 X 变量的个数无关。但 RSS，即 $\sum u_i^2$，却与模型中出现的回归元个数相关。直观上，显而易见，随着 X 变量个数的增加，

$\sum a_i^2$ 很可能减小（至少不会增加）；（7.8.1）所定义的 R^2 也将随之增加。有鉴于此，在比较有相同因变量但有不同个数的 X 变量的回归时，选择有最高 R^2 值的模型必须当心。

要比较两个 R^2 项，必须考虑到模型中出现的 X 变量的个数。如果我们考虑如下的另一种判定系数，这个问题就很容易解决：

$$\overline{R}^2 = 1 - \frac{\sum a_i^2/(n-k)}{\sum y_i^2/(n-1)} \tag{7.8.2}$$

其中 k=模型中包括截距项在内的参数个数。（在三变量回归中 $k=3$。为什么?）如此定义的 R^2，称**调整 R^2**（adjusted R^2），记为 $\overline{R}^2$。“调整”一词指对方程（7.8.1）中的平方和所涉及的自由度进行调整：在一个包含 k 个参数（包括截距项）的模型中，$\sum a_i^2$ 有 $n-k$ 个自由度，而 $\sum y_i^2$ 有 $n-1$ 个自由度。（为什么?）对于三变量情形，我们知道 $\sum a_i^2$ 有 $n-3$ 个自由度。

方程（7.8.2）又可写为：

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2} \tag{7.8.3}$$

其中 $\hat{\sigma}^2$ 是残差方差，是真实 σ^2 的一个无偏估计，而 S_Y^2 是 Y 的样本方差。

容易看出 $\overline{R}^2$ 和 R^2 有一定关系；将方程（7.8.1）代入方程（7.8.2）即得：

$$\overline{R}^2 = 1 - (1-R^2)\frac{n-1}{n-k} \tag{7.8.4}$$

从方程（7.8.4）立即看出：（1）对于 $k>1$，$\overline{R}^2 < R^2$。这意味着，随着 X 变量的个数增加，调整 R^2 比未调整的 R^2 增加得慢些。（2）虽然 R^2 必定是非负的，但 $\overline{R}^2$ 可以是负的。[①] 在应用中，如果遇到 $\overline{R}^2$ 出现负的情形，就把它的值取为零。

实践中应选用哪一个 R^2？如瑟尔所指出的那样：

> ……R^2 对回归拟合的描述，特别是当解释变量的个数相对于观测次数来说不算很少的时候，明显地偏向于乐观，因此，用 $\overline{R}^2$ 而不用 R^2 是一种好的实践。[②]

但瑟尔的观点并没有被人们一致接受，因为他未曾为 $\overline{R}^2$ 的“优越性”提出任何一般理论性的论点。例如，戈德伯格称辩，**修正 R^2**（modified R^2）也是同样的好[③]：

---

① 然而，要注意，若 $R^2=1$，则 $\overline{R}^2=R^2=1$。当 $R^2=0$ 时，$\overline{R}=(1-k)/(n-k)$。此时，如果 $k>1$，$\overline{R}^2$ 就是负的。

② Henri Theil, *Introduction to Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1978, p. 135.

③ Arthur S. Goldberger, *A Course in Econometrics*, Harvard University Press, Cambridge, Mass., 1991, p. 178. 关于 R^2 更多的批判性意见，见 S. Cameron, “Why is the R Squared Adjusted Reported?” *Journal of Quantitative Economics*, vol. 9, no. 1, January 1993, pp. 183-186。作者称辩：“R^2 不是一个检验统计量，而且没有任何清晰直观的理由把它用做描述统计量。最后，我们应该明白，它并非预防数据开采的有效手段。”

$$\text{修正 } R^2 = (1-k/n)R^2 \tag{7.8.5}$$

他建议把 R^2、n 和 k 一起报告出来，让读者决定怎样为说明 n 和 k 的作用而调整 R^2。

且不管这个建议怎样，大多数统计软件包都是把方程（7.8.4）所给的调整 R^2 连同惯用的 R^2 一起报告。读者完全可以把 $\bar{R}^2$ 当作另一个摘要统计量来看待。

顺便提一句，对于儿童死亡率回归（7.6.2），读者可以验证，R^2 为 0.698 1，记住此例中$(n-1)=63$，$(n-k)=60$。恰如所料，$\bar{R}^2$ 0.698 1 小于 $R^2$0.707 7。

除了用 R^2 和调整 R^2 作为拟合优度的度量方法之外，还有其他准则（或判据）也常被用来判断一个回归模型的适用性。其中的两个是**赤池信息准则**（Akaike's Information criterion）和**雨宫预测准则**（Amerniya's Prediction criterion），用以挑选相互媲美的模型。当在第 13 章考虑模型选择问题时，我们将详细讨论这些准则。

□ 比较两个 R^2 值

根据判定系数，不管是用调整的还是未经调整的判定系数来评价两个模型，一定要注意样本容量 n 和因变量都必须相同，而解释变量则可取任何形式。因此，对模型：

$$\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{7.8.6}$$

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i \tag{7.8.7}$$

计算的两个 R^2 是不可比较的。理由如下：按定义，R^2 度量着因变量的变异被（诸）解释变量解释的部分（占 Y 总变异的比例）。因此，在方程（7.8.6）中，R^2 度量着由 X_2 和 X_3 解释的 $\ln Y$ 的变异部分，而在方程（7.8.7）中，R^2 则度量着被解释的 Y 的变异部分，两者不是同一回事。如第 6 章中所指出的，$\ln Y$ 的变化给出 Y 的相对变化或比例变化，而 Y 的变化则指 Y 的绝对变化。所以 $\text{var } \hat{Y}_i/\text{var } Y_i$ 不等于 $\text{var}(\widehat{\ln Y_i})/\text{var}(\ln Y_i)$。也就是说，两个判定系数是不同的。[①]

在回归子的形式不同的两个模型中，如何比较其 R^2 呢？为回答这个问题，让我们首先考虑一个数字例子。

例 7.2　1970—1980 年美国的咖啡消费

考虑表 7—1 中的数据。表中是有关美国 1970—1980 年日均咖啡消费量（Y）和真实零售价格（X）的数据。对这些数据应用 OLS，我们得到如下回归结果：

---

① 由 R^2 的定义可知，对于线性模型：

$$1-R^2 = \frac{\text{RSS}}{\text{TSS}} = \frac{\sum a_i^2}{\sum (Y_i - \bar{Y})^2}$$

对于对数模型：

$$1-R^2 = \frac{\sum a_i^2}{\sum (\ln Y_i - \overline{\ln Y})^2}$$

由于这两个表达式右端的分母各不相同，我们不能直接比较其 R^2 项。

如例 7.2 所示，对于线性形式，RSS=0.149 1（咖啡消费的残差平方和）。而对于对数线性形式，RSS=0.022 6（对数咖啡消费的残差平方和）。这两种残差属于不同的数量级，因而不可直接比较。

$$\hat{Y}_t = 2.691\,1 - 0.479\,5X_t \qquad (7.8.8)$$

$$\text{se} = (0.121\,6)\ (0.114\,0) \qquad \text{RSS} = 0.149\,1 \quad r^2 = 0.662\,8$$

这个结果的经济含义是：随着咖啡价格上涨1单位，日均咖啡消费量平均下降约半杯。约等于0.66的r^2值意味着，咖啡价格变化大约能解释咖啡消费量变化的66%。读者很容易验证，这个方程的斜率系数是统计显著的。

表7—1　　1970—1980年美国咖啡消费（Y）与平均真实零售价格（X）*的关系

年份	Y（人均日消费杯数）	X（美元/磅）
1970	2.57	0.77
1971	2.50	0.74
1972	2.35	0.72
1973	2.30	0.73
1974	2.25	0.76
1975	2.20	0.75
1976	2.11	1.08
1977	1.94	1.81
1978	1.97	1.39
1979	2.06	1.20
1980	2.02	1.17

注：*名义价格除以食品与饮料的消费者价格指数（CPI），1967=100。

资料来源：Y取自*Summary of National Coffee Drinking Study*，Data Group，Elkins Park，Penn.，1981；名义X（即以当前价格表示的X）取自*Nielsen Food Index*，A. C. Nielsen，New York，1981。

感谢斯科特·E·桑德伯格（Scott E. Sandberg）为收集数据所做的工作。

利用同样的数据，也可以估计出如下双对数或常弹性模型：

$$\widehat{\ln Y_t} = 0.777\,4 - 0.253\,0 \ln X_t$$

$$\text{se} = (0.015\,2)(0.049\,4) \quad \text{RSS} = 0.022\,6 \quad r^2 = 0.744\,8 \qquad (7.8.9)$$

由于这是一个双对数模型，所以斜率系数直接给出了价格弹性系数的一个估计值。在目前的例子中，它告诉我们，若咖啡的价格上涨1%，则日咖啡消费量平均下降约0.25个百分点。记住，在线性模型（7.8.8）中，斜率系数只给出了咖啡消费量相对价格的变化率。（你如何在线性模型中估计价格弹性呢？）约等于0.74的r^2值意味着，咖啡价格的对数变化大约能解释咖啡消费量对数变化的74%。读者很容易验证，这个方程的斜率系数是统计显著的。

既然线性模型的r^2值0.662 8比对数线性模型的r^2值0.744 8小，那你可能禁不住要选择r^2值高的后一个模型。但出于前面提到的原因，我们不能这么做。你若非要对这两个r^2值进行比较，那你可以采取如下步骤：

1. 从方程（7.8.9）中计算出每个观测的$\widehat{\ln Y_t}$；即从此模型中得到每个观测值的对数估计值。取其反对数，然后按照方程（3.5.14）所指明的方法计算这些反对数与实际Y_t值之间的r^2。这个r^2值和得自方程（7.8.8）的r^2值便是可比的。

2. 或者，假设所有的Y值都为正，则取其对数便得到$\ln Y$。从线性模型（7.8.8）中得到Y_t的估计值（即$\hat{Y}_t$），然后取这些估计值的对数（即得到$\ln \hat{Y}_t$），并按照方程（3.5.14）所指明的方法计算$\ln Y_t$与$\ln \hat{Y}_t$之间的r^2值。这个r^2值与从方程（7.8.9）得到的r^2值便是可比的。

对我们的咖啡消费一例而言，我们在表 7—2 中给出了计算这些可比较的 r^2 值所必需的原始数据。为了将线性模型（7.8.8）中的 r^2 值与双对数模型（7.8.9）的 r^2 值相比较，我们首先得到 $\hat{Y}_i$ 的对数值 $\ln \hat{Y}_i$ [在表 7—2 中第（6）列给出]，然后得到 Y 实际值的对数 [在表 7—2 中第（5）列给出]，再利用方程（3.5.14）计算这两组数值之间的 r^2。计算出来的 r^2 值为 0.677 9，现在便可与对数线性模型的 r^2 值 0.744 8 相比较了。这两个 r^2 值之间的差别约为 0.07。

表 7—2　　用于比较两个 R^2 值的原始数据

年份	Y_i (1)	$\hat{Y}_i$ (2)	$\widehat{\ln Y_i}$ (3)	$\widehat{\ln Y_i}$ 的反对数 (4)	$\ln Y_i$ (5)	$\ln \hat{Y}_i$ (6)
1970	2.57	2.321 887	0.843 555	2.324 616	0.943 906	0.842 380
1971	2.50	2.336 272	0.853 611	2.348 111	0.916 291	0.848 557
1972	2.35	2.345 863	0.860 544	2.364 447	0.854 415	0.852 653
1973	2.30	2.341 068	0.857 054	2.356 209	0.832 909	0.850 607
1974	2.25	2.326 682	0.846 863	2.332 318	0.810 930	0.844 443
1975	2.20	2.331 477	0.850 214	2.340 149	0.788 457	0.846 502
1976	2.11	2.173 233	0.757 943	2.133 882	0.746 688	0.776 216
1977	1.94	1.823 176	0.627 279	1.872 508	0.662 688	0.600 580
1978	1.97	2.024 579	0.694 089	2.001 884	0.678 034	0.705 362
1979	2.06	2.115 689	0.731 282	2.077 742	0.722 706	0.749 381
1980	2.02	2.130 075	0.737 688	2.091 096	0.703 098	0.756 157

注：第（1）列：来自表 7—1 的 Y 实际值。
第（2）列：来自线性模型（7.8.8）的 Y 估计值。
第（3）列：来自双对数模型（7.8.9）的 $\ln Y$ 估计值。
第（4）列：第（3）列中数值的反对数。
第（5）列：第（1）列中 Y 的对数值。
第（6）列：第（2）列中 $\hat{Y}$ 的对数值。

另一方面，如果我们想将对数线性模型的 r^2 值与线性模型中的 r^2 值相比较，那么，我们首先从方程（7.8.9）计算出每个观测的 $\widehat{\ln Y_i}$ [表 7—2 中第（3）列给出]，再求出其反对数值 [表 7—2 中第（4）列给出]，然后按照公式（3.5.14）计算这些反对数值与 Y 实际值之间的 r^2。计算出来的 r^2 值为 0.718 7，比从线性模型（7.8.8）中所得到的 r^2 值 0.662 8 略高一些。

如此看来，无论用哪一种方法，对数线性模型拟合得总是略好一些。

□ 在回归元之间的分配 R^2

让我们回到儿童死亡率一例。我们在方程（7.6.2）中看到，PGNP 和 FLR 两个回归元解释了儿童死亡率变异中的 0.707 7 或 70.77%。但现在我们再考虑去掉 FLR 变量的回归（7.7.2），r^2 值下降到 0.166 2。这是否意味着 r^2 值的差值 0.541 5（=0.707 7−0.166 2）都是去掉的变量 FLR 所能解释的部分呢？另一方面，如果考虑去掉 PGNP 变量的回归（7.7.3），r^2 值下降到 0.669 6。这是否又意味着 r^2 值的差值 0.038 1（=0.707 7−0.669 6）是去掉的变量 PGNP 所能解释的部分呢？

于是问题是：我们是否能够如此将多元回归的 R^2 值 0.707 7 在 PGNP 和 FLR 这两个回归元之间分配？不幸的是，我们不能这么做，因为正如我们刚才说明的那样，这种分配取决于我们引入回归元的顺序。这里的部分问题在于这两个回归元的相关关系，其相关系数为 0.268 5（用表 6—4 中的数据来验证）。在大多数包含几个回归元的应用研究中，回归元之间的相关都是一个常见问题。当然，若回归元之间存在完全共线性，那么问题就很严重了。

最好的实践忠告就是，试图将 R^2 值在其包含的回归元中进行分配没有什么意义。

□ 关于 $\bar{R}^2$ 最大化的“游戏”

在结束本节讨论时，提出一个警告是合适的：有时一些研究者玩 $\bar{R}^2$ 最大化的游戏。也就是说，要选择有最高 $\bar{R}^2$ 值的模型。但这样做可能是危险的。因为在回归分析中，我们的目的并不是为了得到一个高的 $\bar{R}^2$，而是要得到真实总体回归系数的可靠估计并作出有关的统计推断。在经验分析中，得到一个很高 $\bar{R}^2$ 的同时发现某些回归系数统计上不显著或与先验预期的符号相反的情形屡见不鲜。故研究者应更关心解释变量对因变量的逻辑或理论关系及统计显著性。如果在这一研究过程中，我们得到了一个高的 $\bar{R}^2$ 自然很好；另一方面，如果 $\bar{R}^2$ 偏低，也未必模型是坏的。①

实际上，戈德伯格曾对 R^2 的作用有过告诫：

> 在我们看来，在回归分析中，R^2 作为从一堆数据里拟合一个样本最小二乘回归的优度，只起到一种很平常的作用。在经典回归（经典线性回归）模型中没有哪一点要求 R^2 必须很高。因而，一个高的 R^2 不是肯定模型的证据，而一个低的 R^2 也不是否定模型的证据。
>
> 事实上，关于 R^2 的最重要的事情是，它在经典回归模型中是不重要的。经典回归模型是用来研究一个总体中的参数的，它不关心在一个样本中拟合的好坏，……如果人们坚持要对预测成功（而不是失败）有一个度量，那么有了 σ^2 也许足够了。毕竟，如果用总体回归函数作为预测元，参数 σ^2 就是期望预测误差的平方。换句话说，预测的标准误差的平方……也许对适当取定的 X（回归元）值来说，是富有信息的。②

---

① 一些作者意在抵消对 R^2 用于衡量拟合优度以及比较两个或多个 R^2 值的过分重视。参看 Christopher H. Achen, *Interpreting and Using Regression*, Sage Publications, Beverly Hills, Calif., 1982, pp. 58-67, 以及 C. Granger and P. Newbold, "R^2 and the Transformation of Regression Variables," *Journal of Econometrics*, vol. 4, 1976, pp. 205-210。值得一提的是，根据最高的 R^2 挑选模型的做法，属于一种数据开采，导致了所谓的**预检偏误**（pretest bias）。这种偏误可能破坏了经典线性回归模型 OLS 估计量的某些性质。关于这一论题，读者可参考 George G. Judge, Carter R. Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, John Wiley, New York, 1982, Chapter 21。

② Arther S. Goldberger, op. cit., pp. 177-178.

7.9 柯布-道格拉斯生产函数：函数形式再议

在 6.4 节中，我们说明怎样通过适当的变量代换把非线性关系式转换为线性，以便在经典线性回归模型的框架内考虑问题。当时相对于双变量情形而讨论的各种变换，是能够容易地推广应用到多元回归模型上来的。本节拟通过双变量对数线性模型的多变量推广来讲解变量代换；其他的函数形式则散见于习题以及本书其余部分所讨论的说明性例子中。我们即将讨论的特殊例子，是生产理论中著名的**柯布-道格拉斯生产函数**（Cobb-Douglas production function）。

柯布-道格拉斯生产函数的随机形式可表达为：

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i} \tag{7.9.1}$$

其中Y=产出；

X_2 =劳动投入；

X_3 =资本投入；

u=随机干扰项；

e=自然对数的底。

显然，方程（7.9.1）给出的产出与两种投入之间的关系式是非线性的。然而，通过模型的对数变换，可得到：

$$\begin{aligned}\ln Y_i &= \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \\ &= \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i\end{aligned} \tag{7.9.2}$$

其中 $\beta_0 = \ln \beta_1$。

这种形式的模型对参数 β_0，β_2 和 β_3 是线性的，因而是一个线性回归模型。注意，虽然如此，它对变量 Y 和 X 为非线性的而对这些变量的对数为线性的。简言之，方程（7.9.2）是一个对数—对数、双对数或对数线性模型。它是双变量对数线性模型（6.5.3）在多元回归中的对应形式。

柯布-道格拉斯生产函数的性质是众所周知的：

1. β_2 是产出对劳动投入的（偏）弹性，即它度量着在保持资本投入不变的情况下劳动投入变化 1%时的产出百分比变化（见习题 7.9）。

2. 相似地，β_3 是在保持劳动投入不变情况下产出对资本投入的（偏）弹性。

3. 总和（$\beta_2 + \beta_3$）给出关于规模报酬（returns to scale）的信息，即产出对投入的比例变化的反应。如果此总和为 1，则规模报酬不变（constant returns to scale），即 2 倍的投入将带来 2 倍的产出，3 倍的投入将带来 3 倍的产出，等等。如果总和小于 1，则规模报酬递减（decreasing returns to scale）——2 倍的投入将带来少于 2 倍的产出。最后，如果总和大于 1，则规模报酬递增（increasing returns to scale）——2 倍的投入将带来多于 2 倍的产出。

在继续讨论之前，应看到无论你的对数线性回归模型涉及多少个 X 变量，每个 X 变量的系数都代表因变量对该变量的（偏）弹性。例如，如果你有一个 k 变量对数线性模型：

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \cdots + \beta_k \ln X_{ki} + u_i \qquad (7.9.3)$$

则从 β_2 到 β_k 的每个（偏）回归系数，都是 Y 对从 X_2 到 X_k 变量的（偏）弹性。[①]

例 7.3 美国制造业部门的价值加成、劳动小时数和资本投入

为了说明柯布-道格拉斯生产函数，我们收集了表 7—3 中的数据；这些数据反映了 2005 年美国 50 个州和华盛顿特区的制造业部门数据。

表 7—3　　2005 年美国制造业部门的价值加成、劳动小时数和资本投入

地区	产出价值加成（千美元）Y	劳动投入（千小时）X_2	资本投入（千美元）X_3
亚拉巴马	38 372 840	424 471	2 689 076
阿拉斯加	1 805 427	19 895	57 997
亚利桑那	23 736 129	206 893	2 308 272
阿肯色	26 981 983	304 055	1 376 235
加利福尼亚	217 546 032	1 809 756	13 554 116
科罗拉多	19 462 751	180 366	1 790 751
康涅狄格	28 972 772	224 267	1 210 229
特拉华	14 313 157	54 455	421 064
哥伦比亚特区	159 921	2 029	7 188
佛罗里达	47 289 846	471 211	2 761 281
佐治亚	63 015 125	659 379	3 540 475
夏威夷	1 809 052	17 528	146 371
爱达荷	10 511 786	75 414	848 220
伊利诺伊	105 324 866	963 156	5 870 409
印第安纳	90 120 459	835 083	5 832 503
艾奥瓦	39 079 550	336 159	1 795 976
堪萨斯	22 826 760	246 144	1 595 118
肯塔基	38 686 340	384 484	2 503 693
路易斯安那	69 910 555	216 149	4 726 625
缅因	7 856 947	82 021	415 131
马里兰	21 352 966	174 855	1 729 116
马萨诸塞	46 044 292	355 701	2 706 065
密歇根	92 335 528	943 298	5 294 356
明尼苏达	48 304 274	456 553	2 833 525

① 为了看到这一点，将方程（7.9.3）对每个 X 变量的对数求偏微分。于是就有 $\partial \ln Y / \partial \ln X_2 = (\partial Y / \partial X_2)(X_2/Y) = \beta_2$，根据定义，这就是 Y 对 X_2 的弹性；同样，$\partial \ln Y / \partial \ln X_3 = (\partial Y / \partial X_3)(X_3/Y) = \beta_3$，这就是 Y 对 X_3 的弹性，等等。

续前表

地区	产出价值加成（千美元） Y	劳动投入（千小时） X_2	资本投入（千美元） X_3
密西西比	17 207 903	267 806	1 212 281
密苏里	47 340 157	439 427	2 404 122
蒙大拿	2 644 567	24 167	334 008
内布拉斯加	14 650 080	163 637	627 806
内华达	7 290 360	59 737	522 335
新罕布什尔	9 188 322	96 106	507 488
新泽西	51 298 516	407 076	3 295 056
新墨西哥	20 401 410	43 079	404 749
纽约	87 756 129	727 177	4 260 353
北卡罗来纳	101 268 432	820 013	4 086 558
北达科他	3 556 025	34 723	184 700
俄亥俄	124 986 166	1 174 540	6 301 421
俄克拉何马	20 451 196	201 284	1 327 353
俄勒冈	34 808 109	257 820	1 456 683
宾夕法尼亚	104 858 322	944 998	5 896 392
罗得岛	6 541 356	68 987	297 618
南卡罗来纳	37 668 126	400 317	2 500 071
南达科他	4 988 905	56 524	311 251
田纳西	62 828 100	582 241	4 126 465
得克萨斯	172 960 157	1 120 382	11 588 283
犹他	15 702 637	150 030	762 671
佛蒙特	5 418 786	48 134	276 293
弗吉尼亚	49 166 991	425 346	2 731 669
华盛顿	46 164 427	313 279	1 945 860
西弗吉尼亚	9 185 967	89 639	685 587
威斯康星	66 964 978	694 628	3 902 823
怀俄明	2 979 475	15 221	361 536

资料来源：*2005 Annual Survey of Manufactures*，Sector 31：Supplemental Statistics for U. S..

假定模型（7.9.2）满足经典线性回归模型的假定。[①] 用 OLS 法得到如下回归（计算机打印结果见附录 7A 第 7A.5 节）：

$$\widehat{\ln Y_i} = 3.887\,6 + 0.468\,3 \ln X_{2i} + 0.521\,3 \ln X_{3i} \qquad (7.9.4)$$

$$(0.396\,2)\quad (0.098\,9)\qquad (0.096\,9)$$

$$t = (9.811\,5)\quad (4.734\,2)\qquad (5.380\,3)$$

$$R^2 = 0.964\,2 \qquad \mathrm{df} = 48$$

$$\overline{R}^2 = 0.962\,7$$

① 注意在柯布-道格拉斯生产函数（7.9.1）中，我们以特殊方式引进随机误差项，以使对数变换后得到通常的线性形式。参看 6.9 节。

从方程（7.9.4）我们看到2005年美国制造业产出的劳动和资本弹性分别是0.468 3和0.521 3。换言之，在研究时期，保持资本投入不变，劳动投入增加1%，平均导致产出增加约0.47%。类似地，保持劳动投入不变，资本投入增加1%平均导致产出增加约0.52%。把两个产出弹性相加得到0.99，即为规模报酬参数的取值。看得出来，在此研究期间，美国50个州和哥伦比亚特区的制造业具有规模报酬不变的特征。

从纯粹的统计观点看，所估计的回归线对数据的拟合相当良好。R^2 取值为0.964 2，表示产出（的对数）的变异96%都可由劳动和资本（的对数）来解释。在第8章中，我们将看到怎样用估计的标准误去检验有关美国制造业的柯布-道格拉斯生产函数中参数“真”值的假设。

7.10 多项式回归模型

作为本章的最后一部分，我们现在考虑一类多元回归模型，即**多项式回归模型**（polynomial regression models）。这类模型在有关成本和生产函数的计量经济研究中有广泛的用途。在介绍这些模型的同时，我们进一步扩大了经典线性回归模型的适用范围。

为便于分析，考虑图7—1。该图描述了生产一种商品的短期边际成本（Y）与它的产出水平（X）之间的关系。图中随手画出的、教科书般的U型MC曲线表明了MC和产出之间的关系是非线性的。如果我们要把这种得自给定散点图的关系加以量化，怎么办？换句话说，什么类型的计量经济模型能给出边际成本先降后升的性质？

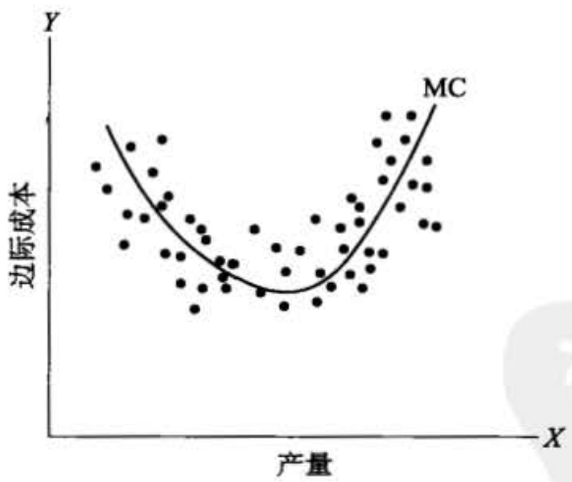


图7—1 U型边际成本曲线

从图形上看，图7—1描绘的MC曲线代表一条抛物线。在数学上，抛物线的表达式是

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \tag{7.10.1}$$

它被称为二次函数或更一般地称为变量 X 的二次多项式——X 的最高次方代表多项式的次数。（如果在上述函数中加进 X^3 项，就称为三次多项式，如此类推。）

方程（7.10.1）的随机形式可写为：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i \tag{7.10.2}$$

此即二阶多项式回归。

k 阶多项式回归（kth degree polynomial regression）可写成

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + u_i \tag{7.10.3}$$

注意，在这类多项式回归中，方程右边只有一个解释变量，但以不同乘方出现，从而使方程成为多元回归模型。顺便提一下，如果 X 被假定为固定的或非随机的，那么带有乘方的各 X_i 项也成为固定的或非随机的。

这种模型会带来什么特殊的估计问题吗？由于二次多项式（7.10.2）或 k 次多项式（7.10.3）对参数 β 而言都是线性的，故可用普通最小二乘法或极大似然法来估计。但会有什么共线性问题吗？既然各个 X 项都是 X 的幂函数，它们会不会高度相关？是的，但应记住，像 X^2、X^3、X^4 等项都是 X 的非线性函数，所以严格地说，并不违反无多重共线性的假定。总之，多项式回归模型没有提出任何新的估计问题，可用本章讲过的方法去估计它们。

例 7.4　估计总成本函数

作为多项式回归的一个例子。考虑表 7—4 给出的短期内某商品产出及其总成本数据。什么类型的回归模型能拟合这些数据呢？为此，我们先作出散点图，如图 7—2 所示。

表 7—4　总成本（Y）与产出（X）

产出	总成本（美元）
1	193
2	226
3	240
4	244
5	257
6	260
7	274
8	297
9	350
10	420

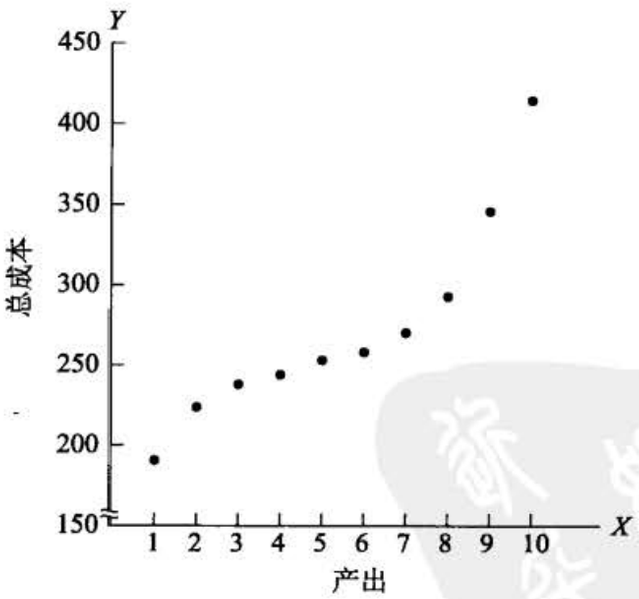


图 7—2　总成本曲线

由此图显而易见，总成本与产出之间的关系像一条拉长的 S 曲线；注意这条总成本曲线先是缓慢上升，然后急剧上升，如同著名的边际报酬递减定律所描述的。总成本曲线的 S 形状可以由

下面的立方或三次多项式来刻画：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \tag{7.10.4}$$

其中 Y=总成本，X=产出。

给定表 7—4 的数据，可用 OLS 估计方程（7.10.4）中的参数。在估计之前，先看看经济理论对短期立方成本函数（7.10.4）是如何描述的。基本价格理论表明，在短期里，典型地说，生产的边际成本（MC）和平均成本（AC）都是 U 型的——最初，随着产出增加，MC 和 AC 都下降，但到了一定产出水平之后，两者均转而升高，再次显示边际报酬递减定律的后果。这可以从图 7—3（还有图 7—1）看出。而由于 MC 和 AC 曲线可以由总成本曲线推导出来，故这些曲线的 U 型性质给总成本曲线（7.10.4）的参数添加了一些约束。事实上，可以证明，如果短期边际成本曲线和平均成本曲线遵循 U 型，方程（7.10.4）的参数必须满足如下约束条件①：

$$\begin{aligned} &1.\ \beta_0,\ \beta_1 \text{ 和 } \beta_3 > 0; \\ &2.\ \beta_2 < 0; \\ &3.\ \beta_2^2 < 3\beta_1\beta_3。 \end{aligned} \tag{7.10.5}$$

所有这些理论探讨也许看来有些令人厌烦，但这种知识在我们分析经验结果时却非常有用，如果经验结果与事先的理论预期不符，那么，假定我们的模型没有设定误差（即没有选用错误的模型），我们就必须修改我们的理论或寻求新的理论，然后重新开始我们的经验研究。但如同引言中所提到的那样，这是任何经验研究的共性。

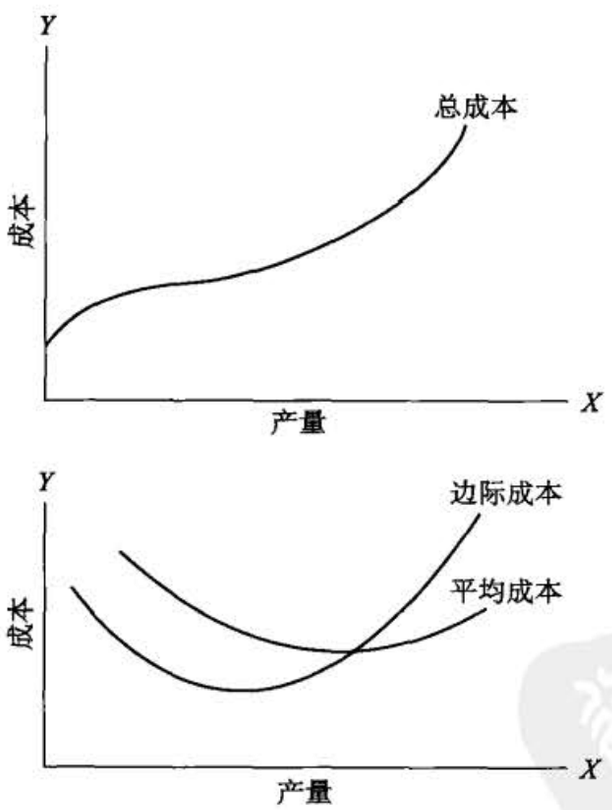


图 7—3　短期成本函数

① 参看 Alpha C. Chiang, *Fundamental Methods of Mathematical Economics*, 3d ed., McGraw-Hill, New York, 1984, pp. 250-252。

实证结果。用三次多项式拟合表 7—4 的数据，得到如下结果：

$$\hat{Y}_i = 141.766\,7 + 63.477\,6\,X_i - 12.961\,5\,X_i^2 + 0.939\,6X_i^3$$

$$(6.375\,3)\quad (4.778\,6)\qquad (0.985\,7)\qquad (0.059\,1)\quad R^2 = 0.998\,3 \qquad (7.10.6)$$

（注：括号中的数字是估计的标准误。）虽然在下章里我们将分析这些结果的统计显著性。但读者能证实它们和方程（7.10.5）所列举的理论预期是一致的。至于怎样解释回归（7.10.6），我们把它作为习题留给读者。

例 7.5　2007 年 190 个国家的 GDP 增长率与相对人均 GDP（以 2000 年十亿美元计）

作为多项式回归模型的另外一个经济例子，考虑如下回归结果：

$$\widehat{\text{GDPG}}_i = 5.534\,7 - 5.578\,8\,\text{RGDP} + 2.837\,8\,\text{RGDP}^2$$

$$\text{se} = (0.243\,5)(1.599\,5)\qquad (1.439\,1) \qquad (7.10.7)$$

$$R^2 = 0.109\,2 \qquad \overline{R}^2 = 0.099\,6$$

其中 GDPG 表示 2007 年 GDP 的百分比增长率，RGDP 表示 2007 年的相对人均 GDP（即占美国 2007 年人均 GDP 的百分比）。调整 R^2 告诉我们，在考虑了回归元的个数之后，该模型也只解释了 GDPG 变动的 9.96%。即便从未经调整的 R^2 来看，0.109 2 也很小。这个值听起来让人气馁，但我们在下一章将会看到，在一个含有大量观测的横截面数据中，经常会遇到这么低的 R^2 值。而且，我们在下一章将会证明，即便一个明显很低的 R^2 值也可能是统计显著的（即异于零）。

---

*7.11　偏相关系数

□ 简单与偏相关系数的释义

在第 3 章中，我们介绍了作为度量两变量之间线性关联程度的相关系数 r。对于三变量回归模型，我们可以算出三个相关系数：r_{12}（Y 与 X_2 之间的相关），r_{13}（Y 与 X_3 之间的相关）和 r_{23}（X_2 与 X_3 之间的相关）；注意，出于记号上的便利，我们令下标 1 代表 Y。这些相关系数可称**毛**（gross）或**简单相关系数**（simple correlation coefficients），或称**零阶相关系数**（correlation coefficients of zero order）。这些系数都能按方程（3.5.13）所给相关系数的定义计算出来。

但考虑下述问题：比方说，如果有第三个变量 X_3 同 Y 和 X_2 都可能相关，r_{12} 果真度量了 Y 与 X_2 之间的“真实”（线性）关联度吗？这个问题又可类比于下述问题：假设真实的回归模型是（7.1.1），但我们从模型中略去了变量 X_3，仅做 Y 对 X_2 的

---

* 选读内容。

回归，并得到斜率系数 b_{12}。这个系数是否等于我们一开始就要估计的模型（7.1.1）的真实系数 β_2 呢？从我们在 7.7 节中的讨论看，答案是明显的。一般来说，r_{12} 在 X_3 出现的情形下不大可能反映 Y 和 X_2 之间的真实关联度。事实上，我们即将看到，它容易给出 Y 与 X_2 之间相关性质的一个错误印象。因此，我们还需要有一个不依赖于 X_3 对 X_2 和 Y 影响（如果这个影响存在的话）的一种相关系数。我们最好把这种系数称为**偏相关系数**（partial correlation coefficient），并且可以计算出来。它类似于偏回归系数。让我们定义：

$r_{12.3}$ = X_3 保持不变下的 Y 和 X_2 的偏相关系数

$r_{13.2}$ = X_2 保持不变下的 Y 和 X_3 的偏相关系数

$r_{23.1}$ = Y 保持不变下的 X_2 和 X_3 的偏相关系数

从简单或零阶相关系数很容易就能计算出这些偏回归系数（证明见习题）[①]：

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \tag{7.11.1}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \tag{7.11.2}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} \tag{7.11.3}$$

由方程（7.11.1）到（7.11.3）给出的偏相关系数被称为**一阶相关系数**（first-order correlation coefficients），这里所说的阶是指次段下标的个数。例如，$r_{12.34}$ 是二阶相关系数，$r_{12.345}$ 是三阶相关系数，等等。前面曾指出，r_{12}、r_{13} 等称为简单或零阶相关系数。$r_{12.34}$（比方说）的含义是在保持 X_3 和 X_4 不变的情况下，Y 和 X_2 之间的相关系数。

□ 简单与偏相关系数的解说

在双变量情形中，简单相关系数 r 有明显的含义：它度量着因变量 Y 与单个解释变量 X 之间的（线性）关联（而非因果关系）程度。可一旦我们超出双变量情形，我们对简单相关系数的解释就需要细心。例如，从方程（7.11.1）我们观察到：

1. 即使 $r_{12}=0$，$r_{12.3}$ 并不为零，除非 r_{13} 或 r_{23} 或两者都为零。

2. 如果 $r_{12}=0$，而 r_{13} 和 r_{23} 均不为零且有相同符号，则 $r_{12.3}$ 为负的，而如果 r_{13} 和 r_{23} 符号相反，则 $r_{12.3}$ 为正的。举一个例子将能说明这一点。令 Y=农作物收成，X_2 =雨量，X_3 =气温。假定 $r_{12}=0$，即作物收成和雨量没有关系。再假定 r_{13} 是正的，而 r_{23} 是负的。这时方程（7.11.1）表明 $r_{12.3}$ 将是正的；也就是说，在保持气温不变的情况下，收成和雨量有正的相关关系。然而这种看起来似乎矛盾的结论并不令人惊奇。因为气温 X_3 同时影响收成 Y 和雨量 X_2。要找出作物收成与雨量之间的净关系，需要去除“讨厌”变量气温的影响。本例表明了人们怎样被简单相关系数

① 绝大多数多元回归分析软件都例行计算这个简单相关系数；因此很容易计算偏相关系数。

误导。

3. $r_{12.3}$ 项和 r_{12} 项（及类似项的比较）不一定同号。

4. 在双变量情形中，r_2 介于 0 与 1 之间。偏相关系数的平方也有同样性质。利用这一事实，读者应能证实从方程（7.11.1）可推出下列表达式：

$$0 \leqslant r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leqslant 1 \tag{7.11.4}$$

这给出了三个零阶相关系数的相互关系。同样的表达式可从方程（7.11.2）和（7.11.3）推导出来。

5. 假设 $r_{13} = r_{23} = 0$，这是否意味着 r_{12} 也等于零？答案可从方程（7.11.4）明显看出。Y 与 X_3 以及 X_2 与 X_3 不相关，并不意味着 Y 与 X_2 不相关。

顺便提一下，表达式 $r_{12.3}^2$ 可称为**偏判定系数**（coefficient of partial determination），并可解释为未被 X_3 解释的 Y 的变异部分由于 X_2 被引进到模型中来而得到解释的比例（参看习题 7.5）。它在概念上类似于 R^2。

在继续讨论之前，请注意 R^2、简单相关系数以及偏相关系数之间存在如下关系：

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} \tag{7.11.5}$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2)r_{13.2}^2 \tag{7.11.6}$$

$$R^2 = r_{13}^2 + (1 - r_{13}^2)r_{12.3}^2 \tag{7.11.7}$$

在结束本节讨论之际，请考虑：前面我们说过，如果在模型中多引进一个解释变量，R^2 必不会减小。这点可由方程（7.11.6）明显看出。该方程指出，由 X_2 和 X_3 联合解释的 Y 的变异部分（比例）是两个部分之和：由 X_2 单独解释的部分（r_{12}^2），以及未被 X_2 解释的部分（$=1 - r_{12}^2$）乘以在保持 X_2 的影响不变下由 X_3 解释的比例。现在，只要 $r_{13.2}^2 > 0$，就有 $R^2 > r_{12}^2$。$r_{13.2}^2$ 最小不过是零，这时 $R^2 = r_{12}^2$。

要点与结论

1. 本章介绍最简单的多元（多变量）线性回归模型，即三变量回归模型。我们默认“线性”一词指对参数为线性，对变量不一定为线性。

2. 虽然三变量回归模型在多个方面都是双变量模型的推广，却涉及一些新的概念，诸如偏回归系数、偏相关系数、多元相关系数、调整与未调整（对自由度）R^2、多重共线性和设定偏误。

3. 本章还考虑多元回归模型的函数形式，如柯布-道格拉斯生产函数和多项式回归模型。

4. 虽然 R^2 和调整 R^2 是对所选模型对给定数据集拟合好坏的总度量，但它们的重要性不可过分夸大。最为关键的是对进入模型的变量的系数，应带有什么先验性符号，从而对这个模型有一个基本的理论预期，以及下章要讲的关于这些系数的统计显著性。

5. 本章所给出的结果，很容易就能推广至涉及任意多个回归元的多元线性回归模型中，但代数运算会变得非常烦琐。使用矩阵代数就可避免这种烦琐性。对于感兴趣的读者，在作为选读用

的附录C中，我们用矩阵代数作出对 k 变量回归模型的推广。一般读者可继续阅读本书的其余部分。

习　题

问答题

7.1　考虑表 7—5 中的数据：

表 7—5

Y	X_2	X_3
1	1	2
3	2	1
8	3	−3

根据这些数据估计以下回归：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_{1i} \tag{1}$$

$$Y_i = \lambda_1 + \lambda_3 X_{3i} + u_{2i} \tag{2}$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{3}$$

注：只估计系数，不用估计标准误。

a. $\alpha_2 = \beta_2$ 吗？为什么？

b. $\lambda_3 = \beta_3$ 吗？为什么？

你能从这道题得出什么重要的结论？

7.2　利用以下数据，估计偏回归系数及其标准误，以及调整与未调整的 R^2 值：

$\overline{Y} = 367.693$　　$\overline{X}_2 = 402.760$　　$\overline{X}_3 = 8.0$

$\sum (Y_i - \overline{Y})^2 = 66\,042.269$

$\sum (X_{2i} - \overline{X}_2)^2 = 84\,855.096$

$\sum (X_{3i} - \overline{X}_3)^2 = 280.000$

$\sum (Y_i - \overline{Y})(X_{2i} - \overline{X}_2) = 74\,778.346$

$\sum (Y_i - \overline{Y})(X_{3i} - \overline{X}_3) = 4\,250.900$

$\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3) = 4\,796.000$

$n = 15$

7.3　证明方程（7.4.7）还可表达为：

$$\hat{\beta}_2 = \frac{\sum y_i (x_{2i} - b_{23} x_{3i})}{\sum (x_{2i} - b_{23} x_{3i})^2}$$

$$= \frac{y \text{ 与 } x_2 \text{ 的净(除去 } x_3\text{) 协方差}}{x_2 \text{ 的净(除去 } x_3\text{) 变异}}$$

其中 b_{23} 是 X_2 对 X_3 回归的斜率系数。（提示：回忆 $b_{23} = \sum x_{2i} x_{3i} / \sum x_{3i}^2$。）

7.4 在一个多元回归模型中，告诉你误差项 u_i 服从如下概率分布，即 $u_i \sim N(0, 4)$。你将如何构造一个蒙特卡罗实验来验证真实方差为 4。

7.5 证明 $r^2_{12.3} = (R^2 - r^2_{13})/(1 - r^2_{13})$，并解释该方程。

7.6 如果关系式 $\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 = 0$ 对 X_1，X_2 和 X_3 的所有值都成立，试求三个偏相关系数的值。

7.7 能从一组数据得到以下结果吗？

a. $r_{23} = 0.9$，　$r_{13} = -0.2$，　$r_{12} = 0.8$

b. $r_{12} = 0.6$，　$r_{23} = -0.9$，　$r_{31} = -0.5$

c. $r_{21} = 0.01$，　$r_{13} = 0.66$，　$r_{23} = -0.7$

7.8 考虑如下模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

其中 X_2 表示教育变量，X_3 表示工作年限变量。假设你漏掉了工作年限变量。预计会出现什么类型的问题或偏误？并口头加以解释。

7.9 证明方程（7.9.2）中的 β_2 和 β_3 确实给出产出的劳动和资本弹性。（不用微积分也能回答此问题；只要回忆弹性系数的定义，并记住如果变化比较小，一个变量的对数变化就是一种相对变化。）

7.10 考虑本章讨论的三变量线性回归模型。

a. 假设你将所有的 X_2 值都乘以 2。这种度量单位的改变对参数估计值及其标准误有什么影响（如果有的话）？

b. 现在假设与（a）不同，你将所有的 Y 值都乘以 2，对所估计的参数及其标准误又将有何影响（如果有的话）？

7.11 一般地说，$R^2 \neq r^2_{12} + r^2_{13}$，但仅当 $r_{23} = 0$ 时，等式成立。试评论并指出这一结论的意义。[提示：参看方程（7.11.5）。]

7.12 考虑以下模型[①]：

模型 A：$Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + u_{1t}$

模型 B：$(Y_t - X_{2t}) = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_{2t}$

a. α_1 和 β_1 的 OLS 估计是否一样？为什么？

b. α_3 和 β_3 的 OLS 估计是否一样？为什么？

c. α_2 和 β_2 有什么关系？

d. 你能比较两个模型的 R^2 吗？为什么？

7.13 假设你估计消费函数和储蓄函数[②]：

$$Y_i = \alpha_1 + \alpha_2 X_i + u_{1i}$$

$$Z_i = \beta_1 + \beta_2 X_i + u_{2i}$$

其中 Y=消费，Z=储蓄，X=收入，并且 $X=Y+Z$，即收入等于消费加储蓄。

a. α_2 和 β_2 存在什么关系（如果有的话）？给出你的计算。

b. 两个模型的残差平方和 RSS 是否一样？作出解释。

---

① 改编自 Wojciech W. Charemza and Derek F. Deadman, *Econometric Practice: General to Specific Modelling, Cointegration and Vector Autogression*, Edward Elgar, Brookfield, Vermont, 1992, p. 18。

② 改编自 Peter Kennedy, *A Guide to Econometrics*, 3d ed., The MIT Press, Cambridge, Massachusetts, 1992, p. 308, Question #9。

c. 你能比较两模型的 R^2 吗？为什么？

7.14 假设你把方程（7.9.1）中给出的柯布-道格拉斯模型表达成如下形式：

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} u_i$$

如果你做这个模型的对数变换，你将在等式右边得到 $\ln u_i$ 作为干扰项。

a. 为了能应用经典正态线性回归模型的性质，你需要对 $\ln u_i$ 做什么概率假设？你会怎样利用表 7—3 中的数据去检验这个假设。

b. 同样的假设也适用于 u_i 吗？为什么？

7.15 **过原点回归**。考虑以下过原点回归：

$$Y_i = \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

a. 你打算怎样估计这些未知数？

b. 对这个模型而言 $\sum \hat{u}_i$ 会是零吗？为什么？

c. 对这个模型会不会有 $\sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = 0$？

d. 什么时候你会使用这样的模型？

e. 你能把你的结果推广到 k 变量模型吗？

（提示：参照第 6 章对双变量情形的讨论。）

实证分析题

7.16 **玫瑰的需求**。[①] 表 7—6 给出如下变量的季度数据：

Y＝售出的玫瑰数量，打；

X_2 ＝玫瑰的平均批发价格，美元/打；

X_3 ＝石竹的平均批发价格，美元/打；

X_4 ＝每周家庭平均可支配收入，美元/周；

X_5 ＝底特律市区从 1971 年第Ⅲ季度到 1975 年第Ⅱ季度的趋势变量，取值 1，2…。

表 7—6　1971 年第Ⅲ季度至 1975 年第Ⅱ季度底特律市区对玫瑰的季度需求数据

年份与季度	Y	X_2	X_3	X_4	X_5
1971—Ⅲ	11 484	2.26	3.49	158.11	1
Ⅳ	9 348	2.54	2.85	173.36	2
1972—Ⅰ	8 429	3.07	4.06	165.26	3
—Ⅱ	10 079	2.91	3.64	172.92	4
—Ⅲ	9 240	2.73	3.21	178.46	5
—Ⅳ	8 862	2.77	3.66	198.62	6
1973—Ⅰ	6 216	3.59	3.76	186.28	7
—Ⅱ	8 253	3.23	3.49	188.98	8
—Ⅲ	8 038	2.60	3.13	180.49	9
—Ⅳ	7 476	2.89	3.20	183.33	10
1974—Ⅰ	5 911	3.77	3.65	181.87	11
—Ⅱ	7 950	3.64	3.60	185.00	12
—Ⅲ	6 134	2.82	2.94	184.00	13
—Ⅳ	5 868	2.96	3.12	188.20	14

① 感谢乔·沃尔什（Joe Walsh）从底特律市区的一个大批发商收集到这些数据并进行了加工。

续前表

年份与季度	Y	X_2	X_3	X_4	X_5
1975—Ⅰ	3 160	4.24	3.58	175.67	15
—Ⅱ	5 872	3.69	3.53	188.00	16

请你考虑如下需求函数：

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \alpha_5 X_{5t} + u_t$$

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + \beta_5 X_{5t} + u_t$$

a. 估计线性模型的参数并解释所得结果。

b. 估计对数线性模型的参数并解释计算结果。

c. β_2，β_3 和 β_4 分别给出需求的自价格弹性，交叉价格弹性和收入弹性。它们的先验符号是什么？你的结果同先验预期相符吗？

d. 你怎样对线性模型计算自价格弹性、交叉价格弹性和收入弹性？

e. 根据你的分析，你会选择哪个模型（如果可选）？为什么？

7.17 **野猫活动**（wild-cat activity）。“野猫”是指为了在一个开发区发现和生产石油或天然气，或者为了在一个已发现的油田里寻找新的产油或产气池，或者为了提高一个已知的产油或产气池的最大产量而冒险钻探的井口。表 7—7 给出以下变量的数据。①

表 7—7　　野猫活动

以千计的野猫数 (Y)	每桶价格（不变美元） (X_2)	国内产量（每天百万桶） (X_3)	GNP（十亿不变美元） (X_4)	时间 (X_5)
8.01	4.89	5.52	487.67	1948=1
9.06	4.83	5.05	490.59	1949=2
10.31	4.68	5.41	533.55	1950=3
11.76	4.42	6.16	576.57	1951=4
12.43	4.36	6.26	598.62	1952=5
13.31	4.55	6.34	621.77	1953=6
13.10	4.66	6.81	613.67	1954=7
14.94	4.54	7.15	654.80	1955=8
16.17	4.44	7.17	668.84	1956=9
14.71	4.75	6.71	681.02	1957=10
13.20	4.56	7.05	679.53	1958=11
13.19	4.29	7.04	720.53	1959=12
11.70	4.19	7.18	736.86	1960=13
10.99	4.17	7.33	755.34	1961=14
10.80	4.11	7.54	799.15	1962=15

① 感谢雷蒙德·萨文诺（Raymond Savino）收集并加工了这些数据。

续前表

以千计的野猫数 (Y)	每桶价格（不变美元） (X_2)	国内产量（每天百万桶） (X_3)	GNP（十亿不变美元） (X_4)	时间 (X_5)
10.66	4.04	7.61	830.70	1963=16
10.75	3.96	7.80	874.29	1964=17
9.47	3.85	8.30	925.86	1965=18
10.31	3.75	8.81	980.98	1966=19
8.88	3.69	8.66	1 007.72	1967=20
8.88	3.56	8.78	1 051.83	1968=21
9.70	3.56	9.18	1 078.76	1969=22
7.69	3.48	9.03	1 075.31	1970=23
6.92	3.53	9.00	1 107.48	1971=24
7.54	3.39	8.78	1 171.10	1972=25
7.47	3.68	8.38	1 234.97	1973=26
8.63	5.92	8.01	1 217.81	1974=27
9.21	6.03	7.78	1 202.36	1975=28
9.23	6.12	7.88	1 271.01	1976=29
9.96	6.05	7.88	1 332.67	1977=30
10.78	5.89	8.67	1 385.10	1978=31

资料来源：Energy Information Administration，1978 Report to Congress.

其中 Y=钻打（探）的井口（野猫）数；

X_2 =前期井地价格（不变美元，1972=100）；

X_3 =国内产量；

X_4 =GNP，（不变美元，1972=100）；

X_5 =趋势变量，1948=1，1949=2，…，1978=31。

看看下面的模型对数据拟合得如何：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 \ln X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + u_t$$

a. 你能对此模型作些合理的先验预期吗？

b. 假定模型可以接受，估计模型的参数及其标准误，并求 R^2 和 $\bar{R}^2$。

c. 按照先验预期的观点评论你的结果。

d. 为了解释野猫活动，你能给出其他的模型吗？理由是什么？

7.18 **1962—1981 年美国国防预算支出**。为了说明美国国防预算，请你考虑如下模型：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + u_t$$

其中Y_t =年度 t 的国防预算支出，十亿美元计；

X_{2t} =年度 t 的 GNP，十亿美元计；

X_{3t} =年度 t 的美国军事销售/援助，十亿美元计；

X_{4t} =太空工业销售，十亿美元计；

X_{5t} =涉及多于十万军人的军事冲突。当军队为 100 000 人或多于 100 000 人时，此变量取值 1；当军队人数小于 100 000 人时，它取值零。

为了检验此模型，现在为你提供表 7—8 中的数据：

a. 估计此模型的参数及其标准误并求 R^2，修正 R^2 和 $\bar{R}^2$。

b. 评论所得结果，同时考虑你对 Y 与各 X 变量之间关系的任何先验预期。

c. 你还想把其他变量包括在这个模型中吗？理由是什么？

表 7—8　　1962—1981 年美国国防预算支出

年份	国防预算支出 Y	GNP X_2	美国军事销售/援助 X_3	太空工业销售 X_4	100 000 人及以上的军事冲突 X_5
1962	51.1	560.3	0.6	16.0	0
1963	52.3	590.5	0.9	16.4	0
1964	53.6	632.4	1.1	16.7	0
1965	49.6	684.9	1.4	17.0	1
1966	56.8	749.9	1.6	20.2	1
1967	70.1	793.9	1.0	23.4	1
1968	80.5	865.0	0.8	25.6	1
1969	81.2	931.4	1.5	24.6	1
1970	80.3	992.7	1.0	24.8	1
1971	77.7	1 077.6	1.5	21.7	1
1972	78.3	1 185.9	2.95	21.5	1
1973	74.5	1 326.4	4.8	24.3	0
1974	77.8	1 434.2	10.3	26.8	0
1975	85.6	1 549.2	16.0	29.5	0
1976	89.4	1 718.0	14.7	30.4	0
1977	97.5	1 918.3	8.3	33.3	0
1978	105.2	2 163.9	11.0	38.0	0
1979	117.7	2 417.8	13.0	46.2	0
1980	135.9	2 633.1	15.3	57.6	0
1981	162.1	2 937.7	18.0	68.9	0

资料来源：数据由艾伯特·卢基诺（Albert Lucchino）从各种政府出版物中搜集。

7.19　**1960—1982 年美国对鸡肉的需求。**为了研究美国人均鸡肉消费量，我们提供表 7—9 中的数据：

表 7—9　　1960—1982 年美国对鸡肉的需求

年份	Y	X_2	X_3	X_4	X_5	X_6
1960	27.8	397.5	42.2	50.7	78.3	65.8
1961	29.9	413.3	38.1	52.0	79.2	66.9
1962	29.8	439.2	40.3	54.0	79.2	67.8
1963	30.8	459.7	39.5	55.3	79.2	69.6
1964	31.2	492.9	37.3	54.7	77.4	68.7
1965	33.3	528.6	38.1	63.7	80.2	73.6
1966	35.6	560.3	39.3	69.8	80.4	76.3
1967	36.4	624.6	37.8	65.9	83.9	77.2

续前表

年份	Y	X_2	X_3	X_4	X_5	X_6
1968	36.7	666.4	38.4	64.5	85.5	78.1
1969	38.4	717.8	40.1	70.0	93.7	84.7
1970	40.4	768.2	38.6	73.2	106.1	93.3
1971	40.3	843.3	39.8	67.8	104.8	89.7
1972	41.8	911.6	39.7	79.1	114.0	100.7
1973	40.4	931.1	52.1	95.4	124.1	113.5
1974	40.7	1 021.5	48.9	94.2	127.6	115.3
1975	40.1	1 165.9	58.3	123.5	142.9	136.7
1976	42.7	1 349.6	57.9	129.9	143.6	139.2
1977	44.1	1 449.4	56.5	117.6	139.2	132.0
1978	46.7	1 575.5	63.7	130.9	165.5	132.1
1979	50.6	1 759.1	61.6	129.8	203.3	154.4
1980	50.1	1 994.2	58.9	128.0	219.6	174.9
1981	51.7	2 258.1	66.4	141.0	221.6	180.8
1982	52.9	2 478.7	70.4	168.2	232.6	189.4

注：实际价格是用食品的消费者价格指数去除名义价格得到的。

资料来源：Y 的数据来自 Citibase，X_2 至 X_6 的数据来自美国农业部。感谢罗伯特·J·费希尔（Robert J. Fisher）收集数据并进行统计分析。

其中 Y＝人均鸡肉消费量，磅；

X_2 ＝人均真实可支配收入，美元；

X_3 ＝每磅鸡肉的真实零售价格，美分；

X_4 ＝每磅猪肉的真实零售价格，美分；

X_5 ＝每磅牛肉的真实零售价格，美分；

X_6 ＝每磅鸡肉替代品的综合真实价格，美分。这是每磅猪肉和牛肉真实零售价格的加权平均。其权数是猪肉和牛肉的总消费量中两者各自的相对消费量。

现考虑下面的需求函数：

$$\ln Y_t = \alpha_1 + \alpha_2 \ln X_{2t} + \alpha_3 \ln X_{3t} + u_t \tag{1}$$

$$\ln Y_t = \gamma_1 + \gamma_2 \ln X_{2t} + \gamma_3 \ln X_{3t} + \gamma_4 \ln X_{4t} + u_t \tag{2}$$

$$\ln Y_t = \lambda_1 + \lambda_2 \ln X_{2t} + \lambda_3 \ln X_{3t} + \lambda_4 \ln X_{5t} + u_t \tag{3}$$

$$\ln Y_t = \theta_1 + \theta_2 \ln X_{2t} + \theta_3 \ln X_{3t} + \theta_4 \ln X_{4t} + \theta_5 \ln X_{5t} + u_t \tag{4}$$

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{6t} + u_t \tag{5}$$

由微观经济学可知，对一种商品的需求通常都依赖于消费者的真实收入、该商品的真实价格，以及替代品或互补品的真实价格。按照这些思路，回答以下问题。

a. 从这里所列举的需求函数中你会选择哪一个？为什么？

b. 你怎样解释这些模型中的 $\ln X_2$ 和 $\ln X_3$ 的系数？

c. 模型（2）和（4）的设定有什么不同？

d. 如果你采用设定（4），你会预见到什么问题：（提示：猪肉和牛肉价格与鸡肉价格一道被引进。）

e. 因为设定（5）包含牛肉和猪肉的综合价格，你会认为需求函数（5）优于函数（4）吗？为

什么？

f. 猪肉和（或）牛肉是鸡肉的竞争或替代产品吗？你怎样知道？

g. 假定函数（5）是“正确”的需求函数。估计此模型的参数。计算它们的标准误，以及 R^2，$\bar{R}^2$ 和修正 R^2。解释你的结果。

h. 假如你使用“不正确”的模型（2）。通过考虑 γ_2 和 γ_3 值分别同 β_2 和 β_3 的关系，评估这一错误设定的后果。（提示：注意 7.7 节的讨论。）

7.20 在一项劳动市场的人员周转研究中，詹姆斯·F·拉根（James F. Ragan）对 1950 年第Ⅰ季度至 1979 年第Ⅳ季度期间的美国经济获得了如下结果。[①]（括号中的数字是估计的 t 统计量。）

$$\widehat{\ln Y_t} = 4.47 - 0.34 \ln X_{2t} + 1.22 \ln X_{3t} + 1.22 \ln X_{4t} + 0.80 \ln X_{5t} - 0.005\,5 X_{6t}$$

(4.28) (−5.31) (3.64) (3.10) (1.10) (−3.09)

$$\bar{R}^2 = 0.537\,0$$

注：我们将在下章讨论 t 统计量。

其中 Y=制造业中的辞职率，定义为每 100 个雇员中自愿离职的人数；

X_2 =成年男性失业率的工具变量或代理变量；

X_3 =年龄小于 25 岁的雇员百分比；

$X_4 = N_{t-1}/N_{t-4}$ = $t-1$ 季度与 $t-4$ 季度的制造业就业比率；

X_5 =妇女雇员百分比；

X_6 =时间趋势（1950—Ⅰ=1）。

a. 解释上面列出的结果。

b. 观测到 Y 的对数和 X_2 的对数之间呈负相关关系，在先验上是否说得过去？

c. 为什么 $\ln X_3$ 的系数是正的？

d. 既然趋势系数是负的，那么辞职率的百分比就存在一个长期下降趋势，为什么出现这样一种下降趋势呢？

e. $\bar{R}^2$ 是否“太”低？

f. 你能从所给数据估计回归系数的标准误吗？为什么？

7.21 下面考虑美国 1980—1998 年间的货币需求函数：

$$M_t = \beta_1 Y_t^{\beta_2} r_t^{\beta_3} e^{u_t}$$

其中 M=真实货币需求，利用货币的 M2 定义；

Y=真实 GDP；

r=利率。

为估计上述货币需求函数，为你提供了表 7—10 中的数据。

表 7—10　　1980—1998 年美国的货币需求

观测	GDP	M2	CPI	LTRATE	TBRATE
1980	2 795.6	1 600.4	82.4	11.27	11.506
1981	3 131.3	1 756.1	90.9	13.45	14.029
1982	3 259.2	1 911.2	96.5	12.76	10.686
1983	3 534.9	2 127.8	99.6	11.18	8.630

① 资料来源：参见 Ragan，“Turnover in the Labor Market：A Study of Quit and Layoff Rates，” *Economic Review*，Federal Reserve Bank of Kansas City，May 1981，pp. 13-22。

续前表

观测	GDP	M2	CPI	LTRATE	TBRATE
1984	3 932.7	2 311.7	103.9	12.41	9.580
1985	4 213.0	2 497.4	107.6	10.79	7.480
1986	4 452.9	2 734.0	109.6	7.78	5.980
1987	4 742.5	2 832.8	113.6	8.59	5.820
1988	5 108.3	2 995.8	118.3	8.96	6.690
1989	5 489.1	3 159.9	124.0	8.45	8.120
1990	5 803.2	3 279.1	130.7	8.61	7.510
1991	5 986.2	3 379.8	136.2	8.14	5.420
1992	6 318.9	3 434.1	140.3	7.67	3.450
1993	6 642.3	3 478.5	144.5	6.59	3.020
1994	7 054.3	3 502.2	148.2	7.37	4.290
1995	7 400.5	3 649.3	152.4	6.88	5.510
1996	7 813.2	3 824.2	156.9	6.71	5.020
1997	8 300.8	4 046.7	160.5	6.61	5.070
1998	8 759.9	4 401.4	163.0	5.58	4.810

注：GDP＝国内生产总值（十亿美元计）；
M2＝M2 货币供给；
CPI＝消费者价格指数（1982—1984＝100）；
LTRATE＝长期利率（30 年期国债）；
TBRATE＝3 月期国债利率（年百分比）。
资料来源：*Economic Report of the President*，2000，Tables B-1，B-58，B-67，B-71.

注：为了把名义变量转换成真实变量，将 M 和 GDP 除以 CPI。利率变量则不必除以 CPI。另外还要注意，这里给出了两个利率，一个是以 3 月期国债利率度量的短期利率，一个是以 30 年期国债收益率度量的长期利率，前面的经验研究已经使用过这两个利率。

a. 给定这些数据，估计货币需求对收入和利率的弹性。

b. 如果你不拟合上述需求函数，而代之以对模型 $(M/Y)_t = \alpha_1 r_t^{\alpha_2} e^{u_t}$ 的拟合，你会怎样解释所得到的结果？说明必需的计算。

c. 你如何决定哪个设定更好？（注：第 8 章将给出一个规范的统计检验。）

7.22　表 7—11 给出希腊 1961－1987 年制造业的数据。

表 7—11　　**希腊工业部门**

观测	产出*	资本	劳动**	资本劳动比
1961	35.858	59.600	637.0	0.093 6
1962	37.504	64.200	643.2	0.099 8
1963	40.378	68.800	651.0	0.105 7
1964	46.147	75.500	685.7	0.110 1
1965	51.047	84.400	710.7	0.118 8
1966	53.871	91.800	724.3	0.126 7
1967	56.834	99.900	735.2	0.135 9

续前表

观测	产出*	资本	劳动**	资本劳动比
1968	65.439	109.100	760.3	0.143 5
1969	74.939	120.700	777.6	0.155 2
1970	80.976	132.000	780.8	0.169 1
1971	90.802	146.600	825.8	0.177 5
1972	101.955	162.700	864.1	0.188 3
1973	114.367	180.600	894.2	0.202 0
1974	101.823	197.100	891.2	0.221 2
1975	107.572	209.600	887.5	0.236 2
1976	117.600	221.900	892.3	0.248 7
1977	123.224	232.500	930.1	0.250 0
1978	130.971	243.500	969.9	0.251 1
1979	138.842	257.700	1 006.9	0.255 9
1980	135.486	274.400	1 020.9	0.268 8
1981	133.441	289.500	1 017.1	0.284 6
1982	130.388	301.900	1 016.1	0.297 1
1983	130.615	314.900	1 008.1	0.312 4
1984	132.244	327.700	985.1	0.332 7
1985	137.318	339.400	977.1	0.347 4
1986	137.468	349.492	1 007.2	0.347 0
1987	135.750	358.231	1 000.0	0.358 2

注：* 以1970年不变价格的十亿德拉克马计。

** 以每千人计。

资料来源：感谢克里斯托弗纽波特大学 George K. Zestos 收集数据。

a. 看看柯布-道格拉斯生产函数是否能拟合表中给出的数据，并解释得到的结果。你能得到什么一般性结论？

b. 现考虑如下模型：

$$\text{产出}/\text{劳动} = A\,(K/L)^{\beta}e^{u}$$

其中回归子代表劳动生产率，回归元代表资本劳动比。这种关系有什么经济含义（如果有的话）？估计此模型并解释其结果。

7.23 **蒙特卡罗实验**：考虑如下模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

若告诉你 $\beta_1 = 262$，$\beta_2 = -0.006$，$\beta_3 = -2.4$，$\sigma^2 = 42$，$u_i \sim N(0, 42)$。从给定正态分布中生成64个观测 u_i 的10组集合，并利用表6—4中给出的64次观测（其中 Y=CM，X_2 =PGNP，X_3 = FLR）生成10个 β 系数的估计值集（每个集合都有三个估计系数）。求出每个 β 系数估计值的平均值，并将其与上面给出这些系数的真实值相联系，你能得到什么总体性结论？

7.24 表7—12给出了1947－2000年间美国真实消费支出、真实收入、真实财富和真实利率数据。这些数据还将用于习题8.35。

表 7—12　　1947—2000 年间美国真实消费支出、真实收入、真实财富和真实利率数据

年份	C	Yd	Wealth	Interest Rate	年份	C	Yd	Wealth	Interest Rate
1947	976.4	1 035.2	5 166.8	−10.351	1974	2 653.7	3 051.9	11 868.8	−1.043
1948	998.1	1 090.0	5 280.8	−4.720	1975	2 710.9	3 108.5	12 634.4	−3.534
1949	1 025.3	1 095.6	5 607.4	1.044	1976	2 868.9	3 243.5	13 456.5	−0.657
1950	1 090.9	1 192.7	5 759.5	0.407	1977	2 992.1	3 360.7	13 786.3	−1.190
1951	1 107.1	1 227.0	6 086.1	−5.283	1978	3 124.7	3 527.5	14 450.5	0.113
1952	1 142.4	1 266.8	6 243.9	−0.277	1979	3 203.2	3 628.6	15 340.0	1.704
1953	1 197.2	1 327.5	6 355.6	0.561	1980	3 193.0	3 658.0	15 965.0	2.298
1954	1 221.9	1 344.0	6 797.0	−0.138	1981	3 236.0	3 741.1	15 965.0	4.704
1955	1 310.4	1 433.8	7 172.2	0.262	1982	3 275.5	3 791.7	16 312.5	4.449
1956	1 348.8	1 502.3	7 375.2	−0.736	1983	3 454.3	3 906.9	16 944.8	4.691
1957	1 381.8	1 539.5	7 315.3	−0.261	1984	3 640.6	4 207.6	17 526.7	5.848
1958	1 393.0	1 553.7	7 870.0	−0.575	1985	3 820.9	4 347.8	19 068.3	4.331
1959	1 470.7	1 623.8	8 188.1	2.296	1986	3 981.2	4 486.6	20 530.0	3.768
1960	1 510.8	1 664.8	8 351.8	1.511	1987	4 113.4	4 582.5	21 235.7	2.819
1961	1 541.2	1 720.0	8 971.9	1.296	1988	4 279.5	4 784.1	22 332.0	3.287
1962	1 617.3	1 803.5	9 091.5	1.396	1989	4 393.7	4 906.5	23 659.8	4.318
1963	1 684.0	1 871.5	9 436.1	2.058	1990	4 474.5	5 014.2	23 105.1	3.595
1964	1 784.8	2 006.9	10 003.4	2.027	1991	4 466.6	5 033.0	24 050.2	1.803
1965	1 897.6	2 131.0	10 562.8	2.112	1992	4 594.5	5 189.3	24 418.2	1.007
1966	2 006.1	2 244.6	10 522.0	2.020	1993	4 748.9	5 261.3	25 092.3	0.625
1967	2 066.2	2 340.5	11 312.1	1.213	1994	4 928.1	5 397.2	25 218.3	2.206
1968	2 184.2	2 448.2	12 145.4	1.055	1995	5 075.6	5 339.1	27 439.7	3.333
1969	2 264.8	2 524.3	11 672.3	1.732	1996	5 237.5	5 677.7	29 448.2	3.083
1970	2 314.5	2 630.0	11 650.0	1.166	1997	5 423.9	5 854.5	32 664.1	3.120
1971	2 405.2	2 745.3	12 312.9	−0.712	1998	5 683.7	6 168.6	35 587.0	3.584
1972	2 550.5	2 874.3	13 499.9	−0.156	1999	5 968.4	6 320.0	39 591.3	3.245
1973	2 675.9	3 072.3	13 081.0	1.414	2000	6 257.8	6 539.2	38 167.7	3.576

注：C=真实消费支出，按链式法则以 1996 年十亿美元计算。

Yd=真实个人可支配收入，按链式法则以 1996 年十亿美元计算。

Wealth=真实财富，按链式法则以 1996 年十亿美元计算。

Interest Rate=3 月期国债年名义收益率－通货膨胀率（用价格指数按链式法则计算的年百分比变化率度量）。

名义和真实财富变量是利用美联储对家庭和非营利组织的资金账户流度量的年末净值数据生成的。而把名义财富变量转换成真实财富变量的价格指数，是当年第四季度与次年第四季度按链式法则计算的价格指数的平均。

资料来源：C、Yd 以及季度和年度链式价格指数（1996=100）：Bureau of Economic Analysis, U. S. Department of Commerce (http://www.bea.doc.gov/bea/dn1.htm)。

3 月期国债名义收益率：Economic Report of the President, 2002。

名义财富=家庭年末名义净财富和非利润（来自美联储资金数据：http://www.federalreserve.gov)。

a. 给定表中数据，利用收入、财富和利率估计线性消费函数。拟合的方程是什么？

b. 系数估计值说明各个变量与消费支出有什么关系？

7.25 **估计高通公司的股票价格。**作为多元回归的一个例子，考虑 1995—2000 年间高通（Qualcomm）公司（无线数字通信的设计者和制造商）的周股票价格数据。全部数据可以在本书网站的表 7—13 中找到。在 20 世纪 90 年代末期，技术股的利润特别高，但什么样的回归模型能够最好地拟合这些数据呢？图 7—4 给出了这些数据的一个基本散点图。

这个散点图看上去确实像加长的 S 曲线；平均股票价格看起来略有上升，但增长的速度在图的最右边急剧上升。随着对专业电话需求的急剧增加和技术革命的兴起，股票价格继续以更快的速度上升。

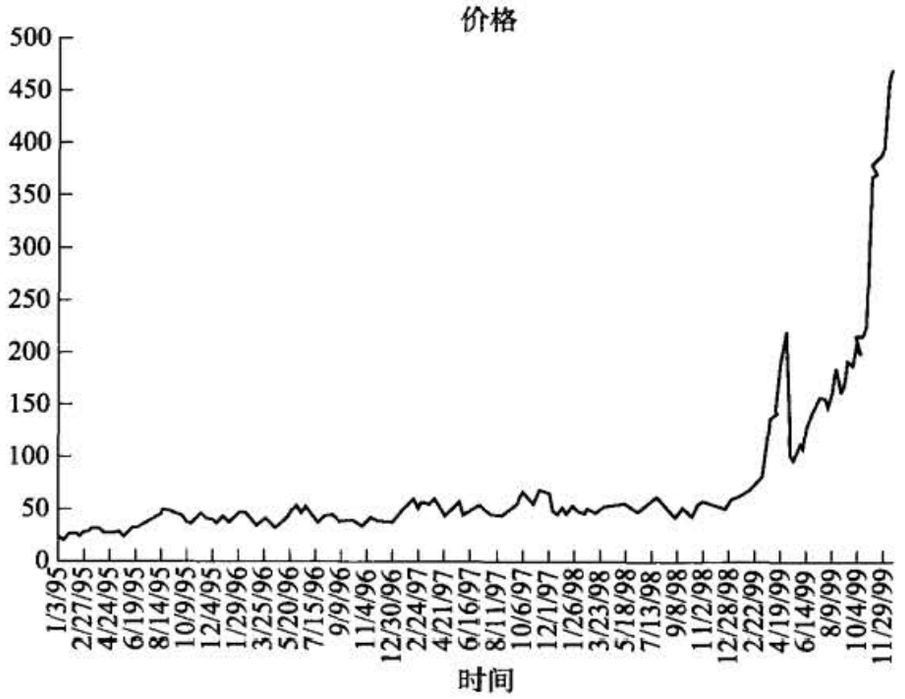


图 7—4　高通公司股票价格的时间变化图

a. 估计一个线性模型，基于时间预测股票收盘价格。这个模型看上去对数据的拟合很好吗？

b. 现在利用时间和时间的平方拟合一个二次模型。这个模型的拟合效果比（a）好吗？

c. 最后，拟合如下立方或三次模型

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

其中 Y=股票价格，X=时间。哪个模型看来是股票价格最好的估计？

附录 7A

□ 7A.1　方程（7.4.3）至（7.4.5）给出 OLS 估计量的推导

将方程

$$\sum a_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2 \qquad (7.4.2)$$

对三个未知数求偏导数，并令所得结果为零，得：

$$\frac{\partial \sum a_i^2}{\partial \hat{\beta}_1} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-1) = 0$$

$$\frac{\partial \sum a_i^2}{\partial \hat{\beta}_2} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{2i}) = 0$$

$$\frac{\partial \sum a_i^2}{\partial \hat{\beta}_3} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{3i}) = 0$$

化简后即得方程（7.4.3）至（7.4.5）。

顺便指出，以上三个方程又可写为：

$$\sum a_i = 0$$

$$\sum a_i X_{2i} = 0 \quad \text{（为什么?）}$$

$$\sum a_i X_{3i} = 0$$

从而表明最小二乘拟合的一些性质，即残差和为零，以及残差与解释变量 X_2 和 X_3 均不相关。

还容易看到，按照类似方法可以得到 k 变量线性回归模型（7.4.20）的 OLS 估计量。因此，先写出：

$$\sum a_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2$$

将此表达式对 k 个未知数的每一个求偏微分并令其结果等式为零，经适当整理后，即得 k 个未知数的 k 个正规方程如下：

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \cdots + \hat{\beta}_k \sum X_{ki}$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} + \cdots + \hat{\beta}_k \sum X_{2i} X_{ki}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \cdots + \hat{\beta}_k \sum X_{3i} X_{ki}$$

……

$$\sum Y_i X_{ki} = \hat{\beta}_1 \sum X_{ki} + \hat{\beta}_2 \sum X_{2i} X_{ki} + \hat{\beta}_3 \sum X_{3i} X_{ki} + \cdots + \hat{\beta}_k \sum X_{ki}^2$$

或者，变为小写字母，将这些方程表达为：

$$\sum y_i x_{2i} = \hat{\beta}_2 \sum x_{2i}^2 + \hat{\beta}_3 \sum x_{2i} x_{3i} + \cdots + \hat{\beta}_k \sum x_{2i} x_{ki}$$

$$\sum y_i x_{3i} = \hat{\beta}_2 \sum x_{2i} x_{3i} + \hat{\beta}_3 \sum x_{3i}^2 + \cdots + \hat{\beta}_k \sum x_{3i} x_{ki}$$

……

$$\sum y_i x_{ki} = \hat{\beta}_2 \sum x_{2i} x_{ki} + \hat{\beta}_3 \sum x_{3i} x_{ki} + \cdots + \hat{\beta}_k \sum x_{ki}^2$$

还应看到，k 变量模型也满足如下方程：

$$\sum a_i = 0$$

$$\sum a_i X_{2i} = \sum a_i X_{3i} = \cdots = \sum a_i X_{ki} = 0$$

□ 7A.2 方程（7.3.5）和（7.6.2）中 PGNP 系数的相等性质

令 Y=CM，X_2=PGNP，X_3=FLR，并用离差形式写成：

$$y_i = b_{13} x_{3i} + a_{1i} \tag{1}$$

$$x_{2i} = b_{23} x_{3i} + a_{2i} \tag{2}$$

现在将 a_i 对 a_2 回归得到：

$$a_1 = \frac{\sum \hat{u}_{1i}\hat{u}_{2i}}{\hat{u}_{2i}^2} = -0.005\ 6 \quad \text{（对本例）} \tag{3}$$

注意，由于 $\hat{u}$ 是残差，所以它们的均值都为零。利用方程（1）和（2），我们可以把方程（3）写成：

$$a_1 = \frac{\sum (y_i - b_{13}x_{3i})(x_{2i} - b_{23}x_{3i})}{\sum (x_{2i} - b_{23}x_{3i})^2} \tag{4}$$

展开上述表达式，并注意到：

$$b_{23} = \frac{\sum x_{2i}x_{3i}}{\sum x_{3i}^2} \tag{5}$$

以及

$$b_{13} = \frac{\sum y_i x_{3i}}{\sum x_{3i}^2} \tag{6}$$

把它们代入方程（4）即得到：

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$$

$$= -0.005\ 6 \quad \text{（对本例）} \tag{7.4.7}$$

□ 7A.3 方程（7.4.19）的推导

回忆：

$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}$$

又可写为：

$$\hat{u}_i = y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$$

其中小写字母指对均值的离差。

但是

$$\begin{aligned}\sum \hat{u}_i^2 &= \sum (\hat{u}_i \hat{u}_i) \\ &= \sum \hat{u}_i (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}) \\ &= \sum \hat{u}_i y_i\end{aligned}$$

这里利用了关系式 $\sum \hat{u}_i x_{2i} = \sum \hat{u}_i x_{3i} = 0$。（为什么?）又因为：

$$\sum \hat{u}_i y_i = \sum y_i \hat{u}_i = \sum y_i (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})$$

即有：

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i} \tag{7.4.19}$$

这就是所要的结果。

□ 7A.4 多元回归模型的极大似然估计法

将第 4 章附录 4A 中的概念加以推广，即可写出 k 变量线性回归模型（7.4.20）的对数似然函数为：

$$\ln L = -\frac{n}{2}\ln \sigma^2 - \frac{n}{2}\ln (2\pi) - \frac{1}{2}\sum \frac{(Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})^2}{\sigma^2}$$

将此函数分别对 β_1，β_2，…，β_k 和 σ^2 求偏微分，便得到以下的 $k+1$ 个方程：

$$\frac{\partial \ln L}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})(-1) \tag{1}$$

$$\frac{\partial \ln L}{\partial \beta_2} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})(-X_{2i}) \tag{2}$$

……

$$\frac{\partial \ln L}{\partial \beta_k} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})(-X_{ki}) \tag{k}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_k X_{ki})^2 \tag{$k+1$}$$

令这些偏导数为零（最优化的一阶条件），并用 $\tilde{\beta}_1$，$\tilde{\beta}_2$，…，$\tilde{\beta}_k$ 和 $\tilde{\sigma}^2$ 表示 ML 估计量，经代数化简后可得（前 k 个方程）：

$$\sum Y_i = n\tilde{\beta}_1 + \tilde{\beta}_2 \sum X_{2i} + \cdots + \tilde{\beta}_k \sum X_{ki}$$

$$\sum Y_i X_{2i} = \tilde{\beta}_1 \sum X_{2i} + \tilde{\beta}_2 \sum X_{2i}^2 + \cdots + \tilde{\beta}_k \sum X_{2i} X_{ki}$$

……

$$\sum Y_i X_{ki} = \tilde{\beta}_1 \sum X_{ki} + \tilde{\beta}_2 \sum X_{2i} X_{ki} + \cdots + \tilde{\beta}_k \sum X_{ki}^2$$

如同我们在附录 7A 第 7A.1 节所看到的那样，这正是最小二乘理论所导出的正规方程。因此，ML 估计量 $\tilde{\beta}$ 就是前面已给出的 OLS 估计量 $\hat{\beta}$。如第 4 章附录 4A 所指出的，两者的相等并非偶然。

把 ML（=OLS）估计量代入刚才推出的第 $k+1$ 个方程，经过化简，便得到 σ^2 的 ML 估计量为：

$$\tilde{\sigma}^2 = \frac{1}{n}\sum (Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_{2i} - \cdots - \tilde{\beta}_k X_{ki})^2$$

$$= \frac{1}{n}\sum \hat{u}_i^2$$

如本书已指出的，此估计量不同于 OLS 估计量 $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-k)$，而由于后者是 σ^2 的无偏估计，故引出结论：ML 估计量 σ^2 是一个有偏误的估计量。然而，容易验证，$\tilde{\sigma}^2$ 是渐近无偏的。

□ 7A.5 柯布-道格拉斯生产函数方程（7.9.4）的 EViews 输出结果

Dependent Variable: Y1
Method: Least Squares
Included observations: 51

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.887600	0.396228	9.811514	0.0000
Y2	0.468332	0.098926	4.734170	0.0000
Y3	0.521279	0.096887	5.380274	0.0000

R-squared	0.964175	Mean dependent var.	16.94139
Adjusted R-squared	0.962683	S.D. dependent var.	1.380870
S.E. of regression	0.266752	Akaike info criterion	0.252028
Sum squared resid.	3.415520	Schwarz criterion	0.365665
Log likelihood	−3.426721	Hannan-Quinn criterion	0.295452
F-statistic	645.9311	Durbin-Watson stat.	1.946387
Prob. (*F*-statistic)	0.000000		

Covariance of Estimates

	C	Y2	Y3
C	0.156997	0.010364	−0.020014
Y2	0.010364	0.009786	−0.009205
Y3	−0.020014	−0.009205	0.009387

Y	X2	X3	Y1	Y2	Y3	Y1HAT	Y1RESID
38,372,840	424,471	2,689,076	17.4629	12.9586	14.8047	17.6739	−0.2110
1,805,427	19,895	57,997	14.4063	9.8982	10.9681	14.2407	0.1656
23,736,129	206,893	2,308,272	16.9825	12.2400	14.6520	17.2577	−0.2752
26,981,983	304,055	1,376,235	17.1107	12.6250	14.1349	17.1685	−0.0578
217,546,032	1,809,756	13,554,116	19.1979	14.4087	16.4222	19.1962	0.0017
19,462,751	180,366	1,790,751	16.7840	12.1027	14.3981	17.0612	−0.2771
28,972,772	224,267	1,210,229	17.1819	12.3206	14.0063	16.9589	0.2229
14,313,157	54,455	421,064	16.4767	10.9051	12.9505	15.7457	0.7310
159,921	2,029	7,188	11.9824	7.6153	8.8802	12.0831	−0.1007
47,289,846	471,211	2,761,281	17.6718	13.0631	14.8312	17.7366	−0.0648
63,015,125	659,379	3,540,475	17.9589	13.3991	15.0798	18.0236	−0.0647
1,809,052	17,528	146,371	14.4083	9.7716	11.8939	14.6640	−0.2557
10,511,786	75,414	848,220	16.1680	11.2307	13.6509	16.2632	−0.0952
105,324,866	963,156	5,870,409	18.4726	13.7780	15.5854	18.4646	0.0079
90,120,459	835,083	5,832,503	18.3167	13.6353	15.5790	18.3944	−0.0778
39,079,550	336,159	1,795,976	17.4811	12.7253	14.4011	17.3543	0.1269
22,826,760	246,144	1,595,118	16.9434	12.4137	14.2825	17.1465	−0.2030
38,686,340	384,484	2,503,693	17.4710	12.8597	14.7333	17.5903	−0.1193
69,910,555	216,149	4,726,625	18.0627	12.2837	15.3687	17.6519	0.4109
7,856,947	82,021	415,131	15.8769	11.3147	12.9363	15.9301	−0.0532
21,352,966	174,855	1,729,116	16.8767	12.0717	14.3631	17.0284	−0.1517
46,044,292	355,701	2,706,065	17.6451	12.7818	14.8110	17.5944	0.0507
92,335,528	943,298	5,294,356	18.3409	13.7571	15.4822	18.4010	−0.0601
48,304,274	456,553	2,833,525	17.6930	13.0315	14.8570	17.7353	−0.0423
17,207,903	267,806	1,212,281	16.6609	12.4980	14.0080	17.0429	−0.3820
47,340,157	439,427	2,404,122	17.6729	12.9932	14.6927	17.6317	0.0411
2,644,567	24,167	334,008	14.7880	10.0927	12.7189	15.2445	−0.4564
14,650,080	163,637	627,806	16.5000	12.0054	13.3500	16.4692	0.0308
7,290,360	59,737	522,335	15.8021	10.9977	13.1661	15.9014	−0.0993
9,188,322	96,106	507,488	16.0334	11.4732	13.1372	16.1090	−0.0756
51,298,516	407,076	3,295,056	17.7532	12.9168	15.0079	17.7603	−0.0071
20,401,410	43,079	404,749	16.8311	10.6708	12.9110	15.6153	1.2158
87,756,129	727,177	4,260,353	18.2901	13.4969	15.2649	18.1659	0.1242
101,268,432	820,013	4,086,558	18.4333	13.6171	15.2232	18.2005	0.2328
3,556,025	34,723	184,700	15.0842	10.4552	12.1265	15.1054	−0.0212
124,986,166	1,174,540	6,301,421	18.6437	13.9764	15.6563	18.5945	0.0492
20,451,196	201,284	1,327,353	16.8336	12.2125	14.0987	16.9564	−0.1229
34,808,109	257,820	1,456,683	17.3654	12.4600	14.1917	17.1208	0.2445
104,858,322	944,998	5,896,392	18.4681	13.7589	15.5899	18.4580	0.0101
6,541,356	68,987	297,618	15.6937	11.1417	12.6036	15.6756	0.0181
37,668,126	400,317	2,500,071	17.4443	12.9000	14.7318	17.6085	−0.1642
4,988,905	56,524	311,251	15.4227	10.9424	12.6484	15.6056	−0.1829
62,828,100	582,241	4,126,465	17.9559	13.2746	15.2329	18.0451	−0.0892
172,960,157	1,120,382	11,588,283	18.9686	13.9292	16.2655	18.8899	0.0786
15,702,637	150,030	762,671	16.5693	11.9186	13.5446	16.5300	0.0394
5,418,786	48,134	276,293	15.5054	10.7817	12.5292	15.4683	0.0371
49,166,991	425,346	2,731,669	17.7107	12.9607	14.8204	17.6831	0.0277
46,164,427	313,279	1,945,860	17.6477	12.6548	14.4812	17.3630	0.2847
9,185,967	89,639	685,587	16.0332	11.4035	13.4380	16.2332	−0.2000
66,964,978	694,628	3,902,823	18.0197	13.4511	15.1772	18.0988	−0.0791
2,979,475	15,221	361,536	14.9073	9.6304	12.7981	15.0692	−0.1620

注：Y1=lnY；Y2=lnX2；Y3=lnX3。

本征值为 3.786 1 和 1 875 269，将在第 10 章给出。

第 8 章 多元回归分析：推断问题

本章为第 5 章的续篇。它把该章所讲的区间估计与假设检验的思想扩展到涉及三个或多个变量的模型上来。虽然第 5 章所论述的概念在许多方面都可直截了当地应用于多元回归，但有少数特点则为多元回归模型所特有。而正是这些特点在本章中受到了更多的关注。

8.1 再议正态性假定

至此我们知道，如果我们唯一的目的是对回归模型的参数进行点估计，则普通最小二乘法将足够使用，并不需要对干扰项 u_i 的概率分布做任何假定。但若我们的目的在于估计和推断两个方面，则如同第 4 章和第 5 章所述，我们还需要假定 u_i 服从某个概率分布。

由于已明确说明过的一些理由，我们曾假定 u_i 服从均值为零、方差 σ^2 为常数的正态分布。我们对多元回归模型继续做同样的假定。有了正态性假定，并参照第 4 章和第 7 章的讨论，我们发现，偏回归系数的 OLS 估计量无异于 ML 估计量，是最优线性无偏估计量（BLUE）。[①] 此外，估计量 $\hat{\beta}_2$、$\hat{\beta}_3$ 和 $\hat{\beta}_1$ 本身也是正态分布的，其均值等于 β_2、β_3 和 β_1，而方差在第 7 章给出了。而且，$(n-3)\hat{\sigma}^2/\sigma^2$ 服从自由度为

① 有了正态性假定，OLS 估计量 $\hat{\beta}_1$、$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 就是在整个无偏估计类中的最小方差估计量，不管它是不是线性估计量。简言之，它们是 BUE（最优无偏估计量）。参看 C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York, 1965, p. 258。

$n-3$ 的 χ^2 分布，并且三个 OLS 估计量均独立于 $\hat{\sigma}^2$ 而分布，其证明类似于附录 3A 所讨论的双变量情形。因此，再参照第 5 章，即可证明，在标准误的计算中，若 σ^2 由其无偏估计 $\hat{\sigma}^2$ 代替，则：

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \tag{8.1.1}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)} \tag{8.1.2}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\text{se}(\hat{\beta}_3)} \tag{8.1.3}$$

均服从自由度为 $n-3$ 的 t 分布。

注意自由度为 $n-3$ 是因为在计算 $\sum \hat{u}_i^2$ 并因此计算 $\hat{\sigma}^2$ 之前，我们要先估计三个偏回归系数，从而给残差平方和的计算加上三个约束。（按照这个逻辑，在四变量情形中将有 $n-4$ 个自由度，依此类推。）于是，t 分布可用于构造关于真实总体偏回归系数的置信区间并检验统计假设。同理，χ^2 分布可用于检验关于真实 σ^2 的假设。我们用下面的说明性例子来阐明具体的操作步骤。

例 8.1　修正儿童死亡率例子

在第 7 章，我们对一个由 64 个国家构成的样本将儿童死亡率（CM）对人均 GNP（即 PGNP）和妇女识字率（FLR）进行回归。方程（7.6.2）中给出的回归结果增加某些信息后复制如下：

$$\begin{aligned}
\widehat{\text{CM}}_i &= 263.641\,6 - 0.005\,6\,\text{PGNP}_i - 2.231\,6\text{FLR}_i \\
\text{se} &= (11.593\,2) \quad (0.001\,9) \quad (0.209\,9) \\
t &= (22.741\,1) \quad (-2.818\,7) \quad (-10.629\,3) \\
p\text{ 值} &= (0.000\,0)^* \quad (0.006\,5) \quad (0.000\,0)^* \\
R^2 &= 0.707\,7 \qquad \bar{R}^2 = 0.698\,1
\end{aligned} \tag{8.1.4}$$

其中*表示极小值。

在方程（8.1.4）中，我们沿袭了方程（5.11.1）中首次引入的格式，其中第一行括号中的数字是估计标准误，第二行表示相关总体系数为零的虚拟假设下的 t 值，第三行表示估计的 p 值，还给出 R^2 和调整 R^2 值。我们已在例 7.1 中解释过这个回归。

所观察到这些结论的统计显著性如何？比如，考虑 PGNP 的系数 −0.005 6。这个系数是统计显著的吗？即它统计显著地异于零吗？类似地，FLR 的系数 −2.231 6 是统计显著的吗？这两个系数都是统计显著的吗？为回答这个问题及与此相关的问题，让我们首先考虑在多元回归模型中可能会遇到的假设检验类型。

8.2　多元回归中的假设检验：总评

一旦我们走出简单的双变量线性回归模型的范围，假设检验就会以多种有趣的

形式出现，诸如：

1. 检验关于个别偏回归系数的假设（8.3 节）。

2. 检验所估计的多元回归模型的总体显著性，也就是要判别是否全部偏斜率系数同时为零（8.4 节）。

3. 检验两个或多个系数是否相等（8.5 节）。

4. 检验偏回归系数是否满足某种约束条件（8.6 节）。

5. 检验所估计的回归模型在时间上或在不同横截面单元上的稳定性（8.7 节）。

6. 检验回归模型的函数形式（8.8 节）。

因为在经验分析中常常出现这些类型的一种或多种检验，我们将分节讨论每一种类型的检验。

8.3 检验关于个别偏回归系数的假设

如在 8.1 节中指出的，引用假定 $u_i \sim N(0, \sigma^2)$，便可用 t 检验统计量对任一个别的偏回归系数的假设进行检验。为了说明操作步骤，考虑我们的数值例子，我们假设：

$$H_0: \beta_2 = 0 \qquad \text{和} \qquad H_1: \beta_2 \neq 0$$

虚拟假设是说，保持 X_3（妇女识字率）不变，X_2（人均 GNP）对 Y（儿童死亡率）无（线性）影响。[①] 为了检验这个虚拟假设，我们利用方程（8.1.2）中给出的 t 检验。参照第 5 章的做法，如果计算的 t 值超过选定显著性水平的临界 t 值，便拒绝假设；否则，就不拒绝它。对于我们的例子，利用方程（8.1.2）并注意到在虚拟假设下 $\beta_2 = 0$，我们得到：

$$t = \frac{-0.005\,6}{0.002\,0} = -2.818\,7 \tag{8.3.1}$$

如方程（8.1.4）所示。

注意我们有 64 个观测。因此本例中的自由度为 61。（为什么?）你若查阅附录 D 中的 t 表，没有与自由度 61 相对应的数据。与之最接近的自由度是 60。若使用自由度 60，并假设显著性水平（即犯第 I 类错误的概率）α 为 5%，则双尾检验的临界 t 值为 2.0（自由度为 60 的 $t_{\alpha/2}$），而单尾检验的临界 t 值为 1.671（自由度为 60 的 t_{α}）。

对于本例，对立假设是双侧的。因此我们使用双尾 t 值。既然计算出来的 t 值 2.818 7（绝对值）超过了临界 t 值 2，那我们就可以拒绝人均 GNP 对儿童死亡率没有影响的虚拟假设。更明确地讲，保持妇女识字率不变，人均 GNP 对儿童死亡率具

① 在大多数经验研究中，虚拟假设都被叙述成这种形式，即采取如下的极端立场（一种稻草人策略）：因变量与所考虑的解释变量之间无任何关系。用意是要从判明两变量之间是否存在一个无关紧要的关系开始。

有显著的（负面）影响，这与先验预期完全一致。图 8—1 从图形上说明了这一情形。

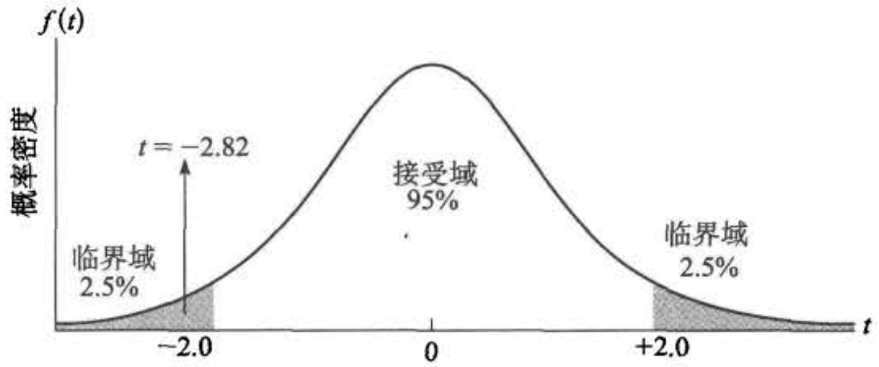


图 8—1　t 的 95%置信区间（自由度为 60）

实践中，人们不必假定一个特定的 α 值来进行假设检验，仅使用方程（8.1.4）中的 p 值即可。本例中的 p 值就是 0.006 5。对这个 p 值（即精确的显著性水平）的解释是，如果虚拟假设正确，得到一个大于等于 2.818 7（在绝对值上）的 t 值的概率仅为 0.006 5 或 0.65%，这确实是一个相当小的概率，比人为选定的 $\alpha=5\%$ 小得多。

本例还为我们提供了一个决定是用单尾还是用双尾 t 检验的机会。既然推测儿童死亡率与人均 GNP 负相关（为什么?），那我们就应该使用单尾检验。即虚拟假设和对立假设应该是：

$$H_0:\beta_2 < 0 \qquad \text{和} \qquad H_1:\beta_2 \geqslant 0$$

读者已经知道，我们在本例中能基于单尾 t 检验拒绝虚拟假设。如果我们在一个双尾检验中拒绝一个虚拟假设，那么，由于在单尾检验中仅在一个方向上使用这个统计量进行判断，我们就有足够的证据拒绝这个假设。

在第 5 章中，我们曾看到假设检验和置信区间估计之间的密切关系。用我们的例子来说，β_2 的 95%置信区间是：

$$\hat{\beta}_2 - t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2) \leqslant \beta_2 \leqslant \hat{\beta}_2 + t_{\alpha/2}\,\mathrm{se}(\hat{\beta}_2)$$

具体到我们的例子就变成：

$$-0.005\,6 - 2\times 0.002\,0 \leqslant \beta_2 \leqslant -0.005\,6 + 2\times 0.002\,0$$

即：

$$-0.009\,6 \leqslant \beta_2 \leqslant -0.001\,6 \qquad (8.3.2)$$

也就是说，β_2 以 95%的置信系数落在 −0.009 6 与 −0.001 6 之间。因此，如果选取了容量同样为 64 的 100 个样本并构造像方程（8.3.2）这样的 100 个置信区间，则我们预期其中的 95 个包含着真实总体参数 β_2。由于虚拟假设的零值不落在区间（8.3.2）内，故能以 95%的置信系数拒绝虚拟假设即真实 $\beta_2 = 0$。

由此，无论我们用方程（8.3.1）中的 t 显著性检验还是用方程（8.3.2）的置信区间估计，我们都得到同样的结论。但鉴于置信区间估计与假设检验之间的密切关

系，这也无足为奇。

按照上述方法，我们可以利用方程（8.2.1）提供的数据，检验关于儿童死亡率回归模型中的其他参数假设。例如，假设我们想检验的假设是保持人均 GNP 的影响不变，妇女识字率对儿童死亡率没有什么影响。我们确信能拒绝该假设，因为在此假设下，得到一个绝对值大于等于 10.6 的 t 值的 p 值实际上是 0。

在继续讨论之前，记住 t 检验的程序是基于误差项 u_i 服从正态分布的假定。尽管我们不能直接观测 u_i，但我们能够观测到它们的代理变量 $\hat{u}_i$，即残差。对儿童死亡率一例而言，残差直方图如图 8—2 所示。

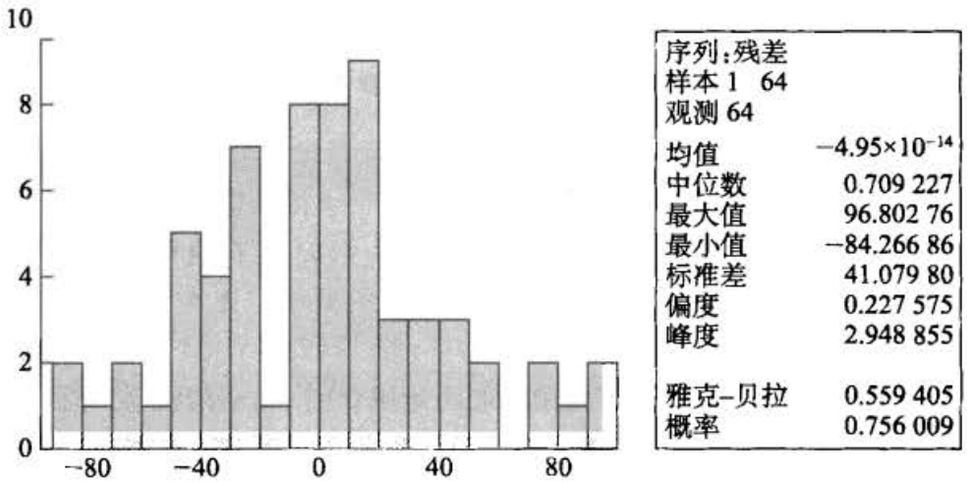


图 8—2　回归（8.1.4）的残差直方图

从直方图来看，残差是正态分布的。我们也可以计算方程（5.12.1）所示的雅克-贝拉（JB）正态性检验。这里的 JB 值为 0.559 4，p 值为 0.76。[①] 因此，本例中的误差项看来服从正态分布。当然也要记住，JB 检验是一个大样本检验，我们在这个例子中只有 64 次观测可能不够多。

8.4　检验样本回归的总显著性

在整个上一节中，我们都只是在每一真实偏回归系数为零的单个假设下讨论偏回归系数估计值的个别显著性的检验问题，而现在我们考虑如下假设：

$$H_0: \beta_2 = \beta_3 = 0 \tag{8.4.1}$$

这个虚拟假设是关于 β_2 和 β_3 联合地或同时地等于零的一个联合假设。对这样一个假设的检验被称作对所观测到的或所估计回归线的**总显著性**（overall significance）

① 对我们的例子而言，偏态值为 0.227 6，峰态值为 2.948 8。而对一个正态分布变量而言，偏态值和峰态值分别为 0 和 3。

检验，也就是检验 Y 是否与 X_2 和 X_3 存在线性关系。

能不能像 8.3 节那样，逐一检验 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的显著性来检验方程（8.4.1）中的联合假设呢？答案是否定的，理由如下。

在 8.3 节，检验一个所观测到的偏回归系数的个别显著性时，我们隐含地假定每一个显著性检验都是根据一个不同的（即独立的）样本进行的。这样，在假设 $\beta_2 = 0$ 下检验 $\hat{\beta}_2$ 的显著性时，我们无形地假定了用于这一检验的样本不同于在假设 $\beta_3 = 0$ 下用来检验 $\hat{\beta}_3$ 的显著性的那个样本。但是，如果我们用同一样本数据去检验方程（8.4.1）中的联合假设，我们就违反了检验方法所依据的基本假定。[①] 这一问题可另行表述如下：在方程（8.3.2）中我们对 β_2 构造了一个 95% 置信区间。如果我们仍按 95% 的置信系数用同样的样本数据构造 β_3 的一个置信区间，我们就不能断言 β_2 和 β_3 同时落在其各自置信区间上的（置信）概率是 $(1-\alpha)(1-\alpha) = 0.95 \times 0.95$。

换言之，虽然下面两个命题

$$\Pr[\hat{\beta}_2 - t_{\alpha/2}\mathrm{se}(\hat{\beta}_2) \leqslant \beta_2 \leqslant \hat{\beta}_2 + t_{\alpha/2}\mathrm{se}(\hat{\beta}_2)] = 1-\alpha$$

$$\Pr[\hat{\beta}_3 - t_{\alpha/2}\mathrm{se}(\hat{\beta}_3) \leqslant \beta_3 \leqslant \hat{\beta}_3 + t_{\alpha/2}\mathrm{se}(\hat{\beta}_3)] = 1-\alpha$$

个别地看是正确的，但认为 β_2 和 β_3 同时落入区间 $[\hat{\beta}_2 \pm t_{\alpha/2}\mathrm{se}(\hat{\beta}_2), \hat{\beta}_3 \pm t_{\alpha/2}\mathrm{se}(\hat{\beta}_3)]$ 的概率是 $(1-\alpha)^2$ 就不对。因为如果用同样的数据去推导这些区间，这些区间就不会是独立的。换句话说，

> ……检验一系列单个假设，不等于联合地检验同样的这些假设。其直观上的理由是，在对几个假设的一个联合检验中，任何一个单个假设都受其他假设所含信息的“影响”[②]。

以上论证的要点在于：对于一个给定的实例（样本），只能得到一个置信区间或者只能作出一个显著性检验。那么，我们怎样来检验联立的虚拟假设 $\beta_2 = \beta_3 = 0$ 呢？答案如下所示。

□ 检验所观测到的多元回归的总显著性的方差分析法：F 检验

这个理由刚才解释过了，我们不能用通常的 t 检验去检验多个真实偏斜率系数同时为零的联合假设。然而，这个联合假设可以用在 5.9 节曾介绍的**方差分析**（analysis of variance，ANOVA）去检验。现说明如下。

回忆恒等式：

$$\sum y_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \sum \hat{u}_i^2 \qquad (8.4.2)$$

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$

① 在任一给定样本中，协方差 $\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3)$ 未必是零；即 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 可能相关。见方程（7.4.17）。

② Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods*, Springer-Verlag, New York, 1984, p. 37.

和平常一样，TSS有 $n-1$ 个自由度，而RSS有 $n-3$ 个自由度，它的理由已经讨论过了。ESS因为是 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的函数所以有2个自由度。因此，按照5.9节讨论的ANOVA程序，我们可以列出表8—1。

表8—1　　三变量回归的ANOVA表

变异来源	SS	df	MSS
来自回归（ESS）	$\hat{\beta}_2\sum y_i x_{2i}+\hat{\beta}_3\sum y_i x_{3i}$	2	$\dfrac{\hat{\beta}_2\sum y_i x_{2i}+\hat{\beta}_3\sum y_i x_{3i}}{2}$
来自残差（RSS）	$\sum \hat{u}_i^2$	$n-3$	$\hat{\sigma}^2=\dfrac{\sum \hat{u}_i^2}{n-3}$
总计	$\sum y_i^2$	$n-1$	

现在可以证明[①]，在 u_i 的正态分布假定下以及在虚拟假设 $\beta_2=\beta_3=0$ 下，变量：

$$F=\frac{(\hat{\beta}_2\sum y_i x_{2i}+\hat{\beta}_3\sum y_i x_{3i})/2}{\sum \hat{u}_i^2/(n-3)}=\frac{\text{ESS/df}}{\text{RSS/df}} \tag{8.4.3}$$

服从自由度为2和 $n-3$ 的 F 分布。

上述 F 有什么用？可以证明[②]，在 $u_i\sim N(0,\sigma^2)$ 的假定下，

$$E\frac{\sum \hat{u}_i^2}{n-3}=E(\hat{\sigma}^2)=\sigma^2 \tag{8.4.4}$$

再加上假定 $\beta_2=\beta_3=0$，便能证明：

$$\frac{E(\hat{\beta}_2\sum y_i x_{2i}+\hat{\beta}_3\sum y_i x_{3i})}{2}=\sigma^2 \tag{8.4.5}$$

因此，如果虚拟假设是真实的，方程（8.4.4）和（8.4.5）都将对真实 σ^2 给出同样的估计。这一命题无足为奇。因为如果 Y 与 X_2 和 X_3 的关系微不足道，则 Y 变异的唯一来源是来自 u_i 所代表的随机势力。然而，如果虚拟假设错误，即 X_2 和 X_3 确实影响 Y，则不能在方程（8.4.4）和（8.4.5）之间画等号。这时，在适当考虑自由度之后，ESS要相对大于RSS。从而，方程（8.4.3）的 F 值对真实斜率系数同时为零的这一虚拟假设提供了一种检验。如果从方程（8.4.3）算出的 F 值大于 $\alpha\%$ 显著水平的 F 表中的 F 临界值，我们就拒绝 H_0；否则就不拒绝它。另一种方法是，如果所观测到的 F 的 p 值足够低，则可拒绝 H_0。

表8—2对 F 检验进行了总结。回到我们的例子，我们得出表8—3。

利用（8.4.3）即得：

$$F=\frac{128\,681.2}{1\,742.88}=73.832\,5 \tag{8.4.6}$$

① 参看K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, New York, 1960, pp. 278-280。

② 同上。

表 8—2　　F 统计量小结

虚拟假设 H_0	对立假设 H_1	拒绝 H_0 的临界域
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{S_1^2}{S_2^2} > F_{\alpha,n\text{ df},d\text{ df}}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{S_1^2}{S_2^2} > F_{\alpha/2,n\text{ df},d\text{ df}}$ 或 $< F_{(1-\alpha/2),n\text{ df},d\text{ df}}$

注：1. σ_1^2 和 σ_2^2 为两个总体方差。
2. s_1^2 和 s_2^2 为两个样本方差。
3. n df 和 d df 分别为分子自由度和分母自由度。
4. 在计算 F 比率时，将较大的 S^2 值放在分子上。
5. F 临界值在最后一列给出，F 的第一个下标是显著性水平，后两个下标为分子和分母自由度。
6. 注意 $F_{(1-\alpha/2),n\text{ df},d\text{ df}} = 1/F_{\alpha/2,n\text{ df},d\text{ df}}$。

表 8—3　　儿童死亡率一例的 ANOVA 表

变异来源	SS	df	MSS
来自回归	257 362.4	2	128 681.2
来自残差	106 315.6	61	1 742.88
总计	363 678	63	

得到一个大于或等于 73.832 5 的 F 值的 p 值几乎是 0，从而拒绝 PGNP 和 FLR 同时对儿童死亡率没有影响的假设。如果使用惯常的 5%的显著性水平，分子自由度为 2 和分母自由度为 60（实际自由度为 61）的 F 临界值约为 3.15，若用 1%的显著性水平，F 临界值则约为 4.98。显然，观察到约为 74 的 F 值比这些 F 临界值中的任何一个都大得多。

我们可将上述 F 检验方法推广到一般情形。

□ 检验多元回归的总显著性：F 检验

决策规则

给定 k 变量回归模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

检验假设：

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

（即全部斜率系数同时为零），相对于

$$H_1: \text{并非全部斜率系数同时为零}$$

计算

$$F = \frac{\text{ESS/df}}{\text{RSS/df}} = \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \tag{8.4.7}$$

如果 $F > F_\alpha(k-1, n-k)$，则拒绝 H_0；否则不拒绝它，其中 $F_\alpha(k-1, n-k)$ 是显著性水平为 α、分子自由度为 $k-1$ 和分母自由度为 $n-k$ 时的 F 临界值。换言之，如果由方程（8.4.7）得到的 F 值的 p 值足够低，即可拒绝 H_0。

毋庸赘言，在三变量情形(Y 和 X_2、X_3)中 $k=3$，在四变量情形中 $k=4$，如此类推。

顺便一提，大多数回归软件都例行把（由方差分析表给出的）F 值连同通常的回归结果诸如系数估计值及其标准误（差）和 t 值等一起算出。在 t 值的计算中，通常都把虚拟假设取为 $\beta_i=0$。

个别与联合假设检验对比。在 8.3 节中，我们讨论了个别偏回归系数的显著性检验，并在 8.4 节中我们又讨论了整个回归（即全部系数同时为零）的总显著性检验。**我们要重申这两类检验是不相同的。**因此，有可能根据 t 检验或（8.3 节的）置信区间而接受某一系数 β_k 为零的假设，但另一方面却拒绝全部系数为零的联合假设。

> 应吸取的经验教训是，个别的（多个）置信区间所提供的联合“信息”，不能代替假设的联合检验和联合置信命题中所蕴涵的联合置信区域（蕴涵于 F 检验中）。①

□ R^2 和 F 之间的一个重要关系式

判定系数 R^2 与方差分析中所用的 F 检验之间存在密切关系。假定干扰项 u_i 为正态分布，并且虚拟假设 $\beta_2=\beta_3=0$ 成立，我们曾看到：

$$F=\frac{\text{ESS}/2}{\text{RSS}/(n-3)} \tag{8.4.8}$$

服从自由度为 2 和 $n-3$ 的 F 分布。

推广到 k 变量情形（包括截距），如果假定干扰项是正态分布的，而且有虚拟假设：

$$H_0:\beta_2=\beta_3=\cdots=\beta_k=0 \tag{8.4.9}$$

则随之有：

$$F=\frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \tag{8.4.7) = (8.4.10}$$

服从自由度为 $k-1$ 和 $n-k$ 的 F 分布。（注：待估计的参数个数是 k，其中之一为截距项。）

让我们对方程（8.4.10）进行如下推导：

$$\begin{aligned}F&=\frac{n-k}{k-1}\cdot\frac{\text{ESS}}{\text{RSS}}\\&=\frac{n-k}{k-1}\cdot\frac{\text{ESS}}{\text{TSS}-\text{ESS}}\\&=\frac{n-k}{k-1}\cdot\frac{\text{ESS/TSS}}{1-(\text{ESS/TSS})}\\&=\frac{n-k}{k-1}\cdot\frac{R^2}{1-R^2}\\&=\frac{R^2/(k-1)}{(1-R^2)/(n-k)}\end{aligned} \tag{8.4.11}$$

① Fomby et al., op. cit., p. 42.

其中我们用到了定义 $R^2=\text{ESS/TSS}$。方程（8.4.11）表明 F 和 R^2 是何种关系。两者是同向变化的。当 $R^2=0$ 时，F 随之等于零。R^2 越大，F 值也越大。在极限处，当 $R=1$ 时，F 无限大。因此，F 检验既是所估计的回归的总显著性的一个度量，也是 R^2 的一个显著性检验。换句话说，检验虚拟假设（8.4.9）等价于检验（总体）R^2 等于零的虚拟假设。

对于三变量情形，方程（8.4.11）变为：

$$F=\frac{R^2/2}{(1-R^2)/(n-3)} \tag{8.4.12}$$

利用 F 与 R^2 之间的紧密联系，可把 ANOVA 表 8—1 重新设计成表 8—4。

表 8—4　　用 R^2 表示的 ANOVA 表

变异来源	SS	df	MSS*
来自回归	$R^2(\sum y_i^2)$	2	$R^2(\sum y_i^2)/2$
来自残差	$(1-R^2)(\sum y_i^2)$	$n-3$	$(1-R^2)(\sum y_i^2)/(n-3)$
总计	$\sum y_i^2$	$n-1$	

注：* 在计算 F 值时无需用 $\sum y_i^2$ 乘以 R^2 和 $1-R^2$，因为如方程（8.4.12）所示，它将被消掉。

对于我们的说明性例子，利用方程（8.4.12）得到：

$$F=\frac{0.707\ 7/2}{(1-0.707\ 7)/61}=73.872\ 6$$

除四舍五入的误差外，这个 F 值与前面得到的 F 值相同。

用 R^2 来表示 F 检验的一个优点在于计算上的便利：所需要知道的仅仅是 R^2 值而已。因此，方程（8.4.7）所给的总显著性 F 检验可重新用 R^2 表达，如表 8—4 所示。

□ 用 R^2 表述的多元回归总显著性检验

决策规则

用 R^2 表述的多元回归总显著性检验：另一个等价于方程（8.4.7）的检验。

给定 k 变量回归模型：

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\cdots+\beta_kX_{ki}+u_i$$

为了检验假设

$$H_0:\beta_2=\beta_3=\cdots=\beta_k=0$$

相对于

$$H_1:\text{非全部斜率系数同时为零}$$

计算：

$$F=\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{8.4.13}$$

如果 $F > F_\alpha(k-1, n-k)$，则拒绝 H_0；否则可接受 H_0，其中 $F_\alpha(k-1, n-k)$ 是显著性水平为 α、分子自由度为 $k-1$ 和分母自由度为 $n-k$ 时的 F 临界值。换言之，如果由方程（8.4.13）所得到的 F 值的 p 值足够低，即可拒绝 H_0。

在继续讨论之前，回头看一下第 7 章的例 7.5。我们从（7.10.7）看到，RGDP（相对人均 GDP）和 RGDP 的平方只解释了 190 个国家构成的样本中 GDPG（GDP 增长率）变异的 10.92%。$R^2 = 0.109\ 2$ 看起来是个“低”值。它真的在统计上异于零吗？我们如何说明？

回想我们前面在“R^2 和 F 之间的一个重要关系式”专题中，对方程（8.4.11）或两个回归元特殊情形下的方程（8.4.12）给出的 R^2 和 F 值之间的关系展开了讨论。前面曾指出，若 R^2 为零，则 F 因此为零，若回归元对回归子没有影响便是如此。因此，若我们将 $R^2 = 0.109\ 2$ 代入公式（8.4.12），我们便得到

$$F = \frac{0.109\ 2/2}{(1-0.109\ 2)/187} = 11.461\ 8$$

在虚拟假设 $R^2 = 0$ 下，上述 F 值服从自由度为 2 和 187 的 F 分布。（注意，有 190 个观测和 2 个回归元。）我们从 F 表看出，这个 F 值在 5%的显著性水平上仍是显著的，p 值实际上是 0.000 02。因此，尽管 R^2 只有 0.109 2，但我们仍能拒绝两个回归元对回归子没有影响的虚拟假设。

本例引申出的一个重要的经验是，在涉及几个变量观测值的横截面数据中，由于横截面单元的多样性，所以得到的 R^2 一般都很低。所以，对横截面回归中得到低 R^2 值不用感到吃惊或着急。重要的是，正确地设定模型，回归元具有正确（即理论预期）的符号，以及统计显著（希望如此）的回归系数。读者应该在 5%或更好（即低于 5%）的显著性水平上验证方程（7.10.7）中的两个回归元都是个别统计显著的。

□ 一个解释变量的“增量”或“边际”贡献

我们在第 7 章中说过，我们一般不能将 R^2 值在各个回归元之间分配。在我们儿童死亡率的例子中，我们发现 R^2 为 0.707 7，但由于这两个回归元在我们手头的样本中可能相关，所以我们不知道哪些值属于回归元 PGNP 的功劳，哪些又属于 FLR 的功劳。利用方差分析的方法，我们可以对此有更深入的了解。

对我们的说明性例子而言，我们发现 X_2（即 PGNP）和 X_3（即 FLR）基于（单独进行的）t 检验都是个别统计显著的。我们还发现，基于总体的 F 检验，这两个回归元共同对回归子 Y（儿童死亡率）产生明显影响。

现在假设我们单独引入 PGNP 和 FLR；即首先将儿童死亡率对 PGNP 回归并评价其显著性，然后在模型中增加 FLR，以判明它是否有任何贡献（当然，可以对调 X_2 和 X_3 进入的次序）。所谓贡献，我们意指在模型中增加一个变量，是否相对于

RSS“显著”地增加了 ESS（从而影响 R^2）。把这一贡献称作一个解释变量的**增量**（incremental）或**边际**（marginal）贡献，也许是适当的。

在实践中，增量贡献的问题是一个重要的论题。在大多数经验研究工作中，研究者对一个已含有若干个变量的模型是否值得再添加一个新的 X 变量，是没有足够把握的。研究者不愿意加入那些对 ESS 贡献很少的变量。同样的道理，研究者不愿意排除一些能实质上增加 ESS 的变量。但怎样决定一个 X 变量的引进是否能显著地减少 RSS 呢？容易通过方差分析的应用来回答这个问题。

假设我们先做儿童死亡率对 PGNP 的回归，并得到下述结果：

$$\widehat{CM}_i = 157.424\ 4 - 0.011\ 4\ PGNP \qquad (8.4.14)$$
$$t = (15.989\ 4)(-3.515\ 6) \qquad r^2 = 0.166\ 2$$
$$p\ 值 = (0.000\ 0)(0.000\ 8) \qquad 调整\ r^2 = 0.152\ 8$$

如这些结论所示，PGNP 对 CM 具有明显影响。表 8—5 给出了上述回归的 ANOVA 表。

表 8—5　　回归方程（8.4.14）的 ANOVA 表

变异来源	SS	df	MSS
ESS（来自 PGNP）	60 449.5	1	60 449.5
RSS	303 228.5	62	4 890.782 2
总计	363 678	63	

假定干扰项 u_i 是正态分布的，并且在 PGNP 对 CM 没有影响的虚拟假设下，我们知道

$$F = \frac{60\ 449.5}{4\ 890.782\ 2} = 12.359\ 8 \qquad (8.4.15)$$

服从自由度为 1 和 62 的 F 分布。因为计算出来的 p 值为 0.000 8，所以这个 F 值是高度显著的。因此，和前面一样，我们拒绝 PGNP 对 CM 没有影响的假设。顺带指出，$t^2 = (-3.515\ 6)^2 = 12.359\ 4$。这近似等于方程（8.4.15）中的 F 值，其中 t 值是从方程（8.4.14）中得到的。但这个结果也无足为奇，因为如第 5 章第一次证明的那样，在同样的虚拟假设和显著性水平下，自由度为 n 的 t 统计量等于自由度为 1 和 n 的 F 值。本例中 n=64。

假如在做完回归（8.4.14）之后，我们决定把 X_3（即 FLR）增加到模型中来，并得到多元回归（8.1.4）。我们需要回答的问题是：

1. 知道 PGNP 位于模型中并且和 CM 有显著关系，FLR 的边际或增量贡献是什么？

2. FLR 的增量贡献在统计上显著吗？

3. 根据什么准则把变量加入模型？

可通过 ANOVA 技术来回答上述问题。为了说明这一点，让我们构造表 8—6。此表中 X_2 表示 PGNP，X_3 表示 FLR。

表 8—6　　用于评价变量增量贡献的 ANOVA 表

变异来源	SS	df	MSS
仅由于 X_2 的 ESS	$Q_1 = \hat{\beta}_{12}^2 \sum x_2^2$	1	$\frac{Q_1}{1}$
仅由于增加的 X_3 的 ESS	$Q_2 = Q_3 - Q_1$	1	$\frac{Q_2}{1}$
X_2 和 X_3 共同的 ESS	$Q_3 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\frac{Q_3}{2}$
RSS	$Q_4 = Q_5 - Q_3$	$n-3$	$\frac{Q_4}{n-3}$
总计	$Q_5 = \sum y_i^2$	$n-1$	

为了评估在扣除 X_2 的贡献后 X_3 的增量贡献，我们构造：

$$F = \frac{Q_2/\mathrm{df}}{Q_4/\mathrm{df}} = \frac{(\mathrm{ESS}_{\mathrm{new}} - \mathrm{ESS}_{\mathrm{old}})/\text{新增回归元个数}}{\mathrm{RSS}_{\mathrm{new}}/\mathrm{df}(=n-\text{新模型中的参数个数})} \tag{8.4.16}$$

$$= \frac{Q_2/1}{Q_4/61} \quad \text{（对于本例）}$$

其中 $\mathrm{ESS}_{\mathrm{new}}$ = 新模型的 ESS（指增加新回归元后的 Q_3），$\mathrm{ESS}_{\mathrm{old}}$ = 原有模型的 ESS（$=Q_1$），和 $\mathrm{RSS}_{\mathrm{new}}$ = 新模型的 RSS（指扣除所有回归元的贡献后的 Q_4）。对于我们的说明性例子，结果如表 8—7 所示。

表 8—7　　说明性例子的 ANOVA 表：增量分析

变异来源	SS	df	MSS
仅由于 PGNP 的 ESS	60 449.5	1	60 449.5
仅由于增加的 FLR 的 ESS	196 912.9	1	196 912.9
PGNP 和 FLR 共同的 ESS	257 362.4	2	128 681.2
RSS	106 315.6	61	1 742.878 6
总计	363 678	63	

现在应用方程（8.4.16），我们得到：

$$F = \frac{196\,912.9}{1\,742.878\,6} = 112.981\,4 \tag{8.4.17}$$

在 u_i 的通常的假定下，这个 F 值服从自由度为 1 和 62 的 F 分布。读者应该能够验证，这个 F 值是高度显著的，表明模型中增加 FLR 明显提高了 ESS 并因此提高 R^2 值。因此，模型中应该增加 FLR。同样注意到，如果你将多元回归（8.1.4）中的 FLR 系数值平方，即 $(-10.629\,3)^2$，在允许存在四舍五入误差的情况下，你将得到方程（8.4.17）中的 F 值。

另外，如同我们在方程（8.4.13）中所做的那样，方程（8.4.16）的 F 比率还可仅用 R^2 值重新表达出来。如习题 8.2 所表明的那样，方程（8.4.16）的 F 比率等

价于如下 F 比率[①]：

$$F=\frac{(R^2_{\text{new}}-R^2_{\text{old}})/\text{df}}{(1-R^2_{\text{new}})/\text{df}}$$
$$=\frac{(R^2_{\text{new}}-R^2_{\text{old}})/\text{新回归元个数}}{(1-R^2_{\text{new}})/\text{df}(=n-\text{新模型中的参数个数})} \tag{8.4.18}$$

此 F 比率服从有适当分子和分母自由度（在我们的例子中分别是 1 和 61）的 F 分布。

对我们的例子而言，$R^2_{\text{new}}=0.707\ 7$［由方程（8.1.4）］和 $R^2_{\text{old}}=0.166\ 2$［由方程（8.4.14）］。因此，

$$F=\frac{(0.707\ 7-0.166\ 2)/1}{(1-0.707\ 7)/61}=113.05 \tag{8.4.19}$$

除近似计算中的误差外，和方程（8.4.17）中的 F 值差不多一样。这个 F 值高度显著，强化了变量 FLR 应当位于此模型中的早期结论。

提醒注意：你若使用方程（8.4.11）中给出的 R^2 型 F 检验，要保证新老模型中因变量是相同的。如果它们不同，则使用方程（8.4.16）中给出的 F 检验。

何时加进一个新变量。刚才描述的 F 检验程序，为决定是否增加一个变量到回归模型中来提供了一个程式化方法。研究者常常遇到从几个不相上下的模型中挑选一个模型的问题，这些模型**有同一因变量**但有不同的解释变量。作为一种见机行事的选择（因为分析的理论基础薄弱），研究者常常选择有最高 $\bar{R}^2$ 值的模型。这样一来，只要增加一个变量能增加 $\bar{R}^2$ 的值，就把它保留在模型中，即使它在统计意义上并不显著地减少 RSS。于是问题变为：$\bar{R}^2$ 值在什么情况下增加？可以证明，如果新增变量的系数的 t 值在绝对值上大于 1，$\bar{R}^2$ 就会增加。这里的 t 值，是在所指系数的总体值为零的假设下计算的［即在真实 β 值为零的假设下由方程（5.3.2）计算出来的 t 值］。[②] 上述准则又可用不同的方式叙述为：仅当一个新增解释变量的 F（$=t^2$）值大于 1 时，它的引进才使 $\bar{R}^2$ 增大。

不管用哪一准则，儿童死亡率一例中 FLR 变量的 t 值为 $-10.629\ 3$ 或 F 值为 112.981 4，均表明 $\bar{R}^2$ 将增大。的确，当 FLR 加入模型时，$\bar{R}^2$ 从 0.152 8 增加到 0.698 1。

何时加进一组变量。能不能找到类似的规则，以决定是否值得把一组变量加入到模型中来，或从模型中排除出去？从方程（8.4.18）应能看到答案：如果一组变量的加入（排除）给出一个大（小）于 1 的 F 值，$\bar{R}^2$ 将增加（减小）。当然，从方程（8.4.18）容易看出一组变量的加入（排除）是否显著地增加（减少）了回归模型的解释能力。

---

① 如下的 F 检验是 8.6 节中方程（8.6.9）或（8.6.10）所给的更为一般的 F 检验的一个特殊情形。

② 证明见 Dennis J. Aigner, *Basic Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1971, pp. 91-92。

8.5 检验两个回归系数是否相等

假如在多元回归

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \tag{8.5.1}$$

中，我们要检验假设

$$H_0: \beta_3 = \beta_4 \quad 或 \quad (\beta_3 - \beta_4) = 0$$
$$H_1: \beta_3 \neq \beta_4 \quad 或 \quad (\beta_3 - \beta_4) \neq 0 \tag{8.5.2}$$

即两个斜率系数 β_3 和 β_4 相等。

这样的虚拟假设没有实际意义。例如，令方程（8.5.1）代表对某商品的需求函数，其中 Y=某商品的需求量，X_2=该商品的价格，X_3=消费者的收入，X_4=消费者的财富。这时，虚拟假设意味着，收入系数与财富系数相等。或者，如果把 Y 和 X 表达为对数形式，则方程（8.5.2）中的虚拟假设意味着消费的收入弹性与财富弹性相同。（为什么?）

怎样检验这种虚拟假设呢？在经典假设下，可以证明：

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\text{se}(\hat{\beta}_3 - \hat{\beta}_4)} \tag{8.5.3}$$

服从自由度为 $n-4$ 的 t 分布，因为方程（8.5.1）是四变量模型或更一般地，自由度为 $n-k$ 的模型，其中 k 为包括截距项在内待估计参数的总个数。标准误 $\text{se}(\hat{\beta}_3 - \hat{\beta}_4)$ 则可从以下熟知的公式得到（详见本书附录 A）：

$$\text{se}(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{\text{var}\,(\hat{\beta}_3) + \text{var}\,(\hat{\beta}_4) - 2\text{cov}\,(\hat{\beta}_3, \hat{\beta}_4)} \tag{8.5.4}$$

如果将虚拟假设和 $\text{se}(\hat{\beta}_3 - \hat{\beta}_4)$ 的表达式代入方程（8.5.3），我们的检验统计量就变为：

$$t = \frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\text{var}\,(\hat{\beta}_3) + \text{var}\,(\hat{\beta}_4) - 2\text{cov}\,(\hat{\beta}_3, \hat{\beta}_4)}} \tag{8.5.5}$$

于是，检验方法包括如下步骤：

1. 估计 $\hat{\beta}_3$ 和 $\hat{\beta}_4$。任何标准计算机软件都能做到。

2. 大多数标准计算机软件都照常算出所估计参数的方差与协方差。[①] 从这些（方差与协方差）估计值很容易算得方程（8.5.5）分母中的标准误。

3. 从方程（8.5.5）算出 t 比率。注意本例中的虚拟假设是 $\beta_3 - \beta_4 = 0$。

4. 如果从方程（8.5.5）计算出来的 t 变量超过给定自由度下在指定显著性水平

① 协方差公式的代数表达式颇为复杂。用矩阵符号简洁地描述这个表达式，则在附录 C 中给出。

上的 t 临界值，则可拒绝虚拟假设；否则不拒绝。换言之，如得自方程（8.5.5）的 t 统计量的 p 值足够低，就可拒绝虚拟假设。注意，p 值越低，拒绝虚拟假设的证据就越强。因此，当我们说一个 p 值低或足够低，那就意味着它低于显著性水平 10%、5%或 1%。在此决策中包含个人判断。

例 8.2　立方成本函数再议

回顾 7.10 节例 7.4 所估计的立方总成本函数，为方便起见，将其重写如下：

$$\hat{Y}_i = 141.766\ 7 + 63.477\ 7X_i - 12.961\ 5X_i^2 + 0.939\ 6X_i^3$$
$$\text{se} = (6.375\ 3)\quad (4.778\ 6)\quad (0.985\ 7)\quad (0.059\ 1)$$
$$\text{cov}(\hat{\beta}_3, \hat{\beta}_4) = -0.057\ 6 \qquad R^2 = 0.998\ 3 \tag{7.10.6}$$

其中 Y 为总成本，而 X 为产出。括号中的数字代表估计的标准误。

假如我们要检验假设：立方成本函数中的 X^2 和 X^3 项的系数相等，即 $\beta_3 = \beta_4$ 或 $\beta_3 - \beta_4 = 0$。在回归（7.10.6）中，我们已具备用于方程（8.5.5）中的 t 检验所需要的全部数字结果。具体的演算步骤如下：

$$\begin{aligned} t &= \frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\text{var}(\hat{\beta}_3) + \text{var}(\hat{\beta}_4) - 2\text{cov}(\hat{\beta}_3, \hat{\beta}_4)}} \\ &= \frac{-12.961\ 5 - 0.939\ 6}{\sqrt{(0.986\ 7)^2 + (0.059\ 1)^2 - 2\times(-0.057\ 6)}} \\ &= \frac{-13.901\ 1}{1.044\ 2} = -13.313\ 0 \end{aligned} \tag{8.5.6}$$

读者可以验证，对于 6 个自由度（为什么?）即使在 0.002（或 0.2%）的显著水平上（双尾检验），所观测到的 t 值也超过了 t 临界值；p 值极小，仅为 0.000 006。从而我们拒绝假设：立方成本函数中的 X^2 和 X^3 有相同的系数。

8.6　受约束的最小二乘法：检验线性等式约束条件

经济理论有时会提出某一回归模型中的系数满足一些线性等式约束条件。例如，考虑柯布-道格拉斯生产函数：

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i} \qquad (7.9.1) = (8.6.1)$$

其中 Y=产出，X_2=劳动力投入和 X_3=资本投入。写成对数形式，方程就变为：

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \tag{8.6.2}$$

其中 $\beta_0 = \ln\beta_1$。

现在如果规模报酬不变（投入的同比例变化导致产出也同比例变化），经济理论将提出：

$$\beta_2 + \beta_3 = 1 \tag{8.6.3}$$

这就是线性等式约束之一例。[①]

怎样判知规模报酬是否不变，即约束条件（8.6.3）正确？有两种方法回答这个问题。

□ t 检验方法

最简单的程序，是先不明显考虑约束条件（8.6.3），而是按照通常的方式估计方程（8.6.2）。即做所谓**无约束或无限制回归**（unrestricted or unconstrained regression）。一旦估计了 β_2 和 β_3（比方说，用 OLS），就可通过方程（8.5.3）的 t 检验来检验假设或约束（8.6.3），即：

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{\text{se}(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{\sqrt{\text{var}(\beta_2) + \text{var}(\beta_3) + 2\text{cov}(\beta_2, \beta_3)}} \tag{8.6.4}$$

其中在虚拟假设下，$\beta_2 + \beta_3 = 1$，而分母是 $\hat{\beta}_2 + \hat{\beta}_3$ 的标准误。然后参照 8.5 节，如果从方程（8.6.4）计算的 t 值超过在选定显著性水平上的 t 临界值，则拒绝不变规模报酬假设；否则不拒绝。

□ F 检验法：受约束最小二乘法

前述 t 检验是一种静观后效法，因为我们是在估计“无约束”回归之后再分析线性约束是否得到满足。一种直接方法则是一开始便把约束条件（8.6.3）纳入估计过程中。在本例中，这种过程不难实现。由方程（8.6.3）可得：

$$\beta_2 = 1 - \beta_3 \tag{8.6.5}$$

或

$$\beta_3 = 1 - \beta_2 \tag{8.6.6}$$

因此，利用两等式之一便可消去方程（8.6.2）中的一个系数 β，然后估计所得的方程。于是，我们利用方程（8.6.5）把柯布-道格拉斯生产函数写为：

$$\begin{aligned} \ln Y_i &= \beta_0 + (1 - \beta_3) \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \\ &= \beta_0 + \ln X_{2i} + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i \end{aligned}$$

或

$$\ln Y_i - \ln X_{2i} = \beta_0 + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i \tag{8.6.7}$$

或

$$\ln (Y_i / X_{2i}) = \beta_0 + \beta_3 \ln (X_{3i} / X_{2i}) + u_i \tag{8.6.8}$$

其中 Y_i / X_{2i} = 产出/劳动力比率和 X_{3i} / X_{2i} = 资本/劳动力比率，两者都是有重大经济意义的数量。

，留意原始方程（8.6.2）经过了何种变换。一旦我们从方程（8.6.7）或

① 如果 $\beta_2 + \beta_3 < 1$，此关系式将构成线性不等式约束。为处理这种约束，需要用到数学规划技术。

(8.6.8) 估计出 β_3，β_2 就容易从关系式 (8.6.5) 算出。不言而喻，这种估计程序保证了所估计的两个投入系数之和必然等于 1。方程 (8.6.7) 或 (8.6.8) 所描述的程序被称为**受约束最小二乘** (restricted least squares，RLS)。此程序可推广到含有任意多个解释变量，以及包含多于一个线性等式约束的模型。推广方法见瑟尔 (1971)。[1] (还可参见下面的一般 F 检验法。)

怎样比较无约束和受约束的两个最小二乘回归呢？换句话说，我们怎么知道，比方说，约束 (8.6.3) 是否站得住脚呢？这个问题可通过应用如下 F 检验来回答。令：

$\sum \hat{u}_{UR}^2$ =无约束回归 (8.6.2) 的 RSS；

$\sum \hat{u}_R^2$ =受约束回归 (8.6.7) 的 RSS；

m=线性约束个数 (本例中是 1)；

k=无约束回归中的参数个数；

n=观测次 (个) 数。

于是，

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} = \frac{(\sum \hat{u}_R^2 - \sum \hat{u}_{UR}^2)/m}{\sum \hat{u}_{UR}^2/(n-k)} \tag{8.6.9}$$

服从自由度为 m 和 $n-k$ 的 F 分布。(注：UR 和 R 分别表示无约束和受约束。)

上述 F 检验还可通过 R^2 表达如下：

$$F = \frac{(R_{UR}^2 - R_R^2)/m}{(1-R_{UR}^2)/(n-k)} \tag{8.6.10}$$

其中 R_{UR}^2 和 R_R^2 分别是得自无约束和受约束回归的 R^2 值。即得自回归 (8.6.2) 和 (8.6.7) 的 R^2 值。应注意到：

$$R_{UR}^2 \geqslant R_R^2 \tag{8.6.11}$$

以及

$$\sum \hat{u}_{UR}^2 \leqslant \sum \hat{u}_R^2 \tag{8.6.12}$$

习题 8.4 要求你对这些命题做出解释。

提醒注意：在使用方程 (8.6.10) 时，要记住，如果在受约束和无约束两个模型中因变量不相同，则 R_{UR}^2 和 R_R^2 不可直接比较，此时，可用第 7 章介绍的程序把两个 R^2 值转化为可比的 (参看例 8.3)，或使用方程 (8.6.9) 给出的 F 检验。

例 8.3　1955—1974 年墨西哥经济的柯布-道格拉斯生产函数

为了说明上述讨论，我们考虑表 8—8 中给出的数据。尝试对这些数据拟合柯布-道格拉斯生

① Henri Theil, *Principles of Econometrics*, John Wiley & Sons, New York, 1971, pp. 43-45.

产函数得到如下结果：

$$\begin{aligned}\widehat{\ln \text{GDP}_t} &= -1.652\,4 + 0.339\,7 \ln \text{Labor}_t + 0.846\,0 \ln \text{Capital}_t \\ t &= (-2.725\,9) \quad (1.829\,5) \quad (9.062\,5) \\ p\text{值} &= (0.014\,4) \quad (0.084\,9) \quad (0.000\,0) \\ R^2 &= 0.995\,1 \\ \text{RSS}_{\text{UR}} &= 0.013\,6\end{aligned} \tag{8.6.13}$$

其中 RSS_{UR} 因我们在估计方程（8.6.13）时没有施加限制而成为无约束的 RSS。

表 8—8　　墨西哥的真实 GDP、就业和真实固定资本

年份	GDP（1960 年百万比索）	就业（千人）	固定资本（1960 年百万比索）
1955	114 043	8 310	182 113
1956	120 410	8 529	193 749
1957	129 187	8 738	205 192
1958	134 705	8 952	215 130
1959	139 960	9 171	225 021
1960	150 511	9 569	237 026
1961	157 897	9 527	248 897
1962	165 286	9 662	260 661
1963	178 491	10 334	275 466
1964	199 457	10 981	295 378
1965	212 323	11 746	315 715
1966	226 977	11 521	337 642
1967	241 194	11 540	363 599
1968	260 881	12 066	391 847
1969	277 498	12 297	422 382
1970	296 530	12 955	455 049
1971	306 712	13 338	484 677
1972	329 030	13 738	520 553
1973	354 057	15 924	561 531
1974	374 977	14 154	609 825

资料来源：Victor J. Elias, *Sources of Growth: A Study of Seven Latin American Economies*, International Center for Economic Growth, ICS Press, San Francisco, 1992. Data from Tables E5, E12 and E14.

我们在第 7 章已经看到如何解释柯布-道格拉斯生产函数的系数。如你所见，产出/劳动弹性约为 0.34，而产出/资本弹性约为 0.85。如果我们把这些系数相加则得到 1.19，表明考察期内墨西哥经济可能正经历着规模报酬递增的阶段。当然我们不知道 1.19 是否显著异于 1。

为看出是否如此，让我们施加规模报酬不变的约束，并给出如下回归：

$$\begin{aligned}\widehat{\ln (\text{GDP/Labor})_t} &= -0.494\,7 + 1.015\,3 \ln (\text{Capital/Labor})_t \\ t &= (-4.061\,2) \quad (28.105\,6) \\ p\text{值} &= (0.000\,7) \quad (0.000\,0) \\ R^2_{\text{R}} &= 0.977\,7 \qquad \text{RSS}_{\text{R}} = 0.016\,6\end{aligned} \tag{8.6.14}$$

因为我们已经施加了规模报酬不变的限制，其中 RSS_R 为约束 RSS。

由于上面两个回归的因变量不同，所以我们必须使用方程（8.6.9）中给出的 F 检验；我们有得到 F 值所需要的数据。

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)}$$
$$= \frac{(0.016\,6 - 0.013\,6)/1}{(0.013\,6)/(20-3)}$$
$$= 3.75$$

注意，因为我们只施加了一个约束，所以 $m=1$；而因为我们有 20 个观测，且在无约束回归中有 3 个参数，所以 $n-k=17$。

此 F 值服从分子自由度为 1 和分母自由度为 17 的 F 分布。读者很容易验证，即使在 5%的显著性水平上，这个 F 值仍不显著。（见附录 D 表 D—3。）

于是结论就是，墨西哥经济在样本期内可能仍具有规模报酬不变的特征，因此采用方程（8.6.14）中给出的约束回归没有坏处。此回归表明，若资本/劳动比提高 1%，则劳动生产率也平均上升 1%。

一般的 F 检验方法[①]

方程（8.6.10）中的 F 检验或与它等价的方程（8.6.9），为检验有关 k 变量回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{8.6.15}$$

中的一个或多个参数的假设，提供了一般方法。方程（8.4.16）中的 F 检验或方程（8.5.3）中的 t 检验不过是方程（8.6.10）的一个应用特例。例如，如同：

$$H_0: \beta_2 = \beta_3 \tag{8.6.16}$$

$$H_0: \beta_3 + \beta_4 + \beta_5 = 3 \tag{8.6.17}$$

这种涉及 k 变量模型参数的一些线性约束的假设，或者如同：

$$H_0: \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \tag{8.6.18}$$

从而意味着某些回归元在模型中并不出现的假设，都可通过方程（8.6.10）中的 F 检验方法来检验。

从 8.4 节和 8.6 节的讨论，读者一定已察知 F 检验方法的一般策略是：首先有一个较大的模型，如无约束模型（8.6.15），然后通过从中删除某些变量如（8.6.18），或通过对较大模型中的一个或多个参数加以某种线性约束如方程（8.6.16）或（8.6.17），而有一个较小的受约束或受限制模型。

然后，分别用无约束和受约束模型拟合数据，以获得判定系数 R^2_{UR} 和 R^2_R。注意，无约束模型的自由度为 $n-k$，而对受约束模型，自由度为 m，其中 m 为线性约束［如方程（8.6.16）或（8.6.17）］个数，或为从模型中省略的回归元个数。［例如，

① 如果是用极大似然法进行估计，则一个类似的即将讨论的方法是似然比检验，因它较复杂，故放在本章附录中讨论。进一步的讨论，见于 Theil, op. cit., pp. 179-184。

如果方程（8.6.18）成立，则 $m=4$，因为该方程假定 4 个回归元不在模型中出现。]然后按照方程（8.6.9）或（8.6.10）计算 F，并使用决策规则：如果计算的 F 超过 $F_\alpha(m, n-k)$，其中 $F_\alpha(m, n-k)$ 是显著性水平为 α 的 F 临界值，我们就拒绝虚拟假设；否则不拒绝。

让我们进行说明。

例 8.4　1960—1982 年美国鸡肉需求

习题 7.19 中的一问是要求读者考虑以下对鸡肉的需求函数：

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + \beta_5 \ln X_{5t} + u_i \tag{8.6.19}$$

其中 Y=每人鸡肉消费量（磅），X_2=每人实际可支配收入（美元），X_3=每磅鸡肉实际零售价格（美分），X_4=每磅猪肉实际零售价格（美分）和 X_5=每磅牛肉实际零售价格（美分）。

在此模型中，β_2、β_3、β_4 和 β_5 分别是收入弹性、自价格弹性、与猪肉的交叉价格弹性和与牛肉的交叉价格弹性。（为什么?）根据经济理论，我们预期：

$$\begin{aligned}
&\beta_2 > 0 \\
&\beta_3 < 0 \\
&\beta_4 > 0, \text{ 如果鸡肉和猪肉是替代品} \\
&\quad < 0, \text{ 如果鸡肉和猪肉是互补品} \\
&\quad = 0, \text{ 如果鸡肉和猪肉是无关产品} \\
&\beta_5 > 0, \text{ 如果鸡肉和牛肉是替代品} \\
&\quad < 0, \text{ 如果鸡肉和牛肉是互补品} \\
&\quad = 0, \text{ 如果鸡肉和牛肉是无关产品}
\end{aligned} \tag{8.6.20}$$

假如某人执意认为鸡肉与猪肉和牛肉为无关产品，即鸡肉的消费不受猪肉和牛肉价格的影响。简单地表示，即：

$$H_0: \beta_4 = \beta_5 = 0 \tag{8.6.21}$$

从而有受约束回归：

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + u_t \tag{8.6.22}$$

当然，方程（8.6.19）是无约束回归。

利用习题 7.19 中的数据，我们得到以下方程：

无约束回归：

$$\begin{aligned}
\widehat{\ln Y_t} = {} & 2.189\,8 + 0.342\,5 \ln X_{2t} - 0.504\,6 \ln X_{3t} + 0.148\,5 \ln X_{4t} + 0.091\,1 \ln X_{5t} \\
& (0.155\,7)\ (0.083\,3) \qquad (0.110\,9) \qquad (0.099\,7) \qquad (0.100\,7) \\
& R^2_{UR} = 0.982\,3
\end{aligned} \tag{8.6.23}$$

受约束回归：

$$\begin{aligned}
\widehat{\ln Y_t} = {} & 2.032\,8 + 0.451\,5 \ln X_{2t} - 0.377\,2 \ln X_{3t} \\
& (0.116\,2)\ \ (0.024\,7) \qquad (0.063\,5) \\
& R^2_R = 0.980\,1
\end{aligned} \tag{8.6.24}$$

其中括号内的数字是估计的标准误。注：因两模型有相同的因变量，故方程（8.6.23）和（8.6.24）中的两个 R^2 值是可比的。

现在检验假设（8.6.21）的 F 比率是：

$$F=\frac{(R^2_{UR}-R^2_R)/m}{(1-R^2_{UR})/(n-k)} \tag{8.6.10}$$

因为在本例中涉及两个约束：$\beta_4=0$ 和 $\beta_5=0$，故 m 值为 2，而由于 $n=23$ 和 $k=5$（5 个 β 系数），分母自由度 $n-k$ 为 18，因此 F 比率是：

$$F=\frac{(0.982\ 3-0.980\ 1)/2}{(1-0.982\ 3)/18}=1.122\ 4 \tag{8.6.25}$$

它服从自由度为 2 和 18 的 F 分布。

显然，在 5%的显著性水平上，这个 F 值不是统计显著的 [$F_{0.5}(2,\ 18)=3.55$]。p 值是 0.347 2。因此没有理由拒绝虚拟假设，即对鸡肉的需求不依赖于猪肉和牛肉价格。简言之，我们可以接受受约束回归（8.6.24）作为鸡肉需求函数的表达式。

注意，在自价格弹性为负和收入弹性为正的意义上，需求函数符合先验的经济预期。然而，估计的价格弹性的绝对值在统计上小于 1，这意味着鸡肉需求是缺乏价格弹性的。（为什么？）而且，收入弹性虽是正数，在统计上仍然小于 1。这表明鸡肉不是奢侈品；按照惯例，如果一种商品的收入弹性大于 1，它就被称作奢侈品。

8.7 检验回归模型的结构或参数稳定性：邹至庄检验

在我们使用一个涉及时间序列数据的回归时，回归子 Y 和回归元之间的关系可能会出现**结构变动**（structural change）。结构变动意味着，模型中的参数值在整个期间内不能保持相同。结构变动有时源于外部力量（如 1973 年和 1979 年 OPEC 石油卡特尔提出的石油涨价或 1990—1991 年的海湾战争），或源于政策变化（如 1973 年附近从固定汇率制向浮动汇率制的转换），或国会所采取的行动（如里根总统在其两任任期内的税收变化或最小工资率的变化），或一系列其他原因。

我们如何发现结构变动确实存在呢？具体而言，考虑表 8—9 中给出的数据。此表给出了美国 1970—1995 年个人可支配收入和个人储蓄的数据（以十亿美元计）。假如我们想估计储蓄（Y）与个人可支配收入（X）之间的简单储蓄函数。既然我们有数据，那就能得到 Y 对 X 的 OLS 回归。但如果我们那么做，那我们就认为储蓄和个人可支配收入（disposable personal income，DPI）的关系在 26 年间没有多大变化。这是一个难以置信的假定。比如，众所周知，美国 1982 年遭受了其和平时期最大的衰退。城市失业率当年达到了自 1948 年以来的最高水平 9.7%。这种事件可能会破坏储蓄和 DPI 之间的关系。为看出是否如此，我们把样本数据分为两个时期：1970—1981 年的衰退前时期和 1982—1995 年的衰退后时期。

我们现在有三个可能的回归：

时期 1970—1981：$Y_t=\lambda_1+\lambda_2X_t+u_{1t}\quad n_1=12$ (8.7.1)

时期 1982—1995：$Y_t=\gamma_1+\gamma_2X_t+u_{2t}\quad n_2=14$ (8.7.2)

时期 1970—1995：$Y_t = \alpha_1 + \alpha_2 X_t + u_t \quad n = n_1 + n_2 = 26$ (8.7.3)

回归（8.7.3）假定这两个时期之间没有区别，因此对 26 个观测构成的整个时期估计储蓄和 DPI 之间的关系。换言之，此回归假定截距和斜率系数在整个期间保持不变；即不存在结构变动。若确实如此，则 $\alpha_1 = \lambda_1 = \gamma_1$，$\alpha_2 = \lambda_2 = \gamma_2$。

表 8—9　　1970—1995 年美国储蓄和个人可支配收入　　（单位：十亿美元）

观测	储蓄	收入	观测	储蓄	收入
1970	61.0	727.1	1983	167.0	2 522.4
1971	68.6	790.2	1984	235.7	2 810.0
1972	63.6	855.3	1985	206.2	3 002.0
1973	89.6	965.0	1986	196.5	3 187.6
1974	97.6	1 054.2	1987	168.4	3 363.1
1975	104.4	1 159.2	1988	189.1	3 640.8
1976	96.4	1 273.0	1989	187.8	3 894.5
1977	92.5	1 401.4	1990	208.7	4 166.8
1978	112.6	1 580.1	1991	246.4	4 343.7
1979	130.1	1 769.5	1992	272.6	4 613.7
1980	161.8	1 973.3	1993	214.4	4 790.2
1981	199.1	2 200.2	1994	189.4	5 021.7
1982	205.5	2 347.3	1995	249.3	5 320.8

资料来源：*Economic Report of the President*，1997，Table B-28，p. 332.

回归（8.7.1）和（8.7.2）假定这两个时期的回归不同，即截距和斜率参数如带下标的参数所示都不相同。在上述回归中，u 表示误差项，n 表示观测次数。

针对表 8—9 中给出的数据，上述三个回归的经验结果如下：

$$\hat{Y}_t = 1.016\ 1 + 0.080\ 3\ X_t$$
$$t = (0.087\ 3)\ (9.601\ 5)$$
$$R^2 = 0.902\ 1 \quad RSS_1 = 1\ 785.032 \quad df = 10 \qquad (8.7.1a)$$
$$\hat{Y}_t = 153.494\ 7 + 0.014\ 8X_t$$
$$t = (4.692\ 2)\ (1.770\ 7)$$
$$R^2 = 0.297\ 1 \quad RSS_2 = 10\ 005.22 \quad df = 12 \qquad (8.7.2a)$$
$$\hat{Y}_t = 62.422\ 6 + 0.037\ 6X_t$$
$$t = (4.891\ 7)\ (8.893\ 7)$$
$$R_2 = 0.767\ 2 \quad RSS_3 = 23\ 248.30 \quad df = 24 \qquad (8.7.3a)$$

在上述回归中，RSS 表示残差平方和，括号中的数字都是估计的 t 值。

粗看之下，所估计的回归表明储蓄和 DPI 之间的关系在这两个子时期并不相同。上述储蓄—收入回归中的斜率表示**边际储蓄倾向**（marginal propensity to save，MPS），即个人可支配收入增加一美元导致储蓄的（平均）变化。在 1970—1981 年期间，MPS 约为 0.08，但在 1982—1995 年期间，MPS 约为 0.02。这种变化是否源于里根总统所追求的经济政策很难说。这进一步表明，无视两个时期的差异而将 26 个观测放在一起做一个通常的回归即**混合回归**（pooled regression）（8.7.3a）可能

不适当。当然，上述判断仍需要由适当的统计检验来支持。顺便指出，散点图和估计的回归线如图 8—3 所示。

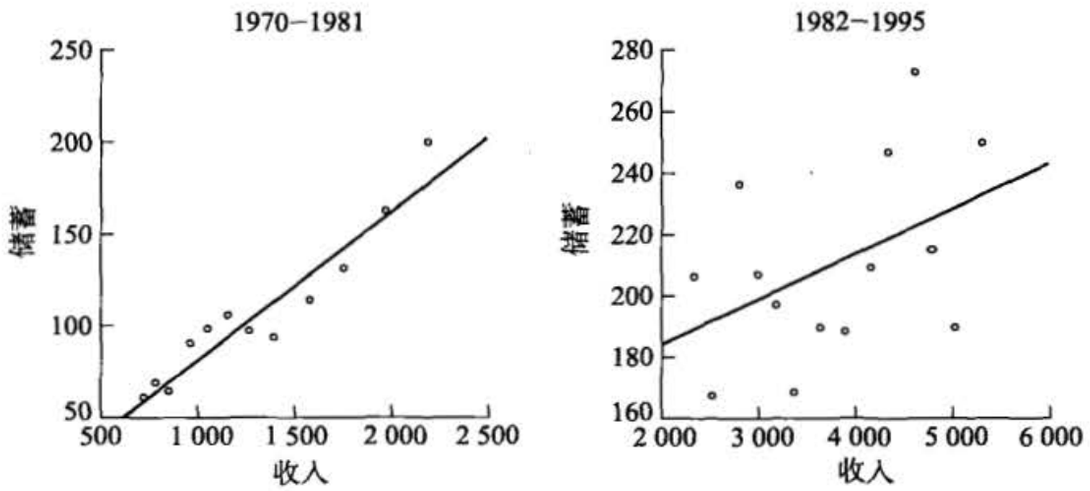


图 8—3

现在可能的区别即结构变动可能因截距或斜率或二者共同所致。我们如何找出其差别呢？从图 8—3 中可对此得到直觉，但规范的检验会更有帮助。

邹至庄检验（Chow test）适逢所需。[①] 该检验假定：

1. $u_{1t} \sim N(0, \sigma^2)$ 和 $u_{2t} \sim N(0, \sigma^2)$，也就是说，两个子期间回归的误差项是有相同方差 σ^2 的（同方差性）正态分布变量。

2. 两个误差项 u_{1t} 和 u_{2t} 是独立分布的。

邹至庄检验的机制如下：

1. 估计回归（8.7.3），若无参数不稳定性，则为适当估计，并得到 RSS_3，df $= n_1 + n_2 - k$，其中 k 为所估计的参数个数，在本例中为 2。对我们的例子而言，$RSS_3 = 23\ 248.30$，我们称 RSS_3 为**约束残差平方和**（restricted residual sum of squares，RSS_R），因为它是通过施加 $\lambda_1 = \gamma_1$ 和 $\lambda_2 = \gamma_2$（即子期间回归没有不同）的约束后得到。

2. 估计方程（8.7.1）并在 df$= n_1 - k$ 下得到其残差平方和 RSS_1。本例中 $RSS_1 = 1\ 785.032$，df$=10$。

3. 估计方程（8.7.2）并在 df$= n_2 - k$ 下得到其残差平方和 RSS_2。本例中 $RSS_2 = 10\ 005.22$，df$=12$。

4. 既然这两个样本集被视为独立，我们就能把 RSS_1 和 RSS_2 相加得到所谓**无约束残差平方和**（unrestricted residual sum of squares，RSS_{UR}），即

$$RSS_{UR} = RSS_1 + RSS_2 \qquad df = n_1 + n_2 - 2k$$

在本例中，

$$RSS_{UR} = 1\ 785.032 + 10\ 005.22 = 11\ 790.252$$

① Gregory C. Chow, "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, vol. 28, no. 3, 1960, pp. 591-605.

5. 藏在邹至庄检验背后的思想是，若不存在结构变动［即回归（8.7.1）和（8.7.2）实质相同］，则 RSS_R 和 RSS_{UR} 在统计上不应该不同。因此，若我们构造如下 F 比率

$$F = \frac{(RSS_R - RSS_{UR})/k}{(RSS_{UR})/(n_1 + n_2 - 2k)} \sim F_{[k,(n_1+n_2-2k)]} \tag{8.7.4}$$

邹至庄已经证明，在回归（8.7.1）和（8.7.2）（在统计上）相同（即没有结构变动或转折）的虚拟假设下，以上给出的 F 比率服从分子和分母自由度分别为 k 和 $n_1 + n_2 - 2k$ 的 F 分布。

6. 因此，若在应用中计算出的 F 值没有超过 F 表中在选定显著性水平（或 p 值）上的 F 临界值，则不能拒绝参数稳定（即没有结构变动）的虚拟假设。此时使用混合（约束?）回归（8.7.3）就是合理的。相反，若计算出的 F 值超过了 F 临界值，则拒绝参数稳定的假设，并断定回归（8.7.1）和（8.7.2）是不同的，此时混合回归（8.7.3）至少是没有把握的。

回到我们的例子中，我们发现

$$F = \frac{(23\ 248.30 - 11\ 790.252)/2}{11\ 790.252/22} = 10.69 \tag{8.7.5}$$

我们在 F 表中发现，自由度为 2 和 22 的 1%F 临界值为 5.72。因此，得到一个大于等于 10.69 的 F 值的概率远小于 1%；实际的 p 值只有 0.000 57。

因此，邹至庄检验看起来支持我们前面的预感，假设该检验背后的假定都满足，美国 1970—1995 年期间的储蓄—收入关系已经历了一次结构变动。稍后我们对此还有补充。

顺便一提，注意邹至庄检验可轻而易举地推广到不止一次结构变动的情形。比如，若我们相信，储蓄—收入关系在克林顿总统 1992 年 1 月入主白宫后发生了变化，那我们就会把样本分为三个时期：1970—1981 年，1982—1991 年，1992—1995 年，并进行邹至庄检验。当然，我们将有 4 个 RSS 项，每个子期间一个，混合数据一个，但检验的逻辑仍然是一样的。直至 2007 年的数据现在都可以利用，所以最后一个时期可延伸到 2007 年。

必须牢记关于邹至庄检验的一些警告：

1. 必须满足该检验背后的假定。比如，必须弄清楚回归（8.7.1）和（8.7.2）中的误差方差是否相同。这点我们稍后再谈。

2. 邹至庄检验只是告诉我们回归（8.7.1）和（8.7.2）是否有差别，并没有告诉我们差别是来自截距、斜率还是二者都有。但在讨论虚拟变量的第 9 章，我们将看到如何回答这个问题。

3. 邹至庄检验假定我们知道结构转折点。在我们的例子中，我们假定是 1982 年。但若不能确定结构变动何时发生，我们就必须使用其他方法。[①]

① 至于详尽讨论，参见 William H. Greene, *Econometric Analysis*, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 2000, pp. 293-297。

在我们结束对邹至庄检验和储蓄—收入回归的讨论之前，让我们考查邹至庄检验背后的一个假定，即两个时期的误差方差相同。因为我们不能观测到真实的误差方差，所以我们能从回归（8.7.1a）和（8.7.2a）中给出的 RSS 得到它们的估计值，即

$$\hat{\sigma}_1^2 = \frac{\text{RSS}_1}{n_1 - 2} = \frac{1\,785.032}{10} = 178.503\,2 \tag{8.7.6}$$

$$\hat{\sigma}_2^2 = \frac{\text{RSS}_2}{n_2 - 2} = \frac{10\,005.22}{14 - 2} = 833.768\,3 \tag{8.7.7}$$

注意，由于每个方程中都有两个估计参数，所以我们从观测数中减去 2 得到自由度。给定邹至庄检验背后的假定，则 $\hat{\sigma}_1^2$ 和 $\hat{\sigma}_2^2$ 为两个子期间真实方差的无偏估计量。因此，可以证明，若 $\sigma_1^2 = \sigma_2^2$（即两个子总体的方差相同，如邹至庄检验所假定的那样），则

$$\frac{(\hat{\sigma}_1^2/\sigma_1^2)}{(\hat{\sigma}_2^2/\sigma_2^2)} \sim F_{(n_1-k),(n_2-k)} \tag{8.7.8}$$

服从分子和分母自由度分别为 $n_1 - k$ 和 $n_2 - k$ 的 F 分布，在我们的例子中，由于在每个子回归中都只有两个参数，所以 $k = 2$。

当然，$\sigma_1^2 = \sigma_2^2$ 使上述 F 检验简化为计算

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \tag{8.7.9}$$

注：按惯例，我们把两个估计方差中较大的一个放在分子中。（关于 F 检验和其他概率分布的详细知识，可参见附录 A。）

在应用中，计算这个 F 值并与适当自由度的 F 临界值相比较，就能决定是否拒绝两个子总体的方差相同的虚拟假设。若虚拟假设未被拒绝，则可以使用邹至庄检验。

回到我们的储蓄—收入回归，我们得到如下结果

$$F = \frac{833.768\,3}{178.503\,2} = 4.670\,1 \tag{8.7.10}$$

在两个子总体的方差相等的虚拟假设下，此 F 值服从分子和分母自由度分别是 12 和 10 的 F 分布。（注：我们已经将较大的估计方差放在分子上。）从附录 D 中的表 D—3 看到，自由度为 12 和 10 的 5%和 1%F 临界值分别是 2.91 和 4.71。计算出来的 F 值在 5%的显著水平上是显著的，在 1%的水平上也几乎是显著的。于是，我们的结论将是两个子总体方差并不相同，因此，严格地讲，我们不应该使用邹至庄检验。

我们在这里的目的是要说明，应用研究中经常用到的邹至庄检验的机制。若两个子总体的误差方差不同，则邹至庄检验可进行修正。但这一程序超出了本书的范围。[①]

① 在异方差条件下对邹至庄检验的讨论，参见 William H. Greene, *Econometric Analysis*, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 2000, pp. 292-293, and Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar, U. K., 1994, p. 51。

我们前面提到的另外一点是，邹至庄检验对回归参数可能发生变化的时间的选择十分敏感。我们在例子中假定变化可能发生在出现衰退的 1982 年。如果我们当时假定它是罗纳德·里根开始执政的 1981 年，那我们可能会发现，计算出来的 F 值并不相同。事实上，习题 8.34 要求读者验证这一点。

如果我们不想选择结构关系可能出现转折的时点，那我们可以选用其他方法，如**递归残差检验**（recursive residual test）。在有关模型设定分析的第 13 章，我们将讨论这个问题。

8.8 用多元回归做预测

在 5.10 节我们曾说明如何用双变量回归模型作（1）均值预测（mean prediction），即预测总体回归函数（或总体回归线）上的点，以及作（2）个值预测（individual prediction），即对回归元 X 的特定值 $X=X_0$ 预测 Y 的个值。

多元回归的估计结果也可用于同样目的，并且预测程序明显地是双变量情形的一个直接推广，只不过用于估计预测值的方差或标准误的公式［与方程（5.10.2）和（5.10.6）相比］更为复杂，故宜于放到描述矩阵方法的附录 C 中处理。当然，大多数标准回归软件包可例行做到这一点，所以没有必要查找矩阵表述。附录 C 对喜欢数学的学生有好处，并给出了一个完整的示例。

*8.9 假设检验三联体：似然比、瓦尔德与拉格朗日乘数检验[①]

大体上说，在本章和前面的章节中，我们曾用 t、F 和 χ^2 检验对线性（对参数而言）回归范围内的各种假设进行了检验。但是我们一旦跳出线性回归模型的这个多少还算理想的背景，我们还需要检验关于线性或非线性回归模型的假设检验方法。

著名的三位一体：**似然比**（likelihood ratio，LR）、**瓦尔德**（Wald，W）和**拉格朗日乘数**（Lagrange multipler，LM）检验的使用能达到此目的。值得注意的是，这三种检验在渐近（即大样本）意义上都是等价的，因为每一种检验的检验统计量都服从 χ^2 分布。

虽然我们将在本章附录中讨论**似然比检验**，但出于实用方面的原因，本书将不

* 选读内容。

① 一篇易读的论文，见 A. Buse，"The Likelihood Ratio，Wald and Lagrange Multiplier Tests：An Expository Note"，*American Statistician*，vol. 36，1982，pp. 153-157。

使用这些检验，因为大多数研究者遇到的样本，不幸都是小样本或有限样本，而对于小样本，我们至今一直在使用的 F 检验已经足够了。且看戴维森（Davidson）和麦金农（MacKinnon）是怎样说的：

> 对于线性回归模型，不管它的误差是不是正态分布的，当然都不需要过问 LM、W 和 LR，因为我们不能从这些统计量中得到任何不为 F 检验所包含的信息。①

* 8.10 检验回归的函数形式：在线性与对数线性回归模型之间进行选择

线性回归模型（回归子是回归元的线性函数）或对数线性回归模型（回归子的对数是回归元的对数的线性函数）之间的选择，是经验分析中由来已久的一个问题。我们可用麦金农、怀特（White）和戴维森提出的一种检验，简称为 **MWD 检验**（MWD test），在上述两个模型之间进行选择。②

为说明这种检验，假定：

H_0：线性模型：Y 是回归元 X 的线性函数。

H_1：对数线性模型：$\ln Y$ 是回归元 X 的对数（即 X 的对数 $\ln X$）的线性函数。

其中，和平常一样，H_0 和 H_1 指虚拟假设和对立假设。

MWD 检验可分为以下几个步骤③：

步骤 1 估计线性模型并获得 Y 的估计值，且记为 Yf（即 $\hat{Y}$）。

步骤 2 估计对数线性模型并获得 $\ln Y$ 的估计值，且记为 $\ln f$（即 $\widehat{\ln Y}$）。

步骤 3 计算 $Z_1 = \ln Yf - \ln f$。

步骤 4 做 Y 对 X 和得自步骤 3 的 Z_1 的回归。如果按照通常的 t 检验，Z_1 的系数是统计显著的，就拒绝 H_0。

步骤 5 计算 $Z_2 = \ln f$ 的反对数 $- Yf$。

步骤 6 做 Y 的对数对 X 的对数和 Z_2 的回归。如果按照通常的 t 检验，Z_2 的系数是统计显著的，就拒绝 H_1。

---

① Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993, p. 456.

② J. MacKinnon, H. White, and R. Davidson, "Tests for Model Specification in the Presence of Alternative Hypothesis: Some Further Results," *Journal of Econometrics*, vol. 21, 1983, pp. 53-70. 相似的检验有 A. K. Bera and C. M. Jarque, "Model Specification Tests: A Simultaneous Approach," *Journal of Econometrics*, vol. 20, 1982, pp. 59-82。

③ William H. Greene, *ET. The Econometrics Toolkit Version 3*, Econometric Software, Bellport, New York, 1992, pp. 245-246.

MWD检验虽然看似复杂，其实这个检验的逻辑很简单，如果线性模型确实是正确的模型，步骤4的构造变量就不会是统计显著的，因为这时从线性模型估计出的Y值和从对数线性模型估计出来的Y值（为了比较而取反对数之后的）不会有什么差别。同样的评语也适用于对立假设H_1。

例8.5 玫瑰需求

参照习题7.16所给的1971年第Ⅲ季度至1975年第Ⅱ季度底特律市区对玫瑰需求的季度数据，为便于说明，我们将把对玫瑰的需求仅看作玫瑰和石竹两种价格的函数，而且暂不考虑收入变量。现在考虑以下模型：

$$\text{线性模型：} Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + u_t \tag{8.10.1}$$

$$\text{对数线性模型：} \ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + u_t \tag{8.10.2}$$

其中Y是玫瑰需求量（打），X_2是玫瑰平均批发价格（美元/打），而X_3是石竹平均批发价格（美元/打）。先验地，预期α_2和β_2为负（为什么?），而α_3和β_3为正（为什么?）。我们知道，对数线性模型中的斜率系数是弹性系数。

回归的结果如下：

$$\begin{aligned} \hat{Y}_t &= 9\,734.217\,6 - 3\,782.195\,6\,X_{2t} + 2\,815.251\,5\,X_{3t} \\ t &= (3.370\,5) \quad (-6.606\,9) \quad (2.971\,2) \\ F &= 21.84 \quad R^2 = 0.770\,96 \end{aligned} \tag{8.10.3}$$

$$\begin{aligned} \widehat{\ln Y_t} &= 9.227\,8 - 1.760\,7 \ln X_{2t} + 1.339\,8 \ln X_{3t} \\ t &= (16.234\,9)\ (-5.904\,4) \quad (2.540\,7) \\ F &= 17.50 \quad R^2 = 0.729\,2 \end{aligned} \tag{8.10.4}$$

这些结果表明，似乎线性模型和对数线性模型均对数据拟合良好：参数有预期符号且t值和R^2值均在统计上显著。

为了要根据MWD检验在两个模型之间作出选择，我们先检验真实模型是线性的假设。然后按照检验的步骤4，算得以下回归：

$$\begin{aligned} \hat{Y}_t &= 9\,727.568\,5 - 3\,783.062\,3\,X_{2t} + 2\,817.715\,7\,X_{3t} + 85.231\,9\,Z_{1t} \\ t &= (3.217\,8) \quad (-6.333\,7) \quad (2.836\,6) \quad (0.020\,7) \\ F &= 13.44 \quad R^2 = 0.770\,7 \end{aligned} \tag{8.10.5}$$

由于Z_1的系数在统计上不显著（p值是0.98），故我们不拒绝真实模型是线性的假设。

假如我们调换一下假设。假设真实模型是对数线性的。按照MWD检验的步骤6，得到以下回归结果：

$$\begin{aligned} \widehat{\ln Y_t} &= 9.148\,6 - 1.969\,9 \ln X_t + 1.589\,1 \ln X_{2t} - 0.001\,3\,Z_{2t} \\ t &= (17.082\,5)\ (-6.418\,9) \quad (3.072\,8) \quad (-1.661\,2) \\ F &= 14.17 \quad R^2 = 0.779\,8 \end{aligned} \tag{8.10.6}$$

Z_2的系数约在12%的水平上统计显著（p值是0.122 5），因此可在这一显著水平上拒绝真实模型是对数线性的假设。当然，如果我们要墨守成规地引用1%或5%的显著水平，则还不能拒绝真实模型是对数线性的假设。本例表明，在一定情况下，有可能对任一模型都不能拒绝。

要点与结论

1. 本章推广并细致地分析了最先在第 5 章中对双变量线性回归模型引进的区间估计与假设检验的思想。

2. 在一个多元回归中，检验一个偏回归系数的个别显著性（用 t 检验）和检验回归的总显著性（即 H_0：全部偏斜率系数为零或 $R^2=0$）是不相同的。

3. 特别地，在个别 t 检验的基础上发现了一个或多个偏回归系数在统计上不显著，并不意味着全部偏回归系数在统计上也是（集体地）不显著的。后一假设只能用 F 统计量加以检验。

4. F 检验是丰富多彩的。它可用于检验各种各样的假设。例如：（1）个别的回归系数是否统计显著，（2）是否全部偏斜率系数为零，（3）两个或多个系数是否统计上相等，（4）一些系数是否满足某些线性约束条件，以及（5）回归模型是否是结构稳定的。

5. 和双变量情形一样，多元回归模型可用于均值和/或个值预测的目的。

习　　题

问答题

8.1　假如你要研究某产品，比如说汽车，在某些年里的销售情况，有人建议你试用下面的模型：

$$Y_t=\beta_0+\beta_1 t$$

$$Y_t=\alpha_0+\alpha_1 t+\alpha_2 t^2$$

其中 Y_t=时间 t 的销售量，t=时间（以年计）。第一个模型假设销售量是时间的线性函数，而第二个模型把它表述为时间的二次函数。

a. 讨论这些模型的性质。

b. 你会如何在两个模型之间做出选择？

c. 在什么情况下二次模型是有用的？

d. 试图找到美国在过去 20 年里的汽车销售量数据，并看哪个模型对数据拟合的较好。

8.2　证明方程（8.4.16）中的 F 比率等于方程（8.4.18）中的 F 比率。（提示：ESS/TSS=R^2。）

8.3　证明方程（8.4.18）和（8.6.10）中的 F 检验是等价的。

8.4　证明命题（8.6.11）和（8.6.12）。

8.5　考虑柯布-道格拉斯生产函数：

$$Y=\beta_1 L^{\beta_2} K^{\beta_3} \tag{1}$$

其中 Y=产出，L=劳动力投入，K=资本投入。把方程（1）的两边同时除以 K 得到：

$$(Y/K) = \beta_1 (L/K)^{\beta_2} K^{\beta_2+\beta_3-1} \tag{2}$$

取（2）的自然对数得：

$$\ln (Y/K) = \beta_0 + \beta_2 \ln(L/K) + (\beta_2 + \beta_3 - 1)\ln K + u_i \tag{3}$$

其中 $\beta_0 = \ln\beta_1$。

a. 假如你有做回归（3）的数据，你会怎样检验规模报酬不变即 $\beta_2 + \beta_3 = 1$ 这个假设？

b. 如果有规模报酬不变情形，你会怎样解释回归（3）？

c. 用 L 而不用 K 去除方程（1），会有什么不同吗？

8.6 **当 $R^2=0$ 时的 R^2 临界值**。方程（8.4.11）给出在全部偏斜率系数同时为零（即 $R^2=0$）的假设下 F 与 R^2 的关系。正如我们能从 F 表求出在显著性水平 α 上的 F 临界值，我们能通过以下关系式求出 R^2 临界值：

$$R^2 = \frac{(k-1)F}{(k-1)F + (n-k)}$$

其中 k 是回归模型中包括截距在内的参数个数，而 F 是在显著性水平 α 上的 F 临界值。如果所测的 R^2 超过从上述公式计算出来的临界 R^2 值，就可拒绝真实 R^2 为零的假设。

证明上述公式并求出（在 $\alpha=5\%$ 处）回归（8.1.4）的 R^2 临界值。

8.7 根据 1968—1987 年年度数据得到如下回归结果：

$$\hat{Y}_t = -859.92 + 0.647\,0X_{2t} - 23.195X_{3t} \qquad R^2 = 0.977\,6 \tag{1}$$

$$\hat{Y}_t = -261.09 + 0.245\,2X_{2t} \qquad R^2 = 0.938\,8 \tag{2}$$

其中 Y=美国进口商品支出（1982 年十亿美元），X_2=个人可支配收入（1982 年十亿美元），X_3=趋势变量。判断方程（1）中 X_3 的标准误是否为 4.275 0。说明你的计算。（提示：利用 R^2、F 与 t 的关系。）

8.8 假设回归

$$\ln (Y_i/X_{2i}) = \alpha_1 + \alpha_2 \ln X_{2i} + \alpha_3 \ln X_{3i} + u_i$$

中的回归系数及其标准误均已知[①]，你如何估计以下回归模型参数及其标准误？

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

8.9 假定：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i}X_{3i} + u_i$$

其中 Y 是个人消费支出，X_2 是个人收入，X_3 是个人财富。[②]（X_{2i}，X_{3i}）被称为**交互作用项**（interaction term）。此表达式的含义是什么？你会怎样检验边际消费倾向（即 β_2）独立于消费者财富的假设？

8.10 给定如下回归结果：

$$\hat{Y}_t = 16\,899 - 2\,978.5\,X_{2t} \qquad R^2 = 0.614\,9$$

$$t = (8.515\,2)\ (-4.728\,0)$$

$$\hat{Y}_t = 9\,734.2 - 3\,782.2X_{2t} + 2\,815X_{3t} \qquad R^2 = 0.770\,6$$

$$t = (3.370\,5)\quad (-6.607\,0)\quad (2.971\,2)$$

你能求出这些结果所依据的样本容量吗？（提示：利用 R^2、F 与 t 值的关系。）

8.11 根据我们对分别以 t 和 F 检验为基础的个别检验和联合检验的讨论，以下哪些情况看来比较可能？

---

① 取自 Peter Kennedy, *A Guide to Econometrics*, the MIT Press, 3d ed., Cambridge, Mass., 1992, p. 310。

② Ibid., p. 327.

a. 拒绝基于 F 统计量的联合虚拟假设，但不拒绝基于个别 t 检验的每个独立的虚拟假设。

b. 拒绝基于 F 统计量的联合虚拟假设，且拒绝基于 t 检验的一个个别假设，而不拒绝基于 t 检验的其他个别假设。

c. 拒绝基于 F 统计量的联合虚拟假设，并且拒绝基于个别 t 检验的每个独立的虚拟假设。

d. 不拒绝基于 F 统计量的联合虚拟假设，并且不拒绝基于个别 t 检验的每个独立的虚拟假设。

e. 不拒绝基于 F 统计量的联合虚拟假设，拒绝基于 t 检验的一个个别假设，而不拒绝基于 t 检验的其他个别假设。

f. 不拒绝基于 F 统计量的联合虚拟假设，但拒绝基于个别 t 检验的每个独立的虚拟假设。[①]

实证分析题

8.12 参照习题 7.21。

a. 真实现金需求的实际收入弹性和利率弹性是什么？

b. 上述弹性是个别统计显著的吗？

c. 检验所估计回归的总显著性。

d. 对真实现金需求的收入弹性显著地异于 1 吗？

e. 应该把利率变量留在模型中吗？为什么？

8.13 根据美国 1992 年 46 个州的数据，巴尔塔泽（Baltagi）得到如下回归结果[②]：

$$\widehat{\ln C} = 4.30 - 1.34 \log P + 0.17 \log Y$$
$$\text{se} = (0.91)\ (0.32) \qquad (0.20) \qquad \bar{R}^2 = 0.27$$

其中C=香烟消费（每年的包数计）；

P=每包香烟的真实价格；

Y=真实人均可支配收入。

a. 香烟需求的价格弹性是多少？它统计显著吗？若显著，它在统计上异于 1 吗？

b. 香烟需求的收入弹性是多少？它显著吗？若不显著，其原因是什么？

c. 你如何根据上面给出的**调整** R^2 来得到 R^2？

8.14 伍德里奇从 209 个企业的样本得到如下回归结果[③]：

$$\widehat{\log(\text{salary})} = 4.32 + 0.280 \log(\text{sales}) + 0.017\,4\ \text{roe} + 0.000\,24\ \text{ros}$$
$$\text{se} = (0.32) \quad (0.035) \qquad (0.004\,1) \qquad (0.000\,54) \qquad R^2 = 0.283$$

其中 salary=CEO 薪水；

sales=企业年销售额；

roe=股权百分比收益；

ros=企业股票回报。

log 表示自然对数。括号中的数字为估计的标准误。

a. 根据你对各个系数符号的先验预期，解释上述回归。

b. 哪个系数在 5%的显著性水平上是个别统计显著的？

---

① 取自 Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley, Reading, Mass., 1991, p. 79。

② 参见 Badi H. Baltagi, *Econometrics*, Springer-Verlag, New York, 1998, p. 111。

③ 参见 Jeffrey M. Wooldridge, *Introductory Econometrics*, South-Western Publishing Co., 2000, pp. 154-155。

c. 回归的总显著性如何？你使用何种检验方式？为什么？

d. 你能把 roe 和 ros 的系数解释成弹性系数吗？为什么？

8.15 假定 Y 和 X_2，X_3，…，X_k 是联合正态分布的，并假定虚拟假设为各个总体偏相关系数等于零，费希尔（R. A. Fisher）曾证明：

$$t = \frac{r_{12.34\cdots k}\sqrt{n-k-2}}{\sqrt{1-r^2_{12.34\cdots k}}}$$

服从 $n-k-2$ 个自由度的 t 分布，其中 k 指第 k 阶偏相关系数，而 n 是观测值的总个数。（注：$r_{12.3}$ 是一阶偏相关系数，$r_{12.34}$ 是二阶偏相关系数，如此类推。）参照习题 7.2。假定 Y，X_2 和 X_3 是联合正态分布的，计算三个偏相关 $r_{12.3}$，$r_{13.2}$，和 $r_{23.1}$，并在相应的总体相关系数为零的假设下检验它们的显著性。

8.16 在研究 1921—1941 年和 1948—1957 年两时期美国对农用拖拉机的需求中，格里利切斯（Griliches）[①] 得到如下结果：

$$\widehat{\log Y_t} = C - 0.519 \log X_{2t} - 4.933 \log X_{3t} \qquad R^2 = 0.793$$
$$(0.231) \qquad (0.477)$$

其中 Y_t = 每年 1 月 1 日农场拥有拖拉机存量的价值，以 1935—1939 年的美元价值度量；X_2 = 时间 $t-1$ 的拖拉机支付价格指数除以全部农作物收取价格指数；X_3 = 第 $t-1$ 年的利率；C = 常数。log 表示自然对数。而括号中的数字是估计的标准误。

a. 解释上述回归。

b. 所估计的斜率系数个别地看在统计上显著吗？它们显著地异于 1 吗？

c. 使用方差分析方法检验整个回归的显著性。提示：利用 ANOVA 方法的 R^2 形式。

d. 如何计算农用拖拉机需求的利率弹性？

e. 如何检验所估计的 R^2 的显著性？

8.17 考虑 1950—1969 年间英国经济[②]的如下工资决定方程：

$$\hat{W}_t = 8.582 + 0.364(\mathrm{PF})_t + 0.004(\mathrm{PF})_{t-1} - 2.560U_t$$
$$(1.129)\ (0.080) \qquad (0.072) \qquad (0.658)$$
$$R^2 = 0.873 \quad \mathrm{df} = 15$$

其中 W = 平均每个雇员的工资和薪水；

PF = 最终产品的要素成本价格；

U = 表示英国失业人数占雇员总人数的百分比；

t = 时间。

（括号内的数字是估计的标准误。）

a. 解释上述方程。

b. 所估计的系数个别地看在统计上显著吗？

c. 引进 $(\mathrm{PF})_{t-1}$ 的合理性何在？

d. 是否应把变量 $(\mathrm{PF})_{t-1}$ 从模型中删去？为什么？

---

① Z. Griliches, "The Demand for a Durable Input: Farm Tractors in the United States, 1921—1957," in *The Demand for Durable Goods*, Arnold C. Harberger (ed.), The University of Chicago Press, Chicago, 1960, Table 1, p. 192.

② 取自 *Prices and Earnings in* 1951—1969: *An Econometric Assessment*, Dept. of Employment, HMSO, 1971, Eq. (19), p. 35。

e. 怎样计算雇员的工资和薪水对失业率的弹性？

8.18 习题 8.17 所给的工资决定方程的一个变型有如下式[①]：

$$\hat{W}_t = 1.073 + 5.288V_t - 0.116X_t + 0.054M_t + 0.046M_{t-1}$$
$$(0.797)\ (0.812)\quad (0.111)\quad (0.022)\quad (0.019)$$
$$R^2 = 0.934 \quad df = 14$$

其中 W ＝平均每个雇员的工资和薪水；

V ＝表示英国岗位空缺占英国雇员总人数的百分比；

X ＝平均每个就业人员的国内生产总值；

M ＝进口价格；

M_{t-1} ＝上（或滞后）年的进口价格。

（括号内的数字是估计的标准误。）

a. 解释上述方程。

b. 哪些估计系数是个别统计显著的？

c. 引进 X 变量的合理性何在？在先验预期上，X 的预期符号应为负吗？

d. 在模型中同时引进 M_t 和 M_{t-1} 用意何在？

e. 哪些变量可从模型中删去？

f. 检验所观测回归的总显著性。

8.19 对于方程（8.6.24）中估计的鸡肉需求函数：所估计的收入弹性等于 1 吗？价格弹性等于 −1 吗？

8.20 对于需求函数（8.6.24）你怎样检验收入弹性与价格弹性数值相同而符号相反的假设？说明必要的计算。[注：$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.001\ 42$。]

8.21 参照习题 7.16 的玫瑰需求函数。仅考虑对数设定形式。

a. 所估计的需求自价格弹性（即对玫瑰价格的弹性）是什么？

b. 它是统计显著的吗？

c. 如果是，它是否在统计上异于 1？

d. 先验预期上，X_3（石竹价格）和 X_4（收入）的预期符号是什么？经验结果和这些预期相符吗？

e. 如果 X_3 和 X_4 的系数在统计意义上不显著，可能是什么原因？

8.22 参照关于野猫活动的习题 7.17。

a. 所估计的每个斜率系数在 5%的显著水平上都是个别统计显著的吗？

b. 你会拒绝假设 $R^2 = 0$ 吗？

c. 1948—1978 年期间野猫活动的瞬时增长率是什么？相应的复合增长率呢？

8.23 参照习题 7.18 所估计的美国国防预算支出回归。

a. 对所估计的回归结果作一般性评论。

b. 构造 ANOVA 表并检验全部斜率系数为零的假设。

8.24 下面列出的所谓**超越生产函数**（transcendental production function，TPF），是著名的柯布-道格拉斯生产函数的一个推广：

$$Y_i = \beta_1 L^{\beta_2} K^{\beta_3} e^{\beta_4 L + \beta_5 K}$$

其中 Y＝产出，L＝劳动力投入，K＝资本投入。

取对数并加入随机干扰项便得到随机的 TPF：

---

① Ibid，Eq.(67)，p.37.

$$\ln Y_i = \beta_0 + \beta_2 \ln L_i + \beta_3 \ln K_i + \beta_4 L_i + \beta_5 K_i + u_i$$

其中 $\beta_0 = \ln \beta_1$。

a. 此函数具有什么性质?

b. 要使 TPF 化为柯布-道格拉斯生产函数，β_4 和 β_5 的值必须是什么?

c. 如果你拥有数据，你会怎样判明 TPF 是否可简化为柯布-道格拉斯生产函数? 你会用什么检验方法?

d. TPF 对表 8—8 中的数据拟合得怎样? 说明你的计算。

8.25 **1948—1978 年美国能源价格与资本形成。**为了检验假设：相对于产出的能源价格上升现有资本与劳动力资源的生产力下降，塔托姆（John A. Tatom）估计了 1948 年第Ⅰ季度至 1978 年第Ⅱ季度时期的美国生产函数。①

$$\widehat{\ln(y/k)} = \underset{(16.33)}{1.549\ 2} + \underset{(21.69)}{0.713\ 5}\ln(h/k) - \underset{(-6.42)}{0.108\ 1}\ln(P_e/P) + \underset{(15.86)}{0.004\ 5}\,t$$

$$R^2 = 0.98$$

其中 y =私有企业部门的真实产出；

k =资本服务流量的一种度量；

h =私有企业部门的工时（人员小时）；

P_e=燃料及相关产品的生产者价格；

P =私有企业部门价格缩减因子；

t =时间。

括号中的数字是 t 统计量。

a. 这些结果是否支持了作者的假设?

b. 在 1972—1977 年间，能源相对价格 P_e/P 增加了 60%。按照估算的回归，生产力损失了多少?

c. 除去 h/k 和 P_e/P 的变化后，在样本期间里生产力的趋势增长率如何?

d. 你会怎样解释系数 0.713 5?

e. 每个偏斜率系数估计值都个别统计显著的这一事实（为什么?），意味着我们可以拒绝假设 $R^2=0$ 吗? 为什么?

8.26 **电缆需求。**表 8—10 给出一个电缆制造商用来预测 1968—1983 年间对一主要用户的销售量数据。②

表 8—10 **回归变量**

年份	X_2 GNP（十亿美元）	X_3 新房动工数（千套）	X_4 失业率（%）	X_5 滞后 6 个月的最惠利率	X_6 用户用线增量（%）	Y 年销售量（MPF）
1968	1 051.8	1 503.6	3.6	5.8	5.9	5 873
1969	1 078.8	1 486.7	3.5	6.7	4.5	7 852
1970	1 075.3	1 434.8	5.0	8.4	4.2	8 189

① 见 "Energy Prices and Capital Formation: 1972－1977," *Review*, Federal Reserve Bank of St. Louis, vol. 61, no. 5, May 1979, p. 4。

② 感谢丹尼尔·J·里尔登（Daniel J. Reardon）收集并加工了这些数据。

续前表

年份	X_2 GNP（十亿美元）	X_3 新房动工数（千套）	X_4 失业率（%）	X_5 滞后6个月的最惠利率	X_6 用户用线增量（%）	Y 年销售量（MPF）
1971	1 107.5	2 035.6	6.0	6.2	4.2	7 497
1972	1 171.1	2 360.8	5.6	5.4	4.9	8 534
1973	1 235.0	2 043.9	4.9	5.9	5.0	8 688
1974	1 217.8	1 331.9	5.6	9.4	4.1	7 270
1975	1 202.3	1 160.0	8.5	9.4	3.4	5 020
1976	1 271.0	1 535.0	7.7	7.2	4.2	6 035
1977	1 332.7	1 961.8	7.0	6.6	4.5	7 425
1978	1 399.2	2 009.3	6.0	7.6	3.9	9 400
1979	1 431.6	1 721.9	6.0	10.6	4.4	9 350
1980	1 480.7	1 298.0	7.2	14.9	3.9	6 540
1981	1 510.3	1 100.0	7.6	16.6	3.1	7 675
1982	1 492.2	1 039.0	9.2	17.5	0.6	7 419
1983	1 535.4	1 200.0	8.8	16.0	1.5	7 923

表中变量定义如下：

Y =年销售量，百万英尺双线（million paired feet，MPF）；

X_2=国民生产总值（GNP），十亿美元；

X_3=新房动工数，千套；

X_4=失业率，%；

X_5=滞后6个月的最惠利率；

X_6=用户用线增量，%。

考虑以下模型：

$$Y_i = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + u_t$$

a. 估计以上回归。

b. 此模型中各系数的预期符号是什么？

c. 经验结果与先验预期一致吗？

d. 这些估计的偏回归系数个别地看在5%的显著水平上是统计显著的吗？

e. 假如你先做 Y 对 X_2、X_3 和 X_4 的回归，然后决定是否再加进变量 X_5 和 X_6。你如何知道值不值得把 X_5 和 X_6 加进来呢？你用哪一种检验？说明必要的计算。

8.27 纳洛夫（M. Nerlove）曾估计如下的电力产生的成本函数[①]：

$$Y = AX^{\beta}P_1^{\alpha_1}P_2^{\alpha_2}P_3^{\alpha_3}u \qquad (1)$$

其中Y=总生产成本；

X=千瓦小时产出；

① Marc Nerlove, "Returns to Scale in Electric Supply," in Carl Christ, ed., *Measurement in Economics*, Stanford University Press, Palo Alto, Calif., 1963. 符号有所变化。

P_1＝劳动力投入价格；

P_2＝资本投入价格；

P_3＝燃料价格；

u＝干扰项。

理论上，预期价格弹性之和为 1，即 $\alpha_1 + \alpha_2 + \alpha_3 = 1$。引进这一约束，上述成本函数就可写为：

$$(Y/P_3) = AX^{\beta}(P_1/P_3)^{\alpha_1}(P_2/P_3)^{\alpha_2}u \tag{2}$$

换言之，(1) 是无约束成本函数，而 (2) 是受约束成本函数。

根据 29 个中等厂家的一个样本并通过对数变换，纳洛夫得到如下回归结果：

$$\widehat{\ln Y_i} = -4.93 + 0.94 \ln X_i + 0.31 \ln P_1 - 0.26 \ln P_2 + 0.44 \ln P_3$$

$$\text{se} = (1.96) \quad (0.11) \qquad (0.23) \qquad (0.29) \qquad (0.07)$$

$$\text{RSS} = 0.336 \tag{3}$$

$$\widehat{\ln (Y/P_3)} = -6.55 + 0.91 \ln X + 0.51 \ln(P_1/P_3) + 0.09 \ln (P_2/P_3)$$

$$\text{se} = (0.16)(0.11) \qquad (0.19) \qquad (0.16)$$

$$\text{RSS} = 0.364 \tag{4}$$

a. 解释方程 (3) 和 (4)。

b. 你怎样判断约束 $\alpha_1 + \alpha_2 + \alpha_3 = 1$ 是否正确？说明你的计算。

8.28 **估计资本资产定价模型 (CAPM)**。在 6.1 节中我们简要地叙述了现代证券组合理论中著名的资本资产定价模型。在经验分析中，CAPM 的估计分为两阶段：

阶段Ⅰ (时间序列回归)。对样本所含 N 种证券的每一种，做如下的一个时间回归：

$$R_{it} = \hat{\alpha}_i + \hat{\beta}_i R_{mt} + e_{it} \tag{1}$$

其中 R_{it} 和 R_{mt} 是年度 t 第 i 种证券和市场证券组合（比如，S&P 500）的回报率；β_i，如本书其他地方讲过的，是第 i 种证券的 β 或市场波动系数，而 e_{it} 是干扰项。一共有 N 个这种回归（每一种证券有一个），从而给出 β_i 的 N 个估计值。

阶段Ⅱ (横截面回归)。在这一阶段，我们在 N 种证券上做以下回归：

$$\bar{R}_i = \hat{\gamma}_1 + \hat{\gamma}_2 \hat{\beta}_i + u_i \tag{2}$$

其中 $\bar{R}_i$ 是在阶段Ⅰ所覆盖的样本时期内算出的第 i 种证券的平均或均值回报率，$\hat{\beta}_i$ 是从阶段Ⅰ回归估计出来的系数，而 u_i 是干扰项。

将阶段Ⅱ的回归 (2) 和 CAPM 方程 (6.1.2) 相比，写成：

$$\text{ER}_i = r_f + \beta_i(\text{ER}_m - r_f) \tag{3}$$

其中 r_f 代表无风险回报率，我们即看到 $\hat{\gamma}_1$ 是 r_f 的一个估计值，$\hat{\gamma}_2$ 是市场风险溢价 $\text{ER}_m - r_f$ 的一个估计值。

因此，在 CAPM 的经验检验中，$\bar{R}_i$ 和 $\hat{\beta}_i$ 被分别地用作 ER_i 和 β_i 的估计量。现在，如果 CAPM 成立，则在统计意义上：

$$\hat{\gamma}_1 = r_f$$

$$\hat{\gamma}_2 = R_m - r_f, \quad \text{ER}_m - r_f \text{ 的估计量}$$

其次，考虑另一可选择模型：

$$\bar{R}_i = \hat{\gamma}_1 + \hat{\gamma}_2 \hat{\beta}_i + \hat{\gamma}_3 s_{e_i}^2 + u_i \tag{4}$$

其中 $s_{e_i}^2$ 是得自阶段Ⅰ回归的第 i 种证券的残差方差。那么，如果 CAPM 正确，$\hat{\gamma}_3$ 就不会显著地异于零。

为了检验 CAPM，利维 (Levy) 根据 1948—1968 年 101 种股票的一个样本，做回归 (2) 和

(4)，并得到如下结果[①]：

$$\hat{R}_i = 0.109 + 0.037\beta_i$$
$$(0.009)\ (0.008) \qquad (2)'$$
$$t = (12.0) \quad (5.1) \quad R^2 = 0.21$$

$$\hat{R}_i = 0.106 + 0.002\,4\hat{\beta}_i + 0.201\,s_{e_i}^2$$
$$(0.008)\ (0.007) \quad (0.038) \qquad (4)'$$
$$t = (13.2) \quad (3.3) \quad (5.3) \quad R^2 = 0.39$$

a. 这些结果支持了 CAPM 吗？

b. 是否值得把变量 $s_{e_i}^2$ 加进模型中来？你怎样知道？

c. 如果 CAPM 成立，方程 (2)′中的 $\hat{\gamma}_1$ 应接近无风险利率 r_f 的均值。这个估计值是 10.9%。此值像不像是观测期间（1948—1968 年）无风险回报率的一个合理估计呢？（不妨考虑国债或类似的较无风险的资产的回报率。）

d. 如果 CAPM 成立，由方程 (2)′得到的市场风险溢价（$\bar{R}_m - r_f$）将是 3.7%。如果假定 r_f 为 10.9%，则意味着在样本期间 $\bar{R}_m$ 约为 14.6%。这像不像是一个合理的估计呢？

e. 你对 CAPM 能做什么一般性评论吗？

8.29 参照习题 7.21c。现在，你有了必要的工具，你将用哪种检验来在两个模型之间做出选择。给出必要的计算。注意，两个模型的因变量不同。

8.30 参照习题 8.3。利用方程（8.6.4）中所给出的 t 检验，说明墨西哥经济在研究期内是否存在规模报酬不变的情况。

8.31 回到我们曾几次讨论的儿童死亡率一例。在回归 (7.6.2) 中，我们将儿童死亡率 (CM) 对人均 GNP（即 PGNP）和妇女识字率（FLR）回归。现在我们通过增加总生育率 (TFR) 来扩展模型。表 6—4 中包含所有这些数据。重做回归 (7.6.2) 并给出扩展模型的回归结果如下：

1. $$\widehat{CM}_i = 263.641\,6 - 0.005\,6\,PGNP_i - 2.231\,6\,FLR_i$$
$$se = (11.593\,2) \quad (0.001\,9) \quad (0.209\,9) \quad R^2 = 0.707\,7 \qquad (7.6.2)$$

2. $$\widehat{CM}_i = 168.306\,7 - 0.005\,5\,PGNP_i - 1.768\,0\,FLR_i + 12.868\,6\,TFR_i$$
$$se = (32.891\,6) \quad (0.001\,8) \quad (0.248\,0) \quad (?) \quad R^2 = 0.747\,4$$

a. 你如何解释 TFR 的系数？据经验，预期 CM 和 TFR 之间的关系是正还是负？给出你回答的理由。

b. 在这两个方程之间 PGNP 和 FLR 的系数有变化吗？若有，变化的原因是什么？所观测的差别是统计显著的吗？你使用哪个检验，为什么？

c. 你如何在模型 1 和 2 之间做出选择？你用哪个检验来回答这个问题？给出必要的计算。

d. 我们没有给出 TFR 系数的标准误。你能求出它吗？（提示：回忆 t 和 F 分布之间的关系。）

8.32 回到习题 1.7，它给出 21 个企业在广告印象和广告支出方面的数据。习题 5.11 要你对这些数据描点，并决定二者关系的适当模型。令 Y 表示保留的印象数，X 为广告支出，则得到如下回归：

① H. Levy, "Equilibrium in an Imperfect Market: A Constraint on the Number of Securities in the Portfolio," *American Economic Review*, vol. 68, no. 4, September 1978, pp. 643-658.

模型Ⅰ： $\hat{Y}_i = 22.163 + 0.363\,1\,X_i$

se = (7.089) (0.097 1)　　$r^2 = 0.424$

模型Ⅱ： $\hat{Y}_i = 7.059 + 1.084\,7\,X_i - 0.004\,0\,X_i^2$

se = (9.986) (0.369 9)　(0.001 9)　　$R^2 = 0.53$

a. 解释这两个模型。

b. 哪个模型更好？为什么？

c. 你用哪个或哪些检验来选择模型？

d. 广告支出存在“收益递减”吗，即在一定的广告支出水平（饱和水平）后就不再支出广告费吗？你能求出这个支出水平吗？给出必要的计算。

8.33　在回归（7.9.4）中，我们给出用柯布-道格拉斯生产函数来拟合2005年美国50个州和华盛顿特区制造业部门的结果。基于此回归，说明是否存在规模报酬不变的情况。

a. 使用方程（8.6.4）中给出的 t 检验来说明。并告诉你两个斜率估计量之间的协方差为 −0.038 43。

b. 用方程（8.6.9）中给出的 F 检验来说明。

c. 这两种检验结果是否不同？对于50个州和华盛顿特区制造业部门在样本期内的规模报酬，你有什么结论？

8.34　重新考虑8.7节中的储蓄—收入回归。假设我们把样本分为1970—1982年和1983—1995年两个时期，利用邹至庄检验判断储蓄—收入回归在两个时期是否有结构变动。将你的结论与8.7节中给出的结论相比较，对邹至庄检验对样本分成两（或多）期转折点的敏感性，你有什么总体结论？

8.35　参考习题7.24和表7.12中美国1947—2000年间四个经济变量的数据。

a. 基于消费支出对真实收入、真实财富和真实利率的回归，看哪些回归系数在5%的显著水平上是个别统计显著的。估计系数的符号与经济理论一致吗？

b. 基于（a）中的结论，你如何估计收入弹性、财富弹性和利率弹性？你是否还需要哪些额外信息来计算这些弹性？

c. 你如何检验收入弹性与财富弹性相同的假设？给出必要的计算。

d. 假设不再使用（a）中估计的线性消费函数，你把消费支出的对数对收入的对数、财富的对数和利率进行回归。给出回归结果。你如何解释这些结果？

e. （d）中估计的收入弹性和财富弹性是多少？你如何解释（d）中估计的利率系数？

f. 在（d）的回归中，你能使用利率的对数而不是利率本身吗？为什么？

g. 你如何比较（b）和（d）中估计的弹性？

h. 在（a）和（d）估计的回归模型之间，你更喜欢哪一个？为什么？

i. 假设不是估计（d）中给出的模型，你只是将消费支出的对数对收入的对数进行回归。你如何确定是否值得在模型中增加财富的对数？你又如何确定是否值得在模型中同时增加财富的对数和利率？给出必要的计算。

8.36　参考8.8节的内容和表8—9中有关1970—1995年间个人可支配收入和个人储蓄的数据。在该节中，我们引入邹至庄检验来分析数据在两个时期之间是否发生了结构变化。表8—11给出了1970—2005年间的更新数据。根据美国国民经济研究局（National Bureau of Economic Research）的看法，美国最近一次的经济衰退期在2001年底结束。把这些数据分成三个时期：（1）1970—1981年，（2）1982—2001年，（3）2002—2005年。

表 8—11　　1970—2005 年美国储蓄和个人可支配收入　　（单位：十亿美元）

观测	储蓄	收入	观测	储蓄	收入
1970	69.5	735.7	1988	272.9	3 748.7
1971	80.6	801.8	1989	287.1	4 021.7
1972	77.2	869.1	1990	299.4	4 285.8
1973	102.7	978.3	1991	324.2	4 464.3
1974	113.6	1 071.6	1992	366.0	4 751.4
1975	125.6	1 187.4	1993	284.0	4 911.9
1976	122.3	1 302.5	1994	249.5	5 151.8
1977	125.3	1 435.7	1995	250.9	5 408.2
1978	142.5	1 608.3	1996	228.4	5 688.5
1979	159.1	1 793.5	1997	218.3	5 988.8
1980	201.4	2 009.0	1998	276.8	6 395.9
1981	244.3	2 246.1	1999	158.6	6 695.0
1982	270.8	2 421.2	2000	168.5	7 194.0
1983	233.6	2 608.4	2001	132.3	7 486.8
1984	314.8	2 912.0	2002	184.7	7 830.1
1985	280.0	3 109.3	2003	174.9	8 162.5
1986	268.4	3 285.1	2004	174.3	8 681.6
1987	241.4	3 458.3	2005	34.8	9 036.1

资料来源：Department of Commerce, Bureau of Economic Analysis.

a. 估计整个数据集（1970—2005 年）和第三个时期（2002 年以后）的模型。利用邹至庄检验，判断第三个时期与整个数据期之间是否存在明显的变化？

b. 利用表 8—11 中的新数据，判断第一个时期（1970—1981 年）与整个数据集之间在有更多观测可以使用的情况下是否仍存在显著差异。

c. 对第二个时期（1982—2001 年）与整个数据集进行邹至庄检验，看这个时期与其余数据之间是否有明显不同。

* 附录 8A

似然比检验

在附录 4A 中我们讨论极大似然（ML）原理时，说明了怎样获得双变量回归模型的 ML 估计量；似然比（LR）检验是以此 ML 原理为根据的。该原理可直接推广应用到多元回归模型中。在干扰项 u_i 为正态分布的假定下，我们证明对双变量模型言，回归系数的 OLS 和 ML 估计量是相同的。但所估计的误差方差却不相同。σ^2 的 OLS 估计量是 $\sum a_i^2/(n-2)$，而 ML 估计量是 $\sum a_i^2/n$。前者是无偏的，但后者有偏误，尽管这种偏误在大样本中将消失。

* 选读内容。

这些对多元回归而言也是正确的，为便于说明，且考虑三变量回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{1}$$

对应于附录4A中的方程（5），模型（1）的对数似然函数可写成：

$$\ln \text{LF} = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i})^2 \tag{2}$$

如附录4A所示，将此函数对 β_1，β_2，β_3 和 σ^2 微分并令其结果表达式为零，解方程组即得到这些参数的ML估计量；β_1，β_2 和 β_3 的ML估计量将无异于OLS估计量［后者已见于方程（7.4.6）至（7.4.8）］，而误差方差则有所不同，因为残差平方和将被除以 n，而不是OLS情形中的（$n-3$）。

现假定虚拟假设 H_0 是：变量 X_3 的系数 β_3 为零。这时，由方程（2）给出的对数似然函数变为：

$$\ln \text{LF} = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_{2i})^2 \tag{3}$$

由于方程（3）是在先验约束 $\beta_3 = 0$ 下估计的，故称**受约束对数似然函数**（restricted log-likelihood function，RLLF），而方程（1）则无参数方面的先验约束，故可称**无约束对数似然函数**（unrestricted log LF，ULLF）。为了检验先验约束 β_3 为零的真实性，LR检验使用如下的检验统计量：

$$\lambda = 2(\text{ULLF} - \text{RLLF}) \tag{4}$$①

其中ULLF和RLLF分别是无约束对数似然函数［方程（2）］和受约束对数似然函数［方程（3）］。可以证明，在大样本中，由方程（4）给出的检验统计量 λ 服从自由度等于虚拟假设中所加约束个数的 χ^2 分布。本例中此个数为1。

LR检验的基本思想是简单的：如果先验约束真实，则受约束与无约束（对数）LF不应有差异。这时方程（4）中的 λ 将是零。但如果先验约束不真实，则两个LF必定有差异。而我们知道，在大样本中 λ 服从 χ^2 分布，于是能找出这个差异在（比方说）1%或5%显著水平上是否统计显著的。此外，我们还能找出 λ 估计值的 p 值。

现在让我们用儿童死亡率的例子来说明LR检验。如果我们像在方程（8.1.4）中那样将儿童死亡率对人均GNP和妇女识字率回归，我们就得到ULLF为－328.101 2，但如果我们只将CM对PGNP回归，则得到RLLF为－361.639 6。从绝对值看（即不考虑符号），前者较小，由于我们在前面一个模型中增加一个变量，所以这也讲得通。

现在的问题是，是否值得增加FLR变量。若不值得，则约束和无约束LLF就不应该有大的差别，但若值得，二者就会有所差别。为了看出这个差别在统计上是否显著，我们现在利用（4）式给出的LR检验：

$$\lambda = 2 \times [-328.101\ 2 - (-361.639\ 6)] = 67.076\ 8$$

此值在渐近意义上服从1个自由度（因从完整模型中去掉变量FLR从而只有1个约束）的 χ^2 分布。得到这样一个 χ^2 值的 p 值几乎为0，从而得到变量FLR不应该从模型中去掉的结论。换言之，约束回归在目前的情况下不能成立。

令RRSS和URSS表示约束残差平方和和无约束残差平方和，方程（4）也可以表示成

$$-2\ln\lambda = n\ (\ln \text{RRSS} - \ln \text{URSS}) \tag{5}$$

它服从自由度为 r 的 χ^2 分布，其中 r 表示对模型施加的约束个数（即从原模型中去掉的系数个数 r）。

尽管我们不会深入探讨W检验和LM检验，但这些检验还是可以实施如下：

① 此式又可表达为－2(RLLF－ULLF) 或－2 ln(RLF/ULF)。

$$\text{瓦尔德统计量（W）} = \frac{(n-k)(\text{RRSS}-\text{URSS})}{\text{URSS}} \sim \chi_r^2 \tag{6}$$

$$\text{拉格朗日乘数统计量（LM）} = \frac{(n-k+r)(\text{RRSS}-\text{URSS})}{\text{RRSS}} \sim \chi_r^2 \tag{7}$$

其中 k 表示无约束模型中回归元个数，而 r 表示约束个数。

你由上述方程可以看到，所有这三个检验都是渐近（即在大样本中）等价的，也就是说，它们给出类似的答案。不过，在小样本中，它们的答案可能有所不同。这些统计量之间一个有意思的关系是，可以证明：

$$W \geqslant LR \geqslant LM$$

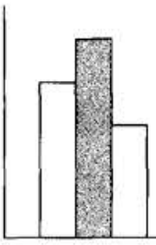
因此，在小样本中，一个假设可能被 W 统计量拒绝但不被 LM 统计量拒绝。①

书中曾指出，对我们的大多数目的而言，t 检验和 F 检验就足够了。但由于上述讨论的三个检验可用于检验线性模型的非线性假设，或用于检验对方差—协方差矩阵的约束，因此它们具有更广泛的适用性。它们还可用于误差正态分布的假定站不住脚的情形。

由于 W 检验和 LM 检验的数学复杂性，我们在此不予深究。但像刚刚指出的那样，渐近地看，LR 检验、W 检验和 LM 检验给出一致的结果，所以对检验方法的选择完全取决于计算上的便利性。

① 对此的解释，参见 G. S. Maddala, *Introduction to Econometrics*, 3d ed., John Wiley & Sons, New York, 2001, p. 177。

第 9 章



虚拟变量回归模型

我们在第 1 章简单讨论了经验分析中通常会遇到的四种变量类型：**比率尺度**、**区间尺度**、**序数尺度**和**名义尺度**。我们在前面几章曾遇到的变量类型基本上都是比率尺度。但这不应该给我们留下回归模型只能处理比率尺度变量的印象。回归模型也可以处理前面提到的其他几种数据类型。我们在本章不仅要考虑涉及比率尺度变量的模型而且要考虑涉及名义尺度变量的模型。这种变量也被称为**指标变量**（indicator variables）、**分类变量**（categorical variables）、**定性变量**（qualitative variables）或**虚拟变量**（dummy variables）。[①]

9.1 虚拟变量的性质

在回归分析中，因变量或回归子不仅经常受到比率尺度变量（如收入、产出、价格、成本、身高、温度）的影响，还会受到定性变量或名义尺度变量的影响，如性别、种族、肤色、宗教、国籍、地区、政治动乱和党派等。例如，保持所有其他因素不变，我们发现女性工人比相应男性工人挣得少，而白人比非白人挣得多。[②] 这种情况可能是性别或种族歧视所致，但不论如何，诸如性别和种族之类的定性变量看来都能影响回归子，而且明显应该包含在解释变量或回归元之中。

① 我们将在第 15 章讨论序数尺度变量。

② 对这方面证据的一个综述，可参见 Bruce E. Kaufman and Julie L. Hotchkiss, *The Economics of Labor Markets*, 5th ed., Dryden Press, New York, 2000。

由于这种变量通常都标志着出现或不出现某种“品质”或属性，如男性或女性、黑人或白人、天主教或非天主教、民主党或共和党等，所以它们基本上都是名义尺度变量。我们能量化这种属性的途径之一，就是构造一个取值 1 或 0 的人为变量，1 表示出现（或具备）那种属性，0 表示没有那种属性。比如，1 可能标志着一个人是女性，而 0 则标志着男性；或者 1 标志着一个人是大学毕业生，而 0 标志着不是，等等。假定这种取值 0 和 1 的变量被称为**虚拟变量**（dummy variables）。[①] 这种变量实质上就是一个将数据区分为相互排斥类别（如男性或女性）的工具。

虚拟变量也可以像定量变量那样轻而易举地放到回归模型中。事实上，一个回归模型所包含的回归元可以都是虚拟或定性变量。这种模型被称为**方差分析**（analysis of variance，ANOVA）模型。[②]

9.2 ANOVA 模型

为说明 ANOVA 模型，考虑如下例子。

例 9.1 不同地区公立学校教师的薪水

表 9—1 给出了 2005—2006 学年度 50 个州和哥伦比亚特区公立学校教师的平均薪水（美元）数据。这 51 个地区被分为三个地理区域：(1) 东北和中北部（共 21 个地区）；(2) 南部（共 17 个地区）；及 (3) 西部（共 13 个地区）。目前，暂不考虑表的格式及其中的其他数据。

假设我们想知道，公立学校和教师平均年薪（AAS）在这个国家的三个地区之间是否有所不同。如果你仅对这三个地区中教师的平均年薪进行简单的算术平均，那么这三个地区的平均值分别是 49 538.71 美元（东北和中北部），46 293.59 美元（南部）和 48 104.62 美元（西部）。这些数字看起来不同，但它们在统计上也彼此不同吗？有各种统计方法来比较两个和多个均值，通常是进行**方差分析**（analysis of variance）。[③] 但在回归分析的框架下也能做到这一点。

为看出这一点，考虑如下模型

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \tag{9.2.1}$$

其中 Y_i = 第 i 个州公立学校教师的平均薪水；

D_{2i} = 1，若该州位于东北和中北部；

① 虚拟变量取值 0 和 1 绝非必须。通过一个诸如 $Z=a+bD$（其中 $b\neq 0$）之类的线性函数，就可以把数对 (0，1) 转换成任意一个其他的数对，这里 a 和 b 都是常数，而且 $D=1$ 或 0。当 $D=1$ 时，$Z=a+b$，而当 $D=0$ 时，$Z=a$，于是数对 (0，1) 就变成了数对 $(a, a+b)$。比如 $a=1$ 和 $b=2$，虚拟变量就变成了 (1，3)。这个表达式表明，定性或虚拟变量不具有一个自然度量尺度。这就是把它们描述成名义尺度变量的原因。

② ANOVA 模型被用于评价定量回归子和定性或虚拟回归元之关系的统计显著性。它们通常可以用于比较两组或多组（或类别）均值的差别，因此比只用于比较两组均值的 t 检验更一般。

③ 至于应用方面的讨论，可参见 John Fox，*Applied Regression Analysis，Linear Models，and Related Methods*，Sage Publications，1997，Chapter 8。

=0，若该州位于美国其他地区；

D_{3i}=1，若该州位于南部；

=0，若该州位于美国其他地区。

应看到，除了不是定量回归元而是定性或虚拟回归元（若观测值属于某特定组则取值为1，若它不属于那一组则取值0）之外，方程（9.2.1）与前面考虑的任何一个多元回归模型都是一样的。此后，我们所有的虚拟变量都用字母 D 表示。表9—1中的虚拟变量就是这样构造的。

表9—1　　2005—2006年公立学校教师的平均薪水

	薪水	支出	D_2	D_3		薪水	支出	D_2	D_3
康涅狄格	60 822	12 436	1	0	佐治亚	49 905	8 534	0	1
伊利诺伊	58 246	9 275	1	0	肯塔基	43 646	8 300	0	1
印第安纳	47 831	8 935	1	0	路易斯安那	42 816	8 519	0	1
艾奥瓦	43 130	7 807	1	0	马里兰	56 927	9 771	0	1
堪萨斯	43 334	8 373	1	0	密西西比	40 182	7 215	0	1
缅因	41 596	11 285	1	0	北卡罗来纳	46 410	7 675	0	1
马萨诸塞	58 624	12 596	1	0	俄克拉何马	42 379	6 944	0	1
密歇根	54 895	9 880	1	0	南卡罗来纳	44 133	8 377	0	1
明尼苏达	49 634	9 675	1	0	田纳西	43 816	6 979	0	1
密苏里	41 839	7 840	1	0	得克萨斯	44 897	7 547	0	1
内布拉斯加	42 044	7 900	1	0	弗吉尼亚	44 727	9 275	0	1
新罕布什尔	46 527	10 206	1	0	西弗吉尼亚	40 531	9 886	0	1
新泽西	59 920	13 781	1	0	阿拉斯加	54 658	10 171	0	0
纽约	58 537	13 551	1	0	亚利桑那	45 941	5 585	0	0
北达科他	38 822	7 807	1	0	加利福尼亚	63 640	8 486	0	0
俄亥俄	51 937	10 034	1	0	科罗拉多	45 833	8 861	0	0
宾夕法尼亚	54 970	10 711	1	0	夏威夷	51 922	9 879	0	0
罗得岛	55 956	11 089	1	0	爱达荷	42 798	7 042	0	0
南达科他	35 378	7 911	1	0	蒙大拿	41 225	8 361	0	0
佛蒙特	48 370	12 475	1	0	内华达	45 342	6 755	0	0
威斯康星	47 901	9 965	1	0	新墨西哥	42 780	8 622	0	0
阿拉巴马	43 389	7 706	0	1	俄勒冈	50 911	8 649	0	0
阿肯色	44 245	8 402	0	1	犹他	40 566	5 347	0	0
特拉华	54 680	12 036	0	1	华盛顿特区	47 882	7 958	0	0
哥伦比亚特区	59 000	15 508	0	1	怀俄明	50 692	11 596	0	0
佛罗里达	45 308	7 762	0	1					

注：D_2=1表示该州位于美国东北或中北部；=0表示其他地区。

D_3=1表示该州位于美国南部；=0表示其他地区。

资料来源：National Educational Association，as reported in 2007.

模型（9.2.1）告诉了我们什么呢？假定误差项满足通常的OLS假定，通过对方程（9.2.1）的两边同时取期望，我们得到：

东北和中北部公立学校教师薪水的均值为：

$$E(Y_i \mid D_{2i}=1,\ D_{3i}=0)=\beta_1+\beta_2 \tag{9.2.2}$$

南部公立学校教师薪水的均值为：

$$E(Y_i \mid D_{2i}=0,\ D_{3i}=1)=\beta_1+\beta_3 \tag{9.2.3}$$

你可能想知道，我们如何求西部教师薪水的均值。若你猜测它等于 β_1，那就完全正确，因为

$$E(Y_i \mid D_{2i}=0,\ D_{3i}=0)=\beta_1 \tag{9.2.4}$$

换句话说，西部公立学校教师薪水的均值由多元回归（9.2.1）中的截距 β_1 给出，而“斜率”系数 β_2 和 β_3 则告诉我们，东北、中北部和南部教师薪水的均值与西部相比有多大的差别。但我们如何才能知道这种差别在统计上是否显著呢？在回答这个问题之前，让我们先给出回归（9.2.1）的结果。利用表 9—1 中给出的数据，我们得到如下结果：

$$\begin{aligned}
&\hat{Y}_i = 48\,014.615 + 1\,524.099\,D_{2i} - 1\,721.027\,D_{3i} \\
&\text{se}=(1\,857.024)\quad (2\,363.139)\quad (2\,467.151) \\
&t=(25.853)\quad (0.645)\quad (-0.698) \\
&\quad (0.000\,0)^*\quad (0.522\,0)^*\quad (0.488\,8)^* \qquad R^2=0.044\,0
\end{aligned} \tag{9.2.5}$$

其中 * 表示 p 值。

如这些回归结果所示，西部教师薪水的均值约为 48 015 美元，而东北和中北部教师薪水的均值则约高 1 524 美元，南部教师薪水均值则约低 1 721 美元。如方程（9.2.3）和（9.2.4）所示，将西部教师薪水加上这些地区之间的薪水差异，很容易就能得到后两个地区实际的薪水均值。我们如此可以得到后两个地区的薪水均值约为 49 539 美元和 46 294 美元。

但我们如何才能知道，这些薪水均值是否与参照组西部地区教师薪水的均值在统计上有所差异？这相当容易，我们需要做的只是辨别方程（9.2.5）中的每一个斜率系数是否统计显著。从这个回归中可以看出，东北和中北部的估计系数在统计上是不显著的，因为其 p 值为 52%，而南部地区的估计系数也不是统计显著的，因为其 p 值约为 49%。因此总体结论便是，西部与东北和中北部以及南部公立学校教师薪水的均值大致相同。从图上看，这种情形如图 9—1 所示。

在解释这些差异时，我们必须给予警告。这些虚拟变量只是简单地指出了可能存在的这些差异，但不能给出导致这些差异的原因。受教育水平、生活成本指数、性别和种族上的差异都可能对所观测到的差异具有某种影响。因此，除非我们能考虑所有其他可能影响教师薪水的变量，否则我们就不能确定导致这些差异的原因。

从上述讨论明显可见，我们所要做的只是看附属于各个虚拟变量的系数是不是个别统计显著的。此例还表明，在回归模型中包含定性或虚拟回归元是多么容易。

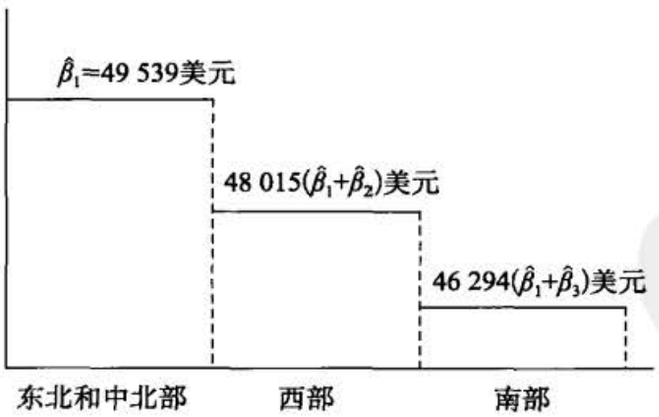


图 9—1　三个地区公立学校教师的平均薪水（以美元计）

□ 对使用虚拟变量的告诫

尽管在回归模型中包含虚拟变量很容易，但在使用它们时仍必须小心。具体而言，需考虑如下方面：

1. 我们在例 9.1 中为区分三个区域而只使用了两个虚拟变量 D_2 和 D_3。我们为什么不用三个虚拟变量来区分这三个区域呢？假设我们这样做并将模型（9.2.1）写成：

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \tag{9.2.6}$$

其中 D_{1i}对西部的观测取值 1，对其他观测取值 0。于是我们现在对每个地理区域都有了一个虚拟变量。利用表 9—1 中的数据，如果你做回归（9.2.6），计算机将“拒绝”做这个回归（不妨试试看）。[①] 为什么呢？原因在于，在方程（9.2.6）的背景下，你既有每个类别或组的虚拟变量，又有一个截距，这样你就遇到了**完全共线性**（perfect collinearity）的情况，即变量之间存在完全线性关系。为什么？回到表 9—1，设想我们现在增加了 D_1 列：如果该州处于西部就取值 1，否则取值 0。现在，如果你将这三个 D 列水平相加就得到由 51 个 1 构成的一列。但由于截距 α 对每个观测都（隐含地）为 1，所以你又得到一个由 51 个 1 构成的一列。换言之，这三个 D 列之和再次生成了截距列，由此导致了完全共线性。在这种情况下，估计模型（9.2.6）是不可能的。

这里的信息是：**若定性变量有 m 个类别，则只需引入 $m-1$ 个虚拟变量**。在我们的例子中，定性变量“区域”有三类，所以我们只需引入两个虚拟变量。如果你不遵守这个规则，那你就会陷入所谓的**虚拟变量陷阱**（dummy variable trap），即完全共线性或完全多重共线性（若变量之间存在不止一个精确的关系）情形。在模型中有不止一个定性变量时这个规则依然适用，稍后会给出一个例子。因此我们应该将前述规则重新表述为：**对每个定性回归元而言，所引入的虚拟变量个数必须比该变量的类别数少一个。**因此，如果我们在例 9.1 中有教师性别的信息，那我们就应该再使用一个（而非两个）虚拟变量，对女性取值 1 和对男性取值 0 或相反。

2. 不指定其虚拟变量的那一组被称为**基（base）组、基准（benchmark）组、控制（control）组、比较（comparison）组、参照（reference）组或省略（omitted）组**。所有其他的组都与基准组进行比较。

3. 截距值（β_1）代表了基准组的均值。在例 9.1 中基准组为西部地区。因此，在回归（9.2.5）中约为 48 015 的截距值代表着西部各州教师薪水的均值。

4. 附属于方程（9.2.1）中虚拟变量的系数被称为**级差截距系数**（differential intercept coefficient），因为它们告诉我们，取值为 1 的地区的截距值与基准组的截距系数之间的差别。比如在方程（9.2.5）中，约为 1 524 的系数值意味着，与作为基准组的西部地区的薪水均值 48 015 美元相比，东北和中北部教师的薪水均值约高 1 524 美元。

5. 如果像在我们的说明性例子中那样定性变量不止一类，那么，基准组的选择完全取决于研究者。基准组的选择有时候受所研究的特殊问题的支配。在我们的说明性例子中，我们可以选择南部作为基准组。在那种情况下，方程（9.2.5）中的回

① 实际上你将得到数据矩阵退化的提示信息。

归结果将有所变化，因为现在都是与南部做比较。当然，这不会改变此例中的总体结论（为什么?）。此时，截距值将是南部教师薪水的均值，约为 46 294 美元。

6. 我们前面对虚拟变量陷阱做过警告。如果我们在这种模型中不使用截距项，那么引入与变量的类别相同数量的虚拟变量就能够回避虚拟变量陷阱的问题。因此，如果我们从方程（9.2.6）中去掉截距项，并考虑如下模型

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \tag{9.2.7}$$

由于此时没有完全共线性，所以就不会陷入虚拟变量陷阱。**但要确定当你做这个回归时，一定要使用回归软件包中的无截距选项。**

我们如何解释回归（9.2.7）呢？如果你对方程（9.2.7）求期望，你将会发现：

β_1 = 西部教师薪水的均值

β_2 = 东北和中北部教师薪水的均值

β_3 = 南部教师薪水的均值

换言之，**去掉截距项并容许每一类别都有一个虚拟变量，我们就直接得到不同群组的均值。**在我们说明性的例子中，方程（9.2.7）的结果如下：

$$\hat{Y}_i = 48\,014.62\,D_{1i} + 49\,538.71\,D_{2i} + 46\,293.59\,D_{3i}$$

$$\begin{aligned} \text{se} &= (1\,857.204) \quad (1\,461.240) \quad (1\,624.077) \\ t &= (25.853)^* \quad (33.902)^* \quad (28.505)^* \end{aligned} \tag{9.2.8}$$

$$R^2 = 0.044$$

其中 * 表示这些 t 比率的 p 值很小。

如你所见，虚拟变量的系数直接给出了西部、东北和中北部及南部三个地区的（薪水）均值。

7. 如下引入虚拟变量的方法中哪种更好呢：（1）为每个类别都引入一个虚拟变量并省略截距项；或（2）引入截距项和 $m-1$ 个虚拟变量，其中 m 为虚拟变量的类别数？如肯尼迪（Kennedy）所指出：

> 大多数研究者认为，在一个含有截距的方程中，他们能更容易地处理他们通常最感兴趣的问题：是否有某个组与基准组有所不同以及有多大的不同，所以在方程中包括截距更方便。为了检查分组是否得当，也可通过将虚拟变量的系数相对 0 做 t 检验（或者更一般地，对适当的虚拟变量系数集做一个 F 检验），就可以检验分类是否适当。*①

9.3 含有两个定性变量的 ANOVA 模型

我们在上一节考虑了含有一个三类别定性变量的 ANOVA 模型。我们在本节

* 因为分组的人可能预料在基准组与其他组之间存在统计上的显著差异。——译者注

① Peter Kennedy, *A Guide to Econometrics*, 4th ed., MIT Press, Cambridge, Mass., 1998, p. 223.

将考虑含有两个定性变量的 ANOVA 模型，并揭示虚拟变量的某些其他特点。

例 9.2　小时工资与婚姻状况和居住地的关系

从 1985 年 5 月的一个 528 人的样本中得到如下回归结论[①]：

$$\begin{aligned} \hat{Y}_i &= 8.814\,8 + 1.099\,7\,D_{2i} - 1.672\,9D_{3i} \\ \text{se} &= (0.401\,5) \quad (0.464\,2) \quad (0.485\,4) \\ t &= (21.952\,8) \quad (2.368\,8) \quad (-3.446\,2) \\ &\quad (0.000\,0)^* \quad (0.018\,2)^* \quad (0.000\,6)^* \\ R^2 &= 0.032\,2 \end{aligned} \tag{9.3.1}$$

其中Y ＝小时工资，美元；

D_2＝婚姻状况：1＝已婚，0＝其他；

D_3＝居住地：1＝南部，0＝其他。

回归中 * 表示 p 值。

在此例中，我们有两个定性回归元，且每个回归元分为两个类别。因此，我们为每个定性回归元都指定一个虚拟变量。

这里的基准组是什么呢？显然是未婚的非南部居民组。换句话说，不住在南部的未婚人士属于被省略组。于是所有的组都与这个组进行比较。此基准组的小时工资均值约为 8.81 美元。与其相比，已婚者的平均小时工资约高 1.10 美元，即实际的平均工资为 9.91 美元(＝8.81＋1.10)。相比之下，那些住在南部的人的平均小时工资约低 1.67 美元，实际小时工资为 7.14 美元。

上述小时工资与基组相比在统计上有差异吗？是的，因为所有的级差截距都是统计显著的，且 p 值都相当小。

在此例中值得注意的一点是：一旦遇到多于一个定性变量，你就必须密切注意基组，因为所有其他组都是与基组进行比较。在有几个定性回归元，而且每个回归元又有几个类别时这一点就特别重要。但现在怎样操作几个定性变量应该清楚了。

9.4　同时含有定性和定量回归元的回归：ANCOVA 模型

前面两节讨论的那种 ANOVA 模型，尽管在社会学、心理学、教育学和市场研究等领域很常见，但在经济学中并不普遍。在多数典型的经济研究中，回归模型的解释变量既有一些定性的，又有一些定量的。同时包含定性和定量变量的回归模型被称为**协方差分析**（analysis of covariance，ANCOVA）模型。ANCOVA 模型是对 ANOVA 模型的推广，在一个同时包括定量和定性（或虚拟）回归元的模型中，这

① 数据得自下书的数据盘：Arthur S. Goldberger，*Introductory Econometrics*，Harvard University Press，Cambridge，Mass.，1998。我们已经在第 2 章考虑过这些数据。

种模型提供了一种方法，能在统计上控制定量回归元［被称为**协变量**（covariates）或**控制变量**（control variables）］的影响。我们现在就来说明 ANCOVA 模型。

例 9.3　教师薪水与地区及对公立学校生均拨款的关系

为了说明为什么要进行这种分析，让我们重新考虑例 9.1，并设想三个地区公立学校教师的平均薪水可能本来就没有什么不同，因为我们应考虑到有些变量无法在不同地区使之标准化，比如考虑地方政府对公立学校的支出这种变量（因为公共教育基本上是地方和州政府的事情）。为了看出是否如此，我们提出如下模型：

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i \tag{9.4.1}$$

其中Y_i＝公立学校教师的州平均薪水，美元；

X_i＝对公立学校每个学生的支出，美元；

D_{2i} ＝1，若该州位于东北和中北部；

　　＝0，其他；

D_{3i} ＝1，若该州位于南部；

　　＝0，其他。

X 的数据在表 9—1 中给出。记住我们把西部作为基准组。除两个定性回归元之外，我们还有一个定量变量 X，前面曾提到，在 ANCOVA 模型的背景下，X 被称为**协变量**。

从表 9—1 中的数据得到模型（9.4.1）的结论如下：

$$\begin{aligned}
\hat{Y}_i &= 28\,694.918 - 2\,954.127 D_{2i} - 3\,112.194 D_{3i} + 2.340\,4 X_i \\
\text{se} &= (3\,262.521) \quad (1\,862.576) \quad (1\,819.873) \quad (0.359\,2) \\
t &= (8.795)^{*} \quad (-1.586)^{**} \quad (-1.710)^{**} \quad (6.515)^{*} \\
&\qquad\qquad\qquad\qquad\qquad R^2 = 0.497\,7
\end{aligned} \tag{9.4.2}$$

其中 * 表示 p 值低于 5%，** 表示 p 值高于 5%。

如这些结论所示，在其他条件不变的情况下：公共教育支出每增加 1 美元，公立学校教师的薪水约上升 2.34 美元。控制教育支出后，我们现在看到，东北和中北部以及南部的级差截距系数都不显著。这些结论与方程（9.2.5）中的结论不同。但这无足为奇，因为我们在方程（9.2.5）中没有包含对每个学生的公共教育支出这个协变量，如图 9—2 所示。

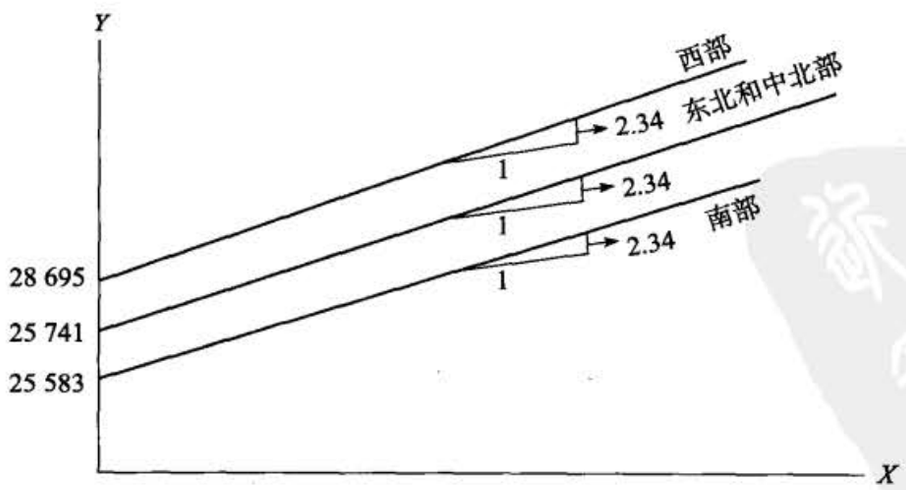


图 9—2　公立学校教师薪水（Y）与对每个学生教育支出（X）的关系

注意，尽管我们对三个地区给出三个回归线，但从统计上看，西部和南部的回归线是一样的。还可以注意到，这三个回归线是平行的。(为什么?)

9.5 邹至庄检验的虚拟变量方法[①]

我们在 8.8 节讨论过邹至庄检验，以考察一个回归模型的结构稳定性。我们在那里讨论的例子涉及美国 1970—1995 年间储蓄—收入之间的关系。我们将样本期间一分为二，1970—1981 年和 1982—1995 年，邹至庄检验表明，储蓄对收入的回归在这两个区间存在着差异。

然而，我们不知道这两个回归的差异是源于截距项、斜率系数还是二者兼而有之。这种知识本身常常是十分有用的。

参照方程 (8.7.1) 和 (8.7.2)，我们看到四种可能性，如图 9—3 所示。

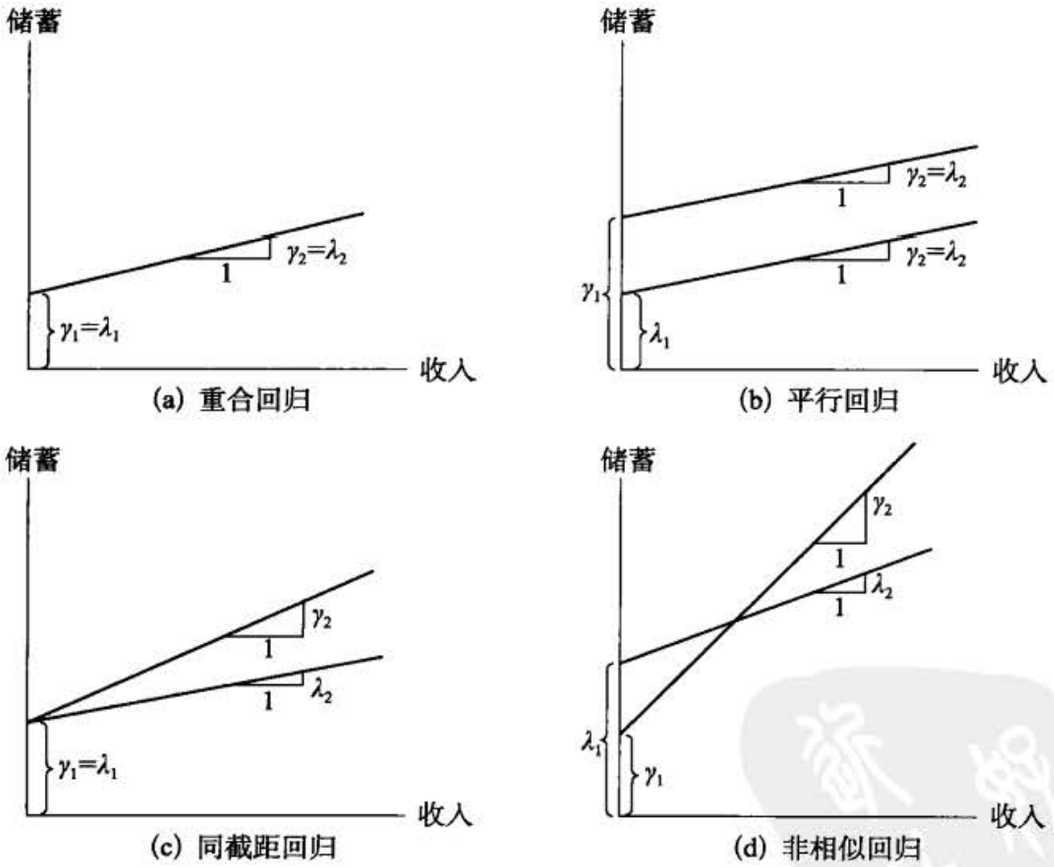


图 9—3 合理的储蓄—收入回归

① 本节内容取自作者的如下论文："Use of Dummy Variables in Testing for Equality between Sets of Coefficients in Two Linear Regressions: A Note," 和 "Use of Dummy Variables··· A Generalization," 均发表在 *American Statistician*, vol. 24, nos. 1 and 5, 1970, pp. 50-52 and 18-21。

1. 两个回归的截距和斜率都相同。这种**重合回归**（coincident regressions）的情形如图 9—3a 所示。

2. 两个回归的斜率相同但截距不同。这种**平行回归**（parallel regressions）的情形如图 9—3b 所示。

3. 两个回归的截距相同但斜率不同。这种**同截距回归**（concurrent regressions）的情形如图 9—3c 所示。

4. 两个回归的截距和斜率都不同。这种**非相似回归**（dissimilar regressions）的情形如图 9—3d 所示。

前面提到，8.7 节中所讨论的多步骤邹至庄检验程序只告诉我们两个（或多个）回归是否不同，但没有告诉我们这种不同来自哪里。通过将所有观测（共 26 个）混合起来，如果存在着差异，只需做如下多元回归便能探明这种差异的根源[①]：

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \tag{9.5.1}$$

其中Y=储蓄；

X=收入；

t=时间；

D =1，1982—1995 年之间的观测；

=0，其他（即 1970—1981 年之间的观测）。

表 9—2 说明了数据矩阵的结构。

表 9—2　　1970—1995 年间美国的储蓄与收入数据

观测	储蓄	收入	虚拟变量	观测	储蓄	收入	虚拟变量
1970	61	727.1	0	1983	167	2 522.4	1
1971	68.6	790.2	0	1984	235.7	2 810	1
1972	63.6	855.3	0	1985	206.2	3 002	1
1973	89.6	965	0	1986	196.5	3 187.6	1
1974	97.6	1 054.2	0	1987	168.4	3 363.1	1
1975	104.4	1 159.2	0	1988	189.1	3 640.8	1
1976	96.4	1 273	0	1989	187.8	3 894.5	1
1977	92.5	1 401.4	0	1990	208.7	4 166.8	1
1978	112.6	1 580.1	0	1991	246.4	4 343.7	1
1979	130.1	1 769.5	0	1992	272.6	4 613.7	1
1980	161.8	1 973.3	0	1993	214.4	4 790.2	1
1981	199.1	2 200.2	0	1994	189.4	5 021.7	1
1982	205.5	2 347.3	1	1995	249.3	5 320.8	1

注：虚拟变量取值 1 表示始于 1982 年的观测；取值 0 表示其他观测。收入和储蓄数据都以十亿美元计。

资料来源：*Economic Report of the President*，1997，Table B-28，p. 332.

为了看出方程（9.5.1）的含义，像通常一样假定 $E(u_i)=0$，我们得到：

1970—1981 年的平均储蓄函数：

① 如在邹至庄检验中一样，混合法假定同方差性，即 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 。

$$E(Y_t \mid D_t = 0, X_t) = \alpha_1 + \beta_1 X_t \tag{9.5.2}$$

1982—1995 年的平均储蓄函数：

$$E(Y_t \mid D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t \tag{9.5.3}$$

读者将会注意到，它们是与方程（8.7.1）和（8.7.2）相同的函数，其中 $\lambda_1 = \alpha_1$，$\lambda_2 = \beta_1$，$\gamma_1 = \alpha_1 + \alpha_2$ 和 $\gamma_2 = \beta_1 + \beta_2$。因此，估计方程（9.5.1）就等同于逐一估计两个储蓄函数（8.7.1）和（8.7.2）。

在方程（9.5.1）中，和前面一样，α_2 是**级差截距**，而**级差斜率系数**也被称为**斜率漂移因子**（slope drifter），β_2 表示的是第二个期间储蓄函数（虚拟变量取值为 1 的那一组）的斜率与第一个期间相比有多大的差异。注意，以交互或**相乘形式**（interactive or multiplicative form）引入虚拟变量 D（即 D 乘以 X），如何使我们能区别两个期间的斜率系数，与我们以**相加形式**（additive form）引入虚拟变量来区分两个期间的截距殊途同归。

例 9.4　美国储蓄—收入回归中的结构差异：虚拟变量方法

在继续讨论之前，让我们先给出应用美国储蓄—收入数据估计模型（9.5.1）得到的回归结果：

$$\begin{aligned} \hat{Y}_t &= 1.016\,1 \quad + \quad 152.478\,6\,D_t + 0.080\,3X_t - 0.065\,5(D_tX_t) \\ \text{se} &= (20.164\,8) \quad (33.082\,4) \quad (0.014\,4) \quad (0.015\,9) \\ t &= (0.050\,4)^{**} \quad (4.609\,0)^{*} \quad (5.541\,3)^{*} \quad (-4.096\,3)^{*} \\ & \qquad\qquad\qquad\qquad\qquad R^2 = 0.881\,9 \end{aligned} \tag{9.5.4}$$

其中，* 表示 p 值低于 5%，而 ** 表示 p 值高于 5%。

如这些回归结果所示，级差截距和斜率系数都是统计显著的，这强烈地表明，两个期间的储蓄—收入回归是如图 9—3d 那样的。

从方程（9.5.4）我们可以推导出方程（9.5.2）和（9.5.3）：

1970—1981 年的储蓄—收入函数：

$$\hat{Y}_t = 1.016\,1 + 0.080\,3\,X_t \tag{9.5.5}$$

1982—1995 年的储蓄—收入函数：

$$\begin{aligned} \hat{Y}_t &= (1.016\,1 + 152.478\,6) + (0.080\,3 - 0.065\,5)\,X_t \\ &= 153.494\,7 + 0.014\,8\,X_t \end{aligned} \tag{9.5.6}$$

它们正是我们在方程（8.7.1a）和（8.7.2a）中所得到的结果，这无足为奇。这些回归已如图 8—3 所示。

现在很容易看出，虚拟变量方法［即估计方程（9.5.1）］相对邹至庄检验［即估计（8.7.1）、（8.7.2）和（8.7.3）三个回归］具有如下优势：

1. 我们只需要做一个回归，因为单个回归很容易就能以方程（9.5.2）和（9.5.3）所示的方式推导出来。

2. 单一回归（9.5.1）可用于检验各种假设。因此，如果级差截距系数 α_2 是统计非显著的，那我们或许可以接受这两个回归具有相同截距的同截距假设（见图 9—3c）。类似地，如果级差斜率系数 β_2 统计不显著而 α_2 是显著的，那我们或许就不能拒绝这两个回归具有相同斜率的平行回归假设（比较图 9—3b）。用通常的 F 检验（回想受约束最小二乘 F 检验）能对整个回归的稳定性

（即 α_2 和 β_2 同时为零）进行检验。如果不能拒绝这个假设，那么回归线将像图 9—3a 那样重合。

3. 邹至庄检验不能明确告诉我们截距和斜率系数中到底是哪个不同，还是都不相同。也就是说，只是截距不同、只是斜率不同或二者都不同皆可得到一个显著的邹至庄检验。换言之，我们不能从邹至庄检验中得知，在给定情形中，图 9—3 中的四种可能性到底存在哪一种。就此看来，虚拟变量方法具有明显的优势，因为它不仅能告诉我们两个回归是否不同，而且还能确定这种差别的来源——是源于截距、斜率还是二者皆有。在实践中，了解两个回归到底是哪个系数不同，其重要性绝不亚于对这两个回归不同的了解。

4. 最后，由于数据混合（即在一个回归中包括所有的观测）增加了自由度，这可能会提高估计系数的相对精度。当然，要记住的是，每增加一个虚拟变量都消耗一个自由度。

9.6 使用虚拟变量的交互效应

虚拟变量是一个能处理一系列有趣问题的灵活工具。为看出这一点，考虑如下模型：

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \tag{9.6.1}$$

其中Y＝以美元计的小时工资；

X＝受教育水平（读书年数）；

D_2＝1，若为女性；＝0，其他；

D_3＝1，若既非白人又非西班牙裔人；＝0，其他。

在这个模型中，性别和种族为定性回归元，而受教育水平为定量变量。[①] 这个模型中的暗含假定是，性别虚拟变量 D_2 的差别影响对两个种族类别而言是一样的，而种族虚拟变量 D_3 的差别影响对两个性别而言也是一样的。这就是说，若男性工资的均值比女性高，则不论是对哪一个种族来讲都是如此。同理，若既非白人又非西班牙裔人的工资均值较低，则不论他们是男性还是女性都是如此。

在许多应用中，这种假定是不能成立的。在既非白人又非西班牙裔人的种族中，女性可能比男性挣得少。换句话说，两个定性变量 D_2 和 D_3 之间可能会**相互影响**（interaction）。因此，它们对 Y 均值的影响可能不像方程（9.6.1）那样单纯是**相加形式的**，还有可能像如下模型一样是**乘积形式的**：

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i \tag{9.6.2}$$

其中变量定义与模型（9.6.1）相同。

我们从方程（9.6.2）得到：

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i \tag{9.6.3}$$

① 如果我们用高中以下学历、高中文化和高中以上学历来表示受教育程度，那我们就只需要两个虚拟变量来表示这三个类别。

这就是既非白人又非西班牙裔人的女性平均小时工资。观察到

α_2＝作为女性的级差效应；

α_3＝作为非白人又非西班牙裔人的级差效应；

α_4＝作为非白人又非西班牙裔人女性的级差效应。

它表明，女性非白人又非西班牙裔人的小时工资均值与女性或非白人又非西班牙裔人男性有所不同（差别为 α_4）。例如，如果所有三个级差系数都为负，那就意味着，在与基组（在本例中即男性白人或男性西班牙裔人组）相比时，非白人又非西班牙裔人女性比女性或非白人又非西班牙裔人男性挣得更少。

读者现在可以看出，**交互虚拟变量**（interaction dummy，即两个定性或虚拟变量之积）如何能改变两个单独考虑的影响因素的作用（即将它们的影响相加）。

例 9.5　平均小时工资与受教育水平、性别和种族的关系

让我们首先给出基于模型（9.6.1）的回归结论。利用用于估计回归（9.3.1）的数据，我们得到如下结果：

$$\hat{Y}_i = -0.261\,0 - 2.360\,6D_{2i} - 1.732\,7D_{3i} + 0.802\,8X_i \tag{9.6.4}$$
$$t=(-0.235\,7)^{**}\ (-5.487\,3)^{*}\ (-2.180\,3)^{*}\ (9.909\,4)^{*}$$
$$R^2=0.203\,2,\quad n=528$$

其中 * 表示 p 值低于 5%，** 表示 p 值高于 5%。

读者可以验证，级差截距系数都是统计显著的，并具有预期的符号（为什么?），而且作为一个无足为奇的结论，受教育水平对小时工资有很大的影响。

如方程（9.6.4）所示，在其他条件不变的情况下，女性的平均小时工资约低 2.36 美元，非白人又非西班牙裔人男性的平均小时工资也约低 1.73 美元。

我们现在来考虑包括交互虚拟变量的模型（9.6.2）的结论。

$$\hat{Y}_i = -0.261\,0 - 2.360\,6D_{2i} - 1.732\,7D_{3i} + 2.128\,9D_{2i}D_{3i} + 0.802\,8X_i \tag{9.6.5}$$
$$t=(-0.235\,7)^{**}\ (-5.487\,3)^{*}\ (-2.180\,3)^{*}\ (1.742\,0)^{**}\ (9.909\,5)^{**}$$
$$R^2=0.203\,2,\quad n=528$$

其中 * 表示 p 值低于 5%，** 表示 p 值高于 5%。

如你所见，相加的两个虚拟变量仍是统计显著的，但交互虚拟变量在通常 5%的显著性水平上不是统计显著的；其实际的 p 值为 8%。若你认为这是一个足够低的概率，那么方程（9.6.5）中的结论就可做如下解释：保持受教育水平不变，若将三个虚拟系数相加则得到 −1.964（＝−2.360 5−1.732 7+2.128 9），这意味着，非白人又非西班牙裔人女性的小时工资均值约低 1.96 美元，这个数字介于−2.360 5（单纯性别差异）和−1.732 7（单纯种族不同）之间。

上例明显地揭示了模型中包含两个或多个虚拟变量时交互虚拟变量的作用。重要的是要注意到，我们在模型（9.6.5）中假定，小时工资相对受教育水平增长率（多受一年教育小时工资约增加 80 美分）对不同性别和不同种族而言没有什么不同。但情况可能并非如此，若你想对此加以检验，则必须引入级差斜率系数（见习题 9.25）。

9.7 季节分析中虚拟变量的使用

许多基于月度或季度数据的经济时间序列都表现出季节特征（规则地摆动）。比如在圣诞或其他重大的节假日期间百货公司的销售额、节假日家庭对货币（或现金余额）的需求、夏天对冰淇淋和软饮料的需求、谷物在收割季节过后的价格、搭乘飞机旅行的需求，等等。从一个时间序列中去掉季节因素或成分，使我们能专注于诸如趋势之类的其他因素，通常都是好事情。[①] 从时间序列中去除季节成分的过程被称为**除季节性**（deseasonalization）或**季节调整**（seasonal adjustment），由此得到的时间序列被称为**除季节性**（deseasonalized）或**季节调整后的**（seasonally adjusted）时间序列。诸如失业率、消费者价格指数（CPI）、生产者价格指数（PPI）和工业生产指数等重要的经济时间序列通常都以季节调整后的形式公布。

虽然有几种方法能用于去除一个时间序列中的季节性，但我们只考虑这些方法中的一种，即虚拟变量方法。[②] 为了说明如何用虚拟变量去除经济时间序列中的季节性，考虑表 9—3 中给出的数据。此表给出 1978—1985 年四种主要厨具销售数量的季度数据，这四种厨具是洗碗机、污物碾碎机、冰箱和洗衣机，均以千台计。此表还给出以 1982 年十亿美元计的耐用品支出数据。

为了说明虚拟变量方法，我们将只考虑样本期内冰箱的销售数据。我们首先看一下数据，如图 9—4 所示。这个数据表明，数据中可能存在与各个季节相联系的季节性。为了解是否如此，考虑如下模型：

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t \tag{9.7.1}$$

其中 Y_t＝冰箱销售数量（以千计），D 是虚拟变量，分别在相应季节取值 1，在其他季节取值 0。注意，为避免虚拟变量陷阱，我们为一年中的每个季度都指定一个虚拟变量，但不要截距项。若某给定季度中存在某种季节效应，将会由该季度虚拟变量系数的统计显著的 t 值表现出来。[③]

表 9—3　1978 年第Ⅰ季度到 1985 年第Ⅳ季度厨具销售和耐用品支出的季度数据

DISH	DISP	FRIG	WASH	DUR	DISH	DISP	FRIG	WASH	DUR
841	798	1 317	1 271	252.6	480	706	943	1 036	247.7
957	837	1 615	1 295	272.4	530	582	1 175	1 019	249.1

① 一个时间序列可以包含四个成分：**季节性**（seasonal）、**周期性**（cyclical）、**趋势**（trend）和严格的随机部分。

② 关于季节调整的各种模型，可参见，例如：Francis X. Diebod, *Elements of Forecasting*, 2d ed., South-Western Publishing, 2001, Chapter 5。

③ 注意一个技术上的问题。为每个季度指定一个虚拟变量的方法，假定季节因素（若存在的话）是确定的而非随机的。当我们在本书第 5 篇讨论时间序列计量经济学时还会重新回到这个问题上来。

续前表

DISH	DISP	FRIG	WASH	DUR	DISH	DISP	FRIG	WASH	DUR
999	821	1 662	1 313	270.9	557	659	1 269	1 047	251.8
960	858	1 295	1 150	273.9	602	837	973	918	262
894	837	1 271	1 289	268.9	658	867	1 102	1 137	263.3
851	838	1 555	1 245	262.9	749	860	1 344	1 167	280
863	832	1 639	1 270	270.9	827	918	1 641	1 230	288.5
878	818	1 238	1 103	263.4	858	1 017	1 225	1 081	300.5
792	868	1 277	1 273	260.6	808	1 063	1 429	1 326	312.6
589	623	1 258	1 031	231.9	840	955	1 699	1 228	322.5
657	662	1 417	1 143	242.7	893	973	1 749	1 297	324.3
699	822	1 185	1 101	248.6	950	1 096	1 117	1 198	333.1
675	871	1 196	1 181	258.7	838	1 086	1 242	1 292	344.8
652	791	1 410	1 116	248.4	884	990	1 684	1 342	350.3
628	759	1 417	1 190	255.5	905	1 028	1 764	1 323	369.1
529	734	919	1 125	240.4	909	1 003	1 328	1 274	356.4

注：DISH=洗碗机，DISP=污物碾碎机，FRIG=冰箱，WASH=洗衣机，这四个数据以千台计；DUR=耐用品支出，以1982年十亿美元计。

资料来源：*Business Statistics and Survey of Current Business*，Department of Commerce（various issues）.

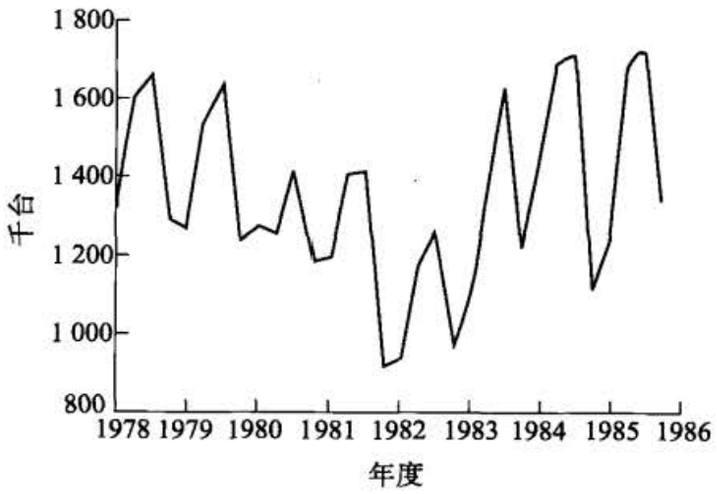


图 9—4　1978—1985 年冰箱销售数量（季度数据）

注意，我们在方程（9.7.1）中实际上将 Y 对一个截距进行回归，不同之处在于，我们容许不同的季节（即季度）有不同的截距。结果，每个虚拟变量的系数都将给出相应季度或季节里冰箱销售的均值。（为什么?）

例 9.6　冰箱销售中的季节性

从表 9—4 中给出的冰箱销售数据，我们得到如下回归结果：

$$\hat{Y}_t = 1\,222.125D_{1t} + 1\,467.500D_{2t} + 1\,569.750D_{3t} + 1\,160.000D_{4t} \quad (9.7.2)$$

$$t = (20.372\,0)\quad (24.462\,2)\quad (26.166\,6)\quad (19.336\,4)$$

$$R^2 = 0.531\,7$$

注：我们没有给出估计系数的标准误，因为所有的虚拟变量都只取值1或0，所以每个标准误都等于59.990 4。

方程（9.7.2）中α的系数估计值表示了每个季节（即季度）冰箱销售（以千台计）的平均数量或均值。于是，第一季度冰箱的平均销售量约为1 222千台，第二季度约1 468千台，第三季度约1 570千台，第四季度约1 160千台。

顺便指出，为避免虚拟变量陷阱，我们曾为每个季度指定一个虚拟变量并省略截距项，我们也可以在包括截距项的同时只指定三个虚拟变量。假设我们把第一季度作为参照季度，并为第二、三和四季度指定虚拟变量，就得到如下回归结果（见表9—4中的数据背景）：

$$\hat{Y}_t = 1\,222.1\,250 + 245.375\,0D_{2t} + 347.625\,0D_{3t} - 62.125\,0D_{4t}$$

$$t = (20.372\,0)^{*} \quad (2.892\,2)^{*} \quad (4.097\,4)^{*} \quad (-0.732\,2)^{**} \qquad (9.7.3)$$

$$R^2 = 0.531\,8$$

其中*表示p值低于5%，**表示p值高于5%。

表9—4　　1978—1985年美国各季度冰箱销售数据

FRIG	DUR	D_2	D_3	D_4	FRIG	DUR	D_2	D_3	D_4
1 317	252.6	0	0	0	943	247.7	0	0	0
1 615	272.4	1	0	0	1 175	249.1	1	0	0
1 662	270.9	0	1	0	1 269	251.8	0	1	0
1 295	273.9	0	0	1	973	262.0	0	0	1
1 271	268.9	0	0	0	1 102	263.3	0	0	0
1 555	262.9	1	0	0	1 344	280.0	1	0	0
1 639	270.9	0	1	0	1 641	288.5	0	1	0
1 238	263.4	0	0	1	1 225	300.5	0	0	1
1 277	260.6	0	0	0	1 429	312.6	0	0	0
1 258	231.9	1	0	0	1 699	322.5	1	0	0
1 417	242.7	0	1	0	1 749	324.3	0	1	0
1 185	248.6	0	0	1	1 117	333.1	0	0	1
1 196	258.7	0	0	0	1 242	344.8	0	0	0
1 410	248.4	1	0	0	1 684	350.3	1	0	0
1 417	255.5	0	1	0	1 764	369.1	0	1	0
919	240.4	0	0	1	1 328	356.4	0	0	1

注：FRIG＝冰箱销售数量，以千台计；DUR＝耐用品支出，以1982年十亿美元计；D_2取值1表示第二季度，否则取值0；D_3取值1表示第三季度，否则取值0；D_4取值1表示第四季度，否则取值0。

资料来源：*Business Statistics and Survey of Current Business*, Department of Commerce (various issues).

因为我们把第一季度视为基准组，所以各季度虚拟变量的系数现在就是级差截距，表示在虚拟变量取值为1的那个季度里，Y的平均值与基准季度相比有多大的差异。换言之，季节虚拟变量的系数将给出Y的平均值相对基准季度的增加或减少。如果你将各个级差截距值与基准平均值1 222.125相加，你就会得到各个季度的平均值。如此一来，除四舍五入的误差外，你将恰好重新得到方程（9.7.2）。

但现在你将看到将一个季度作为基准季度的价值，方程（9.7.3）表明第四季度Y的平均值并非统计上异于第一季度的平均值，因为第四季度的虚拟变量系数并非统计显著。当然，你的结论将随着哪一组作为基准组而变化，但总体结论将不会改变。

我们如何能得到冰箱销售的除季节性时间序列呢？这很容易做到。用模型（9.7.2）［或(9.7.3)］对每个观测估计Y值，然后用每个实际Y值减去Y的估计值便得到回归（9.7.2）的残

差（$Y_t - \hat{Y}_t$）。我们在表 9—5 中给出这些残差。①

表 9—5　　冰箱销售回归：实际值、拟合值和残差［方程（9.7.3）］

	实际值	拟合值	残差	残差图
1978—Ⅰ	1 317	1 222.12	94.875	
1978—Ⅱ	1 615	1 467.50	147.500	
1978—Ⅲ	1 662	1 569.75	92.250	
1978—Ⅳ	1 295	1 160.00	135.000	
1979—Ⅰ	1 271	1 222.12	48.875	
1979—Ⅱ	1 555	1 467.50	87.500	
1979—Ⅲ	1 639	1 569.75	69.250	
1979—Ⅳ	1 238	1 160.00	78.000	
1980—Ⅰ	1 277	1 222.12	54.875	
1980—Ⅱ	1 258	1 467.50	−209.500	
1980—Ⅲ	1 417	1 569.75	−152.750	
1980—Ⅳ	1 185	1 160.00	25.000	
1981—Ⅰ	1 196	1 222.12	−26.125	
1981—Ⅱ	1 410	1 467.50	−57.500	
1981—Ⅲ	1 417	1 569.75	−152.750	
1981—Ⅳ	919	1 160.00	−241.000	
1982—Ⅰ	943	1 222.12	−279.125	
1982—Ⅱ	1 175	1 467.50	−292.500	
1982—Ⅲ	1 269	1 569.75	−300.750	
1982—Ⅳ	973	1 160.00	−187.000	
1983—Ⅰ	1 102	1 222.12	−120.125	
1983—Ⅱ	1 344	1 467.50	−123.500	
1983—Ⅲ	1 641	1 569.75	71.250	
1983—Ⅳ	1 225	1 160.00	65.000	
1984—Ⅰ	1 429	1 222.12	206.875	
1984—Ⅱ	1 699	1 467.50	231.500	
1984—Ⅲ	1 749	1 569.75	179.250	
1984—Ⅳ	1 117	1 160.00	−43.000	
1985—Ⅰ	1 242	1 222.12	19.875	
1985—Ⅱ	1 684	1 467.50	216.500	
1985—Ⅲ	1 764	1 569.75	194.250	
1985—Ⅳ	1 328	1 160.00	168.000	
				− 0 +

① 当然，这假定了虚拟变量方法是除去时间序列中季节性的适当方法，而且假定时间序列（TS）又能表示成：TS=$s+c+t+u$，其中 s 表示季节性，c 表示周期性，t 表示趋势，u 表示随机部分。但如果时间序列的形式是 TS=$(s)(c)(t)(u)$，即四个成分是相乘的关系，则上述除季节性的方法就不适用，因为上述方法假定时间序列的四个成分是相加的关系。我们在时间序列计量经济学的章节中还会对此谈得更多。

这些残差代表着什么呢？它们代表着冰箱销售时间序列中除去季节因素后剩余的成分，即趋势、周期和随机几种成分（但参见上页注释①中的忠告）。

既然模型（9.7.2）和（9.7.3）都没有包含任何协变量，那我们在模型中引入定量回归元会改变这种情况吗？由于耐用品支出对冰箱需求有重要影响，所以让我们通过引入这个变量来扩展我们的模型（9.7.3）。以1982年十亿美元计的耐用品支出数据已经在表9—3中给出。这就是我们模型中的（定量）X变量。回归结果如下

$$\hat{Y}_t = 456.244\ 0 + 242.497\ 6D_{2t} + 325.264\ 3D_{3t} - 86.080\ 4D_{4t} + 2.773\ 4X_t$$
$$t = (2.559\ 3)^{*}\quad (3.695\ 1)^{*}\quad (4.942\ 1)^{*}\quad (-1.307\ 3)^{**}\quad (4.449\ 6)^{*} \qquad (9.7.4)$$
$$R^2 = 0.729\ 8$$

其中*表示p值低于5%，**表示p值高于5%。

同样记住，我们仍把第一季度作为基组。与在方程（9.7.3）中一样，我们看到第二季度和第三季度的级差截距系数都统计显著地异于第一季度，但第四季度的截距与第一季度的截距在统计上大致相同。约为2.77的X（耐用品支出）系数告诉我们，容许季节效应，若耐用品支出增加1个单位，则冰箱销售平均上升约2.77个单位，即近似为3个单位；记住冰箱销售以千台计，而X则以1982年十亿美元计。

这里一个有趣的问题是：正如冰箱销售所表现出来的季节性类型一样，耐用品支出不也表现出季节性类型吗？那我们如何考虑X的季节性呢？方程（9.7.4）中有趣的是，模型中的虚拟变量不仅去掉了Y中的季节性，还去掉了X中的季节性（若存在的话）。[这来自于统计学中的一个著名定理，即**弗里希-沃夫定理**（Frisch-Waugh theorem）。①] 真可谓是一石（虚拟变量方法）二鸟（除去两个时间序列中的季节性）！

若你想得到上述命题的非正式证明，则只需按如下步骤：（1）如方程（9.7.2）或（9.7.3）那样将Y对虚拟变量回归并保留残差S_1，这些残差代表除趋势后的Y。（2）对X做一个类似的回归并得到此回归的残差S_2；这些残差代表除趋势后的X。（3）将S_1对S_2回归。你将发现此回归中的斜率系数恰好就是回归（9.7.4）中X的系数。

9.8 分段线性回归

为了说明虚拟变量的另一种用处，考虑图9—5，它表明一个假想的公司是如何酬劳其销售代表的。其支付佣金的方式取决于销售量的一个目标或临界值X^*，低于那个值，就使用一种（随机）佣金结构，高于那个水平就使用另一种佣金结构。（注：除销售量外，还有其他因素影响销售佣金。假定所有这些其他因素都由随机干扰项表示。）更具体而言，假定销售佣金在临界值X^*之前随销售量线性地增加，在

① 其证明可参见 Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar, Lyme, U. K., 1995, pp. 150-152。

这个临界值之后仍线性地增加，只是斜率更大。于是，我们得到一个由两段或两部分构成的**分段线性回归**（piecewise linear regression），在图 9—5 中用Ⅰ和Ⅱ表示这两段，而且销售佣金是在临界值处改变斜率的。给定佣金、销售额和临界值 X^* 的数据，就能用虚拟变量的方法估计图 9—5 中所示的分段线性回归两个线段的（不同）斜率。我们可以如下进行：

$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i \tag{9.8.1}$$

其中Y_i=销售佣金；

X_i=销售员带来的销售量；

X^*=销售临界值，也被称为**结点**（knot）（事先已知）[①]；

D_i=1，若 $X_i > X^*$；

=0，若 $X_i < X^*$。

假定 $E(u_i)=0$，我们就立即看到

$$E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i \tag{9.8.2}$$

它给出目标销售量 X^* 之前销售佣金的均值，而

$$E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i \tag{9.8.3}$$

则给出目标销售量 X^* 之后销售佣金的均值。

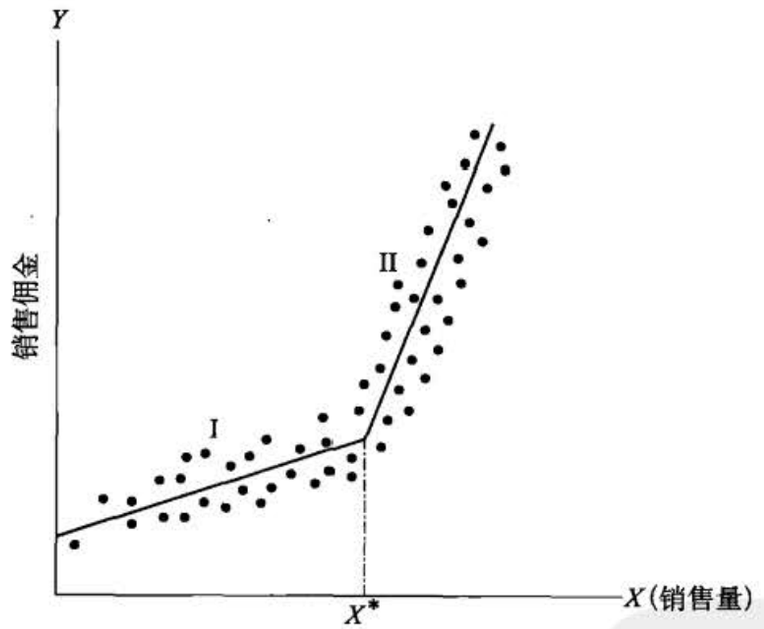


图 9—5　假想销售佣金与销售量之间的关系

注：Y 轴上的截距表示确保的最低佣金。

① 可临界值并不总是那么明显。作为权宜之计，可以将因变量相对解释变量描点，并观察二者之间的关系是否在 X 的某给定值（即 X^*）前后有明显的变化。发现转折点的分析方法在所谓**转换回归模型**（switching regression models）中可以见到。但这是一个高深的专题，对此进行讨论的教材可参见 Thomas Fomby, R. Carter Hill, and Stanley Johnson, *Advanced Econometric Methods*, Springer-Verlag, New York, 1984, Chapter 14。

于是，对图 9—5 中的分段线性回归而言，β_1 就给出了线段 I 中回归线的斜率，而 $\beta_1+\beta_2$ 则给出了线段 II 中回归线的斜率。要检验回归在临界值 X^* 处没有转折的假设，通过了解所估计的级差斜率系数 $\hat{\beta}_2$ 的统计显著性很容易就能做到（见图 9—6）。

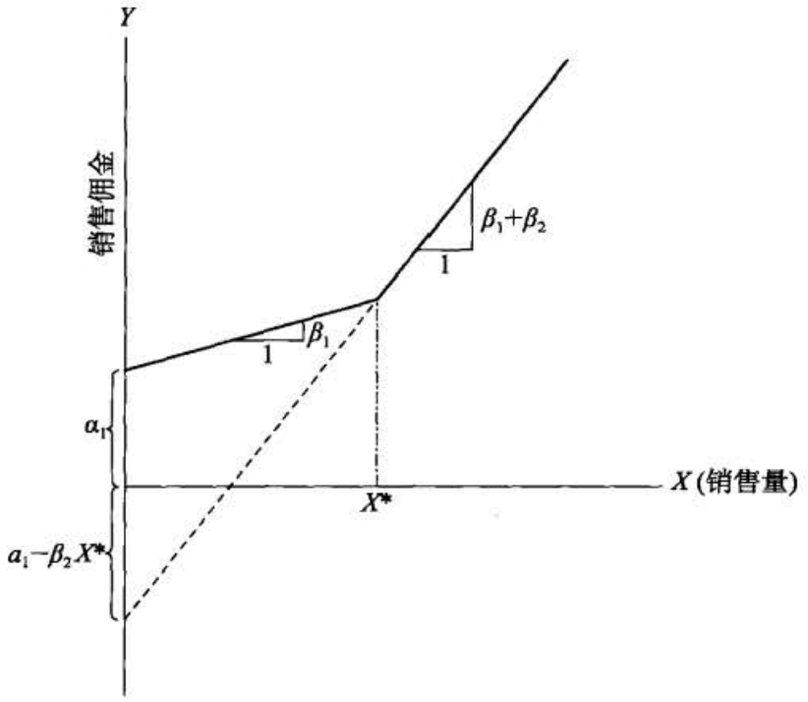


图 9—6　分段线性回归的参数

顺便指出，我们刚才所讨论的分段线性回归，只是所谓**间断函数**（spline functions）这种更一般函数类型中的一个例子。①

例 9.7　总成本与产出之间的关系

作为分段线性回归的一个应用例子，考虑表 9—6 中所给出的假想的总成本—总产出数据。我们还被告知，总成本在产出为 5 500 个单位时可能会改变其斜率。

在方程（9.8.4）以 Y 表示总成本和 X 表示总产出，我们得到如下结果：

$$\hat{Y}_i = -145.72 + 0.279\,1X_i + 0.094\,5(X_i - X_i^*)D_i$$
$$t = (-0.824\,5)\ (6.066\,9)\quad (1.144\,7) \qquad (9.8.4)$$
$$R^2 = 0.973\,7,\ X^* = 5\,500$$

如这些结论所示，生产的边际成本约为每单位产出 28 美分，而当产出高于 5 500 单位后，边际成本则约为 37 美分（=28+9），由于虚拟变量在例如 5%的显著性水平上仍不显著，所以这两个斜率之间的差别不是统计显著的。于是，在所有的实践中，都可以去掉虚拟变量而将总成本直接对总产出进行回归。

① 对间断回归的一个容易切入的讨论例如，可参见 Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, 3d ed., New York, 2001, pp. 228-230。

表 9—6　　总产出与总成本的假想数据

总成本（美元）	总产出（单位）
256	1 000
414	2 000
634	3 000
778	4 000
1 003	5 000
1 839	6 000
2 081	7 000
2 423	8 000
2 734	9 000
2 914	10 000

9.9　面板数据回归模型

记得我们在第 1 章曾讨论过一系列可用于经验分析的数据，如横截面数据、时间序列数据、混合数据（时间序列数据与横截面数据的综合）和面板数据等。虚拟变量方法很容易就能扩展应用到混合数据和面板数据。由于面板数据的使用在应用研究中日益常见，所以我们将在第 16 章更详尽地讨论这个专题。

9.10　虚拟变量方法的某些技术问题

□ 在半对数回归中对虚拟变量的解释

我们在第 6 章讨论过线性到对数模型，其中的回归子为对数形式，而回归元为线性形式。在这样的模型中，回归元的斜率系数给出了半弹性的解释，即回归元的单位变化导致回归子的百分比变化。这只在回归元为定量变量时如此。如果回归元是一个虚拟变量又会怎么样？为明确起见，考虑如下模型

$$\ln Y_i = \beta_1 + \beta_2 D_i + u_i \tag{9.10.1}$$

其中 Y 表示小时工资率（美元），D 取值 1 表示女性，取值 0 表示男性。

我们如何解释这一模型呢？假定 $E(u_i)=0$，我们得到

男性工人的工资函数：

$$E(\ln Y_i \mid D_i = 0) = \beta_1 \tag{9.10.2}$$

女性工人的工资函数：

$$E(\ln Y_i \mid D_i = 1) = \beta_1 + \beta_2 \tag{9.10.3}$$

因此，截距 β_1 给出了男性工人小时工资对数的均值，而“斜率”系数则给出了男女之间小时工资对数均值的差别。这是一种相当笨拙的表述方法。但如果我们对 β_1 取反对数，所得到的就不是男性工人小时工资的均值，而是他们工资的**中位数**（median）。如你所知，均值、中位数和众数是度量一个随机变量中心趋势的三种方法。如果我们对 $\beta_1+\beta_2$ 取反对数，则得到女性工人小时工资的中位数。

例 9.8　小时工资的对数与性别的关系

为了说明方程（9.10.1），我们使用例 9.2 背后的数据，基于 528 个观测所得到的回归结果如下：

$$\widehat{\ln Y_i} = 2.176\,3 \quad - \quad 0.243\,7D_i \tag{9.10.4}$$
$$t = (72.294\,3)^* \quad (-5.504\,8)^*$$
$$R^2 = 0.054\,4$$

其中 * 表示 p 值实际上是零。

取 2.176 3 的反对数便得到男性工人小时工资的中位数 8.813 6 美元，而取[(2.176 3－0.243 7)＝1.928 57]的反对数则得到女性工人小时工资的中位数 6.879 6 美元。因此，与男性工人相比，女性工人小时工资的中位数约低 21.94%[＝(8.813 6－6.879 6)/8.813 6]。

有趣的是，按照哈沃森（Halvorsen）和帕姆奎斯特（Palmquist）所建议的方法，我们可直接得到一个虚拟回归元的半弹性。[①] 对所估计的虚拟系数取反对数（以 e 为底）并减去 1，再将差值乘以 100 即可。（至于其背后的原因，可参见附录 9A.1。）因此，取－0.243 7 的反对数得到 0.783 66，减去 1 后得到－0.216 3，再乘以 100 便得到－21.63%。这就表明一个女性工人（D＝1）薪水的中位数比同等男性工人薪水的中位数低约 21.63%，除四舍五入的误差，与我们前面所得到的结论相同。

□ 虚拟变量与异方差性

让我们重新回到例子 1970—1981 年期间和 1982—1995 年期间及整个 1970—1995 年期间的美国储蓄—收入回归。在用虚拟变量方法检验结构稳定性时，我们假定误差 $\text{var}(u_{1i})=\text{var}(u_{2i})=\sigma^2$，即两个期间的误差方差相同。这也是邹至庄检验所需要的一个假定。如果这个假定不成立——即两个子期间的误差方差不同，那就很有可能得到具有误导性的结论。所以，必须首先利用合适的统计方法来验证两个子期间方差的相等性。尽管我们在异方差性那一章中将更全面地讨论这个专题，但我们在第 8 章还是看到如何用 F 检验来解决这个问题。[②]（参见我们在第 8 章中对邹至

① Robert Halvorsen and Raymond Palmquist, “The Interpretation of Dummy Variables in Semilogarithmic Equations,” *American Economic Review*, vol. 70, no. 3, pp. 474-475.

② 即便在出现异方差性的情况下，也能使用邹至庄检验程序，但那就必须使用**瓦尔德检验**（Wald test）。此检验背后涉及的数学多少有些复杂。在有关异方差性的章节，我们会再次讨论这个问题。

庄检验的讨论。）我们在那里曾指出，两个期间的误差方差看似有所不同。因此，前面给出的邹至庄检验和虚拟变量方法的结论可能都不是那么可靠。当然，我们这里只是说明可以用于解决一个问题（如结构稳定性的问题）的各种方法。在某个特定的研究中，这些方法可能站不住脚，但大多数统计方法都是如此。当然，我们在稍后有关异方差性的一章中将会看到，也可以采用一些适当的补救措施来解决这个问题（参见习题 9.28）。

□ 虚拟变量与自相关

除异方差性外，经典线性回归模型还假定回归模型中的误差项不相关。但若这些误差项（特别是涉及虚拟回归元的模型）相关，结果会怎么样呢？由于我们在关于自相关的一章中将深入讨论自相关这个专题，所以我们把对这个问题的回答留到那个时候。

□ 若因变量是一个虚拟变量会怎么样？

在我们到目前为止所讨论的模型中，回归子都是定量的，而回归元可以是定性的，也可以是定量的，也可以二者都有。但回归子偶尔也可能是定性或虚拟变量。比如，考虑一个工人是否参加劳动力市场的决策。参与的决策是一个是否命题，如果决定参与则回答是，决定不参与则回答否。当然，参与劳动力市场的决策还取决于其他几个因素，如起始工资率、受教育水平和劳动市场状况（由失业率度量）等。

我们还能用 OLS 估计回归子是虚拟变量的回归模型吗？是的，我们仍可机械地这么做。但在这种模型中会遇到几个统计问题。而且既然还有一些不同于 OLS 估计的方法不存在这些问题，那我们就把这个专题留到稍后的一章中讨论（见关于 logit 和 probit 模型的第 15 章）。我们在那一章中还将讨论回归子不止两种类型的模型；比如，是自己开车、乘公共汽车还是坐地铁去上班的决策，或者选择兼职工作、全职工作还是根本就不工作的决策。与因变量只有两种类型的**二值因变量模型**（dichotomous dependent variable models）相比，这种模型被称为**多值因变量模型**（polytomous dependent variable models）。

9.11 进一步研究的专题

文献中讨论的几个与虚拟变量相关的专题相当高深，包括（1）**随机或变参数模型**（random or varying parameters models），（2）**转换回归模型**（switching regression models），及（3）**非均衡模型**（disequilibrium models）。

在本书所讨论的回归模型中，假定参数 β 都是未知但固定的。随机系数模型及其

几个变形则假定 β 也可以是随机的。此领域一个重要的参考文献就是斯瓦迷（Swamy）的一本书。[①]

在同时使用级差截距与级差斜率的虚拟变量模型中，暗含地假定了我们知道转折点。于是，在我们 1970—1995 年期间的储蓄—收入例子中，深信 1982 年的衰退改变了储蓄与收入之间的关系，我们将期间分为 1970—1981 年和 1982—1995 年两个（衰退前和衰退后）子区间。有时候不容易确定转折什么时候发生。**转换回归模型**就是在这种情况下提出的方法。转换回归模型将转折点视为一个随机变量，并通过一个迭代过程来决定转折实际上在什么时候发生。此领域的开山之作当属戈德菲尔德（Goldfeld）和匡特（Quandt）的那本书。[②]

欲处理市场未出清（即需求不等于供给）的**非均衡状态**（disequilibrium situations），则需要一些特殊的估计方法。经典的例子是一种商品的供给与需求。对一种商品的需求是其价格和其他一些变量的函数，而对一种商品的供给也是其价格和其他一些变量的函数，但这些变量与进入需求函数的变量不完全相同。现在，实际购买和卖出的商品数量并不一定等于供求相等时的数量，因此导致非均衡。对**非均衡模型**的全面讨论，读者可参见匡特的著作。[③]

9.12 一个结束性例子

我们用一个例子来结束本章的讨论，用以说明本章得到的一些论点。表 9—7 提供了 1990 年印度南部一个工业小镇中 261 个工人的样本数据

表 9—7　　1990 年的一个印度工人样本

WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}	WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}
120	57	0	0	0	0	0	120	21	0	0	0	0	0
224	48	0	0	1	1	0	25	18	0	0	0	0	1
132	38	0	0	0	0	0	25	11	0	0	0	0	1
75	27	0	1	0	0	0	30	38	0	0	0	1	1
111	23	0	1	0	0	1	30	17	0	0	0	1	1
127	22	0	1	0	0	0	122	20	0	0	0	0	0
30	18	0	0	0	0	0	288	50	0	1	0	1	0
24	12	0	0	0	0	0	75	45	0	0	0	0	1
119	38	0	0	0	1	0	79	60	0	0	0	0	0
75	55	0	0	0	0	0	85.3	26	1	0	0	0	1

① P. A. V. B. *Swamy*, *Statistical Inference in Random Coefficient Regression Models*, Springer-Verlag, Berlin, 1971.

② S. Goldfeld and R. Quandt, *Nonlinear Methods in Econometrics*, North Holland, Amsterdam, 1972.

③ Richard E. Quandt, *The Econometrics of Disequilibrium*, Basil Blackwell, New York, 1988.

续前表

WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}	WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}
324	26	0	1	0	0	0	350	42	0	1	0	1	0
42	18	0	0	0	0	0	54	62	0	0	0	1	0
100	32	0	0	0	0	0	110	23	0	0	0	0	0
136	41	0	0	0	0	0	342	56	0	0	0	1	0
107	48	0	0	0	0	0	77.5	19	0	0	0	1	0
50	16	1	0	0	0	1	370	46	0	0	0	0	0
90	45	0	0	0	0	0	156	26	0	0	0	1	0
377	46	0	0	0	1	0	261	23	0	0	0	0	0
150	30	0	1	0	0	0	54	16	0	1	0	0	0
162	40	0	0	0	0	0	130	33	0	0	0	0	0
18	19	1	0	0	0	0	112	27	1	0	0	0	0
128	25	1	0	0	0	0	82	22	1	0	0	0	0
47.5	46	0	0	0	0	1	385	30	0	1	0	1	0
135	25	0	1	0	0	0	94.3	22	0	0	1	1	1
400	57	0	0	0	1	0	350	57	0	0	0	1	0
91.8	35	0	0	1	1	0	108	26	0	0	0	0	0
140	44	0	0	0	1	0	20	14	0	0	0	0	0
49.2	22	0	0	0	0	0	53.8	14	0	0	0	0	1
30	19	1	0	0	0	0	427	55	0	0	0	1	0
40.5	37	0	0	0	0	1	18	12	0	0	0	0	0
81	20	0	0	0	0	0	120	38	0	0	0	0	0
105	40	0	0	0	0	0	40.5	17	0	0	0	0	0
200	30	0	0	0	0	0	375	42	1	0	0	1	0
140	30	0	0	0	1	0	120	34	0	0	0	0	0
80	26	0	0	0	0	0	175	33	1	0	0	1	0
47	41	0	0	0	0	1	50	26	0	0	0	0	1
125	22	0	0	0	0	0	100	33	1	0	0	1	0
500	21	0	0	0	0	0	25	22	0	0	0	1	1
100	19	0	0	0	0	0	40	15	0	0	0	1	0
105	35	0	0	0	0	0	65	14	0	0	0	1	0
300	35	0	1	0	1	0	47.5	25	0	0	0	1	1
115	33	0	1	0	1	1	163	25	0	0	0	1	0
103	27	0	0	1	1	1	175	50	0	0	0	1	1
190	62	1	0	0	0	0	150	24	0	0	0	1	1
62.5	18	0	1	0	0	0	163	28	0	0	0	1	0
50	25	1	0	0	0	0	163	30	1	0	0	1	0
273	43	0	0	1	1	1	50	25	0	0	0	1	1
175	40	0	1	0	1	0	395	45	0	1	0	1	0
117	26	1	0	0	1	0	175	40	0	0	0	1	1
950	47	0	0	1	0	0	87.5	25	1	0	0	0	0
100	30	0	0	0	0	0	75	18	0	0	0	0	0
140	30	0	0	0	0	0	163	24	0	0	0	1	0

续前表

WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}	WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}
97	25	0	1	0	0	0	325	55	0	0	0	1	0
150	36	0	0	0	0	0	121	27	0	1	0	0	0
25	28	0	0	0	0	1	600	35	1	0	0	0	0
15	13	0	0	0	0	1	52	19	0	0	0	0	0
131	55	0	0	0	0	0	117	28	1	0	0	0	0

变量定义如下：

WI＝以卢比度量的周收入；

Age＝年龄；

D_{SEX}＝1，表示男性工人；＝0，表示女性工人；

DE_2＝1，表示一个工人受过初等教育；

DE_3＝1，表示一个工人受过中等教育；

DE_4＝1，表示一个工人受过高等教育；

DPT＝1，表示一个工人拥有一份永久性工作；＝0，表示工作是暂时性的。

参照组是没接受过初等教育并拥有暂时性工作的男性工人。我们想弄清楚周收入与年龄、性别、受教育程度和工作性质之间的关系。为此，我们估计如下回归模型：

$$\ln WI_i=\beta_1+\beta_2 AGE_i+\beta_3 D_{SEX}+\beta_4 DE_2+\beta_5 DE_3+\beta_6 DE_4+\beta_6 DPT+u_i$$

根据劳动经济学文献，我们把工资对数表示成解释变量的函数。在第 6 章曾提到，像工资这种变量的密度分布倾向于偏斜；对这种变量进行对数变换既能降低其偏斜程度，又能减小其异方差性。

利用 EViews 6，我们得到如下回归结果。

```
Dependent Variable: Ln(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261
```

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.706872	0.113845	32.56055	0.0000
AGE	0.026549	0.003117	8.516848	0.0000
D_{SEX}	-0.656338	0.088796	-7.391529	0.0000
DE_2	0.113862	0.098542	1.155473	0.2490
DE_3	0.412589	0.096383	4.280732	0.0000
DE_4	0.554129	0.155224	3.569862	0.0004
DPT	0.558348	0.079990	6.980248	0.0000

R-squared	0.534969	Mean dependent var.	4.793390
Adjusted R-squared	0.523984	S.D. dependent var.	0.834277
S.E. of regression	0.575600	Akaike info criterion	1.759648
Sum squared resid.	84.15421	Schwarz criterion	1.855248
Log likelihood	-222.6340	Hannan-Quinn criter.	1.798076
F-statistic	48.70008	Durbin-Watson stat.	1.853361
Prob(F-statistic)	0.000000		

这些结论表明，周收入的对数与年龄、受教育程度和工作性质正相关，但与性别负相关，这是一个无足为奇的结论。尽管受过初等教育的工人与没有受过初等教育的工人相比，周收入看似没有实质性差异，但受过中等教育的工人的周薪却更高一些，而且受过高等教育的工人，其周收入还要高出更多。

虚拟变量的系数可解释为与参照组的差值。因此，*DPT* 变量的系数表明，那些拥有永久性工作的工人比那些拥有暂时工作的工人平均而言要挣更多的钱。

我们从第 6 章知道，在一个线性到对数模型（因变量是对数形式，解释变量是线性形式）中，解释变量的系数代表着半弹性，即给出了这个解释变量值的单位变化导致因变量的相对变化或百分比变化。但正如文中提到的那样，当解释变量是虚拟变量时，我们必须非常小心。这里我们必须取虚拟变量系数估计值的反对数并减去 1，再将结果乘以 100。因此，为了得到正式工人的周收入相对临时工人的周收入的百分比变化，我们要将 *DPT* 系数 0.558 348 取反对数并减去 1，然后乘以 100。在本例中，结果是 $(e^{0.558\,348}-1)=1.747\,78-1=0.747\,78$，或约 75%。建议读者对模型中包含的其他虚拟变量计算这种百分比变化。

其结果表明，性别和受教育程度对周收入具有明显影响。性别与受教育程度之间有可能存在交互影响吗？受过高等教育的男性工人比受过高等教育的女性工人挣的周收入更高吗？为了考察这种可能性，我们可以把上述工资回归方程加以扩展，引进性别与受教育程度的交互项。回归结果如下：

Dependent Variable: Ln(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	3.717540	0.114536	32.45734	0.0000
AGE	0.027051	0.003133	8.634553	0.0000
D_{SEX}	-0.758975	0.110410	-6.874148	0.0000
DE_2	0.088923	0.106827	0.832402	0.4060
DE_3	0.350574	0.104309	3.360913	0.0009
DE_4	0.438673	0.186996	2.345898	0.0198
$D_{SEX}*DE_2$	0.114908	0.275039	0.417788	0.6765
$D_{SEX}*DE_3$	0.391052	0.259261	1.508337	0.1327
$D_{SEX}*DE_4$	0.369520	0.313503	1.178681	0.2396
DPT	0.551658	0.080076	6.889198	0.0000

R-squared	0.540810	Mean dependent var.	4.793390
Adjusted R-squared	0.524345	S.D. dependent var.	0.834277
S.E. of regression	0.575382	Akaike info criterion	1.769997
Sum squared resid.	83.09731	Schwarz criterion	1.906569
Log likelihood	-220.9847	Hannan-Quinn criter.	1.824895
F-statistic	32.84603	Durbin-Watson stat.	1.856488
Prob (*F*-statistic)	0.000000		

尽管虚拟变量的交互项表明，性别与受教育程度之间存在一定的交互影响，但

这种影响在统计上并不显著，因为所有交互项的系数都不是个别统计显著的。

有意思的是，如果我们去掉受教育程度虚拟变量，而保留这些虚拟变量交互项，我们得到如下结果：

```
Dependent Variable: LOG(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261
```

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.836483	0.106785	35.92725	0.0000
AGE	0.025990	0.003170	8.197991	0.0000
D_{SEX}	-0.868617	0.106429	-8.161508	0.0000
$D_{SEX}*DE_2$	0.200823	0.259511	0.773851	0.4397
$D_{SEX}*DE_3$	0.716722	0.245021	2.925140	0.0038
$D_{SEX}*DE_4$	0.752652	0.265975	2.829789	0.0050
DPT	0.627272	0.078869	7.953332	0.0000

R-squared	0.514449	Mean dependent var.	4.793390
Adjusted R-squared	0.502979	S.D. dependent var.	0.834277
S.E. of regression	0.588163	Akaike info criterion	1.802828
Sum squared resid.	87.86766	Schwarz criterion	1.898429
Log likelihood	-228.2691	Hannan-Quinn criter.	1.841257
F-statistic	44.85284	Durbin-Watson stat.	1.873421
Prob (*F*-statistic)	0.000000		

现在看上去受教育程度本身对工人的周收入没有影响，但却以交互形式产生影响。这表明，我们在使用虚拟变量时必须小心。要弄清楚受教育程度与 *DPT* 是否存在交互影响，作为一个练习留给读者。

要点与结论

1. 取值 1 或 0（或其线性变换）的虚拟变量是在回归模型中引入定性回归元的一种手段。

2. 虚拟变量是一种基于性质或属性（性别、婚姻状况、种族和宗教信仰等）而将一个样本分为不同子群的数据分类方法，并暗含地容许对每个子群分别进行回归。如果回归子对各子群中定性变量的变化有不同的响应，那就由各子群回归的截距或斜率或二者的差别反映出来。

3. 尽管虚拟变量方法是一种通用方法，但仍需仔细处理。首先，如果回归中包含了截距项，那么，虚拟变量的个数就必须比每个定性变量的类别数少一个。其次，附属于虚拟变量的系数必须总是相对基组或参照组（即虚拟变量取值为 0 的组）来解释。基组的选择取决于所进行研究的目的。最后，如果一个模型含有几个均可分为几类的定性变量，那么，引入虚拟变量就会占用很大的自由度。因此，总是要根据分析中可供使用的观测数来决定要引入的虚拟变量数。

4. 在虚拟变量的诸多应用中，本章仅考虑了如下为数不多的几种：（1）将两个或多个回归进行比较，（2）除去时间序列数据中的季节性，（3）交互虚拟变量，（4）在半对数模型中解释虚拟

变量，及（5）分段线性回归。

5. 我们对在异方差性和自相关情形下使用虚拟变量提出了郑重警告。但由于我们在以后的章节中将会详尽讨论这些专题，所以还是留待以后再探个究竟。

习　　题

问答题

9.1　如果你有连续几年的月度数据，为检验如下假设，需要引入多少个虚拟变量：

a. 一年中所有 12 个月份都表现出季节类型。

b. 只有 2、4、6、8、10、12 月表现出季节类型。

9.2　考虑如下回归结果（t 比率放在括号内）[①]：

$$\begin{aligned}\hat{Y}_i &= 1\,286 + 104.97X_{2i} - 0.026X_{3i} + 1.20X_{4i} + 0.69X_{5i} \\ t &= (4.67)\quad (3.70)\quad (-3.80)\quad (0.24)\quad (0.08) \\ &\quad -19.47X_{6i} + 266.06X_{7i} - 118.64X_{8i} - 110.61X_{9i} \\ &\quad (-0.40)\quad (6.94)\quad (-3.04)\quad (-6.14) \\ &\quad R^2 = 0.383,\ n = 1\,543\end{aligned}$$

其中Y＝妻子希望每年花在工作上的小时数，以每年工作的小时数加上花在找工作的时间之和计算；

X_2＝妻子税后真实小时收入；

X_3＝丈夫在上一年度税后真实收入；

X_4＝妻子的年龄；

X_5＝妻子的受教育年数；

X_6＝态度变量，若被调查者愿意工作而且她丈夫也同意她工作则取值 1，否则取值 0；

X_7＝态度变量，若被调查者的丈夫支持她工作则取值 1，否则取值 0；

X_8＝年龄低于 6 岁的子女数；

X_9＝年龄在 6～13 岁之间的子女数。

a. 各非虚拟回归元系数的符号有经济含义吗？说明你的观点。

b. 你如何解释虚拟变量 X_6 和 X_7？这些虚拟变量统计显著吗？由于样本相当大，你可以“2－t”经验法则来回答问题。

c. 在这项研究中，一位妇女的年龄和受教育程度不是影响其劳动力参与决策的显著因素，你认为这是为什么？

9.3　考虑如下回归结果。[②]（实际数据在表 9—8 中。）

$$\begin{aligned}\widehat{\mathrm{UN}}_t &= 2.749\,1 + 1.150\,7D_t - 1.529\,4V_t - 0.851\,1\,(D_tV_t) \\ t &= (26.896)\ (3.628\,8)\ (-12.555\,2)\ (-1.981\,9) \\ &\quad R^2 = 0.912\,8\end{aligned}$$

---

① Jane Leuthold, “The Effect of Taxation on the Hours Worked by Married Women,” *Industrial and Labor Relations Review*, no. 4, July 1978, pp. 520-526.（为适合我们的格式符号有所变化。）

② Damodar Gujarati, “The Behaviour of Unemployment and Unfilled Vacancies: Great British, 1958—1971,” *The Economic Journal*, vol. 82, March 1972, pp. 195-202.

其中 UN =失业率，%；

V =岗位空缺率，%；

D =1，对1966年第Ⅳ季度之后的期间；

=0，对1966年第Ⅳ季度之前的期间；

t = 时间，以季度度量。

注：1966年第Ⅳ季度，当时英国的劳工党政府放松了对国民保险法案的限制，以定额年金和（先前）根据收入确定补贴的混合体制取代了短期失业救济的定额年金制，从而提高了失业救济金水平。

表 9—8　　习题 9.3 中所做回归的数据矩阵

年份和季度	失业率 UN,%	岗位空缺率 V,%	D	DV	年份和季度	失业率 UN,%	岗位空缺率 V,%	D	DV
1980—Ⅳ	1.915	0.510	0	0	1965—Ⅰ	1.201	0.997	0	0
1959—Ⅰ	1.876	0.541	0	0	—Ⅱ	1.192	1.035	0	0
—Ⅱ	1.842	0.541	0	0	—Ⅲ	1.259	1.040	0	0
—Ⅲ	1.750	0.690	0	0	—Ⅳ	1.192	1.086	0	0
—Ⅳ	1.648	0.771	0	0	1966—Ⅰ	1.089	1.101	0	0
1960—Ⅰ	1.450	0.836	0	0	—Ⅱ	1.101	1.058	0	0
—Ⅱ	1.393	0.908	0	0	—Ⅲ	1.243	0.987	0	0
—Ⅲ	1.322	0.968	0	0	—Ⅳ	1.623	0.819	1	0.819
—Ⅳ	1.260	0.998	0	0	1967—Ⅰ	1.821	0.740	1	0.740
1961—Ⅰ	1.171	0.968	0	0	—Ⅱ	1.990	0.661	1	0.661
—Ⅱ	1.182	0.964	0	0	—Ⅲ	2.114	0.660	1	0.660
—Ⅲ	1.221	0.952	0	0	—Ⅳ	2.115	0.698	1	0.698
—Ⅳ	1.340	0.849	0	0	1968—Ⅰ	2.150	0.695	1	0.695
1962—Ⅰ	1.411	0.748	0	0	—Ⅱ	2.141	0.732	1	0.732
—Ⅱ	1.600	0.658	0	0	—Ⅲ	2.167	0.749	1	0.749
—Ⅲ	1.780	0.562	0	0	—Ⅳ	2.107	0.800	1	0.800
—Ⅳ	1.941	0.510	0	0	1969—Ⅰ	2.104	0.783	1	0.783
1963—Ⅰ	2.178	0.510	0	0	—Ⅱ	2.056	0.800	1	0.800
—Ⅱ	2.067	0.544	0	0	—Ⅲ	2.170	0.794	1	0.794
—Ⅲ	1.942	0.568	0	0	—Ⅳ	2.161	0.790	1	0.790
—Ⅳ	1.764	0.677	0	0	1970—Ⅰ	2.225	0.757	1	0.757
1964—Ⅰ	1.532	0.794	0	0	—Ⅱ	2.241	0.746	1	0.746
—Ⅱ	1.455	0.838	0	0	—Ⅲ	2.366	0.739	1	0.739
—Ⅲ	1.409	0.885	0	0	—Ⅳ	2.324	0.707	1	0.707
—Ⅳ	1.296	0.987	0	0	1971—Ⅰ	2.516*	0.583*	1	0.583*
					—Ⅱ	2.909*	0.524*	1	0.524*

注：* 为初步估计值。

资料来源：Damodar Gujarati, "The Behavior of Unemployment and Unfilled Vacancies: Great Britain, 1958—1971," *The Economic Journal*, vol. 82, March 1972, p. 202.

a. 你对失业率和岗位空缺率之间的关系有何先验预期？

b. 保持岗位空缺率不变，在从1966年第Ⅳ季度开始的期间内，平均失业率为多少？它与1966年第Ⅳ季度之前的期间有显著不同吗？你何以知道？

c. 1966 年第Ⅳ季度之前和之后的斜率在统计上不同吗？你何以知道？

d. 从这项研究能安全地断定慷慨的失业救济金导致更高的失业率吗？这在经济上讲得通吗？

9.4 威廉·诺德豪斯（William Nordhaus）根据 1972—1979 年的年度数据估计了如下模型，用以解释 OPEC 的石油价格行为（括号中数字为标准误）。①

$$\hat{y}_t = 0.3x_{1t} + 5.22x_{2t}$$

$$\text{se} = (0.03)\quad (0.50)$$

其中y＝当年与前一年的价差，美元/桶；

x_1＝当年现货价格与前一年 OPEC 价格之差；

x_2＝虚拟变量，在 1974 年取值 1，否则取值 0。

解释这个结果并利用图形说明结论。这些结论表明 OPEC 有什么样的垄断势力？

9.5 考虑如下模型

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + u_i$$

其中 Y＝一位大学教授的年薪；

X＝从教年限；

D＝性别虚拟变量。

考虑定义虚拟变量的三种方式：

a. D 对男性取值 1，对女性取值 0。

b. D 对女性取值 1，对男性取值 2。

c. D 对女性取值 1，对男性取值−1。

对每种虚拟变量定义解释上述回归模型。是否有某个方法比另外一个方法更好？说明你的理由。

9.6 参考回归（9.7.3）。你如何检验 D_2 和 D_3 的系数相同的假设？如何检验 D_2 和 D_4 的系数相同的假设？如果 D_3 的系数在统计上与 D_2 的系数不同，而且 D_4 的系数与 D_2 的系数也不同，那是否意味着 D_3 和 D_4 的系数不同？提示：$\text{var}(A \pm B) = \text{var}(A) + \text{var}(B) \pm 2\text{cov}(A, B)$

9.7 参考 9.5 节中讨论的美国储蓄—收入一例。

a. 你如何从混合回归（9.5.4）中得到方程（9.5.5）和（9.5.6）中回归系数的标准误？

b. 为了得到数值结果，可能的话，还需要什么附加信息？

9.8 米勒（R. J. Miller）研究联邦存款保险公司（Federal Deposit Insurance Corporation，FDIC）在检查 91 家银行所耗费劳动时间的研究中，估计了如下函数②：

$$\widehat{\ln Y} = 2.41 + 0.367\,4\ln X_1 + 0.221\,7\ln X_2 + 0.080\,3\ln X_3$$
$$(0.047\,7)\qquad (0.062\,8)\qquad (0.028\,7)$$
$$-0.175\,5D_1 + 0.279\,9D_2 + 0.563\,4D_3 - 0.257\,2\,D_4$$
$$(0.290\,5)\qquad (0.104\,4)\qquad (0.165\,7)\qquad (0.078\,7)$$
$$R^2 = 0.766$$

其中 Y ＝FDIC 检查者的劳动时间；

X_1＝银行的总资产；

---

① "Oil and Economic Performance in Industrial Countries," *Brookings Papers on Economic Activity*, 1980, pp. 341-388.

② "Examination of Man-Hour Cost for Independent, Joint, and Divided Examination Programs," *Journal of Bank Research*, vol. 11, 1980, pp. 28-35. 注：为了与我们保持一致，符号有所变化。

X_2＝银行的办公室总数；

X_3＝银行分级贷款占总贷款的比例；

D_1＝1，若管理评级为“优”；

D_2＝1，若管理评级为“普通”；

D_3＝1，若管理评级为“满意”；

D_4＝1，若公司的检查与政府的检查同时进行。

括号中的数字为估计的标准误。

a. 解释这些结论。

b. 在解释模型中的虚拟变量时，因 Y 以对数形式出现而带来什么问题？

c. 你如何解释虚拟变量的系数？

9.9　为了评价美联储自 1979 年 7 月以来放松利率管制的政策效果，我们的一个学生悉尼·兰格（Sidney Langer）利用 1975 年第三季度至 1983 年第二季度期间的季度数据估计了如下模型[①]：

$$\hat{Y}_t = 8.587\,1 - 0.132\,8P_t - 0.710\,2\mathrm{Un}_t - 0.238\,9M_t$$
$$\text{se}=(1.956\,3)\quad(0.099\,2)\qquad(0.190\,9)\qquad(0.072\,7)$$
$$+0.659\,2Y_{t-1} + 2.583\,1\mathrm{Dum}_t \qquad R^2=0.915\,6$$
$$(0.103\,6)\qquad(0.754\,9)$$

其中Y ＝3 月期国债利率；

P＝预期通货膨胀率；

Un＝进行季节调整后的失业率；

M＝货币基础的变化；

Dum＝虚拟变量，对 1979 年 7 月 1 日之后的观测都取值 1。

a. 解释这一模型的估计结果。

b. 放松利率管制有什么样的影响？这些结果在经济上讲得通吗？

c. P_t、Un_t 和 M_t 的系数都为负，你能给出经济学上的根据吗？

9.10　参考本章中讨论的分段回归。假设在 X^* 处不仅斜率系数发生了变化，而且回归线还发生了跳跃，如图 9—7 所示。你将如何修正方程（9.8.1），以考虑回归线在 X^* 处的跳跃。

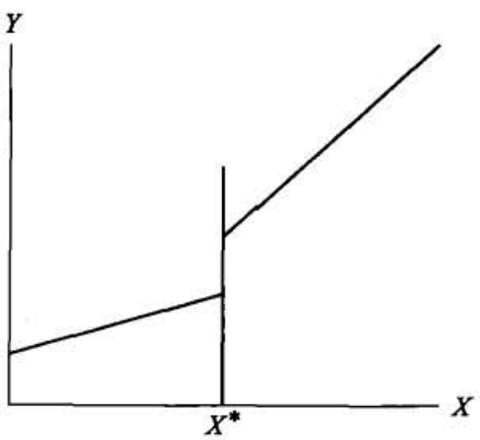


图 9—7　不连续的分段线性回归

① Sidney Langer, “Interest Rate Deregulation and Short-Term Interest Rates,” unpublished term paper.

9.11 **每盎司可乐价格的决定因素。**我的一个学生凯茜·谢弗（Cathy Schaefer）对77个横截面观测数据估计了如下回归[①]：

$$P_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \mu_i$$

其中P_i＝每盎司可乐的价格；

D_{1i}＝001，若是折扣商店；

＝010，若是连锁商店；

＝100，若是便利店；

D_{2i}＝10，若是品牌可乐；

＝01，若是无品牌可乐；

D_{3i}＝0001，若67.6盎司（2升）瓶装；

＝0010，若28～33.8盎司瓶装；

＝0100，若16盎司瓶装；

＝1000，若12盎司听装。

估计的结果如下：

$$\hat{P}_i = 0.014\ 3 - 0.000\ 004 D_{1i} + 0.009\ 0 D_{2i} + 0.000\ 01 D_{3i}$$

$$\text{se} = \quad (0.000\ 01) \quad (0.000\ 11) \quad (0.000\ 00)$$

$$t = \quad (-0.383\ 7) \quad (8.392\ 7) \quad (5.812\ 5)$$

$$R^2 = 0.603\ 3$$

注：标准误只保留到小数点后5位。

a. 对模型中引入虚拟变量的方式进行评论。

b. 假定虚拟变量的分类是可以接受的，你将如何对上述结果进行解释？

c. D_3的系数为正且显著，你如何合理地解释这个结论？

9.12 根据20世纪70年代早期101个国家以美元计人均收入（X）和以年计预期寿命（Y）的数据，森（Sen）和斯里瓦斯特瓦（Srivastava）得到如下回归结果[②]：

$$\hat{Y}_i = -2.40 + 9.39 \ln X_i - 3.36\ [D_i\ (\ln X_i - 7)]$$

$$\text{se} = (4.73)\quad (0.859)\quad (2.42)\qquad R^2 = 0.752$$

其中D_i在$\ln X_i$大于7时取值1，否则取值0。注：当$\ln X_i = 7$时，X＝1 097美元（近似）。

a. 以对数形式引入收入变量的原因是什么？

b. 你如何解释$\ln X_i$的系数9.39？

c. 引入回归元$D_i\ (\ln X_i - 7)$的理由是什么？你如何解释这个回归元？你又如何解释这个回归元的系数−3.36？（提示：线性分段回归。）

d. 假定穷国与富国之间的分界线为人均收入1 097美元，你如何推导出收入低于1 097美元国家的回归线和收入高于1 097美元国家的回归线？

e. 你从这个问题给出的结果中能得到什么一般结论？

9.13 考虑如下模型：

$$Y_i = \beta_1 + \beta_2 D_i + u_i$$

---

① Cathy Schaefer, "Price Per Ounce of Cola Beverage as a Function of Place of Purchase, Size of Container, and Branded or Unbranded Product," unpublished term project.

② Ashish Sen and Muni Srivastava, *Regression Analysis: Theory, Methods, and Applications*, Springer-Verlag, New York, 1990, p. 92. 符号有所改变。

其中 D_i 对前 20 个观测取值 0，而对后 30 个观测取值 1。并告诉你 var $(u_i^2)=300$。

a. 你如何解释 β_1 和 β_2？

b. 这两组的均值分别是多少？

c. 你如何计算 $\hat{\beta}_1+\hat{\beta}_2$ 的方差？注：已知 cov $(\hat{\beta}_1,\hat{\beta}_2)=-15$。

9.14 为了评价工作权法案（该法案不把工会会员作为就业的先决条件）对工会关系的影响，利用美国 1982 年 50 个州的数据估计出如下回归结果①：

$$\widehat{\text{PVT}}_i = 19.806\,6 - 9.391\,7\text{RTW}_i$$

$$t = (17.035\,2)\quad(-5.108\,6)$$

$$r^2 = 0.352\,2$$

其中 PVT=1982 年私人部门雇员加入工会的百分比；RTW 为虚拟变量，若工作权法案生效则 RTW 取值 1，否则取值 0。注：1982 年美国有 20 个州颁布工作权法案。

a. 据经验，预期 PVT 和 RTW 之间有什么样的关系？

b. 回归结果支持先验预期吗？

c. 解释回归结果。

d. 在那些没有颁布工作权法案的州，私人部门雇员加入工会的平均百分比是多少？

9.15 在如下回归模型中：

$$Y_i = \beta_1 + \beta_2 D_i + u_i$$

Y 表示以美元度量的小时工资，D 为虚拟变量，对大学毕业生取值 1，对高中毕业生取值 0。利用第 3 章中的 OLS 公式，证明 $\hat{\beta}_1=\bar{Y}_{hg}$ 和 $\hat{\beta}_2=\bar{Y}_{cg}-\bar{Y}_{hg}$，其中下标有如下含义：hg 表示高中毕业，cg 表示大学毕业。总共有 n_1 个高中毕业生和 n_2 个大学毕业生，总样本为 $n=n_1+n_2$。

9.16 为了研究伯利兹（Belize）在 1970—1992 年间的人口增长率，马克杰（Mukherjee）等人估计了如下模型②：

模型Ⅰ：$\widehat{\ln(\text{Pop})_t} = 4.73 + 0.024t$

$$t = (781.25)\quad(54.71)$$

模型Ⅱ：$\widehat{\ln(\text{Pop})_t} = 4.77 + 0.015\,t - 0.075D_t + 0.011\,(D_t\,t)$

$$t = (2\,477.92)\quad(34.01)\quad(-17.03)\quad(25.54)$$

其中 Pop=以百万计的人口数量；t=趋势变量；D_t 对 1978 年开始的观测取值 1，对此前的观测取值 0；ln 表示自然对数。

a. 在模型Ⅰ中，伯利兹人口在样本期的增长率是多少？

b. 1978 年之前和之后的人口增长率在统计上不同吗？你如何知道？若不同，1972—1977 年和 1978—1992 年期间的增长率各为多少？

实证分析题

9.17 利用表 9—8 中给出的数据，检验 1958 年第Ⅳ季度至 1966 年第Ⅲ季度和 1966 年第Ⅳ季度至 1971 年第Ⅱ季度两个子区间误差方差相同的假设。

9.18 利用第 8 章所讨论的方法，比较无约束和有约束回归（9.7.3）和（9.7.4）；即检验所

---

① 回归结果中所用数据取自于 N. M. Meltz，"Interstate and interprovincial Differences in Union Density," *Industrial Relations*，vol. 28，no. 2，1989，pp. 142-158。

② Chandan Mukherjee，Howard White，and Marc Wuyts，*Econometrics and Data Analysis for Developing Countries*，Routledge，London，1998，pp. 372-375. 符号有所变化。

施加约束的有效性。

9.19 在本章讨论的美国储蓄—收入回归（9.5.4）中，假设虚拟变量的取值不再是1和0，而是用虚拟变量 $Z_i=a+bD_i$，其中 $D_i=1$ 和0，$a=2$，$b=3$。比较你的结论。

9.20 继续储蓄—收入回归（9.5.4），假设你让 D_i 对第二区间的观测取值0，而对第一区间的观测取值1。方程（9.5.4）所示的结论有何变化？

9.21 利用表9—2中给出的数据，考虑如下模型：

$$\ln \text{Savings}_i=\beta_1+\beta_2 \ln \text{Income}_i+\beta_3 \ln D_i+u_i$$

其中 ln 表示自然对数，D_i 对1970—1981年取值1，对1982—1995年取值10。

a. 如此确定虚拟变量的根据是什么？

b. 估计上述模型并解释你的结论。

c. 两个子期间储蓄函数的截距值是多少？你如何解释它们？

9.22 参考表9—3给出的厨具销售季度数据，并考虑如下模型：

$$\text{Sales}_i=\alpha_1+\alpha_2 D_{2i}+\alpha_3 D_{3i}+\alpha_4 D_{4i}+u_i$$

其中 D 是第Ⅱ季度至第Ⅳ季度取值1和0的虚拟变量。

a. 分别对洗碗机、污物碾碎机和洗衣机估计上述模型。

b. 你如何解释估计的斜率系数？

c. 你如何应用所估计的 α 对各个厨具销售数据除去季节变化？

9.23 增加耐用品支出这个回归元后重新估计习题9.22中的模型。

a. 这两道题的回归结果有所不同吗？如果有，什么因素能解释这个差别？

b. 如果耐用品支出的数据中也存在季节性，你如何解释它？

9.24 表9—9给出了美国1916—2004年四年一次的总统选举数据。①

表9—9　　美国总统选举：1916—2004年

观测	Year	V	W	D	G	I	N	P
1	1916	0.516 8	0	1	2.229	1	3	4.252
2	1920	0.361 2	1	0	−11.46	1	5	16.535
3	1924	0.417 6	0	−1	−3.872	−1	10	5.161
4	1928	0.411 8	0	0	4.623	−1	7	0.183
5	1932	0.591 6	0	−1	−14.9	−1	4	7.069
6	1936	0.624 6	0	1	11.921	1	9	2.362
7	1940	0.55	0	1	3.708	1	8	0.028
8	1944	0.537 7	1	1	4.119	1	14	5.678
9	1948	0.523 7	1	1	1.849	1	5	8.722
10	1952	0.446	0	0	0.627	1	6	2.288
11	1956	0.422 4	0	−1	−1.527	−1	5	1.936
12	1960	0.500 9	0	0	0.114	−1	5	1.932
13	1964	0.613 4	0	1	5.054	1	10	1.247
14	1968	0.496	0	0	4.836	1	7	3.215

① 这些数据最初由耶鲁大学已预测总统选举结果多年的雷·费尔（Ray Fair）编制。我们复制于 Samprit Chatterjee, Ali S. Hadi, and Bertram Price, *Regression Analysis by Example*, 3rd ed., John Wiley & Sons, New York, 2000, pp. 150-151，并且更新自 http://fairmodel.econ.yale.edu/rayfair/pdf/2006CHTM.HTM。

续前表

观测	Year	V	W	D	G	I	N	P
15	1972	0.382 1	0	−1	6.278	−1	4	4.766
16	1976	0.510 5	0	0	3.663	−1	4	7.657
17	1980	0.447	0	1	−3.789	1	5	8.093
18	1984	0.408 3	0	−1	5.387	−1	7	5.403
19	1988	0.461	0	0	2.068	−1	6	3.272
20	1992	0.534 5	0	−1	2.293	−1	1	3.692
21	1996	0.547 4	0	1	2.918	1	3	2.268
22	2000	0.502 65	0	0	1.219	1	8	1.605
23	2004	0.512 33	0	1	2.69	−1	1	2.325

注：Year 表示选举年份。

V 表示两党总统投票中民主党所占的份额。

W 为指标变量，对 1920 年、1944 年和 1948 年的选举取值 1，其他取值 0。

D 为指标变量，1 表示民主党在任总统参加竞选，−1 表示共和党在任总统参加竞选，其他情况下为 0。

G 表示选举年份前三个季度真实人均 GDP 的增长率。

I 为指标变量，1 表示选举时在位总统为民主党成员，−1 表示选举时在位总统为共和党成员。

N 表示现任政府在前 15 个季度中真实人均 GDP 增长率超过 3.2%的季度数。

P 表示现任政府前 15 个季度 GDP 缩减指数增长率的绝对值。

a. 利用表 9—9 中给出的数据，提出一个合适模型来预测民主党在两党总统投票中所占的份额。

b. 你如何用这个模型去预测总统选举的结果？

c. 查特吉（Chatterjee）等人建议考虑如下实验模型预测总统选举：

$$V=\beta_0+\beta_1 I+\beta_2 D+\beta_3 W+\beta_4(GI)+\beta_5 P+\beta_6 N+u$$

估计这个模型，并评论其结果与你所选用模型得到的结果之间的关系。

9.25 参考回归（9.6.4）。检验平均小时收入随受教育水平的增长率因性别和种族不同而不同。（提示：使用多个虚拟变量。）

9.26 参考回归（9.3.1）。为了发现性别和居住地这两个虚拟变量之间是否存在某种交互作用，你该如何修正此模型？给出基于此模型的结果，并与方程（9.3.1）中给出的结果相比较。

9.27 在模型 $Y_i=\beta_1+\beta_2 D_i+u_i$ 中，令 D_i 对前 40 个观测取值 0，而对其余 60 个观测取值 1。已知 u_i 的均值为 0，方差为 100。这两个观测集的均值和方差各为多少？①

9.28 参考本章讨论的美国储蓄—收入回归。与方程（9.5.1）不同，考虑如下模型：

$$\ln Y_t=\beta_1+\beta_2 D_t+\beta_3 X_t+\beta_4(D_t X_t)+u_t$$

其中 Y 为储蓄，X 为收入。

a. 估计上述模型，并与方程（9.5.4）的结论相比较。哪个模型更好？

b. 你如何解释此模型中虚拟变量的系数？

c. 如我们在有关异方差性的章节中将看到的那样，对因变量取对数常常会减小数据中的异方差性。分两个期间将 Y 的对数对 X 做回归，看本例中是否如此？并看一下两个期间的误差方差在

① 这个例子选自 Peter Kennedy, *A Guide to Econometrics*, 4th ed., MIT Press, Cambridge, Mass., 1998, p. 347。

统计上是否相同。若相同，则可以按照本章中给出的方法将数据混合，再用邹至庄检验。

9.29 参考印度工人样本（9.12节）和表9—7中的数据。[①] 提醒变量定义如下：

WI=以卢比度量的周收入；

Age=年龄；

$D_{SEX}=1$，表示男性工人；=0，表示女性工人；

$DE_2=1$，表示一个工人受过初等教育；

$DE_3=1$，表示一个工人受过中等教育；

$DE_4=1$，表示一个工人受过高等教育；

$DPT=1$，表示一个工人拥有一份永久性工作；=0，表示工作是暂时性的。

参照组是没有受过初等教育并拥有暂时性工作的女性工人。

9.12节使用了教育程度虚拟变量（DE_2、DE_3 和 DE_4）与性别虚拟变量（D_{sex}）的交互项。如果我们使用教育程度虚拟变量与工人工作性质虚拟变量（DPT）的交互项，结果会怎么样？

a. 估计 ln WI 的模型，其中包括年龄、性别、教育程度虚拟变量，以及三个新引入的交互项：$DE_2\times DPT$、$DE_3\times DPT$ 和 $DE_4\times DPT$。这些新引入的交互项看上去有显著的交互影响吗？

b. 受过初等教育的工人与没有受过初等教育的工人存在明显差异吗？根据教育程度虚拟变量和交互项对此进行评价，并解释你的结果。受过中等教育的工人与没有受过初等教育的工人差别如何？受过高等教育的工人与没有受过初等教育的工人的差别又将如何？

c. 现在从模型中去掉教育虚拟变量，并对结果进行评价。交互项的显著性有所改变吗？

附录 9A 含虚拟回归元的半对数回归

我们在9.10节注意到，在如下形式的模型中

$$\ln Y_i=\beta_1+\beta_2 D_i \tag{1}$$

对于取值1或0的虚拟变量，Y 的相对变化（即半弹性）可如下得到：（β_2 估计值的反对数−1）×100，即

$$(e^{\beta_2}-1)\times 100 \tag{2}$$

证明如下：由于对数和指数互为反函数，所以我们可以把方程（1）写成

$$\ln Y_i=\beta_1+\ln(e^{\beta_2 D_i}) \tag{3}$$

现在，当 $D=0$ 时，$e^{\beta_2 D_i}=1$，当 $D=1$ 时，$e^{\beta_2 D_i}=e^{\beta_2}$。因此，从状态0到状态1，$\ln Y_i$ 变化了 $(e^{\beta_2}-1)$。但一个变量对数的变化只是相对变化，乘以100后就得到百分数变化。因此百分数变化就如所要证明的那样为 $(e^{\beta_2}-1)\times 100$。（注：$\ln_e e=1$，即以e为底e的对数等于1，就像以10为底10的对数等于1一样。记住，以e为底的对数被称为自然对数，而以10为底的对数被称为常用对数。）

① 数据来自 Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries*, Routledge Press, London, 1998。

第2篇

放松经典模型的假定

在第 1 篇里，我们详细考虑了经典正态线性回归模型，并说明了如何用它来处理估计与假设检验这两个统计推断问题，以及预测问题。但必须牢记，这个模型建立在一些简单化了的假定的基础之上。这些假定是：

假定 1：回归模型是参数的线性函数。

假定 2：回归元 X 的值是固定的，或者 X 值独立于误差项。这里，这个假定就意味着我们要求 u_i 与每个 X 变量之间的协方差为零。

假定 3：给定 X，干扰项 u_i 的均值为零。

假定 4：给定 X，干扰项 u_i 的方差恒定不变。

假定 5：给定 X，干扰项之间不存在自相关或序列相关。

假定 6：观测次数 n 必须大于待估计的参数个数。

假定 7：X 变量的取值必须有足够的变异。

在第 1 篇里我们还引入了如下 3 个假定：

假定 8：X 变量之间不存在准确的线性关系。

假定 9：模型被正确设定，即不存在设定偏误。

假定 10：随机项（干扰项）u_i 是正态分布的。

在继续讲下去之前，让我们指出，在大多数教科书中，没有列出这 10 个假定。例如假定 6 和假定 7 通常被认为是理所当然的，因而没有明确列出。我们之所以把它们明确列出，因为区分普通最小二乘法（OLS）具备一些理想的统计性质（比如 BLUE）所需要的假定与 OLS 可用所需要的条件非常有意义。比如，即使不满足假定 7，OLS 估计量仍是 BLUE。但在这种情况下，OLS 估计量的标准误相对其系数而言较大（即 t 比率相对较小），以致难以评价一个或多个回归元对解释平方和的贡献。

如韦瑟里尔（Wetherill）指出的那样，在应用经典线性回归模型时，实际上有两类主要问题：（1）关于模型设定和干扰项 u_i 的问题；（2）关于对数据的假定问题。[①] 假定 1、2、3、4、5、9 和 10 属于第一类，假定 6、7 和 8 则属于第二类。此外，诸如异常观测（不常见或非典型观测）和数据中的测量误差等数据问题也属于第二类。

与干扰项和模型设定的假定有关的问题主要有三个：（1）要偏离一个具体的假定多远才会导致不容忽视的差别？比如，如果 u_i 不完全是正态分布的，那么，在认为 OLS 估计量的 BLUE 性质遭到破坏之前，我们能容忍多大程度对正态性假定的偏离呢？（2）在一个具体的问题中，我们何以发现某个假定是否被破坏呢？因而，在一个具体的应用中，我们如何弄清

---

① G. Barrie Wetherill, *Regression Analysis with Applications*, Chapman and Hall, New York, 1986, pp. 14-15.

楚干扰项是否是正态分布的呢？我们已经讨论过正态性的**安德森-达琳 A^2 检验和雅克-贝拉检验**。(3) 如果某些假定是错误的，我们又能采取哪些补救措施呢？比如，如果在一个应用中发现同方差假定是错误的，那我们该怎么办？

对数据的假定，我们也遇到类似问题：(1) 一个特定的问题有多严重？比如，多重共线性问题严重到非常难以进行估计和推断的程度了吗？(2) 我们如何弄清楚数据问题的严重性？比如，我们如何判断包含或去掉一些或许异常的观测是否会导致分析上的巨大差别？(3) 某些数据问题能轻而易举地得到解决吗？比如，我们能否在原始数据中弄清楚测量误差的来源？

遗憾的是，我们无法对所有这些问题都给出令人满意的回答。在接下来的第 2 篇里，我们将更深入地分析某些假定，但不是对所有假定都进行全面的探讨。具体而言，我们将不再深究假定 2、3 和 10。理由如下：

假定 2：固定回归元与随机回归元

记得我们回归分析所依据的假定是，回归元是非随机的，并假定在重复抽样中取固定不变的数值。采用这种做法有一个很好的理由。就像在第 1 章中指出的那样，经济学家与物理学家不一样，他们通常无法控制他们所用的数据。经济学家常常使用二手数据，即诸如政府和私人机构搜集来的数据。因此，实际可行的策略是，就算解释变量的值本质上是随机的，但就要分析的问题而言，可以假定它们是给定的。因而，回归分析的结果是以这些给定的解释变量值为条件的。

但假如我们不能把这些 X 看成是非随机的或固定的，这就是**随机回归元**（stochastic regressors）情形。这种情况相当复杂。根据假定，u_i 是随机的，如果 X 也是随机的，那我们就必须明确 X 和 u_i 是如何分布的。如果我们愿意做假定 2（即尽管是随机的，但其分布独立于 u_i，或至少与 u_i 不相关），那么实际上我们可以继续把 X 看成非随机的。如克曼塔（Kmenta）所说：

> 因此，放弃 X 的非随机假定，以其取代 X 随机但独立于 $[u]$ 的假定，不至于改变最小二乘估计的优良性质与可行性。[①]

因此，直至我们在第 4 篇讨论联立方程之前，我们都将保留假定 2。[②] 此外，在第 13 章将对非随机回归元进行简要的讨论。

---

① Jan Kmenta, *Elements of Econometrics*, 2d ed., Macmillan, New York, 1986, p. 338.（原文中加重语句。）

② 这里或许要指出一个技术性要点。不使用 X 和 u 独立这个较强的假定，我们或许可以使用 X 变量值与 u 同期（即在同一时点）不相关这个更弱的假定。此时 OLS 估计量可能有偏误，但是一致的，即随着样本容量无限增大，估计量收敛于其真值。而如果 X 和 u 同期相关，则 OLS 估计量既有偏误，又是不一致的。我们在第 17 章将说明，有时在这种情况下，如何用**工具变量法**来求一致估计量。

假定 3：u_i 的均值为零

记得在 k 变量线性回归模型中：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{1}$$

让我们现在假定

$$E(u_i \mid X_{2i}, X_{3i}, \cdots, X_{ki}) = w \tag{2}$$

其中 w 是一个常数；注意在标准模型中 $w=0$，但现在我们令它等于任意常数。

对方程（1）取条件期望，我们便得到

$$\begin{aligned} E(Y_i \mid X_{2i}, X_{3i}, \cdots, X_{ki}) &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + w \\ &= (\beta_1 + w) + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} \\ &= \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} \end{aligned} \tag{3}$$

其中 $\alpha = \beta_1 + w$，而且在取期望时，应该注意到 X 被视为常数。（为什么？）

因此，如果假定 3 不满足，我们将看到我们无法估计原来的截距项 β_1；我们得到的是包含了 β_1 和 $E(u_i)=w$ 的 α。简言之，我们得到的 β_1 的估计值是有偏误的。

但正如我们在许多场合指出的那样，在许多实际情形中，截距项 β_1 无关紧要；斜率系数是更有意义的结果，而即使假定 3 不满足，斜率系数也不受影响。[①] 此外，在许多应用中，截距项并无实质性意义。

假定 10：u 的正态性

如果我们的目的仅在于估计，则此假定并不是非有不可。正如在第 3 章曾指出的那样，无论 u_i 是否正态分布，OLS 估计量都是 BLUE。但有了正态性假定之后，我们就能够证明回归系数的 OLS 估计量服从正态分布，即 $(n-k)\hat{\sigma}^2/\sigma^2$ 服从 χ^2 分布，而研究者就可以无论样本容量是大是小都能利用 t 检验和 F 检验对各种统计假设进行检验。

但如果 u_i 不是正态分布的会怎么样呢？那我们就要依赖对中心极限定理的如下推广；记得我们曾用中心极限定理来说明正态性假定的合理性：

> 如果干扰项 $[u_i]$ 是独立同分布的，且均值为 0 和［不变的］方差为 σ^2，而且如果解释变量在重复抽样中保持不变，那么［O］LS 系数估计量就是渐近正态分布的，且均值等于相应的 β。[②]

---

① 此命题仅当对每个 i 都有 $E(u_i)=w$ 时才是正确的，指出这一点非常重要。不过，如果 $E(u_i)=w_i$，即对每个 i，$E(u_i)$ 是彼此不同的常数，那么偏斜率系数就是有偏误和不一致的。此时假定 3 是否成立将至关重要。证明和更多细节，参见 Peter Schmidt, *Econometrics*, Marcel Dekker, New York, 1976, pp. 36-39。

② Henri Theil, *Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, NJ, 1978, p. 240. 必须指出，对于这个结论而言，X 值固定和 σ^2 为常数的假定非常关键。

因此，通常的检验程序——t 检验和 F 检验——仍然是渐近有效的，也就是说，在大样本而非小样本或有限样本中是有效的。

如果干扰项不是正态分布的，则 OLS 估计量（在同方差和固定 X 的假定下）仍渐近服从正态分布，这一事实难以抚慰那些通常难以获得昂贵的大样本数据的实际经济学家。因此，对假设检验和预测而言，正态性假定就极为重要。于是，考虑到估计和假设检验这两个方面的问题，而且给定在绝大多数经济分析中小样本情形是常规情形而非例外，我们将继续使用正态性假定。[①]（但参见第 13 章 13.12 节。）

当然，这就意味着，当我们在使用小样本时，我们必须明确地检验正态性假定。我们已经考虑过**安德森-达琳正态性检验**和**雅克-贝拉正态性检验**。我们强烈建议读者对回归残差进行正态性检验。记住，在有限样本中，没有正态性假定，通常的 t 和 F 统计量就不服从 t 和 F 分布。

现在就剩下假定 1、4、5、6、7、8 和 9。假定 6、7 和 8 彼此密切相关，并将在共线性一章（第 10 章）中进行讨论。假定 4 在异方差性一章（第 11 章）中讨论。假定 5 在自相关一章（第 12 章）中讨论。假定 9 在模型设定和诊断检验一章（第 13 章）中讨论。由于其特殊性质和数学上的需要，假定 1 作为第 3 篇的一个专题（第 14 章）进行讨论。

为便于教学，我们在讲解这些内容的每一章都采用一个共同的格式，即（1）明确问题的性质；（2）分析它的影响；（3）提出侦察它的方法；（4）考虑补救措施，从而可能得到具有第 1 篇中讨论的优良统计性质的估计量。

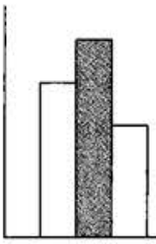
现在有必要给出一点告诫。前面曾指出，对于因违背经典线性回归模型而导致的所有问题，令人满意的答案并不存在。而且，对于一个特定的问题，解决方法可能不止一个，而且我们并不清楚哪个方法最好。此外，在一个具体应用中，对经典线性回归模型的违背可能不止一个方面。因此，设定偏误、多重共线性和异方差性可能同时存在，所以没有一个万能的检验能同时解决所有的问题。[②] 最后，在一个时期曾广为使用的某种检验方法，由于后来有人发现它原先的弊端而不一定继续流行。但这正是科学进步的过程。计量经济学也不例外。

---

① 顺便指出，文献中常在**稳健估计**（robust estimation）的标题下讨论偏离正态性假定的后果及相关论题，而这个专题超出了本书论述的范围。

② 这并非缺少尝试。见 A. K. Bera and C. M. Jarque, "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence," *Economic Letters*, vol. 7, 1981, pp. 313-318。

第 10 章 多重共线性：回归元相关会怎么样？



名词“多重共线性”在计量经济学教科书中和在应用文献中被误用的情况之多，再没有其他名词能比得上。我们的许多解释变量本是高度共线性的，而这就是生活。毫无疑问，有实验的设计 $\mathbf{X}'\mathbf{X}$（即数据矩阵），会远远优于自然实验已为我们提供的设计（即手中的样本）。但怨天尤人完全无济于事。对一个坏的设计采取就事论事的治疗方法，诸如逐步回归（stepwise regression）或脊回归（ridge regression），可能招致灾难性的后果。正确的做法是，宁可接受事实：我们的非实验数据［即不是从经过设计的实验而得到的数据］有时不能对我们感兴趣的参数提供多少信息。①

经典线性回归模型（CLRM）的假定 8 说，包含在回归模型中的各个回归元之间无**多重共线性**（multicollionearity）。本章中，为了寻找下述问题的答案，我们对多重共线性假定进行严格的分析：

1. 多重共线性的性质是什么？
2. 多重共线性真的是问题吗？
3. 它会引起一些什么实际后果？
4. 怎样去发现它？
5. 为了缓解多重共线性问题，能采取哪些补救措施？

我们在本章中还讨论 CLRM 的假定 6 和假定 7。假定 6 要求样本观测次数必须大于回归元的个数，假定 7 要求回归元的取值必须有足够的变异，它们与不存在多

① Edward E. Leamer, “Model Choice and Specification Analysis,” in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, vol. 1, North Holland Publishing Company, Amsterdam, 1983, pp. 300-301.

重共线性的假定有着内在的联系。戈德伯格已经将假定 6 命名为**微数缺测性**（micronumerosity）问题[①]，也就是小样本容量的问题。

10.1 多重共线性的性质

多重共线性一词由弗里希（Ragnar Frisch）引入。[②] 它原先的含义是指一个回归模型中的一些或全部解释变量之间存在一种“完全”或准确的线性关系。[③] 对解释变量 X_1，X_2，…，X_k（其中，为了把截距项考虑进来，在一切观测中取 $X_1=1$）这 k 个变量，如果满足如下条件，我们说它存在一个准确的线性关系：

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \tag{10.1.1}$$

其中 λ_1，λ_2，…，λ_k 为常数，但不同时为零。[④]

然而，现在用多重共线性一词，有更广泛的含义，既包括方程（10.1.1）所示的完全多重共线性情形，还包括 X 变量之间彼此相关，但又不完全相关的如下情形[⑤]：

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + v_i = 0 \tag{10.1.2}$$

其中 v_i 是随机误差项。

为了区分完全（perfect）和不完全（less than perfect）多重共线性，暂且假定 $\lambda_2 \neq 0$。于是，方程（10.1.1）可写为：

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} \tag{10.1.3}$$

这就表明 X_2 与其他变量有准确的线性关系，或者它能从其他 X 变量的线性组合推出。这时，变量 X_2 与方程（10.1.3）右端的线性组合之间的相关系数必定是 1。

类似地，如果 $\lambda_2 \neq 0$，则方程（10.1.2）可写为：

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i \tag{10.1.4}$$

这表明 X_2 不是其他 X 的一个准确的线性组合，因为它还取决于随机误差项 v_i。

作为一个数值例子，考虑如下假设数据：

---

① 见他写的 *A Course in Econometrics*，Harvard University Press，Cambridge，Mass.，1991，p. 249。

② Ragnar Frisch，*Statistical Confluence Analysis by Means of Complete Regression Systems*，Institute of Economics，Oslo University，publ. no. 5，1934.

③ 严格地说，多重共线性是指存在多于一个准确的线性关系。而共线性，则指存在有单个线性关系。但在实践中，这种区分很少得到遵循，从而多重共线性兼指两种情形。

④ 在实践中要得到一个样本，它的回归元的数值是按这种方式联系起来的，这样的机会确实是很小的，除非有意设计。比如，当观测次数小于回归元个数时，或者研究者陷入了第 9 章讨论的“虚拟变量陷阱”时，就会出现方程（10.1.1）这种关系。见习题 10.2。

⑤ 如果只有两个解释变量，交互相关就可由零阶或简单相关系数来度量。但若有多于两个 X 变量，则交互相关由偏相关系数或一个 X 变量对其余所有 X 变量多元回归的相关系数 R 来度量。

X_2	X_3	X_3^*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

很明显，$X_{3i}=5X_{2i}$。因此 X_2 与 X_3 之间存在完全共线性并且相关系数 r_{23} 是 1。变量 X_3^* 无非是把从随机数表生成的数字 2，0，7，9，2 加到 X_3 上而得到的。但现在 X_2 与 X_3^* 之间不再有完全共线性。然而，它们之间的相关系数是 0.995 9，所以两者是高度相关的。

以上对多重共线性的代数处理方法，可通过巴伦坦图（回顾图 3—8，在图 10—1 中重新给出）作简明的描述。在该图中圆圈 Y，X_2 和 X_3 分别代表 Y（因变量）与 X_2 和 X_3（解释变量）的变异。共线性的程度可用 X_2 和 X_3 两圆圈的重叠程度来衡量。在图 10—1a 中 X_2 与 X_3 无重叠，因而无共线性。从图 10—1b 到图 10—1e 共线性程度由低渐高——X_2 与 X_3 的重叠部分越来越大（即阴影部分面积越来越大），共线性程度也越来越高。极端时，X_2 与 X_3 完全重合（或者 X_2 完全落在 X_3 内，或相反），就出现了完全共线性。

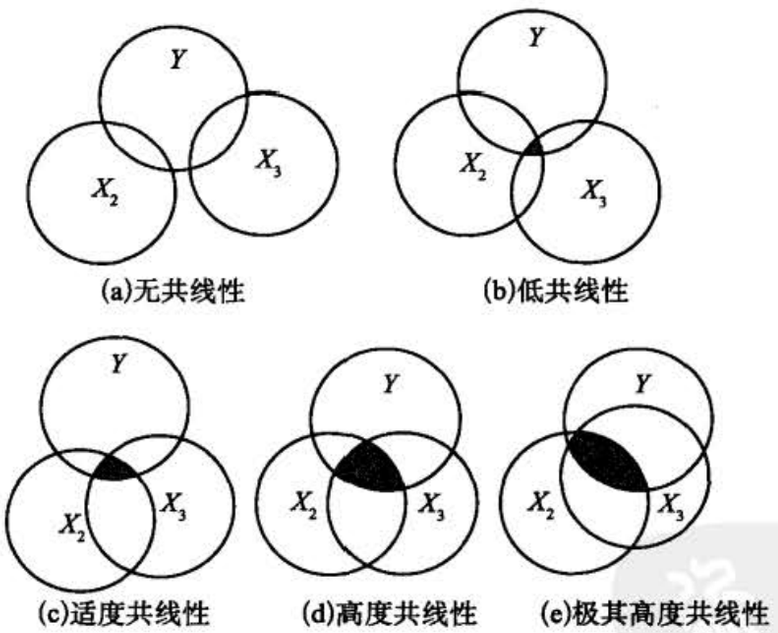


图 10—1　多重共线性的巴伦坦图

顺便指出，我们定义的多重共线性仅对 X 变量之间的线性关系而言。此外，还可能有它们之间的非线性关系。例如，考虑以下回归模型：

$$Y_i=\beta_0+\beta_1X_i+\beta_2X_i^2+\beta_3X_i^3+u_i \tag{10.1.5}$$

其中，比如说，Y＝生产总成本，而 X＝产出。变量 X_i^2（产出的平方）和 X_i^3（产出的

立方）显然与 X_i 有函数关系，但这种关系是非线性的。因此，严格地说，像方程（10.1.5）这样的模型并不违反无多重共线性假定。然而，在具体的应用中，通常测算的相关系数将表明 X_i、X_i^2 和 X_i^3 是高度相关的。如我们即将表明的，这种情形将使得我们难以较准确地（即以较小标准误）估计方程（10.1.5）的参数。

为什么经典线性回归模型要假定 X 之间无多重共线性呢？可以这样去理解：**如果多重共线性是完全的，如方程（10.1.1）所示，则 X 变量的回归系数将是不确定的，并且它们的标准误为无穷大。如果多重共线性是完全的，像方程（10.1.2）那样，那么，虽然回归系数可以确定，却有较大的标准误（相对于系数本身来说），也就是说，系数不能以很高的精度或准确度加以估计。**随后的几节将对此作出证明。

多重共线性有多种来源。按蒙哥马利（Montgomery）和佩克（Peck）的提法，多重共线性可能由以下因素导致。[①]

1. 数据采集所用的方法。例如，抽样限于总体中回归元取值的一个有限范围。

2. 模型或从中取样的总体受到约束。例如，在做电力消费对收入（X_2）和住房面积（X_3）的回归时，总体中有这样的一种有形的约束，即一般地说收入较高的家庭比收入较低的家庭有较大的住房。

3. 模型设定。例如在回归中添加多项式项，尤其当 X 变量的变化范围（极差）较小时。

4. 一个过度决定的模型。这种情况出现在模型的回归元个数大于观测次数时。例如，在医药研究中可能只有少数病人，但却要在他们身上收集大量的变量信息。

多重共线性的另外一个原因（特别是在时间序列数据中）可能是模型中所包含的回归元具有相同的时间趋势，即它们同时随着时间而增减。于是，在消费支出对收入、财富和人口的回归中，回归元收入、财富和人口可能都以多少有些一致的速度递增，从而导致这些变量之间的共线性。

10.2 出现完全多重共线性时的估计问题

前面说过，在完全多重共线性的情形中，回归系数是不确定的，并且其标准误是无穷大的。这一事实容易通过三变量回归模型加以说明。利用离差形式把三个变量都表示为偏离它们各自样本均值的离差，我们就能把三变量回归模型写为：

① Douglas Montgomery and Elizabeth Peck, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York, 1982, pp. 289-290. See also R. L. Mason, R. F. Gunst, and J. T. Webster, "Regression Analysis and Problems of Multicollinearity," *Communications in Statistics A*, vol. 4, no. 3, 1975, pp. 277-292; R. F. Gunst, and R. L. Mason, "Advantages of Examining Multicollinearities in Regression Analysis," *Biometrics*, vol. 33, 1977, pp. 249-260.

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \tag{10.2.1}$$

现在从第 7 章我们得到：

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{7.4.7}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{7.4.8}$$

假定 $X_{3i} = \lambda X_{2i}$，其中 λ 是非零常数（比如 2、4、1.8 等）。以此代入方程（7.4.7）可得：

$$\begin{aligned} \hat{\beta}_2 &= \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} \\ &= \frac{0}{0} \end{aligned} \tag{10.2.2}$$

这是一个不定式。读者容易验证 $\hat{\beta}_3$ 也是不确定的。[①]

我们为什么会得到像方程（10.2.2）那样的结果呢？回想一下 $\hat{\beta}_2$ 的意义：它是在保持 X_3 不变的情况下，当 X_2 每改变一单位时 Y 的平均值的变化率。但如果 X_3 和 X_2 是完全共线性的，就没有任何方法能保持 X_3 不变：随着 X_2 改变，X_3 也按一定的倍数因子 λ 改变。这就意味着没有任何方法能从所给的样本中把 X_2 和 X_3 的各自影响分解开来。从实际方面考虑，X_2 和 X_3 是不可区分的。在应用计量经济学中，我们的宗旨就是要把每个 X 对因变量的偏影响分离开来，所以这个问题是最具破坏性的。

从另一个角度看这个问题，把 $X_{3i} = \lambda X_{2i}$ 代入方程（10.2.1）得到以下方程［另见（7.1.12）］：

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 (\lambda x_{2i}) + \hat{u}_i \\ &= (\hat{\beta}_2 + \lambda \hat{\beta}_3) x_{2i} + \hat{u}_i \\ &= \hat{\alpha} x_{2i} + \hat{u}_i \end{aligned} \tag{10.2.3}$$

其中：

$$\hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3 \tag{10.2.4}$$

对方程（10.2.3）应用平常的 OLS 公式可得：

$$\hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3 = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2} \tag{10.2.5}$$

因此，虽然 $\hat{\alpha}$ 可以唯一地估计出来，却无法唯一地估计 β_2 和 β_3；数学上，

$$\hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3 \tag{10.2.6}$$

---

① 说明此问题的另一方法是：按定义 X_2 和 X_3 的相关系数 $r_{23} = \sum x_{2i} x_{3i} / \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}$。如果 $r_{23}^2 = 1$，即 X_2 和 X_3 完全共线性，则方程（7.4.7）的分母将为零，从而不可能估计出 β_2（或 β_3）。

是一个方程有两个未知数（注意 λ 是给定的），对给定的 $\hat{\alpha}$ 和 λ 值，方程（10.2.6）便有无穷多个解。为了把这个概念说得更具体，令 $\hat{\alpha}=0.8$ 和 $\lambda=2$。这样就得到：

$$0.8=\hat{\beta}_2+2\hat{\beta}_3 \tag{10.2.7}$$

或者：

$$\hat{\beta}_2=0.8-2\hat{\beta}_3 \tag{10.2.8}$$

现在任意选一个 $\hat{\beta}_3$ 的值，将得到 $\hat{\beta}_2$ 的一个解。选择另一个 $\hat{\beta}_3$ 的值又得到 $\hat{\beta}_2$ 的另一个解。不管你怎样尝试，都没有 $\hat{\beta}_2$ 的唯一值。

以上讨论的要点在于：对于完全多重共线性情形，我们无法得到个别回归系数的唯一解。但应注意到，我们能够得到这些系数线性组合的唯一解。给定 λ，线性组合 $\beta_2+\lambda\beta_3$ 的唯一估计值是 α。[①]

顺便指出，对于完全多重共线性的情形，$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差和标准误都是无穷大（见习题 10.21）。

10.3 出现“高度”但“不完全”多重共线性时的估计问题

完全多重共线性情形只不过是一种极端的隐忧。通常，尤其是在涉及经济时间序列的数据中，X 变量之间并无准确的线性关系。拿方程（10.2.1）所给的离差形式的三变量模型来看，我们有的不是准确的多重共线性，而是：

$$x_{3i}=\lambda x_{2i}+v_i \tag{10.3.1}$$

其中 $\lambda\neq 0$ 并且 v_i 是具有性质 $\sum x_{2i}v_i=0$ 的随机误差项。（为什么?）

顺便提一下，图 10—1b 至图 10—1e 的巴伦坦图都代表不完全共线性的情形。

对于这种情形，回归系数 β_2 和 β_3 的估计是可能的。例如将方程（10.3.1）代入方程（7.4.7）得：

$$\hat{\beta}_2=\frac{\sum(y_ix_{2i})(\lambda^2\sum x_{2i}^2+\sum v_i^2)-(\lambda\sum y_ix_{2i}+\sum y_iv_i)(\lambda\sum x_{2i}^2)}{\sum x_{2i}^2(\lambda^2\sum x_{2i}^2+\sum v_i^2)-(\lambda\sum x_{2i}^2)^2} \tag{10.3.2}$$

其中利用了关系式 $\sum x_{2i}v_i=0$。对 $\hat{\beta}_3$ 也可推出类似的表达式。

现在，不同于方程（10.2.2），没有理由先验地认为方程（10.3.2）不可估计。当然，如果 v_i 充分地小，以至非常接近于零，则方程（10.3.1）表示几乎完全共线性。这时我们又将回到方程（10.2.2）的不确定情形。

① 在计量经济学文献中，称类似于 $\beta_2+\lambda\beta_3$ 的函数为**可估函数**（estimable function）。

10.4 多重共线性：是庸人自扰吗？多重共线性的理论后果

回想一下，如果经典模型的假定得到满足，则回归系数的 OLS 估计量是 BLUE（或 BUE，如果加上正态性假定）。而现在可以证明，即使多重共线性是非常高的，如近似多重共线性，那么，OLS 估计量仍保持 BLUE 性质。[①] 那么，大谈特谈多重共线性究竟为了什么？如克里斯托弗·阿肯（Christopher Achen）所说［并参考本章开头所引用利莫尔（Leamer）的话］：

> 初次接触方法论的学生有时担心他们的自变量有相关关系，即所谓多重共线性问题。但多重共线性并不违反回归假定。无偏的、一致的估计值仍将出现，并且对它们的标准误仍将有正确的估计。多重共线性的唯一影响是难于得到标准误较小的系数估计值。然而，仅有少量的观测次数时也会出现这种影响，就好比自变量的方差较小所造成的影响那样。（事实上，从理论的高度看，多重共线性、过少的观测次数以及过小的自变量方差，实质上是同一问题。）因此，"遇到多重共线性我该怎么办"这个问题无异于"如果我没有很多的观测值该怎么办"。统计上的答案是不存在的。[②]

为了彻底弄明白样本容量的重要性，戈德伯格构造**微数缺测性**一词以应对古怪的多音节名称：多重共线性。按照戈德伯格所说的，**准确的微数缺测性**（与准确的多重共线性相对照），是指样本容量 n 等于零的情形。这时，任何种类的估计都是不可能的。近似微数缺测性，则好比近似完全多重共线性，指观测次数刚刚超过待估计参数个数的情形。

利莫尔、阿肯和戈德伯格埋怨人们过少地注意样本大小的问题，而过多地注意多重共线性的问题。没有错，可惜的是，在应用研究工作中用到第二手资料（指由他人收集的数据资料，诸如由政府收集的 GNP 数据）的个人研究者，对样本数据的多少似乎无能为力，而只好"把多重共线性视为对经典线性回归的破坏，以正视估计问题"[③]。

第一，诚然，即使是近似多重共线性的情形，OLS 估计量仍然是无偏的。但无偏性是一种多维样本（multisample）或重复抽样的性质。意思是说，如果我们在 X 变量取固定值的情况下反复抽取样本，并对每一样本计算 OLS 估计量，那么，随着

---

① 由于近似多重共线性本身并不违背第 7 章所列的其他假定，所以 OLS 估计仍是 BLUE。

② Christopher H. Achen, *Interpreting and Using Regression*, Sage Publications, Beverly Hills, Calif., 1982, pp. 82-83.

③ Peter Kennedy, *A Guide to Econometrics*, 3d ed., The MIT Press, Cambridge, Mass., 1992, p. 177.

样本容量的增加，估计量样本值的均值将收敛于它们的真实总体值。

第二，说共线性并不破坏最小方差性质，也没错。在所有线性无偏估计量中，OLS 估计量有最小方差。也就是说，它们是有效的。但这并不意味着，在任一给定的样本中，一个 OLS 估计量的方差一定是小的（相对于估计量的值而言），这点我们即将予以说明。

第三，多重共线性本质上是一种样本（回归）现象。意思是说，即使在总体中 X 变量没有线性关系，但在具体获得的样本中仍可能存在线性关系：当我们设想一个理论或总体回归函数时，我们相信，所有包含在模型中的 X 变量对 Y 都有各自的独立影响。但有可能在任给的一个用以检验 PRF 的样本中，一些或全部 X 变量之间的共线性却是如此之高，以致我们无法区分它们对 Y 的各自影响。看来，是样本把我们难住了。尽管理论告诉我们，所有的 X 变量都重要，但我们的样本还是没有“富裕”到足以在分析中容纳全部 X 变量的境况。

作为说明，再考虑第 3 章的消费—收入例子（例 3.1）。经济学家从理论上推知除收入（Income）外，消费者的财富（Wealth）也是消费支出（Consumption）的重要决定因素。于是我们可以写成：

$$\text{Consumption}_i = \beta_1 + \beta_2 \text{Income}_i + \beta_3 \text{Wealth}_i + u_i$$

但可能出现这样的情形：当我们获得收入和财富的数据时，这两个变量可能高度相关（即使不是完全相关）：较富有的人们一般倾向于有较高的收入。因此，虽然理论上收入和财富都是解释消费支出行为的合理备选变量，但实际上（即在样本中）要分开收入和财富对消费支出的影响，也许是困难的。

理想地，为了分别评价财富和收入对消费支出的影响，我们需要对财富多而收入低以及财富少而收入高的人们进行足够多的样本观测（回顾假定 7）。虽然在横截面研究中（通过增加样本的方法）有可能做到这一点，但在加总时间序列操作中却非常难以实现。

考虑到所有这些理由，多重共线性虽不影响 OLS 估计量的 BLUE 性质，但这一点在实际上并没有什么值得令人宽慰的。我们必须意识到在任给的一个样本中会出现什么情况。这就是下节要讨论的议题。

10.5 多重共线性的实际后果

近似或高度多重共线性可能招致以下后果：

1. 虽然 OLS 估计量是 BLUE，但其方差和协方差偏大，故难以做出精确的估计。

2. 由于上述原因，置信区间要宽得多，以致接受“虚拟假设”（即真实总体系数为零的假设）更为容易。

3. 仍由于上述第一个原因，一个或多个系数的 t 比率倾向于统计上不显著。

4. 虽然一个或多个系数的 t 比率在统计意义上不显著，但总的拟合优度 R^2 仍可能非常之高。

5. OLS 估计量及其标准误对数据的微小变化也会非常敏感。

上述后果可解释如下。

□ OLS 估计量的方差与协方差偏大

方差和协方差偏大可从模型（10.2.1）所给的 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差和协方差公式看到：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)} \tag{7.4.12}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)} \tag{7.4.15}$$

$$\text{cov}(\hat{\beta}_2,\hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \tag{7.4.17}$$

其中 r_{23} 是 X_2 与 X_3 之间的相关系数。

从方程（7.4.12）和（7.4.15）显然可见，随着 r_{23} 趋于 1，即随着共线性的加剧，两估计量的方差也不断增加。在达到极限 $r_{23}=1$ 时，方差变为无穷大。同样，由方程（7.4.17）显然可见，随着 r_{23} 朝着 1 增大，两估计量之间的协方差在绝对值上也不断增大。[注：$\text{cov}(\hat{\beta}_2,\hat{\beta}_3) \equiv \text{cov}(\hat{\beta}_3,\hat{\beta}_2)$。]

方差和协方差增大的速度可由如下定义的**方差膨胀因子**（variance-inflating factor，VIF）看出：

$$\text{VIF} = \frac{1}{(1-r_{23}^2)} \tag{10.5.1}$$

VIF 表明，估计量的方差由于多重共线性的出现而膨胀（inflated）。随着 r_{23} 趋于 1，VIF 趋于无穷大，即随着共线性程度的增加，估计量的方差也增加，并且在达到极限时，它可以变到无穷大。还容易看到，如果 X_2 与 X_3 之间无共线性，VIF 将是 1。

利用这一定义，可将方程（7.4.12）和（7.4.15）表达为：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF} \tag{10.5.2}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2}\text{VIF} \tag{10.5.3}$$

从而表明 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差与 VIF 成正比关系。

为了对方差和协方差随 r_{23} 增加而增加的速度有所认识，表 10—1 对选定的 r_{23} 值计算方差和协方差。如表所示，r_{23} 的增加对 OLS 估计量的方差和协方差估计值有剧烈的影响。当 $r_{23}=0.50$ 时，$\text{var}(\hat{\beta}_2)$ 1.33 倍于 r_{23} 为零时的方差，但当 $r_{23}=0.95$

时，它将近 10 倍于没有共线性的情形。再看，当 r_{23} 从 0.95 增加到 0.995 时，估计的方差值跃升到无共线性情形的 100 倍。所估计的协方差也存在类似的剧烈影响。从图 10—2 中也能看出这一点。

表 10—1　　r_{23} 增加对 var($\hat{\beta}_2$) 和 cov($\hat{\beta}_2$,$\hat{\beta}_3$) 的影响

r_{23} 值 (1)	VIF (2)	var($\hat{\beta}_2$) (3)*	$\frac{\text{var}(\hat{\beta}_2)(r_{23\neq 0})}{\text{var}(\hat{\beta}_2)(r_{23=0})}$ (4)	cov ($\hat{\beta}_2$,$\hat{\beta}_3$) (5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2}=A$	—	0
0.50	1.33	$1.33\times A$	1.33	$0.67\times B$
0.70	1.96	$1.96\times A$	1.96	$1.37\times B$
0.80	2.78	$2.78\times A$	2.78	$2.22\times B$
0.90	5.26	$5.26\times A$	5.26	$4.73\times B$
0.95	10.26	$10.26\times A$	10.26	$9.74\times B$
0.97	16.92	$16.92\times A$	16.92	$16.41\times B$
0.99	50.25	$50.25\times A$	50.25	$49.75\times B$
0.995	100.00	$100.00\times A$	100.00	$99.50\times B$
0.999	500.00	$500.00\times A$	500.00	$499.50\times B$

注：$A=\frac{\sigma^2}{\sum x_{2i}^2}$；

$B=\frac{-\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$；

×＝倍（乘号）。

*要分析 r_{23} 增加对 var($\hat{\beta}_3$) 的影响，只需注意当 $r_{23}=0$ 时，$A=\sigma^2/\sum x_{3i}^2$，但方差和协方差的膨胀因子是一样的。

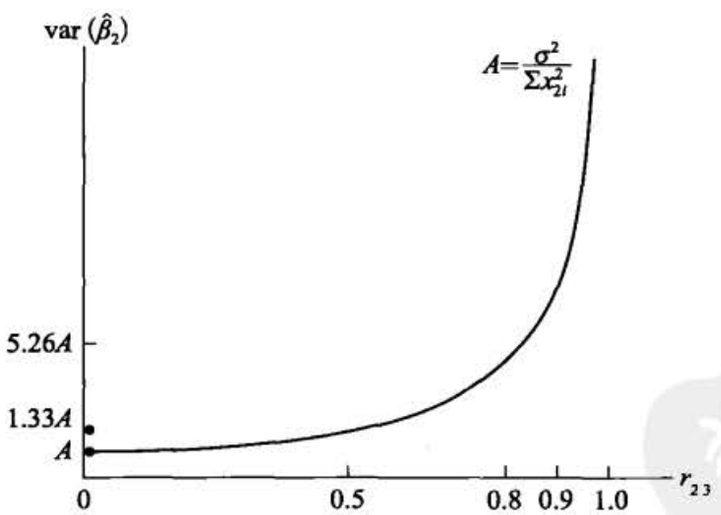


图 10—2　var($\hat{\beta}_2$) 作为 r_{23} 的一个函数的性态

刚才讨论的结果很容易推广到 k 变量模型。在这一模型中，第 k 个系数的方差可如方程（7.5.6）那样表示为

$$\text{var}(\hat{\beta}_j)=\frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right) \qquad (7.5.6)$$

其中 $\hat{\beta}_j$ 表示回归元 X_j 的（估计）偏回归系数；

R_j^2 表示 X_j 对其余 $k-2$ 个回归元进行回归的 R^2；（注：在 k 个变量的回归模型中有 $k-1$ 个回归元。）

$\sum x_j^2 = \sum (X_j - \overline{X}_j)^2$。

我们还可以把方程（7.5.6）写成

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \text{VIF}_j \tag{10.5.4}$$

如你从这个表达式所见，$\text{var}(\hat{\beta}_j)$ 与 σ^2 和 VIF 成正比，但与 $\sum x_j^2$ 成反比。因此，$\text{var}(\hat{\beta}_j)$ 的大小取决于三个部分：（1）σ^2；（2）VIF；和（3）$\sum x_j^2$。最后一个部分与经典模型的假定 8 相联系，它说明回归元的变异越大，在假定其他两个因素不变的情况下，该回归元系数的方差就越小，因此用它估计系数就越准确。

在进一步深入讨论之前，注意 VIF 的倒数被称为**容许度**（tolerance，TOL）。即

$$\text{TOL}_j = \frac{1}{\text{VIF}_j} = (1 - R_j^2) \tag{10.5.5}$$

当 $R_j^2 = 1$（即完全共线性）时，$\text{TOL}_j = 0$；当 $R_j^2 = 0$（即不存在共线性）时，$\text{TOL}_j = 1$。由于 VIF 和 TOL 之间有密切关系，所以可以将它们互换使用。

□ 更宽的置信区间

由于标准误偏大，有关总体参数的置信区间也随之变大。这可由表 10—2 看出，例如，当 $r_{23} = 0.95$ 时，β_2 的置信区间要比 $r_{23} = 0$ 时宽 $\sqrt{10.26}$ 倍或约 3 倍。

表 10—2　　增加共线性对 β_2 的 95%置信区间 "$\hat{\beta}_2 \pm 1.96\text{se}(\hat{\beta}_2)$" 的影响

r_{23} 值	β_2 的 95%置信区间
0.00	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{1.33}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.95	$\hat{\beta}_2 \pm 1.96\sqrt{10.26}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.995	$\hat{\beta}_2 \pm 1.96\sqrt{100}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.999	$\hat{\beta}_2 \pm 1.96\sqrt{500}\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$

注：为方便起见，假定 σ^2 已知，因此可用正态分布，从而用 1.96 作为正态分布下的 95%置信因子。与各 r_{23} 值相对应的标准误取自表 10—1。

因此，在高度多重共线性的情形中，样本数据可能与分歧很大的一些假设均无矛盾，这样就增加了接受错误假设（即犯第 II 类错误）的概率。

□ “不显著”的 t 比率

记得在检验虚拟假设（比方说）$\beta_2=0$ 时，我们使用 t 比率即 $\hat{\beta}_2/\text{se}(\hat{\beta}_2)$，并将估计的 t 值同从 t 表查出的 t 临界值相比。但我们已经看到，在高度共线性情形中，估计的标准误增加奇快，从而使 t 值迅速变小。因此，在此情形中，我们会越来越多地接受有关真实总体值为零的虚拟假设。[①]

□ R^2 值高而少且显著的 t 比率

考虑 k 变量线性回归模型：

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\cdots+\beta_kX_{ki}+u_i$$

如同我们刚才说过的，在高度共线性情形中，有可能会发现一个或多个偏斜率系数基于 t 检验不是个别统计显著的，然而这时 R^2 却高达（比如说）0.9 以上，从而根据 F 检验，可令人信服地拒绝 $\beta_2=\beta_3=\cdots=\beta_k=0$ 的假设。其实，这就是多重共线性的一个信号——不显著的 t 值却带有一个高的总 R^2 值（并因而有一个显著的 F 值）！

下节我们将阐释这个信号。不过鉴于我们在第 8 章中关于个别检验与联合检验的讨论，这种信号的出现并没有什么奇怪。你也许会想到，这里的真正问题在于估计量之间的协方差。如公式（7.4.17）所表明的那样，这些协方差是同回归元之间的相关性有关系的。

□ OLS 估计量及其标准误对数据微小变化的敏感性

只要多重共线性还不是完全的，就有可能估计出回归系数。然而，估计值及其标准误对数据中的哪怕是微小变化，也会非常敏感。

为了看清楚这一点，考虑表 10—3。根据这些数据，我们得到如下多元回归：

$$\begin{aligned}\hat{Y}_i&=1.193\,9+0.446\,3X_{2i}+0.003\,0X_{3i}\\&\quad(0.773\,7)\quad(0.184\,8)\quad(0.085\,1)\\t&=(1.543\,1)\quad(2.415\,1)\quad(0.035\,8)\qquad(10.5.6)\\R^2&=0.810\,1\qquad r_{23}=0.552\,3\\\text{cov}(\hat{\beta}_2,\hat{\beta}_3)&=-0.008\,68\qquad \text{df}=2\end{aligned}$$

表 10—3　　Y，X_2 和 X_3 的人为影响

Y	X_2	X_3
1	2	4
2	0	2
3	4	12

① 用置信区间表示，随着共线性程度的增大，$\beta_2=0$ 这个值将越来越多地落入接受域内。

续前表

Y	X_2	X_3
4	6	0
5	8	16

回归（10.5.6）表明，个别地看，没有一个回归系数在通常1%或5%的显著性水平上是显著的，尽管$\hat{\beta}_2$的单尾t检验在10%的显著性水平上是显著的。

再看表10—4，它和表10—3的差别仅在于：X_3的第3个值和第4个值互相对调了。现在由表10—4中的数据我们得到：

$$\begin{aligned} \hat{Y}_i &= 1.210\,8 + 0.401\,4X_{2i} + 0.027\,0X_{3i} \\ &\quad (0.748\,0)\ (0.272\,1)\quad (0.125\,2) \\ t &= (1.618\,7)\ (1.475\,2)\quad (0.215\,8) \qquad (10.5.7) \\ &\qquad R^2 = 0.814\,3 \qquad r_{23} = 0.828\,5 \\ &\qquad \operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.028\,2 \qquad \mathrm{df} = 2 \end{aligned}$$

表10—4　　Y，X_2和X_3的人为数据

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

数据微小变化的结果是，原先在10%显著性水平上为统计显著的$\hat{\beta}_2$，现在在该水平上也不再显著了。还可注意到方程（10.5.6）中的$\operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.008\,68$，而在方程（10.5.7）中它是$-0.028\,2$，增加了3倍以上。所有这些都可归因于增加了的多重共线性：在方程（10.5.6）中，$r_{23} = 0.552\,3$，而在方程（10.5.7）中，它是0.828 5。类似地，$\hat{\beta}_2$和$\hat{\beta}_3$的标准误都在增大。这正是共线性的通常象征。

我们前面说过，在出现高度共线性时，我们无法精确估计个别回归系数，但可以较精确地估计这些系数的某些线性组合。这一事实可从回归（10.5.6）和（10.5.7）中得到证实。在第一个回归中，两个偏斜率系数之和为0.449 3，而在第二个回归中，此和为0.428 4，基本一致。不仅如此，它们的标准差也实际上相差不大，分别是0.155 0和0.182 3。[①] 然而，要看到，X_3的系数已从0.003急剧地变化到0.027。

① 这些标准误得自公式：

$$\mathrm{se}(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\operatorname{var}(\hat{\beta}_2) + \operatorname{var}(\hat{\beta}_3) + 2\operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3)}$$

注意，增加共线性，$\hat{\beta}_2$和$\hat{\beta}_3$的方差也随之增加。但若两者有较大的负协方差，则如同我们的结果所表明的那样，这些方差可能被抵消。

□ 微数缺测性的后果

仿照多重共线性的俚语，以鹦鹉学话般的方式，戈德伯格根据他对过小样本的分析，引出完全类似的微数缺测性的后果。[①] 建议读者参阅戈德伯格本人进行的分析，看看他为什么把微数缺测性看成同多重共线性一样重要的概念。

10.6 说明性的例子

例 10.1 消费支出与收入和财富的关系

为了说明前面讨论过的种种观点，让我们再来考虑引言中的消费—收入一例。在表 10—5 中包含了消费、收入和消费者的财富数据。如果我们假定消费与收入和财富存在线性关系，则根据表 10—5 我们便得到如下回归结果：

$$\hat{Y}_i = 24.774\,7 + 0.941\,5X_{2i} - 0.042\,4X_{3i}$$
$$(6.752\,5)\quad (0.822\,9)\quad (0.080\,7)$$
$$t = (3.669\,0)\quad (1.144\,2)\quad (-0.526\,1) \qquad (10.6.1)$$
$$R^2 = 0.963\,5\quad \overline{R}^2 = 0.953\,1\quad \mathrm{df} = 7$$

表 10—5　关于消费 Y，收入 X_2 和财富 X_3 的假想数据　（单位：美元）

Y	X_2	X_3
70	80	810
65	100	1 009
90	120	1 273
95	140	1 425
110	160	1 633
115	180	1 876
120	200	2 052
140	220	2 201
155	240	2 435
150	260	2 686

回归（10.6.1）表明收入和财富一起解释了消费变异中的约 96%。然而没有一个斜率系数是个别统计显著的。不但如此，财富变量不仅统计上不显著，而且带有错误的符号。人们会先验地预料到消费与财富之间有正的关系。虽然 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 个别地看都不是统计显著的，但如果我们同时检验假设 $\beta_2 = \beta_3 = 0$，如表 10—6 所示，我们就可以拒绝此假设。在通常的假定下，我们得到：

$$F = \frac{4\,282.777\,0}{46.349\,4} = 92.401\,9 \qquad (10.6.2)$$

① Goldberger, op. cit, pp. 248-250.

显然这个 F 值是高度显著的。

表 10—6　　消费—收入—财富一例的 ANOVA 表

变异来源	SS	df	MSS
来自回归	8 565.554 1	2	4 282.777 0
来自残差	342.445 9	7	46.349 4

分析这一结果的几何意义是有趣的（见图 10—3）。根据回归（10.6.1），我们按照第 8 章讨论的通常程序构造 β_2 和 β_3 各自的 95%置信区间。这些区间表明，个别地看，每一区间都包含着零值。因此，个别而论，我们可以接受两个偏斜率都是零的假设。但当我们构造联合置信区域以检验假设 $\beta_2 = \beta_3 = 0$ 时，由于这个联合置信（区）间实际上是一个椭圆形的，并且不含有原点，就不能接受此假设。[①] 如同前面已指出的那样，在出现高度共线性时，对个别回归元的检验是不可靠的。这时要用总的 F 检验来观察 Y 是否与各个回归元有关。

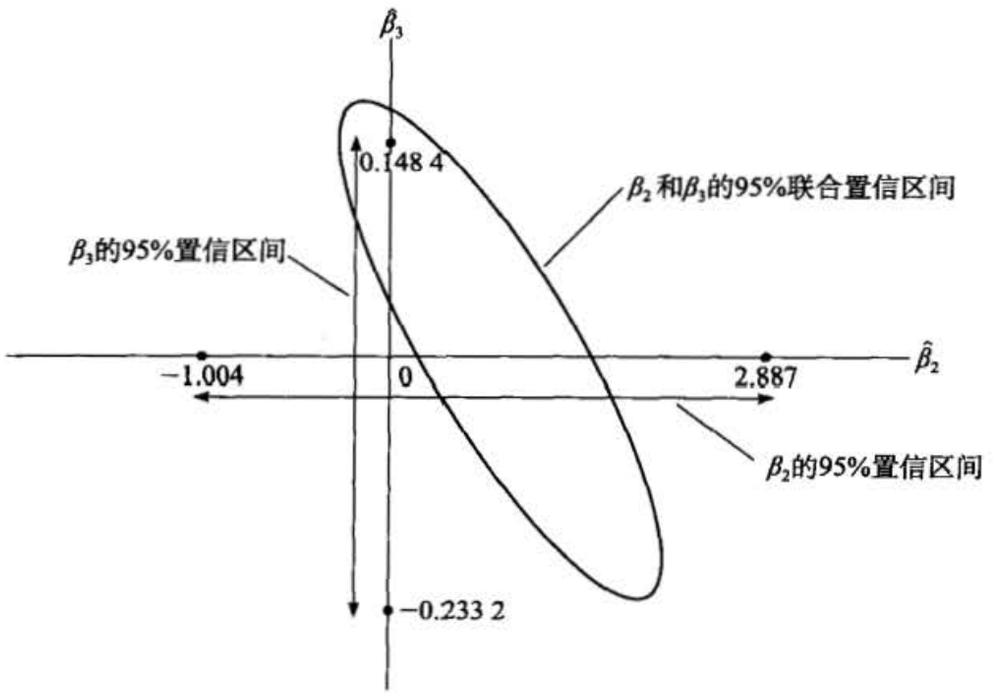


图 10—3　β_2 和 β_3 的个别置信区间与 β_2 和 β_3 的联合置信区间（椭圆形）

我们的例子生动地说明了多重共线性是怎么一回事。F 检验是显著的，而 X_2 和 X_3 的 t 值个别地看又不是显著的，这一事实本身就说明两变量的相关程度如此之高，以致无法区分收入或财富各自对消费的影响。事实上，如果我们做 X_3 对 X_2 的回归便得到：

$$\hat{X}_{3i} = 7.545\,4 + 10.190\,9X_{2i} \qquad (10.6.3)$$
$$(29.475\,8)\quad (0.164\,3)$$
$$t = (0.256\,0)\quad (62.040\,5)\qquad R^2 = 0.997\,9$$

这表明 X_3 和 X_2 之间有着几乎完全的共线性。

现在让我们做 Y 仅对 X_2 的回归，看看会出现什么情况：

① 5.3 节已指出，联合置信域的问题比较复杂，有兴趣的读者可参考那里所引的文献。

$$\hat{Y}_i = 24.454\,5 + 0.509\,1X_{2i}$$
$$(6.413\,8)\quad (0.035\,7) \qquad (10.6.4)$$
$$t=(3.812\,8)\quad (14.243\,2)\qquad R^2=0.962\,1$$

在方程（10.6.1）中，收入变量是统计上不显著的，而现在则是高度显著的。如果不做 Y 对 X_2 的回归而做 Y 对 X_3 的回归，则得到：

$$\hat{Y}_i = 24.411 + 0.049\,8X_{3i}$$
$$(6.874)\quad (0.003\,7) \qquad (10.6.5)$$
$$t=(3.551)\quad (13.29)\qquad R^2=0.956\,7$$

我们看到财富现在对消费支出也有显著的影响，而在方程（10.6.1）中它却没有显著影响。

回归（10.6.4）和（10.6.5）非常明显地表示，在极端多重共线性的情况下，去掉一个高度共线性的变量常常会使另一个 X 变量变成统计显著的。这个结果提示我们，解决极端共线性的一个方法，是扔掉共线性的变量。然而，关于这点在 10.8 节中我们还有话要说。

例 10.2　1947—2000 年美国消费函数

我们现在来考虑一个具体的数据集，即美国 1947—2000 年间有关真实消费支出（C）、真实个人可支配收入（Yd）、真实财富（W）和真实利率（I）的数据，数据在表 10—7 中给出。

表 10—7　　1947—2000 年间美国的消费支出

年份	C	Yd	W	I
1947	976.4	1 035.2	5 166.815	−10.350 94
1948	998.1	1 090	5 280.757	−4.719 804
1949	1 025.3	1 095.6	5 607.351	1.044 063
1950	1 090.9	1 192.7	5 759.515	0.407 346
1951	1 107.1	1 227	6 086.056	−5.283 152
1952	1 142.4	1 266.8	6 243.864	−0.277 011
1953	1 197.2	1 327.5	6 355.613	0.561 137
1954	1 221.9	1 344	6 797.027	−0.138 476
1955	1 310.4	1 433.8	7 172.242	0.261 997
1956	1 348.8	1 502.3	7 375.18	−0.736 124
1957	1 381.8	1 539.5	7 315.286	−0.260 683
1958	1 393	1 553.7	7 869.975	−0.574 63
1959	1 470.7	1 623.8	8 188.054	2.295 943
1960	1 510.8	1 664.8	8 351.757	1.511 181
1961	1 541.2	1 720	8 971.872	1.296 432
1962	1 617.3	1 803.5	9 091.545	1.395 922
1963	1 684	1 871.5	9 436.097	2.057 616
1964	1 784.8	2 006.9	10 003.4	2.026 599
1965	1 897.6	2 131	10 562.81	2.111 669
1966	2 006.1	2 244.6	10 522.04	2.020 251
1967	2 066.2	2 340.5	11 312.07	1.212 616
1968	2 184.2	2 448.2	12 145.41	1.054 986

续前表

年份	C	Yd	W	I
1969	2 264.8	2 524.3	11 672.25	1.732 154
1970	2 317.5	2 630	11 650.04	1.166 228
1971	2 405.2	2 745.3	12 312.92	−0.712 241
1972	2 550.5	2 874.3	13 499.92	−0.155 737
1973	2 675.9	3 072.3	13 080.96	1.413 839
1974	2 653.7	3 051.9	11 868.79	−1.042 571
1975	2 710.9	3 108.5	12 634.36	−3.533 585
1976	2 868.9	3 243.5	13 456.78	−0.656 766
1977	2 992.1	3 360.7	13 786.31	−1.190 427
1978	3 124.7	3 527.5	14 450.5	0.113 048
1979	3 203.2	3 628.6	15 340	1.704 21
1980	3 193	3 658	15 964.95	2.298 496
1981	3 236	3 741.1	15 964.99	4.703 847
1982	3 275.5	3 791.7	16 312.51	4.449 027
1983	3 454.3	3 906.9	16 944.85	4.690 972
1984	3 640.6	4 207.6	17 526.75	5.848 332
1985	3 820.9	4 347.8	19 068.35	4.330 504
1986	3 981.2	4 486.6	20 530.04	3.768 031
1987	4 113.4	4 582.5	21 235.69	2.819 469
1988	4 279.5	4 784.1	22 331.99	3.287 061
1989	4 393.7	4 906.5	23 659.8	4.317 956
1990	4 474.5	5 014.2	23 105.13	3.595 025
1991	4 466.6	5 033	24 050.21	1.802 757
1992	4 594.5	5 189.3	24 418.2	1.007 439
1993	4 748.9	5 261.3	25 092.33	0.624 79
1994	4 928.1	5 397.2	25 218.6	2.206 002
1995	5 075.6	5 539.1	27 439.73	3.333 143
1996	5 237.5	5 677.7	29 448.19	3.083 201
1997	5 423.9	5 854.5	32 664.07	3.12
1998	5 683.7	6 168.6	35 587.02	3.583 909
1999	5 968.4	6 320	39 591.26	3.245 271
2000	6 257.8	6 539.2	38 167.72	3.575 97

资料来源：见表 7—12。

我们使用如下回归模型进行分析

$$\ln C_t = \beta_1 + \beta_2 \ln Yd_t + \beta_3 \ln W_t + \beta_4 I_t + u_t \qquad (10.6.6)$$

其中 ln 表示自然对数。

在这个模型中，系数 β_2 和 β_3 分别给出收入弹性和财富弹性（为什么?），而 β_4 则给出半弹性。（为什么?）回归（10.6.6）的结果如下所示：

Dependent Variable: LOG (C)
Method: Least Squares
Sample: 1947-2000
Included observations: 54

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	-0.467711	0.042778	-10.93343	0.0000
LOG (YD)	0.804873	0.017498	45.99836	0.0000
LOG (WEALTH)	0.201270	0.017593	11.44060	0.0000
INTEREST	-0.002689	0.000762	-3.529265	0.0009

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted *R*-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	Akaike info criterion	-5.947703
Sum squared resid.	0.007121	Schwarz criterion	-5.800371
Log likelihood	164.5880	Hannan-Quinn cariter.	-5.890883
F-statistic	37832.59	Durbin-Watson stat.	1.289219
Prob(*F*-statistic)	0.000000		

注：Log代表自然对数。

这些结果表明，所有系数估计值都是高度统计显著的，因为它们的 p 值都极小。对系数估计值的解释如下。收入弹性约等于0.80，即在其他变量保持不变的情况下，如果收入增加1%，则消费支出平均增加约0.8%。财富系数约等于0.20，意味着如果财富增加1%，同样在保持其他变量不变的情况下，平均消费将增加约0.2%。利率变量的系数告诉我们，如果利率上调1个百分点，在其他条件不变的情况下，消费支出将下降约0.26%。

所有回归元的符号都与先验预期相一致，即收入与财富对消费有正影响，而利率对消费有负影响。

在这个例子中，我们有必要担心多重共线性的问题吗？显然没有必要，因为所有系数都具有正确的符号，而且每个系数又都是个别统计显著的，F 值也高度统计显著，从而表明所有这些变量一起对消费支出具有明显的影响。R^2 值也相当高。

当然，经济变量之间通常都有一定的共线性。只要不是完全共线性，我们仍然能够估计出模型参数。目前我们所能说的是，在本例中，就算存在共线性，看来也不是十分严重。但我们在10.7节将给出一些侦察共线性的诊断检验，并重新考察美国消费函数，以便确定它是否受到共线性问题的干扰。

10.7 多重共线性的侦察

在研究多重共线性的性质与后果之后，自然要问：在任一给定的情况下，特别是在涉及多于两个解释变量的模型中，我们怎样才能知道有没有共线性呢？这里最好记住克曼塔的一个忠告：

1. 多重共线性是一个程度问题而不是有无的问题。有意义的区分不在于有与无之间，而在于它的不同程度。

2. 由于多重共线性是对被假定为非随机的解释变量的情况而言，所以它是一种样本而非总体特征。

因此，我们不做“多重共线性的检验”，但如果我们愿意，可以测度它在任一具体样本中显现的程度。①

由于多重共线性本质上是一种样本现象，它来源于大多数社会科学中所收集的基本上是非实验性质的数据。我们没有侦察或度量其强度的唯一方法。我们拥有的是一些经验规则；正式的也好，非正式的也好，同样是经验规则。现在我们来考虑这些规则中的一些内容。

1. **R^2 值高，少且显著的 t 比率**。如同前面已注意到的那样，这是多重共线性的“经典”征兆。如果 R^2 值高，比方说，超过 0.8，F 检验在大多数情形中都会拒绝所有偏斜率系数同时为零的假设，但个别的 t 检验却表明，没有或很少有偏斜率系数是统计上异于零的。我们的消费—收入—财富一例已清楚地说明这一事实。

虽然这种诊断是可以理解的，但它过于强调“多重共线性的危害，仅仅在于（它使我们）无法把解释变量对 Y 的全部影响加以分解”②。

2. **回归元之间有高度的两两相关**。另一个可以提出的经验规则是，如果每两个回归元的零阶相关系数很高，比方说超过 0.8，则多重共线性问题是严重的。这一准则却带来一个疑问，虽然零阶相关系数很高表明了共线性，但并非所有情形都是，只有高的零阶相关才带来共线性。更技术性地说，高的零阶相关是多重共线性存在的充分条件，而不是必要条件。即使零阶或简单相关系数比较低（比方说，低于 0.5），多重共线性也可能存在。为了弄清楚这种关系，假设我们有一个四变量模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

并且假定有：

$$X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i}$$

其中 λ_2 和 λ_3 为常数，但不同时为零。显然，X_4 是 X_2 和 X_3 的一个准确的线性组合，从而给出 X_4 对 X_2 和 X_3 回归中的判定系数 $R^2_{4.23} = 1$。

回顾第 7 章中的公式（7.11.5），我们可以写：

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}} \tag{10.7.1}$$

但由于完全共线性，$R^2_{4.23} = 1$，于是：

$$1 = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}} \tag{10.7.2}$$

① Jan Kmenta, *Elements of Econometrics*, 2d ed., Macmillan, New York, 1986, p. 431.

② Ibid., p. 439.

不难看出，取 $r_{42}=0.5$，$r_{43}=0.5$ 及 $r_{23}=-0.5$ 这些不是很高的值，就能满足方程(10.7.2)。

因此，在涉及多于两个解释变量的模型中，简单或零阶相关并不提供判别多重共线性的一个准确无误的指南。当然，如果只有两个解释变量，零阶相关也就够了。

3. **检查偏相关系数**。由于刚才指出仅仅依靠零阶相关所带来的问题，法勒(Farrar)和格劳伯(Glauber)建议我们去检查偏相关系数。[①] 例如，在做 Y 对 X_2、X_3 和 X_4 的回归中，发现 $R^2_{1.234}$ 很高，而 $r^2_{12.34}$、$r^2_{13.24}$ 和 $r^2_{14.23}$ 都比较低时，可能表示变量 X_2、X_3 和 X_4 是高度交互相关的，并且至少其中一个变量是多余的。

虽然对偏相关系数的检查会有一定的用处，但不能保证偏相关系数能对多重共线性提供一个准确无误的指南，因为有可能 R^2 和全部偏相关系数足够高，仍出现多重共线性。但更为重要的是，罗伯特·威克斯(C. Robert Wichers)已证明[②]法勒-格劳伯的偏相关检验在下述意义上是无效的：一种给定的偏相关关系可能与不同的多重共线性模式都没有矛盾。法勒-格劳伯检验还受到库马(T. Krishna Kumar)[③]、约翰·奥黑根(John O'Hagan)和布伦丹·麦凯布(Brendan McCabe)[④]的严厉批评。

4. **辅助回归**。由于多重共线性来自某些回归元是其余回归元的准确或近似线性组合，为了找出究竟哪一个 X 变量和其余 X 变量存在这种关系，方法之一是做每一 X_i 对其余 X 变量的回归，并算出相应的 R^2，记为 R^2_i；这样的回归叫做**辅助回归**(auxiliary regression)，以辅助 Y 对诸 X 的主回归。然后，按照方程(8.4.11)中建立的 F 与 R^2 之间的关系，变量

$$F_i=\frac{R^2_{x_1\cdot x_2x_3\cdots x_k}/(k-2)}{(1-R^2_{x_1\cdot x_2x_3\cdots x_k})/(n-k+1)} \tag{10.7.3}$$

服从自由度分别为 $k-2$ 和 $n-k-1$ 的 F 分布。在方程(10.7.3)中，n 表示样本容量，k 表示包括截距项在内的变量个数，而 $R^2_{x_1\cdot x_2x_3\cdots x_k}$ 表示 X_i 变量与其余 X 变量之间回归的判定系数。[⑤]

如果计算值 F_i 超过选定显著水平的 F 临界值，我们就把它看作这个特定的 X_i 和其余 X 存在共线性；如果它不超过 F 临界值，就说 X_i 和其余 X 无共线性。这时可把该变量 X_i 保留在模型中。但如果 F_i 是统计上显著的，则 X_i 的去留问题仍待解

---

① D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited", *Review of Economics and Statistics*, vol. 49, 1967, pp. 92-107.

② "The Detection of Multicollinearity: A Comment," *Review of Economics and Statistics*, vol. 57, 1975, pp. 365-366.

③ "Multicollinearity in Regression Analysis," *Review of Economics and Statistics*, vol. 57, 1975, pp. 366-368.

④ "Tests for the Severity of Multicollinearity in Regression Analysis: A Comment," *Review of Economics and Statistics*, vol. 57, 1975, pp. 368-370.

⑤ 例如，$R^2_{r_2}$ 可通过做 X_{2i} 的如下回归得到：$X_{2i}=a_1+a_3X_{3i}+a_4X_{4i}+\cdots+a_kX_{ki}+a_i$。

决。在 10.8 节中我们将再回到此问题。

但是这种方法并非没有缺点，因为

> ……如果多重共线性仅涉及少数变量，辅助回归还不至于对广泛的多重共线关系有应接不暇之苦，那么所估计的系数也许能揭露回归元之间线性关系的性质。不幸的是，如果遇上几个复杂的线性相关，做这种曲线拟合的练习就不一定有多少价值；要辨别各个不同的交互关系仍是困难的。[①]

除了对所有辅助 R^2 值做形式检验外，还可采取**克莱因的经验法则**（Klein's rule of thumb）：仅当来自一个辅助回归的 R^2 大于得自 Y 对全部回归元的回归中的总 R^2 值时，多重共线性才算是一个麻烦的问题。[②] 当然，和其他经验法则一样，不可把这个经验法则当作法定的规则来运用。

5. **本征值与病态指数**。利用 EViews 和 Stata，我们可以得到诊断多重共线性的本征值（eigenvalues）和病态指数（condition index，CI）*。这里我们不准备讨论本征值，否则将被引到超出本书范围的矩阵代数的专题讨论。但是通过本征值，我们可导出我们要讲的**病态数**（condition number）k，其定义为：

$$k = \frac{\text{最大本征值}}{\text{最小本征值}}$$

以及**病态指数**，其定义为：

$$\text{CI} = \sqrt{\frac{\text{最大本征值}}{\text{最小本征值}}} = \sqrt{k}$$

于是，我们有这样的经验法则：如果 k 在 100 和 1 000 之间，就算有中等强度的多重共线性；如果 k 大于 1 000，就算有严重的多重共线性。另一算法是：如果 CI（$=\sqrt{k}$）介于 10 与 30 之间，就算有中等强度的多重共线性，而如果 CI 大于 30，就算是严重的多重共线性。

对于附录 7A.5 中说明性的例子，最小的本征值为 3.786，最大的本征值为 187.526 9，从而 k=187.526 9/3.786 或约等于 49.53。因此 CI= $\sqrt{49.53}$ =7.037 7；因此 k 和 CI 都表明多重共线性不是很严重。顺便指出，低的本征值（相对于最大本征值而言）一般都表明数据中有近似线性相关性。

有些作者认为病态指数是诊断多重共线性的现有方法中最好的一个。但这一观点并未被广泛赞同。我们则认为，CI 也仅是一种经验法则，也许略为成熟。要进一步了解细节，可阅读参考文献。[③]

---

* 这里 condition 一词其实指 ill-conditioned。——译者注

① George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, John Wiley & Sons, New York, 1982, p. 621.

② Lawrence R. Klein, *An Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, NJ, 1962, p. 101.

③ 特别是参考 D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York, 1980, Chapter 3。但此书不适于初学者阅读。

6. **容许度与方差膨胀因子**。我们已经介绍过 TOL 和 VIF。R_j^2 是 X_j 对其余 $k-2$ 个回归元的（辅助）回归中的判定系数，随着 R_j^2 趋近 1，也就是随着 X_j 与其他回归元的共线性增加，VIF 也增加，并且以无穷大为其极限。

因此，一些作者用 VIF 作为多重共线性的一个指标：VIF_j 值越大，变量 X_j 越“麻烦”或共线性越大。但究竟 VIF 要多高才会使回归陷入麻烦的地步呢？**作为一种经验法则**，如果一个变量的 VIF 超过 10（当 R_j^2 超过 0.90 时将发生这种情况），则认为该变量是高度共线性的。[①]

当然，鉴于 TOL_j 与 VIF_j 之间的密切联系，也能用 TOL_j 来度量多重共线性。TOL_j 越接近于 0，该变量与其他回归元之间的共线性程度就越大。另一方面，TOL_j 越接近于 1，则 X_j 与其他回归元之间没有共线性的证据就越充分。

用 VIF（或容许度）去度量共线性也难免受到批评。如方程（10.5.4）所示，$\text{var}(\hat{\beta}_j)$ 依赖于 3 个因子：σ^2，$\sum x_j^2$ 和 VIF_j。一个高的 VIF 可被一个低的 σ^2 或一个高的 $\sum x_j^2$ 所抵消。换句话说，一个高的 VIF 既不是导致高的方差和高的标准误的必要条件，也不是其充分条件。因此，一个高的 VIF 度量出来的高多重共线性不一定就是高标准误的原因。在所有这些讨论中，高和低这两个用语都是相对而言的。

7. **散点图**。用散点图来看一个回归模型中各个变量之间如何相关是一个好办法。图 10—4 给出了上一节讨论的美国消费—支出一例的散点图（例 10.2）。由于模型中有 4 个变量——1 个因变量和 3 个解释变量（真实个人可支配收入 Yd、真实财富 W 和真实利率 I），所以我们使用一个 4×4 的盒状图。

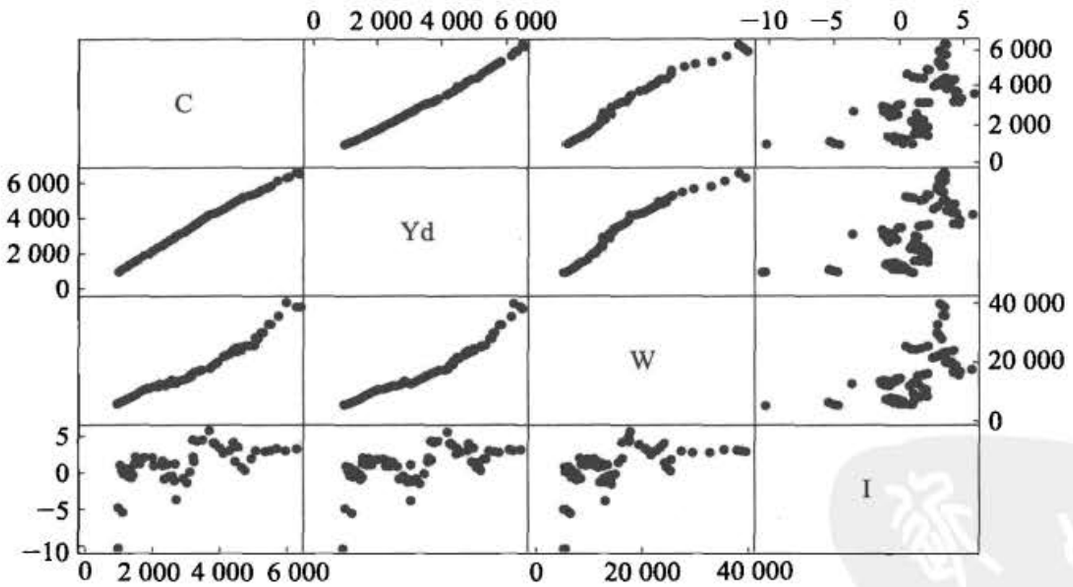


图 10—4 例 10.2 中（表 10—7）数据的散点图

① 参看 David G. Kleinbaum, Lawrence L. Kupper, and Keith E. Muller, *Applied Regression Analysis and Other Multivariate Methods*, 2d ed., PWS-Kent, Boston, Mass, 1988, p. 210。

首先考虑从左上角到右下角的主对角线。主对角线上这四个方格中没有散点图。如果要画的话，我们得到的相关系数也是1，因为它们都是一个变量与自身的相关关系。主对角线之外的方格则给出了变量之间的相关关系。比如，以财富W为例。此图表明，财富与收入Yd高度相关（二者之间的相关系数为0.97），但并非完全相关。如果它们完全相关（即它们之间的相关系数等于1），那我们就不能估计回归(10.6.6)，因为财富和收入之间存在着精确的线性关系。此散点图还表明利率与其他三个变量不是高度相关。

既然有些统计软件已经包含了散点图函数，所以这一诊断方法应该与前面讨论过的方法一起考虑。但需记住，前面曾指出，两个变量之间的简单相关不是存在共线性的一个确切指标。

在结束我们关于侦察多重共线性的讨论之际，我们强调指出，本节所讨论的各种方法实质上都类似垂钓。我们无法知道一种方法在任一特定的应用中是否灵验。真遗憾，想不出有什么好的办法，因为多重共线性就出现在研究者对于一个给定的样本施加不了多少控制的场合之中，尤其是数据本质上是非实验性质的——社会科学研究者所遇到的共同命运。

作为多重共线性的一种仿效，戈德伯格想出许多侦察微数缺测性的方法，比如，制定样本容量的n^*临界值：仅当实际样本容量n小于n^*时，微数缺测性才会成为一个问题。戈德伯格仿效的意义在于，强调小样本容量以及缺少变异的解释变量会导致至少和多重共线性同等严重的问题。

10.8 补救措施

如果多重共线性严重，怎么办？我们有两种选择：(1) 无为而治；或 (2) 采用某些经验法则。

□ 无为而治

布兰查德 (Blanchard) 将“无为而治”流派表述如下[①]：

> 当学生在做他们的第一个普通最小二乘 (OLS) 回归时，他们通常遇到的第一个问题就是多重共线性的问题。他们中许多人断定OLS有些问题；有些人求助于新的通常也有创造性的方法来避免这个问题。但我们告诉他们，那样做是错的，多重共线性是上帝的意志，而不是OLS或其他一般性统计方法的

① O. J. Blanchard, Comment, *Journal of Business and Economic Statistics*, vol. 5, 1967, pp. 449-451. 所引内容复制于 Peter Kennedy, *A Guide to Econometrics*, 4th ed., MIT Press, Cambridge, Mass., 1998, p. 190。

问题。

布兰查德所说的是，多重共线性实质上是一个数据不足的问题（微数缺测性），而我们有时候无法选择能用于经验分析的数据。

同样，并非回归模型中所有的系数都是统计非显著的。此外，即使我们不能更准确地估计一个或多个回归系数，但可以相对有效地估计它们的一个线性组合（即可估计的函数）。如我们在方程（10.2.3）中所见，即使我们不能分别估计 α 的两个部分，但可以从整体上估计出 α。有时候，这是我们在给定数据集的情况下最好的做法。[①]

□ 经验程序

你也可以尝试用如下的经验法则来解决多重共线性问题，其成功与否取决于共线性问题的严重程度。

1. **先验信息**。假使我们考虑模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

其中 Y=消费，X_2 =收入，X_3 =财富。如前所见，收入与财富有高度共线性的趋势。但若先验地认为 $\beta_3 = 0.10\beta_2$；也就是说，消费对财富的变化率是对收入相应变化率的 1/10。这样一来，我们就可进行下面的回归：

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned}$$

其中 $X_i = X_{2i} + 0.10X_{3i}$。一旦估算出 $\hat{\beta}_2$，便可从想象中的 β_2 与 β_3 的关系式估计出 $\hat{\beta}_3$。

怎样获得先验信息呢？它可以来自此前遇到同样严重的共线性问题的经验研究工作，或者来自该研究领域的有关基础理论。例如，在柯布-道格拉斯生产函数（7.9.1）中，如果人们预期规模报酬不变成立，则有 $\beta_2 + \beta_3 = 1$。这样就能做回归（8.6.14），即做产出/劳动比对资本/劳动比的回归。如果劳动和资本之间存在共线性，好比大多数样本数据一般都会遇到的情形那样，这一变换就减轻或消除了共线性问题。但这里提出一个关于施加此类先验约束的忠告是适宜的：

> ……因为，一般地说，我们宁愿检验经济理论上的先验预期而不是单纯地把这些未必合适的预期施加于数据之上。[②]

不管怎样，我们从 8.6 节知道了怎样去明确地检验这些约束的真实性。

2. **横截面数据与时间序列数据并用**。外部信息或先验信息法的一个变种，是横截面数据与时间序列数据的组合，称数据并用（pooling the data）。假如我们要研究

---

① 对此一个有意思的讨论可参见 J. Conlisk, "When Collinearity is Desirable," *Western Economic Journal*, vol. 9, 1971, pp. 393-407。

② Mark B. Stewart and Kenneth F. Wallis, *Introductory Econometrics*, 2d ed., John Wiley & Sons, A Halstead Press Book, New York, 1981, p. 154.

美国的汽车需求，并假定我们拥有车辆出售数、车辆平均价格和消费者收入的时间序列数据，还设定：

$$\ln Y_t = \beta_1 + \beta_2 \ln P_t + \beta_3 \ln I_t + u_t$$

其中 Y=车辆出售数，P=平均价格，I=收入，t=时间。我们的目的是要估计价格弹性 β_2 和收入弹性 β_3。

在时间序列数据中，价格和收入变量一般都有高度共线性的趋势。因此，如果我们做上述回归，我们将遇到通常的多重共线性问题。解决此问题的一个方法曾由托宾（Tobin）提出。[①] 他说，如果我们拥有横截面数据［例如，由消费者定点追踪（consumer panels）产生的数据，或各种私人和政府机构举办的预算研究］，我们就能相当可靠地估计收入弹性 β_3。因为这些数据都产生于一定时点上，价格还不至于有多大变化。令收入弹性的横截面估计为 $\hat{\beta}_3$。利用这一估计值，就可将前述的时间序列回归写为：

$$Y_t^* = \beta_1 + \beta_2 \ln P_t + u_t$$

其中 $Y^* = \ln Y - \hat{\beta}_3 \ln I$，即 Y^* 代表除去收入效应之后的 Y 值。现在就可从上面的回归得到价格弹性的 β_2 估计值。

虽然时间序列和横截面数据并用看来是一个很不错的方法，但刚才的做法可能引起解释方面的问题。因为这样做无形地假定了收入弹性的横截面估计和从纯粹的时间序列分析中得到的估计是一样的。[②] 不管怎样，数据并用技术已经在多种应用中使用，而且当横截面估计在不同截面之间变化不大时是一个值得考虑的方法。这种方法的一个例子见习题 10.26。

3. **剔除变量与设定偏误**。面对严重多重共线性，最简单的做法之一是剔除共线性的变量之一。例如，在我们的消费—收入—财富一例中，当我们剔除财富变量时，得到回归（10.6.4），表明收入变量在原模型中不是统计显著的，而现在则是“高度”显著的。

但从模型中剔除一个变量，可能导致**设定偏误**或**设定误差**。设定偏误指在分析中使用的模型被不正确地设定。比如，假如经济理论告诉我们，在解释消费支出的模型中应同时包括收入和财富，那么剔除财富变量就会构成设定偏误。

虽然我们将在第 13 章中专题讨论设定偏误问题，但在 7.7 节中也曾对它有过思考。我们曾看到如果真实模型是：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

而我们错误地拟合了模型：

$$Y_i = b_1 + b_{12} X_{2i} + a_i \tag{10.8.1}$$

那么，可以证明（见附录 13A.1）：

---

① J. Tobin, “A Statistical Demand Function for Food in the U. S. A.,” *Journal of the Royal Statistical Society*, Ser. A, 1950, pp. 113-141.

② 关于数据并用技术的一个透彻的讨论与应用，参见 Edwin Kuh, *Capital Stock Growth: A Micro-Econometric Approach*, North-Holland Publishing Company, Amsterdam, 1963, Chapters 5 and 6。

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \tag{10.8.2}$$

其中 $b_{32} = X_3$ 对 X_2 回归中的斜率系数。由方程（10.8.2）明显可见，只要 b_{32} 不为零（我们假定 β_3 异于零，否则在原始模型中包括 X_3 是没有意义的），b_{12} 就必定是 β_2 的一个有偏误的估计。[①] 当然，如果 b_{32} 是零，我们本来就没有多重共线性问题。从方程（10.8.2）还明显看到，如果 b_{32} 和 β_3 都是正的（或都是负的），$E(b_{12})$ 将大于 β_2；从而平均而言，b_{32} 高估了 β_2，即导致一个正的偏误；同理，如果乘积 $b_{32}\beta_3$ 是负的，则平均而言，b_{12} 将低估了 β_2，即导致一个负的偏误。

由上述讨论可见，从模型中除掉一个变量以缓解多重共线性的问题可能会导致设定上的偏误。因此，在某些情形中，医治也许比疾病更糟糕，多重共线性虽有碍于对模型参数的准确估计，但剔除变量则对参数的真值有严重的误导。应记得，在近似共线性情形下，OLS 估计量仍是 BLUE。

4. **变量代换**。假使我们拥有消费支出、收入和财富的时间序列数据，数据中收入与财富有高度多重共线性的一个理由是，随着时间的演变这两个变量都朝同一方向变动。减少这种相依性的一个方法是按以下方法去做。

如果关系式：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{10.8.3}$$

在时间 t 成立，那么它在时间 $t-1$ 也成立，因为时间原点是任意的。因此又有：

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \tag{10.8.4}$$

如果从方程（10.8.3）中减去方程（10.8.4），就得到：

$$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t \tag{10.8.5}$$

其中 $v_t = u_t - u_{t-1}$。因为我们不是对原始变量做回归，而是对这些变量的相继差异做回归，方程（10.8.5）被称为**一阶差分形式**（first difference form）。

一阶差分模型常常减轻了多重共线性的严重程度，因为尽管 X_2 与 X_3 存在相当高的相关性，也没有先验的理由相信它们的差分仍然高度相关。

如我们在**时间序列计量经济学**（time series econometrics）的章节中所见，一阶差分变换的一个附带优点在于，它可以使非平稳时间序列变得平稳。我们在那些章节中将会看到平稳时间序列的重要性。在第 1 章曾指出，粗略地讲，如果一个时间序列 Y_t 的均值和方差不随时间而系统地变化，那它就是平稳的。

实践中另外一个常用的变换是**比率变换**（ratio transformation）。考虑模型：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{10.8.6}$$

其中 Y 为以真实价格表示的消费支出，X_2 为 GDP，X_3 为总人口。由于 GDP 和总人口都随时间而增长，所以它们可能会相关。对此问题的一种“解决办法”是，通过将方程（10.8.6）除以 X_3 得到以人均量为基础的模型：

$$\frac{Y_t}{X_{3t}} = \beta_1\left(\frac{1}{X_{3t}}\right) + \beta_2\left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \tag{10.8.7}$$

① 再者，如果 b_{32} 不随样本无限增大而趋于零，则 b_{12} 不仅有偏误，而且没有一致性。

这样的变换可能会减少原有变量的共线性。

但一阶差分或比率变换都不是没有问题。例如，方程（10.8.5）中的误差项 v_t 可能不满足经典线性回归模型的一个假定，即干扰项的序列不相关性。我们在第 12 章将会看到，如果原来的干扰项 u_t 是序列无关的，那么上面得到的误差项 v_t 在多数情况下将会序列相关。因此，治疗比疾病更糟糕。而且，还会因为差分过程而减少一个观测，并因此减少一个自由度。在小样本中，这可能是你起码要考虑的一个因素。另外，一阶差分程序在横截面数据中可能不太适合，因为横截面数据的观测不存在逻辑上的顺序。

类似地，在比率模型（10.8.7）中，如果误差项 u_t 是同方差的，那么误差项：

$$\frac{u_t}{X_{3t}}$$

将是异方差的。我们在第 11 章将会看到这一点。同样，补救的办法比原来的问题更糟糕。

总之，在应用一阶差分或比率变换来解决多重共线性问题时应该尤其小心。

5. **补充新数据**。由于多重共线性是一个样本特性，故有可能在关于同样变量的另一样本中共线性没有第一个样本那么严重。有时只需增大样本容量（如果可能的话）就能减轻共线性问题。例如，在三变量模型中，我们看到：

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)}$$

现在，随着样本增加，$\sum x_{2i}^2$ 一般地说都会增加。（为什么?）因此，对任何给定的 r_{23}^2，$\hat{\beta}_2$ 的方差将减小，从而降低标准误，以使我们能更准确地估计 β_2。

作为一个说明，考虑以下根据 10 次观测的消费 Y 对收入 X_2 和财富 X_3 的回归[①]：

$$\hat{Y}_i = 24.377 + 0.871\,6X_{2i} - 0.034\,9X_{3i} \qquad (10.8.8)$$

$$t=(3.875)\ (2.772\,6)\ (-1.159\,5) \qquad R^2=0.968\,2$$

回归中的财富系数不仅在 5%水平上不是统计显著的，而且有错误的符号。但当样本容量增加到 40 次观测时（微数缺测性?），我们得到如下结果：

$$\hat{Y}_i = 2.090\,7 + 0.729\,9X_{2i} + 0.060\,5X_{3i} \qquad (10.8.9)$$

$$t=(0.871\,3)\ (6.001\,4) \qquad (2.001\,4) \qquad R^2=0.967\,2$$

现在，财富系数不仅具有正确的符号，而且在 5%水平上是统计显著的。

要获得补充数据或“更好”的数据，并不总是那么容易的。贾奇（Judge）等人曾说：

> 不幸的是，经济学家很少能取得补充数据而不花大本钱。而要选取他们所希望的解释变量的值就更难了。此外，在非控制的情况下增加新变量，我们必须警惕，新增观测值的生成过程不同于原来数据集的生成过程；就是说，我们

① 感谢艾伯特·朱克（Albert Zucker）提供下述回归结果。

必须有把握看到，与新观测值相对应的经济结构和原来的结构是一样的。[①]

6. **在多项式回归中降低共线性**。在 7.10 节中，我们曾讨论多项式回归模型。这种模型的一个特点是解释变量以不同的幂出现。例如，在总成本对产出、产出的平方和产出的三次方的回归即所谓立方总成本函数（7.10.4）中，各产出项将是相关的，以致难以准确估计各个斜率系数。[②] 然而，在实践中，我们发现，如果将解释变量表达为离差形式（即对均值的离差），多重共线性就可大为降低。但即使如此，问题仍然存在。[③] 这时也许还可考虑诸如**正交多项式**（orthogonal polynomials）之类的方法。[④]

7. **拯救多重共线性的其他方法**。多元统计技术诸如**因子分析**（factor analysis）、**主元法**（principal components）或**脊回归**（ridge regression）常被用来“解决”多重共线性问题。可惜这些技术都要利用矩阵代数才便于讨论。但这样做就超出了本书的范围。[⑤]

10.9 多重共线性一定是坏事吗？如果预测是唯一目的，就未必如此

前面说过，如果回归分析的唯一目的是预测或预报，则多重共线性就不是一个严重的问题。因为 R^2 值越高，预测越准。[⑥] 但是，这也许是“……只要预测值所对应的解释变量值和原始的设计［数据］矩阵 X 都遵从同样近似于准确的相依关系”[⑦]。比方说，如果在回归的估计中发现 $X_2 = 2X_3$ 近似地成立，那么在一个用以预测 Y 的未来值的样本中，X_2 也应近似地等于 $2X_3$。但这是一个实际上难以满足的条件（参看第 338 页注释①）。由此可见，预测将变得越来越不确定。[⑧] 此外，如果分

① Judge et al., op. cit., p. 625. 还参见 10.9 节。

② 如已指出的那样，因 X、X^2 和 X^3 是非线性关系，故严格地说，多项式回归并不违反经典模型中无多重共线性的假定。

③ 参看 R. A. Bradley and S. S. Srivastava, “Correlation and Polynomial Regression,” *American Statistician*, vol. 33, 1979, pp. 11-14.

④ 参看 Norman Draper and Harry Smith, *Applied Regression Analysis*, 2d ed., John Wiley & Sons, New York, 1981, pp. 266-274.

⑤ 一种具有可读性的、从应用观点说明这些技术的读物是 Samprit Chatterjee and Bertram Price, *Regression Analysis by Example*, John Wiley & Sons, New York, 1977, Chapters 7 and 8。还参阅 H. D. Vinod, “A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares,” *Review of Economics and Statistics*, vol. 60, February 1978, pp. 121-131。

⑥ 参看 R. C. Geary, “Some Results about Relations between Stochastic Variables: A Discussion Document,” *Review of International Statistical Institute*, vol. 31, 1963, pp. 163-181。

⑦ Judge et. al., op. cit., p. 619。你还将看到，在此页上证明了为什么尽管有共线性，但如果现有的共线性结构继续存在于未来的样本中，人们就能得到较好的均值预测。

⑧ 更精彩的讨论，参见 E. Malinvaud, *Statistical Methods of Econometrics*, 2d ed., North-Holland Publishing Company, Amsterdam, 1970, pp. 220-221。

析的目的不仅在于预测，而且还在于参数的可靠估计，那么，严重的多重共线性将成为一个问题，因为我们已经看到它导致估计量的标准误偏大。

然而有一种情形，多重共线性不会成为一个严重的问题，这就是 R^2 高，同时回归系数由于较高的 t 值也都表现为个别显著的情形。毕竟，多重共线性诊断（如病态指数）表明了数据中有严重的共线性。那么，什么时候会出现这种“不成为严重问题”的情形呢？如约翰斯顿（Johnston）所说：

> 如果每个个别系数都正好在数值上大大超过真值，那么尽管标准误膨胀了，效应依然显示出来，和/或真值本来就是如此之大，即使估计值过低，仍然表现为显著的。[①]

10.10 一个引申的例子：朗利数据

我们以分析朗利（Longley）所搜集的数据结束本章。[②] 尽管最初搜集的朗利数据只是为了在几个计算机程序中评价普通最小二乘估计值的计算精度，但现在又担负着说明包括多重共线在内的几个计量经济学问题的重任。表 10—8 复制了这些数据。这些数据是 1947—1962 年间的时间序列，Y ＝被雇佣人数（以千人计）；X_1 ＝GNP 暗含的价格缩减指数；X_2 ＝GNP（以百万美元计）；X_3 ＝失业人数（以千人计）；X_4 ＝军队中的人数；X_5 ＝14 岁以上的非机构人口数；X_6 ＝年份：1947 年取值 1，1948 年取值 2，……，1962 年取值 16。

表 10—8　　朗利数据

观测	Y	X_1	X_2	X_3	X_4	X_5	时间
1947	60 323	830	234 289	2 356	1 590	107 608	1
1948	61 122	885	259 426	2 325	1 456	108 632	2
1949	60 171	882	258 054	3 682	1 616	109 773	3
1950	61 187	895	284 599	3 351	1 650	110 929	4
1951	63 221	962	328 975	2 099	3 099	112 075	5
1952	63 639	981	346 999	1 932	3 594	113 270	6
1953	64 989	990	365 385	1 870	3 547	115 094	7
1954	63 761	1 000	363 112	3 578	3 350	116 219	8
1955	66 019	1 012	397 469	2 904	3 048	117 388	9
1956	67 857	1 046	419 180	2 822	2 857	118 734	10
1957	68 169	1 084	442 769	2 936	2 798	120 445	11

① J. Johnston, *Econometric Methods*, 3d ed., McGraw-Hill, New York, 1984, p. 249.

② J. Longley, "An Appraisal of Least-Squares Programs from the Point of the User," *Journal of the American Statistical Association*, vol. 62, 1967, pp. 819-841.

续前表

观测	Y	X_1	X_2	X_3	X_4	X_5	时间
1958	66 513	1 108	444 546	4 681	2 637	121 950	12
1959	68 655	1 126	482 704	3 813	2 552	123 366	13
1960	69 564	1 142	502 601	3 931	2 514	125 368	14
1961	69 331	1 157	518 173	4 806	2 572	127 852	15
1962	70 551	1 169	554 894	4 007	2 827	130 081	16

资料来源：J. Longley，"An Appraisal of Least-Squares Programs from the Point of the User," *Journal of the American Statistical Association*，vol. 62，1967，pp. 819-941.

假定我们的目标是基于这 6 个 X 变量来预测 Y。利用 EViews 6 软件，我们得到如下回归结果：

Dependent Variable: Y
Sample: 1947-1962

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3482259.	890420.4	-3.910803	0.0036
X_1	15.06187	84.91493	0.177376	0.8631
X_2	-0.035819	0.033491	-1.069516	0.3127
X_3	-2.020230	0.488400	-4.136427	0.0025
X_4	-1.033227	0.214274	-4.821985	0.0009
X_5	-0.051104	0.226073	-0.226051	0.8262
X_6	1829.151	455.4785	4.015890	0.0030

R-squared	0.995479	Mean dependent var.	65317.00
Adjusted R-squared	0.992465	S.D. dependent var.	3511.968
S.E. of regression	304.8541	Akaike info criterion	14.57718
Sum squared resid.	836424.1	Schwarz criterion	14.91519
Log likelihood	-109.6174	F-statistic	330.2853
Durbin-Watson stat.	2.559488	Prob(F-statistic)	0.000000

从结果一眼就看出存在多重共线性问题，因为 R^2 的值很高，但有几个变量不是统计显著的（X_1，X_2 和 X_5），这是多重共线性的典型特征。为了更清楚地说明这一点，我们在表 10—9 中给出这 6 个回归元之间的相关关系。

表 10—9　　相关关系

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000 000	0.991 589	0.620 633	0.464 744	0.979 163	0.991 149
X_2	0.991 589	1.000 000	0.604 261	0.446 437	0.991 090	0.995 273
X_3	0.620 633	0.604 261	1.000 000	−0.177 421	0.686 552	0.668 257
X_4	0.464 744	0.446 437	−0.177 421	1.000 000	0.364 416	0.417 245
X_5	0.979 163	0.991 090	0.686 552	0.364 416	1.000 000	0.993 953
X_{16}	0.991 149	0.995 273	0.668 257	0.417 245	0.993 953	1.000 000

此表给出了所谓的**相关矩阵**（correlation matrix）。此表中主对角线上的数字（从左上角到右下角）给出了一个变量与其自身的相关系数，根据定义，都应该是 1，

而主对角线之外的数字给出了 X 变量两两之间的相关系数。此表的第一行给出了 X_1 与其他变量之间的相关系数。比如 0.991 589 就是 X_1 与 X_2 之间的相关系数，0.620 633 是 X_1 与 X_3 之间的相关系数，等等。

可以看到，这两两之间的相关系数有几个很高，表明可能存在着严重的共线性问题。当然，记住前面给过的警告，这种两两相关可能是存在多重共线性的充分但非必要条件。

为了进一步了解多重共线性的性质，我们做辅助回归，即将每个 X 变量都对其余的 X 变量进行回归，为节省篇幅，我们只给出从这些回归所得到的 R^2 值，由表 10—10 给出。由于辅助回归的 R^2 值很高（X_4 的回归可能例外），看来确实存在严重的共线性问题。从容许度因子能得到相同的信息。前面曾指出，容许度因子越接近零，共线性的证据就越大。

表 10—10　　辅助回归的 R^2 值

因变量	R^2 值	容许度（TOL）$=1-R^2$
X_1	0.992 6	0.007 4
X_2	0.999 4	0.000 6
X_3	0.970 2	0.029 8
X_4	0.721 3	0.278 7
X_5	0.997 0	0.003 0
X_6	0.998 6	0.001 4

应用克莱因的经验法则，我们看到，6 个辅助回归中有 3 个回归得到的 R^2 值超过了总体 R^2 值（即从 Y 对所有 X 变量回归得到的 R^2 值）0.995 4，再次表明朗利数据确实被多重共线性问题所困扰。顺便提一句，应用方程（10.7.3）中给出的 F 检验，读者应该能够验证上表中给出的 R^2 值都是统计显著异于零的。

我们在前面曾指出，OLS 估计量及其标准误对数据的微小变化都很敏感。习题 10.32 要求读者在去掉最后一个观测的情况下重做 Y 对所有 6 个 X 变量的回归，即对 1947—1961 年期间做回归。你将看到，仅去掉一年的观测，回归结果会如何变化。

既然我们已经证实存在多重共线性问题，那我们能采取什么“补救”措施呢？让我们考虑原来的模型。首先，我们可以不用名义 GNP 而用真实 GNP，将名义 GNP 除以 GNP 暗含的价格缩减指数 X_1 即可。其次，由于 14 岁以上非机构人口数因人口数自然增长而随时间不断增长，所以它与我们模型中的时间变量 X_6 高度相关。因此，不再同时采用这两个变量，我们将留下 X_5 并去掉 X_6。第三，没有充分有力的理由把失业人数 X_3 包括进来；可能失业率是劳动力市场状况的一个更好的度量指标，但我们没有这方面的数据，故去掉变量 X_3。经过这些变化，我们得到如下回归结果（RGNP=真实 GNP）[①]：

① X_5 和 X_6 之间的相关系数约为 0.993 9，实际上是很高的相关度。

Dependent Variable: Y
Sample: 1947-1962

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	65720.37	10624.81	6.185558	0.0000
RGNP	9.736496	1.791552	5.434671	0.0002
X_4	-0.687966	0.322238	-2.134965	0.0541
X_5	-0.299537	0.141761	-2.112965	0.0562

R-squared	0.981404	Mean dependent var.	65317.00
Adjusted R-squared	0.976755	S.D. dependent var.	3511.968
S.E. of regression	535.4492	Akaike info criterion	15.61641
Sum squared resid.	3440470.	Schwarz criterion	15.80955
Log likelihood	-120.9313	F-statistic	211.0972
Durbin-Watson stat.	1.654069	Prob(F-statistic)	0.000000

尽管 R^2 值与原来的 R^2 值相比略有下降，但仍然很高。现在，所有的估计系数都是统计显著的，系数的符号也都符合其经济含义。

我们让读者自己构想另外一个模型并分析结果的变化。仍须记住以前我们听到对数据进行比率变换来解决共线性问题的警告。我们在第 11 章将再次讨论这个问题。

要点与结论

1. 经典线性回归模型的假定之一，是解释变量 X 之间无多重共线性。大致地说，多重共线性指的是 X 变量之间有准确的或近似准确的线性关系。

2. 多重共线性有如下后果：如果 X 之间有完全的共线性，则它们的回归系数是不确定的，并且它们的标准误没有定义。如果共线性是高度的而不是完全的，则回归系数的估计是可能的，但趋向于有很大的标准误。其结果是，系数的总体值不能准确地加以估计。然而，如果目的在于估计这些系数的线性组合即所谓可估函数（estimable functions），则虽有完全多重共线性也无妨。

3. 虽然没有识破共线性的十拿九稳的方法，却有如下几种指标可以利用：

(a) 多重共线性的最明显信号是 R^2 异常高而回归系数在通常 t 检验的基础上却没有一个是统计上显著的。当然，这是一个极端情形。

(b) 在仅有两个解释变量的模型中，检查两个变量之间的零阶或简单相关系数，会得到对共线性的一个相当好的认识。如果此相关值高，则通常可归咎于多重共线性。

(c) 然而，当模型涉及多于两个 X 解释变量时，由于可能低的零阶相关却高的多重共线性，模型中的零阶相关系数可能被误导。对于这种情形，也许有必要检查偏相关系数。

(d) 如果 R^2 高而偏相关系数低，则多重共线性是可能的。这时一个或多个变量可能是多余的。但若 R^2 高且偏相关系数也高，则多重共线性也许不易识破。而且，如威克斯、库马、奥黑根和麦凯布等人所指出的那样，法勒和格劳伯建议的偏相关系数检验有一些统计上的毛病。

(e) 因此，不妨拿模型中的每个 X_i 变量对所有其余 X 变量做一个回归，并求出相应的判定系数 R_i^2。一个高的 R_i^2 将表明 X_i 和其余的 X 高度相关，从而可考虑把 X_i 从模型中清除出去，如果

这样做不致引起严重的设定偏误。

4. 多重共线性的侦察仅是整个战役的一半，另一半则是怎样解决多重共线性的问题。同样没有什么十拿九稳的办法。只有那么几条经验法则，其中包括：（1）利用外部或先验信息；（2）横截面数据与时间序列数据并用；（3）剔除一个高度共线性的变量；（4）数据转换；（5）获取或补充新的数据。当然，哪条法则在实践中灵验，要看数据的性质和共线性问题的严重程度。

5. 我们曾看到多重共线性在预测中的作用，并指出除非共线性的结构继续存在于未来的样本之中，否则，利用受到多重共线性困扰的回归估计进行预报，将是冒险的。

6. 虽然多重共线性在文献中备受关注（有人甚至认为关注过多），但在经验研究中遇到的一个同等重要的问题是微数缺测性，即样本（容量）的微小性。按照戈德伯格的意见，当一篇研究论文在抱怨多重共线性时，读者应把“多重共线性”换为“微数缺测性”，看看这种抱怨是否有道理。[①] 他建议，读者需决定观测次数 n 要多小，才能认为遇上了样本微小性问题，正如人们需决定辅助回归中的 R_i^2 值要多高，才能宣称一个共线性问题是非常严重的。

习　　题

问答题

10.1　在 k 变量模型中有 k 个正规方程用以估计 k 个未知数。这些正规方程见于附录 C。假定 X_k 是其余 X 变量的一个完全的线性组合，你如何说明在这种情形中不可能估计这 k 个回归系数？

10.2　考虑表 10—11 中的一组假想数据。假如你要用如下模型拟合数据：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

a. 你能估计这三个未知数吗？为什么？

b. 如果不能，那么你能估计这些参数的线性组合，即可估函数是什么？说明必要的计算。

表 10—11

Y	X_2	X_3
−10	1	1
−8	2	3
−6	3	5
−4	4	7
−2	5	9
0	6	11
2	7	13
4	8	15
6	9	17
8	10	19
10	11	21

10.3　参照第 8 章中讨论的儿童死亡率的例子。此例涉及儿童死亡率（CM）对人均 GNP

① Goldberger, op. cit., p. 250.

（即 PGNP）和妇女识字率（FLR）的回归。现在假设我们增加变量总人口出生率（TFR），得到如下回归结果。

Dependent Variable: CM

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	168.3067	32.89165	5.117003	0.0000
PGNP	-0.005511	0.001878	-2.934275	0.0047
FLR	-1.768029	0.248017	-7.128663	0.0000
TFR	12.86864	4.190533	3.070883	0.0032

R-squared	0.747372	Mean dependent var.	141.5000
Adjusted R-squared	0.734740	S.D. dependent var.	75.97807
S.E. of regression	39.13127	Akaike info criterion	10.23218
Sum squared resid.	91875.38	Schwarz criterion	10.36711
Log likelihood	-323.4298	F-statistic	59.16767
Durbin-Watson stat.	2.170318	Prob(F-statistic)	0.000000

a. 将这些回归结果与方程（8.1.4）中给出的结果相比较。你看到了什么变化？你又如何解释这些变化？

b. 值得在模型中增加变量 TFR 吗？为什么？

c. 既然所有的 t 系数都是个别统计显著的，我们能否说此时不存在共线性问题？

10.4 如果关系式 $\lambda_1 X_{1i}+\lambda_2 X_{2i}+\lambda_3 X_{3i}=0$ 对所有 λ_1，λ_2 和 λ_3 值都成立，试估计 $r_{12.3}$，$r_{13.2}$ 和 $r_{23.1}$。再求 $R^2_{1.23}$，$R^2_{2.13}$ 和 $R^2_{3.12}$。在此情形中多重共线性的程度如何？注：$R^2_{1.23}$ 是 Y 对 X_2 和 X_3 回归中的判定系数。类似地解释其他 R^2 值。

10.5 考虑以下模型：

$$Y_t=\beta_1+\beta_2 X_t+\beta_3 X_{t-1}+\beta_4 X_{t-2}+\beta_5 X_{t-3}+\beta_6 X_{t-4}+u_t$$

其中 Y=消费，X=收入，t=时间。上述模型假定 t 时期的消费不仅是 t 时期收入的函数，而且是以前多期收入的函数。例如，2000 年第一季度的消费支出是同季度收入和 1999 年四个季度收入的函数。这类模型叫做**分布滞后模型**（distributed lag models）。我们将在后面的一章加以讨论。

a. 你预期在这类模型中存在多重共线性吗？为什么？

b. 如果预期有多重共线性，你会怎样解决这个问题？

10.6 考虑 10.6 节的说明性例子（例 10.1）。你会怎样解释方程（10.6.1）和（10.6.4）所得到的边际消费倾向中的差异？

10.7 在涉及诸如 GNP、货币供给、价格、收入、失业等时间序列的数据中，一般都疑虑存在多重共线性，为什么？

10.8 设想在如下模型中

$$Y_i=\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+u_i$$

其中 X_2 和 X_3 之间的相关系数 r_{23} 为零。因此，某人建议你做如下回归：

$$Y_i=\alpha_1+\alpha_2 X_{2i}+u_{1i}$$

$$Y_i=\gamma_1+\gamma_3 X_{3i}+u_{2i}$$

a. 会不会有 $\hat{\alpha}_2=\hat{\beta}_2$ 且 $\hat{\gamma}_3=\hat{\beta}_3$ 呢？为什么？

b. $\hat{\beta}_1$ 会等于 $\hat{\alpha}_1$ 或 $\hat{\gamma}_1$ 或两者的某个线性组合吗？

c. 是否 var $(\hat{\beta}_2)$ = var $(\hat{\alpha}_2)$ 且 var $(\hat{\beta}_3)$ = var $(\hat{\gamma}_3)$？

10.9 参照第7章的说明性例子。在该例中，我们对2005年美国所有50个州和华盛顿特区的制造业部门拟合了柯布-道格拉斯生产函数。由方程（7.9.4）给出的回归结果表明，劳动和资本的系数都是个别统计显著的。

a. 判明劳动和资本两个变量是否高度相关。

b. 如果你对（a）的回答是肯定的，你会不会从模型中剔除劳动变量（比方说），而仅对资本投入作产出变量的回归呢？

c. 如果你这样做，你将犯哪一种设定偏误？找出这种偏误的性质。

10.10 参照例7.4。这个问题的相关矩阵如下：

	X_i	X_i^2	X_i^3
X_i	1	0.974 2	0.928 4
X_i^2		1.0	0.987 2
X_i^3			1.0

a. 对“由于零阶相关非常之高，必定有严重多重共线性”的说法加以评论。

b. 你会从模型中剔除 X_i^2 和 X_i^3 吗？

c. 如果你把它们剔除，X_i 的系数值将会出现什么情况？

10.11 **逐步回归**。为决定一个回归模型的“最优”解释变量集，研究者常用逐步回归的方法。在此方法中，既可采取每次引进一个 X 变量**逐步向前回归**（stepwise forward regression）的程序，也可先把所有可能的 X 变量都放在一个多元回归中，然后逐一地把它们剔除**逐步向后回归**（stepwise backward regression）。加进或剔除一个变量，通常是根据 F 检验看它对ESS的贡献而作出决定的。根据你现在对多重共线性的认识，你赞成某种逐步程序吗？为什么？①

10.12 判断如下命题是正确、错误还是不确定，并说明理由。

a. 尽管有完全多重共线性，OLS估计量仍然是BLUE。

b. 在高度多重共线性的情形中，要评价一个或多个偏回归系数的个别显著性是不可能的。

c. 如果有某一辅助回归显示出高的 R_i^2 值，则高度共线性的存在便确定无疑。

d. 变量的两两高度相关并不表示高度多重共线性。

e. 如果分析的目的仅仅是预测，则多重共线性是无害的。

f. 其他条件不变，VIF越高，OLS估计量的方差越大。

g. 和VIF相比，容许度（TOL）是多重共线性的更好度量指标。

h. 如果在多元回归中，根据通常的 t 检验，全部偏斜率系数都是个别统计上不显著的，你就不会得到一个高的 R^2 值。

i. 在 Y 对 X_2 和 X_3 的回归中，假如 X_3 的值很少变化，就会使 $\text{var}(\hat{\beta}_3)$ 增大，在极端的情形下，如果全部 X_3 值都相同，$\text{var}(\hat{\beta}_3)$ 将是无穷大。

10.13

a. 证明：如果对 $i=2, 3, \cdots, k$，$r_{1i}=0$，则：

$$R_{1.23\cdots k} = 0$$

b. 对于变量 $X_1(=Y)$ 对 $X_2, X_3, \cdots, X_k$ 的回归来说，这一发现有什么重要意义？

10.14 假如 $X_1(=Y), X_2, \cdots, X_k$ 的全部零阶相关系数都等于 r。

① 参看 Arthur S. Goldberger and D. B. Jochems, “Note on Stepwise Least-Squares,” *Journal of the American Statistical Association*, vol. 56, March 1961, pp. 105-110。比较一下你的理解是否和这些作者一致。

a. $R^2_{1.23\cdots k}$ 值是多少？

b. 一阶相关系数的值是多少？

* 10.15　可以证明，用矩阵表述有（见附录 C）：

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

a. 当 X 变量之间有完全共线性时，$\hat{\boldsymbol{\beta}}$ 会发生什么情况？

b. 你怎样知道有没有完全共线性？

* 10.16　用矩阵符号表示，我们可以证明方差协方差矩阵：

$$\text{var}-\text{cov}\ (\hat{\boldsymbol{\beta}})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

在（a）有完全多重共线性和（b）高度但并非完全共线性的情况下，上述方差协方差矩阵会分别出现什么情况？

* 10.17　考虑如下的**相关矩阵**（correlation matrix）：

$$\mathbf{R}=\begin{array}{c} \\ X_2 \\ X_3 \\ \cdots \\ X_k \end{array}\begin{array}{c} \begin{array}{cccc} X_2 & X_3 & \cdots & X_k \end{array} \\ \begin{bmatrix} 1 & r_{23} & \cdots & r_{2k} \\ r_{32} & 1 & \cdots & r_{3k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k2} & r_{k3} & \cdots & 1 \end{bmatrix} \end{array}$$

你怎样从相关矩阵看出是否（a）有完全共线性，（b）有不完全的共线性，以及（c）各个 X 不相关。

提示：你可利用 $\mathbf{R}$ 的行列式 $|\mathbf{R}|$ 来回答这些问题。

* 10.18　**正交解释变量**。假设在如下模型中：

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\cdots+\beta_kX_{ki}+u_i$$

从 X_2 到 X_k 各不相关。这样的变量叫做**正交变量**（orthogonal variables）。在这种情形中，

a. $(\mathbf{X}'\mathbf{X})$ 矩阵的结构将是怎样的？

b. 你将怎样求 $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$？

c. $\hat{\boldsymbol{\beta}}$ 的方差协方差矩阵具有何种性质？

d. 假如你在做完回归之后，想再引进另一正交变量 X_{k+1} 到模型中来，你需要重新计算先前的系数 $\hat{\beta}_1$ 至 $\hat{\beta}_k$ 吗？为什么？

10.19　考虑如下模型：

$$\text{GNP}_t=\beta_1+\beta_2\text{M}_t+\beta_3\text{M}_{t-1}+\beta_4(\text{M}_t-\text{M}_{t-1})+u_t$$

其中，$\text{GNP}_t=t$ 时期的 GNP，$\text{M}_t=t$ 时期的货币供给，$\text{M}_{t-1}=t-1$ 期的货币供给，$\text{M}_t-\text{M}_{t-1}=$从 $t-1$ 期到 t 时期货币供给的变化。也就是，此模型设想 t 时期的 GNP 是 t 时期和 $t-1$ 期的货币供给以及此期间货币供给变化量的函数。

a. 假定你拥有估计上述模型的数据，你能成功地估计出模型的全部系数吗？为什么？

b. 如果不能，那么什么系数可以估计？

c. 假使 $\beta_3\text{M}_{t-1}$ 一项不在模型中出现，你对（a）的回答仍然一样吗？

d. 重做（c），但现在假定 $\beta_2\text{M}_t$ 不出现。

10.20　证明方程（7.4.7）和（7.4.8）还可表示为：

$$\hat{\beta}_2=\frac{(\sum y_ix_{2i})(\sum x_{3i}^2)-(\sum y_ix_{3i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2)(1-r_{23}^2)}$$

---

* 选做题。

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2)(1 - r_{23}^2)}$$

其中 r_{23} 是 X_2 和 X_3 的相关系数。

10.21 利用方程（7.4.12）和（7.4.15），证明当存在完全共线性时，$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差是无穷大。

10.22 证明由方程（10.5.6）和（10.5.7）估计的斜率系数总和的标准误分别是 0.154 9 和 0.182 5。（见 10.5 节。）

10.23 对于 k 变量回归模型，可以证明方程（7.5.6）中给出的第 k 个（$k = 2, 3, \cdots, k$）偏回归系数 $\hat{\beta}_k$ 的方差可表示为[①]：

$$\text{var}(\hat{\beta}_k) = \frac{1}{n-k} \frac{\sigma_y^2}{\sigma_k^2} \left(\frac{1 - R^2}{1 - R_k^2} \right)$$

其中 $\sigma_y^2 = Y$ 的方差，σ_k^2 =第 k 个解释变量的方差，$R_k^2 = X_k$ 对其余 X 变量的回归中的判定系数，R^2 =Y 对全部 X 变量的回归中的判定系数。

a. 其他情况不变，如果 σ_k^2 增加，var($\hat{\beta}_k$) 会出现什么情况？这时多重共线性问题有什么含义？

b. 如果共线性是完全的，上述公式会出现什么情况？

c. 判断正误："$\hat{\beta}_k$ 的方差随 R^2 上升而下降，因此由高的 R_k^2 产生的影响可由高的 R^2 来抵消。"

10.24 根据 1899—1922 年美国制造业部门的年度数据，多尔蒂（Dougherty）获得如下回归结果[②]：

$$\widehat{\log Y} = 2.81 - 0.53 \log K + 0.91 \log L + 0.047t \qquad (1)$$

$$\text{se} = (1.38) \quad (0.34) \quad (0.14) \quad (0.021)$$

$$R^2 = 0.97 \qquad F = 189.8$$

其中 Y =真实产出指数，K =真实资本投入指数，L =真实劳动投入指数，t=时间或趋势。log 表示自然对数。

利用同样数据，他又获得以下回归：

$$\widehat{\log(Y/L)} = -0.11 + 0.11 \log (K/L) + 0.006t \qquad (2)$$

$$\text{se} = (0.03) \quad (0.15) \quad (0.006)$$

$$R^2 = 0.65 \qquad F = 19.5$$

a. 回归（1）中有没有多重共线性？你怎样知道？

b. 在回归（1）中，log K 的先验符号是什么？结果是否与预期相一致？为什么？

c. 你怎样替回归的函数形式（1）做辩护？（提示：柯布-道格拉斯生产函数。）

d. 解释回归（1）。在此回归中趋势变量有什么作用？

e. 回归（2）的道理何在？

f. 如果原先的回归（1）有多重共线性，是否已被回归（2）减弱？你怎样知道？

g. 如果回归（2）被看作回归（1）的一个受约束形式，作者施加的约束是什么呢？（提示：规模报酬。）你怎样知道这个约束是否正确？你用哪一种检验？说明你的计算。

h. 两个回归的 R^2 值是可比的吗？为什么？如果它们现在的形式不可比，你会怎样使得它们可比？

---

① 此公式见于 R. Stone, "The Analysis of Market Demand," *Journal of the Royal Statistical Society*, vol. B7, 1945, p. 297，还可回顾方程（7.5.6）。进一步的讨论，见于 Peter Kennedy, *A Guide to Econometrics*, 2d ed., The MIT Press, Cambridge, Mass., 1985, p. 156。

② Christopher Dougherty, *Introduction to Econometrics*, Oxford University Press, New York, 1992, pp. 159-160.

10.25 批判性地评价如下命题：

a. “多重共线性实际上不是一个建模的错误，而是数据不充分的一种状况。”①

b. “如果不能得到更多的数据，那就必须接受数据包含有限信息量的事实并相应地设定模型。试图估计过分复杂的模型是经验丰富的应用计量经济学家最常见的错误之一。”②

c. “研究者通常认为，只要在回归结果中没有看到他们预先假设的符号，他们先验推定重要的变量具有不显著的 t 值，或者去掉一个解释变量会导致各种回归结果都明显变化，那就是多重共线性在作怪。不幸的是，这些条件中没有一个是存在共线性的充分或必要条件，而且对于解决他们提出的估计问题需要什么样的额外信息没有提供任何有用的建议。”③

d. “……任何包含多于四个自变量的时间序列回归都会带来垃圾。”④

实证分析题

10.26 克莱因和戈德伯格试图对美国经济拟合如下回归模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

其中 Y =消费，X_2 =工资收入，X_3 =非工资、非农场收入，X_4 =农场收入。但他们预料 X_2，X_3 和 X_4 高度共线，因此通过横截面分析把 β_3 和 β_4 估计为 $\beta_3 = 0.75\beta_2$ 和 $\beta_4 = 0.625\beta_2$。利用这些估计，他们重新建立他们的消费函数如下：

$$Y_i = \beta_1 + \beta_2(X_{2i} + 0.75X_{3i} + 0.625X_{4i}) + u_i = \beta_1 + \beta_2 Z_i + u_i$$

其中，$Z_i = X_{2i} + 0.75X_{3i} + 0.625X_{4i}$。

a. 用这个修改的模型去拟合表 10—12 所附数据，并估计 β_1 至 β_4。

b. 你会怎样解释变量 Z？

表 10—12

年份	Y	X_2	X_3	X_4	年份	Y	X_2	X_3	X_4
1936	62.8	43.41	17.10	3.96	1946	95.7	76.73	28.26	9.76
1937	65.0	46.44	18.65	5.48	1947	98.3	75.91	27.91	9.31
1938	63.9	44.35	17.09	4.37	1948	100.3	77.62	32.30	9.85
1939	67.5	47.82	19.28	4.51	1949	103.2	78.01	31.39	7.21
1940	71.3	51.02	23.24	4.88	1950	108.9	83.57	35.61	7.39
1941	76.6	58.71	28.11	6.37	1951	108.5	90.59	37.58	7.98
1945*	86.3	87.69	30.29	8.96	1952	111.4	95.47	35.17	7.42

注：*战争年代 1942—1944 年的数据缺失。其他年份的数据以 1939 年十亿美元计。

资料来源：L. R. Klein and A. S. Goldberger, *An Economic Model of the United States, 1929—1952*, North Holland Publishing Company, Amsterdam, 1964, p. 131.

10.27 表 10—13 给出 1975—2005 年期间美国进口（Imports）、GDP 和消费者价格指数（CPI）数据。

---

① Samprit Chatterjee, Ali S. Hadi, and Bertram Price, *Regression Analysis by Example*, 3d ed., John Wiley & Sons, New York, 2000, p. 226.

② Russel Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993, p. 186.

③ Peter Kennedy, *A Guide to Econometrics*, 4th ed., MIT Press, Cambridge, Mass., 1998, p. 187.

④ 此段引文是已故计量经济学家兹维·格里利谢斯所说，引自 Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary*, Addison Wesley, Reading, Mass., 1991, p. 224。

表 10—13　　1975—2005 年美国商品进口（Imports）、GDP 和 CPI

年份	CPI	GDP	Imports	年份	CPI	GDP	Imports
1975	53.8	1 638.3	98 185	1991	136.2	5 995.9	491 020
1976	56.9	1 825.3	124 228	1992	140.3	6 337.7	536 528
1977	60.6	2 030.9	151 907	1993	144.5	6 657.4	589 394
1978	65.2	2 294.7	176 002	1994	148.2	7 072.2	668 690
1979	72.6	2 563.3	212 007	1995	152.4	7 397.7	749 374
1980	82.4	2 789.5	249 750	1996	156.9	7 816.9	803 113
1981	90.9	3 128.4	265 067	1997	160.5	8 304.3	876 470
1982	96.5	3 225.0	247 642	1998	163.0	8 747.0	917 103
1983	99.6	3 536.7	268 901	1999	166.6	9 268.4	1 029 980
1984	103.9	3 933.2	332 418	2000	172.2	9 817.0	1 224 408
1985	107.6	4 220.3	338 088	2001	177.1	10 128.0	1 145 900
1986	109.6	4 462.8	368 425	2002	179.9	10 469.6	1 164 720
1987	113.6	4 739.5	409 765	2003	184.0	10 960.8	1 260 717
1988	118.3	5 103.8	447 189	2004	188.9	11 712.5	1 472 926
1989	124.0	5 484.4	477 665	2005	195.3	12 455.8	1 677 371
1990	130.7	5 803.1	498 438				

注：数据针对所有城镇消费者；除非特别指出，以 1982—1984 年为基年，1982—1984＝100。
资料来源：Department of Labor，Bureau of Labor Statistics.

请你考虑以下模型：

$$\ln \text{Imports}_t = \beta_1 + \beta_2 \ln \text{GNP}_t + \beta_3 \ln \text{CPI}_t + u_t$$

a. 用表中数据估计此模型的参数。

b. 你猜想数据中有多重共线性吗？

c. 利用病态指数，分析共线性的性质。

d. 做回归：(1) $\ln \text{Imports}_t = A_1 + A_2 \ln \text{GDP}_t$

(2) $\ln \text{Imports}_t = B_1 + B_2 \ln \text{CPI}_t$

(3) $\ln \text{GDP}_t = C_1 + C_2 \ln \text{CPI}_t$

根据这些回归你能对数据中多重共线性的性质说些什么？

e. 假使数据有多重共线性，但 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 在 5%水平上是个别显著的，并且总的 F 检验也是显著的。对这样的情形，我们用不用考虑共线性问题？

10.28　参考习题 7.19 关于美国对鸡肉的需求函数。

a. 利用对数线性（双对数）模型估计各种辅助回归，一共有多少个这样的回归？

b. 你怎样从这些辅助回归决定哪些回归元是高度共线性的？你用哪一种检验？说明你的计算细节。

c. 如果数据中有显著的共线性，你会剔除哪个（些）变量以减少共线性问题的严重性？如果你这样做，你会遇到什么计量经济学问题？

d. 除了剔除变量之外，你有什么建议可以缓解共线性问题？作出解释。

10.29　表 10—14 给出作为若干变量的函数的美国新客车出售数据。

a. 给出一个适当的线性或对数线性模型，以估计美国对汽车的需求函数。

b. 如果你决定用表中全部回归元作为解释变量，你预料会遇到多重共线性的问题吗？为什么？

c. 如果你这样预期，你准备怎样解决这个问题？明确你的假定并说明全部计算。

表 10—14

年份	Y	X_2	X_3	X_4	X_5	X_6
1971	10 227	112.0	121.3	776.8	4.89	79 367
1972	10 872	111.0	125.3	839.6	4.55	82 153
1973	11 350	111.1	133.1	949.8	7.38	85 064
1974	8 775	117.5	147.7	1 038.4	8.61	86 794
1975	8 539	127.6	161.2	1 142.8	6.16	85 846
1976	9 994	135.7	170.5	1 252.6	5.22	88 752
1977	11 046	142.9	181.5	1 379.3	5.50	92 017
1978	11 164	153.8	195.3	1 551.2	7.78	96 048
1979	10 559	166.0	217.7	1 729.3	10.25	98 824
1980	8 979	179.3	247.0	1 918.0	11.28	99 303
1981	8 535	190.2	272.3	2 127.6	13.73	100 397
1982	7 980	197.6	286.6	2 261.4	11.20	99 526
1983	9 179	202.6	297.4	2 428.1	8.69	100 834
1984	10 394	208.5	307.6	2 670.6	9.65	105 005
1985	11 039	215.2	318.5	2 841.1	7.75	107 150
1986	11 450	224.4	323.4	3 022.1	6.31	109 597

注：Y＝新客车出售量，千辆，未经季节调整数据；

X_2 ＝ 新车，消费者价格指数，1967＝100，未经季节调整；

X_3 ＝ 消费者价格指数，全部项目，全部城市消费者，1967＝100，未经季节调整；

X_4 ＝ 个人可支配收入（PDI），十亿美元，未经季节调整；

X_5 ＝ 利率，百分数，直接使用金融公司票据（placed directly）；

X_6 ＝ 民间就业劳动人数，千人，未经季节调整。

资料来源：*Business Statistics, 1986*, A Supplement to the *Current Survey of Business*, U. S. Department of Commerce.

10.30 为了评价年度最低工资保障（负收入税）政策的可行性，兰德公司（Rand Corporation）进行了一项研究，以评价劳动供给（平均工作小时数）对小时工资提高的反应。[①] 此研究中的数据取自 6 000 户男户主年收入低于 15 000 美元的一个国民样本。这些数据被分成 39 个人口组，并放在表 10—15 中。由于 4 个人口组中的某些变量缺失，所以此表中只给出了 35 个组的数据。用于分析的各个变量的定义在表末给出。

a. 将该年度平均工作小时数对表中变量进行回归，并解释你的回归。

b. 数据中存在多重共线性的证据吗？你如何知道？

c. 计算各个回归元的方差膨胀因子（VIF）和 TOL 指标。

d. 若存在多重共线性问题，那你会采用什么补救措施（如果有的话）？

e. 此研究对负收入税的可行性有何结论？

① D. H. Greenberg and M. Kosters, *Income Guarantees and the Working Poor*, Rand Corporation, R-579-OEO, December 1970.

表 10—15　　35 个人口组的工作小时数及其他数据

观测	Hours	Rate	ERSP	ERNO	NEIN	Assets	Age	DEP	School
1	2 157	2.905	1 121	291	380	7 250	38.5	2.340	10.5
2	2 174	2.970	1 128	301	398	7 744	39.3	2.335	10.5
3	2 062	2.350	1 214	326	185	3 068	40.1	2.851	8.9
4	2 111	2.511	1 203	49	117	1 632	22.4	1.159	11.5
5	2 134	2.791	1 013	594	730	12 710	57.7	1.229	8.8
6	2 185	3.040	1 135	287	382	7 706	38.6	2.602	10.7
7	2 210	3.222	1 100	295	474	9 338	39.0	2.187	11.2
8	2 105	2.493	1 180	310	255	4 730	39.9	2.616	9.3
9	2 267	2.838	1 298	252	431	8 317	38.9	2.024	11.1
10	2 205	2.356	885	264	373	6 789	38.8	2.662	9.5
11	2 121	2.922	1 251	328	312	5 907	39.8	2.287	10.3
12	2 109	2.499	1 207	347	271	5 069	39.7	3.193	8.9
13	2 108	2.796	1 036	300	259	4 614	38.2	2.040	9.2
14	2 047	2.453	1 213	297	139	1 987	40.3	2.545	9.1
15	2 174	3.582	1 141	414	498	10 239	40.0	2.064	11.7
16	2 067	2.909	1 805	290	239	4 439	39.1	2.301	10.5
17	2 159	2.511	1 075	289	308	5 621	39.3	2.486	9.5
18	2 257	2.516	1 093	176	392	7 293	37.9	2.042	10.1
19	1 985	1.423	553	381	146	1 866	40.6	3.833	6.6
20	2 184	3.636	1 091	291	560	11 240	39.1	2.328	11.6
21	2 084	2.983	1 327	331	296	5 653	39.8	2.208	10.2
22	2 051	2.573	1 194	279	172	2 806	40.0	2.362	9.1
23	2 127	3.262	1 226	314	408	8 042	39.5	2.259	10.8
24	2 102	3.234	1 188	414	352	7 557	39.8	2.019	10.7
25	2 098	2.280	973	364	272	4 400	40.6	2.661	8.4
26	2 042	2.304	1 085	328	140	1 739	41.8	2.444	8.2
27	2 181	2.912	1 072	304	383	7 340	39.0	2.337	10.2
28	2 186	3.015	1 122	30	352	7 292	37.2	2.046	10.9
29	2 188	3.010	990	366	374	7 325	38.4	2.847	10.6
30	2 077	1.901	350	209	95	1 370	37.4	4.158	8.2
31	2 196	3.009	947	294	342	6 888	37.5	3.047	10.6
32	2 093	1.899	342	311	120	1 425	37.5	4.512	8.1
33	2 173	2.959	1 116	296	387	7 625	39.2	2.342	10.5
34	2 179	2.971	1 128	312	397	7 779	39.4	2.341	10.5
35	2 200	2.980	1 126	204	393	7 885	39.2	2.341	10.6

注：Hours＝该年度平均工作小时数；
Rate＝平均小时工资，美元；
ERSP＝配偶年均收入，美元；
ERNO＝其他家庭成员的年均收入，美元；
NEIN＝年均非劳动收入；
Assets＝平均家庭资产拥有量（银行存款等），美元；
Age＝被调查者的平均年龄；
DEP＝平均赡养人数；
School＝平均完成的最高年级。

资料来源：D. H. Greenberg and M. Kosters，*Income Guarantees and the Working Poor*，Rand Corporation，R-579-OEO，December 1970.

10.31　表 10—16 给出了 1960 年美国 47 个州的犯罪率数据。试用一个适当的模型来解释犯罪率与表中 14 个社会经济变量的关系。在给出你的模型时，特别注意共线性问题。

表 10—16　　1960 年美国 47 个州的犯罪数据

观测	R	Age	S	ED	EX_0	EX_1	LF	M	N	NW	U_1	U_2	W	X
1	79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
2	163.5	143	0	113	103	95	583	1 012	13	102	96	36	557	194
3	57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
4	196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
5	123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
6	68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
7	96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
8	155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
9	85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
10	70.5	140	0	118	71	68	632	1 029	7	15	100	24	526	174
11	167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
12	84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
13	51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
14	66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
15	79.8	152	1	87	57	53	530	986	30	72	92	43	405	264
16	94.6	142	1	88	81	77	497	956	33	321	116	47	427	247
17	53.9	143	0	110	66	63	537	977	10	6	114	35	487	166
18	92.9	135	1	104	123	115	537	978	31	170	89	34	631	165
19	75.0	130	0	116	128	128	536	934	51	24	78	34	627	135
20	122.5	125	0	108	113	105	567	985	78	94	130	58	626	166
21	74.2	126	0	108	74	67	602	984	34	12	102	33	557	195
22	43.9	157	1	89	47	44	512	962	22	423	97	34	288	276
23	121.6	132	0	96	87	83	564	953	43	92	83	32	513	227
24	96.8	131	0	116	78	73	574	1 038	7	36	142	42	540	176
25	52.3	130	0	116	63	57	641	984	14	26	70	21	486	196
26	199.3	131	0	121	160	143	631	1 071	3	77	102	41	674	152
27	34.2	135	0	109	69	71	540	965	6	4	80	22	564	139
28	121.6	152	0	112	82	76	571	1 018	10	79	103	28	537	215
29	104.3	119	0	107	166	157	521	938	168	89	92	36	637	154
30	69.6	166	1	89	58	54	521	973	46	254	72	26	396	237
31	37.3	140	0	93	55	54	535	1 045	6	20	135	40	453	200
32	75.4	125	0	109	90	81	586	964	97	82	105	43	617	163
33	107.2	147	1	104	63	64	560	972	23	95	76	24	462	233
34	92.3	126	0	118	97	97	542	990	18	21	102	35	589	166
35	65.3	123	0	102	97	87	526	948	113	76	124	50	572	158

续前表

观测	R	Age	S	ED	EX_0	EX_1	LF	M	N	NW	U_1	U_2	W	X
36	127.2	150	0	100	109	98	531	964	9	24	87	38	559	153
37	83.1	177	1	87	58	56	638	974	24	349	76	28	382	254
38	56.6	133	0	104	51	47	599	1 024	7	40	99	27	425	225
39	82.6	149	1	88	61	54	515	953	36	165	86	35	395	251
40	115.1	145	1	104	82	74	560	981	96	126	88	31	488	228
41	88.0	148	0	122	72	66	601	998	9	19	84	20	590	144
42	54.2	141	0	109	56	54	523	968	4	2	107	37	489	170
43	82.3	162	1	99	75	70	522	996	40	208	73	27	496	224
44	103.0	136	0	121	95	96	574	1 012	29	36	111	37	622	162
45	45.5	139	1	88	46	41	480	968	19	49	135	53	457	249
46	50.8	126	0	104	106	97	599	989	40	24	78	25	593	171
47	84.9	130	0	121	90	91	623	1 049	3	22	113	40	588	160

注：变量定义：R=犯罪率，每百万人口中向警察报告的违法次数；
Age=每千人中年龄在14～24岁的男性人数；
S=位于南方与否的指标变量（0=否，1=是）；
ED=25岁及25岁以上人口读书年数的均值乘以10；
EX_0=1960年州和地方政府对警方的人均支出；
EX_1=1959年州和地方政府对警方的人均支出；
LF=每千名14～24岁城镇男性居民的劳动力参与率；
M=每千名女性对应的男性人数；
N=以十万计的州人口规模；
NW=每千人中非白人的人口数；
U_1=每千名14～24岁城镇男性的失业率；
U_2=每千名35～39岁城镇男性的失业率；
W=以十美元计可转换商品和资产或家庭收入的中位数；
X=每千户中挣到中位数收入一半的家庭数；
观测=州（1960年的47个州）。

资料来源：W. Vandaele, "Participation in Illegitimate Activities: Erlich Revisted," in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation*, National Academy of Sciences, 1978, pp. 270-335.

10.32 参照10.10节中给出的朗利数据。去掉1962年的数据重做表中的回归；即做1947—1961年期间的回归。比较这两个回归。从此题中你能得到什么一般性结论？

10.33 **更新的朗利数据**。我们已经扩展了10.10节中给出的数据，使之包含1959—2005年的观测。新数据在表10—17中给出。这些数据包括Y=就业人数（以千计）；X_1=GNP暗含的价格缩减指数；X_2=GNP（以百万美元计）；X_3=失业人数（以千计）；X_4=武装部队中的人数（以千计）；X_5=16岁以上非收容人口数；X_6=年份，1959年等于1，1960年等于2，直至2005年等于47。

a. 根据本章提示的方法生成散点图，以评价各个自变量之间的关系。有很强的相关关系吗？这种关系看似线性的吗？

b. 生成一个相关矩阵。不考虑因变量，哪些变量的彼此相关看起来最为明显？

c. 做一个标准的OLS回归来预测以千计的就业人口数。自变量的系数与你的预期一致吗？

d. 基于以上结论，你相信这些数据存在共线性问题吗？

表 10—17　　更新的朗利数据：1959—2005 年

观测	Y	X_1	X_2	X_3	X_4	X_5	X_6
1959	64 630	82.908	509 300	3 740	2 552	120 287	1
1960	65 778	84.074	529 500	3 852	2 514	121 836	2
1961	65 746	85.015	548 200	4 714	2 573	123 404	3
1962	66 702	86.186	589 700	3 911	2 827	124 864	4
1963	67 762	87.103	622 200	4 070	2 737	127 274	5
1964	69 305	88.438	668 500	3 786	2 738	129 427	6
1965	71 088	90.055	724 400	3 366	2 722	131 541	7
1966	72 895	92.624	792 900	2 875	3 123	133 650	8
1967	74 372	95.491	838 000	2 975	3 446	135 905	9
1968	75 920	99.56	916 100	2 817	3 535	138 171	10
1969	77 902	104.504	990 700	2 832	3 506	140 461	11
1970	78 678	110.046	1 044 900	4 093	3 188	143 070	12
1971	79 367	115.549	1 134 700	5 016	2 816	145 826	13
1972	82 153	120.556	1 246 800	4 882	2 449	148 592	14
1973	85 064	127.307	1 395 300	4 365	2 327	151 476	15
1974	86 794	138.82	1 515 500	5 156	2 229	154 378	16
1975	85 846	151.857	1 651 300	7 929	2 180	157 344	17
1976	88 752	160.68	1 842 100	7 406	2 144	160 319	18
1977	92 017	170.884	2 051 200	6 991	2 133	163 377	19
1978	96 048	182.863	2 316 300	6 202	2 117	166 422	20
1979	98 824	198.077	2 595 300	6 137	2 088	169 440	21
1980	99 303	216.073	2 823 700	7 637	2 102	172 437	22
1981	100 397	236.385	3 161 400	8 273	2 142	174 929	23
1982	99 526	250.798	3 291 500	10 678	2 179	177 176	24
1983	100 834	260.68	3 573 800	10 717	2 199	179 234	25
1984	105 005	270.496	3 969 500	8 539	2 219	181 192	26
1985	107 150	278.759	4 246 800	8 312	2 234	183 174	27
1986	109 597	284.895	4 480 600	8 237	2 244	185 284	28
1987	112 440	292.691	4 757 400	7 425	2 257	187 419	29
1988	114 968	302.68	5 127 400	6 701	2 224	189 233	30
1989	117 342	314.179	5 510 600	6 528	2 208	190 862	31
1990	118 793	326.357	5 837 900	7 047	2 167	192 644	32
1991	117 718	337.747	6 026 300	8 628	2 118	194 936	33
1992	118 492	345.477	6 367 400	9 613	1 966	197 205	34
1993	120 259	353.516	6 689 300	8 940	1 760	199 622	35
1994	123 060	361.026	7 098 400	7 996	1 673	201 970	36
1995	124 900	368.444	7 433 400	7 404	1 579	204 420	37
1996	126 708	375.429	7 851 900	7 236	1 502	207 087	38
1997	129 558	381.663	8 337 300	6 739	1 457	209 846	39

续前表

观测	Y	X_1	X_2	X_3	X_4	X_5	X_6
1998	131 463	385.881	8 768 300	6 210	1 423	212 638	40
1999	133 488	391.452	9 302 200	5 880	1 380	215 404	41
2000	136 891	399.986	9 855 900	5 692	1 405	218 061	42
2001	136 933	409.582	10 171 600	6 801	1 412	220 800	43
2002	136 485	416.704	10 500 200	8 378	1 425	223 532	44
2003	137 736	425.553	11 017 600	8 774	1 423	226 223	45
2004	139 252	437.795	11 762 100	8 149	1 411	228 892	46
2005	141 730	451.946	12 502 400	7 591	1 378	231 552	47

资料来源：Department of Labor，Bureau of Labor Statistics and http：//siadapp. dmdc. osd. mil/personnel/MILITARY/ Miltop. htm.

*10.34 在奶酪熟化的过程中，有几个化学过程决定了最终产品的味道。表10—18中给出的数据就是30个成熟的切达奶酪（cheddar cheeses）样本中各种化学成分的浓度以及对每个样本口感的主观评价指标。变量Acetic和H_2S分别表示醋酸和硫化氢浓度的自然对数。变量Lactic表示乳酸浓度，而且没有进行对数变换。

a. 画出这四个变量的散点图。

b. 将Taste对Acetic和H_2S做一个二元回归并解释你的结果。

c. 将Taste对Lactic和H_2S做一个二元回归并解释你的结果。

d. 将Taste对Acetic、Lactic和H_2S做一个多元回归并解释你的结果。

e. 你对多重共线性有何认识，你对这些回归做何取舍？

f. 根据这些分析，你能得到什么总体结论？

表10—18　　奶酪中的化学成分

观测	Taste	Acetic	H_2S	Lactic
1	12.300 00	4.543 000	3.135 000	0.860 000
2	20.900 00	5.159 000	5.043 000	1.530 000
3	39.000 00	5.366 000	5.438 000	1.570 000
4	47.900 00	5.759 000	7.496 000	1.810 000
5	5.600 000	4.663 000	3.807 000	0.990 000
6	25.900 00	5.697 000	7.601 000	1.090 000
7	37.300 00	5.892 000	8.726 000	1.290 000
8	21.900 00	6.078 000	7.966 000	1.780 000
9	18.100 00	4.898 000	3.850 000	1.290 000
10	21.000 00	5.242 000	4.174 000	1.580 000
11	34.900 00	5.740 000	6.142 000	1.680 000
12	57.200 00	6.446 000	7.908 000	1.900 000
13	0.700 000	4.477 000	2.996 000	1.060 000
14	25.900 00	5.236 000	4.942 000	1.300 000

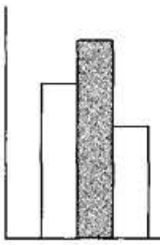
* 选做题。

续前表

观测	Taste	Acetic	H_2S	Lactic
15	54.900 00	6.151 000	6.752 000	1.520 000
16	40.900 00	3.365 000	9.588 000	1.740 000
17	15.900 00	4.787 000	3.912 000	1.160 000
18	6.400 000	5.142 000	4.700 000	1.490 000
19	18.000 00	5.247 000	6.174 000	1.630 000
20	38.900 00	5.438 000	9.064 000	1.990 000
21	14.000 00	4.564 000	4.949 000	1.150 000
22	15.200 00	5.298 000	5.220 000	1.330 000
23	32.000 00	5.455 000	9.242 000	1.440 000
24	56.700 00	5.855 000	10.199 00	2.010 000
25	16.800 00	5.366 000	3.664 000	1.310 000
26	11.600 00	6.043 000	3.219 000	1.460 000
27	26.500 00	6.458 000	6.962 000	1.720 000
28	0.700 000	5.328 000	3.912 000	1.250 000
29	13.400 00	5.802 000	6.685 000	1.080 000
30	5.500 000	6.176 000	4.787 000	1.250 000

资料来源： http：//lib.stat.cmu.edu/DASL/Datafiles/Cheese.html.

第 11 章 异方差性：误差方差不是常数会怎么样？



经典线性回归模型的一个重要假定（假定 4）是，出现在总体回归函数中的干扰项 u_i 是同方差性的；也就是说，它们都有相同的方差。本章中，我们分析这一假定的真实性，并探明如果此假定不成立将会出现什么情况。类似于第 10 章，我们寻求下述问题的答案：

1. 异方差性的性质是什么？
2. 它的后果是什么？
3. 怎样去发现它？
4. 有什么补救措施？

11.1 异方差的性质

如第 3 章中所指出的那样，经典线性回归模型的重要假定之一是，以给定解释变量值为条件的每一干扰项 u_i 的方差是一个等于 σ^2 的常数。这就是同方差性假定。**同方差性**（homoscedasticity）意谓相同的散布，即相等的方差。用符号表示为：

$$E(u_i^2) = \sigma^2 \qquad i = 1,2,\cdots,n \tag{11.1.1}$$

从图形上看，双变量回归模型中的同方差性可表示为图 3—4。为方便起见，将该图重制为图 11—1。如图 11—1 所示，以给定 X_i 为条件的 Y_i 的条件方差（等于 u_i 的条件方差），不管变量 X 取什么值，都保持不变。

与此相对，考虑图 11—2。该图表明，Y_i 的条件方差随 X 增加而增加。这里，

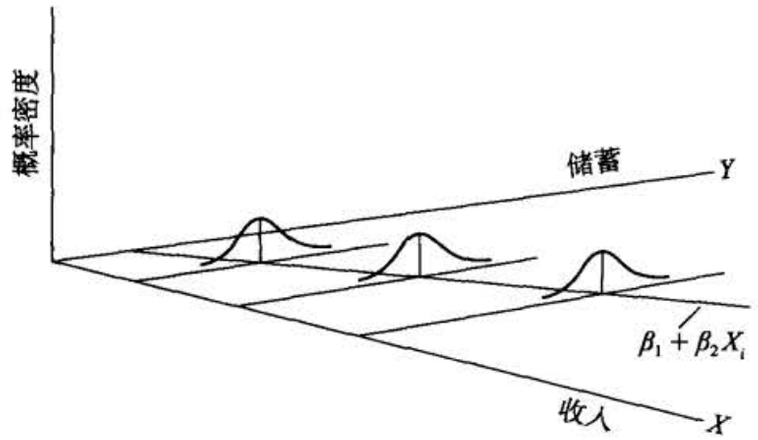


图 11—1　同方差性干扰

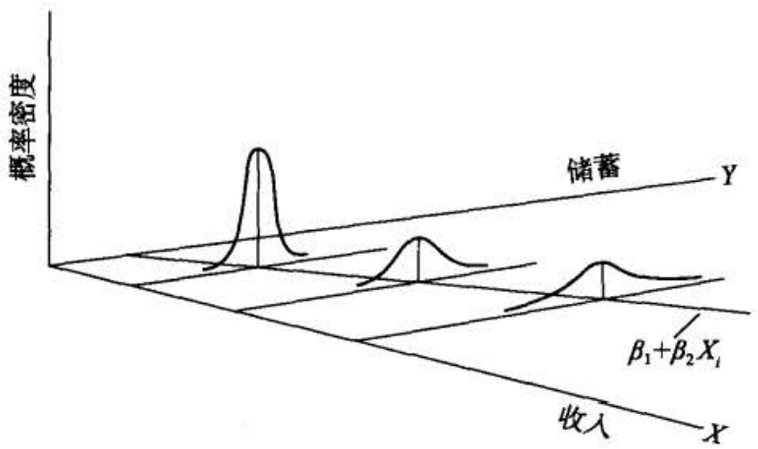


图 11—2　异方差性干扰

Y_i 的方差不再保持不变，从而有异方差性。符号上写为：

$$E(u_i^2)=\sigma_i^2 \tag{11.1.2}$$

注意，σ_i^2 的下标提醒我们，u_i 的条件方差（$=Y_i$ 的条件方差）不再是常数。

为了看清楚同方差性和异方差性的区别，假定在双变量模型 $Y_i=\beta_1+\beta_2X_i+u_i$ 中，Y 代表储蓄和 X 代表收入。图 11—1 和图 11—2 都表明随着收入增加，储蓄平均来说也增加。但在图 11—1 中，储蓄的方差在所有的收入水平上都保持不变。而在图 11—2 中，它却随收入增加而增加。看来在图 11—2 中，较高收入的家庭不仅比低收入的家庭平均而言有更多的储蓄，而且在他们的储蓄中有更大的变异。

有几个理由说明为什么 u_i 的方差可能有变化。其中的一些如下所述[①]：

1. 按照误差学习模型（error-learning models），人们在学习的过程中，其行为误差随时间或错误次数而减少。在这种情形中，预期 σ_i^2 会减小。作为一个例子，考虑图 11—3。该图描绘了一次测验；在给定的一段时间里，打字出错个数与用于打字练习的小时数的关系。如图 11—3 所示，随着打字练习的小时数的增加，不仅平均

① 参见 Stefan Valavanis，*Econometrics*，McGraw-Hill，New York，1959，p. 48。

打字出错个数有所下降，而且打字出错个数的方差也有所下降。

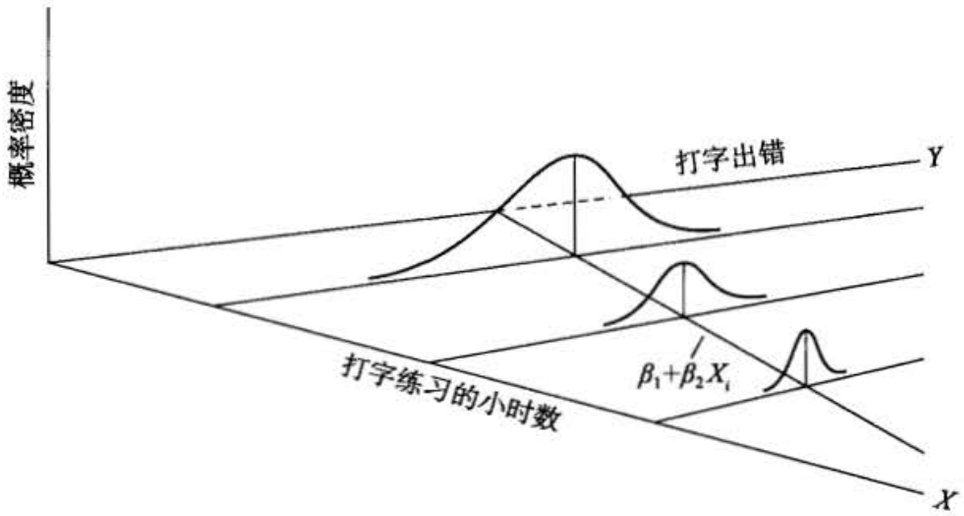


图 11—3　异方差性示例

2. 随着收入增长，人们有更多的可随意支配收入（discretionary income）[①]，从而如何支配他们的收入有更大的选择范围。因此，在做储蓄对收入的回归时，很可能发现，由于人们对其储蓄行为有更多的选择，σ_i^2 与收入俱增（如图 11—2 所示）。同理，利润较丰厚的公司在分红政策方面和利润微薄的公司相比，一般均可预料有较大的变化。而且以增长为导向（growth-oriented）的公司相对于已发展定型的公司，在红利支付方面也可能表现出更多的变异。

3. 随着数据采集技术的改进，σ_i^2 可能减小。例如，有成熟的数据处理设备的银行，在为客户提供的月度或季度报表中，相对于没有这种设备的银行，会出现更少的差错。

4. 异方差性还会因为**异常观测**（outliers）的出现而产生。一个超越正常范围的观测值或称异常观测，是指和其他观测值相比相差很多（非常小或非常大）的观测值。更具体地，异常观测是来自于与产生其余观测值的总体不同的另一个总体。[②] 包括或不包括这样的一个观测值，尤其是样本较小时，会在很大程度上改变回归分析的结果。

作为一个例子，考虑图 11—4 中的散点图。此图根据习题 11.22 中表 11—9 所给数据，对 20 个国家在第二次世界大战后直至 1969 年期间的股票价格（Y）和消费价格（X）的百分比变化进行描点。图中，对智利的观测值 Y 和 X，远大于对其他国家的观测值，故可视为一个异常观测。类似于这种情况，同方差性的假定就难以维持。在习题 11.22 中，我们要求读者考虑，如果在分析中把对智利的观测值除掉，会出现什么样的回归结果。

---

① 像 Valavanis 说的那样："收入增加了，人们现在只在意一元钱怎样用。而过去则在意一角钱怎样用。"参见 Stefan Valavanis，*Econometrics*，McGraw-Hill，New York，1959，p. 48。

② 感谢迈克尔·麦卡利尔（Michael McAleer）向我指出这一点。

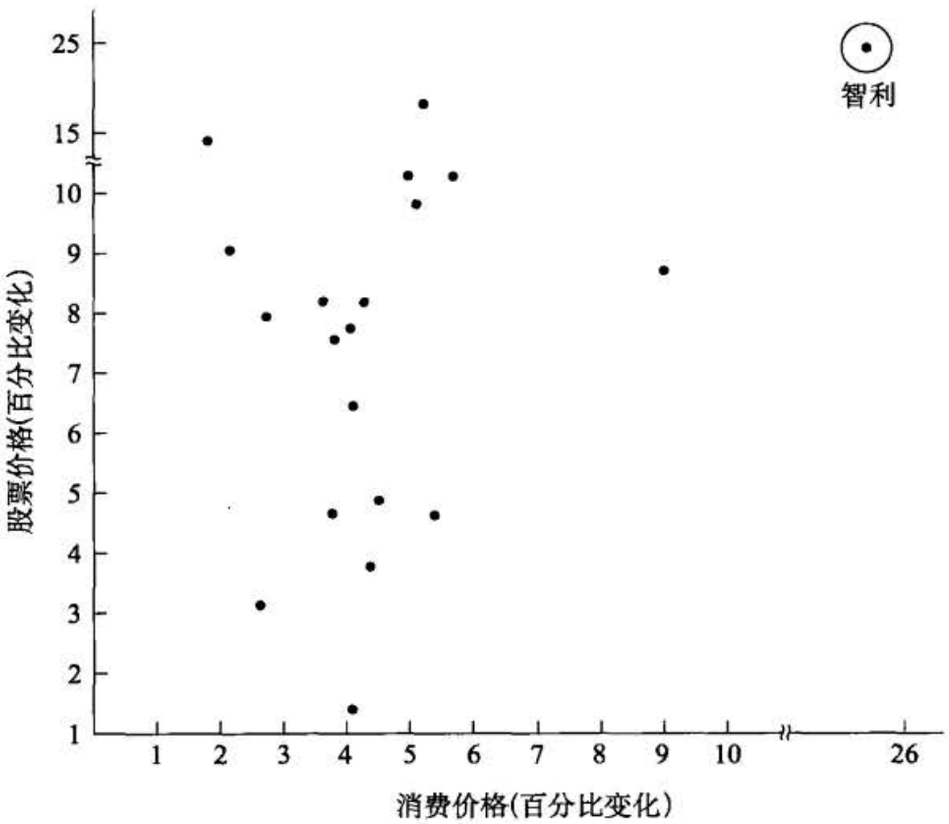


图 11—4　股票价格与消费价格的关系

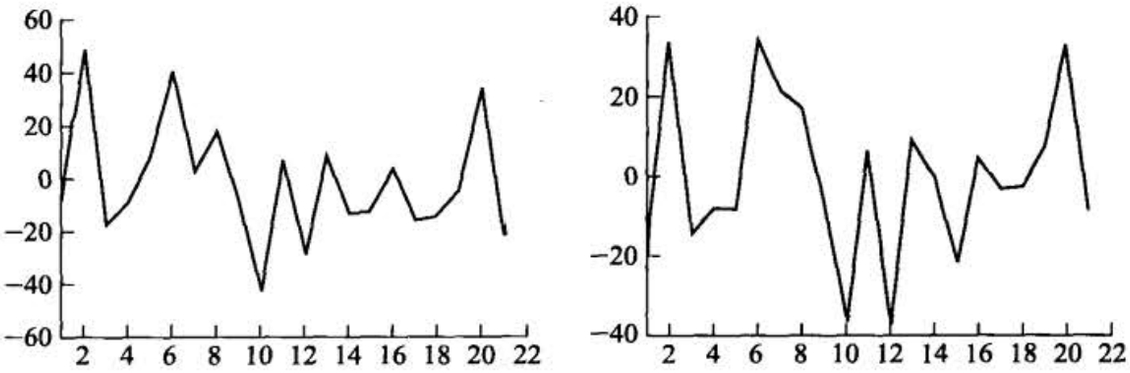
5. 异方差性的另一来源来自于对 CLRM 假定 9 的破坏，即回归模型的设定是不正确的。虽然我们将在第 13 章中对设定偏误的问题做更全面的探讨，但常常看来像是异方差性问题，其实是由于模型中的一些重要变量被忽略了。例如，在一个对商品的需求函数中，如果没有把有关的互补品和（或）替代品的价格包括进来（遗漏变量偏误），则回归残差可能给人以异方差的表面印象；而当模型把所忽略的变量包括进来时，这种印象也许会消失。

作为一个简明的例子，回想我们对广告印象（Y）与广告支出（X）之关系的研究。（参见习题 8.32。）若只将 Y 对 X 回归并观测此回归的残差，你会看到一种类型，但若将 Y 对 X 和 X^2 回归，你又会看到另一种类型，从图 11—5 明显可以看出这一点。我们已经看到，X^2 属于此模型。（参见习题 8.32。）

6. 异方差性的另一个来源是模型中一个或多个回归元的分布**偏态**（skewness）。诸如收入、财富和教育等经济变量都是很好的例子。众所周知，大多数社会中收入和财富的分配都是不匀称的，处在顶端的少数几人拥有大部分的收入和财富。

7. 异方差性的其他来源：如戴维·韩德瑞（David Hendry）所说的那样，由于（1）不正确的数据变形（如比率或一阶差分变换等）和（2）不正确的函数形式（如线性与对数线性模型的变换），同样能导致异方差性。[①]

① David F. Hendry, *Dynamic Econometrics*, Oxford University Press, 1995, p. 45.



(a) 广告印象 Y 对广告支出 X 进行回归的残差　(b) 广告印象 Y 对 X 和 X^2 进行回归的残差

图 11—5

注意，异方差性问题在横截面数据中比在时间序列数据中更为常见。在横截面数据中，人们通常在一个给定的时间点上对总体中的一些成员进行观测，例如对个别的消费者或家庭、厂商、产业或地区（如州、农村或城市）等进行观测。而且，这些成员可能大小不一，例如厂商有大、中、小之分，收入有高、中、低之分。而另一方面，在时间序列数据中，人们经常收集同一实体在一个时期内的数据，例如美国在 1955—2005 年间的 GNP、消费支出、储蓄或就业数据。

作为横截面分析中常会遇到的异方差性的一个说明，且考虑表 11—1。该表给出 1958 年按厂商或企业就业职工人数划分的 10 个非耐用品制造行业平均每个职工的薪金数据。表中还给出 9 个按职工人数分组的平均生产力数字。

表 11—1　　1958 年按厂商职工人数划分的非耐用品制造行业的人均薪金　　（单位：美元）

行业	就业人数（平均职工人数）								
	1～4	5～9	10～19	20～49	50～99	100～249	250～499	500～999	1 000～2 499
食品干果	2 994	3 295	3 565	3 907	4 189	4 486	4 676	4 968	5 342
烟草产品	1 721	2 057	3 336	3 320	2 980	2 848	3 072	2 969	3 822
纺织品	3 600	3 657	3 674	3 437	3 340	3 334	3 225	3 163	3 168
器皿用具	3 494	3 787	3 533	3 215	3 030	2 834	2 750	2 967	3 453
纸张类	3 498	3 847	3 913	4 135	4 445	4 885	5 132	5 342	5 326
印刷与出版	3 611	4 206	4 695	5 083	5 301	5 269	5 182	5 395	5 552
化工产品	3 875	4 660	4 930	5 005	5 114	5 248	5 630	5 870	5 876
石油与煤炭	4 616	5 181	5 317	5 337	5 421	5 710	6 316	6 455	6 347
橡胶与塑料	3 538	3 984	4 014	4 287	4 221	4 539	4 721	4 905	5 481
皮革与皮革制品	3 016	3 196	3 149	3 317	3 414	3 254	3 177	3 346	4 067
平均薪金	3 396	3 787	4 013	4 104	4 146	4 241	4 388	4 538	4 843
薪金标准差	742.2	851.4	727.8	805.06	929.9	1 080.6	1 243.2	1 307.7	1 110.7
平均生产力	9 355	8 584	7 962	8 275	8 389	9 418	9 795	10 281	11 750

资料来源：*The Census of Manufacturers*，U. S. Department of Commerce，1958（computed by author）.

虽然不同行业有不同的产出构成，但表 11—1 清楚地表明，平均而言，大的厂

家比小的厂家支付更多的工资。例如，职工人数在1～4人的厂商平均薪金约3 396美元，而职工人数在1 000～2 499人的厂商平均薪金约4 843美元。但应注意，在不同就业人数的类别之间，如估计的工资收入的标准差所表明的那样，工资有相当大的变异。这点还可以从图11—6看出，图11—6给出了每个职工人数组中薪金标准差和平均薪金。我们清楚地看到，总体上，薪金标准差随着平均薪金的提高而提高。

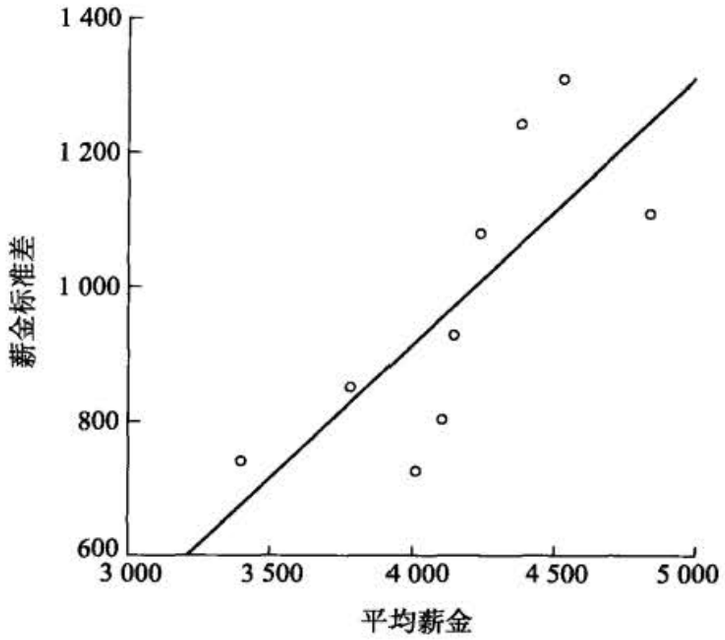


图11—6　薪金标准差与平均薪金

11.2　出现异方差性时的OLS估计

如果引进异方差性 $E(u_i^2)=\sigma_i^2$ 而保留经典模型的所有其他假定，OLS估计量及其方差会出现什么变化呢？为了回答此问题，让我们回到双变量模型：

$$Y_i=\beta_1+\beta_2X_i+u_i$$

按照惯常的公式，β_2 的OLS估计量是：

$$\hat{\beta}_2=\frac{\sum x_iy_i}{\sum x_i^2}=\frac{n\sum X_iY_i-\sum X_i\sum Y_i}{n\sum X_i^2-(\sum X_i)^2} \tag{11.2.1}$$

但现在它的方差是（参看附录11A第11A.1节）：

$$\operatorname{var}(\hat{\beta}_2)=\frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2} \tag{11.2.2}$$

这显然不同于同方差性假定下的常用方差公式：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{11.2.3}$$

当然，如果对每个 i 都有 $\sigma_i^2 = \sigma^2$，那么这两个公式是相同的。（为什么?）

回顾一下，如果经典模型的各个假定（包括同方差性在内）全部成立，则 $\hat{\beta}_2$ 是最优线性无偏估计量（BLUE）。那么，当我们仅除掉同方差性假定而代之以异方差性的假定时，它还会不会是 BLUE 呢？容易证明，$\hat{\beta}_2$ 仍是线性和无偏的。事实上，如附录 3A 第 3A.2 节所证明的那样，为了证明 $\hat{\beta}_2$ 的无偏性，干扰项（u_i）的同方差性并非必要。的确，u_i 的方差、同方差或异方差，与无偏性的证明无关。记得在附录 3A 的 3A.7 节中，我们证明过，在经典线性回归模型的假定之下，$\hat{\beta}_2$ 是一个一致估计量。尽管我们不去证明它，但我们可以证明，$\hat{\beta}_2$ 在异方差情形下是一个一致估计量；即随着样本容量无限扩大，估计的 β_2 收敛于其真实值。而且，还可以证明，在一定的条件（被称为正则性条件）下，$\hat{\beta}_2$ 还是渐近正态分布的（asymptotically normally distributed）。当然，上述结论对多元回归模型中的其他参数也成立。

认定 $\hat{\beta}_2$ 是线性无偏的，那它是不是“有效”或“最优”的，即是否在所有线性无偏估计量中有最小方差呢？并且这个最小方差是由方程（11.2.2）给出的吗？对两个问题的回答都是否定的：$\hat{\beta}_2$ 不再是最优的，而且最小方差也不由方程（11.2.2）给出。那么，在出现异方差性时，什么才是 BLUE 呢？下节给出回答。

11.3 广义最小二乘法

为什么方程（11.2.1）所给的 β_2 的常用 OLS 估计量虽然无偏但非最优呢？直观的理由可从表 11—1 看出。如表所示，各就业组之间的工薪收入有相当大的变异。假如要我们做每个职工的薪金对就业人数的回归，我们就应对薪金的这种组间变异知识加以利用。最理想的是设计出这样一种估计方案：对来自变异较大的总体的观测值赋予较小的权重，而对来自较小变异的总体观测值赋予较大的权重。检查一下表 11—1 便知道，相对于来自 5～9 人和 250～499 人的那些观测值要对来自如同 10～19 人和 20～49 人的就业组的观测值作更大的加权。因为后一种观测值比较紧密地聚集在它们的均值周围，从而能使我们更准确地估计 PRF。

可惜的是，常用的 OLS 方法并不采取这种策略，因而对表 11—1 中的职工薪金这个因变量 Y 的不等变异所包含的信息未加利用：它仍然对每一观测值同样重视或同等加权。而名为**广义最小二乘**（generalized least squares，GLS）的一种估计方法则明确地利用了这一信息，因而能得到 BLUE。为了看清楚怎样做到这一点，让我们继续利用现在已经熟悉的双变量模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{11.3.1}$$

为便于代数上的处理，我们把上述模型写为：

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i \tag{11.3.2}$$

其中对每个 i 均有 $X_{0i}=1$。相信读者能看出这两种写法是完全相同的。

现在假定不同的方差 σ_i^2 已知。将方程（11.3.2）的两边除以 σ_i 得：

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \tag{11.3.3}$$

为了易于阐述，将它写为：

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^* \tag{11.3.4}$$

其中带星号变量或转换变量为原始变量除以（已知的）σ_i。我们用符号 β_1^* 和 β_2^* 表示转换模型的参数，以区别于常用的 OLS 参数 β_1 和 β_2。

转换原始模型的用意何在？为说明这点，可注意变换误差项 u_i^* 的如下特点：

$$\begin{aligned}
\text{var}(u_i^*) = E(u_i^*)^2 &= E\left(\frac{u_i}{\sigma^i}\right)^2 && \text{因为 } E(u_i^*)=0 \\
&= \frac{1}{\sigma_i^2} E(u_i^2) && \text{因为 } \sigma_i^2 \text{ 已知} \\
&= \frac{1}{\sigma_i^2}(\sigma_i^2) && \text{因为 } E(u_i^2) = \sigma_i^2 \\
&= 1
\end{aligned} \tag{11.3.5}$$

这是一常数。就是说，转换干扰项 u_i^* 的方差，现在有了同方差性。因为我们仍保留着经典模型的其他假定，所以 u_i^* 的这一同方差性的发现表明，如果我们把 OLS 应用到转换模型（11.3.3）上，将产生 BLUE 估计量。简言之，这时估计出来的 β_1^* 和 β_2^* 是 BLUE，而 OLS 估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 则不是。

先将原始变量转换成满足经典模型假设的转换变量，然后对它们使用 OLS 程序，叫做广义最小二乘法（GLS）。概括地说，GLS 是对满足标准最小二乘假定的转换变量的OLS。如此得到的估计量叫做 **GLS 估计量**（GLS estimators）。这些估计量是 BLUE。

估计 β_1^* 和 β_2^* 的具体步骤如下：首先写下对应于方程（11.3.3）的 SRF：

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) + \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{\hat{u}_i}{\sigma_i}\right)$$

或者：

$$Y_i^* = \hat{\beta}_1^* X_{0i}^* + \hat{\beta}_2^* X_i^* + \hat{u}_i^* \tag{11.3.6}$$

然后最小化：

$$\sum \hat{u}_i^{2*} = \sum (Y_i^* - \hat{\beta}_1^* X_{0i}^* - \hat{\beta}_2^* X_i^*)^2$$

即：

$$\sum \left(\frac{\hat{u}_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right)\right]^2 \tag{11.3.7}$$

以获得 GLS 估计量。附录 11A 第 11A.2 节给出最小化方程（11.3.7）的标准计算

程序。如该节所表明的，β_2^* 的 GLS 估计量为：

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{11.3.8}$$

它的方差为：

$$\text{var}(\hat{\beta}_2^*) = \frac{\sum w_i}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{11.3.9}$$

其中 $w_i = 1/\sigma_i^2$ 。

□ OLS 和 GLS 的差别

从第 3 章中看到，OLS 要求我们最小化：

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \tag{11.3.10}$$

而 GLS 要求我们最小化表达式（11.3.7），而该式又可写为：

$$\sum w_i \hat{u}_i^2 = \sum w_i (Y_i - \hat{\beta}_1^* X_{0i} - \hat{\beta}_2^* X_i)^2 \tag{11.3.11}$$

其中 $w_i = 1/\sigma_i^2$ [读者可证明方程（11.3.11）和方程（11.3.7）相同]。

可见，在 GLS 中，我们最小化以 $w_i = 1/\sigma_i^2$ 为权重的一个加权残差平方和，而在 OLS 中我们最小化一个未加权或等权（相当于一回事）残差平方和（RSS）。方程（11.3.7）表明，GLS 分配给每一观测的权重与它的 σ_i 成反比，也就是说，在最小化方程（11.3.11）的 RSS 的过程中，来自有较大 σ_i 的总体观测将得到较小的权重，而来自有较小 σ_i 的总体观测将得到较大的权重。为了看清楚 OLS 和 GLS 的差别，且考虑图 11—7 中这个假想的散点图。

在（未加权的）OLS 中，点 A、B 和 C 处的误差在 RSS 的最小化过程中都得到相等的权重。显而易见，这时 C 的误差将支配 RSS。但在 GLS 中，这个极端的观测 C，和另外两个观测值相比，将得到相对小的权重。如前所说，这个策略是正确的，因为，为了更可靠地估计总体回归函数，我们应该对那些紧密围绕其（总体）均值的观测比给那些远离均值的观测赋予更大的权重。

因方程（11.3.11）最小化一个加权的 RSS，故适宜称之为**加权最小二乘**（weighted least squares，WLS），并把由此得到的，并由方程（11.3.8）和（11.3.9）给出的估计量称为 WLS **估计量**（WLS estimators）。但 WLS 只不过是更为一般的估计方法 GLS 的一种特殊情形。在异方差性的讨论中，WLS 和 GLS 两词可交换使用。但在以后的章节中，我们会遇到 GLS 的其他特殊情形。

顺便指出，如果对所有的 i 都有 $w_i = w$，即为一常数，则 $\hat{\beta}_2^*$ 等同于 $\hat{\beta}_2$，并且 $\text{var}(\hat{\beta}_2^*)$ 也等同于由方程（11.2.3）所给的常用（即同方差性的）$\text{var}(\hat{\beta}_2)$，这一点

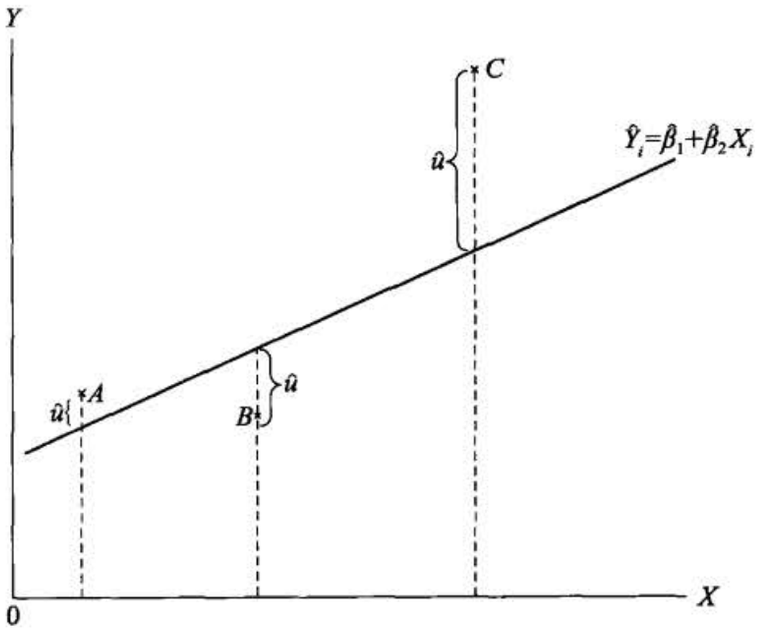


图 11—7　假想的散点图

应该无足为奇。（为什么?）（参看习题 11.8。）

11.4　出现异方差性时使用 OLS 的后果

正如我们已经看到的那样，$\hat{\beta}_2^*$ 和 $\hat{\beta}_2$ 两者都是（线性）无偏估计量：在重复抽样中，平均而言，$\hat{\beta}_2^*$ 和 $\hat{\beta}_2$ 都将等于真实 β_2，也就是说，它们都是无偏估计量。但我们知道 $\hat{\beta}_2^*$ 才是有效的，即有最小方差。那么，如果我们继续使用 OLS 估计量 $\hat{\beta}_2$，我们的置信区间、假设检验以及其他相关事宜会出现什么情况呢？我们区分两种情形讨论。

□ 考虑异方差性的 OLS 估计

假如我们使用 $\hat{\beta}_2$，但又明显考虑有异方差性而使用由方程（11.2.2）给出的方差公式。那么，按照该公式，假定 σ_i^2 为已知，是否就可以利用通常的 t 和 F 检验构造置信区间并进行假设检验呢？一般地说，回答是否定的。因为可以证明 $\text{var}(\hat{\beta}_2^*) \leqslant \text{var}(\hat{\beta}_2)$。[①] 这就是说，根据 $\text{var}(\hat{\beta}_2)$ 构造的置信区间将无谓地过大。其结果是，t 和 F 检验很可能给我们提供了不准确的结果：因为明显过大的 $\text{var}(\hat{\beta}_2)$ 会使本来（如果我们使用由 GLS 程序构造的正确的置信区间的话）显著的系数变成了统计上不显著的（因 t 值过小）。

① 一个正式的证明见 Phoebus J. Dhrymes, *Introductory Econometrics*, Springer-Verlag, New York, 1978, pp. 110-111。顺便指出，$\hat{\beta}_2$ 的效率损失［指 $\text{var}(\hat{\beta}_2)$ 超过 $\text{var}(\hat{\beta}_2^*)$ 多少］取决于 X 变量的样本值以及 σ_i^2 值。

□ 忽视异方差性的 OLS 估计

如果在有或怀疑有异方差性的情形下，我们不但使用了 $\hat{\beta}_2$，而且继续使用方程（11.2.3）所给的常用（同方差性的）方差公式，情况就变得严重了：注意，这是我们所讨论的两种情形中尤为常见的一种。原因是，当我们使用标准的 OLS 回归软件包时，忽略异方差性（或对异方差性无知）就会给出 $\hat{\beta}_2$ 的方差，就像方程（11.2.3）所给出的那样。首先，方程（11.2.3）所给的 $\text{var}(\hat{\beta}_2)$ 是方程（11.2.2）所给 $\text{var}(\hat{\beta}_2)$ 的一个有偏误的估计量。也就是说，平均而言，前者不是高估就是低估了后者。而且，一般地说，我们无法告知这个偏误是正的（过高估计）还是负的（过低估计）。可从方程（11.2.2）清楚地看到，它取决于 σ_i^2 的变化与解释变量 X 的取值之间的关系（参看习题 11.9）。之所以有偏误，是因为当异方差性出现时，σ^2 的惯用估计量 $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-2)$ 不是 σ^2 的无偏估计量（参见附录 11A.3）。忽视异方差性的结果是，我们不能再依赖通常计算的置信区间和通常使用的 t 和 F 检验。[①] **总之，如果我们忽视异方差性而执意使用惯常的检验程序，则无论我们得出什么结论或作出什么推断，都可能产生严重的误导。**

为使问题的讨论更加明朗，我们引用戴维森和麦金农所做的一个**蒙特卡罗**（Monte Carlo）实验。[②] 他们考虑一个简单的模型，可用我们的符号表示如下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{11.4.1}$$

他们假定 $\beta_1 = 1$、$\beta_2 = 1$ 和 $u_i \sim N(0, X_i^{\alpha})$。最后一个式子表明误差方差是异方差性的，并且它的值是回归元 X 值的 α 次方。例如，当 $\alpha = 1$ 时，误差方差与 X 值成正比；当 $\alpha = 2$ 时，误差方差与 X 值的平方成正比，等等。在 11.6 节中我们将讨论这种比例关系的逻辑性。根据 20 000 次重复实验，并令 α 取不同的值，他们得到使用 OLS［见方程（11.2.3）］和使用 OLS 但考虑到有异方差性［见方程（11.2.2）］以及使用 GLS［见方程（11.3.9）］的两个回归系数的标准误。现对所选的一些 α 值，把它们的结果列表如下。

α 值	$\hat{\beta}_1$ 的标准误			$\hat{\beta}_2$ 的标准误		
	OLS	OLS_{het}	GLS	OLS	OLS_{het}	GLS
0.5	0.164	0.134	0.110	0.285	0.277	0.243
1.0	0.142	0.101	0.048	0.246	0.247	0.173
2.0	0.116	0.074	0.007 3	0.200	0.220	0.109
3.0	0.100	0.064	0.001 3	0.173	0.206	0.056
4.0	0.089	0.059	0.000 3	0.154	0.195	0.017

注：OLS_{het} 表示考虑异方差性的 OLS。

① 由方程（5.3.6）我们知道 β_2 的 $100(1-\alpha)\%$ 置信区间是 $[\hat{\beta}_2 \pm t_{\alpha/2}\text{se}(\hat{\beta}_2)]$。但若 $\text{se}(\hat{\beta}_2)$ 不能无偏地加以估计，我们还能对通常计算的置信区间给予多少信赖呢？

② Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993, pp. 549-550.

这些结果最令人注目的特点是，不管是否考虑对异方差性的修正，OLS 一律过高地估计了由（正确的）GLS 程序得到的真实标准误，尤其是在 α 值较大时，从而证实了 GLS 的优越性。这些结果还表明，如果我们不用 GLS 而只用 OLS，不管是否考虑到异方差性，情况都不清晰。常用的 OLS，相对于顾及异方差性的 OLS 来说，其标准误或者偏之于过大（对截距言）或者一般偏之于过小（对斜率系数言）。一条明显的信息是：在出现异方差性时就要用 GLS。然而，在实际中，GLS 并不总是容易使用的，理由见后。同样，我们以后还要提到，除非异方差性很严重，否则，可能不应该放弃 OLS 而使用 GLS 或 WLS。

由上述讨论可见，异方差性是一个潜在的严重问题，研究者需要知道它在某一给定的情况中是否出现。如果发现有异方差性，就可采取纠正措施，比如使用加权最小二乘回归或某些其他方法。然而，在我们考虑各种纠正措施之前，我们必须先判断在某一给定情况中，是否有或很可能有异方差性。这就是下一节要讨论的问题。

□ 一个技术性注解

尽管我们曾经说过，在异方差情形中，GLS 是 BLUE，而 OLS 不是，但在有些例子中 OLS 在异方差情况下仍是 BLUE。[①] 只是这种例子在实践中并不多见。

11.5 异方差性的侦察

和多重共线性一样，一个重要的实际问题是：我们怎样知道在一个具体的情况中是否有异方差性？而且也和多重共线性类似，并不存在有侦察异方差性的硬性规定，只有少数经验规则。但是这种结局是不可避免的，因为除非我们知道对应于选定 X 值的整个 Y 总体，如同表 2—1 或表 11—1 所给的总体那样，否则 σ_i^2 是无从获知的。然而，在经济研究中，这样的数据（总体）可遇不可求。在这方面，计量经济学家不同于诸如农学或生物学等领域的科学家。农学或生物学的研究者们能很好地控制他们的研究主题。而在经济研究中，对应于一个具体的 X 值，多数情形都只有一个样本 Y 值。所以没有任何方法能仅从一个 Y 观测值去获知 σ_i^2。因此，在大多数的计量经济调查研究中，异方差性不过是一种直觉、深思熟虑的猜测、先前经验或纯粹猜想。

① 其原因在于，高斯-马尔可夫定理为 OLS 的有效性提供了充分（而非必要）条件。OLS 为 BLUE 的充分必要条件由**克鲁斯卡尔定理**（Kruskal's theorem）给出，但这个问题超出了本书的范围。感谢迈克尔·麦卡利尔使我注意到这一点。详尽分析可参见 Denzil G. Fiebig, Michael McAleer, and Robert Bartels, "Properties of Ordinary Least Squares Estimators in Regression Models with Nonspherical Disturbances," *Journal of Econometrics*, vol. 54, No. 1-3, Oct.-Dec., 1992, pp. 321-334。对喜欢数学的同学而言，我会在附录 C 中用矩阵代数进一步讨论这个问题。

有了上述告诫，现在就可以列举一些非正式或正式的侦察异方差性的方法。如下面的讨论将要表明的那样，大多数方法都基于对我们所能观测到的 OLS 残差 a_i 的分析，而不是对干扰项 u_i 的分析。我们寄希望于它们是 u_i 的良好估计。当样本容量相当大时，这一希望也许能够实现。

□ 非正式方法

问题的性质。往往根据所考虑问题的性质就能判别是否会遇到异方差性。例如，普雷斯（Prais）和霍撒克（Houthakker）在一项家庭预算研究中发现，围绕消费对收入的回归，残差的方差随收入增加而增加。仿效这一开拓性的研究，现在人们一般都假定在类似的调查中可以预期不同干扰项之间有不相等的方差。[①] 事实上，在涉及异质性调查单位的横截面数据中，异方差性可能比较常见。例如，在投资与销售量、利率等变量之间关系的横截面分析中，如果样本同时包含小、中和大型厂家，一般都预期有异方差性。

事实上，我们已经遇到过这种例子。我们在第 2 章讨论了美国小时工资均值与受教育年数的关系。而且，我们在那里还讨论了印度 55 个家庭的食物支出与总支出之间的关系（见习题 11.16）。

图解法。如果对异方差性的性质没有任何先验或经验信息，实际上，可先在无异方差性的假定下做回归分析，然后对残差的平方 a_i 做一事后检查，看看这些 a_i^2 是否呈现任何系统性的样式。虽然 a_i^2 还不等于 u_i^2，但可作为一个代理变量，特别是当样本容量足够大时。[②] 对 a_i^2 的检查可能出现如图 11—8 中的那些样式。

在图 11—8 中，a_i^2 相对于 $\hat{Y}_i$ 进行描点，$\hat{Y}_i$ 是从回归线读出的 Y_i 的估计值，其用意是要找出 Y 的估计均值是否与残差的平方系统相关。在图 11—8a 中，我们未发现这两个变量之间有任何系统性联系，表明了数据中也许没有异方差性。图 11—8b 到图 11—8e 则呈现一定的模式。例如，图 11—8c 表示 a_i^2 与 $\hat{Y}_i$ 之间的一个线性关系，而图 11—8d 和图 11—8e 则表示 a_i^2 与 $\hat{Y}_i$ 之间的二次关系。利用这种虽然是非正式的知识，我们却有可能对数据进行变换，使得变换后的数据不具有异方差性。在 11.6 节中，我们将分析几种这样的变换。

除了将 a_i^2 对 $\hat{Y}_i$ 描点外，还可将它们对解释变量之一描点，特别是像图 11—8a 那样在对 $\hat{Y}_i$ 的描点结果看不出异方差性的情况下。如图 11—9 所示，对 X 描点的结果会显示出类似于图 11—8 的图样。（在双变量模型的情形中，将 a_i^2 对 $\hat{Y}_i$ 描点等价于将它对 X_i 描点。因此，图 11—9 和图 11—8 必然是相似的。但当我们考虑两个或多个 X 变量的模型时，情况就不同了；这时可将 a_i^2 相对于模型中的任一个 X 变量描点。）

---

① S. J. Prais and H. S. Houthakker, *The Analysis of Family Budgets*, Cambridge University Press, New York, 1955.

② 至于 a_i 与 u_i 之间的关系，参看 E. Malinvaud, *Statistical Methods of Econometrics*, North Holland Publishing Company, Amsterdam, 1970, pp. 88-89。

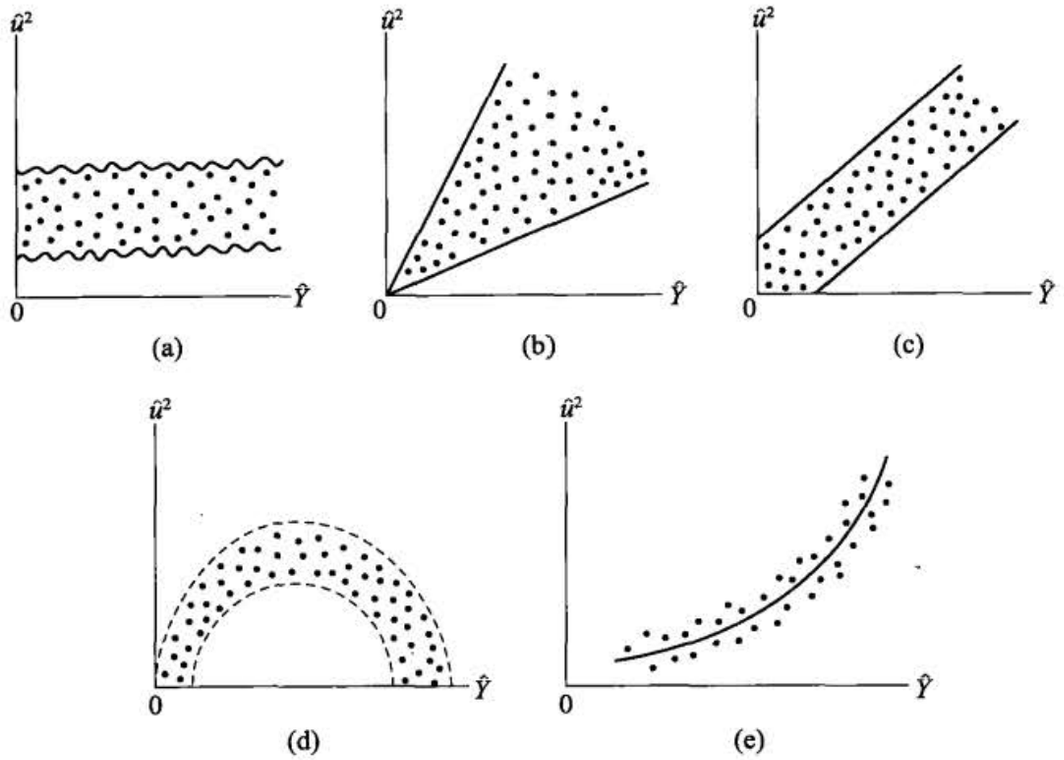


图 11—8　残差平方估计值的假想图样

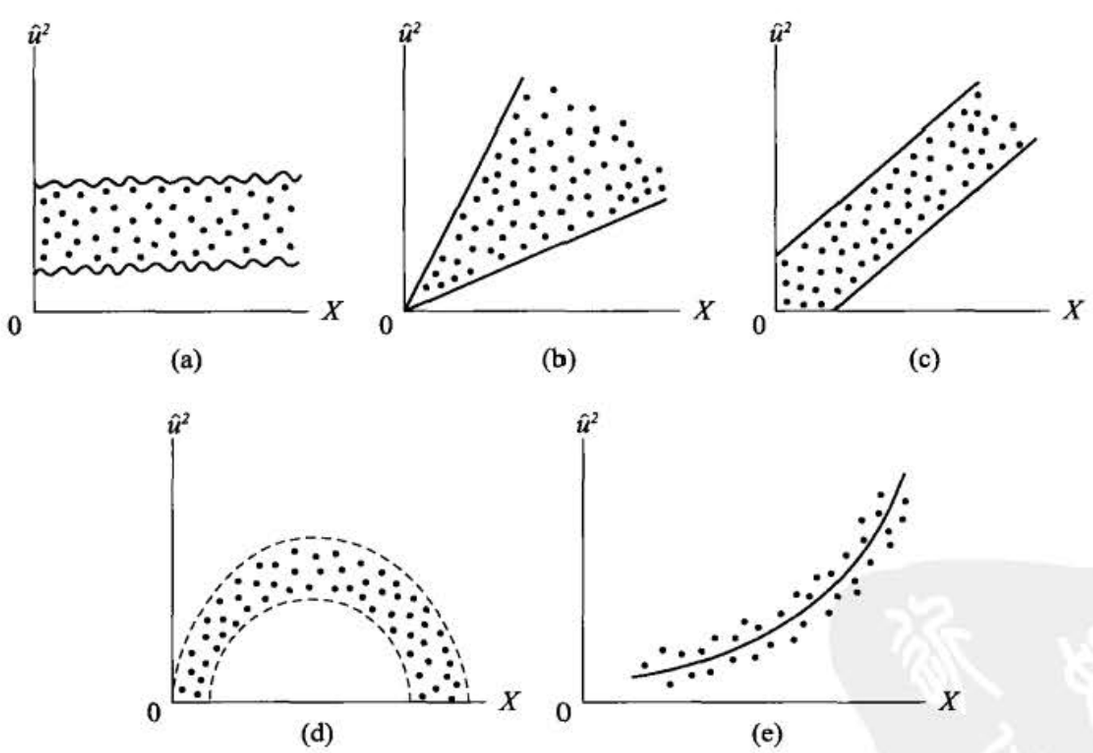


图 11—9　残差平方估计值相对于 X 的散点图

例如，一个类似于图 11—9c 的图形可能表明干扰项的方差与 X 变量有线性关系。因此，如果在储蓄对收入的回归中，发现有如同图 11—9c 那样的图样，就表明异方差可能正比于收入变量的取值。这一知识有助于我们将数据进行变换，使得对

变换后的数据进行回归时，干扰项的方差就变成了同方差的。下节我们还将回到这个问题上。

□ 正式方法

帕克检验。[①] 帕克（Park）通过指出 σ_i^2 是解释变量 X_i 的某个函数，从而把图解法公式化。他所建议的函数形式是：

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$$

或者：

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i \tag{11.5.1}$$

其中 v_i 是随机干扰项。

由于 σ_i^2 通常是未知的，所以帕克建议用 $\hat{u}_i^2$ 作为替代变量并做如下回归：

$$\begin{aligned} \ln \hat{u}_i^2 &= \ln \sigma^2 + \beta \ln X_i + v_i \\ &= a + \beta \ln X_i + v_i \end{aligned} \tag{11.5.2}$$

如果 β 表现为统计显著的，就表明数据中有异方差性。如果它不显著，则可接受同方差性假设。可见，帕克检验是一个两阶段程序。在第一阶段中，我们做 OLS 回归而不考虑异方差性问题。我们从这一回归获得 $\hat{u}_i$，然后在第二阶段中做回归（11.5.2）。

虽然帕克检验从经验上看颇有魅力，却遇到一些问题，戈德菲尔德（Goldfeld）和匡特（Quandt）曾指出，进入方程（11.5.2）的误差项 v_i 可能不满足 OLS 假设，而且本身还可能是异方差的。[②] 然而作为一个纯粹探索性的方法，帕克检验还是可以使用的。

例 11.1　薪金与生产力的关系

为说明帕克方法，我们利用表 11—1 中的数据做如下回归：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

其中 Y=以千美元计的平均酬金，X=以千美元计的平均生产力，i=企业的第 i 类就业规模。回归结果如下：

$$\begin{aligned} \hat{Y}_i &= 1\,992.345\,2 + 0.232\,9 X_i \\ \text{se} &= (936.479\,1) \quad (0.099\,8) \\ t &= (2.127\,5) \quad (2.333) \qquad R^2 = 0.437\,5 \end{aligned} \tag{11.5.3}$$

结果表明斜率系数估计值在单尾 t 检验的基础上达到 5%的显著水平。方程表示劳动生产力每增加比方说 1 美元，劳动报酬平均约增加 23 美分。

---

① R. E. Park, "Estimation with Heteroscedastic Error Terms," *Econometrica*, vol. 34, no. 4, October 1966, p. 888. 帕克检验是哈维（A. C. Harvey）提出的一般检验的一种特殊情形。见 A. C. Harvey, "Estimating Regression Models with Multiplicative Heteroscedasticity," *Econometrica*, vol. 44, no. 3, 1976, pp. 461-465。

② Stephen M. Goldfeld and Richard E. Quandt, *Nonlinear Methods in Econometrics*, North Holland Publishing Company, Amsterdam, 1972, pp. 93-94.

将得自回归（11.5.3）的残差用于方程（11.5.2）中对 X_i 的回归，给出如下结果：

$$\widehat{\ln a_i^2} = 35.817 - 2.809\,9\ln X_i$$
$$\text{se}=(38.319)\quad(4.216) \tag{11.5.4}$$
$$t = (0.934)\quad(-0.667)\qquad R^2=0.059\,5$$

显然，两变量之间无统计上的显著关系。按照帕克检验，便可下结论说，在误差的方差中没有异方差性。①

格莱泽检验。② 格莱泽（Glejser）检验的思想实质上类似于帕克检验。格莱泽建议，在从 OLS 回归取得残差 a_i 之后，用 a_i 的绝对值对被认为与 σ_i^2 密切相关的 X 变量做回归。在他的实验中，他使用如下多种函数形式：

$$|a_i| = \beta_1 + \beta_2 X_i + v_i$$
$$|a_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$
$$|a_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i$$
$$|a_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$
$$|a_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$
$$|a_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

其中 v_i 是误差项。

格莱泽检验仍然可作为一种经验或实际处理方法加以使用。但戈德菲尔德和匡特指出误差项的若干问题，如期望值非零、序列相关（见第 12 章）以及有讽刺意味的异方差性。③ 格莱泽方法的另一困难是，像

$$|a_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$

以及

$$|a_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

这类模型，对参数而言是非线性的，因而不能用平常的 OLS 程序加以估计。

格莱泽发现，在异方差性的侦察中，上列模型中的前四个对大样本来说，一般都能给出令人满意的结果，因此，从实际角度考虑，格莱泽方法可用于大样本。而在小样本中，则仅可作为摸索异方差性的一种定性方法。

例 11.2　薪金与生产力的关系：格莱泽检验

继续例 11.1，将回归（11.5.3）所得到残差的绝对值对平均生产力（X）回归，得到如下

---

① 帕克所选的特殊函数形式仅是他的一种建议。另一种函数形式可能表明有显著关系。例如，不妨试用 a_i^2 代替 $\ln a_i^2$ 作为因变量。

② H. Glejser, "A New Test for Heteroscedasticity," *Journal of the American Statistical Association*, vol. 64, 1969, pp. 316-323.

③ 详见戈德菲尔德与匡特的前引文献第 3 章。

结论：

$$\begin{aligned} \widehat{|a_i|} &= 407.278\,3 - 0.020\,3X_i \\ \text{se} &= (633.162\,1)\quad (0.067\,5)\qquad r^2=0.012\,7 \\ t &= \quad (0.643\,2)\ (-0.301\,2) \end{aligned} \tag{11.5.5}$$

诚如你从此回归中所见，残差的绝对值与回归元平均生产力之间没有关系，这就加强了基于帕克检验所得到的结论。

斯皮尔曼的等级相关经验。在习题 3.8 中，我们曾定义斯皮尔曼（Spearman）的等级相关系数为：

$$r_s = 1-6\left[\frac{\sum d_i^2}{n(n^2-1)}\right] \tag{11.5.6}$$

其中 d_i＝第 i 个单位或现象的两种不同特性所处的等级之差，而 n＝观测单位或现象的级别个数。上述等级相关系数可按下述方法用于侦察异方差性：假定 $Y_i=\beta_0+\beta_1 X_i+u_i$。

步骤 1　对 Y 和 X 的数据做回归拟合并求出残差 $\hat{u}_i$。

步骤 2　忽视 $\hat{u}_i$ 的符号，也就是取其绝对值 $|\hat{u}_i|$，将 $|\hat{u}_i|$ 和 X_i（或 $\hat{Y}_i$）同时按递升或递降次序划分等级，然后计算上述斯皮尔曼的等级相关系数。

步骤 3　假定总体等级相关系数 ρ_s 为零且 $n>8$，样本 r_s 的显著性可通过 t 检验按下述方法加以检验[①]：

$$t=\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \tag{11.5.7}$$

其自由度 df＝$n-2$。

如果计算的 t 值超过 t 临界值就可接受异方差性假设，否则拒绝。如果回归模型涉及多于一个 X 变量，则可在 $|\hat{u}_i|$ 与每一 X 变量之间分别计算 r_s，再用方程（11.5.7）中的 t 检验作统计显著性检验。

例 11.3　等级相关检验的说明

为说明等级相关检验，考虑表 11—2 中的数据。这些数据包含了 10 个共同基金的平均年回报（E，%）及其标准差（σ_i，%）。

投资组合理论中的资本市场线（CML）假定期望收益（E_i）和风险（用标准差 σ 来度量）之间有如下线性关系

$$E_i=\beta_1+\beta_2\sigma_i$$

---

① 见 G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Charles Griffin & Company, London, 1953, p. 455。

表 11—2　　异方差性的等级相关检验

共同基金名称	E_i，平均年回报，%	σ_i，年回报标准差，%	$\hat{E}_i^{\dagger}$	$\lvert a_i \rvert^{\ddagger}$，残差，$\lvert (E_i - \hat{E}_i) \rvert$	$\lvert a_i \rvert$ 的等级	σ_i 的等级	d，两等级之差	d^2
波士顿基金	12.4	12.1	11.37	1.03	9	4	5	25
特拉华基金	14.4	21.4	15.64	1.24	10	9	1	1
权益基金	14.6	18.7	14.40	0.20	4	7	−3	9
基本投资基金	16.0	21.7	15.78	0.22	5	10	−5	25
互有投资基金	11.3	12.5	11.56	0.26	6	5	1	1
卢米斯销售相互投资基金	10.0	10.4	10.59	0.59	7	2	5	25
麻省信托投资基金	16.2	20.8	15.37	0.83	8	8	0	0
新英格兰基金	10.4	10.2	10.50	0.10	3	1	2	4
波士顿普塔姆基金	13.1	16.0	13.16	0.06	2	6	−4	16
惠灵顿基金	11.3	12.0	11.33	0.03	1	3	−2	4
总计							0	110

注：†得自回归：$\hat{E}_i = 5.819\,4 + 0.459\,0\sigma_i$；
‡残差的绝对值。
按递升次序编排等级。

利用表 11—2 中的数据，估计上述模型，并从中计算出残差。由于数据涉及规模与投资目标都不相同的 10 个互助基金，人们先验预料会存在异方差性。为检验此假设，现利用等级相关方法。必要的计算也见表 11—2。

应用公式（11.5.6）得到：

$$r_s = 1 - 6 \times \frac{110}{10 \times (100 - 1)} = 0.333\,3 \tag{11.5.8}$$

再用方程（11.5.7）中的 t 检验得：

$$t = \frac{0.333\,3 \times \sqrt{8}}{\sqrt{1 - 0.111\,0}} = 0.999\,8 \tag{11.5.9}$$

对于 8 个自由度，即使在 10%的显著水平上，这个 t 值也是不显著的；p 值是 0.17。因此，没有迹象表明解释变量与残差绝对值之间有任何系统的联系，故可认为没有异方差性。

戈德菲尔德-匡特检验。[1] 这一广为流传的方法适用于异方差性方差 σ_i^2 与回归模型中解释变量之一有正向关系的情形。为简单起见，考虑通常的双变量模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

假设 σ_i^2 与 X 变量的正向关系为：

$$\sigma_i^2 = \sigma^2 X_i^2 \tag{11.5.10}$$

其中 σ^2 是一常数。[2]

① 戈德菲尔德与匡特的前引文献第 3 章。

② 这仅是一个可取的假定。实际上要求 σ_i^2 与 X_i 存在单调关系。

假定方程（11.5.10）设想 σ_i^2 与 X 变量的平方成正比。普莱斯和霍撒克在其家庭预算研究中曾发现这种假定甚为有用。（见 11.5 节的非正式方法。）

如果方程（11.5.10）得当，则意味着 X_i 值越大，σ_i^2 也越大。如果情况正是如此，则模型中有异方差性是最为可能的。为作出明确的检验，戈德菲尔德和匡特提出如下的步骤：

步骤 1 从最小的 X 值开始，按 X 值的大小顺序将观测值排列。

步骤 2 略去居中的 c 个观测值，其中 c 是预定的，并将其余 $(n-c)$ 个观测值分成两组，每组 $(n-c)/2$ 个。

步骤 3 分别对前 $(n-c)/2$ 个观测值和后 $(n-c)/2$ 个观测值各拟合一个回归，并分别获得残差平方和 RSS_1 和 RSS_2，RSS_1 代表对较小 X_i 值所做回归的 RSS（小方差组），而 RSS_2 代表对较大 X_i 值所做回归的 RSS（大方差组）。这些 RSS 各有：

$$\frac{n-c}{2}-k \quad \text{或} \quad \left(\frac{n-c-2k}{2}\right) \text{个自由度（df）}$$

其中 k 是包括截距在内的待估计参数个数。（为什么?）当然，对于双变量情形，$k=2$。

步骤 4 计算比率

$$\lambda = \frac{\text{RSS}_2/\text{df}}{\text{RSS}_1/\text{df}} \tag{11.5.11}$$

如果假定 u_i 是正态分布的（我们常做这种假定），并且如果同方差性假定成立，则可以证明方程（11.5.11）中的 λ 服从分子和分母自由度均为 $(n-c-2k)/2$ 的 F 分布。

如果在一项应用中，计算的 λ（$=F$）值大于选定显著性水平的 F 临界值，就可拒绝同方差性假设，也就是说很可能出现了异方差性。

在举例阐明此检验前，谈谈省略 c 个居中观测值的做法是适宜的。这些观测值的省略是为了突出或激化小方差组（即 RSS_1）与大方差组（即 RSS_2）之间的差异，但戈德菲尔德-匡特检验之所以能成功地做到这一点，有赖于怎样选好 c。[①] 对于双变量模型，戈德菲尔德和匡特所做的蒙特卡罗实验表明，当样本容量为 30 时，c 约为 8 为宜，当样本容量为 60 时，c 约为 16 为宜。但贾奇等人却提出，在实践中发现，当 $n=30$ 时，取 $c=4$；当 $n=60$ 时，取 $c=10$ 为宜。[②]

在继续讲述之前，还要提请注意，当模型中有多于一个 X 变量时，在检验的步骤 1 中，就可按任一个 X 的大小顺序将观测值排列。例如，在模型 $Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+u_i$ 中，可按这些 X 中的任一个将数据排序。如果我们事先没有把握哪个

① 用专业术语来说，检验的功效依赖于 c 的选择。在统计学中，一个检验的**功效**（power）由拒绝非真虚拟假设的概率［即 1－Prob（犯第Ⅱ类错误）］来衡量。这里，虚拟假设是：两组数据的方差相同，即有同方差性。进一步的讨论，见 M. M. Ali and C. Giaccotto, "A Study of Several New and Existing Tests for Heteroscedasticity in the General Linear Model," *Journal of Econometrics*, vol. 26, 1984, pp. 355-373。

② George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, John Wiley & Sons, New York, 1982, p. 422.

X 变量合适，则可对每一 X 变量进行检验，或者通过对每一个 X 轮流做帕克检验。

例 11.4 戈德菲尔德-匡特检验

为说明戈德菲尔德-匡特检验，我们在表 11—3 中给出一个 30 户家庭的横截面消费与收入数据。假如我们认为消费与收入有线性关系，而数据中有异方差性，并且进一步假设异方差的性质如方程（11.5.10）所设。为了进行检验，我们把重新排序的数据也在表 11—3 中给出。

表 11—3　为说明戈德菲尔德-匡特检验的假想消费 Y（美元）与收入 X（美元）数据

Y	X	按 X 值排序的数据		
		Y	X	
55	80	55	80	
65	100	70	85	
70	85	75	90	
80	110	65	100	
79	120	74	105	
84	115	80	110	
98	130	84	115	
95	140	79	120	
90	125	90	125	
75	90	98	130	
74	105	95	140	
110	160	108	145	
113	150	113	150	
125	165	110	160	
108	145	125	165	居中的 4 个观测值
115	180	115	180	
140	225	130	185	
120	200	135	190	
145	240	120	200	
130	185	140	205	
152	220	144	210	
144	210	152	220	
175	245	140	225	
180	260	137	230	
135	190	145	240	
140	205	175	245	
178	265	189	250	
191	270	180	260	
137	230	178	265	
189	250	191	270	

略去居中的4个观测值后，对开头的13个和末尾的13个观测值分别做OLS回归，并计算相应的残差平方和，具体结果如下（括号中为标准误）。

对前13个观测值做回归：

$$\hat{Y}_i = 3.409\,4 + 0.696\,8X_i$$

(8.704 9)　(0.074 4)　　$r^2=0.888\,7$　　$RSS_1=377.17$　　df=11

对后13个观测值做回归：

$$\hat{Y}_i = -28.027\,2 + 0.794\,1X_i$$

(30.642 1)　(0.131 9)　　$r^2=0.768\,1$　　$RSS_2=1\,536.8$　　df=11

从这些结果我们得到：

$$\lambda = \frac{RSS_2/df}{RSS_1/df} = \frac{1\,536.8/11}{377.17/11}$$

$$\lambda = 4.07$$

对于11个分子自由度和11个分母自由度，5%显著水平的F临界值是2.82。由于估计的$F(=\lambda)$值超过此临界值，故可作误差方差中有异方差性的结论。然而，如果我们把显著性水平定在1%上，则我们未必拒绝同方差性假定。（为什么？）注意，观测到λ的p值是0.014。

布罗施-帕甘-戈弗雷检验。[①] 戈德菲尔德-匡特检验的成功不仅依赖于c值（被省略的居中观测值个数），还依赖于用以排序的X变量的正确识别。如果我们考虑布罗施-帕甘-戈弗雷（Breusch-Pagan-Godfrey，BPG）检验，则可避免这种检验的局限性。

为说明这种检验，考虑k变量线性回归模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{11.5.12}$$

假定误差方差σ_i^2有如下函数关系：

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}) \tag{11.5.13}$$

即σ_i^2是非随机变量Z的某个函数；部分或全部X可用作Z。具体地说，假定：

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} \tag{11.5.14}$$

即σ_i^2是Z的一个线性函数。如果$\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$，则$\sigma_i^2 = \alpha_1$，即为一常数。因此，为了检验$\sigma_i^2$是否同方差，就可检验假设$\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$。这就是布罗施-帕甘-戈弗雷检验的基本思想。具体检验步骤如下：

步骤1　用OLS估计方程（11.5.12）并得到残差$\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_n$。

步骤2　计算$\tilde{\sigma}^2 = \sum \hat{u}_i^2/n$。回顾第4章可知这是$\sigma^2$的极大似然（ML）估计量。[注：OLS估计量是$\sum \hat{u}_i^2/(n-k)$。]

步骤3　按以下定义构造p_i：

$$p_i = \hat{u}_i^2/\tilde{\sigma}^2$$

---

① T. Breusch and A. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, vol. 47, 1979, pp. 1287-1294. 又见 L. Godfrey, "Testing for Multiplicative Heteroscedasticity," *Journal of Econometrics*, vol. 8, 1978, pp. 227-236。由于二者的相似性，故将其命名为异方差性的布罗施-帕甘-戈弗雷检验。

这无非就是将每个平方残差除以 $\tilde{\sigma}^2$ 。

步骤 4 将如此构造的 p_i 对 Z 回归：

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \tag{11.5.15}$$

其中 v_i 是回归的残差项。

步骤 5 从方程（11.5.15）中求出 ESS（解释平方和）并定义：

$$\Theta = \frac{1}{2}(\text{ESS}) \tag{11.5.16}$$

假定 u_i 是正态分布的。可以证明，如果有同方差性，当样本容量 n 无限增大时，则：

$$\Theta \underset{\text{asy}}{\sim} \chi^2_{m-1} \tag{11.5.17}$$

也就是说，Θ 服从自由度为（$m-1$）的 χ^2 分布。（注：asy 意谓渐近地。）

因此，在一项应用中，如果所计算的 Θ（$=\chi^2$）超过选定显著性水平的 χ^2 临界值，就可拒绝同方差性假设；否则不拒绝。

读者可能想知道，为什么 BPG 选取 $\frac{1}{2}$ ESS 作为检验统计量。其原因略为复杂，留待参阅参考文献。[①]

例 11.5 布罗施-帕甘-戈弗雷检验

作为一个例子，我们再回到曾用来说明戈德菲尔德-匡特异方差性检验的数据（表 11—3）。将 Y 对 X 回归得到：

步骤 1

$$\hat{Y}_i = 9.290\,3 + 0.637\,8X_i$$

$$\text{se}=(5.231\,4)\quad(0.028\,6)\qquad \text{RSS}=2\,361.153\quad R^2=0.946\,6 \tag{11.5.18}$$

步骤 2

$$\tilde{\sigma}^2 = \sum \hat{u}_i^2/30 = 2\,361.153/30 = 78.705\,1$$

步骤 3 用 78.705 1 除得自回归（11.5.18）的残差 $\hat{u}_i$，以构造变量 p_i。

步骤 4 假定 p_i 按方程（11.5.14）的假设与 X_i（$=Z_i$）有线性关系，我们算得回归：

$$p_i = -0.742\,6 + 0.010\,1X_i$$

$$\text{se}=(0.752\,9)\quad(0.004\,1)\qquad \text{ESS}=10.428\,0\quad R^2=0.18 \tag{11.5.19}$$

步骤 5

$$\Theta = \frac{1}{2}(\text{ESS}) = 5.214\,0 \tag{11.5.20}$$

在 BPG 检验的假定下，方程（11.5.20）中的 Θ 渐近服从 1 个自由度的 χ^2 分布。[注：方程（11.5.19）中只有一个回归元。] 现在从 χ^2 表我们查到，对于 1 个自由度，5%的 χ^2 临界值是 3.841 4 和 1%的 χ^2 临界值是 6.634 9。由此可知算出来的 χ^2 值 5.214 0 在 5%的显著性水平上显著，但在 1%的显著性水平上不显著。因此，我们得到如同戈德菲尔德-匡特检验一样的结论。但应记住，严格地说，BPG 检验是一种渐近性或大样本检验，而在本例中 30 个观测值还未必构成一

① 参见 Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar, Cheltenham, U. K., 1994, pp. 178-179。

大样本。还应指出，在小样本中，该检验对干扰项 u_i 的正态性假定非常敏感。当然，我们可通过第 5 章讨论的 χ^2 检验或雅克-贝拉检验去检验正态性假定。①

怀特的一般异方差性检验。戈德菲尔德-匡特检验要求按照被认为是引起异方差性的 X 变量把观测值重新排序，而 BGP 检验则易受偏离正态性假定的影响。怀特所提出的检验，不同于这两个检验，并不要求排序也不依赖于正态性假定，而且易于付诸实施。② 为说明其基本思想，考虑如下三变量回归模型（对 k 变量模型的推广也很显然）：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{11.5.21}$$

怀特检验进行如下：

步骤 1 对给定的数据，估计方程（11.5.21）并获得残差 $\hat{u}_i$。

步骤 2 再做如下（辅助）回归：

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \tag{11.5.22}$$③

即将得自原回归的残差平方对原始 X 回归元、其平方项和交叉乘积项做回归，还可引进回归元的高次方。注意方程中有一常数项，即使原始回归不一定包含它。从这个（辅助）回归求 R^2。

步骤 3 在无异方差性的虚拟假设下，可以证明，从辅助回归算得的 R^2 乘以样本容量（n），渐近地服从自由度等于辅助回归中的回归元（不包括常数项）个数的 χ^2 分布，即：

$$n \cdot R^2 \underset{\text{asy}}{\sim} \chi^2_{\text{df}} \tag{11.5.23}$$

其中 df 的定义如前。在本例中，因辅助回归中有 5 个回归元，故有 5 个自由度。

步骤 4 如果方程（11.5.23）中算得的 χ^2 值超过选定显著性水平的 χ^2 临界值，结论就是存在异方差性。如果不超过，就认为没有异方差性，也就是说，在辅助回归（11.5.22）中，$\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$［参看关于回归（11.5.22）的注释］。

例 11.6 怀特异方差性检验

根据 41 个国家的横截面数据，斯蒂芬·刘易斯（Stephen Lewis）估计了如下回归模型④：

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \tag{11.5.24}$$

其中 Y=贸易税收（进口与出口税收）与政府总收入之比，X_2=进出口总和与 GNP 之比，X_3=

① 关于这个问题，参见 R. Koenker, "A Note on Studentizing a Test for Heteroscedasticity," *Journal of Econometrics*, vol. 17, 1981, pp. 1180-1200。

② H. White, "A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity," *Econometrica*, vol. 48, 1980, pp. 817-818.

③ 隐含于这一步骤的假定是误差 u_i 的方差 σ_i^2 与回归元及其平方和交叉乘积项存在函数关系。如果这个回归的全部偏斜率系数同时等于零，则误差方差是一个等于 α_1 的同方差常数。

④ Stephen R. Lewis, "Government Revenue from Foreign Trade," *Manchester School of Economics and Social Studies*, vol. 31, 1963, pp. 39-47.

人均 GNP；ln 表示自然对数。他的假设是 Y 与 X_2 有正向关系（贸易额越高，贸易税收越高），并且 Y 与 X_3 有负向关系（随着收入增加，政府发现直接税——例如所得税——比贸易税更易于征收）。

经验结果支持了这些假设。对我们来说，重要的问题是数据中有没有异方差性。由于数据是涉及多个异质性国家的横截面数据，人们会先验地预期误差方差中存在异方差性。将怀特的异方差性检验应用于从回归（11.5.24）得到的残差，得到如下结果[①]：

$$\hat{u}_i^2 = -5.841\,7 + 2.562\,9 \ln \text{Trade}_i + 0.691\,8 \ln \text{GNP}_i - 0.408\,1 (\ln \text{Trade}_i)^2$$
$$-0.049\,1 (\ln \text{GNP}_i)^2 + 0.001\,5 (\ln \text{Trade}_i)(\ln \text{GNP}_i) \quad R^2 = 0.114\,8 \qquad (11.5.25)$$

注：因标准误不切合我们这里的目的，故未给出。

现在 $n \cdot R^2 = 41 \times 0.114\,8 = 4.706\,8$，它渐近地服从自由度为 5 的 χ^2 分布。（为什么？）对于 5 个自由度，5%的 χ^2 临界值是 11.070 5；10%的临界值是 9.236 3；25%的临界值是 6.625 68。为了一切实际目的，我们都可下结论说，根据怀特检验，这里不存在异方差性。

对怀特检验作一评论是适宜的。如果模型有多个回归元，那么引进所有的回归元及其平方（或更高次方）项和它们的交叉乘积就会迅速消耗掉许多自由度。因此，在使用怀特检验时要保持警觉。[②]

在方程（11.5.2）中给出的怀特统计量统计显著的情形下，异方差性并非必然的原因，也可能是设定误差，在第 13 章中我们将更详细的讨论设定误差（回顾 11.1 节中的第 5 点理由）。换句话说，**怀特检验可能是（纯粹）异方差性的一个检验，或者是设定错误的一个检验，或者两者兼有**。已经被证明，若怀特检验程序中没有出现交叉项，则是对纯粹异方差性的检验；若出现交叉项，则既是对异方差性又是对设定偏误的检验。[③]

其他异方差性检验。还有若干个其他异方差性检验，每个都依赖于一定的假定。有兴趣的读者可参阅有关文献。[④] 我们只提出其中的一种，因为它特别简单。这就是**寇因克-巴塞特检验**（Koenker-Basett，KB test）。与异方差性的帕克检验、布罗施-帕甘-戈弗雷检验和怀特检验相似，KB 检验也是基于残差的平方 $\hat{u}_i^2$，但不是对一个或多个回归元做回归，而是对回归子估计值的平方进行回归。具体而言，若原模型是

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad (11.5.26)$$

---

① 这些结果复制于 William F. Lott and Subhash C. Ray, *Applied Econometrics: Problems with Data Sets*, Instructor's Manual, Chapter 22, pp. 137-140，但符号有所改变。

② 为节省自由度，有时候也可以把这个检验加以修改，见习题 11.18。

③ 参见 Richard Harris, *Using Cointegration Analysis in Econometrics Modelling*, Prentice Hall & Harvester Wheatsheaf, U. K., 1995, p. 68。

④ 参见 M. J. Harrison and B. P. McCabe, "A Test for Heteroscedasticity Based on Ordinary Least Squares Residuals," *Journal of the American Statistical Association*, vol. 74, 1979, pp. 494-499; J. Szroeter, "A Class of Parametric Tests for Heteroscedasticity in Linear Econometric Models," *Econometrica*, vol. 46, 1978, pp. 1311-1327; M. A. Evans and M. L. King, "A Further Class of Tests for Heteroscedasticity," *Journal of Econometrics*, vol. 37, 1988, pp. 265-276; 以及 R. Koenker and G. Bassett, "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, vol. 50, 1982, pp. 43-61。

估计此模型并从中得到 $\hat{u}_i$，然后估计

$$\hat{u}_i^2 = \alpha_1 + \alpha_2(\hat{Y}_i)^2 + v_i \tag{11.5.27}$$

其中 $\hat{Y}_i$ 是从模型（11.5.26）中得到的估计值。虚拟假设是 $\alpha_2 = 0$。若未被拒绝，则可以断定不存在异方差性。利用通常的 t 检验或 F 检验就能检验虚拟假设。（注意 $F_{1,k} = t_k^2$。）若模型（11.5.26）是双对数模型，则将残差平方对 $(\log\hat{Y}_i)^2$ 进行回归。KB 检验的另一个优点在于，即便原模型（11.5.26）中的误差项不是正态分布的，它仍能适用。你若将 KB 检验用于例 11.1，你将发现，将方程（11.5.3）中得到的残差的平方对方程（11.5.3）中估计的 $\hat{Y}_i^2$ 进行回归，斜率系数统计上不异于零，从而加强了帕克检验的结论。由于此例中只有一个回归元，所以这个结果无足为奇，但 KB 检验在一个或多个回归元的情况下都能适用。

对异方差检验的一个注解。我们在本节已经讨论了几个异方差检验。那么，我们如何判断哪个检验最好呢？这是一个不太容易回答的问题，因为这些检验基于的假定彼此不同。在比较这些检验时，我们需要注意检验尺度（显著性水平）、检验功效（拒绝一个错误假设的概率）和对异常观测的敏感程度。

我们已经指出，流行并易于实施的怀特异方差检验也有一些局限。正是由于这些局限，所以它对对立假设的功效降低。此外，在辨别导致异方差性的因素或变量时，怀特检验没有多大帮助。

类似地，布罗施-帕甘-戈弗雷检验对正态性假定过于敏感。相比之下，寇因克-巴塞特检验就不依赖于正态性假定，并因而可能更有功效。[①] 在戈德菲尔德-匡特检验中，如果我们遗漏过多观测，我们也可能降低了检验功效。

对各种异方差检验进行比较分析超出了本书的范围。但感兴趣的读者可以参考约翰·莱昂（John Lyon）和蔡锦良（Chin-Ling Tsai）的论文，以对各种异方差检验的优劣有所认识。[②]

11.6 补救措施

正如已经看到的那样，异方差性虽然不损坏 OLS 估计量的无偏性和一致性，却使它们不再是有效的，甚至不是渐近（即在大样本中）有效的。效率的缺乏使得通常的假设检验程序变得可疑。因此，补救措施显然是需要的。补救方法可分为两种：σ_i^2 已知时的补救方法和 σ_i^2 未知时的补救方法。

① 详细情况参见 William H. Greene, *Econometric Analysis*, 6th ed., Pearson/Prentice-Hall, New Jersey, 2008, pp. 165-167。

② 见他们的论文："A Comparison of Tests of Heteroscedasticity," *The Statistician*, vol. 45, no. 3, 1996, pp. 337-349。

□ σ_i^2 已知时：加权最小二乘法

正如在 11.3 节所看到的那样，如果已知 σ_i^2，纠正异方差性的最明显方法就是采取加权最小二乘法，因为这样一来，得到的估计量是 BLUE。

例 11.7　加权最小二乘法说明

为说明此法，假定我们要针对表 11—1 中的数据研究薪金与就业人数之间的关系。为简单起见，我们用 1 表示 1～4 人职工组，2 表示 5～9 人职工组，…，9 表示 1 000～2 499 人职工组。我们还可用表中各组就业人数的组中值表示就业人数。

现令 Y 代表职工平均薪金（美元），而 X 代表就业人数，我们做如下回归［参看方程（11.3.6)］：

$$Y_i/\sigma_i = \hat{\beta}_1^*(1/\sigma_i) + \hat{\beta}_2^*(X_i/\sigma_i) + (\hat{u}_i/\sigma_i) \tag{11.6.1}$$

其中 σ_i 是表 11—1 中的薪金标准差。计算此回归所必需的原始数据由表 11—4 给出。

表 11—4　加权最小二乘回归的说明

酬金，Y	就业人数，X	σ_i	Y_i/σ_i	X_i/σ_i
3 396	1	742.2	4.566 4	0.001 3
3 787	2	851.4	4.448 0	0.002 3
4 013	3	727.8	5.513 9	0.004 1
4 104	4	805.06	5.097 8	0.005 0
4 146	5	929.9	4.458 5	0.005 4
4 241	6	1 080.6	3.924 7	0.005 5
4 387	7	1 241.2	3.528 8	0.005 6
4 538	8	1 307.7	3.470 2	0.006 1
4 843	9	1 110.7	4.353 2	0.008 1

注：在回归（11.6.2）中，因变量是（Y_i/σ_i）而自变量是（$1/\sigma_i$）和（X_i/σ_i）。

资料来源：Y_i 和 σ_i（薪金的标准差）的数据来自表 11—1。就业人数 1＝1～4 人职工组，2＝5～9 人职工组……其他数据也来自表 11—1。

在讨论回归结果之前，要注意方程（11.6.1）没有截距项。（为什么?）因此，有必要利用过原点回归模型去估计 β_1^* 和 β_2^*，这是第 6 章中已讨论过的一个问题。但当今大多数计算机软件都有去掉截距项的选择（例如参看 Minitab 或 EViews)。还应注意方程（11.6.1）另一个有趣的特点：它有两个解释变量 $1/\sigma_i$ 和 X_i/σ_i，但如果我们在做薪金对就业人数的回归时用 OLS，则回归中只有一个解释变量 X_i。（为什么?）

WLS 回归的结果如下：

$$\widehat{(Y_i/\sigma_i)} = 3\,406.639(1/\sigma_i) + 154.153(X_i/\sigma_i) \tag{11.6.2}$$

$$(80.983) \qquad (16.959)$$

$$t = (42.066) \qquad (9.090) \qquad R^2 = 0.999\,3^{①}$$

为便于比较，我们给出平常的或不加权的 OLS 回归结果如下：

① 如第 152 页注释①所指出的那样，过原点回归的 R^2 和带截距模型的 R^2 不可直接相比。报告的 R^2＝0.999 3 已考虑到差异。（关于如何修正因无截距而造成 R^2 的差异，详见各种软件，也可参阅附录 6A 第 6A.1 节。）

$$\hat{Y}_i = 3\,417.833 + 148.767X_i \qquad (11.6.3)$$
$$(81.136)\quad (14.418)$$
$$t = (42.125)\quad (10.318) \qquad R^2 = 0.938\,3$$

在习题 11.7 中我们要求读者去比较两种回归。

□ σ_i^2 未知时

如前所述，若已知真实的 σ_i^2，我们可用 WLS 法得到 BLUE 估计量。由于真实的 σ_i^2 鲜为人知，是否有某种方法，即使在有异方差性的情形下，也能获得 OLS 估计量的方差和协方差的（统计上）一致性估计呢？答案是肯定的。

怀特的“异方差一致”方差与标准误。怀特曾证明，可以做出这样一种估计，它可以对真实的参数值作出渐近（即大样本）有效的统计推断。[①] 我们将不讨论数学上的细节，以免超出本书的讨论范围。附录第 11A.4 节勾勒了怀特程序。目前有若干计算机软件在给出平常的 OLS 方差和协方差的同时，也给出怀特的异方差校正方差和标准误。[②] 顺便提一句，怀特的异方差校正标准误又被称为**稳健标准误**（robust standard errors）。

例 11.8　怀特程序的说明

作为一个例子，我们引用格林的一些结果如下[③]：

$$\hat{Y}_i = 832.91 - 1\,834.2\,(\text{Income}) + 1\,587.04\,(\text{Income})^2$$
$$\text{OLS se} = (327.3)\quad (829.0)\quad (519.1)$$
$$t = (2.54)\quad (2.21)\quad (3.06) \qquad (11.6.4)$$
$$\text{怀特 se} = (460.9)\quad (1\,243.0)\quad (830.0)$$
$$t = (1.81)\quad (-1.48)\quad (1.91)$$

其中 Y=1979 年各州公立学校人均支出，Income=1979 年美国各州人均收入。样本由美国 50 个州及华盛顿特区构成。

以上数字结果表明，经（怀特）异方差校正的标准误比 OLS 标准误大得多，因而所估计的 t 值比得自 OLS 的要小得多。根据后者，两个回归元在 5%水平上都是统计显著的，而根据怀特估计量则不然。但应指出，怀特的异方差校正标准误可能大于也可能小于未校正的标准误。

由于当今现成的回归软件包都备有方差的怀特异方差一致估计量，建议读者做回归时予以报告。如华莱士（Wallace）和西尔弗（Silver）所说的：

① 参看怀特的前引文献。

② 更专门化的名称是**异方差性一致的协方差矩阵估计量**（heteroscedasticity-consistent covariance matrix estimators，HCCME）。

③ William H. Greene, *Econometric Analysis*, 2d ed., Macmillan, New York, 1993, p. 385.

一般地说，经常地使用［回归程序中备有的］怀特选择（White option）大概是个好主意，也许通过怀特输出同 OLS 输出相比，可以看出在一组特定的数据中异方差性是否构成一个严重的问题。①

关于异方差性模式的可能假定。怀特程序除了本身是一个大样本程序外，还有一个缺点，就是这样得到的估计量不如先按异方差类型进行数据变换之后再做估计来得有效。为说明这一点，让我们再回到双变量回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

我们现在考虑关于异方差性模式的几种假定。

假定 1：误差方差正比于 X_i^2：

$$E(u_i^2) = \sigma^2 X_i^2 \qquad (11.6.5)$$②

如果作为一种“猜测”或通过描图或通过帕克和格莱泽方法，认为 u_i 的方差正比于解释变量 X 的平方（见图 11—10），则可对原模型做如下变换：用 X_i 去除原模型：

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

$$= \beta_1 \frac{1}{X_i} + \beta_2 + v_i \qquad (11.6.6)$$

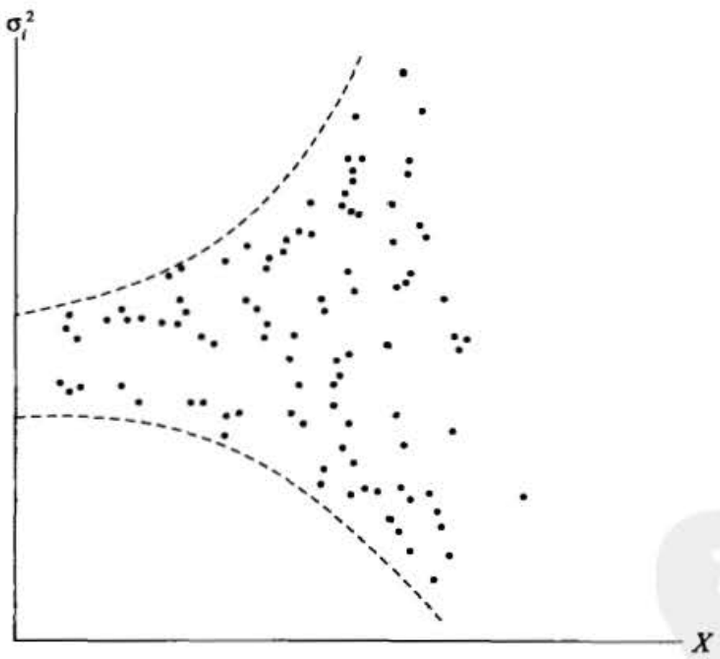


图 11—10　误差方差正比于 X^2

① T. Dudley Wallace and J. Lew Silver, *Econometrics: An Introduction*, Addison-Wesley, Reading, Mass, 1988, p. 265.

② 我们回想在戈德菲尔德-匡特检验的讨论中曾遇见这种假定。

其中 v_i 是变换后的干扰项，等于 u_i/X_i。现在容易验证，

$$E(v_i^2) = E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2}E(u_i^2) = \sigma^2 \qquad \text{利用 (11.6.5)}$$

从而 v_i 的方差是同方差的，并可对变换方程（11.6.6）进行 OLS，做 Y_i/X_i 对 $1/X_i$ 的回归。

注意，在变换后的回归中，截距项 β_2 是原方程的斜率系数，而斜率系数 β_1 则是原方程中的截距项。因此，要回到原来的模型，需用 X_i 乘以估计的方程（11.6.6）。习题 11.20 给出这种变换的一个应用。

假定 2：误差方差正比于 X_i。**平方根变换**：

$$E(u_i^2) = \sigma^2 X_i \qquad (11.6.7)$$

如果认为 u_i 的方差不是正比于 X_i 的平方而是正比于 X_i 本身，就可将原始模型变换如下（参见图 11—11）：

$$\begin{aligned}\frac{Y_i}{\sqrt{X_i}} &= \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} \\ &= \beta_1\frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + v_i \qquad (11.6.8)\end{aligned}$$

其中 $v_i = u_i/\sqrt{X_i}$ 且 $X_i > 0$。

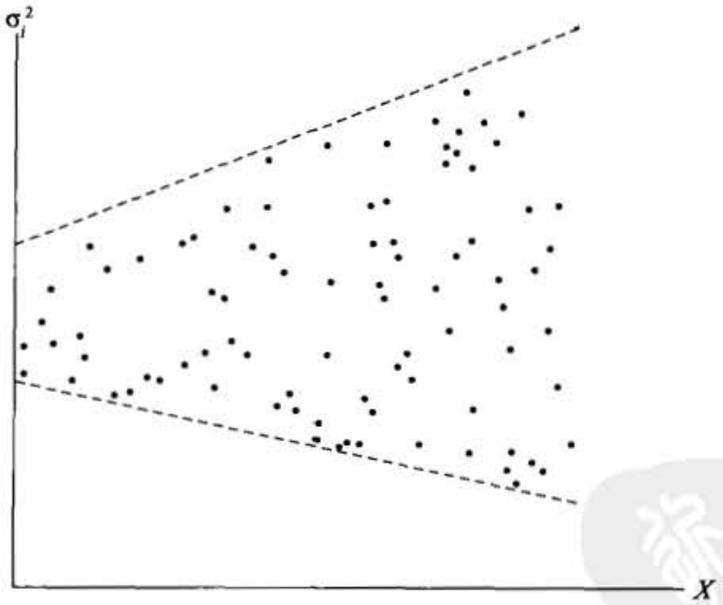


图 11—11　误差方差正比于 X

在假定 2 下容易验证 $E(v_i^2) = \sigma^2$ 为同方差情形，因此，可按 OLS 对方程（11.6.8）进行 $Y_i/\sqrt{X_i}$ 对 $1/\sqrt{X_i}$ 和 $\sqrt{X_i}$ 的回归。

注意变换后模型的一个重要特点：它没有截距项。因此要用过原点回归模型去估计 β_1 和 β_2。在做完回归（11.6.8）之后，只需乘以 $\sqrt{X_i}$，即回到原始模型。

一个有意思的情形是零截距模型 $Y_i = \beta_2 X_i + u_i$。在此情形中，方程（11.6.8）变为

$$\frac{Y_i}{\sqrt{X_i}} = \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} \tag{11.6.8a}$$

而且可以证明

$$\hat{\beta}_2 = \frac{\overline{Y}}{\overline{X}} \tag{11.6.8b}$$

也就是说，加权最小二乘估计量无非就是因变量的均值与解释变量的均值之比。[欲证明方程（11.6.8b），只需使用方程（6.1.6）中给出的过原点回归公式即可。]

假定 3：误差方差正比于 Y 均值的平方。

$$E(u_i^2) = \sigma^2 [E(Y_i)]^2 \tag{11.6.9}$$

方程（11.6.9）假定 u_i 的方差正比于 Y 期望值的平方（参见图 11—8e）。现在

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

因此，若将原模型变换如下：

$$\begin{aligned}\frac{Y_i}{E(Y_i)} &= \frac{\beta_1}{E(Y_i)} + \beta_2\frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)} \\ &= \beta_1\left(\frac{1}{E(Y_i)}\right) + \beta_2\frac{X_i}{E(Y_i)} + v_i\end{aligned} \tag{11.6.10}$$

其中 $v_i = u_i/E(Y_i)$，可以看出 $E(v_i^2) = \sigma^2$；即干扰项 v_i 是同方差的。从而回归（11.6.10）满足经典线性回归模型的同方差假定。

然而，由于 $E(Y_i)$ 依赖于未知的 β_1 和 β_2，变换方程（11.6.10）是无法操作的。当然，我们知道 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$，这是 $E(Y_i)$ 的一个估计量。因此，可按两步进行：第一步，暂且忽略异方差性的问题，作平常的 OLS 回归并获得 $\hat{Y}_i$。然后利用估计的 $\hat{Y}_i$ 作如下的模型变换：

$$\frac{Y_i}{\hat{Y}_i} = \beta_1\left(\frac{1}{\hat{Y}_i}\right) + \beta_2\left(\frac{X_i}{\hat{Y}_i}\right) + v_i \tag{11.6.11}$$

其中 $v_i = (u_i/\hat{Y}_i)$。第二步，我们做回归（11.6.11）。虽然 $\hat{Y}_i$ 并不正好等于 $E(Y_i)$，但 $\hat{Y}_i$ 是一致估计量；当样本无限增大时，它们将趋于 $E(Y_i)$ 的真值。因此，如果样本容量足够大，变换（11.6.11）实际上会有令人满意的表现。

假定 4：和回归 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 相比，

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \tag{11.6.12}$$

这样的对数变换常常能降低异方差性。

之所以出现这种结果，是因为对数变换压缩了测量变量的尺度，把两个值的 10 倍之差降低到约 2 倍之差。例如，数值 80 十倍于数值 8。但ln80（=4.328 0)仅约 2

倍于 ln8（=2.079 4）。

对数变换的另一优点是，斜率系数 β_2 度量了 Y 对 X 的弹性，即对应于 X 的 1% 的变化，Y 的百分比变化。例如，Y 是消费而 X 是收入，方程（11.6.12）中的 β_2 将测出收入弹性，而在原始模型中，β_2 仅测出对应于收入的单位变化，平均消费的变化率。这就是对数模型在经验计量经济学中广为应用的原因之一。（至于对数变换所带来的一些问题，参看习题 11.4。）

在结束我们对补救措施的讨论之时，我们再次强调，以上所有讨论的变换都是一种权宜之计。我们基本上是在猜测 σ_i^2。在所讨论的变换中哪一种能行之有效，要看问题的性质和异方差性的严重程度。应记住，我们所考虑的变换还存在其他的一些问题：

1. 当我们超出双变量模型的范围时，我们也许不能预先知道应选择哪一个 X 变量进行数据变换。①

2. 在假定 4 中讨论的对数变换，当某些 Y 和 X 值为零或负数时便不适用。②

3. 然后，还有一个**谬误相关**（spurious correlation）的问题。该词来自卡尔·皮尔逊（Karl Pearson），指的是即使原始变量是不相关或随机的，但变量的比率却被发现有相关关系的情形。③ 例如在模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 中，Y 和 X 也许不相关。但在变换模型 $Y_i/X_i = \beta_1(1/X_i) + \beta_2$ 中，Y_i/X_i 和 $1/X_i$ 却常常被发现存在相关关系。

4. 当无法直接得知 σ_i^2 而要从前面讨论的一个或多个变换中作出估计时，所有用到的 t 检验、F 检验等检验程序，严格地说，**都只在大样本中有效**。因此，在小样本或有限样本中，如何根据各种变换去解释所得到的结果，必须多加小心。④

11.7 总结性的例子

在结束我们对异方差性的讨论之际，我们用三个例子说明侦察它的各种方法以及对它的一些补救措施。

例 11.9

让我们再次回到曾考虑过几次的儿童死亡率一例。我们从 64 个国家的数据得到方程（8.1.4）

① 然而，一种实际的做法也许是将 a_i^2 对每个变量描图，然后决定用哪个 X 变量去变换数据。（见图 11—9。）

② 有时可用 $\ln(Y_i + k)$ 或 $\ln(X_i + k)$，其中 k 是一个有待选择的正数，目的是要所有的 Y 和 X 值都变换成正数。

③ 例如，如果 X_1，X_2 与 X_3 彼此无关：$r_{12} = r_{13} = r_{23} = 0$，而我们发现比值 X_1/X_3 与 X_2/X_3 相关，则称此为谬误相关，“更一般地讲，如果相关不出现在原始数据中，而是由于数据的处理方法而引发出来的，则这种相关可以说是谬误的。”见 M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms*, Hafner Publishing, New York, 1972, p. 143。

④ 更多细节，见 George G. Judge et al., op. cit., Section 14.4, pp. 415-420。

中所示的回归结果。由于数据是横截面数据，涉及的国家在儿童死亡率上有不同的表现，所以很可能会出现异方差性。为探明究竟，首先看从方程（8.1.4）中得到的残差。这些残差画在图 11—12 中。从图上看，残差没有显示出任何存在异方差性的明显形式。尽管如此，表面现象仍可能有欺骗性。所以，我们用帕克、格莱泽和怀特检验，看是否有异方差性的证据。

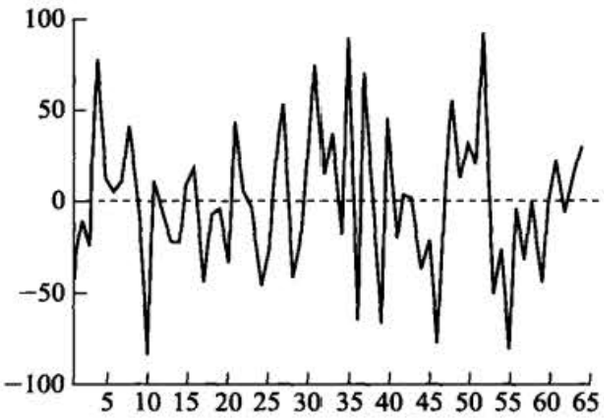


图 11—12　回归（8.1.4）中的残差

帕克检验。由于有两个回归元 GNP 和 FLR，所以我们可以将回归（8.1.4）中残差的平方对其中任意一个回归，或者将它们对回归（8.1.4）中估计出来的 CM 值（$=\widehat{\text{CM}}$）做回归。利用后者，我们得到如下结论。

$$\widehat{u_i^2} = 854.400\,6 + 5.701\,6\,\widehat{\text{CM}_i}$$
$$t = (1.201\,0)\quad (1.242\,8) \qquad r^2 = 0.024 \tag{11.7.1}$$

注：u_i 是从回归（8.1.4）中得到的残差，$\widehat{\text{CM}}$ 是从回归（8.1.4）中估计出来的 CM 值。

如此回归所示，残差的平方与估计的 CM 之间没有系统的关系（为什么?），这就表明同方差性可能站得住脚。顺便一提，将残差平方的对数对 $\widehat{\text{CM}}$ 的对数做回归，也不会改变这个结论。

格莱泽检验。将方程（8.1.4）中所得到残差的绝对值对同一回归所估计的 CM 值做回归，得到如下结论：

$$|\widehat{u_i}| = 22.312\,7 + 0.064\,6\,\widehat{\text{CM}_i}$$
$$t = (2.808\,6)\quad (1.262\,2) \qquad r^2 = 0.025\,0 \tag{11.7.2}$$

同样，由于斜率系数的 t 值并非统计显著，所以残差的绝对值与估计的 CM 值之间没有系统的关系。

怀特检验。应用含交叉项和不含交叉项的怀特异方差检验，我们没有发现任何异方差性证据。我们也重新估计（8.1.4）以得到怀特异方差一致的标准误和 t 值，结论与方程（8.1.4）中给出的那些结论十分相似，从我们前面所做的各种异方差检验来看，无足为奇。

总之，儿童死亡率回归（8.1.4）看来不存在异方差性的问题。

例 11.10　2005 年美国 14 个产业群的 R&D 支出、销售额和利润

表 11—5 给出了美国 14 个产业群的研发支出（R&D）、销售额（Sales）和利润数据（Profits），所有数据都以百万美元计。由于表中横截面数据差别很大，所以在 R&D 对销售额（或利

润）的回归中就可能出现异方差性。回归结果如下：

$$\widehat{R\&D_i} = 1\,338 + 0.043\,7Sales_i$$
$$se = (5\,015) \quad (0.027\,7) \tag{11.7.3}$$
$$t = (0.27) \quad (1.58) \qquad r^2 = 0.172$$

无足为奇，R&D与销售额之间有明显的正相关关系，尽管在传统的显著性水平不是统计显著的。

表 11—5　　2005 年美国不同行业销售额和研发支出数据　　（单位：百万美元）

行业	销售额 Sales	研发支出 R&D	利润 Profits
1. 食品	374 342	2 716	234 662
2. 纺织、服装与皮革	51 639	816	53 510
3. 基础化工	109 899	2 277	75 168
4. 树脂、合成橡胶、纤维和丝制品	132 934	2 294	34 645
5. 医药	273 377	34 839	127 639
6. 玻璃与塑料制品	90 176	1 760	96 162
7. 金属制品	174 165	1 375	155 801
8. 机械	230 941	8 531	143 472
9. 计算机及配件	91 010	4 955	34 004
10. 半导体及其他电子元件	176 054	18 724	81 317
11. 航空、测量、电子治疗器械及其他控制设备	118 648	15 204	73 258
12. 电子设备、家具及其配件	101 398	2 424	54 742
13. 航空航天产品及部件	227 271	15 005	72 090
14. 医疗设备与用品	56 661	4 374	52 443

资料来源：National Science Foundation, Division of Science Resources Statistics, Survey of Industrial Research and Development: 2005 and the U. S. Census Bureau Annual Survey of Manufacturers, 2005.

为了看出回归（11.7.3）是否遇到异方差的问题，我们从上述回归中得到残差 $\hat{u}_i$ 及其平方 $\hat{u}_i^2$，并相对销售额进行描点，如图 11—13 所示。从此图来看，残差及其平方与销售额之间有系统关系，表明可能存在异方差性。为规范地进行检验，我们使用帕克、格莱泽和怀特检验，并给出如下结果：

帕克检验。

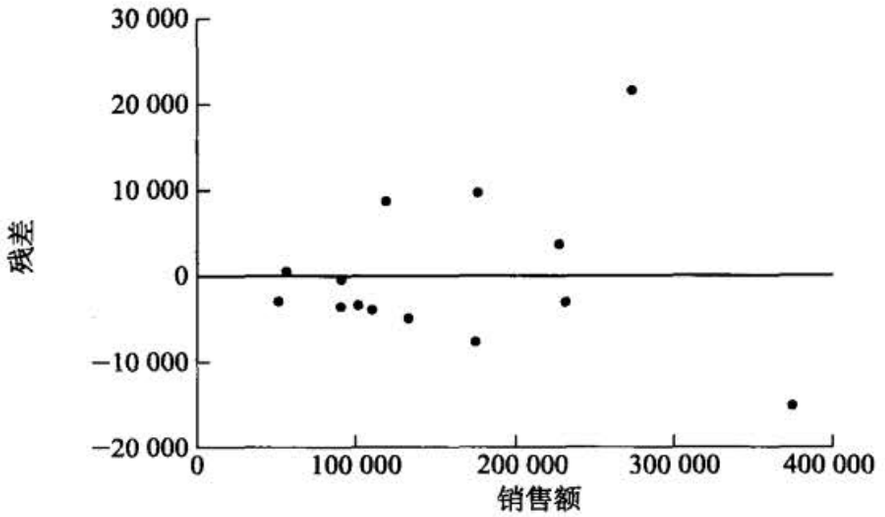
$$\widehat{\hat{u}_i^2} = -72\,493\,719 + 916.1Sales_i$$
$$se = (54\,940\,238) \quad (303.9) \tag{11.7.4}$$
$$t = \quad (-1.32) \quad (3.01) \qquad r^2 = 0.431$$

帕克检验表明，残差平方与销售额之间存在着统计显著的正相关。

格莱泽检验。

$$\widehat{|\hat{u}_i|} = -1\,003 + 0.046\,39\,Sales_i$$
$$se = \quad (2\,316) \quad (0.012\,8) \tag{11.7.5}$$
$$t = (-0.43) \quad (3.62) \quad r^2 = 0.522$$

格莱泽检验也表明，残差的绝对值与销售额之间也有系统的关系，从而增加了回归（11.7.3）存在异方差性的可能性。



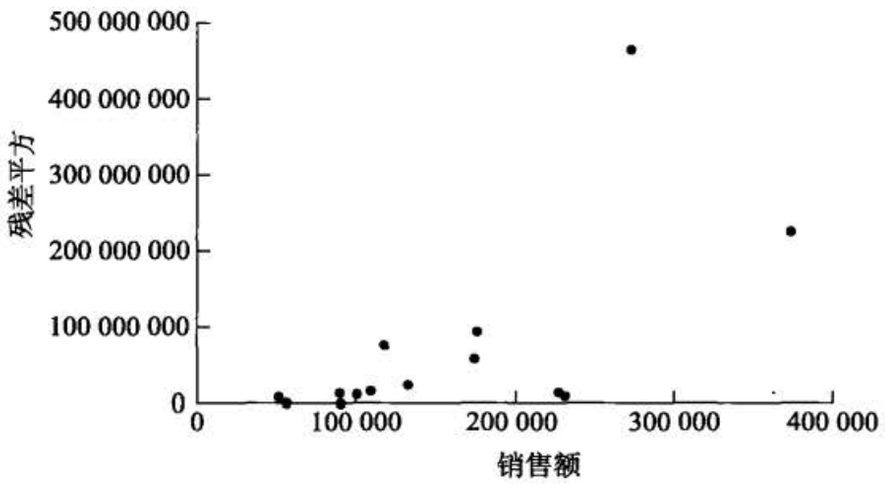


图 11—13　残差及残差平方对销售额的散点图

怀特检验。

$$\widehat{u_i^2} = -46\ 746\ 325 + 578\ \text{Sales}_i + 0.000\ 846\ \text{Sales}_i^2$$

$$\text{se} = (112\ 224\ 348)\quad (1\ 308)\quad (0.003\ 171) \qquad (11.7.6)$$

$$t = \quad (-0.42)\quad (0.44)\quad (0.27)\qquad R^2 = 0.435$$

利用 R^2 值和 $n=14$，我们得到 $nR^2=6.090$，在不存在异方差性的虚拟假设之下，服从自由度为 2 的 χ^2 分布［由于方程（11.7.6）中有两个回归元］。得到大于等于 6.090 的一个 χ^2 值的 p 值约为 0.047 6。既然这个 p 值足够低，那么，怀特检验也表明存在异方差性。

总之，基于残差图及帕克、格莱泽和怀特检验，我们在方程（11.7.3）中所做的 R&D 回归遇到了异方差性的问题。由于真实的误差方差未知，所以欲得到异方差校正的标准误和 t 值，我们还不能使用加权最小二乘法。于是，对于误差方差的性质，我们必须做出某种有经验的猜测。

作为本例的结束，我们如在 11.6 节中所讨论的那样给出如下怀特异方差一致标准误：

$$\widehat{\text{R\&D}_i} = 1\ 337.87 + 0.043\ 7\ \text{Sales}_i$$

$$\text{se} = (4\ 892.447)\quad (0.041\ 1) \qquad (11.7.7)$$

$$t = \quad (0.27)\quad (1.06)\qquad r^2 = 0.172$$

与原回归（11.7.3）（即不对异方差性校正）相比，我们看到，尽管参数的估计值没有变化（与我

们的预料相一致），但截距系数的标准误下降了，而斜率系数的标准误则略有上升。但须牢记，怀特程序是一个严格的大样本程序，而我们这里只有14个观测。

例 11.11

本书网站上的表11—6给出了俄亥俄州西北部94个学区的薪水及有关数据。最初对这些数据估计了如下回归：

$$\ln(\text{Salary})_i = \beta_1 + \beta_2 \ln(\text{Famincome}) + \beta_3 \ln(\text{Propvalue}) + u_i$$

其中 Salary＝教师的平均薪水（美元），Famincome＝该学区的家庭平均收入（美元），而 Propvalue＝该学区的财富均值（美元）。

由于这是一个双对数模型，所以所有斜率系数都表示弹性。基于本书讨论的各种异方差检验，发现上述模型存在异方差问题。因此，我们求了（怀特的）稳健标准误。下表给出了上述回归使用和不使用稳健标准误的回归结果。

变量	系数	OLS标准误	稳健标准误
截距项	7.019 8	0.805 3	0.772 1
		(8.717 1)	(9.090 8)
ln(Famincome)	0.257 5	0.079 9	0.100 9
		(3.223 0)	(2.551 6)
ln(Propvalue)	0.070 4	0.020 7	0.046 0
		(3.397 6)	(1.531 1)
R^2	0.219 8		

注：圆括号中的数字是估计的 t 比率。

尽管无论我们使用OLS还是使用怀特的方法得到的系数值和 R^2 都保持不变，但标准误发生了变化；ln(Propvalue) 的系数标准误变化最明显。通常的OLS表明该变量的系数估计值是高度统计显著的，而怀特的稳健标准误却表明，这个系数即便在10%的显著性水平上也不显著。本例要说明的是，如果存在异方差性，我们应该在估计模型时加以考虑。

11.8 谨防对异方差性反应过度

回到上一节中讨论的R&D的例子，我们看到，当我们对原模型（11.7.3）进行平方根变换来校正其异方差性时，斜率系数的标准误下降了，其 t 值上升了。这一变化显著到值得担心的程度了吗？换言之，我们什么时候应该真正担心异方差性的问题？如一位作者所言："一个好的模型，绝不会因异方差性的原因而被抛弃。"[①]

① N. Gregory Mankiw, "A Quick Refresher Course in Macroeconomics," *Journal of Economic Literature*, vol. XXVIII, December 1990, p. 1648.

这里，牢记约翰·福克斯（John Fox）的警告会有所帮助。

……只有在问题严重的时候，误差方差不相等的问题才值得去修正。

误差方差不是常数，对普通最小二乘估计量的有效性和最小二乘推断的可靠性所产生的影响，取决于多个因素，包括：样本容量、σ_i^2 中变异的程度、X（即回归元）值的结构及误差方差与 X 之间的关系。因此，就异方差所导致的危害而言，不可能得到一个纯粹一般性的结论。①

回顾模型（11.3.1），我们已看到斜率估计量的方差 var $(\hat{\beta}_2)$ 由方程（11.2.3）所示的常用表达式给出。在广义最小二乘法下，斜率估计量的方差 var $(\hat{\beta}_2^*)$ 由方程（11.3.9）给出。我们知道，后者比前者更有效。但二者之间的差别到底要多大，我们才应该真正担心呢？作为一个经验法则，福克斯建议，“当最大方差比最小方差的10倍还大时”②，我们就要担心这个问题。于是，回到前面提到的戴维森和麦金农的蒙特卡罗模拟结果，考虑 $\alpha=2$ 的情况。所估计 β_2 的方差在 OLS 下为 0.04，而在 GLS 下为 0.012，前者约为后者的 3.33 倍。③ 根据福克斯法则，在这种情况下，异方差性的严重程度不足以引起担心。

还要记住，尽管有异方差性的问题，但 OLS 估计量仍是线性无偏和渐近（即在大样本中）正态分布的（在一般条件下）。

在我们讨论其他违背经典线性回归模型之假定情况时，我们将会看到，在本节敲响的警钟作为一个一般规则也是适当的。否则，就有可能反应过度。

要点与结论

1. 经典线性回归模型的一个关键假定是干扰项 u_i 都有相同的方差 σ^2。如果此假定不成立，则说有异方差性。

2. 异方差性并不破坏 OLS 估计量的无偏性和一致性性质。

3. 但这些估计量不再是最小方差或有效的，也就是说，它们不是 BLUE。

4. 如果相异的误差方差 σ_i^2 已知，加权最小二乘法（WLS）可给出 BLUE 估计量。

5. 当异方差性出现时，OLS 估计量的方差不能由常用的 OLS 公式给出。如果我们一味地使用 OLS 公式，则以这些公式为依据的 t 检验和 F 检验可能严重误导，以致得出错误的结论。

6. 列举异方差性的后果容易，侦察异方差的工作困难。现有若干诊断性检验，但无法告知在特定情况中哪一检验能行之有效。

7. 即使异方差性受到怀疑并且被侦察出来，如何纠正并非易事。如果样本足够大，则可获取

① John Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, California, 1997, p. 306.

② Ibid., p. 307.

③ 注意，我们为了得到方差已经把标准误平方了。

OLS估计量的怀特异方差校正标准误并以之作为统计推断的依据。

8. 另外，根据OLS残差，我们可以合理地猜测异方差性的可能模式，以便将原始数据变换成没有异方差性的数据加以使用。

习　　题

问答题

11.1　用简明的理由说明以下命题是正确的、错误的或者不确定的。

a. 当异方差性出现时，OLS估计量是有偏误的和非有效的。

b. 如果出现异方差性，则惯用的 t 检验和 F 检验无效。

c. 在异方差性的情况下，常用的OLS法必定高估了估计量的标准误。

d. 如果OLS回归的残差表现出系统模式，这就说明数据中存在异方差性。

e. 没有任何一般性的异方差性检验能独立于误差项与某一变量相关的假定。

f. 如果一个回归模型误设（比如说，漏掉一个重要变量），则OLS残差必定表现出明显的样式。

g. 如果模型不正确地漏掉一个有非恒定方差的回归元，则（OLS）残差将是异方差性的。

11.2　在对一个含有30个厂商的随机样本做的平均薪金（W）对职工人数（N）的回归中，得到如下回归结果①：

$$\hat{W} = 7.5 + 0.009N$$
$$t = \text{n.a.} \quad (16.10) \qquad R^2 = 0.90 \tag{1}$$
$$\hat{W}/N = 0.008 + 7.8\,(1/N)$$
$$t = (14.43)\ (76.58) \qquad R^2 = 0.99 \tag{2}$$

a. 你怎样解释这两个回归？

b. 从方程（1）到方程（2）作者做了什么假定？他是否担心过异方差性？你怎样知道？

c. 怎样把这两个模型的截距和斜率联系起来？

d. 你能比较两个模型的 R^2 值吗？为什么？

11.3　a. 你能用OLS法估计下列模型中的参数吗？为什么？

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$
$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

b. 如果不能，你能提出一个估计这些模型参数的正式或非正式的方法吗？（参见第14章。）

11.4　虽然如方程（11.6.12）所示的对数模型常常能降低异方差性，但需特别注意这种模型误差项的性质。例如，模型

$$Y_i = \beta_1 X_i^{\beta_2} u_i \tag{1}$$

可以写为：

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + \ln u_i \tag{2}$$

a. 如果 $\ln u_i$ 有零期望值，u_i 的分布必须是什么？

① 参见 Dominick Salvatore, *Managerial Economics*, McGraw-Hill, New York, 1989, p. 157。

b. 如果 $E(u_i)=1$，会不会有 $E(\ln u_i)=0$？为什么？

c. 如果 $E(\ln u_i)$ 不为零，怎样能使它等于零？

11.5 证明方程（11.3.8）中的 β_2^* 还可表达为：

$$\beta_2^* = \frac{\sum w_i y_i^* x_i^*}{\sum w_i x_i^{2*}}$$

以及方程（11.3.9）中的 $\text{var}(\beta_2^*)$ 又可表达为：

$$\text{var}(\beta_2^*) = \frac{1}{\sum w_i x_i^{2*}}$$

其中 $y_i^* = Y_i - \bar{Y}^*$ 和 $x_i^* = X_i - \bar{X}^*$ 代表对加权均值 $\bar{Y}^*$ 和 $\bar{X}^*$ 的离差，其中：

$$\bar{Y}^* = \sum w_i Y_i / \sum w_i$$

$$\bar{X}^* = \sum w_i X_i / \sum w_i$$

11.6 为了教学的目的，哈努谢克（Hanushek）和杰克逊（Jackson）估计了如下模型：

$$C_t = \beta_1 + \beta_2 \text{GNP}_t + \beta_3 D_t + u_i \tag{1}$$

其中 C_t=年度 t 的私人总消费支出，GNP_t=年度 t 的国民生产总值，以及 D_t=年度 t 的国防支出。分析的目的在于研究国防支出对经济中其他支出的影响。

他们假设 $\sigma_i^2 = \sigma^2(\text{GNP}_t)^2$，从而将（1）变换如下，并加以估计：

$$C_t/\text{GNP}_t = \beta_1(1/\text{GNP}_t) + \beta_2 + \beta_3(D_t/\text{GNP}_t) + u_t/\text{GNP}_t \tag{2}$$

根据 1946—1975 年的数据得到的经验结果如下（括号中为标准误）[①]：

$$\hat{C}_t = 26.19 + 0.624\,8\,\text{GNP}_t - 0.439\,8D_t$$
$$(2.73)\quad (0.006\,0)\qquad (0.073\,6)\qquad R^2 = 0.999$$

$$\widehat{C_t/\text{GNP}_t} = 25.92(1/\text{GNP}_t) + 0.624\,6 - 0.431\,5(D_t/\text{GNP}_t)$$
$$(2.22)\qquad (0.006\,8)\ (0.059\,7)\qquad R^2 = 0.875$$

a. 作者对异方差性的性质做了什么假定？你能说明其中的理由吗？

b. 比较两个回归的结果。对原始模型的变换是否使结果有所改进，也就是说，降低了估计的标准误吗？为什么？

c. 你能比较这两个 R^2 值吗？为什么？（提示：检查因变量。）

11.7 参照方程（11.6.2）和（11.6.3）中估计的回归。两个回归的结果十分相近。这是什么原因？

11.8 证明：若对每一个 i，$w_i = w$ 都为常数，则 β_2^* 和 $\hat{\beta}_2$ 以及它们的方差都是相同的。

11.9 参照公式（11.2.2）和（11.2.3）。假定：

$$\sigma_i^2 = \sigma^2 k_i$$

其中 σ^2 是常数，而 k_i 是已知但不一定相等的权数。

利用此假定，证明方程（11.2.2）中的方差可表达为：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \cdot \frac{\sum x_i^2 k_i}{\sum x_i^2}$$

右端第一项即是方程（11.2.3）中的方差公式，也就是在同方差性下的 $\text{var}(\hat{\beta}_2)$。你能说出在异

① Eric A. Hanushek and John E. Jackson, *Statistical Methods for Social Scientists*, Academic, New York, 1977, p. 160.

方差性下的 var $(\hat{\beta}_2)$ 和在同方差性下的 var $(\hat{\beta}_2)$ 之间的关系具有什么性质吗？（提示：分析以上公式的右端第二项。）你能对方程（11.2.2）和（11.2.3）之间的关系得出任何一般性的结论吗？

11.10 在模型（注：没有截距）

$$Y_i = \beta_2 X_i + u_i$$

中，你被告知 $\text{var}(u_i) = \sigma^2 X_i^2$。证明：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2 \sum X_i^4}{(\sum X_i^2)^2}$$

实证分析题

11.11 用表 11—1 的数据做平均薪金 Y 对平均生产力 X 的回归，把就业人数当作观测单元。解释你的结果，并察看你的结果是否和方程（11.5.3）给出的结果一致。

a. 从上面的回归算出残差 $\hat{u}_i$。

b. 按照帕克检验，将 $\ln \hat{u}_i^2$ 对 $\ln X_i$ 回归，并验证回归方程（11.5.4）。

c. 按照格莱泽方法，将 $|\hat{u}_i|$ 对 X_i 回归，再将 $|\hat{u}_i|$ 对 $\sqrt{X_i}$ 回归。然后评述你的结果。

d. 求 $|\hat{u}_i|$ 对 X_i 的等级相关，然后对数据中是否有异方差性以及它的性质进行评论。

11.12 表 11—6 给出 1971 年第Ⅰ季度至 1973 年第Ⅳ季度期间美国制造行业按公司的资产规模分类的销售/现金比率数据。（数据是按季度报告的。）销售/现金比率可看作公司部门收入流转速度的一个度量，也就是一个美元的周转次数。

表 11—6 资产规模 （单位：百万美元）

年度与季度	1～10	10～25	25～50	50～100	100～250	250～1 000	1 000+
1971－Ⅰ	6.696	6.929	6.858	6.966	7.819	7.557	7.860
－Ⅱ	6.826	7.311	7.299	7.081	7.907	7.685	7.351
－Ⅲ	6.338	7.035	7.082	7.145	7.691	7.309	7.088
－Ⅳ	6.272	6.265	6.874	6.485	6.778	7.120	6.765
1972－Ⅰ	6.692	6.236	7.101	7.060	7.104	7.584	6.717
－Ⅱ	6.818	7.010	7.719	7.009	8.064	7.457	7.280
－Ⅲ	6.783	6.934	7.182	6.923	7.784	7.142	6.619
－Ⅳ	6.779	6.988	6.531	7.146	7.279	6.928	6.919
1973－Ⅰ	7.291	7.428	7.272	7.571	7.583	7.053	6.630
－Ⅱ	7.766	9.071	7.818	8.692	8.608	7.571	6.805
－Ⅲ	7.733	8.357	8.090	8.357	7.680	7.654	6.772
－Ⅳ	8.316	7.621	7.766	7.867	7.666	7.380	7.072

资料来源：*Quarterly Financial Report for Manufacturing Corporations*, Federal Trade Commission and the Securities and Exchange Commission, U. S. government, various issues (computed).

a. 对每一资产规模计算销售/现金比的均值与标准差。

b. 将（a）中计算的均值对标准差描点，把资产规模当作观测单元。

c. 通过适当的回归模型，判断比率的标准差是否随均值增加而增加。如果没有这种关系，怎样自圆其说？

d. 如果两者有统计上的显著关系，你会怎样变换数据以使异方差性不复存在？

11.13 **巴特利特（Bartlett）的同方差检验。**[①] 假使有自由度为 f_1，f_2，…，f_k 的 k 个独立样本方差 s_1^2、s_2^2，…，s_k^2，各来自以 u 为均值和 σ_i^2 为方差的正态分布的总体。再假使我们要检验虚拟假设 H_0：$\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2$；即每一样本方差都是同一总体方差 σ^2 的一个估计值。

如果虚拟假设真实，则：

$$s^2=\frac{\sum_{i=1}^{k} f_i s_i^2}{\sum f_i}=\frac{\sum f_i s_i^2}{f}$$

给出总体方差 σ^2 的共同（联合）估计的一个估计值，其中 $f_i=(n_i-1)$，而 n_i 为第 i 组的观测个数，并且 $f=\sum_{i=1}^{k} f_i$。

巴特利特证明，虚拟假设可通过近似于 $k-1$ 个自由度的 χ^2 分布的比率 A/B 加以检验，其中：

$$A=f\ln s^2-\sum(f_i\ln s_i^2)$$

以及

$$B=1+\frac{1}{3(k-1)}\left[\sum\left(\frac{1}{f_i}\right)-\frac{1}{f}\right]$$

对表 11—1 的数据做巴特利特检验并验证在 5%的显著水平上不能拒绝假设：每一厂商的职工人数组都有相同的总体薪金方差。

注：因每一样本（即就业组）的 n_i 都是 10，故每一样本方差的自由度 f_i 都是 9。

11.14 考虑如下过原点的回归模型：

$$Y_i=\beta X_i+u_i \qquad i=1,2$$

告诉你 $u_1\sim N(0,\sigma^2)$ 和 $u_2\sim N(0,2\sigma^2)$，而且它们相互独立。若 $X_1=+1$，$X_2=-1$，计算 β 的加权最小二乘（WLS）估计量及其方差。若你此时不正确地假定了误差方差相同（比方都等于 σ^2），那么 β 的 OLS 估计量是什么？其方差又是多少？与用 WLS 方法得到的估计量相比，你能得到什么一般性结论？[②]

11.15 表 11—7 给出了 81 辆汽车在 MPG（每加仑耗油量行驶的英里数）、HP（发动机马力）、VOL（驾驶空间的立方英尺数）、SP（最高时速）和 WT（以百磅为单位的车身重量）方面的数据。

表 11—7　　客车单位耗油量行驶里程数据

观测	MPG	SP	HP	VOL	WT	观测	MPG	SP	HP	VOL	WT
1	65.4	96	49	89	17.5	8	59.2	98	62	50	22.5
2	56.0	97	55	92	20.0	9	53.3	98	62	50	22.5
3	55.9	97	55	92	20.0	10	43.4	107	80	94	22.5
4	49.0	105	70	92	20.0	11	41.1	103	73	89	22.5
5	46.5	96	53	92	20.0	12	40.9	113	92	50	22.5
6	46.2	105	70	89	20.0	13	40.9	113	92	99	22.5
7	45.4	97	55	92	20.0	14	40.4	103	73	89	22.5

① 参见 "Properties of Sufficiency and Statistical Test," *Proceedings of the Royal Society of London A*, vol. 160, 1937, p. 268。

② 节选自 F. A. F. Seber, *Linear Regression Analysis*, John Wiley & Sons, New York, 1977, p. 64。

续前表

观测	MPG	SP	HP	VOL	WT	观测	MPG	SP	HP	VOL	WT
15	39.6	100	66	89	22.5	49	31.2	120	130	86	30.0
16	39.3	103	73	89	22.5	50	33.7	109	115	101	35.0
17	38.9	106	78	91	22.5	51	32.6	109	115	101	35.0
18	38.8	113	92	50	22.5	52	31.3	109	115	101	35.0
19	38.2	106	78	91	22.5	53	31.3	109	115	124	35.0
20	42.2	109	90	103	25.0	54	30.4	133	180	113	35.0
21	40.9	110	92	99	25.0	55	28.9	125	160	113	35.0
22	40.7	101	74	107	25.0	56	28.0	115	130	124	35.0
23	40.0	111	95	101	25.0	57	28.0	102	96	92	35.0
24	39.3	105	81	96	25.0	58	28.0	109	115	101	35.0
25	38.8	111	95	89	25.0	59	28.0	104	100	94	35.0
26	38.4	110	92	50	25.0	60	28.0	105	100	115	35.0
27	38.4	110	92	117	25.0	61	27.7	120	145	111	35.0
28	38.4	110	92	99	25.0	62	25.6	107	120	116	40.0
29	46.9	90	52	104	27.5	63	25.3	114	140	131	40.0
30	36.3	112	103	107	27.5	64	23.9	114	140	123	40.0
31	36.1	103	84	114	27.5	65	23.6	117	150	121	40.0
32	36.1	103	84	101	27.5	66	23.6	122	165	50	40.0
33	35.4	111	102	97	27.5	67	23.6	122	165	114	40.0
34	35.3	111	102	113	27.5	68	23.6	122	165	127	40.0
35	35.1	102	81	101	27.5	69	23.6	122	165	123	40.0
36	35.1	106	90	98	27.5	70	23.5	148	245	112	40.0
37	35.0	106	90	88	27.5	71	23.4	160	280	50	40.0
38	33.2	109	102	86	30.0	72	23.4	121	162	135	40.0
39	32.9	109	102	86	30.0	73	23.1	121	162	132	40.0
40	32.3	120	130	92	30.0	74	22.9	110	140	160	45.0
41	32.2	106	95	113	30.0	75	22.9	110	140	129	45.0
42	32.2	106	95	106	30.0	76	19.5	121	175	129	45.0
43	32.2	109	102	92	30.0	77	18.1	165	322	50	45.0
44	32.2	106	95	88	30.0	78	17.2	140	238	115	45.0
45	31.5	105	93	102	30.0	79	17.0	147	263	50	45.0
46	31.5	108	100	99	30.0	80	16.7	157	295	119	45.0
47	31.4	108	100	111	30.0	81	13.2	130	236	107	55.0
48	31.4	107	98	103	30.0						

注：VOL=驾驶空间的立方英尺数。
HP=发动机马力。
MPG=每加仑耗油量行驶的英里数。
SP=最高时速，英里/小时。
WT=车身重量，百磅。
观测=汽车观测序号（车名未记）。

资料来源：U. S. Environmental Protection Agency，1991，Report EPA/AA/CTAB/91-02.

a. 考虑如下模型：

$$MPG_i = \beta_1 + \beta_2 SP_i + \beta_3 HP_i + \beta_4 WT_i + u_i$$

估计模型参数并对结论做出解释。这些结论有经济意义吗？

b. 你认为上述模型中的误差方差存在异方差性吗？为什么？

c. 用怀特检验来检查误差方差是否存在异方差性。

d. 求出怀特异方差一致标准误和 t 值，并与从 OLS 得到的结论进行比较。

e. 若证明存在异方差性，你如何对数据变形，使变形后的数据是同方差的？给出必要的计算。

11.16 **印度的食物支出**。在表 2—8 中，我们已经给出印度 55 个家庭的食物支出和总支出数据。

a. 将食物支出对总支出回归，检查回归所得到的残差。

b. 将得到的残差对总支出描点，看是否存在系统关系。

c. 若（b）中描点图显示存在异方差性，用帕克、格莱泽和怀特检验分析这些检验是否支持从图中观察所得到的异方差性印象。

d. 求出怀特异方差一致标准误，并与 OLS 标准误进行比较。判断此例中是否值得校正异方差性。

11.17 重做习题 11.16，只是把食物支出的对数对总支出的对数做回归。如果你在习题 11.16 的线性模型中得到异方差性，但在对数线性模型中没有发现，你会得出什么结论？给出所有必要的计算。

11.18 **怀特检验的简易途径**。正文中提到，如果有 n 个回归元，而且我们引入所有回归元及其平方项和交叉乘积项，怀特检验会消耗太多的自由度。因此，不估计像方程（11.5.22）那样的回归，而简单地做如下回归：

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 \hat{Y}_i + \alpha_3 \hat{Y}_i^2 + v_i$$

其中 $\hat{Y}_i$ 为你从模型中估计出来的 Y 值（回归子的估计值）。毕竟，$\hat{Y}_i$ 无非就是回归元的加权平均，只是以估计的回归系数为权重。

从上述回归中求出 R^2 值，并用方程（11.5.22）检验不存在异方差性的假设。

在习题 11.16 中，对食物支出的例子应用上述检验。

11.19 回到 11.7 节和习题 11.10 中所讨论的 R&D 一例。以利润为回归元重做一遍。据经验，预期你的结果会与以销售额为回归元的结果不同吗？为什么？

11.20 表 11—8 给出了美国研究型大学中统计学正教授在 2007 年薪水的中位数数据。

表 11—8　　2007 年统计学正教授薪水的中位数

任职年限	人数	薪水中位数（美元）
0～1	40	101 478
2～3	24	102 400
4～5	35	124 578
6～7	34	122 850
8～9	33	116 900
10～14	73	119 465
15～19	69	114 900
20～24	54	129 072
25～30	44	131 704
31 年及以上	25	143 000

资料来源：American Statistical Association，“2007 Salary Report.”

a. 将薪水中位数对担任正教授年限（作为工作经验的一种度量）描点。为便于描点，假定薪水中位数对应于任职年限的中点。于是任职 4～5 年的薪水 124 578 美元对应于任职 4.5 年，以此类推。对最后一组，假定其范围是 31～33。

b. 考虑如下回归模型

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i \tag{1}$$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + v_i \tag{2}$$

其中 Y=薪水中位数，X=任职年数（以任职年数的中点数据度量），u 和 v 为误差项。你能证明模型（2）比模型（1）更合适吗？利用所给数据估计这两个模型。

c. 若在模型（1）中观测到异方差性，而在模型（2）中没有，你会得出什么结论？给出必要的计算。

d. 若在模型（2）中观测到异方差性，你将如何对数据做变换以消除异方差性？

11.21　给出以下数据：

从前 30 个观测值算出的 $RSS_1 = 55$，自由度 df=25。

从后 30 个观测值算出的 $RSS_2 = 140$，自由度 df=25。

完成显著性水平为 5%的戈德菲尔德-匡特异方差性检验。

11.22　表 11—9 给出 20 个国家的股票价格 Y 和消费者价格 X 年百分率变化的一个横截面数据。

表 11—9　　第二次世界大战后（直至 1969 年）股票价格与消费者价格

国家	变化率（年百分率变化）	
	股票价格，Y	消费者价格，X
1. 澳大利亚	5.0	4.3
2. 奥地利	11.1	4.6
3. 比利时	3.2	2.4
4. 加拿大	7.9	2.4
5. 智利	25.5	26.4
6. 丹麦	3.8	4.2
7. 芬兰	11.1	5.5
8. 法国	9.9	4.7
9. 德国	13.3	2.2
10. 印度	1.5	4.0
11. 爱尔兰	6.4	4.0
12. 以色列	8.9	8.4
13. 意大利	8.1	3.3
14. 日本	13.5	4.7
15. 墨西哥	4.7	5.2
16. 荷兰	7.5	3.6
17. 新西兰	4.7	3.6
18. 瑞典	8.0	4.0
19. 英国	7.5	3.9
20. 美国	9.0	2.1

资料来源：Phillip Cagan，*Common Stock Values and Inflation：The Historical Record of Many Countries*，National Bureau of Economic Research，Suppl.，March 1974，Table 1，p. 4.

a. 将数据描在散点图上。

b. 将 Y 对 X 回归并分析回归中的残差。你观察到什么？

c. 因智利的数据看来有些异常（异常值?），去掉智利数据后，重作（b）中的回归。分析从此回归得到的残差，你会看到什么？

d. 如果根据（b）的结果你将得到有异方差性的结论，而根据（c）的结果你又得到相反的结论。那么，你能得出什么一般性的结论呢？

11.23 本书网站上的表 11—10 给出了财富 500 强公司的 447 位高管的薪水及相关数据。数据包括 salary=1999 年薪水及红利；totcomp=1999 年 CEO 总报酬；tenure=担任 CEO 年数（不足 6 个月视为 0）；age=CEO 的年龄；sales=1998 年企业的总销售收入；profits=1998 年企业利润；以及 assets=1998 年企业的总资产。

a. 利用这些数据估计如下回归并求布罗施-帕甘-戈弗雷统计量以查验异方差性：

$$\text{salary}_i = \beta_1 + \beta_2\ \text{tenure}_i + \beta_3\ \text{age}_i + \beta_4\ \text{sales}_i + \beta_5\ \text{profits}_i + \beta_6\ \text{assets}_i + u_i$$

看上去有异方差性的问题吗？

b. 现在做以 ln(salary) 为因变量的第二个模型。异方差性有所改善吗？

c. 做薪水与每个自变量的散点图。你能辨别是哪些变量带来的问题吗？为了解决这个问题，你有何建议？你最终会使用哪个模型？

附录 11A

□ 11A.1 方程（11.2.2）的证明

由附录 3A 第 3A.3 节 我们有：

$$\begin{aligned}\text{var}\,(\hat{\beta}_2) &= E(k_1^2u_1^2 + k_2^2u_2^2 + \cdots + k_n^2u_n^2 + 2\text{ 个交叉乘积项}) \\ &= E(k_1^2u_1^2 + k_2^2u_2^2 + \cdots + k_n^2u_n^2)\end{aligned}$$

这是因为由于假定了无序列相关，所以交叉乘积项的期望值一律为零。又由于 k_i 是已知的（为什么?）且 $E(u_i^2) = \sigma_i^2$，所以

$$\text{var}\,(\hat{\beta}_2) = k_1^2E(u_1^2) + k_2^2E(u_2^2) + \cdots + k_n^2E(u_n^2)$$

也即：

$$\text{var}\,(\hat{\beta}_2) = k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + \cdots + k_n^2\sigma_n^2$$

或者：

$$\begin{aligned}\text{var}\,(\hat{\beta}_2) &= \sum k_i^2\sigma_i^2 \\ &= \sum\left[\left(\frac{x_i}{\sum x_i^2}\right)^2\sigma_i^2\right] \qquad \text{因为 } k_i = \frac{x_i}{\sum x_i^2} \qquad (11.2.2) \\ &= \frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}\end{aligned}$$

□ 11A.2 加权最小二乘法

为说明该方法，我们利用双变量模型 $Y_i = \beta_1 + \beta_2X_i + u_i$。不加权的最小二乘法要求最小化：

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \tag{1}$$

而加权最小二乘法要求最小化加权残差平方和：

$$\sum w_i \hat{u}_i^2 = \sum w_i (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)^2 \tag{2}$$

其中 $\hat{\beta}_1^*$ 和 $\hat{\beta}_2^*$ 为加权最小二乘估计量，并且权数 w_i 与给定 X_i 下 u_i 或 Y_i 的条件方差成反比：

$$w_i = \frac{1}{\sigma_i^2} \tag{3}$$

因此不言而喻，$\text{var}(u_i \mid X_i) = \text{var}(Y_i \mid X_i) = \sigma_i^2$ 。

将方程（2）对 $\hat{\beta}_1^*$ 和 $\hat{\beta}_2^*$ 微分得：

$$\frac{\partial \sum w_i \hat{u}_i^2}{\partial \beta_1^*} = 2\sum w_i (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)(-1)$$

$$\frac{\partial \sum w_i \hat{u}_i^2}{\partial \beta_2^*} = 2\sum w_i (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)(-X_i)$$

令上式为零，即得到如下两个正规方程：

$$\sum w_i Y_i = \hat{\beta}_1^* \sum w_i + \hat{\beta}_2^* \sum w_i X_i \tag{4}$$

$$\sum w_i X_i Y_i = \hat{\beta}_1^* \sum w_i X_i + \hat{\beta}_2^* \sum w_i X_i^2 \tag{5}$$

注意这两个正规方程和不加权的最小二乘的正规方程的相似性。

解方程（4）和（5）的联立方程得：

$$\hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}^* \tag{6}$$

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \quad (11.3.8) = (7)$$

欲求方程（11.3.9）中给出的 $\hat{\beta}_2^*$ 的方差，可仿照附录 3A 第 3A.3 节求 $\hat{\beta}_2$ 的方差的方法。

注：$\bar{Y}^* = \sum w_i Y_i / \sum w_i$ 和 $\bar{X}^* = \sum w_i X_i / \sum w_i$ 。容易验证，当对所有 i，$w_i = w$ 为一常数时，这些加权均值便化为平常的或不加权的均值 $\bar{Y}$ 和 $\bar{X}$ 。

□ 11A.3 出现异方差时 $E(\hat{\sigma}^2) \neq \sigma^2$ 的证明

考虑两变量模型

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

其中 $\text{var}(u_i) = \sigma_i^2$。

现在，

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum [\beta_1 + \beta_2 X_i + u_i - \hat{\beta}_1 - \hat{\beta}_2 X_i]^2}{n-2}$$

$$= \frac{\sum [-(\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2) X_i + u_i]^2}{n-2} \tag{2}$$

注意 $(\hat{\beta}_1 - \beta_1) = -(\hat{\beta}_2 - \beta_2)\bar{X} + \bar{u}$，以之代入方程（2）并对两边求期望便得到

$$E(\hat{\sigma}^2) = \frac{1}{n-2} \left\{ -\sum x_i^2 \text{var}(\hat{\beta}_2) + E\left[\sum (u_i - \bar{u})^2\right] \right\}$$

$$= \frac{1}{n-2} \left[-\frac{\sum x_i^2 \sigma_i^2}{\sum x_i^2} + \frac{(n-1)\sum \sigma_i^2}{n} \right] \tag{3}$$

其中用到方程（11.2.2）。

如你从方程（3）中所见，若为同方差性，即对每个 i 都有 $\sigma_i^2 = \sigma^2$，则 $E(\hat{\sigma}^2) = \sigma^2$。因此，惯常计算的期望值 $\hat{\sigma}^2 = \sum \hat{u}^2/(n-2)$ 在异方差情况下不再等于真实的 σ^2。[①]

□ 11A.4 怀特稳健标准误

为了对怀特异方差校正标准误有所了解，考虑两变量回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{var}(u_i) = \sigma_i^2 \tag{1}$$

如方程（11.2.2）所示，

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2} \tag{2}$$

由于 σ_i^2 不能直接观测，所以怀特建议用每个 i 的残差平方 $\hat{u}_i^2$ 取代 σ_i^2，并估计 var $(\hat{\beta}_2)$ 如下：

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \hat{u}_i^2}{(\sum x_i^2)^2} \tag{3}$$

怀特证明，方程（3）是方程（2）的一致估计量，即随着样本容量无限增加，方程（3）收敛于方程（2）。[②]

顺便提醒注意，你的软件中并不包含怀特稳健标准误程序，你可以首先做通常的 OLS 回归，从中得到残差，然后利用（3）式即可。

怀特程序可推广至 k 变量回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{4}$$

任何一个偏回归系数（比方说 $\hat{\beta}_j$）的方差都可如下求得：

$$\text{var}(\hat{\beta}_j) = \frac{\sum \hat{w}_{ji}^2 \hat{u}_i^2}{(\sum \hat{w}_{ji}^2)^2} \tag{5}$$

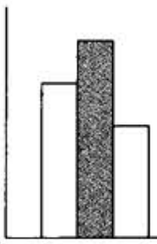
其中 $\hat{u}_i$ 为从原回归（4）中得到的残差，$\hat{w}_j$ 为将回归元 X_j 对方程（4）中其余回归元做（辅助）回归得到的残差。

显而易见，这是一个费时的程序，因为你必须对每个 X 变量都估计方程（5）。当然，你如果有一个统计软件进行例行计算，则所有这些劳动都可以避免：诸如 PC-GIVE、EViews、MICROFIT、SHAZAM、STATA 和 LIMDEP 等现在都能十分轻松地求出怀特的异方差稳健标准误。

① 详细情况可参阅 Jan Kmenta, *Elements of Econometrics*, 2d. ed., Macmillan, New York, 1986, pp. 276-278.

② 更准确地讲，n 乘以方程（3）依概率收敛于 $E[(X_i - \mu_X)^2 u_i^2]/(\sigma_X^2)^2$，即 n 乘以方程（2）的概率极限，其中 n 为样本容量，μ_X 为 X 的期望值，σ_X^2 为 X 的（总体）方差。更详细的情形，参见 Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, South-Western Publishing, 2000, p. 250.

第 12 章



自相关：误差项相关会怎么样？

读者或许记得，经验分析通常有三种数据可用：（1）横截面数据；（2）时间序列数据；（3）横截面数据与时间序列数据的综合，即混合数据。我们在第 1 篇提出经典线性回归模型（CLRM）时做了几个假定，7.1 节有所讨论。不过，我们注意到，不是所有这些假定对每种数据类型都成立。事实上，我们在上一章看到，同方差或误差方差相等的假定在横截面数据中并非总是合理的假定。换言之，横截面数据时常受到异方差问题的纠缠。

不过，由于在横截面研究中，数据的搜集通常都基于横截单位的一个随机样本，比如（消费函数分析中的）家庭或（投资研究分析中的）企业，所以没有先验理由认为，一个家庭或企业的误差项与另一个家庭或企业的误差项相关。如果碰巧在横截单位中观察到了这种相关，则称之为**空间自相关**（spatial autocorrelation），即空间而非时间上的相关。然而，重要的是要记得，在横截面分析中，为了使判断（空间）自相关是否存在的道理讲得过去，数据的顺序必须存在某种逻辑或经济上的意义。

然而，当我们处理时间序列数据时，情况就极为不同了。因为这种数据中的观测服从时间上的一种自然顺序，所以连续的观测很可能表现出内部的相关，特别是在连续观测的时间区间很短时，如一天、一周、一个月而非一年。如果你每天都观测道琼斯或标准普尔 500 等股票价格指数，便发现这些指数接连几天上涨或下降再平常不过了。显然，在这种情形下，CLRM 背后的误差项**不存在自相关或序列相关**的假定就不合常理。

本章中我们将仔细分析这一假定，以期得到对下列问题的回答：

1. 自相关的性质是什么？
2. 自相关会导致什么理论上和实际上的后果？

3. 由于非自相关假定涉及不可观测的干扰项 u_t，怎样才能知道在一个任给的情况中是否有自相关？注意，我们现在用下标 t 来强调我们处理的是时间序列数据。

4. 怎样补救自相关的问题？

读者将发现本章在许多方面类似于上一章对异方差性的讨论。因为**在自相关情况下和在异方差情况下一样，平常的 OLS 估计量虽然仍是线性、无偏和渐近（即在大样本中）正态分布的[①]，但不再是所有线性无偏估计量中方差最小的一个。简言之，相对其他线性无偏估计量而言，它不再是有效的。换言之，OLS 估计量不再是 BLUE。结果，通常的 t，F 和 χ^2 都不再成立。**

12.1 问题的性质

自相关（autocorrelation）一词可定义为“按时间（如在时间序列数据中）或空间（如在横截面数据中）排序的观测序列各成员之间的相关”[②]。在回归的背景中，经典线性回归模型假定各干扰项 u_i 之间不存在自相关，用符号表示：

$$\text{cov}(u_i, u_j \mid x_i, x_j) = E(u_i u_j) = 0 \qquad i \neq j \tag{3.2.5}$$

简单地说，经典模型假定，任一次观测的干扰项都不受任何其他观测干扰项的影响。例如，在做产出对劳动和资本投入的回归中，我们用了季度时间序列数据。如果某一季度的产出受到罢工的影响，没有理由认为这一生产中断会持续到下一季度，也就是说，即使本季度产出下降，却没有理由预期下一季度的产出也下降。同理，如果我们在家庭消费支出对家庭收入的回归中用了横截面数据，那么，一个家庭的收入增加对其消费支出的影响，预期不会波及另一个家庭的消费支出。

然而，如果存在这种相关性，我们就有了自相关。用符号表示：

$$E(u_i u_j) \neq 0 \qquad i \neq j \tag{12.1.1}$$

这时，由本季度罢工引起的生产中断很可能影响下个季度的产出；或者，一个家庭的消费支出增加，很可能促使另一个家庭为了同邻居攀比也随之增大其消费支出。

在我们寻求自相关存在的原因前，有必要澄清一些名词术语方面的问题。虽然现在把名词**自相关**和**序列相关**（serial correlation）看作同义语已成为习惯，但一些作者比较喜欢把两者区分开来。例如，廷特纳（Tintner）定义自相关为“一给定序列同它自身滞后若干期的序列的滞后相关”。而与此同时，他把序列相关一词保留作“两个不同序列的滞后相关”[③]。例如，时间序列 u_1，u_2，…，u_{10} 和它自身滞后一期

① 对此，参见 William H. Greene, *Econometric Analysis*, 4th ed., Prentice Hall, NJ, 2000, Chapter 11, 和 Paul A. Rudd, *An Introduction to Classical Econometric Theory*, Oxford University Press, 2000, Chapter 19。

② Maurice G. Kendall and William R. Buckland, *A Dictionary of Statistical Terms*, Hafner Publishing Company, New York, 1971, p. 8.

③ Gerhard Tintner, *Econometrics*, John Wiley & Sons, New York, 1965.

的序列 u_2，u_3，…，u_{11}之间的相关叫做自相关，而两个不同时间序列 u_1，u_2，…，u_{10}和 v_2，v_3，…，v_{11}（其中 u 和 v 是两个不同的时间序列）之间的相关叫做序列相关。尽管两名词的区分有一定用处，本书中将把它们看作同义语。

让我们想象自相关和无自相关的一些可能模式，如图 12—1 所示。图 12—1a 到图 12—1d 都显示 u_i 中有一个明显的模式。图 12—1a 显示出一个周期模式；图 12—1b 和图 12—1c 分别表明干扰项中有一个上升或下降的线性趋势；而图 12—1d 表明干扰项中既有线性趋势，又有二次趋势。唯有图 12—1e 表示无系统性模式，并符合经典线性回归模型的无相关假定。

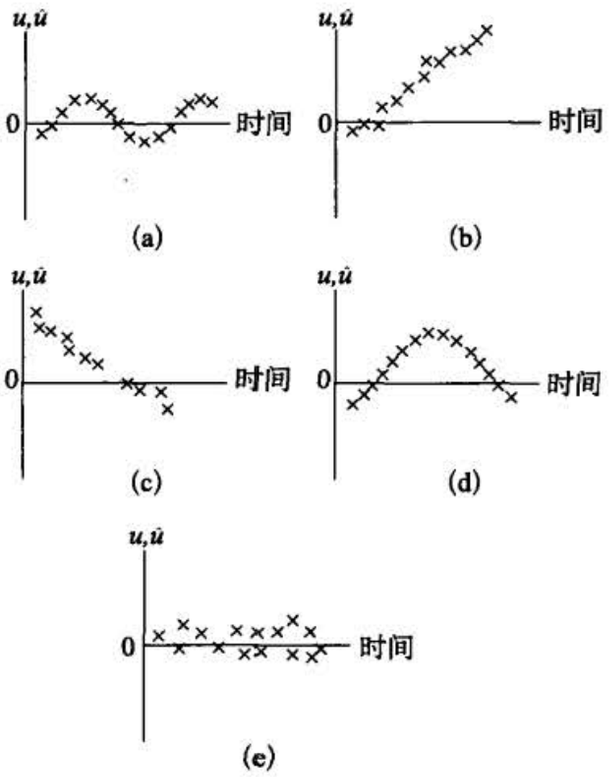


图 12—1　自相关模式与非自相关模式

一个很自然的问题是：为什么会出现序列相关？有种种理由，其中包括：

惯性。大多数经济时间序列都有一个明显的特点，就是它的惯性（inertia）或黏滞。众所周知，GNP、价格指数、生产、就业和失业等时间序列都呈现出一定的（商业）周期。从衰退的谷底开始，当经济开始复苏时，大多数经济序列开始上升。在此上升期间，序列在每一时刻的值都高于前一时刻的值。看来有一种“内在的动力”驱使这一势头继续下去，直至某些情况（如利率或课税或两者同时升高）出现才把它拖慢下来。因此，在涉及时间序列的回归中，相继的观测值很可能是相关的。

设定偏误：应含而未含变量的情形。在经验分析中，研究者常从一个较好但不一定“最好”的回归模型开始。经过回归分析，研究者做事后检查，看结果是否与事先的预期相一致。如不一致，便要开始动手术。例如，研究者可能将拟合回归的残差 $\hat{u}_i$ 描图，并可能观察到类似于图 12—1a 到图 12—1d 那样的模式。这些残差

（为 u_i 的替代变量）也许表明，某些原来待选的、却未被包含的变量，由于种种理由应被包含到模型中来。这就是**应含而未含变量**（excluded variable）的设定偏误。但将这些变量包含进来常常能消除干扰项中所观察到的相关模式。例如，假如我们有如下需求模型：

$$Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+\beta_4X_{4t}+u_t \tag{12.1.2}$$

其中 Y = 牛肉需求量，X_2 = 牛肉价格，X_3 = 消费者收入，X_4 = 猪肉价格，以及 t = 时间。[①] 然而，出于某种原因，我们做了下述回归：

$$Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+v_t \tag{12.1.3}$$

现在，如果方程（12.1.2）是“正确”的模型或“真理”或真实的关系式，则做回归（12.1.3）就无异于令 $v_t=\beta_4X_{4t}+u_t$。于是就猪肉价格影响牛肉消费而言，误差项或干扰项 v 将表现出一种系统模式，从而造成（错误的）自相关。真的有自相关吗？一个简单的检验是，同时做回归（12.1.2）和（12.1.3），看看在做模型（12.1.2）时，原先在模型（12.1.3）中观测到的自相关（如果观测有的话）是否消失。[②]侦察自相关的实际步骤将在 12.6 节中讨论。届时我们将看到，用回归（12.1.2）和（12.1.3）的残差来描图，常常能够反映出自相关的问题。

设定偏误：不正确的函数形式。假使在成本—产出研究中“真实”或“正确”的模型如下：

$$\text{marginal cost}_i=\beta_1+\beta_2\text{output}_i+\beta_3\text{output}_i^2+u_i \tag{12.1.4}$$

其中 marginal cost＝边际成本，output＝产出。但我们拟合了如下模型：

$$\text{marginal cost}_i=\alpha_1+\alpha_2\text{output}_i+v_i \tag{12.1.5}$$

这样，图 12—2 连同“不正确”的线性成本曲线一起，展示了“真实”模型的边际成本曲线。

如图 12—2 所示，在点 A 与点 B 之间的线性边际成本曲线一直在高估真实的边际成本，而在这两点之外则一致地低估真实的边际成本。这种结果是可以预料到的，因为干扰项 v_i 事实上等于产出 $\text{output}_i^2+u_i$，从而包含了 output_i^2 对边际成本的系统影响。在这种情形中，由于错误函数形式的使用，v_i 将要反映出自相关。在第 13 章中，我们将考虑侦察设定偏误的若干方法。

蛛网现象。许多农产品的供给反映出一种所谓的蛛网现象（cobweb phenomenon）。供给对价格的反应要滞后一个时期，因为供给需要经过一定的时间才能实现（有一孕育期）。例如，今年年初的作物种植受去年盛行价格的影响。因此，相关的供给函数是：

$$\text{supply}_t=\beta_1+\beta_2P_{t-1}+u_t \tag{12.1.6}$$

其中 supply＝供给。假使 t 时期的期末价格 P_t 低于 P_{t-1}，农民就很可能决定在 $t+1$ 时期生产比 t 时期更少的东西。显然，在这种情形中，农民由于在 t 时期的过量生产

① 作为一种惯例，我们将用下标 t 表示时间序列数据，而用通常的下标 i 表示横截面数据。

② 如果我们发现真正的问题是设定偏误而不是自相关，则如第 13 章所讲，方程（12.1.3）中参数的 OLS 估计量可能既有偏误，又非一致。

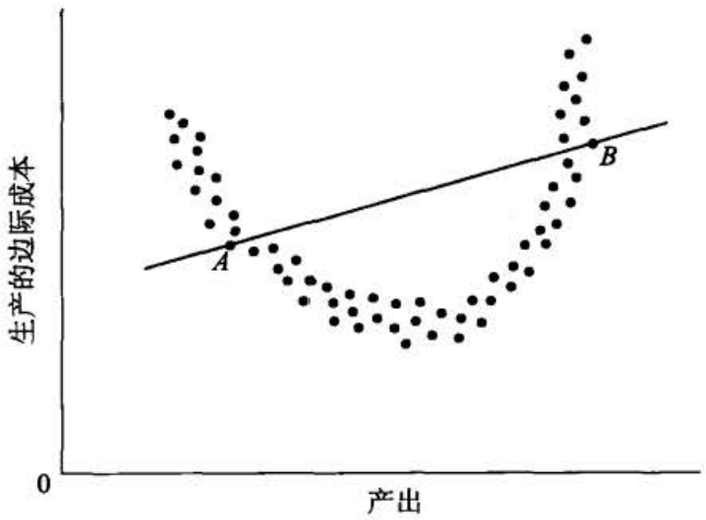


图 12—2　设定偏误：不正确的函数形式

很可能在 $t+1$ 时期削减产量。诸如此类的现象，就不能期望干扰项 u_t 是随机的，从而导致一种蛛网模式。

滞后效应。在一个消费支出对收入的时间序列回归中，人们常常发现当前时期的消费支出除了依赖于其他变量外，还依赖于上一期的消费支出，即：

$$\text{consumption}_t = \beta_1 + \beta_2 \text{income}_t + \beta_3 \text{consumption}_{t-1} + u_t \tag{12.1.7}$$

其中 consumption＝消费支出，income＝收入。像方程（12.1.7）这样的回归，由于解释变量之一是因变量的滞后值而被称为**自回归**。（我们将在第 17 章中研究这类模型。）使用方程（12.1.7）这类模型的理由很简单。由于心理上、技术上以及制度上的原因，消费者不会轻易改变他们的消费习惯。如果现在我们忽略了方程（12.1.7）中的滞后项，所造成的误差项将由于滞后消费对当前消费的影响而反映出一种系统性模式。

数据的“操作”。在经验分析中，原始数据往往是经过“编造的”。例如，在用到季度数据的时间序列回归中，这些数据通常来自月度数据，不过是把 3 个月的观测值加在一起除以 3 罢了。这种平均的计算因减弱了月度数据的波动而导致数据更加匀滑。因此，用季度数据描绘的图形要比用月度数据描绘的图形看来匀滑得多。这种匀滑性本身就能使干扰项中出现系统性样式，从而导致自相关。数据变换的另一来源是数据的**内插**（interpolation）或**外推**（extrapolation）。例如在美国每 10 年进行一次人口普查。最近的一次在 2000 年，而此前的一次在 1990 年。假如现在需要 1990—2000 年两个普查年间的某年数据，通常的做法就是，根据某些特殊的假定进行内插。所有这些数据“糅合”技术都会给数据带来原始数据所没有的系统性样式。[①]

数据变换。作为一个例子，考虑如下模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.1.8}$$

① 关于这一点，参看 William H. Greene，op. cit.，p. 526。

其中 Y=消费支出，X=收入。由于方程（12.1.8）在每个时期都成立，所以它在上一时期（$t-1$）也成立。于是，可以把方程（12.1.8）写成

$$Y_{t-1}=\beta_1+\beta_2 X_{t-1}+u_{t-1} \tag{12.1.9}$$

Y_{t-1}、X_{t-1}和 u_{t-1}分别被称为 Y、X 和 u 的**滞后值**（lagged values），这里滞后一期。我们在本章稍后部分及本书其他几个地方将会看到滞后值的重要性。

现在，如果我们从方程（12.1.8）中减去方程（12.1.9），则得到

$$\Delta Y_t=\beta_2\Delta X_t+\Delta u_t \tag{12.1.10}$$

其中 Δ 表示**一阶差分算子**（first difference operator），表示对所讨论变量连续取差分。因而，$\Delta Y_t=(Y_t-Y_{t-1})$，$\Delta X_t=(X_t-X_{t-1})$，$\Delta u_t=(u_t-u_{t-1})$。为便于实证分析，我们把方程（12.1.10）写成

$$\Delta Y_t=\beta_2\Delta X_t+v_t \tag{12.1.11}$$

其中 $v_t=\Delta u_t=(u_t-u_{t-1})$。

方程（12.1.9）被称为**水平值形式**（level form），方程（12.1.10）被称为**一阶差分形式**（first difference form）。经验分析中，这两种形式都经常使用。比如在方程（12.1.9）中，若 Y 和 X 表示消费支出和收入的对数，则方程（12.1.10）中的 ΔY 和 ΔX 将表示消费支出和收入的对数变化。但如我们所知，一个变量的对数变化是其相对变化或百分比变化（如果将前者乘以 100）。所以，除了研究变量水平值之间的关系外，我们或许还对其增长率之间的关系感兴趣。

现在，如果方程（12.1.8）中的误差项满足标准 OLS 假定，特别是无自相关的假定，那么，可以证明方程（12.1.11）中的误差项 v_t 将会自相关。（附录 12A 第 12A.1 节给出了证明。）这里请注意，像方程（12.1.11）这样的模型被称为**动态回归模型**（dynamic regression models），即含有滞后回归子的模型。我们将在第 17 章深入研究这种模型。

上述例子旨在说明，自相关有时候是作为对原模型进行变换的结果而产生。

非平稳性。我们在第 1 章提到，在处理时间序列数据时，我们必须明确一个给定的时间序列是否平稳。尽管我们在本书第 5 篇全面讨论时间序列计量经济学的章节中将会专题讨论非平稳时间序列，但这里粗略地讲，如果一个时间序列的特征（如均值、方差和协方差）不随时间而变化（time invariant），那它就是一个平稳的时间序列。否则，就是一个非平稳时间序列。

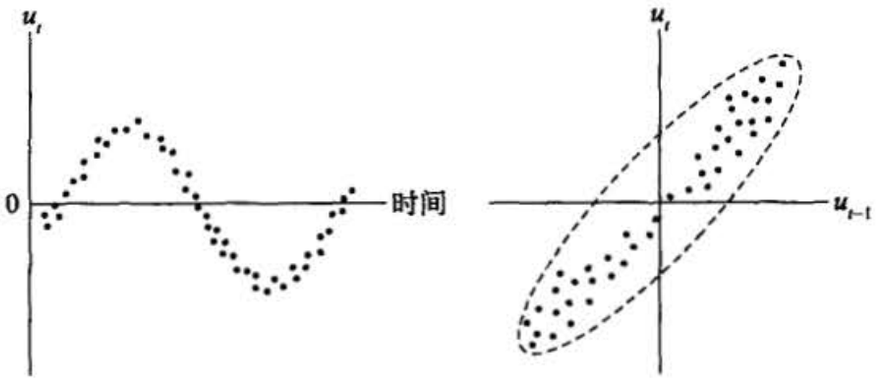
如我们将在第 5 篇所讨论的那样，在一个诸如方程（12.1.8）之类的回归模型中，Y 和 X 很可能都是非平稳的，因此误差 u 也是非平稳的。[①] 此时，误差项将表现出自相关。

总之，有很多原因导致一个回归模型中的误差项自相关。在本章的余下部分，我们较详细地研究自相关所带来的问题及其解决办法。

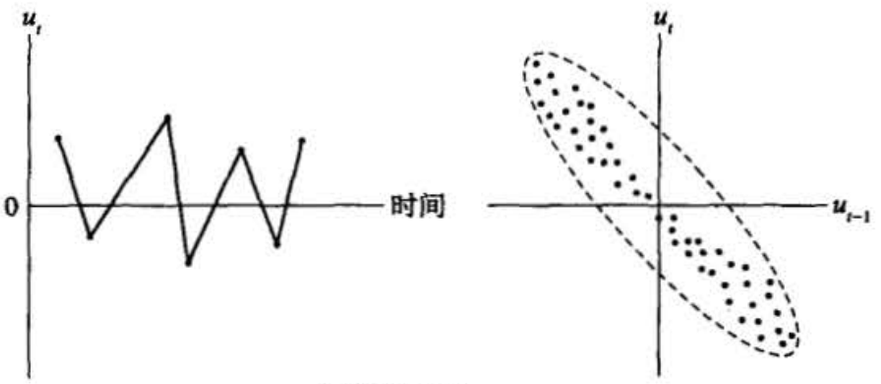
还应指出，虽然多数经济时间序列都因在一个较长时期内或者上升或者下降而

① 我们在第 5 篇也将看到，尽管 Y 和 X 非平稳，可能发现 u 是平稳的。我们以后会解释其含义。

表现出正的自相关（图 12—3a），而不像图 12—3b 那样表现为不断地上下运动，但自相关可以是正的，也可以是负的。



(a)正自相关



(b)负自相关

图 12—3　正的和负的自相关

12.2　出现自相关时的 OLS 估计量

如果我们在干扰项中通过假定 $E(u_t u_{t+s}) \neq 0 (s \neq 0)$ 引进自相关，但保留经典模型的所有其他假定，对 OLS 估计量及其方差来说，会出现什么情况呢？[①] 再次注意到，我们在干扰项右下角用下标 t 来强调我们处理的是时间序列数据。

我们再一次回到双变量模型，以解释所涉及的基本概念。即回到 $Y_t = \beta_1 + \beta_2 X_t + u_t$。要取得任何进展都必须对 u_t 的生成机制做出假定，因为 $E(u_t u_{t+s}) \neq 0 (s \neq 0)$ 这个假定过于一般，不会有任何实际用处。作为一个起点或一阶近似，不妨假定干扰项是这样产生的：

① 若 $s=0$，则得到 $E(u_t^2)$。由于根据假定有 $E(u_t)=0$，所以 $E(u_t^2)$ 表示误差项的方差，显然非零。（为什么？）

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{12.2.1}$$

其中 ρ 被称为**自协方差系数**（coefficient of autocovariance），并且 ε_t 是满足以下标准 OLS 假定的随机干扰项：

$$\begin{aligned} &E(\varepsilon_t) = 0 \\ &\text{var}(\varepsilon_t) = \sigma_\varepsilon^2 \\ &\text{cov}(\varepsilon_t, \varepsilon_{t+s}) = 0 \qquad s \neq 0 \end{aligned} \tag{12.2.2}$$

在工程文献中，具有上述性质的误差项通常被称为**白噪音误差项**（white noise error term）。方程（12.2.1）阐明，t 时期干扰项的值等于 ρ 乘以其上一期干扰值与一个纯粹随机误差项之和。

模式（12.2.1）被称为**马尔可夫一阶自回归模式**（Markov first-order autoregressive scheme）或简称一阶自回归模式，常记为 **AR(1)**。由于方程（12.2.1）可解释为 u_t 对其自身滞后一期的回归，取名为自回归是适宜的。因仅涉及 u_t 及其最近的过去值，即最大滞后是 1，故称一阶。如果模型设为 $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$，它将是 AR(2) 或二阶自回归模式，如此类推。我们在第 5 篇有关时间序列计量经济学的章节中将考查这种高阶模式。

顺便指出，方程（12.2.1）中的自协方差系数又可解释为**一阶自相关系数**（first-order coefficient of autocorrelation），或更准确地称为**滞后 1 期的自相关系数**（coefficient of autocorrelation at lag 1）。①

给定 AR(1) 模式，可以证明（参见附录 12A 第 12A.2 节）：

$$\text{var}(u_t) = E(u_t^2) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{12.2.3}$$

$$\text{cov}(u_t, u_{t+s}) = E(u_t u_{t-s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{12.2.4}$$

$$\text{cor}(u_t, u_{t+s}) = \rho^s \tag{12.2.5}$$

其中 $\text{cov}(u_t, u_{t+s})$ 指相差 s 期的误差项之间的协方差，$\text{cor}(u_t, u_{t+s})$ 指相差 s 期的误差项之间的相关系数。注意，由于协方差和相关系数的对称性，$\text{cov}(u_t, u_{t+s}) = \text{cov}(u_t, u_{t-s})$，而 $\text{cor}(u_t, u_{t+s}) = \text{cor}(u_t, u_{t-s})$。

由于 ρ 是一个介于 -1 与 1 之间的常数，所以方程（12.2.3）表明，在 AR(1) 模式下，u_t 的方差仍是同方差的，但 u_t 不仅与其过去一期的值相关，而且与过去几期的值也相关。关键是注意到，$|\rho| < 1$，即 ρ 的绝对值小于 1。比如，若 $\rho = 1$，上述方差和协方差都没有定义。若 $|\rho| < 1$，我们说方程（12.2.1）中给出的 AR(1) 过程是平稳的；即 u_t 的均值、方差和协方差都不随时间而变化。若 $|\rho| < 1$，则从方程（12.2.4）可见，协方差的值将随着两个误差的时间间隔越远而越小。我们以

① 这个名称是不难解释的。按定义，u_t 与 u_{t-1} 之间的（总体）相关系数是：

$$\rho = \frac{E\{[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]\}}{\sqrt{\text{var}(u_t)}\sqrt{\text{var}(u_{t-1})}} = \frac{E(u_t u_{t-1})}{\text{var}(u_{t-1})}$$

这是因为对每个 t 都有 $E(u_t) = 0$，并且我们保留了同方差性假定而有 $\text{var}(u_t) = \text{var}(u_{t-1})$。读者应能看出，$\rho$ 还是 u_t 对 u_{t-1} 回归中的斜率系数。

后会看到上述结论的作用。

我们使用 AR(1) 过程的原因之一，不仅因为其相对高阶模式较为简单，还因为它在许多应用中都相当有用。此外，对 AR(1) 模式所做的理论和经验研究数量也相当可观。

现在回到我们的双变量回归模型：$Y_t=\beta_1+\beta_2 X_t+u_t$。我们从第 3 章知道，斜率系数的 OLS 估计量为：

$$\hat{\beta}_2=\frac{\sum x_t y_t}{\sum x_t^2} \tag{12.2.6}$$

其方差为：

$$\text{var}\,(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_t^2} \tag{12.2.7}$$

其中小写字母如通常一样表示对均值的离差。

现在，在 AR(1) 模式下，可以证明此估计量的方差为：

$$\text{var}\,(\hat{\beta}_2)_{\text{AR1}}=\frac{\sigma^2}{\sum x_t^2}\left[1+2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2}+2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2}+\cdots+2\rho^{n-1}\frac{x_1 x_n}{\sum x_t^2}\right] \tag{12.2.8}$$

其中 $\text{var}\,(\hat{\beta}_2)_{\text{AR1}}$ 表示 $\hat{\beta}_2$ 在一阶自回归模式下的方差。

方程 (12.2.8) 与 (12.2.7) 的比较表明，前者等于后者乘上方括号中的项，这一项既取决于 ρ，又取决于回归元 X 在各种滞后值之间的样本自相关系数。[①] 一般而言，我们不能预先知道 $\text{var}\,(\hat{\beta}_2)$ 是小于还是大于 $\text{var}\,(\hat{\beta}_2)_{\text{AR1}}$ [但参见下面的方程 (12.4.1)]。当然，若 $\rho=0$，则这两个公式完全一样。(为什么?) 而且，若回归元的连续值之间的相关系数很小，则斜率估计量常用的 OLS 方差将不会严重偏误。但作为一个一般原则，这两个方差应该不同。

为了解方程 (12.2.7) 和 (12.2.8) 中给出的方差之间的差别，假定回归元 X 也服从含有自相关系数 r 的一阶自回归模式。那么，可以证明，方程 (12.2.8) 简化成：

$$\text{var}\,(\hat{\beta}_2)_{\text{AR1}}=\frac{\sigma^2}{\sum x_t^2}\left(\frac{1+r\rho}{1-r\rho}\right)=\text{var}(\hat{\beta}_2)_{\text{OLS}}\left(\frac{1+r\rho}{1-r\rho}\right) \tag{12.2.9}$$

比如，若 $r=0.6$ 和 $\rho=0.8$，我们利用方程 (12.2.9) 可以验证 $\text{var}\,(\hat{\beta}_2)_{\text{AR1}}=2.846\,1\ \text{var}(\hat{\beta}_2)_{\text{OLS}}$。换言之，$\text{var}\,(\hat{\beta}_2)_{\text{OLS}}=\frac{1}{2.846\,1}\text{var}\,(\hat{\beta}_2)_{\text{AR1}}=0.351\,3\ \text{var}\,(\hat{\beta}_2)_{\text{AR1}}$。即通常的 OLS 公式 [即方程 (12.2.7)] 将 $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ 低估了约 65%。你或许意识到，这个答案是给定 r 和 ρ 的值的情况下的特定结果，但这个练习的目的是给你一

① 注意，$r=\sum x_t x_{t+1}/\sum x_t^2$ 为 X_t 与 X_{t+1}(或 X_{t-1}，因为相关系数是对称的) 之间的相关系数；$r^2=\sum x_t x_{t+2}/\sum x_t^2$ 为 X 滞后两期之间的相关系数，等等。

个警告，盲目地使用通常的OLS公式来计算OLS估计量的方差和标准误，可能会给出严重误导性的结论。

假使我们继续使用OLS估计量 $\hat{\beta}_2$，并在平常的方差公式中把AR(1)模式考虑进来而将方差加以适当调整。就是说，我们使用方程（12.2.6）中的 $\hat{\beta}_2$，但使用方程（12.2.8）所给的方差公式。这时 $\hat{\beta}_2$ 的性质如何？容易证明，$\hat{\beta}_2$ 仍然是线性无偏的。事实上，如附录3A第3A.2节所表明，无序列相关性和无异方差性一样，都不是证明 $\hat{\beta}_2$ 的无偏性所必需的。$\hat{\beta}_2$ 是否仍然是BLUE呢？可惜不是；它在线性无偏估计量中不是最小方差的。总之，$\hat{\beta}_2$ 虽然线性无偏，但不再是有效的（当然，相对而言）。读者必定注意到这一发现非常类似于在出现异方差性时 $\hat{\beta}_2$ 较为低效的情形。我们已经看到，在异方差情形中，作为广义最小二乘（GLS）估计量的特殊情形，由方程（11.3.8）给出的加权最小二乘估计量 $\hat{\beta}_2^*$ 是有效的。在自相关的情形中，我们能找到一个BLUE吗？回答是肯定的。这可从下节的讨论中看出。

12.3 自相关出现时的BLUE

继续利用双变量模型并假定AR(1)过程，可以证明 β_2 的BLUE估计量由下式给出[①]：

$$\tilde{\beta}_2^{GLS} = \frac{\sum_{t=2}^{n}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + C \tag{12.3.1}$$

其中 C 是一校正因子，在实际中可以忽略。注意下标从 $t=2$ 变到 $t=n$。从而它的方差是：

$$\text{var}\tilde{\beta}_2^{GLS} = \frac{\sigma^2}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + D \tag{12.3.2}$$

其中 D 也是一个实践中可以忽略的校正因子。（参看习题12.18。）

估计量 $\text{var}\tilde{\beta}_2^{GLS}$ 如其上标所表明的，是由GLS法得到的。在第11章中我们看到，在GLS中我们通过变量变换把额外的信息（异方差性或自相关性）包括到估计程序中去，而在OLS中我们并不直接考虑这种附加信息。读者可以看到，方程（12.3.1）中 β_2 的GLS估计量把自相关参数 ρ 包含在估计公式中，而方程（12.2.6）的OLS公式则对 ρ 干脆不理睬。直观上，这就是为什么GLS估计量是

① 证明见Jan Kmenta，*Elements of Econometrics*，Macmillan，New York，1971，pp. 274-275。校正因子与第一次观测值（Y_1，X_1）有关。关于这个问题，参看习题12.18。

BLUE，而 OLS 估计量不是 BLUE 的理由。——GLS 估计量最大限度地利用了现有的信息。[①] 不言而喻，若 $\rho=0$，就没有额外的信息需要考虑，从而 GLS 和 OLS 两估计量是相同的。

总之，在自相关情形中，由方程（12.3.1）给出的 GLS 估计量才是 BLUE，并且这时的最小方差由方程（12.3.2）给出。方程（12.2.8），尤其是方程（12.2.7），都没有给出最小方差。

一个技术性注解。我们在上一章曾指出，高斯-马尔可夫定理只给出 OLS 为 BLUE 的充分条件。上一章还提到，OLS 为 BLUE 的充要条件由**克鲁斯卡尔定理**（Kruskal's theorem）给出。因此，在有些情况下，尽管存在自相关，OLS 仍可能是 BLUE，只是这种情况在实践中不经常发生而已。

如果我们忽视自相关，一厢情愿地使用 OLS 程序，会出现什么情况？请看下节分析。

12.4 出现自相关时使用 OLS 的后果

如同异方差性情形，在自相关出现时，OLS 估计量仍是线性无偏的和一致性的，但不再是有效的（亦即最小方差）。那么，如果我们继续使用 OLS 估计量，我们平常的假设检验程序会遇到什么问题呢？再次，如同异方差性的情形，我们区分两种情况。为便于教学，我们仍利用双变量模型。虽然如此，把以下的讨论推广到多元回归并无多少困难。[②]

□ 考虑到自相关的 OLS 估计

正如已指出的那样，$\hat{\beta}_2$ 不是 BLUE，即使我们使用 $\text{var}(\hat{\beta}_2)_{\text{ARI}}$，由此得来的置信区间很可能比根据 GLS 程序得到的要宽些。如克曼塔指出的那样，即使样本无限增大，结果也很可能如此。[③] 就是说，$\hat{\beta}_2$ 并非渐近有效的。这一发现对假设检验的含义是明显的：我们很可能宣称一个系数是统计上不显著的（即无异于零），尽管事实上（即根据正确的 GLS 程序）它也许是显著的。这一差别可从图 12—4 明显看出。图中我们做出在真实 $\beta_2=0$ 的假设下 95% 的 OLS［AR(1)］和 GLS 两个置信区间。考虑 β_2 的一个具体的估计值，比如 b_2。由于 b_2 落在 OLS 置信区间内，我们以 95% 的置信度接受真实 $\beta_2=0$ 的假设。但如果我们使用（正确的）GLS 置信区间，由于 b_2 落入拒绝域，我们就会拒绝真实 $\beta_2=0$ 的虚拟假设。

① 关于 $\hat{\beta}_2^{\text{GLS}}$ 是 BLUE 的正式证明，见 Kmenta，ibid。但烦琐的代数证明可利用矩阵符号而大为简化。参看 J. Johnston，*Econometric Methods*，3d ed.，McGraw-Hill，New York，1984，pp. 291-293。

② 但要避免烦琐的代数运算，矩阵代数几乎是必需的。

③ 见 Kmenta，op. cit.，pp. 277-278。

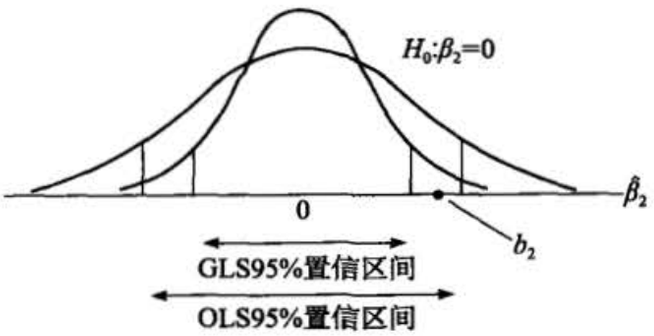


图 12—4　GLS 和 OLS 95%置信区间

一条信息：尽管 OLS 估计量是无偏的和一致性的，但为了构造置信区间并检验假设，要用 GLS 而不用 OLS。（然而，参见 12.11 节。）

□ 忽视自相关的 OLS 估计

如果我们不但使用 $\hat{\beta}_2$ 而且继续使用 $\text{var}(\hat{\beta}_2) = \sigma^2 / \sum x_i^2$，完全不考虑自相关的问题，也就是我们错误地认为关于经典模型的一般假定成立，情况就会变得非常严重。错误将出自多种原因：

1. 残差方差 $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-2)$ 很可能低估了真实的 σ^2。

2. 结果，我们很可能高估了 R^2。

3. 即使没有低估 σ^2，$\text{var}(\hat{\beta}_2)$ 也可能低估了（一阶）自相关情形下的方差 $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ [方程 (12.2.8)]，虽然 $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ 和 $\text{var}(\hat{\beta}_2)^{\text{GLS}}$ 相比是低效的。

4. 因此，通常的 t 和 F 显著性检验都不再可靠。如果仍然使用这些检验，就很可能对所估计的回归系数作出有严重错误的统计显著性结论。

为了证明这些命题，让我们回到双变量模型。由第 3 章知，在经典假定下：

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

给出 σ^2 的一个无偏估计量，即 $E(\hat{\sigma}^2) = \sigma^2$。但若出现 AR(1) 情形的自相关，则可以证明：

$$E(\hat{\sigma}^2) = \frac{\sigma^2\{n - [2/(1-\rho)] - 2\rho r\}}{n-2} \tag{12.4.1}$$

其中 $r = \sum_{t=1}^{n-1} x_t x_{t-1} / \sum_{t=1}^{n} x_t^2$，可解释为 X 的相继观测值之间的（样本）相关系数。[①] 如果 ρ 和 r 都是正的（这对大多数经济时间序列来说都是一个适当的假定），则从方程 (12.4.1) 明显可见 $E(\hat{\sigma}^2) < \sigma^2$；就是说，通常的残差方差公式平均而言低估了真实 σ^2。换言之，$\hat{\sigma}^2$ 偏之于过低。不言而喻，$\hat{\sigma}^2$ 中的这一偏误将被传递到

① 参看 S. M. Goldfield and R. E. Quandt, *Nonlinear Methods in Econometrics*, North Holland Publishing Company, Amsterdam, 1972, p. 183。顺便指出，如果误差项正相关，R^2 趋向于偏误过高，就是说它趋向于比没有自相关时的 R^2 大。

$\text{var}(\hat{\beta}_2)$ 中来，因为实际上后者是由公式 $\hat{\sigma}^2/\sum x_i^2$ 估计的。

但即使 σ^2 未被低估，$\text{var}(\hat{\beta}_2)$ 也是 $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ 的一个偏误估计量。这很容易通过（12.2.7）和（12.2.8）两个不同公式的比较看出。[①] 事实上，如果 ρ 为正（对大多数经济时间序列而言都是对的），并且 X 值是正相关的（大多数经济时间序列也都如此），那么显然有：

$$\text{var}(\hat{\beta}_2) < \text{var}(\hat{\beta}_2)_{\text{AR1}} \tag{12.4.2}$$

就是说，$\hat{\beta}_2$ 的通常的 OLS 方差低估了在 AR(1) 下的 $\hat{\beta}_2$ 的方差［见方程（12.2.9）］。因此，如果我们使用 $\text{var}(\hat{\beta}_2)$，我们就夸大了估计量 $\hat{\beta}_2$ 的精密度（即低估了它的标准误）。结果在 $t = \hat{\beta}_2/\text{se}(\hat{\beta}_2)$ 比率的计算中（在 $\beta_2 = 0$ 的假设下），我们过高地估计了 t 值，从而夸大了 β_2 的估计量的统计显著性。如前所述，如果再加上低估了 σ^2，情况就会变得更糟。

为了看到 σ^2 和 $\hat{\beta}_2$ 的方差通常是如何被 OLS 低估的，让我们做如下**蒙特卡罗实验**（Monte Carlo experiment）。假使在一双变量模型中我们“知道”真实 $\beta_1 = 1$ 和 $\beta_2 = 0.8$，那么这个随机的 PRF 是：

$$Y_t = 1.0 + 0.8X_t + u_t \tag{12.4.3}$$

从而，

$$E(Y_t \mid X_t) = 1.0 + 0.8X_t \tag{12.4.4}$$

给出了真实总体回归线。现假定 u_t 由如下一阶自回归模式生成：

$$u_t = 0.7u_{t-1} + \varepsilon_t \tag{12.4.5}$$

其中 ε_t 满足所有的 OLS 假定。为方便起见，我们进一步假定 ε_t 是均值为 0 方差为 1 的正态变量。方程（12.4.5）假设相继干扰项是正相关的，其自相关系数是一个相当高的值 +0.7。

现在用一张均值为 0 方差为 1 的正态随机数表，按照方程（12.4.5）生成 10 个随机数，如表 12—1 所示。为了启动此实验，还需要给定 u 的一个初始值，比方说，$u_0 = 5$。

表 12—1　　正自相关误差项的一个假设例子

	ε_t	$u_t = 0.7u_{t-1} + \varepsilon_t$
0	0	$u_0 = 5$（假设的）
1	0.464	$u_1 = 0.7 \times 5 + 0.464 = 3.964$
2	2.026 2	$u_2 = 0.7 \times 3.964 + 2.026\ 2 = 4.801\ 0$
3	2.455	$u_3 = 0.7 \times 4.801\ 0 + 2.455 = 5.815\ 7$
4	−0.323	$u_4 = 0.7 \times 5.815\ 7 - 0.323 = 3.748\ 0$
5	−0.068	$u_5 = 0.7 \times 3.748\ 0 - 0.068 = 2.555\ 6$
6	0.296	$u_6 = 0.7 \times 2.555\ 6 + 0.296 = 2.084\ 9$
7	−0.288	$u_7 = 0.7 \times 2.084\ 9 - 0.288 = 1.171\ 4$
8	1.298	$u_8 = 0.7 \times 1.171\ 4 + 1.298 = 2.118\ 0$

① 正式的证明，参看 Kmenta，op. cit.，p. 281。

续前表

	ε_t	$u_t=0.7u_{t-1}+\varepsilon_t$
9	0.241	$u_9=0.7\times2.118\ 0+0.241=1.723\ 6$
10	−0.957	$u_{10}=0.7\times1.723\ 6-0.957=0.249\ 5$

注：ε_t 数据取自 *A Million Random Digits and One Hundred Thousand Deviates*，Rand Corporation，Santa Monica，Calif.，1950。

将表 12—1 中生成的 u_t 描点，我们得到图 12—5。该图表明，开始时每一相继的 u_t 大于它前面的值，后来，一般地说，便小于它前面的值。这种情形通常表明一种正的自相关。

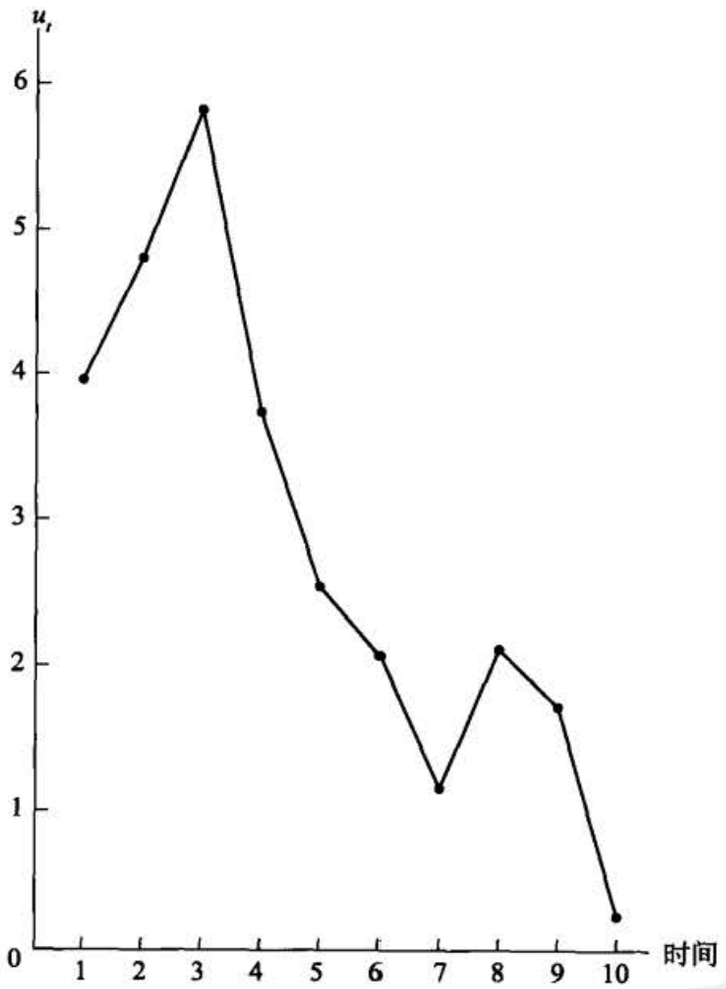


图 12—5　由模式 $u_t=0.7u_{t-1}+\varepsilon_t$ 生成的相关（表 12—1）

现在假如把 X 值固定在 1，2，3，…，10。那么，给定这些 X 值就能按方程 (12.4.3) 生成 10 个 Y 值，连同表 12—1 中所给的 u_t 值，一并放到表 12—2 中，如果我们利用表 12—2 中的数据做 Y 对 X 的回归，就会得到如下（样本）回归：

$$\hat{Y}_t = 6.545\ 2+0.305\ 1\ X_t \tag{12.4.6}$$
$$t = (0.615\ 3)\ (0.099\ 2)$$
$$t = (10.636\ 6)\ (3.076\ 3)$$
$$r^2 = 0.541\ 9 \qquad \hat{\sigma}^2 = 0.811\ 4$$

而真实回归线则由方程（12.4.4）给出。这两条回归线由图 12—6 一并给出。该图清楚地表明，所拟合的回归线在多大程度上歪曲了真实回归线；它严重地低估了真实斜率系数而高估了真实截距。（但应注意，OLS 估计量仍然是无偏的。）

表 12—2　　Y 样本值的生成

X_t	u_t	$Y_t=1.0+0.8X_t+u_t$
1	3.964 0	$Y_1=1.0+0.8\times1+3.964\ 0=5.764\ 0$
2	4.801 0	$Y_2=1.0+0.8\times2+4.801\ 0=7.401\ 0$
3	5.815 7	$Y_3=1.0+0.8\times3+5.815\ 7=9.215\ 7$
4	3.748 0	$Y_4=1.0+0.8\times4+3.748\ 0=7.948\ 0$
5	2.555 6	$Y_5=1.0+0.8\times5+2.555\ 6=7.555\ 6$
6	2.084 9	$Y_6=1.0+0.8\times6+2.084\ 9=7.884\ 9$
7	1.171 4	$Y_7=1.0+0.8\times7+1.171\ 4=7.771\ 4$
8	2.118 0	$Y_8=1.0+0.8\times8+2.118\ 0=9.518\ 0$
9	1.723 6	$Y_9=1.0+0.8\times9+1.723\ 6=9.923\ 6$
10	0.249 5	$Y_{10}=1.0+0.8\times10+0.249\ 5=9.249\ 5$

注：u_t 数据来自表 12—1。

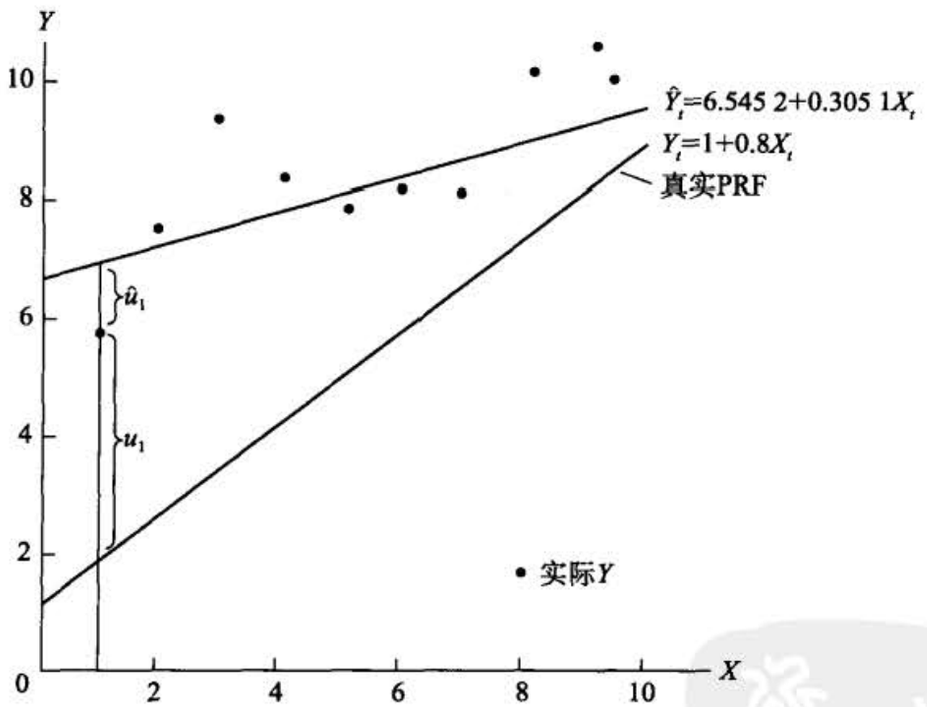


图 12—6　真实 PRF 与表 12—2 所给数据估计的回归线

图 12—6 还表明为什么 u_i 的真实方差常被从 $\hat{u}_i$ 计算得到的估计量 $\hat{\sigma}^2$ 所低估。$\hat{u}_i$ 一般都靠近拟合线（由于 OLS 程序），然而却明显远离了真实的 PRF。因此，它们没有给出 u_i 的正确图像。为了洞察真实 σ^2 被低估的程度，假设我们做另一抽样实验，仍保留表 12—1 和表 12—2 中的 X_t 值和 ε_t 值，但假定 $\rho=0$，即无自相关。由此生成的新样本值如表 12—3 所示。

表 12—3　　零序列相关的 Y 样本值

X_t	$\varepsilon_t=u_t$	$Y_t=1.0+0.8X_t+\varepsilon_t$
1	0.464	2.264
2	2.026	4.626
3	2.455	5.855
4	−0.323	3.877
5	−0.068	4.932
6	0.296	6.096
7	−0.288	6.312
8	1.298	8.698
9	0.241	8.441
10	−0.957	8.043

注：因为没有自相关，所以 u_t 和 ε_t 相同。ε_t 来自表 12—1。

根据表 12—3 得到的回归如下：

$$\begin{aligned}\hat{Y}_t &= 2.534\,5+0.614\,5X_t \\ &\quad (0.679\,6)\ (0.108\,7) \\ t &= (3.791\,0)\ (5.654\,1) \\ r^2 &= 0.799\,7 \qquad \hat{\sigma}^2 = 0.975\,2\end{aligned} \tag{12.4.7}$$

因为现在的 Y 基本上是随机的，故此回归与“真实情况”接近得多。注意，$\hat{\sigma}^2$ 已从 0.811 4（当 $\rho=0.7$ 时）增至 0.975 2（当 $\rho=0$ 时），并且 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的标准误也都已增大。这个结果是与先前考虑的理论结果相一致的。

12.5　1960—2005 年间美国商业部门工资与生产率之间的关系

至此我们已讨论了自相关的后果，一个显然的问题是，我们如何发现这个问题并加以纠正？在讨论这些专题之前，最好先考虑一个简明的例子。表 12—4 给出了美国商业部门在 1960—2005 年间人均真实工资（Y）与人均产出（X）指数数据，这些指数以 1992 年为基年，并取值 100。

表 12—4　　1960—2005 年美国人均真实工资与人均产出指数（数据经过季节调整）

年份	Y	X	年份	Y	X
1960	60.8	48.9	1966	71.7	61.2
1961	62.5	50.6	1967	73.5	62.5
1962	64.6	52.9	1968	76.2	64.7
1963	66.1	55.0	1969	77.3	65.0
1964	67.7	56.8	1970	78.8	66.3
1965	69.1	58.8	1971	80.2	69.0

续前表

年份	Y	X	年份	Y	X
1972	82.6	71.2	1989	95.0	92.4
1973	84.3	73.4	1990	96.2	94.4
1974	83.3	72.3	1991	97.4	95.9
1975	84.1	74.8	1992	100.0	100.0
1976	86.4	77.1	1993	99.7	100.4
1977	87.6	78.5	1994	99.0	101.3
1978	89.1	79.3	1995	98.7	101.5
1979	89.3	79.3	1996	99.4	104.5
1980	89.1	79.2	1997	100.5	106.5
1981	89.3	80.8	1998	105.2	109.5
1982	90.4	80.1	1999	108.0	112.8
1983	90.3	83.0	2000	112.0	116.1
1984	90.7	85.2	2001	113.5	119.1
1985	92.0	87.1	2002	115.7	124.0
1986	94.9	89.7	2003	117.7	128.7
1987	95.2	90.1	2004	119.0	132.7
1988	96.5	91.5	2005	120.2	135.7

注：X = 商业部门每小时产出指数，1992=100。Y = 商业部门每小时真实工资指数，1992=100。
资料来源：Economic Report of the President，2007，Table B-49.

首先把 Y 和 X 的数据描点得到图 12—7。既然预期真实工资与劳动生产率之间的关系为正，那这两个变量相关也就无足为奇。令人吃惊之处在于，二者的关系几乎是线性的，尽管有迹象表明，在生产率的值较高时，二者之间的关系略显非线性。因此，我们决定估计一个线性模型和一个对数线性模型，结论如下：

$$\begin{aligned} \hat{Y}_t &= 32.741\,9 + 0.670\,4\,X_t \\ \text{se} &= (1.394\,0)\ (0.015\,7) \\ t &= (23.487\,4)(42.781\,3) \end{aligned} \tag{12.5.1}$$

$$r^2 = 0.976\,5 \qquad d = 0.173\,9 \qquad \hat{\sigma} = 2.384\,5$$

其中 d 为德宾-沃森统计量，稍后将讨论这个统计量。

$$\begin{aligned} \widehat{\ln Y_t} &= 1.606\,7 + 0.652\,2\ln X_t \\ \text{se} &= (0.054\,7)\quad (0.012\,4) \\ t &= (29.368\,0)\ (52.799\,6) \end{aligned} \tag{12.5.2}$$

$$r^2 = 0.984\,5 \qquad d = 0.217\,6 \qquad \hat{\sigma} = 0.022\,1$$

由于上述模型是双对数模型，所以斜率系数表示弹性。在本例中，我们看到，如果劳动生产率提高 1%，平均工资将提高约 0.65%。

定性而言，这两个模型的结论类似。很高的 t 值表明，在这两种情况下，所估计的系数都“高度”显著。在线性模型中，若生产率指数上升一个单位，则工资指数

平均上升约 0.67 个单位。在对数线性模型中，斜率系数表示弹性（为什么?），我们发现，若生产率指数上升 1%，则真实工资指数平均上升 0.65%。

若存在自相关，方程（12.5.1）和（12.5.2）给出的结论可靠性如何呢? 如前所述，若存在自相关，则估计的标准误就有偏误，因此估计的 t 比率就不可靠。显然，我们需要弄清楚，我们的数据是否受到自相关问题的困扰。在下一节，我们讨论侦察自相关的几种方法。我们只用对数线性模型（12.5.2）来解释这些方法。

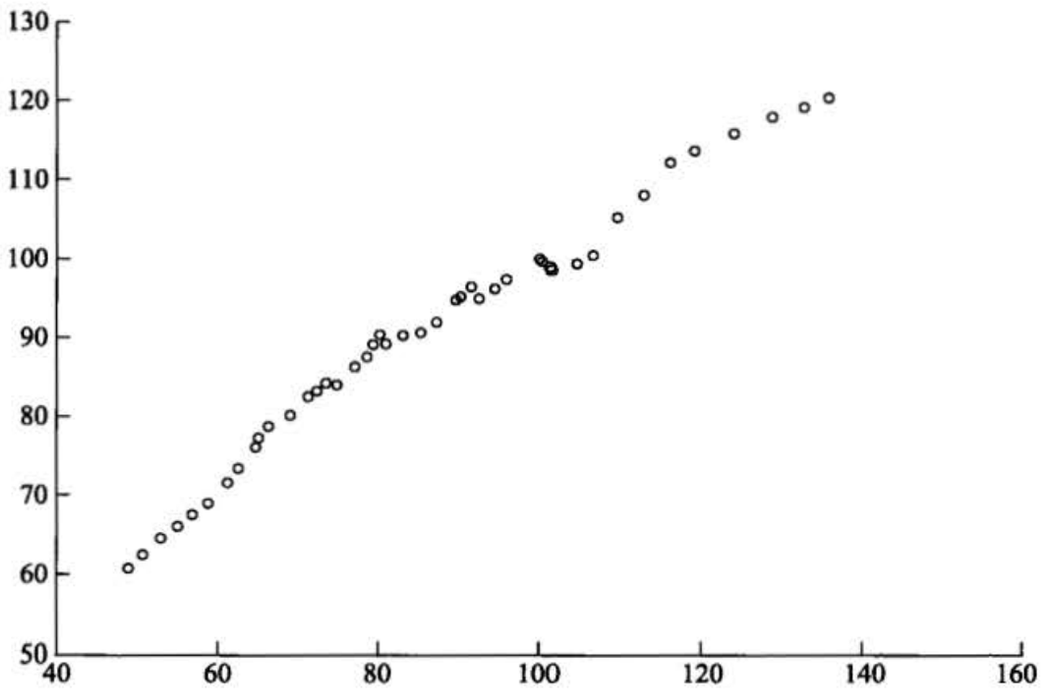


图 12—7　美国工资指数（Y）与生产率指数（X）：1960—2005 年

12.6　侦察自相关

□ I. 图解法

我们说过，经典模型的非自相关（nonautocorrelation）假定是对不可直接观测的总体干扰项 u_t 而言，而我们所能获知的，则是得自通常 OLS 程序的、用以替代 u_t 的残差 $\hat{u}_t$。虽然 $\hat{u}_t$ 不同于 u_t[①]，但对 $\hat{u}_t$ 做一图形检查往往能对 u_t 中可能存有的自相关提供一些线索。其实，对 $\hat{u}_t$ 或 $\hat{u}_t^2$ 的图形检查不仅为自相关而且为异方差性（如我

① 即使干扰项 u_t 是同方差的和不相关的，其估计量 $\hat{u}_t$ 也是异方差和自相关的。关于此问题，参看 G. S. Maddala, *Introduction to Econometrics*, 2d ed., Macmillan, New York, 1992, pp. 480-481。不过，可以证明，随着样本容量无限扩大，残差趋于收敛至其真实值 u_t。对此，可参见 E. Malinvaud, *Statistical Methods of Econometrics*, 2d ed., North-Holland Publishers, Amsterdam, 1970, p. 88。

们在上一章中看到的那样），甚至为模型适宜性或设定偏误（我们将在下一章中看到）都能提供有用的信息。如一位作者所说：

> （残差）图形的生成和分析，作为统计分析的一个标准部分，其重要性无论怎样强调都不过分。除了有时对复杂的问题提供了简明的理解之外，还能在清楚地展现各种个别情形的同时，使我们能从整体上考察数据。①

有多种检查残差的方法。像图 12—8 那样，可以将残差对时间描点而得到一幅**时间顺序图**（time sequence plot）。图 12—8 展现了工资—生产率对数线性回归（12.5.2）的残差。这些残差值与其他数据一起列在表 12—5 中。

表 12—5　　残差的实际值、标准化值及滞后值

观测	S1	SDRES	S1(−1)	观测	S1	SDRES	S1(−1)
1960	−0.036 068	−1.639 433	NA	1983	0.014 416	0.655 291	0.038 719
1961	−0.030 780	−1.399 078	−0.036 068	1984	0.001 774	0.080 626	0.014 416
1962	−0.026 724	−1.214 729	−0.030 780	1985	0.001 620	0.073 640	0.001 774
1963	−0.029 160	−1.325 472	−0.026 724	1986	0.013 471	0.612 317	0.001 620
1964	−0.026 246	−1.193 017	−0.029 160	1987	0.013 725	0.623 875	0.013 471
1965	−0.028 348	−1.288 551	−0.026 246	1988	0.017 232	0.783 269	0.013 725
1966	−0.017 504	−0.795 647	−0.028 348	1989	−0.004 818	−0.219 005	0.017 232
1967	−0.006 419	−0.291 762	−0.017 504	1990	−0.006 232	−0.283 285	−0.004 818
1968	0.007 094	0.322 459	−0.006 419	1991	−0.004 118	−0.187 161	−0.006 232
1969	0.018 409	0.836 791	0.007 094	1992	−0.005 078	−0.230 822	−0.004 118
1970	0.024 713	1.123 311	0.018 409	1993	−0.010 686	−0.485 739	−0.005 078
1971	0.016 289	0.740 413	0.024 713	1994	−0.023 553	−1.070 573	−0.010 686
1972	0.025 305	1.150 208	0.016 289	1995	−0.027 874	−1.266 997	−0.023 553
1973	0.025 829	1.174 049	0.025 305	1996	−0.039 805	−1.809 304	−0.027 874
1974	0.023 744	1.079 278	0.025 829	1997	−0.041 164	−1.871 079	−0.039 805
1975	0.011 131	0.505 948	0.023 744	1998	−0.013 576	−0.617 112	−0.041 164
1976	0.018 359	0.834 515	0.011 131	1999	−0.006 674	−0.303 364	−0.013 576
1977	0.020 416	0.927 990	0.018 359	2000	0.010 887	0.494 846	−0.006 674
1978	0.030 781	1.399 135	0.020 416	2001	0.007 551	0.343 250	0.010 887
1979	0.033 023	1.501 051	0.030 781	2002	0.000 453	0.020 599	0.007 551
1980	0.031 604	1.436 543	0.033 023	2003	−0.006 673	−0.303 298	0.000 453
1981	0.020 801	0.945 516	0.031 604	2004	−0.015 650	−0.711 380	−0.006 673
1982	0.038 719	1.759 960	0.020 801	2005	−0.020 198	−0.918 070	−0.015 650

注：S1＝工资—生产率对数线性回归中得到的残差。
S1(−1) ＝滞后一期的残差值。
SDRES＝标准化残差＝残差/估计值的标准误。

① Stanford Weisberg, *Applied Linear Regression*, John Wiley & Sons, New York, 1980, p. 120.

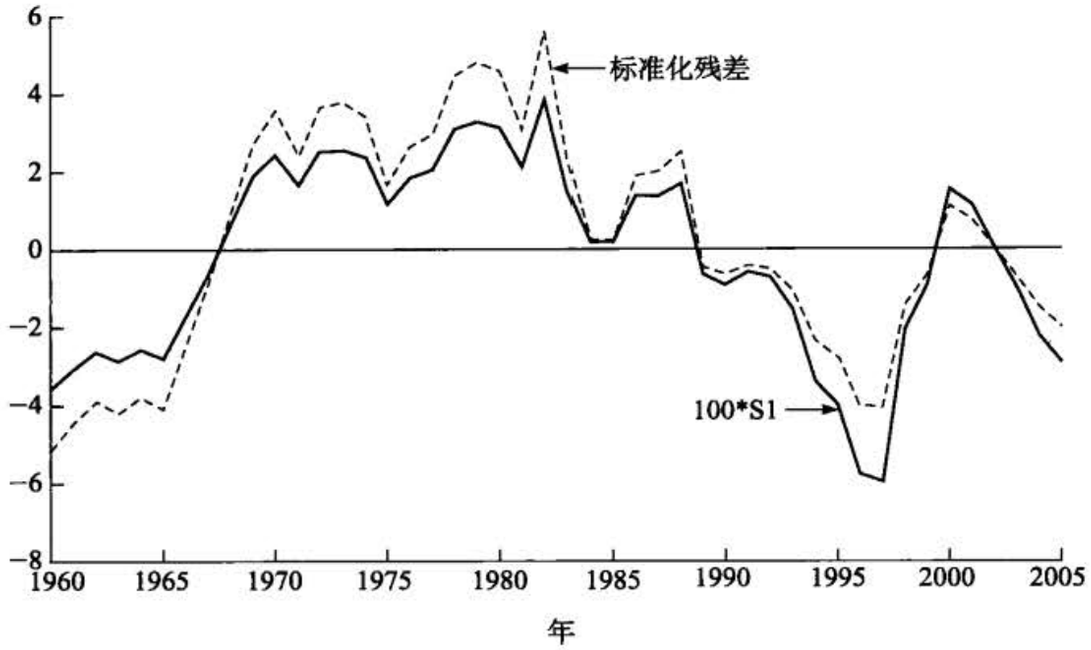


图 12—8　工资—生产率对数线性回归［模型（12.5.2）］中的残差（放大 100 倍）及标准化残差

另一方法是，将**标准化残差**（standardized residual）对时间描图，亦见图 12—8 和表 12—5。标准化残差（$\hat{u}_t/\hat{\sigma}$）无非是 $\hat{u}_t$ 除以估计值的标准误 $\hat{\sigma}(=\sqrt{\hat{\sigma}^2})$。注意到 $\hat{u}_t$ 和 $\hat{\sigma}$ 的测量单位均与回归子 Y 的测量单位相同，这样，$\hat{u}_t/\hat{\sigma}$ 的值就是一个纯数（无测量单位的数），并因此可用来同其他回归的标准化残差相比较。此外，标准化残差和 $\hat{u}_t$ 一样，有零均值（为什么?），并且有近似于 1 的方差。[①] 在大样本中，$\hat{u}_t/\hat{\sigma}$ 近似地服从零均值和单位方差的正态分布。就我们的例子而言，$\hat{\sigma}=2.675\ 5$。

通过分析图 12—8 中的时间顺序图，我们观察到 $\hat{u}_t$ 和标准化残差都呈现一种类似于图 12—1d 的模样，表明 u_t 也许不是随机的。

可从另一角度看待这一问题。将 $\hat{u}_t$ 对 $\hat{u}_{t-1}$ 描点，即 t 时期的残差对它在 $t-1$ 时期的值描点，这是对 AR(1) 方案的一种经验检验。如果残差是非随机的，我们将会得到类似于图 12—3 那样的图形。当我们用上述工资—生产率对数线性回归的 $\hat{u}_t$ 对 $\hat{u}_{t-1}$ 描点时，我们得到的图形如图 12—9 所示，其所依据的数据由表 12—5 给出。该图表明大多数残差都聚集在第一（东北）和第三（西南）象限内，有力地说明残差中有正相关。

刚才讨论的图解法，从性质上看，基本上是主观的或定性的。但有若干定量检验可用来补充这种纯粹定性的方法。现在让我们来介绍其中的一些。

---

① 其实，所谓的**"学生化"残差**（studentized residual），有一单位方差（unit variance），但在实践中，标准化残差给出的图形一般看来都无异于学生化残差所给的图形。因此，我们使用标准化残差也无妨。关于这个问题，参看 Norman Draper and Harry Smith, *Applied Regression Analysis*, 3d ed., John Wiley & Sons, New York, 1998, pp. 207-208。

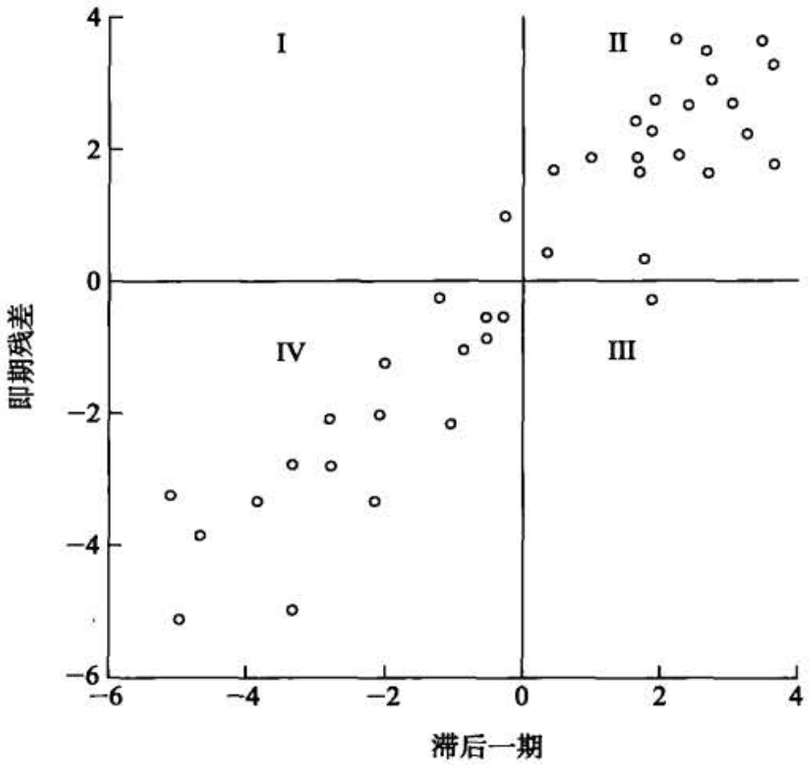


图 12—9　即期残差与滞后残差

□ Ⅱ. 游程检验

如果我们再仔细检查图 12—8，我们会注意到一种奇异的特点：开始时，我们有好几个残差都是负的，然后是一连串正的残差，最后又是几个都是负的残差。如果这些残差是纯粹随机的，我们会观察到这种模式吗？直觉地想，似乎不大可能。这种直觉可通过所谓**游程检验**（runs test）加以核实。这种检验有时又称**吉尔里检验**（Geary test），是一种非参数检验。[①]

为解释这种检验，让我们仅把表 12—5 第一列的工资—生产率回归残差的符号（＋或－）记录下来。

$$(---------)(++++++++++++++++++++++)$$
$$(-----------)(+++)(---) \qquad (12.6.1)$$

即先有 8 个负残差，随之有 21 个正残差，再随之有 11 个负残差，然后是 3 个正残差和 3 个负残差，共 46 次观测。

现在我们定义一个**游程**（run）为同一符号或属性（诸如＋或－）的一个不中断历程（uninterrupted sequence）。我们再定义**游程的长度**（length of a run）为游程中的元素个数。在方程（12.6.1）所示的顺序中共有 5 个游程：1 个 8 次负号游程（长

① 在**非参数**（nonparametric）检验中，我们不对观测值所来自的（概率）分布做任何假定。关于吉尔里检验，参看 R. C. Geary, "Relative Efficiency of Count Sign Changes for Assessing Residual Autoregression in Least Squares Regression," *Biometrika*, vol. 57, 1970, pp. 123-127。

度为 8 的游程），1 个 21 次正号游程（长度为 21 的游程），1 个 11 次负号游程（长度为 10 的游程），1 个 3 次正号游程（长度为 3 的游程），和 1 个 3 次负号游程（长度为 3 的游程）。为了有更好的视觉效果，我们用括号包住各个游程。

在一个严格随机的观测顺序中会出现怎样的游程呢？通过对这一问题的分析，人们能推出游程的一个随机性检验。现在问：在我们的 46 次观测的说明性例子中，我们看到有 5 个游程，如果拿它同一个严格随机的 46 次观测顺序中所预期的游程个数相比，是太多了还是太少了？如果太多，就是说在我们的例子中 a_t 变号太频繁，因而象征着一种负的序列相关（参看图 12—3b）。同理，如果游程太少，则表示可能有正相关（如图 12—3a 所示）。从而，先验地，图 12—8 表示残差中有正相关。

现在令：

N = 总观测个数 $= N_1 + N_2$

N_1 = “+” 个数(即 + 残差)

N_2 = “—” 个数(即 − 残差)

R = 游程个数

于是，在相继结果（这里指残差）互相独立的虚拟假设下，并且假定 $N_1 > 10$ 和 $N_2 > 10$，游程个数将（渐近）服从正态分布，其中：

$$\text{均值：} E(R) = \frac{2N_1N_2}{N} + 1$$

$$\text{方差：} \sigma_R^2 = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)} \tag{12.6.2}$$

注：$N = N_1 + N_2$。

如果随机性假设是可维持的，则可预期在一个问题中所得到的游程个数将以 95% 的置信水平落入如下区间内

$$\text{Prob}[E(R) - 1.96\sigma_R \leqslant R \leqslant E(R) + 1.96\sigma_R] = 0.95 \tag{12.6.3}$$

因此，我们得到这样的规则：

决策规则

在 95% 的置信水平下，若游程个数 R 落在上述置信区间内，就不要拒绝随机性虚拟假设；如果估计的 R 落在此范围外就拒绝虚拟假设。(注意，你可以选择任意一个置信水平。)

回到我们的例子中，负号个数 $N_1 = 24$，正号个数 $N_2 = 22$，$R = 5$。利用方程 (12.6.2) 中的公式我们得到：

$$E(R) = 24$$

$$\sigma_R^2 = 11 \tag{12.6.4}$$

$$\sigma_R = 3.32$$

从而 R 的 95% 置信区间是：

$$(24 \pm 1.96 \times 3.32) = (17.5, 30.5)$$

由于游程个数 5 明显落在此区间之外，按 95%的置信水平，便可拒绝工资—生产率回归中残差的随机性假设。换句话说，残差表现出自相关。作为一般原则，若存在正自相关，则游程数将很小，而若存在负相关，则游程数会很多。当然，我们从方程（12.6.2）可以发现，游程数是太多还是太少。

如果 N_1 或 N_2 小于 20，斯威德（Swed）和艾森哈特（Eisenhart）曾研制出一个专用表，给出在 N 次观测的一个随机顺序中预期游程个数的临界值。此表见附录 D 中的表 D—6。诚望读者利用这些表格证实在我们的工资—生产率回归中，残差确实不是随机的；它们实际上正相关。

□ Ⅲ. 德宾-沃森 d 检验[①]

用以侦察序列相关的最著名的检验是由统计学家德宾（Durbin）和沃森（Watson）提出的。人们普遍称之为**德宾-沃森 d 统计量**（Durbin-Watson d statistic）[②]，其定义如下：

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2} \tag{12.6.5}$$

它无非是相继残差的差异平方和与 RSS 之比。注意，由于取相继差异时损失一个观测值，在 d 统计量的分子中只有 $n-1$ 个观测值。

d 统计量的一大优点是，它仅依赖于估计的残差值，而后者在回归分析中都已例行给出。正因为这一优点，现在常用的做法是把德宾-沃森 d 统计量连同 R^2、调整 R^2、t 比率和 F 值等摘要统计量一起报告。虽然现在 **d 统计量用得很频繁，但记住它的一些基本假定依然很重要：**

1. 回归含有截距项。如果没有截距项，像过原点回归那样，就要重新做带有截距项的回归，以求得 RSS。[③]

2. 解释变量 X 是非随机的，或者在重复抽样中被固定。

3. 干扰项 u_t 是按一阶自回归模式 $u_t = \rho u_{t-1} + \varepsilon_t$ 生成的。因此，它不能用于更高阶自回归模式的侦察。

4. 假定误差项 u_t 服从正态分布。

5. 回归模型不把滞后因变量当作解释变量之一。因此，该检验在如下类型的模型中就不适用：

---

① J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least-Squares Regression," *Biometrika*, vol. 38, 1951, pp. 159-171.

② 常简化为 DW 或 d 统计量。

③ 然而，费尔布拉第（R. W. Farebrother）已计算模型中没有截距项时的 d 值。参看 "The Durbin-Watson Test for Serial Correlation When There Is No Intercept in the Regression," *Econometrica*, vol. 48, 1980, pp. 1553-1563。

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t \qquad (12.6.6)$$

其中 Y_{t-1} 是 Y 的一期滞后值，d 统计量便不适用。这样的模型叫做**自回归模型**（autoregressive model）。在第 17 章中我们将全面地分析它。

6. 没有数据缺失。例如，在我们对 1960—2005 年间的工资—生产率回归中，如果出于某种原因，（比方说）1978 年和 1982 年的观测值缺失，则 d 统计量对这种缺失数据没有补偿办法。①

如德宾和沃森曾指出的那样，由于 d 统计量的分布与出现在给定样本中的 X 值有复杂的关系，所以要推导出其准确的抽样或概率分布就很困难。② 这种困难是可以理解的。因为 d 要从 $\hat{u}_t$ 算出，而 $\hat{u}_t$ 必然依赖于给定的 X。因此，它不同于 t，F 或 χ^2 检验，没有唯一的临界值可以导致拒绝或接受如下虚拟假设：干扰项 u_t 中无一阶序列相关。然而，他们成功地推导出了临界值的一个下限 d_L 和一个上限 d_U，如果从方程（12.6.5）算出的 d 值落在这些临界值的范围之外，就可作出是否有正或负序列相关的判断。此外，这些界限值仅依赖于观测值的个数 n 以及解释变量的个数，却不依赖于这些解释变量取什么值。德宾和沃森已对从 6 到 200 的 n 值和多至 20 个解释变量的临界值编制成表，见附录 D 表 D—5（解释变量可多至 20 个）。

实际检验步骤可借助于图 12—10 作出更好的解释。该图表明 d 的两个极限值是 0 和 4，这可证明如下。将方程（12.6.5）展开得：

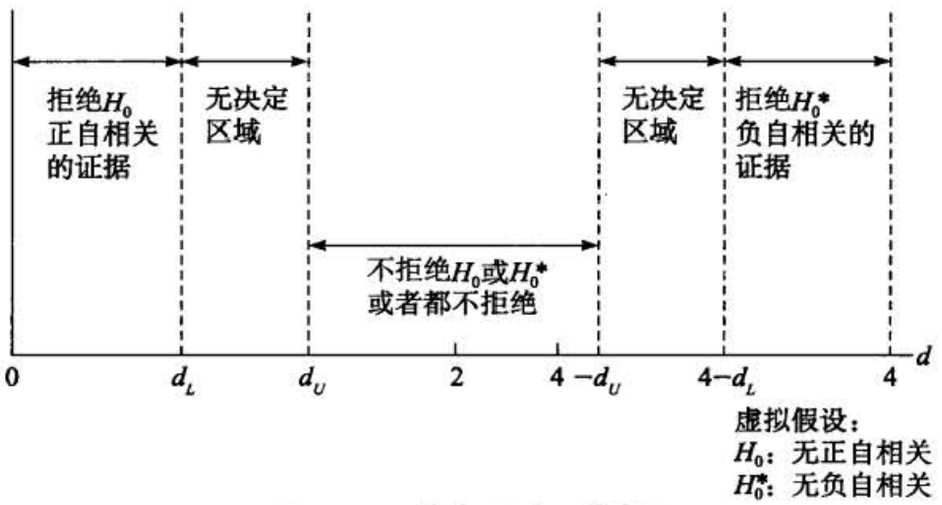


图 12—10　德宾-沃森 d 统计量

$$d = \frac{\sum \hat{u}_t^2 + \sum \hat{u}_{t-1}^2 - 2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \qquad (12.6.7)$$

因 $\sum \hat{u}_t^2$ 和 $\sum \hat{u}_{t-1}^2$ 只有一次观测之差，故可看作近似相等。因此，令：

$$\sum \hat{u}_{t-1}^2 \approx \sum \hat{u}_t^2$$

① 更详尽的细节，参见 Gabor Korosi, Laszlo Matyas, and Istvan P. Szekey, *Practical Econometrics*, Avebury Press, England, 1992, pp. 88-89。

② 参看本节稍后关于“精确”的德宾-沃森检验的讨论。

方程（12.6.7）便可写为：

$$d \approx 2\left[1-\frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right] \tag{12.6.8}$$

其中 $\approx$ 表示近似等于。

现定义

$$\hat{\rho}=\frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \tag{12.6.9}$$

为样本一阶自相关系数，作为 ρ 的一个估计量（参看第 417 页注释①）。利用方程（12.6.9），可将方程（12.6.8）表示为：

$$d \approx 2(1-\hat{\rho}) \tag{12.6.10}$$

但因 $-1 \leqslant \rho \leqslant 1$，故方程（12.6.10）意味着

$$0 \leqslant d \leqslant 4 \tag{12.6.11}$$

这就是 d 的界限；任何一个 d 的估计值必定落入这一界限之内。

由方程（12.6.10）显然可见，若 $\hat{\rho}=0$ 则 $d=2$；就是说，如果没有（一阶）序列相关，则预期 d 约为 2。因此，作为一种经验法则，如果在一项应用中求出 d 等于 2，便可认为没有一阶自相关，不管是正的或负的。如果 $\hat{\rho}=+1$，表示残差中有完全的正相关，则 $d \approx 0$。因此，d 越接近 0，正序列相关的迹象越明显。这种关系还能从方程（12.6.5）中看出。因为如果有正自相关，那些 $\hat{u}_t$ 就会被束缚在一起，以使它们的差异趋于微小，其结果将是分子平方和相对较小，而分母在任一给定的回归中保持着唯一值。

如果 $\hat{\rho}=-1$，即相继残差中有完全的负相关，则 $d \approx 4$。因此，d 越接近 4，负序列相关的迹象越明显。再从方程（12.6.5）来看，这也是可以理解的。因为如果有负自相关，就会有这样一种趋向：一个正的 $\hat{u}_t$ 将随后有一个负的 $\hat{u}_t$；反之亦然，以致 $|\hat{u}_t-\hat{u}_{t-1}|$ 常常比 $|\hat{u}_t|$ 大。因此，相对而言，d 的分子将要大于其分母。

德宾-沃森检验的操作步骤如下：假定该检验的基本假定成立，

1. 做 OLS 回归并获取残差。

2. 按方程（12.6.5）计算 d。（现今的计算机程序大多数都给出 d 值。）

3. 对给定样本容量和给定解释变量个数找出临界值 d_L 和 d_U。

4. 按照表 12—6 的决策规则行事。为了易于参照，再将这些规则描述如图 12—10。

为了说明步骤，再回到我们的工资—生产率回归。由表 12—5 的数据，估计的 d 值为 0.217 5，表明残差中有正的序列相关。（为什么?）由德宾-沃森表我们找出，对于 46 个观测和 1 个解释变量（不包括截距），在 5%水平上，$d_L=1.475$ 和 $d_U=1.566$。由于 d 的估计值 0.217 5 低于 d_L，我们不能拒绝残差中有正序列相关的假设。

表 12—6　　德宾-沃森 d 检验：决策规则

虚拟假设	决策	如果
无正自相关	拒绝	$0<d<d_L$
无正自相关	无决定	$d_L \leqslant d \leqslant d_U$
无负自相关	拒绝	$4-d_L<d<4$
无负自相关	无决定	$4-d_U \leqslant d \leqslant 4-d_L$
无正或负自相关	不拒绝	$d_U<d<4-d_U$

d 检验虽然极为流行，却有一大缺陷：如果它落入**不确定域**（indecisive zone）我们就无法论定一阶自相关是否存在。为解决此问题，一些作者曾提出对德宾-沃森 d 检验的修改建议，但涉及的方法相当复杂，超过了本书的论述范围。[①] 然而，在许多情况中，人们发现上限 d_U 差不多就是真实的显著性界限，因而，如果 d 的估计值落入不确定域，就不妨使用如下**修订的 d 检验**（modified d test）。给定显著性水平 α，

1. H_0：$\rho=0$ 和 H_1：$\rho>0$。如果估计的 $d<d_U$，则在水平 α 上拒绝 H_0；即存在着统计上显著的正自相关。

2. H_0：$\rho=0$ 和 H_1：$\rho<0$。如果估计的 $4-d<d_U$，则在水平 α 上拒绝 H_0；即存在着统计上显著的负相关。

3. H_0：$\rho=0$ 和 H_1：$\rho\neq0$。如果估计的 $d<d_U$ 或 $4-d<d_U$，则在水平 2α 上拒绝 H_0；即存在着统计上显著的正或负自相关。

或许可以指出，随着样本容量的扩大，无决定域逐渐变小，从德宾-沃森表可以清楚地看出这一点。比如，对 4 个回归元和 20 次观测而言，显著性水平为 5%时，d 值的下界和上界分别是 0.894 和 1.828，但当样本容量扩大到 75 时，这些值分别是 1.515 和 1.739。

计算机程序 SHAZAM 能做出（一种）精确的 d 检验（exact d test），即给出 d 计算值的准确概率或 p 值。利用现代计算工具，不难求出计算 d 统计量的 p 值。在我们的工资—生产率回归中，利用 SHAZAM 第 9 版，我们求出 d 计算值 0.217 6 的 p 值实际上为零，因而加强了我们前面基于德宾-沃森表得到的结论。

德宾-沃森 d 检验变得如此神圣，以致实践者常常忘记其背后的假定。特别是假定：(1) 解释变量或回归元是非随机的，(2) 误差项服从正态分布，(3) 回归模型不包括回归子的滞后值，(4) 只考虑一阶序列相关，这些对使用 d 检验都至关重要。

如果一个回归模型包含了回归子的滞后值，这种情形下的 d 值常为 2 左右，从而表明这种模型中不存在（一阶）自相关。因此，在这种模型中有一个有碍于发现序列相关的内在偏误。这并不意味着自回归模型就没有自相关的问题。事实上，德宾曾提出用所谓的 **h 检验**（h test）来检验这种模型中的序列相关。但从统计学意义上看，这一检验不如稍后讨论的**布罗施-戈弗雷检验**（Breusch-Godfrey test）那样有效，所以就没有使用 h 检验的必要。但鉴于其历史上的重要性，在习题 12.36 中将

① 细节见 Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods*, Springer Verlag, New York, 1984, pp. 225-228。

讨论它。

此外，如果误差项 u_t 不是 NIID，那么，惯常使用的 d 检验可能就不可靠。[①] 由此看来，前面讨论过的游程检验因不需要对误差项做任何分布上的假定而占据优势。然而，若样本容量很大（技术性地讲，就是无穷大），那我们就可以使用德宾-沃森 d 检验，因为可以证明[②]：

$$\sqrt{n}\left(1-\frac{1}{2}d\right)\sim N(0,1) \tag{12.6.12}$$

也就是说，在大样本容量下，按方程（12.6.12）变换后的 d 统计量服从标准正态分布。顺便一提，根据 d 和估计的一阶自相关系数 $\hat{\rho}$［见方程（12.6.10）］之间的关系可得到：

$$\sqrt{n}\hat{\rho}\sim N(0,1) \tag{12.6.13}$$

即在大样本情形下，样本容量的平方根与估计的一阶自相关系数之积也服从标准正态分布。

用我们的工资—生产率例子来说明这一检验，我们发现 $n=46$ 时 $d=0.217\ 6$。因此，我们从方程（12.6.12）求出

$$\sqrt{46}\left(1-\frac{0.217\ 6}{2}\right)\approx 6.044\ 7$$

渐近地讲，若零（一阶）自相关的虚拟假设为真，则得到一个大于或等于 6.044 7 的 Z 值（即标准正态变量）的概率极小。记住，对标准正态分布而言，（双侧）5%的 Z 临界值只有 1.96，而 1%的 Z 临界值也约为 2.58。尽管我们的样本容量仅为 46，但实际上足以使用正态近似。结论仍相同，即工资—生产率回归中的残差存在自相关的问题。

但 d 检验最严重的问题是回归元非随机性的假定，即回归元的值在重复抽样中保持不变的假定。若非如此，则无论是有限样本或小样本，还是大样本，d 检验都不成立。[③] 而由于这一假定在涉及时间序列数据的经济模型中通常难以维持，因此有位作者认为，在涉及时间序列数据的计量经济学中，德宾-沃森统计量或许没有用武之地。[④] 按照他的观点，仍有更有用的自相关检验可用，但它们都以大样本为前提。我们下面就讨论一个这种检验：**布罗施-戈弗雷检验**。

□ 自相关的一般性检验：布罗施-戈弗雷检验[⑤]

为了避免自相关的德宾-沃森 d 检验所存在的隐患，统计学家布罗施和戈弗雷提出

---

① 更深入的讨论，参见 Ron C. Mittelhammer, George G. Judge, and Douglas J. Miller, *Econometric Foundations*, Cambridge University Press, New York, 2000, p. 550。

② 参见 James Davidson, *Econometric Theory*, Blackwell Publishers, New York, 2000, p. 161。

③ Davidson, Ibid., p. 161.

④ Fumio Hayashi, *Econometrics*, Princeton University Press, Princeton, NJ, 2000, p. 45.

⑤ 参见 L. G. Godfrey, "Testing against General Autoregressive and Moving Average Error Models When the Regressors Includes Lagged Dependent Variables," *Econometrica*, vol. 46, 1978, pp. 1293-1302。以及 T. S. Breusch, "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, vol. 17, 1978, pp. 334-355。

了一种自相关检验。这种检验容许：(1) 非随机回归元，如回归子的滞后值；(2) 高阶自回归模式，如 AR(1)，AR(2) 等；(3) 白噪音误差项［如方程 (12.2.1) 中的 ε_t］的简单或更高阶**移动平均** (moving averages，MA)[①]，从这种意义上看，它比前面的检验更具有一般性。

尽管从参考文献中可以查到详尽的数学推导，但抛开这些，**BG 检验** (BG test)，也被称为 **LM 检验** (LM test)[②]，具体过程如下：尽管可以在模型中添加许多回归元，但我们仍以双变量回归模型来说明这个检验。此外，回归子的滞后值也可以放到模型中。令：

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.6.14}$$

假定误差项 u_t 服从如下 p 阶自回归 AR(p) 模式：

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \tag{12.6.15}$$

其中 ε_t 为前面讨论过的白噪音项。你将会意识到，这只是对 AR(1) 模式的推广。

欲检验的虚拟假设 H_0 是：

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0 \tag{12.6.16}$$

即不存在任何阶数的序列相关。BG 检验包含如下步骤：

1. 用 OLS 估计方程 (12.6.14) 并得到残差 $\hat{u}_t$。

2. 将 $\hat{u}_t$ 对原 X_t (若原模型中有不止一个 X 变量，则都包括进来) 和第一步所估计的残差滞后值 $\hat{u}_{t-1}$，$\hat{u}_{t-2}$，…，$\hat{u}_{t-p}$做回归。因此，若 $p=4$，则我们就在模型中引入残差的 4 个滞后值作为额外的回归元。注意，在做这个回归时，我们将只有 $n-p$ 次观测。(为什么?) 简言之，做如下回归：

$$\hat{u}_t = \alpha_1 + \alpha_2 X_t + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 \hat{u}_{t-2} + \cdots + \hat{\rho}_p \hat{u}_{t-p} + \varepsilon_t \tag{12.6.17}$$

并从这个 (辅助) 回归中得到 R^2。[③]

3. 若样本容量很大 (技术上讲是无限样本)，则布罗施和戈弗雷证明了

$$(n-p)R^2 \sim \chi^2_p \tag{12.6.18}$$

即 $n-p$ 乘以从辅助回归 (12.6.17) 中得到的 R^2 值服从自由度为 p 的 χ^2 分布。在应用中，若在选定的显著性水平下 $(n-p)R^2$ 超过 χ^2 临界值，我们就拒绝虚拟假设，此时方程 (12.6.15) 中至少有一个 ρ 在统计上显著异于零。

对于 BG 检验，有如下实际要点须注意：

1. 回归模型所包含的回归元中或许有回归子 Y 的滞后值，即 Y_{t-1}，Y_{t-2}等可能作为解释变量而出现。相比之下，德宾-沃森检验则要求回归元中不含回归子滞后值。

2. 前面曾指出，即使干扰项服从 p 阶**移动平均**过程，即 u_t 如下生成：

---

① 比如，在回归 $Y_t = \beta_1 + \beta_2 X_t + u_t$ 中，误差项可表示为 $u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2}$，它表示了白噪音误差项 ε_t 的一个三期移动平均。

② 此检验基于第 8 章曾简要提到的**拉格朗日乘数原理** (Lagrange multiplier principle)。

③ 包含原回归元 X 的原因，是容许 X 可能不是严格非随机的这一事实。但如果它是严格随机的，则可从模型中省掉。对此，参见 Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, South-Western Publishing Co., 2003, p. 386。

$$u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \cdots + \lambda_p \varepsilon_{t-p} \tag{12.6.19}$$

其中 ε_t 为白噪音误差项，即这些误差项满足全部经典假定，BG 检验也可适用。

在有关时间序列计量经济学的章节中，我们将详尽研究 p 阶自回归和移动平均过程。

3. 若方程（12.6.15）中 $p=1$（即一阶自回归），则 BG 检验可称为**德宾的 M 检验**（Durbin's M test）。

4. BG 检验的一个缺陷在于，滞后长度 p 值不能先验设定。这就不可避免对 p 值的多次试验。人们有时候也能用所谓的**赤池**（Akaike）和更有功效的**施瓦茨**（Schwarz）信息准则来筛选滞后长度。我们将在第 13 章及后面有关时间序列计量经济学的章节中讨论这些准则。

5. 给定 X 变量值和 u 的滞后值，该检验假定方程（12.6.15）中的 u 是同方差的。

对 BG 检验的说明：工资—生产率关系

为说明此检验，我们在前面说明性的例子中应用一下这个检验。利用一个 AR(6) 模式，我们得到习题 12.25 中所示的结论。从那里给出的回归结果可以看出，$(n-p)=40$ 和 $R^2=0.7498$。因此，二者相乘则得到一个 χ^2 值 29.992。对 6 个自由度（为什么?）而言，得到一个 χ^2 值大于或等于 29.992 的概率极小；附录 D 表 D—4 中的 χ^2 表表明，得到一个约等于 18.547 6 或更大的 χ^2 值的概率仅为 0.005。因此，得到一个约等于 30 的 χ^2 值的概率确实极小。事实上，实际的 p 值几乎为零。

于是，对我们的例子来说，结论是这 6 个自相关系数中至少有一个非零。

试遍从 1 到 6 的滞后长度，我们发现，只有 AR(1) 系数显著，从而表明没有必要考虑多于一个的滞后。实质上，此时的 BG 检验就是**德宾的 M 检验**。

□ 为什么有这么多的自相关检验?

对这个问题的回答是："……没有某个特定的检验被证明绝对最好（即在统计意义上更有功效），因而分析家在考虑一系列检验程序来侦察自相关的存在或结构时，仍处在十分尴尬的境地。"[①] 当然，上一章讨论异方差性的各种检验时同样存在这个问题。

12.7 发现自相关该怎么办：补救措施

如果我们利用上一节中讨论的一种或多种自相关诊断检验，并发现存在自相关

① Ron C. Mittelhammer et al., op. cit., p. 547. 记住，一个统计检验的功效等于 1 减去犯第 II 类错误的概率，即减去接受一个错误假设的概率。一个检验的最大功效为 1，最小功效为 0。一个检验的功效越接近于 0，该检验就越糟；越接近于 1，就越有效。这些作者所言的本质含义是，不存在单个最有效的自相关检验。

问题，那该怎么办呢？我们有 4 种选择：

1. 尽力查明自相关是否是**纯粹自相关**（pure autocorrelation），而不是模型误设的结果。我们在 12.1 节讨论过，我们有时候观察到的残差形式是模型误设——即排除了某些重要变量——或函数形式不正确所导致的。

2. 若是纯粹自相关，则可对原模型做适当变换，使变换后的模型不存在（纯粹）自相关问题。如同出现异方差时一样，我们必须使用某种**广义最小二乘**（generalized least-square，GLS）。

3. 在大样本下，我们可以用**尼威-威斯特**（Newey-West）方法，以得到 OLS 估计量在对自相关加以修正之后的标准误。这一方法实际上是对我们在上一章中讨论过的怀特异方差一致标准误方法的推广。

4. 在某些情形下，我们可以继续使用 OLS 方法。

鉴于每个专题都很重要，我们将分节加以讨论。

12.8 模型误设与纯粹自相关

让我们回到方程（12.5.2）给出的工资—生产率回归中来。我们在那里看到，d 值为 0.217 6，而我们基于德宾-沃森 d 检验断定，误差项中存在着正的自相关。这种自相关会因我们的模型被不正确地设定而引起吗？由于回归（12.5.1）背后的数据是时间序列数据，所以工资与生产率很可能都表现出时间趋势。若然，我们就必须在模型中包括时间或趋势变量 t，分析工资和生产率在除去趋势因素后的关系。

为对此进行检验，我们在方程（12.5.1）中包含时间趋势变量并得到如下结果：

$$\begin{aligned}\hat{Y}_t &= 0.120\,9 + 1.028\,3X_t - 0.007\,5t \\ \text{se} &= (0.307\,0)\quad (0.077\,6)\quad (0.001\,5) \\ t &= (0.393\,9)\ (13.259\,4)\ (-4.890\,3) \qquad (12.8.1)\\ R^2 &= 0.990\,0; \qquad d = 0.449\,7\end{aligned}$$

对此模型的解释直截了当：真实工资指数逐年递减约 0.75 单位。在容许包含趋势变量后，生产率指数每提高 1 个单位，真实工资指数则平均增加约 1 个单位。有趣的是，即便容许出现趋势变量，d 值依然很低，这表明方程（12.8.1）存在纯粹自相关的问题，而不一定是设定偏误的问题。

我们如何知道方程（12.8.1）被正确地设定了呢？为检验这一点，我们将 Y 对 X 和 X^2 做回归，来检验真实工资指数与生产率指数有非线性相关的可能性。回归结果如下：

$$\begin{aligned}\hat{Y}_t &= -1.784\,3 + 2.196\,3X_t - 0.175\,2\,X_t^2 \\ t &= (-2.771\,3)\ (7.504\,0)\ (-5.278\,5) \qquad (12.8.2)\\ R^2 &= 0.990\,6 \qquad d = 0.356\,1\end{aligned}$$

对这些结论的解释留给读者完成。就目前的问题来看，德宾-沃森统计量的值仍很低，这就表明残差中仍存在正的自相关。

从上述分析可以安全地断定，我们的工资—生产率回归可能遇到纯粹自相关问题，而不一定是设定偏误的问题。知道自相关的后果之后，我们就要采取某种措施来修正它。我们稍后便这么做。

顺便一提，对我们上面给出的所有工资—生产率回归而言，我们用**雅克-贝拉正态性检验**（Jarque-Bera test of normality）发现残差是正态分布的，由于 d 检验假定了误差项的正态性，所以这一结果令人欣慰。

12.9 （纯粹）自相关的修正：广义最小二乘

知道自相关的后果之后，特别是知道 OLS 估计量缺乏效率之后，我们或许要补救这个问题。补救措施取决于对干扰项之间相互依赖的性质的了解，即对自相关结构的了解。

首先，考虑双变量回归模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.9.1}$$

并假定误差项服从 AR(1) 模式，即

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{12.9.2}$$

现在我们考虑两种情况：(1) ρ 已知，(2) ρ 未知并有待估计。

□ ρ 已知

若一阶自相关系数 ρ 已知，则序列相关问题就可轻易解决。如果方程（12.9.1）在时刻 t 成立，则在时刻 $t-1$ 也成立。从而，

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{12.9.3}$$

用 ρ 乘以方程（12.9.3）的两边，得到：

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \tag{12.9.4}$$

从方程（12.9.1）中减去方程（12.9.4）便得到：

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \tag{12.9.5}$$

其中 $\varepsilon_t = u_t - \rho u_{t-1}$。

我们可将方程（12.9.5）表示为：

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \tag{12.9.6}$$

其中 $\beta_1^* = \beta_1(1-\rho)$，$Y_t^* = Y_t - \rho Y_{t-1}$，$X_t^* = X_t - \rho X_{t-1}$，$\beta_2^* = \beta_2$。

由于方程（12.9.6）中的 ε_t 满足全部 OLS 假定，故可直接对转换变量 Y^* 和 X^* 应用 OLS 并获得具有全部最优性质的估计量，即 BLUE。其实，做回归（12.9.6）就等于应用上一章中讨论的广义最小二乘（GLS）。记住，GLS 无非就是把 OLS 用

于变换后满足经典假定的变换模型。

回归（12.9.5）叫做**广义**（generalized）**或准**（quasi）**差分方程**（difference equation）。它不是原来的形式，而是以准差分形式将 Y 对 X 回归。这一差分形式是从一个变量的现期值减去它的前期值的一个比例（$=\rho$）部分而得到的。在这个取差分的过程中，由于第一次观测值没有先前值，所以失去了一次观测。为弥补这一损失，将对 Y 和 X 的第一次观测转换为 $Y_1\sqrt{1-\rho^2}$ 和 $X_1\sqrt{1-\rho^2}$。[①]这一转换被称为**普莱斯-温斯顿变换**（Prais-Winsten transformation）。

□ ρ 未知

方程（12.9.5）所给的广义差分回归的应用虽然直接明了，但因 ρ 实际上鲜为人知，故一般而言难以实现，从而需要另想办法。其中的一些方法如下所述。

一次差分法。因 ρ 落在 0 与±1 之间，故可从两个极端开始尝试。在一个极端上可假定 $\rho=0$ 即无（一阶）序列相关，而在另一个极端上则令 $\rho=\pm1$，即完全正的或负的相关。其实，当我们做一个回归时，我们通常假定没有自相关，然后通过德宾-沃森检验或其他检验以表明这种假定是否合理。而当 $\rho=+1$ 时，广义差分方程（12.9.5）便化为一阶差分方程：

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1})$$

或者：

$$\Delta Y_t = \beta_2 \Delta X_t + \varepsilon_t \tag{12.9.7}$$

其中 Δ 是方程（12.1.10）中曾引入的一阶差分运算子（operator）。

由于方程（12.9.7）中的误差项没有（一阶）序列相关的问题（为什么?），所以为了做回归（12.9.7），唯一要做的就是形成回归子和回归元的一阶差分，并对这些一阶差分做回归。

如果自相关系数很高（比方说大于 0.8）或德宾-沃森 d 统计量很低，那么进行一阶差分变换可能合适。曼德拉曾提出一个粗略的经验法则：只要 $d<R^2$，就能用一阶差分形式。[②] 在我们的工资—生产率对数线性回归（12.5.2）中，$d=0.217\,6$ 和 $r^2=0.984\,5$ 就属于这种情况。对其做一阶差分回归稍后给出。

一阶差分模型（12.9.7）的一个有趣特征是，**它不含有截距项**。因此，为了估计方程（12.9.7），你必须使用**过原点的回归**（regression through the origin，即去掉截距项），现在大多数软件包都能做到这些。不过，如果你忘记从模型中去掉截距项，并估计如下包含截距项的模型

$$\Delta Y_t = \beta_1 + \beta_2 \Delta X_t + \varepsilon_t \tag{12.9.8}$$

---

① 一次观测的损失在大样本中也许关系不大，但在小样本中可能造成明显差异。若不按照指出的方法对第一次观测进行变换，误差方差就不是同方差的。对此，参见 Jeffrey Wooldridge, op. cit., p. 388。关于第一次观测的重要性的某些蒙特卡罗结果，见 Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993, Table 10.1, p. 349。

② Maddala, op. cit., p. 232.

那么，原模型必定有趋势变量，而 β_1 表示它的系数。[①] 因此，在一阶差分模型中引入截距项的一个“意外”好处是，检验原模型中是否出现趋势变量。

回到我们的工资—生产率回归（12.5.2），给定 AR(1) 模式和相对 r^2 来说较低的 d 值，用不含截距的一阶差分形式重做回归（12.5.2）；记住，方程（12.5.2）是水平值形式的。结果如下[②]：

$$\widehat{\Delta Y_t} = 0.653\,9\Delta X_t$$
$$t = (11.404\,2) \qquad r^2 = 0.426\,4 \qquad d = 1.744\,2 \tag{12.9.9}$$

与水平值形式的回归（12.5.2）相比，我们看到，斜率系数没多大变化，但 r^2 值下降相当多。由于取一阶差分时，我们实质上研究的是变量在其（线性）趋势值附近的行为，所以通常都会如此。当然，因为方程（12.9.9）和（12.5.2）这两个模型中的因变量不同，所以我们不能将它们的 r^2 值直接比较。[③] 同时还须注意到，与原来的回归相比，d 值明显提高，这可能表明，在一阶差分回归中没有什么自相关。[④]

一阶差分变换的另一个有趣特征与时间序列背后的平稳性质有关。回到描述 AR(1) 模式的方程（12.2.1）。现在，若 ρ 实际上等于 1，则从方程（12.2.3）和（12.2.4）明显可见，序列 u_t 是非平稳的，因为方差和协方差都变成无穷大。这正是我们为什么在讨论此专题时总施加约束 $|\rho|<1$ 的原因。但从方程（12.2.1）明显可见，若自相关系数实际上为 1，则方程（12.2.1）变成

$$u_t = u_{t-1} + \varepsilon_t$$

或者

$$u_t - u_{t-1} = \Delta u_t = \varepsilon_t \tag{12.9.10}$$

即 u_t 的一阶差分变成平稳序列了，因为它等于白噪音误差项 ε_t。

以上讨论的要点是，若原时间序列是非平稳的，那么其一阶差分很可能变成平稳序列。因此，一阶差分变换起到一箭双雕的作用，既可能会消除（一阶）自相关，又使时间序列变得平稳。我们在更深入讨论时间序列计量经济学的第 5 篇时将会重新探讨这一专题。

我们曾提到，一阶差分变换在 ρ 很高或 d 很低时都适当。严格地讲，一阶差分变换只有在 $\rho = 1$ 时才能成立。事实上，有一个被称为**贝伦布鲁特-韦布检验**（Berenblutt-Webb test）[⑤] 的方法，可用来检验 $\rho=1$ 的假设。其所用的检验统计量被称为 **g 统计量**（g statistic），定义如下：

---

① 这点易于证明。令 $Y_t=\alpha_1+\beta_1 t+\beta_2 X_t+u_t$。因此，$Y_{t-1}=\alpha+\beta_1(t-1)+\beta_2 X_{t-1}+u_{t-1}$，前者减去后者须得到 $\Delta Y_t=\beta_1+\beta_2\Delta X_t+\varepsilon_t$，这就表明此方程中的截距项实际上就是原模型中趋势变量的系数。记住我们假定 $\rho=1$。

② 在习题 12.38 中，要求你包括常数项做这个模型。

③ 比较水平值形式与一阶差分形式的 r^2 略显复杂，对此的进一步讨论，参见 Maddala，op. cit.，Chapter 6。

④ 一阶差分回归中计算出来的 d 值是否能与原水平值回归中的 d 值做同样的解释并不清楚。但用游程检验可以看出，一阶差分回归的残差中没有自相关的证据。

⑤ I. I. Berenblutt and G. I. Webb, “A New Test for Autocorrelated Errors in the Linear Regression Model,” *Journal of the Royal Statistical Society*, Series B, vol. 35, no. 1, 1973, pp. 33-50.

$$g=\frac{\sum_2^n \hat{e}_t^2}{\sum_1^n \hat{a}_t^2} \tag{12.9.11}$$

其中 $\hat{a}_t$ 为原 OLS 回归（即水平值形式的回归）中得到的残差，而 e_t 则为从一阶差分回归中所得到的 OLS 残差。记住，采用一阶差分形式时没有截距项。

为检验 g 统计量的显著性，假定水平形式回归中包含截距项，那我们就可以使用德宾-沃森表，只是现在的虚拟假设是 $\rho=1$，而不是德宾-沃森的假设 $\rho=0$。

重新回到我们的工资—生产率回归，我们从原回归（12.5.2）得到 $\sum \hat{a}_t^2 =$ 0.021 4 和 $\sum \hat{e}_t^2 =0.004\ 6$。代入方程（12.9.11）所示的 g 统计量得到

$$g=\frac{0.004\ 6}{0.021\ 4}=0.214\ 9 \tag{12.9.12}$$

查阅德宾-沃森表（附录 D，表 D—5）中 45 次观测和 1 个解释变量，我们发现 $d_L=$ 1.288，$d_U=1.376$（显著性水平为 5%）。由于所得到的 g 值低于 d 值的下限，所以我们不能拒绝真实的 ρ 为 1 的假设。牢记，尽管我们使用同样的德宾-沃森统计表，但现在的虚拟假设是 $\rho=1$ 而不是 $\rho=0$。鉴于这一发现，方程（12.9.9）所给出的结论或许可以接受。

基于德宾-沃森 d 统计量的 ρ。若因 ρ 与 1 不够接近而不能使用一阶差分变换，那我们从前面在方程（12.6.10）中构建的 d 和 ρ 之间的关系中能找到一个简单的估计方法，我们可以用以下方法估计出 ρ：

$$\hat{\rho}\approx 1-\frac{d}{2} \tag{12.9.13}$$

因此，当样本充分大时，便可从方程（12.9.13）中估计出 ρ 来，并如广义差分方程（12.9.5）那样用它来对数据进行变换。记住，方程（12.9.13）中给出的 ρ 和 d 的关系对小样本情形可能不成立，瑟尔和纳加对此提出了修正意见，可参见习题 12.6。

我们在工资—生产率回归（12.5.2）中得到 $d=0.217\ 6$，代入方程（12.9.13）则得到 $\hat{\rho}\approx 0.891\ 2$。我们可以利用这个估计的 ρ 值来估计方程（12.9.5）。所需做的只是，将 Y 和 X 的当期值都减去其前一期值的 0.891 2 倍，然后按方程（12.9.6）对如此变换的数据做 OLS 回归，其中 $Y_t^* = Y_t - 0.891\ 2Y_{t-1}$，$X_t^* = X_t - 0.891\ 2X_{t-1}$。

从残差中估计出来的 ρ。若 AR(1) 模式 $u_t=\rho u_{t-1}+\epsilon_t$ 成立，一个估计 ρ 的简单方法就是将残差 $\hat{a}_t$ 对 $\hat{a}_{t-1}$ 做回归，因为前面曾指出，$\hat{a}_t$ 是真实 u_t 的一致估计量。即做如下回归：

$$\hat{a}_t=\rho\hat{a}_{t-1}+v_t \tag{12.9.14}$$

其中 $\hat{a}_t$ 为从原（水平值形式）回归中所得到的残差，v_t 为此回归的误差项。注意，在方程（12.9.14）中没有引入截距项的必要，因为我们知道 OLS 残差的总和为零。

方程（12.5.1）中工资—生产率回归的残差已在表 12—5 中给出。利用这些残差，我们可以得到如下回归结果：

$$\begin{aligned}\hat{a}_t &= \quad 0.867\ 8\hat{a}_{t-1}\\ t &= (12.735\ 9) \qquad r^2=0.786\ 3\end{aligned} \tag{12.9.15}$$

如此回归所示，$\hat{\rho}=0.867\ 8$。利用这个估计值，可如同方程（12.9.6）那样对原模型进行变换。由于此程序所估计的 ρ 与从德宾-沃森 d 统计量所得到的 ρ 大致相同，所以利用方程（12.9.15）中的 ρ 进行回归的结果，与利用德宾-沃森 d 统计量估计的 ρ 所得到的回归结果不应该有很大不同。我们留给读者来验证这一点。

估计 ρ 的迭代方法。前面讨论的所有估计 ρ 的方法都只为我们提供了 ρ 的一个估计值。但有些所谓**迭代法**（iterative method）则可多次估计出 ρ 来，即从 ρ 的某个初始值开始，通过逐次逼近，反复估计 ρ 值。这些方法中值得一提的有：**科克伦-奥克特迭代法**（Cochrane-Orcutt iterative procedure）、**科克伦-奥克特两步法**（Cochrane-Orcutt two-step procedure）、**德宾两步法**（Durbin two-step procedure）和**希尔德雷思-卢扫描或搜寻程序**（Hildreth-Lu scanning or search procedure）等。其中，最流行的是科克伦-奥克特迭代法。为节省篇幅，迭代法通过习题加以讨论。记住，这些方法的最终目标是给出 ρ 的一个估计值，使之能用于得到参数的 GLS 估计值。科克伦-奥克特迭代法的优越性之一是，它不仅能用于 AR(1) 模式，也能用于更高阶的自回归模式，比如 AR(2)：$\hat{u}_t=\hat{\rho}_1\hat{u}_{t-1}+\hat{\rho}_2\hat{u}_{t-2}+v_t$。得到两个 ρ 之后，很容易就能推广应用广义差分方程（12.9.6）。当然，计算机可以做到所有这些。

回到工资—生产率回归中来，并假定 AR(1) 模式，我们使用科克伦-奥克特迭代法得到 ρ 的如下估计值：0.887 6、0.994 4 和 0.882 7。最后一个值 0.882 7 现可用于方程（12.9.6）对原模型进行变换，并用 OLS 估计变换后的模型。当然，对变换模型做 OLS 无非就是做 GLS。其结果如下：

Stata 能够估计模型系数与 ρ。例如，如果我们假定 AR(1)，Stata 给出如下结果：

$$\begin{aligned}\hat{Y}_t^* &= 43.104\ 2 + 0.571\ 2X_t \\ \text{se} &= (4.372\ 2)\quad (0.041\ 5) \\ t &= (9.858\ 6)\quad (13.763\ 8)\qquad r^2 = 0.814\ 6\end{aligned} \tag{12.9.16}$$

我们从这些结果中看到，估计的 ρ（即 $\hat{\rho}$）约为 0.882 7，它与方程（12.9.15）中的 $\hat{\rho}$ 没有多大差别。

前面曾指出，由于第一次观测没有前期观测，所以我们在广义差分方程（12.9.6）中丧失一次观测。为了避免这第一次观测的损失，我们可以使用普莱斯-温斯顿变换。利用这一变换和 Stata（第 10 版），我们得到工资一生产率回归的结果：

$$\begin{aligned}\text{Rcompb}_t &= 32.043\ 4 + 0.662\ 8\ \text{Prodb}_t \\ \text{se} &= (3.718\ 2)\quad (0.038\ 6)\qquad r^2 = 0.879\ 9\end{aligned} \tag{12.9.17}$$

在这一变换中，ρ 值约为 0.919 3，经过 13 次迭代而得到。应当指出：如果我们没有对第一次观测进行普莱斯-温斯顿变换而直接把它去掉，那么，尤其在小样本中，有时可能得到极为不同的结果。注意，这里得到的 ρ 与方程（12.9.15）中的 ρ 没有多大差别。

一般评论。用上述各种方法修正自相关，要明确如下几点：

第一，由于在大样本情况下，即便存在自相关问题，OLS 估计量仍是一致的，所

以无论我们是从德宾-沃森 d、从当期残差对前期残差的回归，还是从科克伦-奥克特迭代程序中估计 ρ，都没有多大差别，因为这些方法也都是给出真实 ρ 的一致估计值。

第二，上述方法基本上都是两步法。我们在第一步得到未知 ρ 的一个估计值，第二步用这个估计值变换变量去估计广义差分方程（实质上就是 GLS）。但由于我们用的是 $\hat{\rho}$ 而非真正的 ρ，所以在文献中所有这些估计方法都被称为**可行 GLS**（feasible GLS，FGLS）或**估计 GLS**（estimated GLS，EGLS）。

第三，重要的是要指出，只要我们用 FGLS 或 EGLS 估计变换模型的参数，估计系数都不一定具有通常经典模型所具有的优良性质（比如 BLUE），特别是在小样本情况下。无需复杂的技术性知识，便可总结一个一般原则：只要我们用的是估计量而非真实值，所估计的 OLS 系数在大样本中渐近地具有通常的性质。同时，严格地讲，通常的假设检验程序也是渐近有效的。因此，在小样本下，必须小心地解释估计结果。

第四，在使用 EGLS 时，若不包括第一次观测（就像最初使用科克伦-奥克特程序那样），不仅估计量的数值，就连其有效性都受到不利的影响，特别是当样本容量很小或回归元并非严格随机的情况下。[①] 因此，在小样本中，按照普莱斯-温斯顿的方式保留第一次观测很重要。当然，如果样本容量足够大，EGLS 在有或没有第一次观测的情况下都给出类似结果。顺便指出，以普莱斯-温斯顿方式进行变换的 EGLS 在文献中被称为**完全 EGLS**（full EGLS），或简称为 **FEGLS**。

12.10 修正 OLS 标准误的尼威-威斯特方法

除了使用上一节讨论的 FGLS 方法之外，我们仍可以使用 OLS，只是需要用尼威和威斯特提出的方法对自相关问题修正标准误。[②] 这是对上一章讨论的怀特异方差一致标准误的推广。修正的标准误被称为**异方差—自相关一致标准误**［HAC（heteroscedasticity-and autocorrelation-consistent）standard errors］，或简称**尼威-威斯特标准误**（Newey-West standard errors）。我们不给出尼威-威斯特程序背后复杂的数学知识。[③] 但大多数现代计算机软件现在都能计算尼威-威斯特标准误。但重要的是要指出，尼威-威斯特程序严格地讲只对大样本有效，对小样本可能不合适。但在大样本情形下，我们现在有一种能对自相关修正其标准误的方法，所以就没有必要担心上一节讨论的 EGLS 变换。因此，如果样本足够大，在同时存在自相关和异方差的情况下，就应该使用尼威-威斯特程序来修正 OLS 标准误，因为 HAC 法

① 若回归元如经济数据中极其常见地表现出时间趋势，则尤其如此。

② W. K. Newey and K. West, "A Simple Positive Semi-Definite Heteroscedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, vol. 55, 1987, pp. 703-708.

③ 你若能对付矩阵代数，这种方法在格林的前引文献中有所讨论。

与专为异方差而设计的怀特方法不同，它能同时处理这两种问题。

我们再次回到工资—生产率回归（12.5.1）。我们知道此回归存在自相关的问题。46 次观测样本也足够大，所以我们可用 HAC 程序。我们用 EViews 4 得到如下回归结果：

$$\hat{Y}_t = 32.741\,9 + 0.670\,4X_t$$
$$\text{se} = (2.916\,2)^* \quad (0.030\,2)^* \qquad (12.10.1)$$
$$r^2 = 0.976\,5 \qquad d = 0.171\,9$$

其中 * 表示 HAC 标准误。

将此回归与方程（12.5.1）相比较，我们发现两个方程中的估计系数和 r^2 都相同。但必须注意，HAC 标准误比 OLS 标准误大得多，因此 HAC 的 t 比率比 OLS 的 t 比率小得多。这就表明，OLS 实际上低估了真实标准误。令人惊奇的是，方程（12.5.1）和（12.10.1）中的 d 统计量是一样的。但不必担心，HAC 程序在修正 OLS 标准误时已对此加以考虑。

12.11 OLS 与 FGLS 和 HAC

研究者所面临的实际问题是：在出现自相关问题时，OLS 估计量尽管无偏、一致且渐近正态分布，但仍不是有效的。因此，通常基于 t，F 和 χ^2 的推断程序就不再适合。另一方面，虽然 FGLS 和 HAC 能给出有效的估计量，但这些估计量的有限或小样本性质并没有得到很好地证明。这就意味着，在小样本下，FGLS 和 HAC 实际上可能还不如 OLS。事实上，格里利谢斯（Griliches）和饶（Rao）在一项蒙特卡罗研究中发现[①]，若样本相对较小，且自相关系数 ρ 低于 0.3，则 OLS 至少和 FGLS 一样好。于是，从实践的角度看，在估计的 ρ 低于 0.3 的小样本中，或许应使用 OLS 较好。当然，样本容量是个相对问题，必须从实践中加以判断。如果只有 15～20 次观测，样本就很小，但如果有 50 次以上的观测，样本就足够大了。

12.12 自相关的其他方面

□ 虚拟变量与自相关

我们在第 9 章考虑了虚拟变量回归模型。具体而言，回想我们在方程（9.5.1）

① Z. Griliches, and P. Rao, "Small Sample Properties of Several Two-stage Regression Methods in the Context of Autocorrelated Errors," *Journal of the American Statistical Association*, vol. 64, 1969, pp. 253-272.

中给出的美国 1970—1995 年间储蓄—收入回归模型，为方便起见，复制如下：

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \qquad (12.12.1)$$

其中Y=储蓄；

X=收入；

$D = 1$，1982—1995 年间的观测；

$D = 0$，1970—1981 年间的观测。

基于此模型的回归结果在方程（9.5.4）中给出。当然，此模型是在通常的 OLS 假定下估计的。

但现在假设 u_t 服从一阶自回归 AR(1) 模式，即 $u_t = \rho u_{t-1} + \varepsilon_t$。通常，若 ρ 已知或很容易用上述讨论的方法之一估计出来，我们就能用广义差分方法在没有（一阶）自相关的情况下估计模型参数。但虚拟变量 D 的出现提出了一个特殊问题：注意虚拟变量无非是一次观测被划到第一个或第二个时期。我们如何对此进行变换呢？我们可以按如下程序进行。[①]

1. 在方程（12.12.1）中，第一个时期所有观测的 D 值都是零；第二个时期 D 的**第一个**观测值是 $1/(1-\rho)$ 而不是 1，其余的观测都是 1。

2. 变量 X_t 变换成 $X_t - \rho X_{t-1}$。注意，我们在变换时丢失了一次观测，除非像前面曾指出的那样，用**普莱斯-温斯顿变换**来修复第一次观测。

3. 第一个时期中所有观测的 $D_t X_t$ 值都是 0（注：第一个时期中的 D_t 都是零）；在第二个时期，第一个观测取值 $D_t X_t = X_t$，而其余观测都取值 $D_t X_t - D_t \rho X_{t-1} = X_t - \rho X_{t-1}$。（注：第二个时期中的 D_t 都是 1。）

前面的讨论曾指出，临界观测是第二个时期中的第一次观测。如果像刚才建议的那样考虑到这一点，那么在 AR(1) 自回归约束下估计像方程（12.12.1）这样的回归就不应该有什么问题。在习题 12.37 中，要求读者对第 9 章给出的美国储蓄和收入数据进行这种变换。

□ ARCH 和 GARCH 模型

和在 AR(1) 模式中 t 期的误差项与 $t-1$ 期的误差项相关或 AR(p) 模式中 t 期的误差项与各滞后误差项相关一样，t 期的方差 σ^2 是否也会与其一期或多期滞后值自相关呢？致力于预测股票价格、通货膨胀率和汇率等金融时间序列的研究者已经观察到了这种自相关。若误差方差与前一项误差的平方相关，则有一个吓人的名字**自回归条件异方差**（autoregressive conditional heteroscedasticity，ARCH），若误差方差与过去几期误差项的平方都相关，则被称为**广义自回归条件异方差**（generalized autoregressive conditional heteroscedasticity，GARCH）。由于该专题属于时间序列计量经济学的一般辖域，所以我们将在时间序列计量经济学的有关章节中深入探讨。我们这里的目标只是指出自相关不仅限于当期与过去误差项之间的关系，还包括当

① 见 Maddala，op. cit.，pp. 321-322。

期与过去误差方差之间的相关。

□ 自相关与异方差的共存

如果一个回归模型同时遇到异方差和自相关的问题会怎么样呢？我们能否依次解决问题，即先考虑异方差再考虑自相关呢？事实上，有作者认为“自回归只能在控制了异方差性之后才能侦察出来”[①]。但我们能否给出一种同时解决这两个及其他问题（如模型设定）的万能检验方法呢？回答是肯定的，存在这种检验，但对其讨论离题太远，最好把它们留在参考文献中。[②]不过，正如前面曾指出的那样，如果样本足够大，我们就可以使用 HAC 标准误，因为它们同时考虑了自相关和异方差。

12.13 一个总结性例子

在例 10.2 中，我们给出了美国的消费、收入、财富和利率数据，这些数据都剔除了通货膨胀的影响，是真实项目。基于这些数据，我们估计了美国 1947—2000 年间的消费函数，即将消费的对数对收入和财富的对数进行回归。由于有些真实利率数字是负的，所以我们没有使用利率的对数形式。

Dependent Variable: ln(CONSUMPTION)
Method: Least Squares
Sample: 1947-2000
Included observations: 54

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.467711	0.042778	-10.93343	0.0000
ln(INCOME)	0.804873	0.017498	45.99836	0.0000
ln(WEALTH)	0.201270	0.017593	11.44060	0.0000
INTEREST	-0.002689	0.000762	-3.529265	0.0009

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted R-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	*F*-statistic	37832.59
Sum squared resid.	0.007121	Prob. (*F*-statistic)	0.000000
Log likelihood	164.5880	Durbin-Watson stat.	1.289219

如同所料，收入和财富弹性为正，而利率半弹性为负。尽管各个系数估计值看

① Lois W. Sayrs, *Pooled Time Series Analysis*, Sage Publications, California, 1989, p. 19.

② 参见 Jeffrey M. Wooldridge, op. cit., pp. 402-403，以及 A. K. Bera and C. M. Jarque, “Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence,” *Economic Letters*, vol. 7, 1981, pp. 313-318。

来都是高度统计显著的，但我们还需要检查误差项中存在自相关的可能性。我们知道，在出现自相关的情况下，标准误的估计值可能被低估了。从德宾-沃森 d 统计量看来，消费函数的误差项遇到了（一阶）自相关的问题（请核实）。

为了证实这一判断，我们在容许 AR(1) 自相关的情况下估计了消费函数。结果如下：

```
Dependent Variable: lnCONSUMPTION
Method: Least Squares
Sample (adjusted): 1948-2000
Included observations: 53 after adjustments
Convergence achieved after 11 iterations
```

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.399833	0.070954	-5.635112	0.0000
lnINCOME	0.845854	0.029275	28.89313	0.0000
lnWEALTH	0.159131	0.027462	5.794501	0.0000
INTEREST	0.001214	0.000925	1.312986	0.1954
AR(1)	0.612443	0.100591	6.088462	0.0000

R-squared	0.999688	Mean dependent var.	7.843871
Adjusted R-squared	0.999662	S.D. dependent var.	0.541833
S.E. of regression	0.009954	F-statistic	38503.91
Sum squared resid.	0.004756	Prob. (F-statistic)	0.00000
Log likelihood	171.7381	Durbin-Watson stat.	1.874724

这些结果清楚地表明，我们的回归遇到了自相关问题。利用本章讨论的某种变换消除这种自相关，留给读者自己完成。你可以利用 ρ 的估计值 0.612 4 进行这种变换。下面，我们给出基于考虑了自相关的尼威-威斯特（HAC）标准误的回归结果。

```
Dependent Variable: LCONSUMPTION
Method: Least Squares
Sample: 1947-2000
Included observations: 54
Newey-West HAC Standard Errors & Convariance (lag truncation = 3)
```

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.467714	0.043937	-10.64516	0.0000
LINCOME	0.804871	0.017117	47.02132	0.0000
LWEALTH	0.201272	0.015447	13.02988	0.0000
INTEREST	-0.002689	0.000880	-3.056306	0.0036

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted R-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	F-statistic	37832.71
Sum squared resid.	0.007121	Prob. (F-statistic)	0.000000
		Durbin-Watson stat.	1.289237

上述第一个和最后一个回归的主要区别在于，系数估计值的标准误发生了巨大的变化。尽管如此，斜率系数仍是高度统计显著的。不过，我们并不能保证总是这种情况。

要点与结论

1. 当经典线性回归模型的假定“进入总体回归方程（PRF）的误差项或干扰项 u_t 是随机的或不相关的”不成立时，就有序列相关或自相关的问题。

2. 自相关的出现有种种原因，诸如经济时间序列的惯性或黏滞性，模型遗漏了应包含的重要变量或使用了错误的函数形式所导致的设定偏误、蛛网现象、数据糅合、数据变换，等等。于是，区分纯粹自相关和因刚才提到的一个或多个因素所“引致”的自相关就很重要。

3. 虽然在自相关出现时 OLS 估计量仍是无偏和一致性的，并且是渐近正态分布的，但不再是有效的。结果，常用的显著性 t，F 和 χ^2 检验都不能有效地应用。因此，需要补救措施。

4. 补救措施与干扰项 u_t 中的相依性质有关。而由于这些 u_t 是不可观测的，通常都要假定它们有某种生成机制。

5. 通常假定这种机制是马尔可夫一阶自回归模式，即假定现期的干扰项与前期的干扰项有线性关系，自相关系数 ρ 表示它们之间相互依赖的程度。这种机制被称为 AR(1) 模式。

6. 如果 AR(1) 模式真实且自相关系数已知，则序列相关问题可通过数据变换按照广义差分程序迎刃而解。AR(1) 模式可容易地推广到一个 AR(p) 模式上，还可假定一个移动平均（MA）机制或 AR 与 MA 两种模式的一个混合，叫做自回归移动平均（ARMA）。这个专题将在时间序列计量经济学的有关章节中讨论。

7. 即使在使用一个 AR(1) 模式时，我们也不能先验知道自相关系数 ρ。我们考虑了估计 ρ 的几种方法如：德宾-沃森 d，瑟尔-纳加修正 d，科克伦-奥克特迭代程序，科克伦-奥克特两步程序，以及德宾两步法。在大样本中，这些方法一般地说会得到类似的估计值；而在小样本中，它们表现各不相同。在实践中，科克伦-奥克特迭代法用得很普遍。

8. 使用刚刚讨论过的任何一种方法，我们都能通过 OLS 用广义差分方法估计变换模型的参数，这种方法实质上就是 GLS。但由于我们用了估计的 $\rho(=\hat{\rho})$，所以我们把这种估计方法称为可行或估计 GLS，简记为 FGLS 或 EGLS。

9. 在使用 EGLS 时，去掉第一个观测时必须小心，因为在小样本情形下，保留还是去掉第一个观测对结果有显著影响。因此，在样本容量很小时，建议按照普莱斯-温斯顿程序对第一个观测进行变换。但对大样本而言，是否包括第一个观测没什么差别。

10. 必须强调指出，EGLS 方法只有在大样本条件下才具有通常的优良统计性质。在小样本情况下，OLS 实际上可能比 EGLS 更好，特别是在 $\rho < 0.3$ 时。

11. 我们也可以不用 EGLS，而使用 OLS，只是要用尼威-威斯特（HAC）程序对自相关问题修正其标准误。严格地讲，这一程序只有在大样本条件下才有效。HAC 程序的优势之一是，它不仅修正了自相关，还在出现异方差时修正了异方差。

12. 当然，自相关的侦察工作要先于补救措施。侦察的方法有正式和非正式两种。非正式方法可以对实际或标准化的残差描点，或者将当期残差对历史残差描点。正式的方法，可以使用游程

检验、德宾-沃森 d 检验、渐近正态检验、贝伦布鲁特-韦布检验和布罗施-戈弗雷（BG）检验。其中，最流行而又惯常使用的是德宾-沃森 d 检验，尽管它有荣耀的历史，但仍有几方面的局限性。由于BG检验同时容许AR和MA误差结构，以及容许回归子的滞后值作为解释变量而出现，所以最好使用更一般性的BG检验。但必须牢记，它仍是一个大样本检验。

13. 我们在本章还十分简要地讨论了出现虚拟变量时自相关的侦察。

习　　题

问答题

12.1　判明以下陈述的真伪，简单地陈述你的理由。

a. 当出现自相关时，OLS估计量是偏误的和非有效的。

b. 德宾-沃森 d 检验假定误差项 u_t 的方差有同方差性。

c. 用一阶差分变换消除自相关的方法是假定自相关系数 ρ 为 -1。

d. 如果一个是一阶差分形式的回归，而另一个是水平值形式的回归，那么，这两个模型的 R^2 值是不可直接比较的。

e. 一个显著的德宾-沃森 d 统计量不一定意味着一阶自相关。

f. 在出现自相关时，通常计算的预报值的方差和标准误就不是有效的。

g. 把一个（或多个）重要的变量从回归模型排除出去可能导致一个显著的 d 值。

h. 在AR(1)模式中，假设 $\rho=1$ 既可通过贝伦布鲁特-韦布 g 统计量也可通过德宾-沃森 d 统计量来检验。

i. 如果在 Y 的一阶差分对 X 的一阶差分的回归中有一常数项和一线性趋势项，就意味着在原始模型中有一个线性项和一个二次趋势项。

12.2　给定一个含有50次观测的样本和4个解释变量，在下列情况下你能对自相关的问题说些什么？

a. $d=1.05$。

b. $d=1.40$。

c. $d=2.50$。

d. $d=3.97$。

12.3　在研究生产中的劳动在附加值（value added）中所占份额（即劳动份额）的变动时，古扎拉蒂（Gujarati）考虑如下模型[①]：

模型A：$Y_t=\beta_0+\beta_1 t+u_t$

模型B：$Y_t=\alpha_0+\alpha_1 t+\alpha_2 t^2+u_t$

其中 Y=劳动份额，t=时间。根据1949—1964年数据，对初级金属工业部门得到如下结果：

模型A：$\hat{Y}_t=0.4529-0.0041\,t$　　$R^2=0.5284$　　$d=0.8252$

$$(-3.9608)$$

① Damodar Gujarati, "Labor's Share in Manufacturing Industries," *Industrial and Labor Relations Review*, vol. 23, no. 1, October 1969, pp. 65-75.

模型 B: $\hat{Y}_t = 0.478\,6 - 0.012\,7\,t + 0.000\,5\,t^2$　　$R^2 = 0.662\,9$　　$d = 1.82$

$(-3.272\,4)$　$(2.777\,7)$

其中括号中的数字是 t 比率。

a. 模型 A 中有没有序列相关？模型 B 呢？

b. 怎样证明序列相关的存在？

c. 你会怎样区分“纯粹”自相关和设定偏误？

12.4　**侦察自相关：冯·诺伊曼检验。**[①] 假定残差 a_i 是从正态分布随机抽取的，冯·诺伊曼曾证明，对于大的 n，比率

$$\frac{\delta^2}{s^2} = \frac{\sum (a_i - a_{i-1})^2/(n-1)}{\sum (a_i - \bar{a})^2/n}$$

注：OLS 的 $\bar{a} = 0$。

被称为冯·诺伊曼比率（von Neumann ratio），近似于正态分布，其均值为：

$$E\frac{\delta^2}{s^2} = \frac{2n}{n-1}$$

而方差为：

$$\operatorname{var}\frac{\delta^2}{s^2} = 4n^2\,\frac{n-2}{(n+1)\,(n-1)^3}$$

a. 如果 n 足够大，你会怎样利用冯·诺伊曼比率来检验自相关？

b. 德宾-沃森 d 和此比率有什么关系？

c. d 统计量落在 0 与 4 之间。冯·诺伊曼的相应界限是什么？

d. 此比率依赖于“a 是从正态分布中随机抽取的”这个假定，对 OLS 残差来说，这一假定的真实性如何？

e. 假使在观测次数为 100 的一项应用中发现此比率为 2.88。检验数据中无序列相关的假设。

注：哈特（B. I. Hart）曾对多至 60 次观测的样本容量编制了冯·诺伊曼比率的临界值表。[②]

12.5　在 17 个残差的一个排序中有 11 个正值和 6 个负值。游程个数是 3。这是否表明有自相关的迹象？如果游程个数是 14，你会改变答案吗？

12.6　**根据 d 统计量的瑟尔-纳加 ρ 估计。**瑟尔和纳加曾建议，在小样本中不把 ρ 估计为 $(1-d/2)$，而把它估计为：

$$\hat{\rho} = \frac{n^2(1-d/2) + k^2}{n^2 - k^2}$$

其中 n=观测总个数，d=德宾-沃森 d，及 k=待估系数个数（包括截距）。

证明对于大的 n，ρ 的这个估计值将化为较简单的公式 $(1-d/2)$。

12.7　**估计希尔德雷思-卢扫描或搜寻程序。**[③] 由于在下列一阶自回归模式中

$$u_t = \rho u_{t-1} + \varepsilon_t$$

预期 ρ 落在 -1 与 $+1$ 之间，为确定它的位置，希尔德雷思和卢提出一种系统的“扫描”或搜寻程序。他们建议在 -1 与 $+1$ 之间按一定的间隔，比方说，每隔 0.1 单位试选 ρ 值，并通过广义差分方程（12.6.5）对数据做变换。即把 ρ 选为 -0.9，-0.8，…，0.8，0.9。对每一选取的 ρ 值，做

---

① J. von Neumann, “Distribution of the Ratio of the Mean Square Successive Difference to the Variance,” *Annals of Mathematical Statistics*, vol. 12, 1941, pp. 367-395.

② 此表可见于 Johnston, op. cit., 3d ed., p. 559。

③ G. Hildreth and J. Y. Lu, “Demand Relations with Autocorrelated Disturbances,” Michigan State University, *Agricultural Experiment Station*, Tech. Bull. 276, November 1960.

一个广义差分回归并得到相应的 RSS：$\sum a_t^2$。希尔德雷思和卢建议最后选择使 RSS 最小（从而使 R^2 最大）的 ρ 值。如果需要更精细的结果，则还可采用更小的单位间隔，如每隔 0.01 单位，把 ρ 取为 -0.99，-0.98，…，0.90，0.91，等等。

a. 希尔德雷思-卢程序有何优越性？

b. 怎样知道最后选取的 ρ 值所做的数据转换事实上能保证 $\sum a_t^2$ 最小？

12.8 **估计 ρ：科克伦-奥克特迭代程序。**[1] 作为对此程序的一个说明，考虑双变量模型：

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{1}$$

及 AR(1) 模式

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{2}$$

于是科克伦和奥克特推荐如下步骤来估计 ρ。

1. 用通常的 OLS 方法估计方程（1）并得到残差 a_t。顺便指出，你可以在模型中包含不止一个 X 变量。

2. 利用第 1 步得到的残差做如下回归：

$$a_t = \hat{\rho} a_{t-1} + v_t \tag{3}$$

这是方程（2）在实证中的对应表达式。[2]

3. 利用方程（3）中得到的 $\hat{\rho}$，估计广义差分方程（12.9.6）。

4. 由于事先不知道方程（3）中得到的 $\hat{\rho}$ 是不是 ρ 的最佳估计值，所以把第 3 步中得到的 $\hat{\beta}_1^*$ 和 $\hat{\beta}_2^*$ 值代入原回归（1），并得到新的残差 a_t^* 为

$$a_t^* = Y_t - \hat{\beta}_1^* - \hat{\beta}_2^* X_t \tag{4}$$

由于 Y_t，X_t，$\hat{\beta}_1^*$ 和 $\hat{\beta}_2^*$ 皆已知，故很容易计算出来。

5. 现在估计如下回归

$$a_t^* = \hat{\rho}^* a_{t-1}^* + w_t \tag{5}$$

它类似于方程（3），并给出 ρ 的第二轮估计值。由于我们不知道 ρ 的第二轮估计值是不是真实 ρ 的最佳估计值，所以我们进入第三轮估计，如此等等。这正是科克伦-奥克特程序被称为迭代程序的原因。我们该把这种（愉快的）轮回操作进行到什么程度呢？一般的建议是，当 ρ 的两个相邻估计值相差很小（比如不足 0.01 或 0.005）时，便可停止迭代。在工资—生产率一例中，在停止之前约需要 3 次迭代。

a. 利用科克伦-奥克特迭代程序，估计工资—生产率回归（12.5.2）的 ρ。在得到 ρ 的“最终”估计值之前需要多少次迭代？

b. 利用（a）中得到的 ρ 的最终估计值，在去掉第一次观测和保留第一次观测的情况下，估计工资—生产率回归。结果有何差异？

c. 你认为在变换数据以解决自相关问题时保留第一次观测重要吗？

12.9 **估计 ρ：科克伦-奥克特两步法。**这是科克伦-奥克特迭代程序的一个简化版。第一步，我们从第一次迭代［即上一题中的方程（3）］中估计出 ρ，第二步就用所估计的 ρ 做上一题方程

---

① D. Cochrane and G. H. Orcutt, “Applications of Least-Squares Regressions to Relationships Containing Autocorrelated Error Terms,” *Journal of the American Statistical Association*, vol. 44, 1949, pp. 32-61.

② 注意 $\hat{\rho} = \sum a_t a_{t-1} / \sum a_t^2$。（为什么？）尽管有偏，但 $\hat{\rho}$ 仍是真实 ρ 的一个一致估计量。

(4) 中的广义差分方程。实践中，有时候这种两步法给出的结果，与多次煞费苦心的迭代所得到的结果十分相似。

在正文中给出的说明性工资—生产率回归（12.5.1）中使用科克伦-奥克特两步法，并将结果与迭代法所得到的结果进行比较。特别注意在变换中对第一次观测的处理。

12.10 **估计 ρ：德宾两步法。**[①] 为了解释此方法，我们可以把广义差分方程（12.9.5）等价地写成

$$Y_t = \beta_1(1-\rho) + \beta_2 X_t - \beta_2 \rho X_{t-1} + \rho Y_{t-1} + \varepsilon_t \tag{1}$$

德宾建议用如下两步法估计 ρ：第一步，把方程（1）作为一个多元回归模型将 Y_t 对 X_t，X_{t-1} 和 Y_{t-1} 做回归，并把 Y_{t-1} 回归系数的估计值（$=\hat{\rho}$）作为对 ρ 的一个估计值。第二步，得到 $\hat{\rho}$ 之后，用它估计广义差分方程（12.9.5）或与之等价的方程（12.9.6）中的参数。

a. 在正文中所讨论的工资—生产率例子中应用德宾两步法，并把结果与科克伦-奥克特迭代程序和科克伦-奥克特两步法所得到的结果相比较。评论你所得到结果的“质量”。

b. 若检查上述方程（1），你会观察到 X_{t-1} 的系数等于 X_t 的系数（$=\beta_2$）和 Y_{t-1} 的系数（$=\rho$）之积的-1倍。你如何对这种系数约束进行检验？

12.11 在测量电力供给中的规模报酬时，纳洛夫（Nerlove）利用 1955 年美国 145 个私有公用事业公司的横截面数据，做了对数总成本对对数产出、对数工资、对数资本价格和对数燃料价格的回归。他发现由此回归估计得的残差从德宾-沃森 d 看来呈现“序列”相关。为了寻求补救方法，他将所估计的残差相对于对数产出描图而得到图 12—11。

a. 图 12—11 说明什么？

b. 在上述情况下，你怎样摆脱“序列”相关问题？

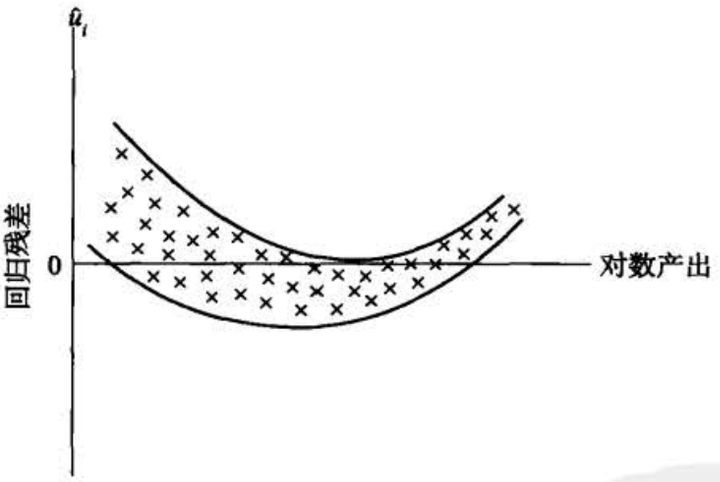


图 12—11 来自纳洛夫研究的残差

资料来源：Marc Nerlove，“Return to Scale in Electric Supply，” in Carl F. Christ et al. *Measurement in Economics*，Stanford University Press，Stanford，Calif.，1963.

12.12 将一个回归的残差对时间描图，得到图 12—12 中的散点图。打上圆圈的一个“极端”残差表示一个异常值（outlier）。所谓异常值，是指远远超出样本中其他观测值的一个观测值，也

① J. Durbin，“Estimation of Parameters in Time-Series Regression Models，” *Journal of the Royal Statistical Society*，series B，vol. 22，1960，pp. 139-153.

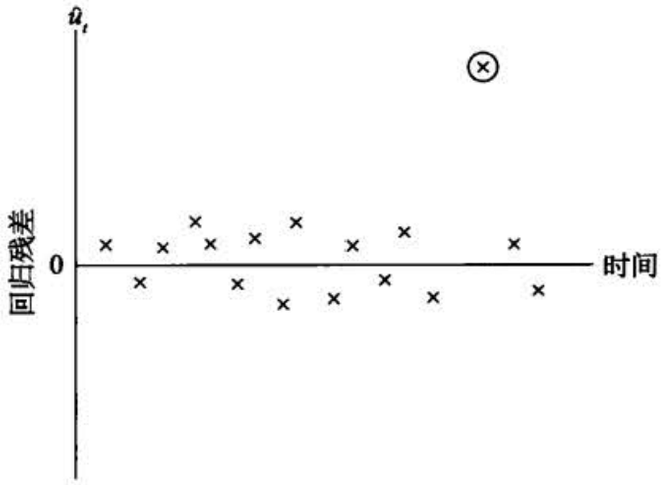


图 12—12　假想回归残差对时间描点

许离开所有观测值的平均值 3～4 个标准差之多。

a. 出现异常值的原因是什么？

b. 如果有异常值（一个或多个），是否应把它（们）剔除，然后再对其余的观测值做回归？

c. 在出现异常值的情况下，德宾-沃森 d 是否适用？

12.13　根据德宾-沃森 d 统计量，你怎样区分“纯粹”自相关和设定偏误？

12.14　假设在下列模型中

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

u_t 确实序列无关。那么，如果我们假定 $u_t = \rho u_{t-1} + \varepsilon_t$，并使用了广义差分回归

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2 X_t - \rho\beta_2 X_{t-1} + \varepsilon_t$$

将会出现什么情况？特别讨论干扰项 ε_t 的性质。

12.15　在对英国如何按要素成本对最终产品定价的一项研究中，根据 1951—1969 年间的年度数据，得到如下结果：

$$\widehat{PF}_t = 2.033 + 0.273W_t - 0.521X_t + 0.256M_t + 0.028M_{t-1} + 0.121PF_{t-1}$$
$$\text{se}=(0.992)\quad(0.127)\quad(0.099)\quad(0.024)\quad(0.039)\quad(0.119)$$
$$R^2 = 0.984 \quad d = 2.54$$

其中 PF＝按要素成本定价的最终产品价格，W ＝ 每个雇员的工薪，X ＝ 每个就业人员的国内总产值，M＝进口价格，M_{t-1} ＝ 滞后一年的进口价格，以及 PF_{t-1} ＝ 前一年按要素成本定价的最终产品价格。①

“因对 18 个观测值和 5 个解释变量，d 值的 5%下限和上限分别是 0.71 和 2.06，而估计的 d 值是 2.54，故表明无正的自相关。”试评论。

12.16　在什么情况下以下各种估计一阶自相关系数 ρ 的方法是合适的：

a. 一阶差分回归。

b. 移动平均回归。

c. 瑟尔-纳加变换。

① 资料来源：*Prices and Earnings in 1951—1969: An Econometric Assessment*, Department of Employment, Her Majesty's Stationery Office, 1971, Table C, p. 37, Eq. 63。

d. 科克伦-奥克特迭代程序。

e. 希尔德雷思-卢搜寻程序。

f. 德宾两步法。

12.17 考虑模型

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

其中

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$$

即误差项服从 AR(2) 模式，其中 ε_t 为白噪音误差项。在考虑二阶自回归情况下，勾勒出估计此模型的步骤。

12.18 把校正因子 C 包括进来，方程（12.3.1）所给的 $\hat{\beta}_2^{GLS}$ 的计算公式是：

$$\hat{\beta}_2^{GLS} = \frac{(1-\rho^2)x_1 y_1 + \sum_{t=2}^{n}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{(1-\rho^2)x_1^2 + \sum_{t=2}^{n}(x_t - \rho x_{t-1})^2}$$

给定此公式和方程（12.3.1），求校正因子 C 的表达式。

12.19 证明估计方程（12.9.5）等价于估计 12.3 节所讨论的不包含对 Y 和 X 的第一次观测的 GLS。

12.20 对于回归（12.9.9），估计的残差值有如下符号，其中每一游程都用括号包住了：

(＋＋＋＋)(－)(＋＋＋＋＋＋＋＋)(－)(＋＋＋＋)(－－) (＋)(－－)(＋)(－－)(＋＋)
(－)(＋)(－－－－－－－－－－)(＋)

说明根据游程检验可以拒绝在这些残差中没有自相关的虚拟假设吗？

*12.21 检验高阶序列相关。假使我们拥有时间序列的季度数据，在涉及季度数据的回归中，更合适的假定将不是方程（12.2.1）所给的 AR(1) 模式，而是如下 AR(4) 模式：

$$u_t = \rho_4 u_{t-4} + \varepsilon_t$$

也就是说，假定当前的干扰项与上年同季度的干扰项相关，而不是与上季度的干扰项相关。

为检验 $\rho_4 = 0$ 的假设，沃利斯（Wallis）[1] 建议做以下的修正德宾-沃森 d 检验：

$$d_4 = \frac{\sum_{t=5}^{n}(\hat{u}_t - \hat{u}_{t-4})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

检验程序与正文中讨论的通常 d 检验步骤相仿。沃利斯曾编制有 d_4 表，可在他的原始论文中找到。

现在假设我们拥有每月数据，能否将德宾-沃森检验加以推广，以适合于这种数据？如果能够，试写出适当的 d_{12} 公式。

12.22 假使你估计以下回归：

$$\Delta \ln Y_t = \beta_1 + \beta_2 \Delta \ln L_t + \beta_3 \Delta \ln K_t + u_t$$

其中 Y = 产出，L = 劳动投入，K = 资本投入，而 Δ 为一阶差分运算子。你怎样解释此模型中的 β_1？可否把它解释为技术变化的一个估计值？说明你的理由。

12.23 正文中曾指出，曼德拉曾建议在德宾-沃森 d 小于 R^2 时应该做一阶差分形式的回归。其背后的逻辑根据是什么？

12.24 参照方程（12.4.1）。假定 $r=0$ 但 $\rho \neq 0$。若（a）$0<\rho<1$，（b）$-1<\rho<0$，其对 $E(\hat{\sigma}^2)$ 有何影响？$\hat{\sigma}^2$ 的偏误何时充分地小？

---

① Kenneth Wallis, "Testing for Fourth Order Autocorrelation in Quarterly Regression Equations," *Econometrica*, vol. 40, 1972, pp. 617-636. d_4 表也可见 Johnston, op. cit., 3d ed., p. 558。

12.25 将方程（12.5.2）给出的工资—生产率回归所得到的残差对其过去 6 期滞后残差［即 AR(6)］做回归，得到如下结果：

Dependent Variable: S1
Method: Least Squares
Sample (adjusted): 1966-2005
Included observations: 40 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
S1(-1)	1.019716	0.170999	5.963275	0.0000
S1(-2)	-0.029679	0.244152	-0.121560	0.9040
S1(-3)	-0.286782	0.241975	-1.185171	0.2442
S1(-4)	0.149212	0.242076	0.616386	0.5417
S1(-5)	-0.071371	0.243386	-0.293240	0.7711
S1(-6)	0.034362	0.167077	0.205663	0.8383

R-squared	0.749857	Mean dependent var.	0.004433
Adjusted R-squared	0.713071	S.D. dependent var.	0.019843
S.E. of regression	0.010629	Durbin-Waston stat.	1.956818
Sum squared resid.	0.003841		

a. 你从上述结果能得出对数工资—生产率数据中自相关的什么性质？

b. 你若认为 AR(1) 机制刻画了数据中的自相关，你会用一阶差分变换处理自相关吗？说明你的理由。

实证分析题

12.26 参考表 12—7 中的铜业数据。

表 12—7　　1951—1980 年美国国内铜价格的决定因素

年份	C	G	I	L	H	A
1951	21.89	330.2	45.1	220.4	1 491.0	19.00
1952	22.29	347.2	50.9	259.5	1 504.0	19.41
1953	19.63	366.1	53.3	256.3	1 438.0	20.93
1954	22.85	366.3	53.6	249.3	1 551.0	21.78
1955	33.77	399.3	54.6	352.3	1 646.0	23.68
1956	39.18	420.7	61.1	329.1	1 349.0	26.01
1957	30.58	442.0	61.9	219.6	1 224.0	27.52
1958	26.30	447.0	57.9	234.8	1 382.0	26.89
1959	30.70	483.0	64.8	237.4	1 553.7	26.85
1960	32.10	506.0	66.2	245.8	1 296.1	27.23
1961	30.00	523.3	66.7	229.2	1 365.0	25.46
1962	30.80	563.8	72.2	233.9	1 492.5	23.88
1963	30.80	594.7	76.5	234.2	1 634.9	22.62
1964	32.60	635.7	81.7	347.0	1 561.0	23.72
1965	35.40	688.1	89.8	468.1	1 509.7	24.50
1966	36.60	753.0	97.8	555.0	1 195.8	24.50
1967	38.60	796.3	100.0	418.0	1 321.9	24.98
1968	42.20	868.5	106.3	525.2	1 545.4	25.58

续前表

年份	C	G	I	L	H	A
1969	47.90	935.5	111.1	620.7	1 499.5	27.18
1970	58.20	982.4	107.8	588.6	1 469.0	28.72
1971	52.00	1 063.4	109.6	444.4	2 084.5	29.00
1972	51.20	1 171.1	119.7	427.8	2 378.5	26.67
1973	59.50	1 306.6	129.8	727.1	2 057.5	25.33
1974	77.30	1 412.9	129.3	877.6	1 352.5	34.06
1975	64.20	1 528.8	117.8	556.6	1 171.4	39.79
1976	69.60	1 700.1	129.8	780.6	1 547.6	44.49
1977	66.80	1 887.2	137.1	750.7	1 989.8	51.23
1978	66.50	2 127.6	145.2	709.8	2 023.3	54.42
1979	98.30	2 628.8	152.5	935.7	1 749.2	61.01
1980	101.40	2 633.1	147.1	940.9	1 298.5	70.87

注：数据来自 *American Metal Market*，*Metals Week* 以及美国商务部出版物，由加里·史密斯（Gary R. Smith）收集。

C=12 个月的平均美国国内钢价（美分/磅）。

G=年度国内生产总值（十亿美元）。

I=12 个月的平均工业生产指数。

L=12 个月的平均伦敦金属交易所钢价（英镑）。

H=每年新房动工数（千套）。

A=12 个月的平均铝价（美分/磅）。

a. 根据这些数据，估计以下回归模型：

$$\ln C_t = \beta_1 + \beta_2 \ln I_t + \beta_3 \ln L_t + \beta_4 \ln H_t + \beta_5 \ln A_t + u_t$$

并解释所得结果。

b. 求出上述回归的残差和标准化残差并作图。你能对这些残差中是否有自回归做些什么猜测？

c. 估计德宾-沃森 d 统计量并对数据中可能出现的自相关性质作出评论。

d. 做游程检验，看你的答案是否不同于刚才在（c）中所得到的结果。

e. 你怎样辨别 AR(p) 过程是否比 AR(1) 过程更好地描述自相关？

注：保留数据作进一步的分析。（见习题 12.28。）

12.27 给定表 12—8 中的数据：

表 12—8

个人消费支出，Y（1958 年十亿美元）	时间，X	Y 的估计值 $\hat{Y}$	残差，$\hat{u}$
281.4	1（=1956）	261.420 8	19.979 1
288.1	2	276.602 6	11.497 3
290.0	3	291.784 4	−1.784 4
307.3	4	306.966 1	0.333 8
316.1	5	322.147 9	−6.047 9
322.5	6	337.329 7	−14.829 7
338.4	7	352.511 5	−14.111 5
353.3	8	367.693 3	−14.393 3

续前表

个人消费支出，Y（1958年十亿美元）	时间，X	Y的估计值$\hat{Y}$	残差，a
373.7	9	382.875 1	−9.175 1
397.7	10	398.056 9	−0.356 9
418.1	11	413.238 6	4.861 3
430.1	12	428.420 6	1.679 5
452.7	13	443.602 2	9.097 7
469.1	14	458.784 0	10.315 9
476.9	15（=1970）	473.965 8	2.934 1

注：$\hat{Y}$数据估计自回归$Y_t=\beta_0+\beta_1 X_t+u_t$。

a. 验证德宾-沃森$d=0.414\ 8$。

b. 干扰项中有没有正的序列相关？

c. 如有，则分别利用下列方法估计ρ：

ⅰ. 瑟尔-纳加方法

ⅱ. 德宾两步程序

ⅲ. 科克伦-奥克特法

d. 利用瑟尔-纳加法变换数据并对变换后的数据做回归。

e.（d）中估计的回归仍显示有自相关吗？如有，如何能将它除去？

12.28 参照习题12.26以及表12—7中的数据。如果该题的结果表明存在序列相关：

a. 用科克伦-奥克特两步程序，估计可行GLS或广义差分回归，并比较你所得的结果。

b. 如果估计自（a）的科克伦-奥克特法的ρ值和从d统计量估计得的结果相差较大，你将选择哪一个估计ρ的方法，为什么？

12.29 参照例7.4。略去变量X^2和X^3，再做回归并检查残差中的“序列”相关。如果发现有序列相关，你会怎样解释它？你会提出什么补救措施？

12.30 回到习题7.21。可以先验地预期在这样的数据中有自相关。因此，建议你以一阶差分形式做真实货币供给的对数对真实国民收入的对数和长期利率的回归。计算此回归并重做原始形式的回归。一阶差分变换所依据的假定是否得到满足？如果不，这种变换很可能导致什么样的偏误？利用现有的数据加以说明。

12.31 **德宾-沃森d用于检验非线性问题。**继续考虑习题12.29。将得自那个回归的残差按X值的递增次序排列。利用方程（12.6.5）所给的公式，从重新排列的残差中估计d。如果计算出来的d值表明有自相关，这将意味着线性模型是不正确的。完整的模型应包含X_i^2和X_i^3项。你能给出这一程序的直觉理由吗？看看你的答案是否同瑟尔所给的相一致。[①]

12.32 回到习题11.22。求出残差并探明残差中是否有自相关。如果发现有序列相关，你会怎样变换数据？在本例中序列相关的意义是什么？

12.33 **蒙特卡罗实验。**参照表12—1和表12—2。利用那里给的ϵ_t和X_t数据，按下列模型生成10个Y值的一个样本：

$$Y_t=3.0+0.5X_t+u_t$$

其中$u_t=0.9\ u_{t-1}+\epsilon_t$。假定$u_0=10$。

① Henri Theil, *Introduction to Econometrics*, Prentice Hall, Englewood Cliffs, NJ, 1978, pp. 307-308.

a. 估计这个方程并评论你的结果。

b. 现在假定 $u_0=17$。重复这个练习 10 次，再评论所得结果。

c. 保持上述结构不变。仅将 $\rho=0.9$ 改为 $\rho=0.3$。然后将你的结果同（b）中所给的结果相比较。

12.34 利用表 12—9 所给数据，估计模型

$$Y_t=\beta_1+\beta_2 X_t+u_t$$

其中 Y=存货，X=销售量，均以十亿美元计。

a. 估计上述回归。

b. 利用（i）德宾-沃森检验和（ii）方程（12.6.13）所给的大样本正态性检验，从估计的残差中探明是否有正的自相关。

c. 如果 ρ 是正的，利用贝伦布鲁特-韦布检验去检验假设 $\rho=1$。

d. 如果你猜测自回归误差结构的阶数是 p，可用布罗施-戈弗雷检验去证实这一点。你会怎样选择阶数 p 呢？

e. 根据此检验的结果，你会怎样转换数据从而把自回归除掉？说明你的全部计算。

f. 重复前面的步骤，但用以下模型：

$$\ln Y_t=\beta_1+\beta_2\ln X_t+u_t$$

g. 你在线性与对数线性两种设定之间如何取舍？说明你的检验方法。

表 12—9　　1950—1991 年美国制造业的存货与销售　　（单位：百万美元）

年份	销售*	存货†	比率	年份	销售*	存货†	比率
1950	46 486	84 646	1.82	1971	224 619	369 374	1.57
1951	50 229	90 560	1.80	1972	236 698	391 212	1.63
1952	53 501	98 145	1.83	1973	242 686	405 073	1.65
1953	52 805	101 599	1.92	1974	239 847	390 950	1.65
1954	55 906	102 567	1.83	1975	250 394	382 510	1.54
1955	63 027	108 121	1.72	1976	242 002	378 762	1.57
1956	72 931	124 499	1.71	1977	251 708	379 706	1.50
1957	84 790	157 625	1.86	1978	269 843	399 970	1.44
1958	86 589	159 708	1.84	1979	289 973	424 843	1.44
1959	98 797	174 636	1.77	1980	299 766	430 518	1.43
1960	113 201	188 378	1.66	1981	319 558	443 622	1.37
1961	126 905	211 691	1.67	1982	324 984	449 083	1.38
1962	143 936	242 157	1.68	1983	335 991	463 563	1.35
1963	154 391	265 215	1.72	1984	350 715	481 633	1.35
1964	168 129	283 413	1.69	1985	330 875	428 108	1.38
1965	163 351	311 852	1.95	1986	326 227	423 082	1.29
1966	172 547	312 379	1.78	1987	334 616	408 226	1.24
1967	190 682	339 516	1.73	1988	359 081	439 821	1.18
1968	194 538	334 749	1.73	1989	394 615	479 106	1.17
1969	194 657	322 654	1.68	1990	411 663	509 902	1.21
1970	206 326	338 109	1.59				

注：* 年度数据为未经季节性调整月度数据的平均。

†从 1982 年开始的数据都是经季节性调整的期末数字，和早期的数字不具可比性。

资料来源：*Economic Report of the President*，1993，Table B-53，p. 408.

12.35　表 12—10 给出了美国 1954—1981 年间普通股即期（t 期）真实回报率（RR_t）、下期（$t+1$ 期）产出增长率（OG_{t+1}）和第 t 期通货膨胀率（Inf_t）的数据，都以百分数表示。

a. 将 RR_t 对 Inf_t 做回归。

b. 将 RR_t 对 OG_{t+1} 和 Inf_t 做回归。

c. 尤金·法马（Eugene Fama）观察到："真实股票回报与通货膨胀之间简单的负相关关系是荒谬的，因为它是两个结构性关系的结果：一个是股票即期真实回报率与预期产出增长率［由 OG_{t+1} 度量］之间的正相关关系，一个是预期产出增长率与即期通货膨胀率之间的负相关关系。"借用这种观点，评论上述两个回归结果。

d. 你预计在回归（a）和（b）中会出现自相关吗？为什么？若你认为会出现，则采取适当的修正措施并给出修正后的结果。

表 12—10　　1954—1981 年间美国真实回报率、产出增长率与通货膨胀率数据

年份	真实回报率	产出增长率	通货膨胀率
1954	53.0	6.7	−0.4
1955	31.2	2.1	0.4
1956	3.7	1.8	2.9
1957	−13.8	−0.4	3.0
1958	41.7	6.0	1.7
1959	10.5	2.1	1.5
1960	−1.3	2.6	1.8
1961	26.1	5.8	0.8
1962	−10.5	4.0	1.8
1963	21.2	5.3	1.6
1964	15.5	6.0	1.0
1965	10.2	6.0	2.3
1966	−13.3	2.7	3.2
1967	21.3	4.6	2.7
1968	6.8	2.8	4.3
1969	−13.5	−0.2	5.0
1970	−0.4	3.4	4.4
1971	10.5	5.7	3.8
1972	15.4	5.8	3.6
1973	−22.6	−0.6	7.9
1974	−37.3	−1.2	10.8
1975	31.2	5.4	6.0
1976	19.1	5.5	4.7
1977	−13.1	5.0	5.9
1978	−1.3	2.8	7.9
1979	8.6	−0.3	9.8
1980	−22.2	2.6	10.2
1981	−12.2	−1.9	7.3

12.36　**德宾 h 统计量**。考虑如下工资决定模型：

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$$

其中 Y=工资=真实小时工资指数；

X=生产率=每小时产出指数。

a. 利用表 12—4 中的数据，估计上述模型并解释你的结论。

b. 由于模型把滞后回归子作为一个回归元包括进来，所以不适合用德宾-沃森 d 来查明数据中是否存在序列相关。对这种所谓自回归模型，德宾曾提出所谓的 **h 统计量**（h statistic）来检验一阶自相关，其定义为[①]：

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n[\text{var}\ (\hat{\beta}_3)]}}$$

其中 n 为样本容量，var $(\hat{\beta}_3)$ 为滞后的 Y_{t-1} 系数的方差，$\hat{\rho}$ 为一阶序列相关的估计值。

在大样本情形（技术上讲，渐近情形）下，德宾已证明，在 $\rho = 0$ 的虚拟假设下，

$$h \sim N(0,1)$$

即 h 统计量服从标准正态分布。我们从正态分布的性质知道，$|h| > 1.96$ 的概率约为 5%。因此，若在一个应用中 $|h| > 1.96$，那我们就可以拒绝 $\rho = 0$ 的虚拟假设，即上述自回归模型中有一阶自相关的证据。

此检验可如下进行：第一步，用 OLS 估计上述模型（目前阶段不必担心任何估计问题）。第二步，记下此模型中的 var $(\hat{\beta}_3)$ 和例行计算出来的 d 统计量。第三步，利用 d 值得到 $\hat{\rho} \approx (1 - d/2)$。注意，有趣的是，尽管我们不能用 d 值来检验此模型中的序列相关，但我们可用它得到 ρ 的一个估计值。第四步，现在可以计算 h 统计量。第五步，若样本容量足够大，且计算的 $|h|$ 超过了 1.96，那我们就可以断定有存在一阶自相关的证据。当然，你可以选用你想选用的任何显著性水平。

在前面给出的自回归工资决定模型中应用 h 检验，并得出适当结论，然后与回归（12.5.1）中所给出的结论相比较。

12.37 **虚拟变量和自相关。**参照第 9 章中讨论的储蓄—收入回归。利用表 9—2 中给出的数据，并假定 AR(1) 模式，在考虑自相关的情况下，重新估计储蓄—收入回归。特别注意对虚拟变量的变换。将结论与第 9 章给出的结论相比较。

12.38 利用表 12—4 中给出的工资—生产率数据，估计模型（12.9.8），并将结论与回归（12.9.9）中给出的结论相比较。你能得到什么结论？

附录 12A

□ 12A.1 方程（12.1.11）中误差项 v_t 自相关的证明

由于 $v_t = u_t - u_{t-1}$，且对每个 t 都有 $E(u) = 0$，所以很容易证明 $E(v_t) = E(u_t - u_{t-1}) = E(u_t) - E(u_{t-1}) = 0$。现在，var (v_t) = var $(u_t - u_{t-1})$ = var (u_t) + var $(u_{t-1}) = 2\sigma^2$，因为每个 u_t 的方差都是 σ^2，且 u 独立分布。因此，v_t 是同方差的。但

① J. Durbin, "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors Are Lagged Dependent Variables," *Econometrica*, vol. 38, pp. 410-421.

$$\text{cov}(v_t, v_{t-1}) = E(v_t v_{t-1}) = E[(u_t - u_{t-1})(u_{t-1} - u_{t-2})] = -\sigma^2$$

则显然非零。因此，尽管 u 不存在自相关，但 v 存在。

□ 12A.2 方程（12.2.3）、（12.2.4）和（12.2.5）的证明

在 AR(1) 模式下，

$$u_t = \rho u_{t-1} + \varepsilon_t \tag{1}$$

因此

$$E(u_t) = \rho E(u_{t-1}) + E(\varepsilon_t) = 0 \tag{2}$$

所以，

$$\text{var}(u_t) = \rho^2 \text{var}(u_{t-1}) + \text{var}(\varepsilon_t) \tag{3}$$

因为 u 和 ε 都不存在自相关。

因为 $\text{var}(u_t) = \text{var}(u_{t-1}) = \sigma^2$，且 $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$，我们得到

$$\text{var}(u_t) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{4}$$

现将方程（1）式两边同时乘以 u_{t-1} 并取期望，则得到

$$\text{cov}(u_t, u_{t-1}) = E(u_t u_{t-1}) = E[\rho u_{t-1}^2 + u_{t-1}\varepsilon_t] = \rho E(u_{t-1}^2)$$

注意到 u_{t-1} 和 ε_t 之间的协方差为零（为什么?），而且 $\text{var}(u_t) = \text{var}(u_{t-1}) = \sigma_\varepsilon^2/(1-\rho^2)$，我们有

$$\text{cov}(u_t, u_{t-1}) = \rho \frac{\sigma_\varepsilon^2}{(1-\rho^2)} \tag{5}$$

以此类推，

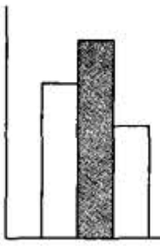
$$\text{cov}(u_t, u_{t-2}) = \rho^2 \frac{\sigma_\varepsilon^2}{(1-\rho^2)}$$

$$\text{cov}(u_t, u_{t-3}) = \rho^3 \frac{\sigma_\varepsilon^2}{(1-\rho^2)}$$

等等。现在，相关系数就是协方差与方差之比，因此，

$$\text{cov}(u_t, u_{t-1}) = \rho \qquad \text{cov}(u_t, u_{t-2}) = \rho^2$$

第 13 章 计量经济建模：模型设定和诊断检验



不能机械地做应用计量经济学，它需要理解、直觉和技巧。[①]

通常我们在驾车通过一座桥梁时并不担心其结构的可靠性，因为我们合理地相信已经有人严格地检查过其工程原理和实践。经济学家做模型时也必须这样，否则，就必须奉送一句警告“使用导致坍塌概不负责”[②]。

经济学家多年来对“真理”的寻求曾给人一种观感：经济学家们就好像在一间黑房子里搜寻一只原本并不存在的黑猫，而计量经济学家还经常声称找到了一只。[③]

经典线性回归模型（CLRM）的假定之一（假定 9）是，分析中所使用的模型被“正确”地设定了：如果模型未被“正确”设定，那我们就遇到**模型设定误差**（model specification error）或**模型设定偏误**（model specification bias）的问题。由于寻找正确设定的模型就像寻找圣杯一样，所以我们在本章将缜密而又严格地考察这个假定。具体而言，我们要考虑如下问题：

1. 我们如何去发现一个“正确”的模型？换言之，在经验分析中选择一个模型的准则有哪些？
2. 我们在实践中容易遇到哪些类型的模型设定误差？
3. 设定误差的后果有哪些？

---

① Keith Cuthbertson, Stephen G. Hall, and Mark P. Taylor, *Applied Econometrics Techniques*, Michigan University Press, 1992, p. X.

② David F. Hendry, *Dynamic Econometrics*, Oxford University Press, U. K., 1995, p. 68.

③ Peter Kennedy, *A Guide to Econometrics*, 3d ed., The MIT Press, Cambridge, Mass., 1992, p. 82.

4. 我们如何侦察设定误差？换言之，我们可以使用哪些诊断工具？

5. 一旦侦察出设定误差，我们能采取哪些补救措施？有什么好处？

6. 我们如何评价几个表现不相上下的备选模型？

模型设定与评价的专题涉及的范围非常广泛，在此领域中已经做过许多经验研究。除此之外，对这个问题还有一些哲学上的论争。尽管我们不能在一章的篇幅内充分说明这个专题，但我们希望能澄清模型设定和模型评价过程中所涉及的一些本质问题。

13.1 模型选择准则

根据韩德瑞（Hendry）和理查德（Richard）的观点，一个用于经验分析的模型应满足如下准则[①]：

1. 数据容纳性；即从模型做出预测必须有逻辑上的可能性。

2. 与理论一致；即必须有好的经济含义。比如，若米尔顿·弗里德曼（Milton Friedman）的**永久收入假说**（permanent income hypothesis）成立，则在永久消费对永久收入的回归中，预期截距项的值应该为零。

3. 回归元的弱外生性；即解释变量或回归元必须与误差项不相关。在一些情况下，回归元的外生性可能是**强外生性**（strictly exogenous）。一个强外生性变量与误差项的当前期、未来期以及滞后期均不相关。

4. 表现出参数的不变性；即参数值的稳定性。否则，预测就很困难。弗里德曼曾指出："对一个假设（模型）有效性唯一重要的检验，就是其预测值与经验值的比较。"[②] 在参数缺乏恒常性时，这种预测就不可靠。

5. 表现出数据的协调性；即从模型中所估计的残差必须完全随机（从技术上讲必须是白噪音）。换言之，若模型适当，则此模型的残差必须是白噪音。否则，模型中就存在着某种形式的设定误差。稍后，我们将阐释设定误差的性质。

6. 模型有一定的包容性；即模型从能解释其结论的意义上讲应该包容或包括所有与之相竞争的模型。简言之，其他模型都不可能再改进我们所选定的模型。

列出一个"好"模型的准则是一方面，但实际上做出一个"好"模型则完全是另一回事，因为我们在实践中很可能会遇到我们在下一节将要讨论的各种各样的设定误差。

① D. F. Hendry and J. F. Richard, "The Econometric Analysis of Economic Time Series," *International Statistical Review*, vol. 51, 1983, pp. 3-33.

② Milton Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics*, University of Chicago Press, Chicago, 1953, p. 7.

13.2 设定误差的类型

假定根据刚才列举的那些准则，我们得到一个我们认为是好的模型。为简明起见，令这个模型为：

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i} \tag{13.2.1}$$

其中 Y_i=生产总成本，X=产出，方程（13.2.1）是教科书中常见的立方总成本函数的例子。

但假使出于某种原因（比如说，不想花工夫去描绘散点图），研究者决定使用以下模型：

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \tag{13.2.2}$$

注意，我们改变了符号以区别于真实模型。

由于方程（13.2.1）被认为是真实的，所以采用方程（13.2.2）就构成一种设定误差，即**漏掉一个有关变量**（omitting a relevant variable）（X_i^3）的误差。因此，方程（13.2.2）中的误差项 u_{2i} 事实上是：

$$u_{2i} = u_{1i} + \beta_4 X_i^3 \tag{13.2.3}$$

我们将很快看到这一关系式的重要性。

现假定另一研究者用了下述模型：

$$Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_i^3 + \lambda_5 X_i^4 + u_{3i} \tag{13.2.4}$$

如果方程（13.2.1）是“真实”的，则方程（13.2.4）也构成一种设定误差，现在是**包含了一个无需或无关变量**（including an unnecessary or irrelevant variable）的误差，意指在真实模型中 λ_5 为零，新误差项其实是

$$\begin{aligned} u_{3i} &= u_{1i} - \lambda_5 X_i^4 \\ &= u_{1i} \qquad \text{因为真实模型中 } \lambda_5 = 0\text{(为什么?)} \end{aligned} \tag{13.2.5}$$

再假定又一研究者拟定以下的模型：

$$\ln Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i^2 + \gamma_4 X_i^3 + u_{4i} \tag{13.2.6}$$

和真实模型（13.2.1）相比，方程（13.2.6）也构成一种设定偏误，其偏误在于使用了**错误的函数形式**（wrong functional form），在方程（13.2.1）中 Y 以线性形式出现，而在方程（13.2.6）中它以线性到对数的形式出现。

最后，考虑有研究者使用如下模型：

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{*2} + \beta_4^* X_i^{*3} + u_i^* \tag{13.2.7}$$

其中 $Y_i^* = Y_i + \varepsilon_i$，$X_i^* = X_i + w_i$，$\varepsilon_i$ 和 w_i 均为测量误差。方程（13.2.7）所表明的是，研究者没有使用真正的 Y_i 和 X_i，却用了含有测量误差的替代变量 Y_i^* 和 X_i^*。因此，在方程（13.2.7）中研究者犯了**测量误差偏误**（errors of measurement bias）。在应用研究中，数据不免受到近似计算、不完全覆盖以及数据缺落等误差的困扰。

在社会科学中我们常用到的是第二手材料，通常无法得知第一手材料收集者是否造成了误差以及误差的类型为何。

另一种设定误差的形式与随机误差 u_i（或 u_t）进入回归模型的方式有关系。比如考虑如下不含截距项的双变量回归模型：

$$Y_i = \beta X_i u_i \tag{13.2.8}$$

其中随机误差项以乘积的形式进入回归方程，并且 $\ln u_i$ 满足 CLRM 的假定，这与误差项以相加的形式进入如下模型是不同的：

$$Y_i = \alpha X_i + u_i \tag{13.2.9}$$

尽管这两个模型中的变量相同，我们已经分别用 β 和 α 来表示方程（13.2.8）和（13.2.9）中的斜率系数。现在问，若方程（13.2.8）是“正确”或“真实”的模型，那么所估计的 α 会给出真实 β 的一个无偏估计值，即 $E(\hat{\alpha}) = \beta$ 吗？若不然，则对误差项不适当的随机设定将构成设定误差的另一个根源。

一个有时被忽视的设定误差是**回归元之间的相互影响**（interaction among the regressors），也就是**一个或多个回归元对回归子的乘积影响**（multiplicative effect of one or more regressors on the regressand）。为说明这一点，考虑如下简化的工资函数：

$$\ln W_i = \beta_1 + \beta_2 \text{Education}_i + \beta_3 \text{Gender}_i + \beta_4 (\text{Education}_i) \cdot (\text{Gender}_i) + u_i \tag{13.2.10}$$

在这个模型中，工资 W 相对于受教育水平（Education）的变化不仅取决于受教育水平本身，还取决于性别（Gender）（$\partial \ln W / \partial \text{Education} = \beta_2 + \beta_4 \text{Gender}$）。同样，工资相对于性别的变化不仅取决于性别本身，还取决于受教育水平。

总之，在提出一个经验模型时，我们很可能会遇到如下一种或多种设定误差：

1. 漏掉一个有关变量；
2. 包含一个无关变量；
3. 采用错误函数形式；
4. 测量误差；
5. 对随机误差项不正确的设定；
6. 误差项正态分布的假定。

在转而详细考察这些设定误差之前，先区分**模型设定误差**（model specification errors）和**模型误设误差**（model mis-specification errors）很有好处。我们以上讨论的前四种误差类型在本质上都属于模型设定误差，因为在我们脑海中都有一个“真实”模型，只是出于某种原因我们没有估计这个正确的模型。在模型误设误差中，我们不知道真实模型是什么。这种情形下，我们或许联想到凯恩斯学派与货币主义学派之间的争论。在解释 GDP 的变化时，货币主义者认为货币是第一位的因素，而凯恩斯主义者则强调政府支出的作用。由此看来，这是两个不相上下的竞争模型。

接下来，我们将首先考虑模型设定误差，然后考察模型误设误差。

13.3 模型设定误差的后果

不管设定误差的来源是什么，会由此造成什么样的后果？为使讨论简单，我们将按三变量模型的框架回答此问题，并在本节只考虑此前讨论过的前两种设定误差形式，(1) **模型拟合不足**（underfitting a model），即漏掉一个有关变量，和 (2) **模型拟合过度**（overfitting a model），即包含了一个无关变量。尽管这里的讨论很容易就能推广到多于两个回归元的情形，可一旦超出三个变量的情形，代数运算就会很烦琐，而且矩阵代数也几乎不可缺少。[①]

□ 模型拟合不足（漏掉一个有关变量）

假如真实模型是：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{13.3.1}$$

但出于某种原因我们拟合了如下模型：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \tag{13.3.2}$$

漏掉 X_3 的后果将是：

1. 如果放弃或漏掉的变量 X_3 与包含进来的变量 X_2 相关，也就是两变量的相关系数 r_{23} 非零，则 $\hat{\alpha}_1$ 和 $\hat{\alpha}_2$ 是有偏误且非一致的。就是说，$E(\hat{\alpha}_1)$ 不等于 β_1，$E(\hat{\alpha}_2)$ 不等于 β_2，而且这种偏误不会随着样本容量的增大而消失。

2. 即使 X_2 与 X_3 不相关（$r_{23}=0$），尽管 $\hat{\alpha}_2$ 现在是无偏的，但 $\hat{\alpha}_1$ 仍有偏误。

3. 误差项（干扰项）的方差 σ^2 将被不正确地估计。

4. 习惯上计算的 $\hat{\alpha}_2$ 的方差（$=\sigma^2/\sum x_{2i}^2$），是真实估计量 $\hat{\beta}_2$ 的方差的一个有偏误的估计量。

5. 后果是，通常的置信区间和假设检验程序对于所估计参数的统计显著性容易导出误导性的结论。

6. 另外一个后果是，基于不正确模型做出的预测及预测（置信）区间都是不可靠的。

虽然对以上命题一一给出证明会使我们离题太远[②]，但我们还是在附录 13A 的 13A.1 节证明了：

$$E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32} \tag{13.3.3}$$

其中 b_{32} 是漏去的变量 X_3 对包含进来的变量 X_2 回归的斜率（$b_{32}=\sum x_{3i}x_{2i}/\sum x_{2i}^2$）。如方

① 但请参见习题 13.32。

② 代数上的处理可参看 Jan Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971, pp. 391-399。熟悉矩阵代数的读者可参考 J. Johnston, *Econometric Methods*, 4th ed., McGraw-Hill, New York, 1997, pp. 119-121。

程（13.3.3）所示，除非 β_3 或 b_{32} 或二者同时为零，否则 $\hat{\alpha}_2$ 就是有偏误的。我们排除了 β_3 为零的可能性，因为如果那样的话，我们就不存在设定误差的问题。系数 b_{32} 在 X_2 与 X_3 无关时为零，这在大多数经济数据中都不太可能。

然而一般而言，偏误的程度将取决于偏误项 $\beta_3 b_{32}$。如果，比方说，β_3 是正的（即 X_3 对 Y 有正的影响），并且 b_{32} 也是正的（即 X_2 与 X_3 正相关）。那么，总体而言，$\hat{\alpha}_2$ 将高估了真实的 β_2（即有正的偏误）。但这种结果并不足为奇，因为 X_2 不仅代表了其对 Y 的直接影响，还包括了其（通过 X_3）对 Y 的间接影响。简单地说，本来应归功于 X_3 的影响，由于 X_3 未“获准”进入模型，所以无从展示其效应，而记在 X_2 的头上。为了做一个简洁的说明，考虑第 7 章曾讨论过的一个例子（例 7.1）。

例 13.1　说明性例子：再谈儿童死亡率一例

将儿童死亡率（CM）对人均 GNP（即 PGNP）和妇女识字率（FLR）做回归，我们得到方程（7.6.2）所示的回归结果，并给出这两个变量的偏斜率系数值分别为 −0.005 6 和 −2.231 6。但如果我们去掉 FLR 变量，则得到方程（7.7.2）所示的结果。若我们把方程（7.6.2）视为正确模型，则方程（7.7.2）因漏掉了有关变量 FLR 而成为一个设定误差模型。现在你可以看到，在正确的模型中，PGNP 变量的系数为 −0.005 6，而在“不正确”的模型（7.7.2）中却为 −0.011 4。

从绝对值来看，与真实模型相比，现在 PGNP 对 CM 有更大的影响。但如果我们把 FLR 对 PGNP 回归（将被排除的变量对包含进来的变量回归），此回归的斜率系数［方程（13.3.3）中的 b_{32}］为 0.002 56。[①] 这就表明，随着 PGNP 每提高 1 个单位，FLR 平均上升约 0.002 56 个单位。但若 FLR 上升这么多单位，则其对 CM 的影响为 $(-2.231\ 6)\times 0.002\ 56 = \hat{\beta}_3 b_{32} = -0.005\ 43$。

因此，我们最后从方程（13.3.3）得到 $\hat{\beta}_2 + \hat{\beta}_3 b_{32} = -0.005\ 6 + (-2.231\ 6)\times 0.002\ 56 \approx -0.011\ 1$，与不正确模型（7.7.2）中得到的 PGNP 系数值大致相等。[②] 此例说明，PGNP 对 CM 的真实影响（−0.005 6）远低于不正确模型（7.7.2）所给出的结果（−0.011 4）。

现在让我们来分析 $\hat{\alpha}_2$ 和 $\hat{\beta}_2$ 的方差：

$$\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \tag{13.3.4}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF} \tag{13.3.5}$$

其中 VIF（对共线性的一种度量）为第 10 章讨论过的方差膨胀因子［$= 1/(1 - r_{23}^2)$］，而 r_{23} 为变量 X_2 和 X_3 之间的相关系数；我们在第 3 章和第 7 章已经熟悉了方程（13.3.4）和（13.3.5）。

① 回归结果为：

$$\widehat{\text{FLR}} = 47.597\ 1 + 0.002\ 56\text{PGNP}$$

$$\text{se} = (3.555\ 3) \quad (0.001\ 1) \quad r^2 = 0.072\ 1$$

② 注意，在真实模型中，$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 都是其真实值的无偏估计值。

由于公式（13.3.4）和（13.3.5）不同，所以 var($\hat{\alpha}_2$) 一般不同于 var($\hat{\beta}_2$)。但我们知道，var($\hat{\beta}_2$) 是无偏的。（为什么?）因此，var($\hat{\alpha}_2$) 就有偏误，从而证实了前面第 4 点论断。由于 $0<r_{23}^2<1$，所以在现在的情形下，可见 var($\hat{\alpha}_2$) < var($\hat{\beta}_2$)。现在，我们面临着一个两难选择：尽管 $\hat{\alpha}_2$ 有偏，但其方差比无偏估计量 $\hat{\beta}_2$ 的方差小（当然，我们现在排除了 $r_{23}=0$ 的情形，因为在实践中回归元之间总是有些相关）。所以，我们这里遇到一个取舍的问题。[①]

但问题还没有结束，因为两个模型的 RSS 不一样，其自由度也不一样，所以从模型（13.3.2）估计的 σ^2 和从真实模型（13.3.1）估计的 σ^2 也就不相同。你或许记得，我们得到 σ^2 的一个估计值为 $\hat{\sigma}^2=\text{RSS/df}$，其大小取决于模型中包含的回归元个数和自由度（$=n-$待估计参数的个数）。* 如果我们现在在模型中增加变量，RSS 通常会减小（记住，随着模型中引入越来越多的变量，R^2 不断变大），但待估计参数增加也使自由度下降。净影响取决于 RSS 的下降是否足以抵消增加回归元所导致的自由度损失。很有可能，一个回归元对回归子具有强烈影响（比如它使 RSS 的减少比在模型中引入此变量所导致的自由度损失大得多），因此，包含这种变量不仅减少偏误，而且还会提高估计量的精确性（即减小标准误）。

另一方面，如果有关的那些变量对回归子只有微弱的影响，而且它们之间高度相关（即 VIF 很大），那么，我们虽然可能减小了模型中所包含变量系数的偏误，但也增大了其标准误（使得效率下降）。事实上，此时在偏误和精确性之间的取舍将取决于各回归元的相对重要性。

为了给这种讨论做出个结论，让我们考虑 $r_{23}=0$ 的特殊情形，即 X_2 和 X_3 无关的情形。这时，b_{32} 将等于零。（为什么?）从而能从方程（13.3.3）看出，$\hat{\alpha}_2$ 现在是无偏的。[②] 而且，似乎从方程（13.3.4）和（13.3.5）看来，$\hat{\alpha}_2$ 和 $\hat{\beta}_2$ 的方差相等。那么，尽管理论上 X_3 是有关变量，是否从模型中略去它也不致有什么害处呢？一般来说，回答是否定的。正如前面所说的，这种情况下因为从方程（13.3.4）估计出来的 var($\hat{\alpha}_2$) 仍是有偏误的，致使我们的假设检验程序仍值得怀疑。[③] 再说，在大多数经济研究中，X_2 和 X_3 都相关，从而产生了以上所讲的问题。**论点已十分清楚：一旦根据相关理论把模型建立起来，切忌从中再忽略掉一个变量。**

□ 包含一个无关变量（模型拟合过度）

现在让我们假定：

$$Y_i=\beta_1+\beta_2X_{2i}+u_i \tag{13.3.6}$$

① 为了避开偏误和有效性之间的取舍关系，可以选择最小化均方误（MSE），因为它同时考虑了偏误和有效性。关于 MSE，可参见统计学方面的附录，即附录 A。也可参见习题 13.6。

* 原文漏掉减号——译者注。

② 但注意 $\hat{\alpha}_1$ 仍是有偏误的，这可如下直观看出：我们知道 $\hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}_2-\hat{\beta}_3\bar{X}_3$，而 $\hat{\alpha}_1=\bar{Y}-\hat{\alpha}_2\bar{X}_2$，即便 $\hat{\alpha}_2=\hat{\beta}_2$，这两个截距估计量也不同。

③ 详细分析可参见 Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar Publisher, 1994, pp. 371-372。

是真实模型，而我们拟合了以下模型：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i \tag{13.3.7}$$

从而导致了在模型中引入一个无关变量的设定误差。

这一假定误差将导致如下后果：

1. “不正确”模型中全部参数的 OLS 估计量都是无偏而又一致的，即 $E(\alpha_1) = \beta_1$，$E(\hat{\alpha}_2) = \beta_2$ 和 $E(\hat{\alpha}_3) = \beta_3 = 0$。

2. 误差方差 σ^2 的估计是正确的。

3. 通常的置信区间和假设检验程序仍然有效。

4. 然而，一般地说，诸 α 系数的估计量将是非有效的，也就是说，它们的方差一般都大于真实模型中 $\hat{\beta}$ 的方差。关于这些命题的部分证明，见附录 13A 第 13A.2 节。这里要考虑的问题是这些 $\hat{\alpha}$ 的相对无效性，这是不难证明的。

我们从通常使用的 OLS 公式得知：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \tag{13.3.8}$$

以及

$$\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \tag{13.3.9}$$

因此，

$$\frac{\text{var}(\hat{\alpha}_2)}{\text{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2} \tag{13.3.10}$$

由于 $0 \leqslant r_{23}^2 \leqslant 1$，推知 $\text{var}(\hat{\alpha}_2) \geqslant \text{var}(\hat{\beta}_2)$，也就是说，虽然平均而言 $\hat{\alpha}_2 = \beta_2$[即 $E(\hat{\alpha}_2) = \beta_2$]，但 $\hat{\alpha}_2$ 的方差一般都大于 $\hat{\beta}_2$ 的方差。

这一发现的含义是，包含无关变量 X_3 将使 $\hat{\alpha}_2$ 的方差不必要地增大，从而使 $\hat{\alpha}_2$ 的精度减小。这对 $\hat{\alpha}_1$ 也是成立的。

注意我们所考虑的两种设定误差的**不对称性**（asymmetry）。如果我们略去了一个有关变量，则留在模型中变量的系数一般地说既有偏误且非一致，误差的估计也是不正确的，从而通常的假设检验程序都是无效的。而另一方面，模型中含有无关变量虽然仍能给出真实模型中系数的无偏且一致估计，而且通常的假设检验方法也仍然有效；但引入多余变量的唯一代价是系数方差的估计值变大了，致使对参数进行概率推断的精度降低了。一个无益的结论似乎是：与其略掉有关变量，不如含有无关变量。但是这种哲理是不值得维护的，因为增加无关变量将导致估计量的效率损失，并且还可以引发多重共线性问题（为什么?），更不用说自由度的损失了。因此：

> 一般而言，最好的方法是，根据理论，仅仅包含那些直接影响因变量而又不能由已被引进的其他变量来代替的解释变量。[①]

① Michael D. Intriligator, *Econometric Models, Techniques and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1978, p. 189。回顾简单性原则。

13.4 设定误差的检验

知道设定误差的后果是一回事，而发现是否犯了这种错误则完全是另一回事。因为没有人要成心去犯这种错误，出现设定误差常常是疏忽所致，或由于基础理论薄弱或由于缺乏检验模型的适当数据，所以我们想尽可能准确地制定模型而又无能为力。如戴维森曾指出："由于经济学的非实验性质，所以我们一向对所观测到的数据的生成机制没有信心。对经济学中任何一个假设的检验，最终都取决于足以设定一个适当节俭的模型的附加假设，而这些假设既可能被证明是合理的，也可能没有被证明是合理的。"①

于是，实际问题不是为什么会造成这种错误（因为通常都会造成），而是如何发现这种错误。一旦发现了有设定误差，也就常常能找出补救的方法。例如，如果能证明模型中不恰当地漏掉了一个变量，明显的补救方法就是把那个变量加到分析中去，当然这里假定能获得有关的数据。

我们在本节中就是要讨论一些可用来侦察设定误差的检验方法。

□ 侦察是否含有无关变量（对过度拟合模型的侦察）

假如为了解释某一现象我们提出一个 k 变量模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{13.4.1}$$

然而，比方说，变量 X_k 是否真的属于模型（所应包含的变量），我们却没有十分把握。一个简单的辨认方法是用通常的 t 检验即 $t = \hat{\beta}_k / \text{se}(\hat{\beta}_k)$ 去检验所估计 β_k 的显著性。又比方说，我们不能肯定 X_3 和 X_4 是否真的属于模型。于是，我们很容易通过第 8 章所讲的 F 检验来判断。因此，侦察模型中是否出现了一个或多个无关变量并不困难。

但至关重要的是，记住在做这些显著性检验时，我们心目中要有一个具体的模型。尽管这个模型在某种程度上是试验性质的，但我们应把它看作**保留假设**（maintained hypothesis）或"真理"。于是，给定该模型，我们可通过平常的 t 和 F 检验去辨认一个或多个回归元是不是真正有关的变量。但要注意，切勿反复使用 t 和 F 检验来建立模型。就是说，我们不可以说，Y 之所以与 X_2 有关，只因为 $\hat{\beta}_2$ 是统计显著的。然后又因为 $\hat{\beta}_3$ 是统计显著的，便把 X_3 包含在模型中，如此等等。这种建模策略被称为**自下而上的方法**（bottom-up approach）（从一个较小的模型开始，然后逐渐扩大模型），或者多少带些轻蔑口吻地称之为**数据挖掘**（data mining）方法、**回归捕捉**（regression fishing）方法、**数据琢磨**（data grubbing）方法、**数据窥探**

① James Davidson, *Econometric Theory*, Blackwell Publishers, Oxford, U. K., 2000, p. 153.

(data snooping) 方法、**数字斟酌** (number crunching) 方法等。

数据挖掘的主要目标是在进行一些诊断检验之后提出一个“最好”的模型，即使最终选定的模型在如下意义上是一个“好”模型：其所有估计系数都具有“正确”的符号，基于 t 和 F 检验都是统计显著的，R^2 值足够高，德宾-沃森 d 统计量的值可以接受（约为 2）等。本专业的纯粹主义者很看不起数据挖掘的实践。用威廉·普尔 (William Pool) 的话说，“……寻求数据基础的经验规律而非经济理论的含义总是很危险的。”[①] “谴责”数据挖掘的原因之一如下。

在数据挖掘情况下的名义与真实显著性水平。不经心的研究者所面临的一种数据挖掘的危险是，诸如 1%、5%或 10%的常用显著性水平 (α) 并非真正的显著性水平。罗维尔 (Lovell) 曾指示，如果有 c 个备用的回归元，根据数据挖掘情况，从中最后选出 k 个 ($k \leqslant c$)，则真实的显著性水平 (α^*) 和名义上的显著性水平 (α) 有如下关系[②]：

$$\alpha^* = 1-(1-\alpha)^{c/k} \tag{13.4.2}$$

或近似地为：

$$\alpha^* \approx (c/k)\alpha \tag{13.4.3}$$

例如，取 $c=15$，$k=5$ 和 $\alpha=5\%$，由方程 (13.4.3)，真实的显著性水平是 $(15/5)\times(5\%)=15\%$。因此，如果一个研究者从数据挖掘中选择 15 个回归中的 5 个，而仅按名义的显著性水平 5%报告这个浓缩模型的结果，并宣称这些结果在统计上是显著的，那么要别人接受这种结果就是一桩有苦难言的事。因为我们知道，真实的显著性水平实际上是 15%。应该指出，若 $c=k$，则不存在数据挖掘的问题，真实和名义的显著性水平也就相同。当然，在实践中，研究者都是仅报告其最后结果而不透露此前是如何通过大量数据挖掘或**预检验** (pretesting) 而得到这些结果的详情。[③]

尽管数据开采有一些明显的缺陷，但仍不断得到承认，特别是在应用计量经济学家中，纯粹主义者（即非数据挖掘）的建模方法完全抵挡不住如此强烈的攻势。如查曼 (Zaman) 指出：

> 不幸的是，运用真实数据集的经验表明，这样一种（纯粹主义）方法既不可行又不理想。之所以说它不可行，是因为很少有经济理论只导致唯一的模型。之所以说它不理想，是因为从数据中了解到的一个关键方面是数据支持和不支持什么类型的模型。即便原模型碰巧表现出优良的拟合性质，但说明和了解数据支持和不支持的模型类型也越来越重要。[④]

---

① William Pool, “Is Inflation Too Low?” the *Cato Journal*, vol. 18, no. 3, Winter 1999, p. 456.

② M. Lovell, “Data Mining,” *Review of Economics and Statistics*, vol. 65, 1983, pp. 1-12.

③ 对预检验的详尽讨论及其可能导致的偏误，参见 T. D. Wallace, “Pretest Estimation in Regression: A Survey,” *American Journal of Agricultural Economics*, vol. 59, 1977, pp. 431-443。

④ Asad Zaman, *Statistical Foundations for Econometric Techniques*, Academic Press, New York, 1996, p. 226.

帕特森（Kerry Patterson）也表达出类似的观点，他指出：

> （数据挖掘）这种方法表明，经济理论与经验设定相互影响而不是各自为阵。[①]

为了避免陷入数据挖掘与纯粹主义关于建模思路的论争，我们可以借鉴彼得·肯尼迪（Peter Kennedy）的观点：

> （模型设定）需要对理论和数据的通盘考虑，设定搜索中所用到的检验程序应按照使数据挖掘成本最小的要求来设计。这方面的例子有：为样本外预测的检验预留下数据、调整显著性水平（罗维尔）和避免使用诸如最大化 R^2 之类的值得怀疑的准则。[②]

如果我们从一个更开阔的视角来看待数据挖掘，把它看成一种寻求经验规律的过程，并能从这些经验规律中判断（现有）理论模型中是否存在错误和/或疏漏，那么，它的作用就太大了。再次引用肯尼迪的话："应用计量经济学家的艺术在于，容许数据驱动理论而又不致陷入太大的数据挖掘的危险。"[③]

□ 对遗漏变量和不正确函数形式的检验

实际上，我们永远不能肯定用来做经验检验的模型是"真理，完全的真理，非真理莫属"。我们是根据理论或洞察力和先前的经验工作建立一个我们认为能抓住问题实质的模型，然后对模型进行经验检验。在获得结果之后，再按照上面讨论过的良好模型准则作事后调查分析。到了这一步我们才会知道所选模型是否适宜。在决定模型的适宜性时，我们着眼于结果的一些概括性特征，诸如 $\bar{R}^2$ 值，估计的 t 比率，估计的系数符号与事先预期的是否一致，德宾-沃森统计量，等等。如果这些诊断特征合理，我们就宣称所选模型是现实的良好代表。同理，如果结果看来不够理想，或由于 $\bar{R}^2$ 值太低，或由于统计上显著或有正确符号的系数太少，或由于德宾-沃森 d 值太低，我们便开始担心模型的适宜性，并着手寻找补救方法。也许我们漏掉某个重要变量，或者用了错误的函数形式，或者没有先求时间序列的差分（以消除序列相关），如此等等。为了帮助我们确定模型不适宜性是否由这些问题的一种或几种所引起，不妨利用下列方法。

残差分析。如在第 12 章中所看到的，对残差的分析曾是侦察自相关和异方差性的一种良好的视觉鉴别法。然而，尤其对横截面数据而言，残差还可以用于分析模型的设定误差，诸如一个重要变量的遗漏或函数形式的误用。事实上，如果存在这种误差，残差图将会显示出明显不同的形状。

---

① Kerry Patterson, *An Introduction to Applied Econometrics*, St. Martin's Press, New York, 2000, p. 10.

② Peter Kennedy, "Sinning in the Basement: What are the Rules? The Ten Commandments of Applied Econometrics," unpublished manuscript.

③ Kennedy, op. cit., p. 13.

为了说明问题，让我们再考虑最先在第 7 章讨论过的立方总生产成本函数，假定真实总成本函数可表述如下：

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \tag{13.4.4}$$

其中 Y=总成本，X=产出，但研究者拟合了以下的二次函数：

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \tag{13.4.5}$$

而另一研究者则拟合下面的线性函数：

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \tag{13.4.6}$$

虽然我们明知两位研究者都造成了设定误差，但为了教学的目的，不妨看看这些模型的残差估计值是何种模样。（表 7—4 给出这些成本—产出数据。）图 13—1 不言而喻：从左至右，逐渐接近真实模型的过程中，不仅残差（在绝对值上）变小，而且与错用模型联系在一起的突出的周期振动也逐渐消失。

由此看到残差图分析的效用：如果有设定误差，残差图必定展现出明显的样式。

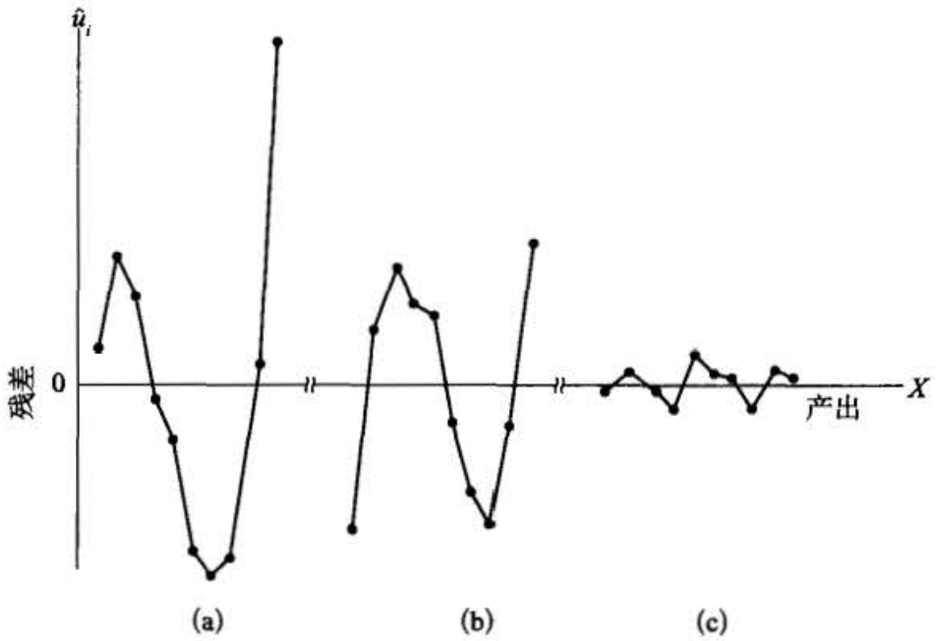


图 13—1　得自（a）线性、（b）二次及（c）立方总成本函数的残差 $\hat{u}_i$

再次使用德宾-沃森 d 统计量。如果我们分析一下表 13—1 中惯常算出的德宾-沃森 d 值，我们便看到，对线性成本函数，估计的 d 是 0.716，表明在估计的残差中有正的“相关”，因为对 $n=10$ 和 $k'=1$，5%的 d 临界值是 $d_L=0.879$ 和 $d_U=1.320$（而 0.716 低于 $d_L=0.879$）。同理，对二次成本函数，估算的 d 是 1.038，则 5%的 d 临界值是 $d_L=0.697$ 和 $d_U=1.641$，从而表明无定论。但若使用修改的 d 检验（见第 12 章），则因估算的 d 值小于 d_U 而可以说残差中有正“相关”。至于立方成本函数这一真实的设定，估计的 d 值表示残差中无任何正“相关”[①]。

① 在本例中，$d=2$ 这个值表示无设定误差。（为什么？）

表 13—1　　从线性、二次以及立方总成本函数估计的残差

观测号	线性模型的 a_i^*	二次模型的 $a_i^\dagger$	立方模型的 a_i^{**}
1	6.600	−23.900	−0.222
2	19.667	9.500	1.607
3	13.733	18.817	−0.915
4	−2.200	13.050	−4.426
5	−9.133	11.200	4.435
6	−26.067	−5.733	1.032
7	−32.000	−16.750	0.726
8	−28.933	−23.850	−4.119
9	4.133	−6.033	1.859
10	54.200	23.700	0.022

注：$^*\hat{Y}_i=166.467+19.933\,X_i$　　$R^2=0.840\,9$

(19.021)　(3.066)　　$\bar{R}^2=0.821\,0$

(8.752)　(6.502)　　$d=0.716$

$^\dagger\hat{Y}_i=222.383-8.025\,0X_i+2.542X_i^2$　　$R^2=0.928\,4$

(23.488)　(9.809)　(0.869)　　$\bar{R}^2=0.907\,9$

(9.468)　(−0.818)　(2.925)　　$d=1.038$

$^{**}\hat{Y}_i=141.767+63.478X_i-12.962X_i^2+0.939X_i^3$　　$R^2=0.998\,3$

(6.375)　(4.778)　(0.985 6)　(0.059 2)　　$\bar{R}^2=0.997\,5$

(22.238)　(13.285)　(−13.151)　(15.861)　　$d=2.70$

当我们拟合线性或二次模型时，所观测的正“相关”并非（一阶）序列相关，而是（模型）设定误差的一种度量。所看到的相关仅反映这样一个事实：本属于模型的某些变量被并入到误差项中去了，故而需从误差中分离出去作为解释变量而单独存在：如果我们从成本函数中排除 X_i^3，则如方程（13.2.3）所示，误设模型（13.2.2）中的误差项实际上是 $u_{1i}+\beta_4 X_i^3$，因而，如果事实上 X_i^3 显著地影响着 Y，它必然展现一种系统性的模式（例如正自相关）。

为了用德宾-沃森检验来侦察模型设定误差，我们采取如下过程：

1. 从假定的模型求得 OLS 残差。

2. 如果认为假定的模型因排除了一个有关的解释变量 Z（比方说），所以是误设的，就可将步骤 1 中所得的残差按 Z 值的递增次序排列。注：Z 变量可以是假定模型所含的 X 变量之一，或该变量的某一函数，如 X^2 或 X^3。

3. 按照这种顺序排列的残差，按通常的 d 公式，即：

$$d=\frac{\sum_{t=2}^{n}(\hat{u}_t-\hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

计算 d 统计量。注意，下标 t 是这里（重排）的观测序号，不一定指时间序列数据。

4. 根据德宾-沃森表，如果 d 值是显著的，就可接受模型误设的假设。当这种情况出现时，补救措施也就寓于其中。

在我们的成本例子中，$Z(=X)$ 变量（产出）已经按从小到大次序排列[①]，因此无需重新计算 d 统计量。正如我们已经看到的，对于线性和二次成本函数，d 统计量都表明了设定误差，如何补救是明显的：在线性成本函数中引进二次项和三次项；在二次函数中引进三次项，简言之，使用立方成本模型。

拉姆齐的 RESET 检验。拉姆齐（Ramsey）曾提出称为回归设定误差检验（regression specification error test，RESET）的一般性设定误差检验。[②] 这里我们仅说明这种检验的最简单情形。为了建立概念，仍用我们的成本—产出例子并假定成本函数对产出是线性的：

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \tag{13.4.6}$$

其中 Y=总成本，X=产出。如果用得自此回归的残差 a_i 对此模型 Y_i 的估计值 $\hat{Y}_i$ 描图，就会得到一个如图 13—2 所示的图形。虽然 $\sum a_i$ 和 $\sum a_i \hat{Y}_i$ 都必然是零（为什么？参看第 3 章），图中的残差仍表明其均值系统地随 $\hat{Y}_i$ 而变化的模式。从而提示我们，如果以某种形式将 $\hat{Y}_i$ 当作回归元引入方程（13.4.6），则应使 R^2 增大。而如果 R^2 的增大是统计上显著的（在第 8 章所讨论的 F 检验的基础上），就表明线性成本函数（13.4.6）是误设的。这就是 RESET 的基本思路。RESET 的操作步骤如下：

1. 从所选模型，例如方程（13.4.6），得到 Y_i 的估计值 $\hat{Y}_i$。

2. 将某种形式的 $\hat{Y}_i$ 作为增补的回归元引入，重做方程（13.4.6）。由图 13—2 我们看出 a_i 与 $\hat{Y}_i$ 之间存在曲线关系，表明可引进 $\hat{Y}_i^2$ 和 $\hat{Y}_i^3$ 作为增补回归元，于是我们做回归：

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i \tag{13.4.7}$$

3. 得自方程（13.4.7）的 R^2 记为 R^2_{new}；得自方程（13.4.6）的 R^2 记为 R^2_{old}，然后利用首次在方程（8.4.18）中引入的 F 检验，即：

$$F = \frac{(R^2_{\text{new}} - R^2_{\text{old}})/\text{新回归元的个数}}{(1 - R^2_{\text{new}})/(n - \text{新模型中的参数个数})} \tag{8.4.18}$$

说明由于方程（13.4.7）的使用，R^2 的增大是不是统计上显著的。

4. 如果所计算的 F，比方说，在 5%水平上显著，就可接受模型（13.4.6）被误设的假设。

回到我们的说明性例子，我们有如下结果（括号中为标准误）：

$$\hat{Y}_i = 166.467 + 19.933 X_i \tag{13.4.8}$$
$$(19.021)\quad (3.066) \qquad R^2 = 0.840\,9$$

$$\hat{Y}_i = 2\,140.722\,3 + 476.655\,7 X_i - 0.091\,87 \hat{Y}_i^2 + 0.000\,119 \hat{Y}_i^3 \tag{13.4.9}$$
$$(132.004\,4)\quad (33.395\,1)\quad (0.006\,20)\quad (0.000\,007\,4) \qquad R^2 = 0.998\,3$$

注：方程（13.4.9）中的 $\hat{Y}_i^2$ 和 $\hat{Y}_i^3$ 得自方程（13.4.8）。

现应用 F 检验求得：

---

① 如果 a_i 按 X_i^2 或 X_i^3 排列也没有关系，因为 X_i^2 和 X_i^3 都是已经排列好的 X_i 的（增）函数。

② J. B. Ramsey, "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society*, series B, vol. 31, 1969, pp. 350-371.

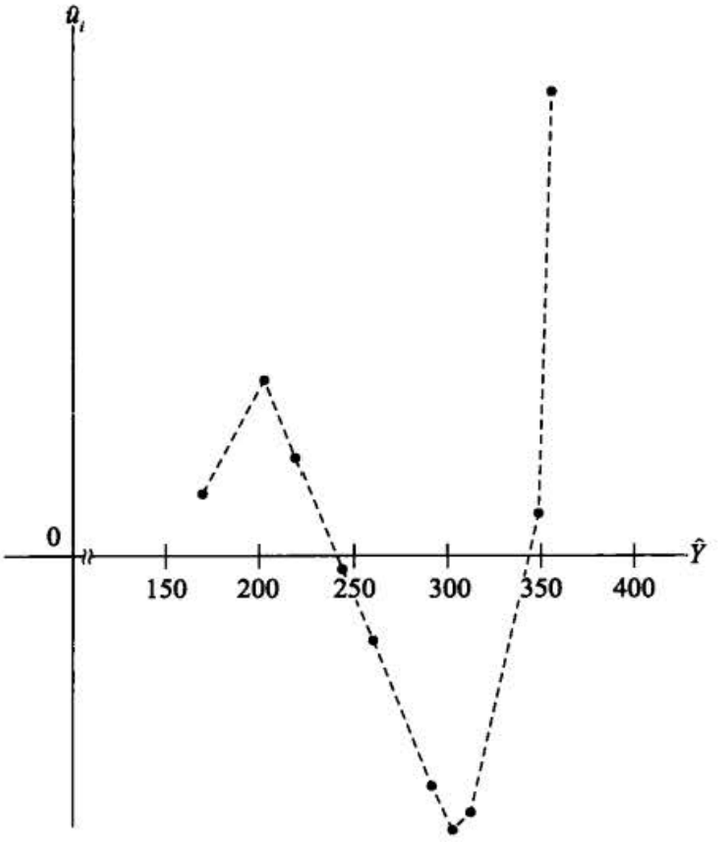


图 13—2　得自线性成本函数 $Y_i=\lambda_1+\lambda_2 X_i+u_{3i}$ 的残差 $\hat{u}_i$ 与 Y 的估计值

$$F=\frac{(0.998\ 3-0.840\ 9)/2}{(1-0.998\ 3)/(10-4)}=284.403\ 5 \tag{13.4.10}$$

读者容易核实此 F 值是高度显著的，表明了模型（13.4.8）是误设的。当然，我们根据残差的视觉分析和德宾-沃森 d 值已得到过同样的结论。还应该指出，由于 $\hat{Y}_i$ 是估计值，所以它是一个随机变量，因此，如果样本足够大，常用的显著性检验还是适用的。

RESET 的优点之一是，它不要求设定对立（alternative）模型，故易于应用。但这同时也是它的缺点，因为即便知道了模型误设也不一定有助于另外选出一个更好的模型。

正如另一位作者所说：

> 实践中，在检查一个建议使用模型的特定对立模型的好坏时，RESET 检验或许不是很好，而它的有用性体现在有某种问题出现时的一般性标志。出于这个原因，像 RESET 这样的检验有时被称为对模型误设的检验，而不是对模型设定的检验。虽然这个区别很微妙，但其基本思想是，设定检验检查一个给定方程的某个特定方面，我们脑海中有明确的虚拟假设和对立假设。另一方面，误设检验则检查出虚拟假设存在某种问题，而对立假设还有很多可能性，这种检验不一定能够给出适当对立假设的明确指引。①

① John Stewart and Len Gill, *Econometrics*, 2d ed., Prentice-Hall Europe, 1998, p. 69.

为增补变量的拉格朗日乘数检验。这是相对于拉姆齐的 RESET 检验的另一检验。为说明此检验，我们继续用前述说明性例子。

如果将线性成本函数（13.4.6）同立方成本函数（13.4.4）相比，前者就是后者的一个受约束形式（restricted version）（回顾我们在第 8 章中关于**约束最小二乘**的讨论）。受约束的回归（13.4.6）假定平方和立方产出项的系数均为零。为检验此假定，拉格朗日乘数（Lagrange multiplier，LM）检验如下进行：

1. 用 OLS 法估计受约束回归（13.4.6）并求得残差 $\hat{u}_i$。

2. 如果无约束的回归（13.4.4）事实上是真实回归，则得自方程（13.4.6）的残差应与平方产出 X_i^2 和立方产出 X_i^3 有关。

3. 这就建议我们用在步骤 1 中得到的 $\hat{u}_i$ 去对全部回归元（包括受限回归中的回归元）做回归，这在本例中是指：

$$\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^3 + v_i \tag{13.4.11}$$

其中 v 是具有通常性质的一个误差项。

4. 对于大样本，恩格尔（Engle）曾证明，从（辅助）回归（13.4.11）估计出来的 R^2 的 n（样本大小）倍遵循自由度等于受约束回归中约束个数的 χ^2 分布，用符号可表示为：

$$nR^2 \underset{\text{asy}}{\sim} \chi^2_{(\text{约束个数})} \tag{13.4.12}$$

其中 asy 表示渐近地，即在大样本中。

5. 如果从方程（13.4.12）得到的 χ^2 值大于选定显著性水平的 χ^2 临界值，就拒绝受限回归；否则不拒绝。

对于我们的例子，回归结果如下：

$$\hat{Y}_i = 166.467 + 19.333X_i \tag{13.4.13}$$

其中 Y 是总成本，而 X 是产出。关于此回归的标准误见表 13—1。

用从方程（13.4.13）得到的残差按方才讲的步骤 3 做回归时，得到如下结果：

$$\begin{aligned} \hat{u}_i &= -24.7 + 43.544\ 3X_i - 12.961\ 5X_i^2 + 0.939\ 6X_i^3 \\ \text{se} &= (6.375)\ (4.779)\quad (0.986)\quad (0.059) \qquad (13.4.14) \\ R^2 &= 0.989\ 6 \end{aligned}$$

虽然我们的样本只有 10 个，谈不上是大样本，但仅仅为了说明 LM 的操作方法，我们算出 $nR^2 = 10 \times 0.989\ 6 = 9.896$。从 χ^2 表我们读出 2 个自由度的 1%的 χ^2 临界值是 9.21，因此，所测的 9.896 这个值在 1%水平上是显著的，从而我们的结论是拒绝约束回归（即线性成本函数）。我们得到与基于拉姆齐的 RESET 检验类似的结论。

13.5 测量误差

我们一直在隐含地假定我们对因变量 Y 和诸解释变量 X 的观测无任何误差。例

如，在消费支出对家庭收入和财富的回归中，我们假定对这些变量的（观测）数据是“准确”的；它们不是由外推、内插或按任何系统的方式进位（如进位到最近似的百分之一美元，等等）而得到的猜测估计（guess estimates）。可惜，这种理想情形由于种种原因，如非应答误差、报道误差和计算误差，实际上是找不到的。且不管什么原因，由于测量误差构成了又一类设定偏误，并带来下面注明的后果，所以它是一个潜在的麻烦问题。

□ 因变量 Y 中的测量误差

考虑以下模型：

$$Y_i^* = \alpha + \beta X_i + u_i \tag{13.5.1}$$

其中Y_i^* =永久性消费支出[①]；

X_i=当前收入；

u_i=随机干扰项。

由于 Y_i^* 不可直接观测，所以我们可能使用了这样的一个可观测变量 Y_i：

$$Y_i = Y_i^* + \varepsilon_i \tag{13.5.2}$$

其中 ε_i 表示 Y_i^* 中的测量误差。于是我们估计的不是方程（13.5.1）而是：

$$\begin{aligned} Y_i &= (\alpha + \beta X_i + u_i) + \varepsilon_i \\ &= \alpha + \beta X_i + (u_i + \varepsilon_i) \\ &= \alpha + \beta X_i + v_i \end{aligned} \tag{13.5.3}$$

其中 $v_i = u_i + \varepsilon_i$ 是一个合成误差项，包含总体干扰项（也可称为方程误差项）和测量误差项。

为简单起见，根据经典线性回归假设假定 $E(u_i)=E(\varepsilon_i)=0$，$\text{cov}(X_i, u_i)=0$（就是说 Y_i^* 的测量误差与 X_i 不相关），$\text{cov}(u_i, \varepsilon_i)=0$（就是说方程误差与测量误差不相关）。有了这些假定就可以证明，从方程（13.5.1）或（13.5.3）估计 β 都将给出真实 β 的一个无偏估计量（见习题 13.7）；就是说，因变量中的测量误差并不破坏OLS估计量的无偏性质。然而，从方程（13.5.1）和（13.5.3）估计的 β 的方差和标准差将是不同的。这是因为按照通常的公式（见第 3 章）我们得到：

$$\text{模型(13.5.1)：} \text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2} \tag{13.5.4}$$

$$\text{模型(13.5.3)：} \text{var}\,(\hat{\beta}) = \frac{\sigma_v^2}{\sum x_i^2} = \frac{\sigma_u^2 + \sigma_\varepsilon^2}{\sum x_i^2} \tag{13.5.5}$$

显然后者大于前者。[②] 因此，**虽然因变量中的测量误差不影响参数估计及其方差的无偏性，但这时所估计的方差却比没有这种测量误差时要大。**

---

① 这一术语出自米尔顿·弗里德曼。还参见习题 13.8。

② 但注意，因在所述条件下合成误差项 $v_i = v_i + \varepsilon_i$ 仍满足最小二乘法的基本假定，此方差仍是无偏的。

□ 解释变量 X 中的测量误差

现假定模型不是（13.5.1）而是：

$$Y_i = \alpha + \beta X_i^* + u_i \tag{13.5.6}$$

其中Y_i＝当前消费支出；

X_i^*＝永久收入；

u_i＝干扰项（方程误差）。

假设我们观测到的不是 X_i^*，而是 X_i：

$$X_i = X_i^* + w_i \tag{13.5.7}$$

其中 w_i 代表 X_i^* 中的测量误差，从而我们估计的不是方程（13.5.6），而是：

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - w_i) + u_i \\ &= \alpha + \beta X_i + (u_i - \beta w_i) \\ &= \alpha + \beta X_i + z_i \end{aligned} \tag{13.5.8}$$

其中 $z_i = u_i - \beta w_i$，是方程误差与观测误差两种误差的一个混合。

现在即使我们假定 w_i 有零均值，序列独立且与 u_i 不相关，我们却不再能假定合成误差项 z_i 独立于解释变量 X_i，因为［假定 $E(z_i)=0$］

$$\begin{aligned} \operatorname{cov}(z_i, X_i) &= E[z_i - E(z_i)][X_i - E(X_i)] \\ &= E(u_i - \beta w_i)(w_i) \qquad \text{利用方程(13.5.7)} \\ &= E(-\beta w_i^2) \\ &= -\beta\sigma_w^2 \end{aligned} \tag{13.5.9}$$

所以，方程（13.5.8）中的解释变量与误差项是相关的，从而违背了经典线性回归模型中的关键性假定：解释变量与随机干扰项无关。如果这一假定被破坏，则可以证明，**OLS 估计量不仅是偏误的而且是非一致性的，即令样本容量 n 无限增大，OLS 估计量仍有偏误。**[①]

附录 13A 第 13A.3 节表明了对于模型（13.5.8）有

$$\operatorname{plim}\hat{\beta} = \beta\left[\frac{1}{1+\sigma_w^2/\sigma_{X^*}^2}\right] \tag{13.5.10}$$

其中 σ_w^2 和 $\sigma_{X^*}^2$ 分别是 w_i 和 X^* 的方差，而 $\operatorname{plim}\hat{\beta}$ 指 β 的概率极限。

因为预期括号内的项会小于 1（为什么?），故方程（13.5.10）表明即使样本容量无限地增大，也不收敛于 β。实际上，若假定 β 为正，$\hat{\beta}$ 将低估 β，也就是它偏向于零。当然，如果 X 中没有测量误差（即 $\sigma_w^2=0$），$\hat{\beta}$ 将给出 β 的一个一致（性）估计量。

因此，当测量误差出现在解释变量中时，将使参数的一致性估计成为不可能，这就给我们提出一个严峻的问题。当然，如我们曾看到的，如果测量误差仅出现于因变量之中，估计量仍是无偏并从而也是一致性的。如果测量误差出现在解释变量

① 如附录 A 所示，$\hat{\beta}$ 是 β 的一致性估计量，随着 n 无限增大，$\hat{\beta}$ 的抽样分布最终收敛到 β。这可技术性地表述为 $\operatorname{plim}_{n\to\infty}\hat{\beta}=\beta$。又如附录 A 所指出，一致性是大样本性质。当一个估计量的有限或小样本性质（如无偏性）不能确定时，一致性常被用来研究该估计量的性态。

中，那怎么办？回答并不容易。一个极端情形是假定 σ_w^2 相对 $\sigma_{X^*}^2$ 较小，以致为了一切实际的目的，我们都可以“假定没有”测量误差并照常进行 OLS 估计。当然，这里的困难在于 σ_w^2 和 $\sigma_{X^*}^2$ 是不易观察或测量的，因此也就无法判断其相对大小。

另一补救建议是利用这样的**工具**（instrumental）或**代理变量**（proxy variables）：它们虽与原始 X 变量高度相关，却与方程误差和测量误差项（即 u_i 和 w_i）都不相关。如果能找到这样的代理变量，我们就能得到 β 的一个一致估计。但这种工作说比做容易得多，实际上不容易找到一个好的代理变量；我们常常陷入一种埋怨天气不好而又无能为力的境况。另外，要弄清楚所选工具变量是否确实独立于误差项 u_i 和 w_i 也是不容易的。

文献中还有解决问题的其他建议。[①] 但大多数都是针对某种给定情况而设计的，并且以限制性很强的假定为基础。对于测量误差的问题，确实没有令人满意的答案，这就是为什么把数据观测得尽可能准确如此重要。

例 13.2　一个例子

我们构造一个例子以突出说明前面的论点，作为本节的结束。

表 13—2 给出真实消费支出 Y^*，真实收入 X^*，观测消费支出 Y 及观测收入 X 的假想数据。此表还说明这些变量是怎样观测的。[②]

表 13—2　真实消费支出 Y^*，真实收入 X^*，观测消费支出 Y 及观测收入 X 的假想数据

（单位：美元）

Y^*	X^*	Y	X	ϵ	w	u
75.466 6	80.00	67.601 1	80.094 0	−7.865 5	0.094 0	2.466 6
74.980 1	100.00	75.443 8	91.572 1	0.463 6	−8.427 9	−10.019 9
102.824 2	120.00	109.695 6	112.140 6	6.871 4	2.140 6	5.824 2
125.765 1	140.00	129.415 9	145.596 9	3.650 9	5.596 9	16.765 1
106.503 5	160.00	104.238 8	168.557 9	−2.264 7	8.557 9	−14.496 5
131.431 8	180.00	125.831 9	171.479 3	−5.599 9	−8.520 7	−1.568 2
149.369 3	200.00	153.992 6	203.536 6	4.623 3	3.536 6	4.369 3
143.862 8	220.00	152.920 8	222.853 3	9.057 9	2.853 3	−13.137 2
177.521 8	240.00	176.334 4	232.987 9	−1.187 4	−7.012 0	8.521 8
182.274 8	260.00	174.525 2	261.181 3	−7.749 6	1.181 3	1.274 8

注：假定 X^* 数据是给定的，在推算其他变量时，我们做了如下假定：(1)$E(u_i) = E(\epsilon_i) = E(w_i) = 0$；(2)$\text{cov}(X,u) = \text{cov}(X,\epsilon) = \text{cov}(u,\epsilon) = \text{cov}(w,u) = \text{cov}(\epsilon,w) = 0$；(3)$\sigma_u^2 = 100$，$\sigma_\epsilon^2 = 36$，和 $\sigma_w^2 = 36$；和(4)$Y_i^* = 25 + 0.6X_i^* + u_i$，$Y_i = Y_i^* + \epsilon_i$，和 $X_i = X_i^* + w_i$。

① 参看 Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods*, Springer-Verlag, New York, 1984, pp. 273-277。还参看 Kennedy, op. cit., pp. 138-140，有关加权回归以及工具变量的讨论。G. S. Maddala, *Introduction to Econometrics*, 3d ed., John Wiley & Sons, New York, 2001, pp. 437-462. Quirino Paris, "Robust Estimators of Errors-in-Variables Models: Part Ⅰ," Working Paper No. 04-007, 200, Department of Agricultural and Resource Economics, University of California at Davis, August 2004.

② 感谢怀特对此例的构造，参看 *Computer Handbook Using SHAZAM*，用以配合 Damodar Gujarati, *Basic Econometrics*, September 1985, pp. 117-121。

仅因变量 Y 有测量误差。根据所给数据，真实消费函数是：

$$\hat{Y}_i^* = 25.00 + 0.600\,0X_i^* \qquad (13.5.11)$$
$$(10.477)\ (0.058\,4)$$
$$t = (2.386\,1)\ (10.276) \quad R^2 = 0.929\,6$$

而与此相比，如果用 Y_i 代替 Y_i^*，则得到：

$$\hat{Y}_i = 25.00 + 0.600\,0X_i^* \qquad (13.5.12)$$
$$(12.218)\ (0.068\,1)$$
$$t = (2.046\,1)\ (8.811\,8) \quad R^2 = 0.906\,6$$

如这些结果所表明，并借助于理论，估计的系数不变。因变量中出现测量误差的唯一影响，是估计的系数标准误有变大的倾向［见方程（13.5.5)］，这从方程（13.5.12）可清楚地看到。顺便指出，方程（13.5.11）和（13.5.12）之所以有相同的回归系数，是因为样本的生成有意地配合了测量误差模型中的那些假定。

X 中的测量误差。已知真实回归是（13.5.11)。现假使我们不用 X_i^* 而用 X_i（注：实际上 X_i^* 近于不可观测)。回归结果如下：

$$\hat{Y}_i^* = 25.992 + 0.594\,2X_i \qquad (13.5.13)$$
$$(11.081\,0)\ (0.061\,7)$$
$$t = (2.345\,7)\ (9.627\,0) \quad R^2 = 0.920\,5$$

这些结果均与理论一致——当解释变量有测量误差时，所估系数是有偏误的。庆幸的是，本例中的偏误比较小——由方程（13.5.10）显见，偏误依赖于 $\sigma_w^2 / \sigma_{X^*}^2$，而在数据的产生中我们假定 $\sigma_w^2 = 36$ 和 $\sigma_{X^*}^2 = 3\,667$，从而导致较小的偏误因子，仅约为 0.98%（=36/3 667)。

当 Y 和 X 都有测量误差时，也就是我们所做的不是 Y_i^* 对 X_i^* 的回归，而是 Y_i 对 X_i 的回归，会出现什么后果？这个问题留给读者去探讨（参见习题 13.23)。

13.6 对随机误差项不正确的设定

研究者所面临的一个常见问题是，对误差项 u_i 进入回归模型的设定。由于误差项不能直接观测到，所以就不容易确定它进入模型的形式。为看出这一点，让我们回到方程（13.2.8）和（13.2.9）所给出的模型中。为便于说明，我们在那里已经假定了模型中不存在截距项。在这里，我们进一步假定方程（13.2.8）中的 u_i 使 $\ln u_i$ 满足通常的 OLS 假定。

如果我们假定方程（13.2.8）是“正确”模型，但我们估计的是（13.2.9)，结果会怎么样呢？附录 13A 第 13A.4 节证明了，若 $\ln u_i \sim N(0, \sigma^2)$，则

$$u_i \sim \text{log normal}[e^{\sigma^2/2}, e^{\sigma^2}(e^{\sigma^2} - 1)] \qquad (13.6.1)$$

因此

$$E(\hat{\alpha}) = \beta e^{\sigma^2/2} \qquad (13.6.2)$$

其中 e 为自然对数的底。

如你所见，$\hat{\alpha}$ 是一个有偏的估计量，因为其均值不等于真实的 β。

在有关非线性参数回归模型的章节中，我们将对随机误差项的设定做更多的探讨。

13.7 嵌套与非嵌套模型

在进行设定检验时，区分**嵌套**（nested）和**非嵌套模型**（non-nested models）很有好处。为说明两者的差别，考虑以下的模型：

$$\text{模型 A：}\quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

$$\text{模型 B：}\quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

我们说模型 B 被嵌套在模型 A 之中，因为模型 B 是模型 A 的一个特殊情形：如果我们估计模型 A，然后检验假设 H_0：$\beta_4 = \beta_5 = 0$，并且不拒绝它（比方说基于 F 检验）[①]，那么模型 A 就简化为模型 B。若我们在模型 B 中增加变量 X_4，那么模型 A 在 $\beta_5 = 0$ 时就简化为模型 B；这里，我们只用 t 检验来检验 X_5 的系数为零的假设。

我们前面讨论过的设定误差检验和第 8 章中讨论过的约束 F 检验在本质上都属于这种嵌套假设检验，只是我们没有这么称呼而已。

现在考虑如下的模型：

$$\text{模型 C：}\quad Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$$

$$\text{模型 D：}\quad Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + v_i$$

其中 X 和 Z 各代表不同的变量。我们说模型 C 和模型 D 是**非嵌套的**（non-nested），因为不能把一个作为另一个的特殊情形推导出来。经济学与其他科学一样，解释同一现象会有多种争持不下的理论。例如货币主义者强调货币在解释 GDP 变化中的作用，而凯恩斯学派则用政府支出的变化去解释 GDP。

这里需要指出，你可以使模型 C 和模型 D 包含相同的回归元，比如模型 D 中可以包含 X_3，而模型 C 中可以包含 Z_2。可即便如此，它们仍是非嵌套模型，因为模型 C 没有包含 Z_3，而模型 D 没有包含 X_2。

即便模型的变量完全一样，函数形式也可能使两个模型成为非嵌套模型。比如，考虑模型

$$\text{模型 E：}\quad Y_i = \beta_1 + \beta_2 \ln Z_{2i} + \beta_3 \ln Z_{3i} + w_i$$

模型 D 和模型 E 仍是非嵌套模型，因为你不能把其中某个模型作为另一个模型的特殊情形而推导出来。

因为我们前面已经看过了对嵌套模型的检验（t 和 F 检验），所以在接下来的一节中，我们将讨论对非嵌套模型（即我们前面提到的模型误设误差）的某些检验。

① 更一般地，可使用我们在第 8 章曾简要讨论过的似然比检验、瓦尔德检验和拉格朗日乘数检验。

13.8 非嵌套假设的检验

根据哈维[①]，检验非嵌套假设的方法大体上分为两类：（1）**判别方法**（discrimination approach），给定两个或多个相争持模型，我们根据某些拟合优度准则选择其一；以及（2）**辨识方法**（discerning approach）（为本书作者用词），在考察一个模型时须顾及其他模型所提供的信息。下面扼要地解释这些方法。

□ 判别方法

考虑 13.7 节中模型 C 和模型 D。由于这两个模型具有相同的因变量，所以我们就可依据诸如我们曾讨论过的 R^2 或调整 R^2 之类的拟合优度准则，在两（或多）个模型之间作出选择。但必须牢记，在比较两（或多）个模型时，回归子必须相同。除这些准则之外，还有其他的准则可以使用，其中包括**赤池信息准则**（Akaike's information criterion，AIC）、**施瓦茨信息准则**（Schwarz's information criterion，SIC）和**马娄斯 C_p 准则**（Mallow's C_p criterion）等。我们将在 13.9 节讨论这些准则。多数现代统计软件包在其例行回归程序中都添加了这些准则中的一个或多个。在本章的最后一节，我们将利用一个引申的例子来说明这些准则。基于这些准则中的一个或多个，最终选择的模型具有最高的 $\bar{R}^2$ 值或最低的 AIC 或 SIC 值等。

□ 辨识方法

非嵌套 F 检验或包容 F 检验。考虑 13.7 节中模型 C 和模型 D，如何在这两个模型之间进行选择呢？为此，假设我们估计如下的嵌套或糅合模型：

$$\text{模型 F：}\quad Y_i = \lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \lambda_4 Z_{2i} + \lambda_5 Z_{3i} + u_i$$

注意模型 F 嵌套或包含了模型 C 和模型 D。但模型 C 并不嵌套于模型 D 中，模型 D 也不嵌套于模型 C 中。因此它们属于非嵌套模型。

现在如果模型 C 是正确的，则 $\lambda_4 = \lambda_5 = 0$，而如果模型 D 是正确的，则 $\lambda_2 = \lambda_3 = 0$。用通常的 F 检验就可以做这个检验，非嵌套 F 检验由此得名。

然而，这种检验程序却带来一些问题。首先，如果 X 与 Z 高度相关，则如在多重共线性一章中所看到的，很可能一个或多个 λ 系数在统计上不显著，尽管基于 F 检验我们有可能拒绝所有斜率系数同时为零的（联合）假设。就此情形，我们无法决定模型 C 抑或模型 D 是正确的模型。其次，还有另一个问题，假使我们选取模型 C 作为参考假设或模型，并发现它的所有系数都是显著的。我们把一个或两个 Z 变量加到模型中，并通过 F 检验发现它对解释平方和（ESS）的增补贡献是不显著的，

① Andrew Harvey, *The Econometric Analysis of Time Series*, 2d ed., The MIT Press, Cambridge, Mass., 1990, Chapter 5.

因此我们就决定选择模型 C。

但假如我们反过来选择模型 D 作为参考模型，并发现它的所有系数也都是显著的，而当我们把一个或两个 X 变量加到此模型并再次用 F 检验时，我们又会发现它对 ESS 的增补贡献也是不显著的，于是我们又会把模型 D 选做正确模型。因此，"参考假设的选择竟能决定模型选择的结果"[①]，尤其是在相互争持的诸回归元中有严重的多重共线性的情况下。最后，人为的嵌套模型 F 可能缺乏经济意义。

例 13.3 一个说明性例子：圣路易斯模型

为了明确名义 GNP 的变化是由货币供给的变化来解释（货币主义），还是由政府支出的变化来解释（凯恩斯主义），我们考虑如下模型：

$$\begin{aligned}\dot{Y}_t &= \alpha+\beta_0\dot{M}_t+\beta_1\dot{M}_{t-1}+\beta_2\dot{M}_{t-2}+\beta_3\dot{M}_{t-3}+\beta_4\dot{M}_{t-4}+u_{1t}\\ &= \alpha+\sum_{i=0}^{4}\beta_i\dot{M}_{t-i}+u_{1t}\end{aligned} \tag{13.8.1}$$

$$\begin{aligned}\dot{Y}_t &= \gamma+\lambda_0\dot{E}_t+\lambda_1\dot{E}_{t-1}+\lambda_2\dot{E}_{t-2}+\lambda_3\dot{E}_{t-3}+\lambda_4\dot{E}_{t-4}+u_{2t}\\ &= \gamma+\sum_{i=0}^{4}\lambda_i\dot{E}_{t-i}+u_{2t}\end{aligned} \tag{13.8.2}$$

其中 $\dot{Y}_t = t$ 时刻名义 GNP 的增长率

$\dot{M}_t = t$ 时刻货币供给量（指 M_1）的增长率

$\dot{E}_t = t$ 时刻充分或高就业下政府支出的增长率

顺便指出方程（13.8.1）和（13.8.2）都属于**分布滞后模型**（distributed-lag models）。这种模型是第 17 章要详加讨论的主题。目前只需知道货币供给或政府支出的单位变化对 GNP 的影响是分布在一段期间而不是瞬时的。

因为不容易先验地在两个模型之间作出选择，故将两者糅合在一起，如下所示：

$$\dot{Y}_t = \text{constant}+\sum_{i=0}^{4}\beta_i\dot{M}_{t-i}+\sum_{i=0}^{4}\lambda_i\dot{E}_{t-i}+u_{3t} \tag{13.8.3}$$

这个嵌套模型就是用以表达并估计著名的圣路易斯联邦储备银行（一个有货币学派倾向的银行）模型的一个形式。此模型给出对于美国在 1953 年第一季度至 1976 年第四季度期间的估计结果如下（括号中是 t 比率）[②]：

系数	估计值		系数	估计值			
β_0	0.40	(2.96)	λ_0	0.08	(2.26)		
β_1	0.41	(5.26)	λ_1	0.06	(2.52)		
β_2	0.25	(2.14)	λ_2	0.00	(0.02)		
β_3	0.06	(0.71)	λ_3	−0.06	(−2.20)		(13.8.4)
β_4	−0.05	(−0.37)	λ_4	−0.07	(−1.83)		
$\sum_{i=0}^{4}\beta_i$	1.06	(5.59)	$\sum_{i=0}^{4}\lambda_i$	0.03	(0.40)	$R^2 = 0.40$	$d = 1.78$

① Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods*, Springer-Verlag, New York, 1984, p. 416.

② Keith M. Carlson, "Does the St. Louis Equation Now Believe in Fiscal Policy?" *Review, Federal Reserve Bank of St. Louis*, vol. 60, no. 2, February 1978, p. 17, table Ⅳ.

这些结果能表明一个模型优于另一个模型吗？如果我们考虑 $\dot{M}$ 和 $\dot{E}$ 的单位变化对 $\dot{Y}$ 的累积效应，我们分别得到 $\sum_{i=0}^{4}\beta_i = 1.06$ 和 $\sum_{i=0}^{4}\lambda_i = 0.03$，前者是统计显著的，而后者不是。这种比较会倾向于支持货币主义者的主张，即货币供给的变化决定着（名义）GNP 的变化。如何严格地评价这一主张，且留给读者作为习题。

戴维森-麦金农 J 检验。[1] 由于刚才列出的非嵌套 F 检验程序中的种种问题，人们提出了另外的检验。其中之一是**戴维森-麦金农 J 检验**（Davidson-MacKinnon J test）。为说明此检验，假使我们要比较假设或模型 C 和模型 D。**J 检验**的步骤如下：

1. 估计模型 D 并由此得到 Y 的估计值 $\hat{Y}_i^D$。

2. 将步骤 1 中得到的估计值作为另一回归元增补到模型 C 中，并随即估计以下模型：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 \hat{Y}_i^D + u_i \tag{13.8.5}$$

其中 $\hat{Y}_i^D$ 的值得自步骤 1。此模型是韩德瑞方法论中的**兼容性原则**（encompassing principle）之一例。

3. 用 t 检验对假设 $\alpha_4 = 0$ 进行检验。

4. 如果假设 $\alpha_4 = 0$ 不被拒绝，就可接受（即不拒绝）模型 C 为真模型。因为方程（13.8.5）代表不为模型 C 所含有的变量影响，也就是说它所包含的 $\hat{Y}_i^D$ 并没有增加模型 C 原有的解释能力。换句话说，模型 D 不含有足以改进模型 C 的任何额外信息，故模型 C 兼容模型 D。类似地推理，如果虚拟假设被拒绝，则模型 C 不会是真模型。（为什么?）

5. 现在把假设或模型 C 和模型 D 的作用颠倒过来，先估计模型 C，并用由此得到的 Y 估计值作为回归元增补到模型 D 中，重复步骤 4，以决定是否认为模型 D 胜过模型 C。更具体而言，我们估计如下模型：

$$Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 \hat{Y}_i^C + u_i \tag{13.8.6}$$

其中 $\hat{Y}_i^C$ 是得自模型 C 的 Y 的估计值。现在假设检验 $\beta_4 = 0$。如该假设不被拒绝，则选择模型 D 而不选模型 C。如假设 $\beta_4 = 0$ 被拒绝，则由于模型 D 没有改进模型 C 的表现*，故选模型 C 而不选模型 D。

J 检验虽然直观上比较可取，却也遇到一些问题。由于方程（13.8.5）和（13.8.6）的两个检验是独立操作的，故有下述可能结局：

	假设：$\alpha_4 = 0$	
假设：$\beta_4 = 0$	不拒绝	拒绝
不拒绝	同时接受 C 和 D	接受 D 而拒绝 C
拒绝	接受 C 而拒绝 D	同时拒绝 C 和 D

---

① R. Davidson and J. G. MacKinnon, "Several Tests for Model Specification in the Presence of Alternative Hypothesis," *Econometrica*, vol. 49, 1981, pp. 781-793.

* 而模型 C 改进了模型 D 的表现。——译者注

如上表所示，如果 J 检验程序导致同时接受或同时拒绝两模型，我们就得不到明确的答案。当两模型均被拒绝时，任一模型都无助于对 Y 行为的解释。同理，若两模型均被接受，则有如克曼塔所说："显然，数据还未充分到足以辨别两个假设（模型）的地步。"①

J 检验的另一个问题是，当我们用 t 统计量去检验模型（13.8.5）和（13.8.6）中估计的 Y 变量的显著性时，t 统计量只是渐近地在大样本中遵从标准正态分布。因此，在小样本中，J 检验会过多地拒绝真假设或真模型，从而它不是（在统计意义上）很有功效的。

例 13.4　个人消费支出与个人可支配收入

为说明 J 检验考虑表 13—3 中的数据，该表给出美国 1970—2005 年期间个人人均消费支出（PPCE）和个人人均可支配收入（PDPI）数据，均以 2008 年美元计算。现考虑以下两个竞争模型：

$$\text{模型 A：}\quad \text{PPCE}_t = \alpha_1 + \alpha_2 \text{PDPI}_t + \alpha_3 \text{PDPI}_{t-1} + u_t \tag{13.8.7}$$

$$\text{模型 B：}\quad \text{PPCE}_t = \beta_1 + \beta_2 \text{PDPI}_t + \beta_3 \text{PPCE}_{t-1} + u_t \tag{13.8.8}$$

表 13—3　1970—2005 年美国个人人均消费支出与个人人均可支配收入　（单位：2008 年美元）

年份	PPCE	PDPI	年份	PPCE	PDPI
1970	3 162	3 587	1988	13 685	15 297
1971	3 379	3 860	1989	14 546	16 257
1972	3 671	4 140	1990	15 349	17 131
1973	4 022	4 616	1991	15 722	17 609
1974	4 364	5 010	1992	16 485	18 494
1975	4 789	5 498	1993	17 204	18 872
1976	5 282	5 972	1994	18 004	19 555
1977	5 804	6 517	1995	18 665	20 287
1978	6 417	7 224	1996	19 490	21 091
1979	7 073	7 967	1997	20 323	21 940
1980	7 716	8 822	1998	21 291	23 161
1981	8 439	9 765	1999	22 491	23 968
1982	8 945	10 426	2000	23 862	25 472
1983	9 775	11 131	2001	24 722	26 235
1984	10 589	12 319	2002	25 501	27 164
1985	11 406	13 037	2003	26 463	28 039
1986	12 048	13 649	2004	27 937	29 536
1987	12 766	14 241	2005	29 468	30 458

资料来源：*Economic Report of the President*，2007.

模型 A 表示 PPCE 依赖于当前的和前期的 PDPI；这是所谓**分布滞后模型**（distributed-lag

① Jan Kmenta，op. cit.，p. 597.

model）之一例（参看第17章）。模型B设想PPCE依赖于当前的PDPI以及前期的PPCE；此模型代表以**自回归模型**（autoregressive model）为名的一个模型（参看第17章）。模型中引进PPCE滞后值的理由是为了反映人们的消费惯性或习惯的持久性。

这两个模型各自的估计结果如下：

$$\text{模型 A：}\quad \widehat{PPCE_t} = -606.6347 + 0.6170\,PDPI_t + 0.3530\,PDPI_{t-1} \tag{13.8.9}$$
$$t = (-3.8334)\ (2.5706)\qquad (1.4377)$$
$$R^2 = 0.9983 \qquad d = 0.2161$$

$$\text{模型 B：}\quad \widehat{PPCE_t} = 76.8947 + 0.2074\,PDPI_t + 0.8104\,PPCE_{t-1} \tag{13.8.10}$$
$$t = (0.7256)\ (2.6734)\qquad (9.7343)$$
$$R^2 = 0.9996 \qquad d = 0.9732$$

如果按照最高R^2准则的判别方法在两者之间进行选择，我们可能会选择模型B（13.8.10），因为它的R^2略高于模型A（13.8.9）；而且，在模型B（13.8.10）中，两个解释变量都是统计显著的，而在模型A（13.8.9）中只有当前PDPI是统计上显著的。（但需当心共线性问题！）但从预测角度看，这两个R^2估计值之间没有多大区别。

为了应用J检验，我们假定模型A是虚拟（null）或维持（maintained）假设，而模型B是对立或备择（alternative）假设。按照上面讨论的J检验的步骤，用得自模型（13.8.10）的PPCE估计值作为模型A中的一个新增回归元，我们将得到以下回归结果：

$$\widehat{PPCE_t} = -35.17 + 0.2762\,PDPI_t - 0.5141\,PDPI_{t-1} + 1.2351\ \widehat{PPCE_t^B} \tag{13.8.11}$$
$$t = (-0.43)\ (2.60)\qquad (-4.05)\qquad (12.06)$$
$$R^2 = 1.00 \qquad d = 1.5205$$

其中方程（13.8.11）右边的$\widehat{PPCE_t^B}$是得自模型B（13.8.10）的PPCE估计值。既然此变量的系数在统计上是显著的（具有极高的t统计量12.06），按照J检验程序，我们必须拒绝模型A而接受模型B。

再假定模型B是维持假设而模型A是备择假设，按照和前面完全一样的程序，我们得到如下结果：

$$\widehat{PPCE_t} = -823.7 + 1.4309\,PDPI_t + 1.0009\,PPCE_{t-1} - 1.4563\ \widehat{PPCE_t^A} \tag{13.8.12}$$
$$t = (-3.45)\ (4.64)\qquad (12.06)\qquad (-4.05)$$
$$R^2 = 1.00 \qquad d = 1.5205$$

其中方程（13.8.12）右边的$\widehat{PPCE_t^A}$是得自模型A（13.8.9）的PPCE估计值。但在此回归中，右端$\widehat{PPCE_t^A}$的系数也是统计显著的（t统计量为-4.05）。这一结果表明，我们现在应拒绝模型B，而接受模型A！

以上分析告诉我们，为了解释1970—2005年美国的个人人均消费支出行为，不见得哪个模型是特别有用的。当然，我们仅仅考虑了两个相互媲美的模型，其实还可以比较多个模型。可以把J检验程序推广到多个模型的比较上，尽管分析上会立即变得复杂起来。

本例生动地表明为什么CLRM要假定分析中所用的回归模型是被正确设定的。显然，在做一个模型时，全力注意“被模型化”的现象是极其重要的。

模型选择的其他检验。刚才讨论的J检验只是模型选择的许多检验中的一种，还有**考克斯检验**（Cox test）、**JA检验**（JA test）、**P检验**（P test）、**米松-理查德兼容性检验**（Mizon-Richard encompassing test），以及这些检验的变型。显然，我们

不能指望一一讨论这些专门的检验，有兴趣的读者可参阅散见于各个注释中所引的文献。[①]

13.9 模型选择准则

在本节，我们讨论几个已用于在互相竞争的模型之间做出选择和/或从预测的角度对模型进行比较的准则。我们在此区分**样本内**（in-sample）预测和**样本外**（out-of-sample）预测。样本内预测本质上告诉我们，所选择的模型在给定样本中对数据拟合得如何。样本外预测则考虑到一个拟合模型在给定回归元值情况下对回归子未来值的预测。

有几个准则用于这一目的。具体而言，我们讨论如下准则：（1）R^2，（2）调整 $R^2(=\bar{R}^2)$，（3）赤池信息准则，（4）施瓦茨信息准则，（5）马娄斯 C_p 准则，及（6）预测 χ^2 准则。所有这些准则都是为了最小化残差平方和（RRS）（或提高 R^2 的值）。但除了第一个准则之外，准则（2）、（3）、（4）和（5）都对包含回归元个数的不断增加进行了惩罚。因此，在模型的拟合优度与其复杂性（由回归元个数来判断）之间有一种权衡取舍的关系。

□ R^2 准则

我们知道，对一个回归模型拟合优度的度量指标之一就是 R^2，其定义为

$$R^2=\frac{\text{ESS}}{\text{TSS}}=1-\frac{\text{RSS}}{\text{TSS}} \tag{13.9.1}$$

如此定义的 R^2 必然介于 0 和 1 之间。R^2 越接近 1，拟合得越好。但 R^2 有一些问题。首先，它度量的是样本内拟合优度，即度量了给定样本中所估计的 Y 值与其实际值有多么接近。它不能保证对样本外观测也能很好地预测。其次，在将两个或多个 R^2 进行比较时，因变量或回归子必须相同。最后，也是最重要的一点，当模型中添加越来越多的变量时，R^2 总不会变小。因此，通过单纯地向模型中添加更多的变量，玩“最大化 R^2”的游戏很诱人。当然，在模型中添加越来越多的变量的确能使 R^2 变大，但同时也使预测误差的方差变大。

□ 调整 R^2 准则

作为对增加回归元来提高 R^2 值的一种惩罚，亨利·瑟尔提出我们在第 7 章中所研究的调整 R^2，记为 $\bar{R}^2$。记得

$$\bar{R}^2=1-\frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}=1-(1-R^2)\frac{n-1}{n-k} \tag{13.9.2}$$

① 也可参见 Badi H. Baltagi, *Econometrics*, Springer, New York, 1998, pp. 209-222。

从这个公式可以看出，$\bar{R}^2 \leqslant R^2$，表明调整 R^2 是如何对增加更多的回归元进行惩罚的。我们在第 8 章曾指出，与 R^2 不同，调整 R^2 只有在所添加变量的 t 值的绝对值大于 1 时才会增加。比较而言，$\bar{R}^2$ 是一个比 R^2 更好的度量指标。但同样记住，为了能进行比较，被比较模型的回归子仍必须相同。

□ 赤池信息准则（AIC）

在 AIC 准则中，进一步对模型中增加回归元进行了惩罚，AIC 的定义为

$$\text{AIC} = e^{2k/n} \frac{\sum a_i^2}{n} = e^{2k/n} \frac{\text{RSS}}{n} \tag{13.9.3}$$

其中 k 为回归元的个数（包括截距项），n 为观测次数。为了数学上方便起见，把方程（13.9.3）写成

$$\ln \text{AIC} = \frac{2k}{n} + \ln\left(\frac{\text{RSS}}{n}\right) \tag{13.9.4}$$

其中 lnAIC 为 AIC 的自然对数，$2k/n$ 为惩罚因子。有些教材和软件只以其对数定义 AIC，所以就没有必要再在 AIC 的前面加上 ln。如你从这个公式中所见，与 $\bar{R}^2$ 相比，AIC 对添加更多回归元施加了更严厉的惩罚。在比较两个或多个模型时，具有最低 AIC 值的模型优先。AIC 的优越性之一在于，它不仅适用于样本内预测，还适用于预测一个回归模型在样本外的表现。此外，它对嵌套和非嵌套模型都适用，甚至还可以用于决定 AR(p) 模型中的滞后长度。

□ 施瓦茨信息准则（SIC）

与 AIC 的思想类似，SIC 准则的定义为

$$\text{SIC} = n^{k/n} \frac{\sum a_i^2}{n} = n^{k/n} \frac{\text{RSS}}{n} \tag{13.9.5}$$

或以对数形式表示为

$$\ln \text{SIC} = \frac{k}{n} \ln n + \ln\left(\frac{\text{RSS}}{n}\right) \tag{13.9.6}$$

其中 $(k/n)\ln n$ 为惩罚因子。通过比较方程（13.9.6）和（13.9.4）明显可以看到，SIC 施加的惩罚比 AIC 更严厉。与 AIC 相似，SIC 的值越低的模型就越好。而且与 AIC 一样，SIC 可以用于比较一个模型在样本内或样本外的预测表现。

□ 马娄斯 C_p 准则

假设我们有一个含有包括截距在内 k 个回归元的模型。和平常一样，令 $\hat{\sigma}^2$ 为真

实 σ^2 的估计量。但假设我们只选择 p（其中 $p \leqslant k$）个回归元，并从使用这 p 个回归元的回归中得到 RSS。令 RSS_p 表示使用 p 个回归元的残差平方和。现在，马娄斯（C. P. Mallows）便提出了模型选择的如下准则，被称为 C_p 准则：

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} - (n - 2p) \tag{13.9.7}$$

其中 n 为观测次数。

我们知道 $E(\hat{\sigma}^2)$ 是真实 σ^2 的一个无偏估计量。现在，如果含有 p 个回归元的模型拟合得足够充分，那么可以证明[①]：$E(\text{RSS}_p) = (n-p)\sigma^2$。因此，近似有

$$E(C_p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n - 2p) \approx p \tag{13.9.8}$$

在根据 C_p 准则选择一个模型时，我们想找到一个 C_p 值很低（约为 p）的模型。换句话说，根据节俭性原则，我们将选择一个含有 p 个回归元（$p < k$）并相当好地拟合数据的模型。

实践中，人们通常将从方程（13.9.7）计算而来的 C_p 对 p 进行描点。一个“充分”的模型将作为一个与 $C_p = p$ 线接近的点而出现，如图 13—3 所示。此图表明，模型 A 因比模型 B 更接近 $C_p = p$ 线而比模型 B 更受欢迎。

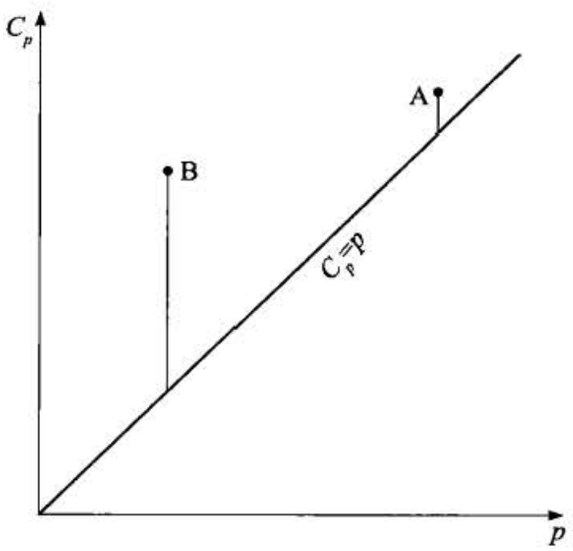


图 13—3　马娄斯 C_p 描点图

□ 对模型选择准则的一句忠告

我们已经讨论了几种模型选择的准则，但我们应该把这些准则看成是对我们在本章中讨论的各种设定检验的补充。以上讨论的某些准则只是纯粹描述性的，或许

① Norman D. Draper and Harry Smith, *Applied Regression Analysis*, 3d ed., John Wiley & Sons, New York, 1998, p. 332. 参见此书中某些关于 C_p 的例子。

没有什么很强的理论性，其中还有些准则易于受到数据挖掘的指控。可尽管如此，这些准则仍频繁地被实践者所使用，所以读者应该对此有所察觉。这些准则中没有哪一个肯定优于其他准则。[①] 大多数现代软件包现在都包括了 R^2、调整 R^2、AIC 和 SIC。尽管马娄斯 C_p 很容易根据定义计算出来，可软件仍没有例行给出它。

□ 用于预测的 χ^2

假使我们有一个基于 n 次观测的回归模型，并假使想利用它来预测回归子在另外 t 次观测中的（均）值。其他地方曾指出，把样本数据留存一部分以分析所估计的模型对未包含进样本的观测（后样本期间的观测）所做的预测如何，这是一个好主意。

现在，预测 χ^2 检验的定义如下：

$$\text{预测 } \chi^2 = \frac{\sum_{n+1}^{n+t} \hat{u}_i^2}{\hat{\sigma}^2} \tag{13.9.9}$$

其中 $\hat{u}_i$ 表示第 i 期（$=n+1$，$n+2$，…，$n+t$）利用所拟合回归得到的参数和后样本期间回归元的值得出的预测误差。$\hat{\sigma}^2$ 是 σ^2 基于所拟合回归的通常 OLS 估计量。

如果我们假设参数值在样本期和后样本期保持不变，则可以证明方程（13.9.9）中给出的统计量服从自由度为 t 的 χ^2 分布，其中 t 表示留作预测的观测次数。如查伦扎（Charemza）和达德曼（Deadman）所指出，预测 χ^2 检验具有弱统计功效（weak statistical power），意味着它正确地拒绝一个错误的虚拟假设的概率很低，因此这个检验应该用作象征性的而非决定性的检验。[②]

13.10 计量经济建模的其他专题

正如本章的引言中所讲，计量经济建模和诊断检验的专题如此广泛而又具有发展前景，所以应该用一本专著来写这个专题。我们在上一节已经触及该领域的一些重大主题。在本节，我们考虑一些研究者可能会发现在实践中很有用的其他性质。具体而言，我们考虑如下专题：（1）**异常数据**（outliers）、**杠杆数据**（leverage）和**有影响力的数据**（influence）；（2）**递归最小二乘法**（recursive least squares，REL）；（3）**邹至庄预测失灵检验**（Chow's prediction failure test）。不可避免，对每个专题的讨论都将十分简单。

① 对这个专题的一个有益讨论可参见 Francis X. Diebold, *Elements of Forecasting*, 2d ed., South Western Publishing, 2001, pp. 83-89。总体上，代伯德（Diebold）建议的是 SIC 准则。

② Wojciech W. Charemza and Derek F. Deadman, *New Directions in Econometric Practice: A General to Specific Modelling, Cointegration and Vector Autoregression*, 2d ed., Edward Elgar Publishers, 1997, p. 30, and pp. 250-252.

□ 异常数据、杠杆数据和有影响力的数据[①]

记得在最小化残差平方和时，OLS 对样本中的每个观测都赋予相等的权重。但由于异常数据、杠杆数据和有影响力的数据这三种特殊类型的数据点的出现，使每个观测对回归结果的影响可能不同。了解它们并掌握它们如何影响回归分析，对我们而言就十分重要。

在回归的背景下，一个**异常数据**可定义为一个具有“很大残差”的观测。记得 $a_i=Y_i-\hat{Y}_i$，即残差表示回归子的实际值与其从回归模型估计出来的估计值之差（或正或负）。当我们说一个残差很大时，总是相对其他的残差而言，而且这个很大的残差会因为它与估计的回归线之间垂直距离很大而立即吸引我们的注意。注意，在一个数据集中，可能有不止一个异常数据。我们在习题 11.22 中已经遇到一个这样的例子，在那个例子中，要求你在 20 个国家构成的样本中，将股票价格的百分比变化（Y）对消费者价格指数的百分比变化（X）做回归。而对智利的观测就是一个异常数据。

如果一个数据点不成比例地远离绝大部分回归元值，那就认为它表现出（高度）**杠杆性**（leverage）。一个**杠杆数据**为什么会带来问题呢？因为它能把回归线向自己拉近，由此改变回归线的斜率，所以它也能带来麻烦。如果这种情况确实发生了，那我们就称这样一个杠杆数据为一个**有影响力的数据**。从样本中去掉这样一个数据会显著地影响回归线。回到习题 11.22 中，你会发现，若你在包含智利数据的情况下将 Y 对 X 回归，那么斜率系数就为正并且“在统计上高度显著”。但如果你去掉智利这个国家的观测数据，斜率系数实际上为零。因此，对智利的观测有杠杆作用，也是一个有影响力的观测。

为了进一步弄清楚异常数据、杠杆数据和有影响力数据的性质，考虑图 13—4。[②]

我们如何处理这种数据点呢？我们应该只是把它们去掉并只考虑其余的数据点吗？德雷珀（Draper）和史密斯（Smith）认为：

> 不假思索地拒绝异常数据并不总是一个明智的选择。异常数据有时候提供了其他数据点不能提供的信息，因为它可能是因环境因素非同寻常的组合所致，而其中有些因素可能具有重要的意义，并要求进一步的研究而不是简单地拒绝。作为一个一般原则，只有在可追踪到诸如记录观测发生错误或仪器没有正确调整（在物理试验中）等原因时，才应该拒绝使用异常数据，否则就需要进行认真的研究。[③]

---

① 接下来的讨论受到如下这本书的影响：Chandan Mukherjee, Howard White, and Marc Wyuts, *Econometrics and Data Analysis for Developing Countries*, Routledge, New York, 1998, pp. 137-148。

② 节选自 John Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, California, 1997, p. 268。

③ Norman R. Draper and Harry Smith, op. cit., p. 76.

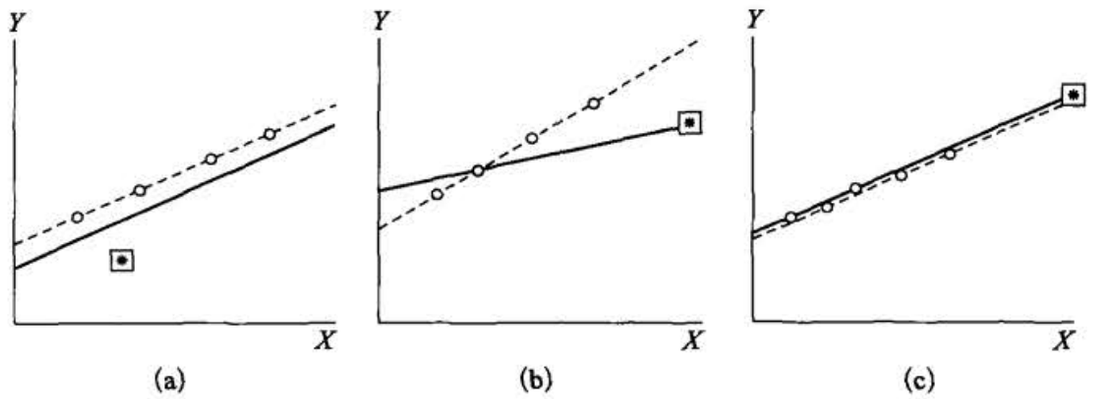


图 13—4　在每个子图中，实线表示对所有数据的 OLS 线，虚线表示去掉异常数据（用 * 表示）后的 OLS 线。在图（a）中，异常数据接近 X 的均值，具有较低的杠杆作用，对回归系数没有什么影响。在图（b）中，异常数据远离 X 的均值，具有很大的杠杆作用，并对回归系数产生明显影响。在图（c）中，异常数据具有很大的杠杆作用，但对回归系数的影响力很低，因为它基本上与其他观测位于同一条直线上。

资料来源：Adapted from John Fox, op. cit., p. 268.

有哪些检验可用于侦察异常数据和杠杆数据点呢？文献中讨论了几种检验，但我们这里不讨论它们，因为这样会使我们离题太远。[①] 诸如 SHAZAM 和 MICROFIT 之类的软件包都有侦察异常数据、杠杆数据和有影响力数据的例行程序。

□ 递归最小二乘法

我们在第 8 章考察了一个回归模型在涉及时间序列数据时的结构稳定性问题，并说明了如何使用**邹至庄检验**。明确地讲，你可能记得我们在那一章中讨论了美国在 1970—2005 年间的一个简单储蓄函数（储蓄作为收入的函数）。我们在那里还看到，储蓄—收入关系可能在 1982 年前后发生了变化。知道了这个结构转折点之后，我们就能用邹至庄检验来确定。

但如果我们不知道结构转折点又会怎么样呢？此时可以使用**递归最小二乘法**（RELS）。RELS 背后的基本思想很简单，并可以用储蓄—收入回归来解释。

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

其中 Y＝储蓄，X＝收入，样本期间为 1970—2005 年。（参见表 8—11 中的数据。）

假设我们首先使用了 1970—1974 年间的数据并估计了这个储蓄函数，得到 β_1 和 β_2 的估计值。然后，我们使用 1970—1975 年间的数据再次估计这个储蓄函数并得到两个参数的估计值。然后我们再使用 1970—1976 年间的数据重新估计储蓄模型。我

① 这里有一些容易获得的文献来源：Alvin C. Rencher, *Linear Models in Statistics*, John Wiley & Sons, New York, 2000, pp. 219-224; A. C. Atkinson, *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford University Press, New York, 1985, Chapter 3; Ashis Sen and Muni Srivastava, *Regression Analysis: Theory, Methods, and Applications*, Springer-Verlag, New York, 1990, Chapter 8; and John Fox, op. cit., Chapter 11。

们以此类推继续增加 Y 和 X 的数据点直至用完全部样本。你可以想象，每组回归都将给出 β_1 和 β_2 的一组新估计值。如果你把这些参数的估计值依次描点，将会看出估计参数是如何变化的。如果所考虑的模型在结构上是稳定的，这两个参数的估计值变化将很小，而且基本上是随机的。然而，如果参数的估计值变化明显，则意味着存在结构转折。因此，RELS 成为时间序列数据中一个常用的例行程序，因为时间序列是按照时间顺序排列的。当横截面数据按照某种“规格”或“规模”变量（如企业就业规模或资产规模等）排序时，RELS 也是一个有用的诊断工具。在习题 13.30 中，要求你对表 8—11 所给的储蓄数据应用 RELS。

诸如 SHAZAM、EViews 和 MICROFIT 之类的软件包现在都例行做递归最小二乘估计。RELS 还生成**递归残差**（recursive residuals），可作为几个诊断检验的基础。[①]

□ 邹至庄预测失灵检验

我们在第 8 章已经讨论过对结构稳定性的邹至庄检验。邹至庄还证明了，他的检验方法略加修改后还可用于对回归模型预测功效的检验。我们再次回到美国 1970—1995 年间的储蓄—收入回归。

假设我们对 1970—1981 年期间估计了储蓄—收入回归，并得到基于 1970—1981 年期间的数据估计的截距和斜率系数 $\hat{\beta}_{1,70-81}$ 和 $\hat{\beta}_{2,70-81}$。现在我们利用 1982—1995 年期间收入的实际值和 1970—1981 年期间的截距和斜率值，来预测 1982—1995 年间每一年的储蓄值。这里的逻辑是，若参数值没有发生严重的结构变化，则基于前一期间系数估计值而估计出来的 1982—1995 年间的储蓄估计值，不应该与后一期间储蓄的实际值有很大不同。当然，若后一期间储蓄的实际值与预测值之间存在巨大差异，则整个数据期间储蓄—收入关系的稳定性就值得怀疑。

储蓄的实际值与估计值之间的差别可通过 F 检验进行：

$$F = \frac{(\sum \hat{u}_t^{*2} - \sum \hat{u}_t^2)/n_2}{\sum \hat{u}_t^2/(n_1 - k)} \tag{13.10.1}$$

其中 n_1 = 初始回归所基于的第一期间（1970—1981 年）中观测的次数，n_2 = 第二期间或预测期间的观测次数，$\sum \hat{u}_t^{*2}$ = 对所有观测（$n_1 + n_2$）估计出来的 RSS，$\sum \hat{u}_t^2$ = 对前 n_1 个观测估计出来的 RSS，k 为待估计参数的个数（此例中为 2）。若误差是独立同正态分布的，则方程（13.10.1）中给出的 F 统计量服从自由度分别为 n_2 和 n_1 的 F 分布。习题 13.31 要求你应用邹至庄预测失灵检验，以分析储蓄—收入关系事实上是否发生了变化。顺便提一句，注意这个检验与前面讨论的预测 χ^2 检验的相似之处。

□ 数据缺失

在应用研究中，有时候样本中缺失一些观测数据也很常见。比如，在时间序列

① 详细情况参见 Jack Johnston and John DiNardo, *Econometric Methods*, 4th ed., McGraw-Hill, New York, 1997, pp. 117—121。

数据中，由于某些特殊情况，有些年代的数据缺乏。第二次世界大战期间有些宏观经济变量的数据就没有，或者出于战略考虑，政府不愿意提供这些数据。在横截面数据中，特别是在通过问卷调查而得到的数据中，缺失某些人的某些变量信息也很常见。在面板数据中，随着时间的推移，有些人退出了，或者有些人没有提供所有问题的信息，这也可能导致数据缺失问题。

无论什么原因，数据缺失是每个研究者时刻都要面对的问题。问题是，我们对数据缺失问题该如何处理呢？我们有对缺失数据**估计数值**（impute value）的方法吗？

这是一个不太容易回答的问题。尽管文献中有一些复杂的解决办法，但由于它们过于复杂，我们在这里就不予深究。不过，我们还是要讨论两种情形。[①] 在第一种情形中，数据缺失的原因与现有观测没有什么关系，这种情形被达奈尔（Darnell）称为“可忽略情形”。在第二种情形中，不仅现有的数据是不完备的，而且缺失的数据可能与现有数据系统性地相关。这是一个更复杂的情形，因为它可能是自选择偏误（self-selection bias）的结果，也就是说，观测数据不是真正随机搜集而来的。

在可忽略情形中，我们或许可以直接忽略缺失数据而使用现有数据。多数统计软件都自动这样处理。当然，在这种情形中，样本容量变小了，而且我们或许不能得到回归系数的精确估计值。但我们使用现有数据或许能够部分地反映出缺失数据的情况。这里我们考虑三种可能性：

1. 在总共 N 个观测中，我们有 N_1 个观测（$N_1<N$）同时具有回归子和 k 个回归元的完备数据，并分别记为 $\boldsymbol{Y}_1$ 和 $\boldsymbol{X}_1$。（$\boldsymbol{Y}_1$ 是一个由 N_1 个观测的回归子构成的列向量，$\boldsymbol{X}_1$ 是一个 k 行和 N_1 列的回归元矩阵。）

2. 有些观测（$N_2<N$）有回归子的完备数据，并记为 $\boldsymbol{Y}_2$，但有些 $\boldsymbol{X}_2$ 的观测不完备（同样是矩阵）。

3. 有些观测（$N_3<N$）没有 $\boldsymbol{Y}$ 的数据，但有 $\boldsymbol{X}$ 的完备数据，并记为 $\boldsymbol{X}_3$。

在第一种情况中，将 $\boldsymbol{Y}_1$ 对 $\boldsymbol{X}_1$ 回归将得到回归系数的无偏但可能非有效的估计值，因为我们忽略了 N_2 和 N_3 的观测。另外两种情况相当复杂，至于其解决办法，读者可从参考文献中获得。[②]

13.11 总结性的例子

我们用能够说明本章中一个或多个论点的例子来结束本章。第一个例子是利用

① 接下来的讨论基于如下著作：Adrian C. Darnell, *A Dictionary of Econometrics*, Edward Elgar Publishing, Lyne, U. K., 1994, pp. 256-258。

② 除了已经引用的参考文献外，还可参见 A. A. Afifi, and R. M. Elashoff, "Missing Observations in Multivariate Statistics," *Journal of the American Statistical Association*, vol. 61, 1966, pp. 595-604, and vol. 62, 1967, pp. 10-29。

横截面数据来确定工资，第二个例子是用时间序列数据来分析美国的真实消费函数。

□ 1. 小时工资的决定模型

为了考察哪些因素决定小时工资，我们考虑劳动经济学家比较喜欢的明瑟（Mincer）模型。这个模型形式如下[①]：

$$\ln \text{wage}_i = \beta_1 + \beta_2 \text{Edu}_i + \beta_3 \text{Exp}_i + \beta_4 \text{Fe}_i + \beta_5 \text{NW}_i + \beta_6 \text{UN}_i + \beta_7 \text{WK}_i + u_i \tag{13.11.1}$$

其中 ln wage＝小时工资的对数（美元）；

Edu＝受教育年数；

Exp＝劳动市场工作年数；

Fe＝1 对女性，＝0 对男性；

NW＝1 对非白人，＝0 对白人；

UN＝1 对工会成员，＝0 对非工会成员；

WK＝1 对不是按小时付酬的工人，＝0 对按小时付酬的工人。

对于那些不是按小时付酬的工人，小时工资是用周薪除以每周工作小时数计算的。

这个模型中还可以包含更多的变量。其中有宗教信仰、婚姻状况、6 岁以下子女数以及财富或非劳动收入等。现在，我们主要考虑方程（13.11.1）中所示的模型。

数据由 1985 年三月调查的 1 289 人构成，它是美国人口普查局（U. S. Census Bureau）定期进行的《当代人口普查》（*Current Population Survey*，*CPS*）中的一部分。这些数据最早由保罗·鲁迪（Paul Rudd）搜集而来。[②]

据经验，我们预期受教育程度和工作经历对工作有正面影响。如果存在一定程度的歧视，预期虚拟变量 Fe 和 NW 对工资有负面影响，而由于收入的不确定性，预期 UN 对工资有正面影响。

当所有虚拟变量都取值为 0 时，方程（13.11.1）便简化成

$$\ln \text{wage}_i = \beta_1 + \beta_2 \text{Edu}_i + \beta_3 \text{Exp}_i + u_i \tag{13.11.2}$$

它表示一个没有加入工会组织并按小时取酬的男性工人的工资函数。这就是基组或参照组。

现在我们来介绍回归结果（见表 13—4）并加以讨论。

首先注意到，由于 p 值都很低，所以个别地看所有的估计系数都是高度显著的。F 值也很高，从而表明所有变量是联合统计显著的。

与参照组的工人相比，一个女性工人和非白人工人的平均工资偏低。而平均而言，加入工会组织和按周计酬的工人的工资更高。

给定我们所考虑的变量，模型（13.11.1）的完备性如何？非白人女性工人可能比白人工人挣得更少吗？没有加入工会组织的非白人女性工人可能比没有加入工会

① 参见 J. Mincer，*School，Experience and Earnings*，Columbia University Press，New York，1974。

② Paul A. Rudd，*An Introduction to Classical Econometric Theory*，Oxford University Press，New York，2000. 我们没有使用年龄数据，因为它与工作年数存在高度共线性。

组织的白人女性挣得少吗？换言之，定量回归元与虚拟变量之间有某种交互影响吗？

表 13—4　　基于方程（13.11.1）的 EViews 回归结果

Dependent Variable: LW
Method: Least Squares
Sample: 1-1,289
Included observations: 1,289

	Coefficient	Std. Error	t Statistic	Prob.
C	1.037880	0.074370	13.95563	0.0000
EDU	0.084037	0.005110	16.44509	0.0000
EXP	0.011152	0.001163	9.591954	0.0000
FE	-0.234934	0.026071	-9.011170	0.0000
NW	-0.124447	0.036340	-3.424498	0.0006
UN	0.207508	0.036265	5.721963	0.0000
WK	0.228725	0.028939	7.903647	0.0000

R-squared	0.376053	Mean dependent var.	2.342416
Adjusted *R*-squared	0.373133	S.D. dependent var.	0.586356
S.E. of regression	0.464247	Akaike info criterion	1.308614
Sum squared resid.	276.3030	Schwarz criterion	1.336645
Log likelihood	-836.4018	Hannan-Quinn criter.	1.319136
F-statistic	128.7771	Durbin-Watson stat.	1.977004
Prob. (*F*-statistic)	0.000000		

统计软件已经能够例行回答这些问题。比如，EViews 就有这种功能。在估计了一个模型之后，如果你认为有些变量可以放入模型中但又不能确定它们的重要性，那么你可以进行遗漏变量检验（test of omitted variable）。

为了说明这一点，假设我们估计方程（13.11.1），又想知道是否应该在模型中引入 Fe 和 NW、Fe 和 UN 以及 Fe 和 WK 的乘积，以便考虑解释变量之间的相互影响。利用 EViews 6 例行程序，我们得到下面的回答，其中虚拟假设是新增的这三个变量对我们所估计的模型没有影响。

如你所料，我们可以利用 F 检验（第 8 章讨论过）来评价这些新增变量的额外或边际影响，并对虚拟假设进行检验。就我们的例子而言，结果如表 13—5 所示：

表 13—5　　使用交互项的部分 EViews 结果

Omitted Variables: FE*NW FE*UN FE*WK

F-statistic	0.805344	Prob. *F*(3,1279)	0.4909
Log likelihood ratio	2.432625	Prob. chi-square (3)	0.4876

既然 0.805 3 的 F 估计值不是统计显著的，而且 p 值也只有约 49%，所以我们不能拒绝如下虚拟假设：女性与非白人、女性与工会会员关系以及女性与按周取酬的交互项对表 13—4 中给出的估计模型没有影响。

读者也可自行尝试回归元的其他组合，并评价它们对原模型的贡献。

在进一步讨论之前，我们注意到模型（13.11.1）表明，工作年数对对数工资的影响是线性的，即保持其他变量不变，工作年数的逐年增加导致工资的相对增加（牢记回归子是对数形式）保持不变。这个假定在一定的工作年限之内可能是正确的，但基本的劳动经济学理论表明，随着工人的年龄越来越大，工资增长的比率是递减的。为了看出在我们的例子中是否如此，我们在原模型中增加工作年数的平方项，得到的结果如表 13—6 所示：

表 13—6　　使用工作经历的平方项得到的 EViews 结果

Dependent Variable: LW
Method: Least Squares
Sample: 1-1,289
Included observations: 1,289

	Coefficient	Std. Error	t Statistic	Prob.
C	0.912279	0.075151	12.13922	0.0000
EDU	0.079867	0.005051	15.81218	0.0000
EXP	0.036659	0.003800	9.647230	0.0000
FE	-0.228848	0.025606	-8.937218	0.0000
NW	-0.121805	0.035673	-3.414458	0.0007
UN	0.199957	0.035614	5.614579	0.0000
WK	0.222549	0.028420	7.830675	0.0000
EXP*EXP	-0.000611	8.68E-05	-7.037304	0.0000

R-squared	0.399277	Mean dependent var.	2.342416
Adjusted *R*-squared	0.395995	S.D. dependent var.	0.586356
S.E. of regression	0.455703	Akaike info criterion	1.272234
Sum squared resid.	266.0186	Schwarz criterion	1.304269
Log likelihood	-811.9549	Hannan-Quinn criter.	1.284259
F-statistic	121.6331	Durbin-Watson stat.	1.971753
Prob. (*F*-statistic)	0.000000		

工作经历的平方项不仅符号为负，而且是高度统计显著的。而且，它还与劳动市场上的表现相一致；随着时间的推移，工资增长的比率越来越慢（$\partial \mathrm{LW}/\partial \mathrm{EXP}=0.036\ 6-0.001\ 2\mathrm{EXP}$）。

我们利用这个机会再来讨论一下赤池信息准则和施瓦茨信息准则。就像 R^2 一样，它们都是对估计模型拟合优度的检验；其差别在于，在 R^2 准则下，其值越高，模型对回归子表现的解释就越好。而在赤池信息准则和施瓦茨信息准则下，这些统计量的数值越低，模型就越好。

当然，如果我们想比较两个或多个模型的好坏，所有这些准则都是可以的。因此，如果你想比较表 13—4 中的模型和表 13—6 中增加了工作年数的平方这个回归元的模型，那么，基于这些准则，我们将看到表 13—6 中的模型比表 13—4 中的模型更好。

顺便指出，这两个模型中的 R^2 值看上去都很“低”，但在具有大量观测的横截面数据中经常看到这么低的 R^2。不过也要注意到，这个“很低”的 R^2 值是统计显

著的，因为在这两个模型中计算出来的 F 统计量都是高度显著的（回顾第 8 章讨论的 F 和 R^2 之间的关系）。

让我们继续表 13—6 中给出的扩展模型。尽管这个模型看上去令人满意，但还是让我们继续探讨一些重要问题。首先，既然我们在处理横截面数据，那就很可能会遇到异方差问题。所以，我们必须弄清楚是否如此。我们使用第 11 章曾讨论过的几个异方差检验，发现这个模型确实存在着异方差性。读者应该能够自行验证这一点。

为了修正所观测到的异方差性，我们可以求第 11 章曾讨论过的怀特的异方差一致标准误。结论在表 13—7 中给出。

表 13—7　　使用怀特的修正标准误而得到的 EViews 结果

```
Dependent Variable: LW
Method: Least Squares
Sample: 1-1,289
Included observations: 1,289
White's Heteroscedasticity-Consistent Standard Errors
  and Covariance
```

	Coefficient	Std. Error	t Statistic	Prob.
C	0.912279	0.077524	11.76777	0.0000
EDU	0.079867	0.005640	14.15988	0.0000
EXP	0.036659	0.003789	9.675724	0.0000
FE	-0.228848	0.025764	-8.882625	0.0000
NW	-0.121805	0.033698	-3.614573	0.0003
UN	0.199957	0.029985	6.668458	0.0000
WK	0.222549	0.031301	7.110051	0.0000
EXP*EXP	-0.000611	9.44E-05	-6.470218	0.0000

R-squared	0.399277	Mean dependent var.	2.342416
Adjusted R-squared	0.395995	S.D. dependent var.	0.586356
S.E. of regression	0.455703	Akaike info criterion	1.272234
Sum squared resid.	266.0186	Schwarz criterion	1.304269
Log likelihood	-811.9549	Hannan-Quinn criter.	1.284259
F-statistic	121.6331	Durbin-Watson stat.	1.971753
Prob. (F-statistic)	0.000000		

如同所料，估计标准误还是有所变化，当然在解释相对工资的行为时，这些变化不足以改变从单个和总体来看所有回归元都很重要的结论。

现在我们再来看误差项是否是正态分布。图 13—5 给出了表 13—7 中模型残差的直方图。雅克-贝拉统计量拒绝了误差正态分布的假设，因为雅克-贝拉统计量很高，而且 p 值几乎为零。注意，对于一个正态分布变量而言，偏度系数和峰度系数分别是 0 和 3。

现在怎么办呢？迄今为止，我们的假设检验程序一直都以回归模型的干扰项或误差项服从正态分布的假定为前提。这是否意味着我们不能在工资回归中合理地使用 t 检验和 F 检验来检验各种假设呢？

回答是**否定**的。就像在前文中提到的那样，OLS 估计量是渐近正态分布的，而且

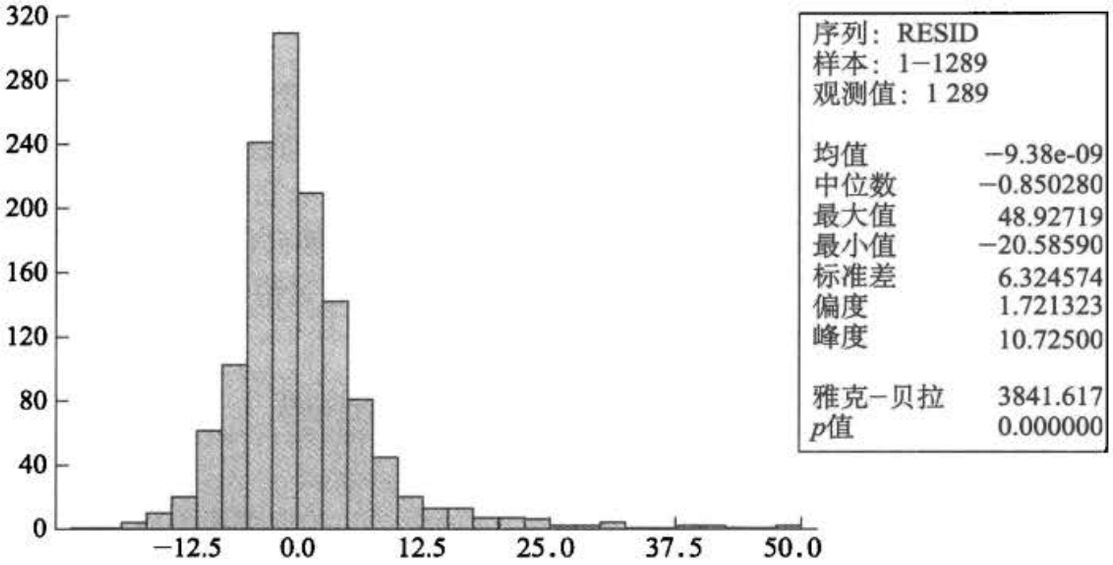


图 13—5　表 13—7 中回归模型残差的直方图

我们还强调指出，误差项具有有限方差并且是同方差的，而且在给定解释变量值的情况下，误差项的均值为 0。因此，在样本足够大的情况下，我们可以继续使用通常的 t 检验和 F 检验。顺便还要指出，我们在得到 OLS 估计量时不需要正态性假定。即使没有正态性假定，在高斯-马尔可夫假定下，OLS 估计量仍是最优线性无偏的（BLUE）。

样本到底多大才算大样本呢？对这个问题没有一个确定的答案，但在我们的工资回归中由1 289个观测构成的样本看上去足够大了。

在我们的工资回归中有“异常观测”吗？图 13—6 给出了因变量（工资的对数）的估计值和残差（即回归子的实际值与估计值之差），此图能够帮助我们澄清有无异常观测的某些看法。

尽管残差的均值恒为 0（为什么？），但图 13—6 表明有些残差与多数残差相比悬

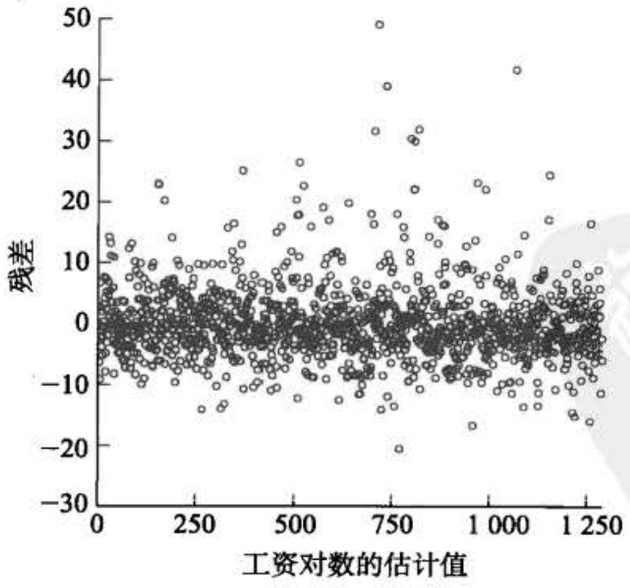


图 13—6　残差与因变量（工资的对数）的估计值

殊（在绝对值上）。可能数据中存在一些异常观测。我们在表 13—8 中给出三个定量变量的原始统计量来帮助读者判断是否确实存在异常观测。

表 13—8

Sample: 1-1,289

	W	EDU	EXP
Mean	12.36585	13.14507	18.78976
Median	10.08000	12.00000	18.00000
Maximum	64.08000	20.00000	56.00000
Minimum	0.840000	0.000000	0.000000
Std. Dev.	7.896350	2.813823	11.66284
Skewness	1.848114	-0.290381	0.375669
Kurtosis	7.836565	5.977464	2.327946
Jarque-Bera	1990.134	494.2552	54.57664
Probability	0.000000	0.000000	0.000000
Sum	15939.58	16944.00	24220.00
Sum Sq. Dev.	80309.82	10197.87	175196.0
Observations	1,289	1,289	1,289

□ 2. 1947—2000 年间美国的真实消费函数

我们在第 10 章曾考虑了美国 1947—2000 年间的消费函数。我们在那里考虑的消费函数的具体形式为：

$$\ln TC_t = \beta_1 + \beta_2 \ln YD_t + \beta_3 \ln W_t + \beta_4 \text{Interest}_t + u_t \qquad (13.11.3)$$

其中 TC、YD、W 和 Interest 分别是真实总消费支出、真实个人可支配收入、真实财富和真实利率。基于我们的数据而得到的结果如表 13—9 所示：

表 13—9　　方程（13.11.3）的回归结果

Method: Least Squares
Sample: 1947-2000
Included observations: 54

	Coefficient	Std. Error	t Statistic	Prob.
C	-0.467711	0.042778	-10.93343	0.0000
LOG(YD)	0.804873	0.017498	45.99836	0.0000
LOG(WEALTH)	0.201270	0.017593	11.44060	0.0000
INTEREST	-0.002689	0.000762	-3.529265	0.0009

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted *R*-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	Akaike info criterion	-5.947703
Sum squared resid.	0.007121	Schwarz criterion	-5.800371
Log likelihood	164.5880	Hannan-Quinn criter.	-5.890883
F-statistic	37832.59	Durbin-Watson stat.	1.289219
Prob. (*F*-statistic)	0.000000		

由于 TC、YD 和 W 都采用对数形式，所以 YD 和 W 的斜率系数估计值分别是收入和财富弹性。如你所料，这些弹性为正且高度统计显著。从数值上看，收入和财富弹性约为 0.80 和 0.20。利率变量的系数表示半弹性。（为什么?）保持其他变量不变，这些结果表明，如果利率提高 1 个百分点，那么平均而言，真实消费支出将下降约 0.27%。注意，估计出来的这个半弹性也是高度统计显著的。

再来看一下总结性的统计量。R^2 值很高，几乎达到 100%。F 值也高度统计显著，这就表明，所有解释变量不仅单个来看而且总体来看对消费支出都有明显的影响。

不过，德宾-沃森统计量表明模型误差是序列相关的。如果我们查阅德宾-沃森统计表（附录 D 中的表 D—5），对于 55 个观测（最接近的数字是 54）和 3 个解释变量的情况，5%的 d 临界值的下限和上限分别是 1.452 和 1.681。由于我们在本例中观察到的 d 值 1.289 2 低于下临界值，所以我们可以断定我们的消费函数中的误差是正相关的。既然大多数时间序列回归都会遇到自相关的问题，所以这也无足为奇。

但在我们接受这个结论之前，还是让我们先弄清楚是否存在设定误差。因为我们知道，有时候自相关的出现是因为我们遗漏了一些重要变量。为了看出是否属于这种情况，我们考虑表 13—10 中的回归。

表 13—10

Dependent Variable: LTC
Method: Least Squares
Sample: 1947-2000
Included observations: 54

	Coefficient	Std. Error	*t* Statistic	Prob.
C	2.689644	0.566034	4.751737	0.0000
LYD	0.512836	0.054056	9.487076	0.0000
LW	-0.205281	0.074068	-2.771510	0.0079
INTEREST	-0.001162	0.000661	-1.759143	0.0848
LYD*LW	0.039901	0.007141	5.587986	0.0000

R-squared	0.999731	Mean dependent var.	7.826093
Adjusted *R*-squared	0.999709	S.D. dependent var.	0.552368
S.E. of regression	0.009421	Akaike info criterion	-6.403689
Sum squared resid.	0.004349	Schwarz criterion	-6.219524
Log likelihood	177.8996	Hannan-Quinn criter.	-6.332663
F-statistic	45534.94	Durbin-Watson Stat.	1.530268
Prob. (*F*-statistic)	0.000000		

这个模型中新增的变量是可支配收入的对数与财富的对数的乘积项。这个乘积项也是高度显著的。注意，现在利率变量尽管保持了其符号，但变得不是那么显著了（p 值约为 8%）。但现在的德宾-沃森 d 值却从 1.28 提高到 1.53。

现在，显著性水平为 5%的 d 临界值是 1.378 和 1.721。我们所观察到的 d 值 1.53 正好介于其间，从而表明基于德宾-沃森统计量，我们不能确定是否存在自相

关。不过，所观察到的 d 值更接近于 d 值的上限。正如前面讨论自相关的章节中所指出的那样，有些作者建议利用这个 d 统计量的上限近似作为真正的显著性界限，因此，如果计算得到的 d 值低于这个上限，还算是存在正相关的证据。根据这一准则，在当前的情况下，我们断定我们的模型还是存在正自相关的问题。

我们还可以使用第 12 章讨论过的布罗施-戈弗雷自相关检验。在表 13—9 的模型中增加方程（12.6.15）中残差估计值的二阶滞后项，我们得到如下结果（见表 13—11）：

表 13—11

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	3.254131	Prob. *F*(2,48)	0.0473
Obs*R-squared	6.447576	Prob. chi-square (2)	0.0398

Dependent Variable: RESID
Method: Least Squares
Sample: 1947-2000
Included observations: 54
Presample missing value lagged residuals set to zero.

	Coefficient	Std. Error	t Statistic	Prob.
C	-0.006514	0.041528	-0.156851	0.8760
LYD	-0.004197	0.017158	-0.244619	0.8078
LW	0.004191	0.017271	0.242674	0.8093
INTEREST	0.000116	0.000736	0.156964	0.8759
RESID(-1)	0.385190	0.151581	2.541147	0.0143
RESID(-2)	-0.165609	0.154695	-1.070556	0.2897

R-squared	0.119400	Mean dependent var.	-9.02E-17
Adjusted *R*-squared	0.027670	S.D. dependent var.	0.011591
S.E. of regression	0.011430	Akaike info criterion	-6.000781
Sum squared resid.	0.006271	Schwarz criterion	-5.779782
Log likelihood	168.0211	Hannan-Quinn criter.	-5.915550
F-statistic	1.301653	Durbin-Watson Stat.	1.848014
Prob. (*F*-statistic)	0.279040		

表 13—11 顶端报告的 F 检验检验了模型中所包含的两个滞后残差系数都为 0 的虚拟假设。由于这个 F 在 5%的显著性水平上是显著的，所以拒绝上述假设。

总而言之，看来误差项是自相关的。我们可以利用第 12 章讨论的一个或多个程序来消除自相关。但为了节省篇幅，把这个任务留给读者来完成。

我们在表 13—12 中报告了给出考虑自相关的 HAC 或尼威-威斯特标准误的回归分析结果。我们 54 个观测的样本容量足以使用 HAC 标准误。

如果你把这些结果与表 13—9 中的结果进行比较，你会发现，回归系数仍然相同，只是标准误有些不同。

表 13—12

Dependent Variable: LTC
Method: Least Squares
Sample: 1947-2000
Included observations: 54
Newey-West HAC Standard Errors and Covariance (lag truncation = 3)

	Coefficient	Std. Error	t Statistic	Prob.
C	-0.467714	0.043937	-10.64516	0.0000
LYD	0.804871	0.017117	47.02132	0.0000
LW	0.201272	0.015447	13.02988	0.0000
INTEREST	-0.002689	0.000880	-3.056306	0.0036

R-squared	0.999560	Mean dependent var.	7.826093
Adjusted *R*-squared	0.999533	S.D. dependent var.	0.552368
S.E. of regression	0.011934	Akaike info criterion	-5.947707
Sum squared resid.	0.007121	Schwarz criterion	-5.800374
Log likelihood.	164.5881	Hannan-Quinn criter.	-5.890886
F-statistic	37832.71	Durbin-Watson Stat.	1.289237
Prob. (*F*-statistic)	0.000000		

我们在本章还讨论了邹至庄预测失灵检验。我们的样本期间为 1947—2000 年。在此期间，包含了几个经济周期，多数周期持续时间很短。比如，1990 年和 2000 年都曾发生过经济衰退。在经济衰退期，消费支出与收入、财富和利率的关系是否有所不同？

为了考虑这个问题，让我们考虑 1990 年的经济衰退，并使用邹至庄的预测失灵检验。前面已经详尽地讨论了这一检验。在 EViews 6 版本中利用邹至庄的预测失灵检验，我们得到表 13—13 中给出的结果。

表 13—13 顶部给出的 F 统计量表明，1990 年之前和之后的消费函数可能存在明显差异，因为它的 p 值在 5%的显著性水平上是不显著的。但如果你选择 10%的显著性水平，那这个 F 值又是统计显著的。

我们可以换个角度来看这个问题。我们在第 8 章讨论了参数稳定性检验。为了看出消费函数的回归系数是否存在统计显著的变化，我们使用第 8 章 8.7 节讨论过的邹至庄检验，并得到表 13—14 给出的结果。

很明显，1990 年之前和之后的消费函数在统计上是不同的，因为根据方程 (8.7.4) 计算出来的 F 统计量是高度显著的，其 p 值只有 0.005 2。

读者可以利用邹至庄的参数稳定性检验和预测失灵检验来判断消费函数在 2000 年之前和之后是否有所变化。为此，你必须把数据扩充到 2000 年之后。还要注意，为了使用这些检验，观测次数必须大于待估计的系数个数。

我们已经尝试了我们的消费数据能够使用的所有诊断检验。到目前为止的分析应该为你如何使用各种检验提供了很好的直觉。

表 13—13　　邹至庄预测失灵检验

Chow's Forecast Test: Forecast from 1991 to 2000

F-statistic	1.957745	Prob. *F* (10,40)	0.0652
Log likelihood ratio	21.51348	Prob. chi-square (10)	0.0178

Dependent Variable: LTC
Method: Least Squares
Sample: 1947-1990
Included observations: 44

	Coefficient	Std. Error	*t* Statistic	Prob.
C	-0.287952	0.095089	-3.028236	0.0043
LYD	0.853172	0.028473	29.96474	0.0000
LW	0.141513	0.033085	4.277239	0.0001
INTEREST	-0.002060	0.000804	-2.562790	0.0143

R-squared	0.999496	Mean dependent var.	7.659729
Adjusted *R*-squared	0.999458	S.D. dependent var.	0.469580
S.E. of regression	0.010933	Akaike info criterion	-6.107640
Sum squared resid.	0.004781	Schwarz criterion	-5.945441
Log likelihood	138.3681	Hannan-Quinn criter.	-6.047489
F-statistic	26430.49	Durbin-Watson Stat.	1.262748
Prob. (*F*-statistic)	0.000000		

表 13—14　　参数稳定性的邹至庄检验

Chow Breakpoint Test: 1990
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1947-2000

F-statistic	4.254054	Prob. *F*(4,46)	0.0052
Log likelihood ratio	16.99654	Prob. chi-square (4)	0.0019
Wald statistic	17.01622	Prob. chi-square (4)	0.0019

13.12　非正态误差与随机回归元

我们在本节将要讨论误差项的非正态分布和随机回归元这两个专题。它们在实践中极其重要，本质上，这些内容有些艰深。

□ 1. 如果误差项不是正态分布的，结果会怎么样?

在第 4 章讨论的经典正态线性回归模型（CNLRM）中，我们假定误差项 u 服从正态分布。我们援引**中心极限定理**（central limit theorem，CLT）来说明正态假定

的合理性。因为有了这个假定，我们就能够证明 OLS 估计量也是正态分布的。因此，无论样本容量有多大，我们都能利用 t 检验和 F 检验来进行假设检验。我们还讨论了利用**雅克-贝拉正态性检验**和**安德森-达琳正态性检验**（Anderson-Darling normality test），来判断一个具体应用研究中估计出来的误差项到底是不是正态分布的。

如果误差不是正态分布的，结果又会怎么样呢？可以证明，OLS 估计量仍是最优线性无偏估计量，也就是说，它们是无偏的，而且在所有线性估计量中具有最小的方差。从直觉上讲，这也无足为奇，因为我们在证明高斯-马尔可夫定理时根本就不需要正态性假定。

那么，问题在哪儿呢？

问题是，我们需要 OLS 估计量的**抽样**（sampling）或**概率分布**（probability distributions）。没有这种分布，我们就无法进行与这些估计量的真值有关的任何假设检验。就像第 3 章和第 7 章提到的那样，OLS 估计量是因变量 Y 的线性函数，而如果解释变量在重复抽样时是非随机的或固定的，那么 Y 本身又是随机误差项 u 的线性函数。因此，我们最终还是需要 u 的概率分布。

正如前面指出的那样，经典正态线性回归模型假定误差项服从正态分布（即均值为 0 且方差保持不变）。利用**中心极限定理**说明了误差项正态分布的合理性之后，我们就能够证明 OLS 估计量本身也是正态分布的，其均值和方差我们在第 4 章和第 7 章曾给予充分讨论。这转而又使得我们能够在小样本或有限样本的情况下，就像在大样本的情况下一样，利用 t 统计量和 F 统计量进行假设检验。因此，正态性假定的作用，特别是在小样本的情况下，是非常重要的。

但如果基于各种正态性检验我们不能保证正态性假定，又会怎么样呢？我们又该怎么办呢？我们有两个选择。第一个办法是**自助法**（bootstrapping），第二个办法是援引**大样本**（large）或**渐近样本理论**（asymptotic sample theory）。

对自助法的讨论会不知不觉地把我们带到应用计量经济学领域，而这将离题太远。自助法背后的基本思想是不断地搅拌（或翻腾）一个给定的样本，并得到我们感兴趣的（OLS 估计量）参数的抽样分布。具体做法最好留给读者查阅参考文献。[①] 顺便一提，自助法一词源于“通过自己的靴襻把自己拉起来”的俗话。

处理非正态误差项的另一种方法就是使用渐近或大样本理论。事实上，我们在第 3 章附录 3A.7 节中证明 OLS 估计量是**一致**估计量的时候，曾对此略提一二。就像在本书附录 A 中讨论的那样，如果一个估计量随着样本容量越来越大而逐渐接近该估计量的真值，那么，它就是一致估计量（见附录 A 中的图 A.11）。

但这对我们的假设检验有何帮助呢？我们仍能使用 t 检验和 F 检验吗？可以证明，在高斯-马尔可夫假定下，OLS 估计量服从**渐近正态分布**（asymptotically nor-

① 规范讨论，参见 Christopher Z. Mooney and Robert D. Duval, *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Sage University Press, California, 1993。更规范的讨论可参见 Russell Davidson and James G. MacKinnon, *Econometric Theory and Methods*, Oxford University Press, New York, 2004, pp. 159-166。

mally distributed)，其均值和方差就是在第 4 章和第 7 章讨论的均值和方差。[①] 因此，在正态假定条件下提出的 t 检验和 F 检验在大样本中是渐近成立的。随着样本容量的扩大，近似效果会相当不错。[②]

□ 2. 如果误差项不是正态分布的，结果会怎么样?

在第 3 章，我们在一些简化条件下介绍了经典线性（于参数）回归模型。其中假定之一就是解释变量或回归元要么固定不变，要么是非随机的，即使是随机的，也要求它们独立于误差项。我们把前面那种情形称为固定回归元情形（fixed regressor case），而把后面那种情形称为随机回归元情形（random regressor case）。

在固定回归元情形中，我们已经知道 OLS 估计量的性质（见第 5 章和第 8 章）。在随机回归元情形中，如果我们继续假定我们的分析是以回归元的给定值为条件，那么，我们在固定回归元情形中研究的 OLS 估计量的性质都继续成立。

如果我们在随机回归元情形中假定这些回归元和误差项是独立分布的，那么，OLS 估计量仍是无偏的，但它们不再是有效的。[③]

如果误差项不是正态分布的，或者回归元不是随机的，或者二者兼而有之，问题就变得复杂了。这里，很难得到 OLS 估计量的有限样本性质的一般结论。不过，在一定的条件下，我们可以援引中心极限定理来证明 OLS 估计量的渐近正态性质。尽管超出了本书的范围，但其证明在其他地方还是可以找到。[④]

13.13 向实际工作者进一言

我们在本章已经取得了很大进展。毫无疑问，建模既是科学又是艺术。实践中的研究者可能被这些理论上的精妙之处和一系列的诊断检验搞得晕头转向。但最好记住马丁·费尔德斯坦（Martin Feldstein）的忠告："应用计量经济学家与理论家一样，很快就会从实践中发现，一个有用的模型不是真实的或现实的模型，而是节俭的、比较合理的和有信息含量的模型。"[⑤]

---

① 高斯-马尔可夫假定误差项的期望值为 0，误差项与每个解释变量都相互独立，误差方差是同方差的，而且误差项中不存在自相关。它还假定解释变量的方差—协方差矩阵是有限的。我们也可以放松误差项与回归元独立的条件，并假定它们不相关这个更弱的条件。

② OLS 估计量渐近正态分布的证明超出了本书的范围，参见 James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 2d ed., Pearson/Addison Wesley, Boston, 2007, pp. 710-711。

③ 技术性的细节，可参见 William H. Greene, *Econometric Analysis*, 6th ed., Pearson/Prentice-Hall, New Jersey, 2008, pp. 49-50。

④ 参见 Greene, op. cit.。

⑤ Martin S. Feldstein, "Inflation, Tax Rules and Investment: Some Econometric Evidence," *Econometrica*, vol. 30, 1982, p. 829.

加拿大西蒙·弗雷泽大学（Simon Fraser University）的彼得·肯尼迪（Peter Kennedy）给出如下“应用计量经济学的十大告诫”[①]：

1. 你应该使用常识和经济理论。

2. 你应该询问正确的问题（即实用性胜于数学上的优美）。

3. 你应该了解背景（不要做无知的统计分析）。

4. 你应该对数据进行审查。

5. 你不应该信奉复杂性，而应使用 **KISS 原则，即保持尽可能简单**（keep it stochastically simple）。

6. 你应该充分而又严格地看待结果。

7. 你应该当心数据挖掘的成本。

8. 你应该准备着妥协（不要信奉教科书中的方法）。

9. 你不应该把显著性和重要性混淆（不要混淆统计显著性和实际显著性）。

10. 当出现敏感性时你应该坦白（即准备接受批评）。

你可能想详细阅读肯尼迪的文章，以体会他所热衷的上述十大告诫的说服力。虽然有些告诫听起来不可当真，但每一条可能都包含着苦涩的真理。

要点与结论

1. CLRM 假定用于分析的计量经济模型是正确设定的，这个假定有两层意义。一是没有**方程设定误差**，二是没有**模型设定误差**，本章主要考虑方程设定误差。

2. 本章讨论的方程设定误差包括：（1）重要变量的遗漏，（2）多余变量的引入，（3）错误函数形式的采用，（4）误差项 u_i 的非正确设定，以及（5）回归子和回归元中的测量误差。

3. 当模型漏掉真实的变量时，后果可能是很严重的，模型中所保留变量的系数的 OLS 估计量不仅是有偏误的，而且不是一致性的。此外，这些系数的方差和标准误的估计都是不正确的，从而使通常的假设检验程序失效。

4. 模型含有无关变量的后果幸而不那么严重：有关和“无关”变量的系数的估计量仍然是无偏的，并且是一致性的。误差方差 σ^2 仍然被正确地估计。唯一的问题是所估计的方差倾向于过大，从而使参数的估计较不准确，即置信区间无必要地扩大。

5. 为了侦察方程设定误差，我们考虑了几种检验，诸如：（1）残差分析，（2）德宾-沃森 d 统计量，（3）拉姆齐 RESET 检验，以及（4）拉格朗日乘数检验。

6. 一类特殊的设定误差是回归子和回归元取值的测量误差。如果仅回归子中有测量误差，则 OLS 估计量是无偏且一致的，但效率较低。如果回归元中有测量误差，则 OLS 估计量是偏误且非一致的。

7. 即使察觉到或猜测到有测量误差，如何补救常常是不容易的。工具或代理变量的使用虽然理论上诱人，但不很实际。因此，在实践中研究者应详细说明他（或她）的数据来源，收集的方

① Peter Kennedy, op. cit., pp. 17-18.

法，所用的定义，等等。官方收集的数据常常附有多个注释，研究者应把这些告知读者。

8. 模型误设误差可能与方程设定误差一样严重。特别地，我们区分了嵌套和非嵌套模型。为了判别适当的模型，我们讨论了非嵌套或包容性的 F 检验和戴维森-麦金农 J 检验，并指出每个检验的局限性。

9. 在实践中选择一个经验模型时，研究者使用过一系列准则。我们讨论了其中的一些准则，比如赤池和施瓦茨信息准则、马娄斯 C_p 准则和预测 χ^2 准则。我们讨论了这些准则的优缺点，并警告读者这些准则不是绝对的，对仔细的设定分析只能起到辅助作用。

10. 我们还讨论了一些附加的专题：(1) 异常数据、杠杆数据和有影响力的数据；(2) 递归最小二乘法；(3) 邹至庄预测失灵检验。我们讨论了它们在应用研究中的作用。

11. 我们还简要地讨论了随机误差项的非正态性和随机回归元这两种特殊情形，以及在 OLS 估计量的小样本或有限样本性质无法证明的情况下渐近或大样本理论的作用。

12. 我们以彼得·肯尼迪对“应用计量经济学家的十大告诫”结束本章。这些告诫的关键所在是要求研究者考虑到计量经济学纯粹技术方面之外的东西。

习　　题

问答题

13.1　参照方程（8.6.23）中所估计的鸡肉需求函数，考虑到 13.1 节所讨论的好模型的属性，你会说这个需求函数是“正确”的设定吗？

13.2　假设真实模型是：

$$Y_i = \beta_1 X_i + u_i \tag{1}$$

但你没有去拟合这个过原点回归，却例行地拟合了通常带有截距的模型：

$$Y_i = \alpha_0 + \alpha_1 X_i + v_i \tag{2}$$

评述这一设定误差的后果。

13.3　继续做习题 13.2，但假定真实模型是 (2)，讨论拟合误设模型 (1) 的后果。

13.4　假设“真实”模型是：

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

而我们增加了一个“无关”变量 X_3 到模型中去（“无关”是指变量 X_3 的真实系数 β_3 为零），并估计了

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + v_i \tag{2}$$

a. 模型 (2) 的 R^2 和调整 R^2 会不会比模型 (1) 的大？

b. 从模型 (2) 得到的 β_1 和 β_2 的估计值是无偏的吗？

c. “无关”变量 X_3 的引入对 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差有影响吗？

13.5　考虑如下“真实”（柯布-道格拉斯）生产函数：

$$\ln Y_i = \alpha_0 + \alpha_1 \ln L_{1i} + \alpha_2 \ln L_{2i} + \alpha_3 \ln K_i + u_i$$

其中Y=产出；

L_1=生产性劳动；

L_2=非生产性劳动；

K=资本。

但若在经验研究中实际用的回归是：

$$\ln Y_i = \beta_0 + \beta_1 \ln L_{1i} + \beta_2 \ln K_i + u_i$$

假定你拥有有关变量的横截面数据，

a. 我们会得到 $E(\hat{\beta}_1) = \alpha_1$ 和 $E(\hat{\beta}_2) = \alpha_3$ 吗？

b. 如果知道 L_2 是生产函数中的一个无关变量，(a) 中的答案能否成立？给出必要的推导。

13.6 参照方程 (13.3.4) 和 (13.3.5)。你会看到 $\hat{\alpha}_2$ 虽然有偏误，却比无偏的 $\hat{\beta}_2$ 有更小的方差。你会怎样在偏误与较小方差之间做出权衡？提示：两种估计量的 MSE（均方误）可表达为：

$$\text{MSE}(\hat{\alpha}_2) = (\sigma^2 / \sum x_{2i}^2) + \beta_3^2 b_{32}^2$$

$$= \text{抽样方差} + \text{平方偏误}$$

$$\text{MSE}(\hat{\beta}_2) = \sigma^2 / \sum x_2^2 (1 - r_{23}^2)$$

关于 MSE，参看附录 A。

13.7 证明从方程 (13.5.1) 或 (13.5.3) 估计的 β 都是真实 β 的一个无偏估计。

13.8 按照弗里德曼的永久收入假说，我们可写出：

$$Y_i^* = \alpha + \beta X_i^* \tag{1}$$

其中 Y_i^* = “永久”消费支出，X_i^* = “永久”收入，但我们观测到的不是“永久”变量，而是：

$$Y_i = Y_i^* + u_i$$

$$X_i = X_i^* + v_i$$

其中 Y_i 和 X_i 是可以观测或测量到的数量，而 u_i 和 v_i 分别是 Y_i^* 和 X_i^* 中的测量误差。

利用可观测的数量，可把消费函数写为：

$$Y_i = \alpha + \beta(X_i - v_i) + u_i$$

$$= \alpha + \beta X_i + (u_i - \beta v_i) \tag{2}$$

假定：(1) $E(u_i) = E(v_i) = 0$，(2) $\text{var}(u_i) = \sigma_u^2$，$\text{var}(v_i) = \sigma_v^2$，(3) $\text{cov}(Y_i^*, u_i) = 0$，$\text{cov}(X_i^*, v_i) = 0$，以及 (4) $\text{cov}(u_i, X_i^*) = \text{cov}(v_i, Y_i^*) = \text{cov}(u_i, v_i) = 0$，证明在大样本中从模型 (2) 估计的 β 可表示为：

$$\text{plim}(\hat{\beta}) = \frac{\beta}{1 + (\sigma_v^2 / \sigma_{X^*}^2)}$$

a. 关于 $\hat{\beta}$ 的偏误性质，你能谈些什么？

b. 如果样本无限地加大，估计的 β 值会不会倾向于与真实 β 相等？

13.9 **资本资产定价模型。**近代投资理论中的资本资产定价模型（CAPM）设定，一定时期内证券（普通股）的平均回报率与证券的波动性即所谓 β 系数（波动性是对风险的度量）有如下关系：

$$\bar{R}_i = \alpha_1 + \alpha_2(\beta_i) + u_i \tag{1}$$

其中 $\bar{R}_i$ =证券 i 的平均回报率；

β_i =证券 i 的真实 β 系数；

u_i =随机干扰项。

真实 β_i 不可直接观测而是按下式估算：

$$r_{it} = \alpha_1 + \beta^* r_{m_t} + e_t \tag{2}$$

其中 r_{it} =时间 t 证券 i 的回报率；

r_{m_t} =时间 t 的市场回报率（指某个广泛的市场指数的回报率，如工业证券 S&P 指数的回报

率）；

e_t＝残差项。

并且其中 β^* 是“真实”β 系数的一个估计值。因此，我们实际上估计的不是模型（1）而是：

$$\bar{R}_i = \alpha_1 + \alpha_2(\beta_i^*) + u_i \tag{3}$$

其中 β_i^* 是从回归（2）得到的。但因 β_i^* 是估计值，真实 β 与 β^* 之间的关系可写为：

$$\beta_i^* = \beta_i + v_i \tag{4}$$

其中 v_i 可称为测量误差。

a. 这一测量误差对 α_2 的估计会有什么影响？

b. 从方程（3）估计的 α_2 会是真实 α_2 的一个无偏估计吗？如果不是，它是 α_2 的一致估计吗？如果不是，你建议使用什么样的补救措施？

13.10　考虑模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{1}$$

为了找出此模型是否因为漏掉变量 X_3 而成为一个误设的模型，你决定用模型（1）给出的残差仅仅对 X_3 一个变量做回归（注：在此回归中有一截距项）。然而，拉格朗日乘数（LM）检验要求你用方程（1）的残差兼对 X_2 和 X_3 及一常数项做回归。为什么你用的程序很可能是不适当的？[①]

13.11　考虑模型：

$$Y_i = \beta_1 + \beta_2 X_i^* + u_i$$

而实际上我们用以度量 X_i^* 的是这样的 X_i：

a. $X_i = X_i^* + 5$。

b. $X_i = 3X_i^*$。

c. $X_i = (X_i^* + \epsilon_i)$，其中 ϵ_i 是具有通常性质的一个纯随机项。

这些测量误差对真实 β_1 和 β_2 的估计将有什么影响？

13.12　参照回归方程（13.3.1）和（13.3.2），用类似于（13.3.3）的方法说明

$$E(\hat{\alpha}_1) = \beta_1 + \beta_3(\bar{X}_3 - b_{32}\bar{X}_2)$$

其中 b_{32} 是遗漏变量 X_3 对所含变量 X_2 回归中的斜率系数。

13.13　批判性地评述利莫尔（Leamer）的下述观点[②]：

> 我对超越统计学（metastatistics）［指实际上来自数据的推断理论］的兴趣源于我对工作中的经济学家的观察。在经济专业中认为计量经济理论无关紧要的人占有一个令人迷惑不解的大部分。计量经济理论与计量经济实践之间的鸿沟足以引起职业者的不安。事实上，一种平静的均衡渗透着我们的期刊和［专业］会议，我们心安理得地把我们自己划分为一方面是孤独虔诚的牧师团——统计理论家，另一方面是一群顽固的罪过者——数据分析者。牧师们被授权列出罪过的清单，并因他们所表现的特殊才能而受到敬畏。罪过者不被期望能避免罪过；但要求他们公开坦白他们的罪过。

13.14　评价亨利·瑟尔所做的如下陈述[③]：

> 就这门艺术的现状而论，最切合实际的做法是对置信系数和显著性作自由灵活的解释，

① Maddala, op. cit., p. 477.

② Edward E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, John Wiley & Sons, New York, 1978, p. vi.

③ Henry Theil, *Principles of Econometrics*, John Wiley & Sons, New York, 1971, p. 605-606.

如果置信区间和检验统计量是从“按习惯的回归策略得到的最后回归”计算出来的话。意思是说，一个95%的置信系数，也许实际上是一个80%的置信系数，而一个1%的显著性水平，也许实际上是一个10%的显著性水平。

13.15 关于20世纪50年代和60年代初所采取的计量经济学方法论，布劳格（M. Blaug）作过如下评述[①]：

……大多数［经验研究］像是放下球网打网球：现代经济学家过多地满足于证明真实世界符合他们的预测，而无意去拒绝本可检验的预测，从而用容易的验证去替代困难的反驳［按照Popper的说法］。

你同意这种观点吗？你不妨浏览布劳格的书，以更好地了解他的观点。

13.16 按照布劳格的意见，“没有证明的逻辑，但有证反的逻辑。”[②] 他这句话的意思是什么？

13.17 回到正文中讨论的圣路易斯模型，参照嵌套F检验所涉及的问题，严格评论回归(13.8.4)所展现的结果。

13.18 假设真实模型是

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$$

但你估计了

$$Y_i = \alpha_1 + \alpha_2 X_i + v_i$$

如果你利用Y在$X=-3$、-2、-1、0、1、2、3处的观测并估计了“不正确”的模型，这些估计值将出现什么偏误？[③]

13.19 为了看出X_i^2是否应属于模型$Y_i=\beta_1+\beta_2X_i+u_i$，拉姆齐RESET检验将估计这个线性模型，并从模型中得到Y_i的估计值［即$\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2X_i$］，然后估计模型$Y_i=\alpha_1+\alpha_2X_i+\alpha_3\hat{Y}_i^2+v_i$，并检验$\alpha_3$的显著性。试证明，若最终表明$\hat{\alpha}_3$在上述（RESET）方程中是统计显著的，则这就等同于直接估计如下模型：$Y_i=\beta_1+\beta_2X_i+\beta_3X_i^2+u_i$。（提示：在RESET回归中代入$\hat{Y}_i$。）[④]

13.20 如下命题正确与否？说明原因。[⑤]

a. 一个观测可能具有影响力但不是异常数据。

b. 一个观测可能是异常数据但不具有影响力。

c. 一个观测可能同时既是异常数据又具有影响力。

d. 如果在模型$Y_i=\beta_1+\beta_2X_i+\beta_3X_i^2+u_i$中，发现$\beta_3$的估计值是统计显著的，那么，即便$\beta_2$的估计值在统计上不显著，我们也应该在模型中保留线性项X_i。

e. 若你用OLS估计了模型$Y_i=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+u_i$或$Y_i=\alpha_1+\beta_2x_{2i}+\beta_3x_{3i}+u_i$，则估计的回归线应该相同，其中$x_{2i}=X_{2i}-\bar{X}_2$和$x_{3i}=X_{3i}-\bar{X}_3$。

实证分析题

13.21 利用习题7.19所给的对鸡肉的需求数据。假使你被告知真实的需求函数是：

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_6 \ln X_{6t} + u_t \qquad (1)$$

---

① M. Blaug, *The Methodology of Economics. Or How Economists Explain*, Cambridge University Press, New York, 1980, p. 256.

② Ibid., p. 14.

③ 节选自G. A. F., Sebeir, *Linear Regression Analysis*, John Wiley & Sons, New York, 1977, p. 176.

④ 节选自Kerry Peterson, op. cit., pp. 184-185.

⑤ 节选自Norman R. Draper and Harry Smith, op. cit., pp. 606-607.

而你有不同的看法并估计了以下的需求函数：

$$\ln Y_t = \alpha_1 + \alpha_2 \ln X_{2t} + \alpha_3 \ln X_{3t} + v_t \tag{2}$$

其中Y＝人均鸡肉消费量（磅）；

X_2＝实际人均可支配收入；

X_3＝实际鸡肉零售价格；

X_6＝鸡肉替代品的实际复合价格。

a. 假定需求函数（1）是真实的，做设定误差的 RESET 和 LM 检验。

b. 假使我们发现方程（1）中的 $\hat{\beta}_6$ 在统计上不显著。这是否意味着用方程（2）去拟合数据就没有设定误差？

c. 如果我们发现 $\hat{\beta}_6$ 不显著，这是否意味着我们不应把替代品的价格作为变量引入到需求函数中来？

13.22　继续习题 13.21。纯粹出于教学的目的，假定模型（2）是真实需求函数。

a. 如果现在我们估计了模型（1），这时我们犯了什么类型的设定错误？

b. 这种设定误差的理论性后果有哪些？用你掌握的数据作出说明。

13.23　假设真实模型为：

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i \tag{1}$$

而由于测量误差你估计了

$$Y_i = \alpha_1 + \alpha_2 X_i + v_i \tag{2}$$

其中 $Y_i = Y_i^* + \epsilon_i$，$X_i = X_i^* + w_i$，ϵ_i 和 w_i 则表示测量误差。

利用表 13—2 中给出的数据，罗列出我们估计方程（2）而没有估计真实模型（1）所导致的后果。

13.24　**蒙特卡罗实验。**[①] 10 个人的每周永久收入为：200、220、240、260、280、300、320、340、380 和 400（美元）。永久消费（Y_i^*）和永久收入（X_i^*）的关系为：

$$Y_i^* = 0.8X_i^* \tag{1}$$

每个人的临时收入等于均值为 0 方差为 1 的正态总体（即标准正态变量）的随机抽样值 u_i 的 100 倍。假定消费中没有临时成分。就是说，观测消费和永久消费相同。

a. 从零均值和单位方差的正态总体中抽取 10 个随机数并得到 10 个观测收入 X_i（$= X_i^* + 100u_i$）。

b. 利用（a）中得到的数据做永久（＝观测）消费对观测收入的回归，并将你的结果同方程（1）所展示的结果相比较。先验地，截距应为零。（为什么？）你的结果是不是这样？为什么？

c. 重复（a）100 次，从而得到 100 个（b）那样的回归。然后用你的结果同真实回归（1）相比。你能得出什么一般性结论。

13.25　参照习题 8.26，按照那里给的变量定义，考虑以下两个解释 Y 的模型：

模型 A：　$Y_t = \alpha_1 + \alpha_2 X_{3t} + \alpha_3 X_{4t} + \alpha_4 X_{6t} + u_t$

模型 B：　$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{5t} + \beta_4 X_{6t} + u_t$

用嵌套 F 检验，你将怎样在这两个模型之间进行选择？

13.26　继续习题 13.25，用 J 检验，你会怎样在这两个模型之间做出选择？

13.27　回到习题 7.19，那是关于美国对鸡肉的需求问题，在那里给出了 5 个模型。

---

① 节选自 Christopher Dougherty, *Introduction to Econometrics*, Oxford University Press, New York, 1992, pp. 253-256.

a. 模型 1 和模型 2 有什么差别？如果模型 2 是正确的，而你估计了模型 1，那么你所犯的是什么类型的错误？你会用哪一种检验——方程设定误差（检验）或模型选择误差（检验）？给出必要的计算。

b. 你会怎样在模型 1 和模型 5 之间做出选择？你会用哪些检验，为什么？

13.28 参照表 8—11，它给出 1970—2005 年间个人储蓄（Y）和个人可支配收入（X）数据。现考虑下述模型：

模型 A：$Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 X_{t-1} + u_t$

模型 B：$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$

你怎样在这两个模型之间做出选择？明确地陈述你用的检验程序及全部计算。假使有人争辩利率变量属于此储蓄函数，你会怎样对此做检验？收集 3 月期的国债利率作为利率的代理变量，并说明你的答案。

13.29 利用习题 13.28 中的数据。为了熟悉递归最小二乘法，对 1970—1981 年、1970—1985 年、1970—1990 年和 1970—1995 年期间估计储蓄函数。评论储蓄函数中估计系数的稳定性。

13.30 继续习题 13.29，但现在使用表 8—10 中更新后的数据。

a. 假设你针对 1970—1981 年估计了储蓄函数。利用如此估计得到的参数和 1982—2000 年个人可支配收入的数据，对后一期间估计预测储蓄，并利用邹至庄预测失灵检验，说明它是否拒绝储蓄函数在这两个期间没有发生变化的假设。

b. 现在估计 2000—2005 年数据的储蓄函数。将结果与用（a）中同样的方法对 1982—2000 年期间数据估计的函数进行比较（邹至庄预测失灵检验）。这两个期间的储蓄函数有明显不同吗？

13.31 **在 k 变量回归模型中遗漏一个变量。**参照方程（13.3.3），它给出了从模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 中漏掉变量 X_3 的偏误。对此可做如下推广：在一个 k 变量模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$ 中，假设我们遗漏了变量 X_k。则可以证明，所包含变量 X_j 的斜率系数的偏误为：

$$E(\hat{\beta}_j) = \beta_j + \beta_k b_{kj} \quad j = 2, 3, \cdots, k-1$$

其中 b_{kj} 为被排除变量 X_k 对模型中包含的所有其他解释变量的辅助回归中 X_j 的（偏）斜率系数。①

回到习题 13.21。当我们从方程（1）中排除掉变量 $\ln X_6$ 时，求出方程（1）中系数偏误的大小。这种排除严重吗？给出必要的计算。

附录 13A

□ 13A.1 $E(b_{12}) = \beta_2 + \beta_3 b_{32}$ [方程（13.3.3）] 的证明

三变量总体回归模型的离差形式可以写成

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u}) \tag{1}$$

首先将两边同时乘以 x_2 并求和，然后将两边同时乘以 x_3 并求和，通常的正规方程就是

$$\sum y_i x_{2i} = \beta_2 \sum x_{2i}^2 + \beta_3 \sum x_{2i} x_{3i} + \sum x_{2i}(u_i - \bar{u}) \tag{2}$$

① 还可以推广到从模型中排除不止一个有关 X 变量的情形。对此，可参见 Chandan Mukherjee et al., op. cit., p. 215。

$$\sum y_i x_{3i} = \beta_2 \sum x_{2i}x_{3i} + \beta_3 \sum x_{3i}^2 + \sum x_{3i}(u_i - \bar{u}) \tag{3}$$

将方程（2）两边同时除以 $\sum x_{2i}^2$，我们得到

$$\frac{\sum y_i x_{2i}}{\sum x_{2i}^2} = \beta_2 + \beta_3 \frac{\sum x_{2i}x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \tag{4}$$

现在记得

$$b_{12} = \frac{\sum y_i x_{2i}}{\sum x_{2i}^2}$$

$$b_{32} = \frac{\sum x_{2i}x_{3i}}{\sum x_{2i}^2}$$

方程（4）便可写成

$$b_{12} = \beta_2 + \beta_3 b_{32} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \tag{5}$$

对方程（5）两边同时取期望，我们最后得到

$$E(b_{12}) = \beta_2 + \beta_3 b_{32}$$

其中用到（a）对于给定样本，b_{32}为已知量；（b）β_2 和 β_3 都是常数；以及（c）u_i 与 X_{2i}（X_{3i}也一样）不相关。

□ 13A.2 含有无关变量的后果：无偏性质

对真实模型（13.3.6）我们有：

$$\hat{\beta}_2 = \frac{\sum yx_2}{\sum x_2^2} \tag{1}$$

并且我们知道它是无偏的。

对模型（13.3.7）我们得到：

$$\hat{\alpha}_2 = \frac{(\sum yx_2)(\sum x_3^2) - (\sum yx_3)(\sum x_2x_3)}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \tag{2}$$

现在真实模型的离差形式是：

$$y_i = \beta_2 x_2 + (u_i - \bar{u}) \tag{3}$$

用方程（3）中的 y_i 代入方程（2）并化简，得：

$$E(\hat{\alpha}_2) = \beta_2 \frac{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} = \beta_2 \tag{4}$$

即 $\hat{\alpha}_2$ 仍是无偏的。

我们还得到：

$$\hat{\alpha}_3 = \frac{(\sum yx_3)(\sum x_2^2) - (\sum yx_2)(\sum x_2x_3)}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} \tag{5}$$

用方程（3）中的 y_i 代入方程（5）并化简，得：

$$E(\hat{\alpha}_3) = \beta_2 \frac{(\sum x_2x_3)(\sum x_2^2) - (\sum x_2x_3)(\sum x_2^2)}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2} = 0 \tag{6}$$

由于真实模型中没有 X_3，这便是它在真实模型中的值。

□ 13A.3 方程（13.5.10）的证明

我们有：

$$Y_i = \alpha + \beta X_i^* + u_i \tag{1}$$

$$X_i = X_i^* + w_i \tag{2}$$

因此，我们有离差形式：

$$y_i = \beta x_i^* + (u_i - \bar{u}) \tag{3}$$

$$x_i = x_i^* + (w_i - \bar{w}) \tag{4}$$

现在，如果我们使用：

$$Y_i = \alpha + \beta X_i + u_i \tag{5}$$

就得到：

$$\hat{\beta} = \frac{\sum yx}{\sum x^2}$$

$$= \frac{\sum [\beta x^* + (u - \bar{u})][x^* + (w - \bar{w})]}{\sum [x^* + (w - \bar{w})]^2} \qquad \text{利用方程(3)和(4)}$$

$$= \frac{\beta \sum x^{*2} + \beta \sum x^* (w - \bar{w}) + \sum x^* (u - \bar{u}) + \sum (u - \bar{u})(w - \bar{w})}{\sum x^{*2} + 2\sum x^* (w - \bar{w}) + \sum (w - \bar{w})^2}$$

由于两个变量之比的期望值不等于它们的期望值之比，我们不能取这个表达式的期望值（注：期望值运算子 E 是一个线性运算子）。因此，我们先用 n 除分子和分母的每一项，然后取概率极限，即取下式的 plim（关于 plim 的详细内容，见附录 A）：

$$\hat{\beta} = \frac{(1/n)[\beta \sum x^{*2} + \beta \sum x^* (w - \bar{w}) + \sum x^* (u - \bar{u}) + \sum (u - \bar{u})(w - \bar{w})]}{(1/n)[\sum x^{*2} + 2\sum x^* (w - \bar{w}) + \sum (w - \bar{w})^2]}$$

因两个变量之比的概率极限就是它们的概率极限之比，故对每项取概率极限得：

$$\text{plim}\hat{\beta} = \frac{\beta \sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_w^2}$$

其中 $\sigma_{X^*}^2$ 和 σ_w^2 是随样本无限增大时 X^* 和 w 的方差。这里我们利用了这样一个事实，即随着样本的无限增大，误差 u 和 w 之间以及它们和真实 X^* 之间都不存在相关关系。由以上表达式，我们最后得到：

$$\text{plim}\hat{\beta} = \beta \left[\frac{1}{1 + (\sigma_w^2 / \sigma_{X^*}^2)} \right]$$

这正是我们要证明的结论。

□ 13A.4 方程（13.6.2）的证明

由于此模型中没有截距项，所以根据过原点回归的公式得到 α 的估计值如下：

$$\hat{\alpha} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{1}$$

代入真实模型（13.2.8）中的 Y，我们得到

$$\hat{\alpha} = \frac{\sum X_i(\beta X_i u_i)}{\sum X_i^2} = \beta \frac{\sum X_i^2 u_i}{\sum X_i^2} \tag{2}$$

统计理论表明，若 $\ln u_i \sim N(0, \sigma^2)$，则

$$u_i = \text{log normal}[e^{\sigma^2/2}, e^{\sigma^2}(e^{\sigma^2-1})] \tag{3}$$

因此

$$\begin{aligned} E(\hat{\alpha}) &= \beta E\left(\frac{\sum X_i^2 u_i}{\sum X_i^2}\right) \\ &= \beta\left(E\frac{X_1^2 u_1 + X_2^2 u_2 + \cdots + X_n^2 u_n}{\sum X_i^2}\right) \\ &= \beta e^{\sigma^2/2}\left(\frac{\sum X_i^2}{\sum X_i^2}\right) = \beta e^{\sigma^2/2} \end{aligned}$$

其中使用到 X 都是非随机的以及每个 u_i 的期望值都是 $e^{\sigma^2/2}$ 的事实。

由于 $E(\hat{\alpha}) \neq \beta$，所以 $\hat{\alpha}$ 是 β 的一个有偏误的估计量。

Mc Graw Hill Education

“十一五”国家重点图书出版规划项目

·经/济/科/学/译/丛·

Basic Econometrics

(Fifth Edition)

计量经济学基础 下册

（第五版）

达摩达尔·N·古扎拉蒂（Damodar N. Gujarati）
唐·C·波特（Dawn C. Porter） 著

中国人民大学出版社