

前 言

本书是为高等学校经济管理类各专业本科的计量经济学课程编写的教材。中国高等学校开设计量经济学课程已有 20 多年的历史，起初只是部分学校的少数专业开设，1998 年经教育部全国经济学教学指导委员会讨论决定，把计量经济学纳入了经济学类所有专业必修的核心课程，此后计量经济学更加受到经济学类各专业的普遍重视，全国各高校不仅经济学类专业已普遍开设了计量经济学，而且一些管理类专业也十分重视这门课程的学习。经过 20 多年的努力，中国高等学校的计量经济学教学已经有了长足的进步。目前，不仅引进或翻译了许多国外的计量经济学教材，而且国内也编写了不少教材。与 20 年前刚开设计量经济学课程时教材奇缺的状况相比，已经有了很大改善。但是，从中国高等学校经济管理类各专业学生的实际出发，作为各专业的共同基础课，应该怎样合理地组织教学内容，怎样用有限的课时使学生既掌握计量经济学的基本理论和方法，又具备运用计量经济学知识分析实际经济问题的能力，还需要认真地加以研究。现在编写计量经济学教材，已不是解决教材的有无问题，而是要在总结多年教学经验的基础上，努力提高教材的质量，编写出最适合于经济管理类专业本科教学使用的教材。

本书充分借鉴了国内外教材的优点，总结了作者 20 多年来在财经院校从事计量经济学教学的经验和体会，是在对过去多次编写的教材反复思考、多方提炼的基础上，重新编写而成的。目标是力图做到“教师最好用，学生最好读”。与其他教材相比，本书有一些明显的特点：

1、从经济管理类各专业的实际出发，精选了教学内容。本科阶段的计量经济学课程的目标，应当定位在使学生掌握计量经济研究的最基本方法，并能够运用这些方法解决实际的经济问题。大学本科的计量经济学课程一般都不到 60 学时，是计量经济学的入门课程，只能以经典计量经济学的内容为主，适当概要性地介绍一些新发展的方向。经典计量经济学应用最为普遍，也是更高层次计量经济学课程的重要基础，符合财经院校绝大多数本科专业教学的实际要求，更多的非经典计量经济学的内容应该放到更高层次的教材中去。本书中未用“*”号注明的部分，是本科计量经济学教学的最基本要求。考虑到全国各学校、各专业的教学要求有一定差异，本书也安排了部分选讲内容，以“*”号和脚注标出供教学中选择，但跳过这些内容并不影响对计量经济学基础内容的系统学习。

2、坚持“重思想、重方法、重应用”的原则，特别注重基本思想、经济背景、基本方

法和实际应用。计量经济学是一门经济学课程，并不是数学课。多年来学生反映计量经济学课程较难，教材看不懂，其原因是教学内容和教材的写法过于数学化。本书尽可能地避免了繁琐的数学推导，少数必要的数学推导和证明也是放到附录中供选择阅读，使之更加适应更多经济管理专业学生的要求。

3、为教学创造良好的条件和环境，根据我们的教学体会，在每一章的开始都设置了从实际经济背景出发提出的“引子”，目的是从实际应用的角度提出本章将要讨论的主要问题，而不是从概念到概念，不是抽象地讲理论和方法。每一章的最后一节都是“案例分析”，通过实际经济问题的案例说明本章讨论的主要方法如何通过 EViews 计算机软件去实际运用。计量经济学中概念和公式较多，为有利于教师对本章的学习内容作总结和学生复习，每一章的最后除了思考题和练习题以外，还提供了本章主要内容的小结，并以表格形式列出了本章的主要公式。

4、本书与普遍应用的 EViews 计算机软件紧密结合，书中所讲的所有方法都要求在 EViews 软件上实现。改变了过去单独介绍软件的做法，而是将 EViews 软件的学习与各章案例分析有机结合，使学生在实际运用中去学习 EViews 的操作方法。

5、许多学生反映学习了计量经济学后不知该怎么运用，对计算的结果难以作出合理的解释。为了培养学生应用计量经济学方法独立解决实际经济问题的能力和素质，本书改变了其他教材介绍若干宏观经济应用模型的作法。而是在最后一章专门讨论应用计量经济学方法作实际项目研究的一般方式，指导学生通过完成“课程论文”，去自己体验计量经济学方法的实际应用，并提高计量经济分析的实际应用能力。从 2000 年起我们就在计量经济学本科的教学中全面采用了这种教学方式，取得了较好的效果。在计量经济学教材中作这样的改革，是在总结教学实践经验基础上的一种探索。

本书的第一、二、三章由西南财经大学教授庞皓编写、第四章由中南财经大学教授徐映梅博士编写、第五、十一章由西南财经大学教授李南成博士编写、第六章由中南财经大学副教授李占风编写、第八、九、十二章由西南财经大学教授黎实博士编写、第七、十章由西南财经大学教授史代敏博士编写，最后庞皓教授对全书作了修改。本书的主审、山西财经大学杭斌教授认真地审阅了全书，并提出了许多很好的修改意见。

本书适合于作为财经院校经济管理类本科各专业“计量经济学”课程的教材，如果适当考虑供教学中选择的内容，也可作为非数量经济、非统计专业研究生的辅助教材。同时，配合本书所附光盘的教学资源，本书还特别适合自学计量经济学的读者阅读。

由于作者水平有限，书中定有错漏之处，恳请广大读者批评指正。

目 录

第一章 导论

第一节 什么是计量经济学

- 一、计量经济学的产生与发展
- 二、计量经济学的性质
- 三、计量经济学与其他学科的关系

第二节 计量经济学的研究步骤

- 一、模型设定
- 二、估计参数
- 三、模型检验
- 四、模型应用

第三节 变量、参数、数据与模型

- 一、计量经济模型中的变量
- 二、参数估计的方法
- 三、计量经济学中应用的数据
- 四、计量经济模型的建立

第一章小结

思考题

第二章 简单线性回归模型

第一节 回归分析与回归函数

- 一、相关分析与回归分析
- 二、总体回归函数 (PRF)
- 三、随机扰动项 u
- 四、样本回归函数 (SRF)

第二节 简单线性回归模型参数的估计

- 一、简单线性回归的基本假定
- 二、普通最小二乘法
- 三、OLS 回归线的性质
- 四、最小二乘估计式的统计性质

第三节 拟合优度的度量

- 一、总变差的分解
- 二、可决系数
- 三、可决系数与相关系数的关系

第四节 回归系数的区间估计和假设检验

- 一、OLS 估计的分布性质
- 二、回归系数的区间估计
- 三、回归系数的假设检验

第五节 回归模型预测

- 一、回归分析结果的报告
- 二、被解释变量平均值预测
- 三、被解释变量个别值预测

第五节 案例分析

第二章小结

第二章主要公式表

思考题与练习题

第二章附录

第三章 多元线性回归模型

第一节 多元线性回归模型及古典假定

- 一、多元线性回归模型
- 二、多元线性回归模型的矩阵形式
- 三、多元线性回归模型的古典假定

第二节 多元线性回归模型的估计

- 一、多元线性回归性参数的最小二乘估计
- 二、参数最小二乘估计的性质
- 三、随机扰动项方差的估计
- 四、多元线性回归模型参数的区间估计

第三节 多元线性回归模型的检验

- 一、拟合优度检验
- 二、回归方程的显著性检验（F-检验）
- 三、回归参数的显著性检验（t-检验）

第四节 多元线性回归模型的预测

- 一、点预测
- 二、平均值的区间预测
- 三、个别值的区间预测

第五节 案例分析

第三章小结

第三章主要公式表

思考题与练习题

第三章附录

第四章 多重共线性

第一节 什么是多重共线性

- 一、多重共线性的含义
- 二、产生多重共线性的背景

第二节 多重共线性产生的后果

- 一、完全多重共线性产生的后果
- 二、不完全多重共线性下产生的后果

第三节 多重共线性的检验

- 一、简单相关系数检验法
- 二、方差扩大因子法
- 三、直观判断法
- 四、逐步回归检测法

*五、特征值与病态指数

第四节 多重共线性的补救措施

- 一、修正多重共线性的经验方法
- 二、逐步回归法

*三、岭回归法简介

第五节 案例分析

第四章小节

第四章主要公式表

思考题与练习题

第五章 异方差性

第一节 异方差性的概念

一、异方差性的实质

二、产生异方差的原因

第二节 异方差性的后果

一、对参数估计式统计特性的影响

二、对参数显著性检验的影响

三、对预测的影响

第三节 异方差性的检验

一、图示检验法

二、戈德菲尔德-夸特 (Goldfeld-Quanadt) 检验

三、White 检验

四、ARCH 检验

五、Glejser 检验

第四节 异方差性的补救措施

一、对模型变换

二、加权最小二乘法

三、模型的对数变换

第五节 案例分析

第五章小结

第五章主要公式表

思考题与练习题

第五章附录

第六章 自相关

第一节 什么是自相关

一、自相关的概念

二、自相关产生的原因

三、自相关的表现形式

第二节 自相关的后果

一、一阶自回归形式的性质

二、自相关对参数估计的影响

三、自相关对模型检验的影响

四、自相关对模型预测的影响

第三节 自相关的检验

一、图示检验法

二、DW 检验法

第四节 自相关的补救

一、广义差分法

二、科克伦—奥克特迭代法

三、其它方法简介

第五节 案例分析

第六章小结

第六章主要公式表

思考题与练习题

第六章附录：

第七章 分布滞后模型与自回归模型

第一节 滞后效应与滞后变量模型

一、经济活动中的滞后现象

二、滞后效应产生的原因

三、滞后变量模型

第二节 分布滞后模型的估计

一、分布滞后模型估计的困难

二、经验加权估计法

三、阿尔蒙法

第三节 自回归模型的构建

一、库伊克模型

二、自适应预期模型

三、局部调整模型

第四节 自回归模型的估计

一、自回归模型估计的困难

二、工具变量法

三、德宾 h-检验

第五节 案例分析

第七章小结

第七章主要公式表

思考题与练习题

第八章 虚拟变量回归

第一节 虚拟变量

一、虚拟变量的基本概念

二、虚拟变量的设置规则

三、虚拟变量的作用

第二节 虚拟解释变量的回归

- 一、用虚拟变量表示不同截距的回归——加法类型
- 二、用虚拟变量表示不同斜率的回归——乘法类型

*第三节 虚拟被解释变量

- 一、线性概率模型 (LPM)
- 二、对数单位模型 (Logit 模型)

第四节 案例分析

第八章小结

第八章主要公式表

思考题与练习题

*第九章 设定误差与测量误差

第一节 设定误差

- 一、设定误差的类型
- 二、变量设定误差的后果

第二节 设定误差的检验

- 一、DW 检验
- 二、拉格朗日乘数 (LM) 检验
- *三、一般性检验 (RESET)

*第三节 测量误差

- 一、模型变量的测量误差
- 二、测量误差的检验

第四节 案例分析

第九章小结

第九章主要公式表

思考题与练习题

第九章附录

*第十章 时间序列计量经济模型

第一节 时间序列计量经济分析的基本概念

- 一、伪回归问题
- 二、随机过程的概念
- 三、时间序列的平稳性

第二节 时间序列平稳性的单位根检验

- 一、单位根过程
- 二、DF 检验
- 三、ADF 检验

第三节 协整

- 一、协整的概念
- 二、协整检验
- 三、误差校正模型

第四节 案例分析

第十章小结

第十章主要公式表

思考题与练习题

第十一章 联立方程组模型

第一节 联立方程模型及其偏倚

一、联立方程模型的性质	
二、联立方程模型中变量的类型	
三、联立方程模型的偏倚性	
第二节 联立方程模型的识别	
一、对模型识别的理解	
二、联立方程模型识别的类型	
三、联立方程模型识别的方法	
第三节 联立方程模型的估计	
一、联立方程模型估计方法的选择	
二、递归模型的估计——OLS 法	
三、恰好识别模型的估计 ——间接最小二乘法	
四、过度识别模型的估计——二段最小二乘法	
第四节 案例分析	
第十一章小结	
第十一章主要公式表	
思考题与练习题	
第十一章附录	
第十二章 实证项目的计量经济研究——课程论文分析	
第一节 实证项目研究的选题	
一、问题的提出	
二、研究题目的选择	
三、文献资料的利用、综述与评价	
第二节 模型设定与数据处理	
一、建模的基本思路	
二、模型设定的要求	
三、模型变量与函数形式的设定	
四、数据的收集与处理	
第三节 计量经济分析	
一、模型的估计	
二、模型的检验	
三、模型的调整	
四、模型计量结果的分析	
五、研究结果的报告	
第十二章附录——实证项目研究（课程论文）示例	
附录	
一、标准正态分布表	
二、t 分布表	
三、 χ^2 分布表	
四、F 分布表	
五、D.W 检验上下界表	
六、DF 分布百分位数表	
七、协整性检验临界值表	
参考文献	

第一章 导 论

引子

“第二次世界大战后的经济学是计量经济学的时代。”

——P. 萨缪尔森 (P. Samuelson)

“在大多数大学和学院中，计量经济学的讲授已成为经济学课程表中最有权威的一部分。”

——R. 克莱因 (R. Klein)

第一节 什么是计量经济学

计量经济学是现代经济学的重要分支。为了深入学习计量经济学的理论与方法，有必要首先从整体上对计量经济学作一些概略性的认识，了解计量经济学的性质、沿革、研究方法以及若干常用的基本概念。

一、计量经济学的产生与发展

计量经济学 (Econometrics) 这个词是 1926 年挪威经济学家、第一届诺贝尔经济学奖获得者弗瑞希 (R.Frisch) 在《论纯经济问题》一文中，按照“生物计量学”(Biometrics) 一词的结构仿造出来的。Econometrics 一词的本意是指“经济度量”，研究对经济现象和经济关系的计量方法，因此 Econometrics 有时也译为“经济计量学”。将 Econometrics 译为计量经济学，是为了强调计量经济学是一门经济学科，不仅要研究经济现象的计量方法，而且要研究经济现象发展变化的数量规律。

计量经济学的产生源于对经济问题的定量研究，这是社会经济发展到一定阶段的客观需要。经济现象本来就充满着数量关系，人们很早就探索用定量的方式研究经济问题。早在 17 世纪英国经济学家、统计学家威廉·配第在《政治算术》中就运用统计方法研究社会经济问题，主张用“数字、重量和尺度”来阐明经济现象。以后的相当一段时间，经济学家们也力图运用数学方法研究经济活动，用数学语言和公式去表达经济范畴和经济规律。但这都还没有形成计量经济学。计量经济学作为经济学的一门独立学科被正式确立，其标志一般认为是 1930 年 12 月弗瑞希和丁伯根 (J.Tinbergen) 等经济学家发起在美国克里富兰成立国际计量经济学会。

第二次世界大战以后，计量经济学在西方各国的影响迅速扩大，发展成为经济学的重要分支。特别是从 20 世纪 40 年代到 60 年代，经典计量经济学逐步完善并得到广泛应用。美国著名经济学家、诺贝尔经济学奖获得者萨缪尔森（P.Samuelson）认为：“第二次世界大战后的经济学是计量经济学的时代”。事实上，在世界诺贝尔经济学奖获得者中，相当一部分都是计量经济学家。

20 世纪 70 年代以来，计量经济学的理论和应用又进入一个新的阶段。首先是计算机的广泛应用和新的计算方法大量提出，所使用的计量经济模型的规模越来越大。更重要的是非经典计量经济学的理论和应用有了新的突破。微观计量经济学、非参数计量经济学、时间序列计量经济学和动态计量经济学等的提出，使计量经济学产生了新的理论体系，协整理论、面板数据、对策论、贝叶斯方法等理论在计量经济学中的应用已成为新的研究课题。

应该看到，计量经济学的发展是与现代科学技术成就结合在一起的，它反映了社会化大生产对各种经济因素和经济活动进行数量分析的客观要求。经济学从定性研究向定量分析的发展，是经济学逐步向更加精密、更加科学发展的表现。正如马克思强调的：一种科学只有成功地运用了数学以后，才算达到了完善的地步。因此另一获得诺贝尔经济学奖的经济学家克莱因（R.Klein）认为：“计量经济学已经在经济学科中居于最重要的地位”。

计量经济学的一个重要特点是它自身并没有固定的经济理论，计量经济学中的各种计量方法和技术，大多来自数学和统计学。我们只要坚持以科学的经济理论为指导，紧密结合中国经济的实际，就能够使计量经济学的理论与方法在中国的经济理论研究和现代化建设中发挥重要的作用。

二、计量经济学的性质

计量经济学的奠基人弗瑞希指出：计量经济学“是统计学、经济学和数学的结合”，“三者结合起来，就有力量，这种结合便构成了计量经济学”。

美国现代经济词典认为：计量经济学是用数学语言来表达经济理论，以便通过统计方法来论述这些理论的一门经济学分支。

萨缪尔逊、库普曼斯、斯通等三位著名经济学家在 1954 年计量经济学家评审委员会的报告中认为：“计量经济学可定义为：根据理论和观测的事实，运用合适的推理方法使之联系起来同时推导，对实际经济现象进行的数量分析。”

尽管这些经济学家对计量经济学定义的表述各不相同，但可以看出，计量经济学不是对经济的一般度量，它与经济理论、统计学、数学都有密切的关系。事实上，计量经济学是以经济理论和经济数据的事实为依据，运用数学、统计学的方法，通过建立数学模型来研究经

济数量关系和规律的一门经济学科。应当注意，计量经济学所研究的主体是经济现象及其发展变化的规律，所以它是一门经济学科。计量经济学当然会运用大量的数学方法，特别是许多数理统计方法，但数学在这里只是工具，而不是研究的主体。

计量经济学的目的是要把实际经验的内容纳入经济理论，确定表现各种经济关系的经济参数，从而验证经济理论，预测经济发展的趋势，为制定经济政策提供依据。为此计量经济学不仅要寻求经济计量分析的方法，而且要对实际经济问题加以研究，要解决达到上述目的的理论和方法论问题。这样，计量经济学分成了两种类型：即理论计量经济学和应用计量经济学。

理论计量经济学研究如何建立合适的方法去测定由计量经济模型所确定的经济关系。现实的经济活动和经济关系异常复杂，一般来说各种经济变量之间并不是精确的函数关系，经济变量间的数量关系不是那么确定，也就是说模型中往往包含一些随机的无法直接控制的因素，所以理论计量经济学要较多地依赖数理统计学方法。除了介绍计量经济模型普遍应用的参数估计方法与检验方法以外，由于经济现象的复杂性，各种实际的经济关系不一定都服从一般的统计规律，理论计量经济学还须研究当一般的统计假定条件不完全满足时将会产生的结果，并寻求解决这些问题的专门方法，也就是说还会面临许多特殊的经济问题，形成一些专门的计量经济方法。所以理论计量经济学是适合于经济关系计量的方法论学科。

应用计量经济学是运用理论计量经济学提供的工具，研究经济学中某些特定领域的经济数量问题，例如生产函数、消费函数、投资函数、供给函数、劳动就业，等等。应用计量经济学以建立应用计量经济学模型为主要内容，强调应用模型的经济学和经济统计学基础，侧重于建立与应用模型过程中实际问题的处理。应用计量经济学研究的是具体的经济现象和经济关系，研究它们在数量上的联系及其变动规律性。除了计量经济方法以外，应用计量经济学更多地要依据经济学理论所确定的经济规律，而且要依据经济统计提供的反映现实经济现象和经济关系的观测数据，运用计量经济模型分析经济结构，预测经济的发展趋势，对经济政策作定量的评价。

三、计量经济学与其他学科的关系

从前面的讨论可以看出，计量经济学是与经济学、经济统计学及数理统计学都有关系的交叉学科。但计量经济学又不是这些学科的简单结合，它与这些学科既有联系又有区别。

计量经济学研究的主体是经济现象和经济关系的数量规律，这决定了计量经济学应当以经济学提供的理论原则和揭示的经济规律为依据。经济学理论所说明的经济规律，是计量经济学分析经济数量关系的理论依据。离开了经济理论的指导，计量经济学就可能无的放矢，

计量经济学的应用也可能会步入歧途。

但是计量经济学并不是盲目地重复经济理论，计量经济学研究是把经济理论与客观现实联系起来分析，计量经济分析的成果或者是对经济理论确定的原则加以验证与充实，或者可以否定某些经济理论原则，而作出补充或修改。计量经济学与经济学的明显区别，在于一般的理论经济学主要根据逻辑推理得出结论，主要用文字说明经济现象和过程的本质与规律，大多具有定性的性质。理论经济学有时也会涉及经济现象的数量关系，例如说明价格与商品需求量及供应量成正比或反比的关系，但经济理论并不提供这类经济关系数量上的度量，并不说明价格的变动将会使供应量和需求量具体增加或降低多少。计量经济学则要对经济理论所确定的经济关系作出定量的估计，也就是对经济理论提供经验的内容。

经济统计学也研究对经济现象的计量，只不过是侧重于对社会经济现象的描述。经济统计提供的数据，是计量经济学据以估计参数、验证理论的基本依据。离开了经济统计，任何对实际经济问题的经济计量分析都会寸步难行。计量经济学对经济统计的这种依赖性是由经济活动的特殊性决定的。经济现象是人所从事的社会性活动，它不可能像对自然现象的物理实验和化学实验那样，可以在实验室中严格控制其他条件不变，去反复观测某种因素变动对所研究现象的影响。经济现象不可能人为地控制“其他条件不变”，能够做的只是被动地观测客观经济活动的既成事实，也就是分析对实际经济现象观测所得的统计数据。

计量经济学所研究的经济现象并不都是呈现为精确的函数关系，计量经济模型中包含了随机误差项，这样模型中的一些变量和所估计的参数都成为了随机变量。数理统计学是研究随机变量统计规律性的学科，所以数理统计学中的回归分析、参数估计、假设检验、方差分析等方法在计量经济学中得到了全面运用，可以说数理统计学是计量经济学的方法论基础。然而，数理统计学只是抽象地研究一般随机变量的统计规律，主要讨论在一定假设条件下一般随机变量的概率分布性质，以及特征值的估计与推断。而计量经济学是从具体的经济模型出发，其参数都具有特定的经济意义，研究对模型参数的估计与推断时，不仅要看在数学原理上是否通得过，还要看与实际的经济内容是否一致。而且，在实际经济问题的计量中，数理统计中一些标准的假定经常不能满足，还需要建立许多专门的经济计量方法。所以，计量经济学并不只是对数理统计方法的简单应用。

作为对计量经济学与其他相关学科关系的总结，可以引述早在 1933 年 R. 弗瑞希为《计量经济学》杂志写的发刊词中的一段话：“对经济的数量研究可以从好几个方面着手，但其中任何一个方面就其本身来说都不应该与计量经济学混为一谈。因此，计量经济学与经济统计学决非一码事；它也不同于我们所说的一般经济理论，尽管经济理论大部分都具有一定的

数量特征；计量经济学也不应视为数学应用于经济学的同义语。经验表明，统计学、经济理论和数学这三者对于实际理解现代经济生活中的数量关系来说，都是必要的。但任何一种观点本身都不是充分条件。三者结合起来才是强有力的，正是这种结合才构成了计量经济学。”

第二节 计量经济学的研究步骤

运用计量经济学研究经济问题，一般可分为四个步骤：即确定变量和数学关系式——模型设定；分析变量间具体的数量关系——估计参数；检验所得结论的可靠性——模型检验；作经济分析和经济预测——模型应用。

一、模型设定

所谓经济模型是指对经济现象或过程的一种数学模拟。社会经济现象和过程是非常复杂的，影响因素众多，经济模型只能把所研究的主要经济因素（表现为经济变量）之间的关系，用适当的数学关系式近似地、简化地表达出来。例如，为了研究居民的消费行为，根据经济学中关于消费行为的理论，认为居民消费支出与其收入成正比例，可将二者的关系表示为如下消费函数：

$$Y = \alpha + \beta X \quad (1.1)$$

其中：Y 为居民消费支出；X 为居民家庭收入； α 和 β 为参数。

（1.1）式中的 β 实际是经济学中的边际消费倾向（MPC）， β 作为斜率系数是消费增量 ΔY 与收入增加量 ΔX 的比例，即 $\beta = \Delta Y / \Delta X$ 。然而，在现实的经济生活中，居民消费支出并不像（1.1）那样，是家庭收入的精确函数。由于还有许多其他未加入模型的因素也会影响居民的消费行为，相同收入的家庭，其消费支出不一定完全相同，所以（1.1）式那样的模型还不是适于对实际经济活动作计量分析的计量经济模型。为了把实际居民消费与实际收入水平的关系表现出来，还需要在模型中引入一个随机误差项，即

$$Y = \alpha + \beta X + u \quad (1.2)$$

其中的 u 是随机误差项，也称随机扰动项。像（1.2）式那样，包含了经济变量、待确定的参数 α 和 β ，并包含了随机误差项 u 的方程式，才是适于对实际经济活动作计量分析的计量经济模型。计量经济模型可以如（1.2）式那样只是一个方程式，这称为单一方程模型。

在有的情况下，需要用相互联系的若干个方程构成的方程组去描述更为复杂的经济关系，这种计量模型称为联立方程模型。

显然，在建立计量经济模型时，为了简化和计量的方便，通常不可能把所有的因素都列入模型，而只能抓住主要影响因素和主要特征，而不得不舍弃某些因素。同时，模型中变量之间的关系可能设计为线性关系，也可能设计为其他非线性关系。建立模型时，模型中变量的取舍及相互关系形式的设计，一定程度上是决定于研究者的主观认识，当不同的研究者对所研究经济问题的认识有差异时，所使用的模型可能会不完全相同。所以，模型的具体形式是需要研究者去设定的问题。设定模型是计量经济研究的关键步骤，设定计量经济模型既是一门科学，又是一门艺术。建立一个好的计量模型，要靠丰富的专业知识，要有适当的方法，更要靠对建模实践的不断总结。

一般说来，设定一个合理的计量经济模型，主要应注意以下几个方面的问题：

1、要有科学的理论依据

建立经济模型是为了反映实际经济活动的规律性，必须对所研究的经济现象的相互关系作科学的理论分析，尽可能使模型真实地反映经济现象实际的依存关系。对别人成功应用过的计量经济模型，也要从经济机理上具体分析，注意模型的应用条件是否符合所研究问题的实际，不应简单地生搬硬套。

2、模型要选择适当的数学形式。

模型的数学形式可以是单一方程，也可以是联立方程，每一个方程可以表现为线性形式，也可以表现为非线性形式。这要根据研究的目的、所研究经济问题的复杂程度以及所掌握的数据资料来决定。可以利用经济学和数理经济学的成果，或利用样本数据绘制变量之间关系的图形作参考。在实际建立模型的过程中，应根据所研究现象相互关系的性质，通过对实际统计资料的试验和分析，经过反复比较，选择尽可能合理的模型数学形式。另外要注意所构造的方程必须是有解的，特别是在建立联立方程模型时，要使内生变量的数目与方程个数相适应。在选择模型数学形式时还应注意，在能够达到研究目的的前提下，应当尽可能选择更为简捷的数学形式，不应只是盲目追求数学形式上的“完美性”。

3、方程中的变量要具有可观测性。因为只有可观测的变量才可能取得实际的统计数据，也才可能对模型中的参数作出具体的估计。

二、估计参数

参数与变量不同，它是计量经济模型中表现经济变量相互依存程度的那些因素，通常参数在模型中是一些相对稳定的量。计量经济模型中的参数决定着变量之间的数量关系，一旦

参数确定了，整个经济系统的基本结构就确定了。例如（1.2）式中作为边际消费倾向的参数 β ，决定着收入与消费的基本结构关系，这种反映经济结构特性的参数也称为结构参数。

在经济总体中，反映经济结构的参数与变量不同，一般来说参数不能直接观测而且是未知的。我们能够获得的，往往只是所研究总体中变量的若干样本观测数据。由于随机误差项的存在，变量之间的数量关系并不呈现为确定的函数关系，通常也不可能精确地去计算参数的数值。如何通过变量的样本观测数据正确地估计总体模型的参数，这是计量经济学研究的核心内容。

经过实际样本信息估计出的参数数值称为参数的估计值，但是由于样本毕竟不等于总体，参数的样本估计值并不一定等于总体参数的真实值。如果用一定的方法能够获得对参数估计过程的公式，这种公式则称为参数的估计式。参数估计式是模型中变量样本观测值的代数式，只要将变量的样本观测值直接代入估计式，即可得到参数的估计值。如何去确定满足计量经济要求的参数估计式，是理论计量经济学的主要内容之一。

三、模型检验

模型中的参数被估计以后，一般说来这样的模型还不能直接加以应用，还需要对估计的计量经济模型作某些检验，其原因是多方面的。首先，在设定模型时，对所研究经济现象规律性的认识可能并不充分，所依据的经济理论对所研究对象也许还不能作出正确的解释和说明。或者虽然经济理论是正确的，但可能我们对问题的认识只是从某些局部出发，或者只是考察了某些特殊的样本，以局部去说明全局的变化规律，可能导致偏差。其次，我们用以估计参数的统计数据或其他信息可能并不十分可靠，或者较多地采用了经济突变时期的数据，不能真实代表所研究的经济关系，或者由于样本太小，所估计的参数只是抽样的某种偶然结果。此外，我们所建立的模型、采用的方法、所用的统计数据，都有可能违反计量经济的基本假定，这也可能导出错误的结论。所谓模型检验，就是要对模型和所估计的参数加以评判，判定在理论上是否有意义，在统计上是否有足够的可靠性。

对计量经济模型的检验主要应从以下几方面进行：

1、经济意义的检验

模型中的变量和参数都有特定的经济意义，经济理论通常对这些变量以及参数的符号和取值范围作出了理论说明，如果所估计的模型与经济理论完全相符，则说明我们所观测的事实证实了这种理论；如果所估计的模型与理论说明不相符，一般来说应当舍弃所估计的模型，设法从模型设定、估计方法、统计数据等方面找出导致错误结论的原因。

但是应强调，实践是检验真理的唯一标准，任何经济学理论，只有当它成功地解释了过去，才能为人们所接受。如果经过反复研究，证明计量经济模型和估计的参数完全正确，而是经济理论本身不完备，这时则应提出修正经济理论的建议。所以，计量经济学模型对检验经济理论、发现和发展经济理论也有重要意义。

2、统计推断检验

模型的参数是用变量的观测值估计的，为了检验参数估计值是否抽样的偶然结果，需要运用数理统计中的统计推断方法，对模型及参数的统计可靠性作出说明。对计量经济模型的统计推断检验，包括对模型的拟合优度的检验、用假设检验和方差分析方法对变量显著性的检验等。

3、计量经济学检验

计量经济学检验主要是检验模型是否符合计量经济方法的基本假定，例如检验模型是否存在多重共线性，检验模型中的随机扰动项是否存在自相关和异方差性，检验模型的可识别性，检验模型中经济变量的平稳性，等等。当模型违反计量经济方法的基本假定时，通常的计量经济方法将失去效用或将导致错误的结论，这时必须对模型作必要的处理，并重新估计模型的参数。

4、模型预测检验

这是指将估计了参数的模型用于实际经济活动的预测，然后将模型预测的结果与经济运行的实际结果相对比，以此检验模型的有效性。

四、模型应用

经过估计参数和模型检验，确认为可靠的计量经济模型，才可以用于实际的经济计量分析。计量经济模型主要可以用于经济结构分析、经济预测和政策评价等几个方面。

所谓经济结构分析，是指用已经估计出参数的模型，对所研究的经济关系进行定量的考察，以说明经济变量之间的数量比例关系。也就是说分析当其他条件不变时，模型体系中的解释变量发生一定的变动，对被解释变量的影响程度。常用的经济结构分析方法有边际分析、弹性分析、乘数分析、比较静力学分析等等。例如，说明一个国家或地区国民总收入 Y 与总消费支出 C 关系的消费函数模型：

$$C = \alpha + \beta Y + u \quad (1.3)$$

通常这是一条斜率为正值且小于 1 的直线。模型中的参数 β 的经济意义是边际消费倾向 $MPC = \Delta C / \Delta Y$ ，假如估计的参数为 $\beta = 0.8$ ，这说明国民总收入每增加 1 亿元，总消费支

出将增加 0.8 亿元，这是宏观经济结构分析中很有意义的数。在此基础上还可进行边际储蓄趋向和收入增长的乘数分析，边际储蓄趋向 $\Delta S/\Delta Y = 1 - MPC$ ，因为收入增长乘数 M 为：

$$M = 1 / (1 - MPC) \tag{1.4}$$

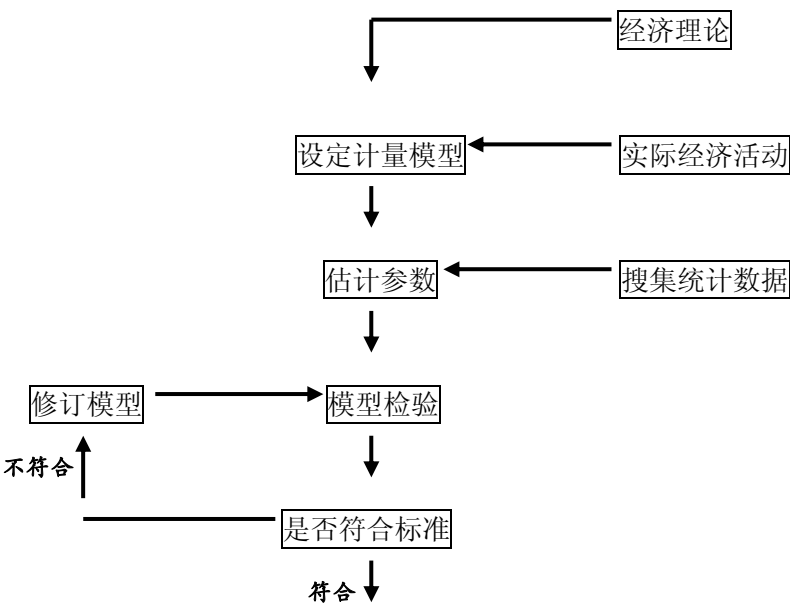
若已估计出 $\beta = MPC = 0.8$ ，则乘数 $M = 5$ ，这说明当投资增加 1 个单位时，将导致国民总收入增加 5 个单位，这又为经济分析提供了重要的定量信息。

所谓经济预测，是指利用估计了参数的计量经济模型，由已知的或预先测定的解释变量，去预测被解释变量在所观测的样本数据以外的数值。计量经济模型本身就是试图从已经发生的经济活动中找出变化规律，然后把这种规律用于样本以外数据的预测。经济预测可以是对被解释变量未来时期的动态预测，也可以是对被解释变量在不同空间状况的空间预测。

所谓政策评价，是利用计量经济模型对各种可供选择的政策方案的实施后果进行模拟测算，从而对各种政策方案作出评价。在这种情况下，我们是把计量经济模型当作经济运行的“实验室”，去模拟所研究的经济体系，分析整个经济体系对各种假设的政策条件的反映。在实际的政策评价时，经常把模型中的某些变量或参数视为可用政策调整的“政策变量”，然后分析“政策变量”的变动对被解释变量的影响。

显然，计量经济模型应用的经济结构分析、经济预测和政策评价三个方面有密切关系，经济结构分析的结果，可用于经济预测，经济预测的结果是政策评价的依据，而政策评价本身，实际就是一种条件预测。

综上所述，完整的计量经济研究过程可以表示为图 1.1：



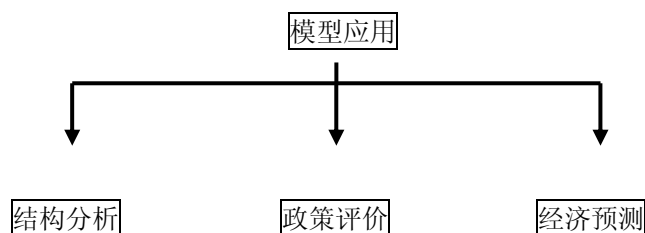


图 1.1 计量经济学的研究过程

第三节 变量、参数、数据与模型

一、计量经济模型中的变量

计量经济模型有多种构成因素，其中一些在不同的时间或空间有不同的状态，会取不同的数值，并且是可以观测的因素，这类因素称为经济变量。例如(1.2)式中的居民家庭收入 X 和居民消费支出 Y 都是经济变量。

计量经济模型中的变量可分为若干类型。从所描述的经济活动形态看，经济变量可分为流量和存量。某些变量具有时间维度，是按一定时期测度的总量，它们是一定时期内累计发生的数量，如国内生产总值、投资量、消费量等，这类变量反映的是经济活动的“流量”。另一些变量不具时间维度，是在一定时点上测度的总量，它们表明某一时点所存在状态的总量，如金融资产、金融负债等，这类总量反映的是经济“存量”。

从变量的因果关系上，可分为解释变量 (Explanatory variable) 和被解释变量 (Explained variable)。在模型中，解释变量是变动的原因，被解释变量是变动的结果。被解释变量是模型要分析研究的对象，也常称为“应变量” (Dependent variable)、“回归子” (Regressand) 等，例如 (1.2) 式中的消费支出 Y 。解释变量也常称为“自变量” (Independent variable)、“回归元” (Regressor) 等，是说明应变量变动主要原因的变量，例如 (1.2) 式中的居民家庭收入 X 。表述被解释变量和解释变量变的术语较多¹，为了表述上尽量一致，避免产生混淆，本书中统一使用“解释变量”表示应变量变动原因的变量；而用“应变量”或“被解释变量”表示分析研究的对象，即作为变动结果的变量。

从变量的性质，又可把变量分为内生变量和外生变量。内生变量是其数值由模型所决定的变量，内生变量是模型求解的结果，例如 (1.2) 式中的消费支出 Y 。外生变量是其数值

¹见古扎拉蒂《计量经济学》上册，中国人民大学出版社，第9页。被解释变量 (Explained variable) 有时也被称为因变量，为了与作为变动原因的变量相区别，本书中称为应变量 (Dependent variable)。

由模型以外决定的变量，例如（1.2）式模型中的家庭收入。在计量经济模型中，外生变量数值的变化能够影响内生变量的变化，而内生变量却不能反过来影响外生变量。在内生变量中，有一些是过去时期的内生变量或称滞后的内生变量，例如在研究消费—收入模型时可能涉及上一期的收入或上一期的消费支出，又如在研究某地区居民收入对消费的影响时可能涉及全国居民的收入量，这种过去时期的、滞后的或更大范围的内生变量，不受本模型研究范围的内生变量的影响，但能够影响我们所研究的本期的内生变量，这种内生变量称为前定内生变量。在模型中前定内生变量的作用视同于外生变量，并与外生变量一起称为前定变量。在单一方程模型中，前定变量一般作为解释变量，内生变量一般作为被解释变量或应变量，而在联立方程模型中内生变量既可作为应变量，又可作为解释变量。

确定模型中的变量，是建立计量经济模型的重要环节，应变量要选择最能反映所研究对象变动情况的变量，例如最能反映居民消费行为的是“居民消费支出”，最能够反映全社会生产成果总量的是“国内生产总值（GDP）”，等等。解释变量则应尽量选择最能够说明应变量变动的主要原因，并能够独立影响应变量的那些变量，次要的变动原因应被归入到随机扰动项中。对变量的选择还要考虑可观测性，有的因素虽然对因变量有重要影响，但是无法获取其观测值，例如家庭财产可能对家庭消费有影响，但家庭的财产数量很难取得数据，这类无法观测的因素不能实际度量，也不宜作为变量列入计量经济模型。

二、参数估计的方法

计量经济模型中的参数一般是未知的，需要根据样本信息去加以估计。估计模型中参数的方法有很多种。例如对于单一方程模型，最常用的是普通最小二乘法、极大似然估计法等。对于联立方程模型常用二段最小二乘法和三段最小二乘法等去估计参数。这些估计方法都是建立在一定假设前提的基础上的，当估计条件不完全满足时，还需要一些特殊的估计方法。由于抽样波动的存在，加之前提条件、估计方法及所确定的估计式不一定那么完备，所得到的参数估计值与总体参数的真实值并不一致，这就要求所得到的参数估计值应符合“尽可能地接近总体参数真实值”的准则。在各种条件下如何寻求模型参数合理的估计方法，是计量经济学研究的主要内容。不过，在理论计量经济学中并不侧重于直接研究参数估计值本身，而是着重于论述所导出的参数估计式是否符合“尽可能地接近总体参数真实值”这样的准则。通常选择参数估计式时应考察其无偏性、最小方差性等统计性质，或者考察大样本时的统计性质。

在实际的计量经济研究中，运用样本数据对参数的估计与检验，将面临很大的计算工作量。现在这方面的工作已经可以由各种有关的计算机应用软件来实现，计算机应用软件已成

为学习计量经济学必不可少的部分。本书采用的计算机应用软件是应用十分广泛的EViews(Econometrics Views)，将结合每一章的内容介绍EViews的使用方法。

三、计量经济学中应用的数据

估计计量经济模型参数的基本依据，是通过对所研究经济变量实际观测所取得的数据。数据是对客观事物信息的一种反映，这种信息如以某种量的标志显现出来就称其为数据。在计量经济研究中使用的数据，主要是各种经济统计数据，也可以是通过专门调查取得的数据，还可以是人为构造的数据。可用于估计参数的数据主要有以下几类：

1、时间序列数据（Time Series Data）

把反映某一总体特征的同一指标的数据，按照一定的时间顺序和时间间隔（如月度、季度、年度）排列起来，这样的统计数据称为时间序列数据。例如逐年的国内生产总值和消费支出、逐月的物价指数……等等。时间序列数据可以是时期数据，也可以是时点数据。

2、截面数据(Cross-Section Data)

同一时间（时期或时点）某个指标在不同空间的观测数据，称为截面数据。“不同的空间”可以是指不同的地理区域，也可以是指不同的行业、部门或个人。例如，同一时间不同家庭的收入和消费支出、某一年各个省（市）的国内生产总值，等等。

3、面板数据（Panel Data）

面板数据指时间序列数据和截面数据相结合的数据，例如在居民收支调查中收集的对各个固定调查户在不同时期的调查数据，又如全国各省市不同年份的经济发展状况的统计数据，就都是面板数据。

4、虚拟变量数据(Dummy Variables Data)

时间序列数据和截面数据都是反映定量事实的数据，这是计量经济分析中用得最多的最基本的数据。但是还有一些定性的事实，不能直接用一般的数量去计量，例如政府政策的变动、自然灾害、政治因素、战争与和平状态……等等。在计量经济研究中常发现，某些客观存在的定性现象确实对所研究的经济变量有明显的影响，需要把它们引入计量经济模型中，这时常用人造的虚拟变量去表示这类客观存在的定性现象“非此即彼”的状态。通常以1去表示某种状态发生，以0表示该种状态不发生。这样的虚拟变量虽然是人为构造的，但反映了客观存在的定性现象，也可以视为一种数据用作模型参数的估计和检验。

以上各种数据虽然都可用于计量经济模型的估计和检验，但是应注意，由于这些数据的性质各不相同，在具体运用时可能不满足某些假定条件而给计量经济分析带来一些影响。例如时间序列数据若是非平稳的，可能造成“伪回归”；截面数据往往存在异方差；利用面板

数据的计量经济模型已成为计量经济学研究的专门问题。

除了模型的正确设定以外，能否取得用于实际计量的适合的样本数据，是计量经济研究成败的关键。计量经济分析中使用的数据主要来自于经济统计，常用的数据可从各种统计年鉴等出版物中取得，特殊的数据则需进行专门的调查才能得到。计量经济研究中使用的数据，要力求真实、可靠、完整，数据的质量直接关系到所估计参数的可靠性。对明显失真的数据，应当予以剔除。在经济结构发生重大变革的时期，其统计数据往往不能反映经济变动的真实趋势和规律，这种“经济突变”时期的数据也不宜直接用于估计模型的参数。此外，我们有时很难直接找到模型所需要的数据，这时可能还需要对能够得到的数据作重新加工，或者试验寻求与所研究的变量高度相关的代用数据，或者对模型的变量与结构加以调整。

计量经济学中利用的数据是可能获得的统计数据，实际的统计数据可能会有观测误差，也可能数据的数量无法满足估计参数的要求，这些数据还可能不满足参数估计方法的基本假定。这样一来，由于数据可能引发诸如自由度问题、多重共线性、序列相关、异方差性等一系列问题。如何设法解决由数据引起的问题，也是理论计量经济学和应用计量经济学要专门研究的内容。

四、计量经济模型的建立

在计量经济研究中，模型是对实际经济现象或过程的一种数学模拟，再完美的模型也不可能将所有的因素都纳入其中，模型只不过是对可计量的复杂经济现象的一种简化与抽象。因此模型只能在一定的假设前提下，忽略众多次要因素，而突出若干所关注的主要经济变量，把有关经济变量的相互依存关系表现为方程式。模型的建立主要靠对现实经济问题的深入研究，要遵循科学的理论原则，也要运用适当的方法。某些经济变量的相互关系通常可以利用来建立计量经济模型，这些关系主要有：

1、行为关系

行为关系指描述决策者经济行为的某些变量与其它变量的关系。例如居民消费行为与其收入、物价水平等的关系。利用行为关系建立的模型称为行为方程式。

2、技术（工艺）关系

这是反映由科学技术水平决定的经济变量间的数量关系，例如说明投入的生产要素与产出的生产成果的技术关系，如著名的柯柏—道格拉斯生产函数，产量 Q 与资本投入量 K 、劳动投入产出量 L 的关系为：

$$Q = AK^{\alpha}L^{\beta}e^u \quad (1.5)$$

其中 A 、 α 、 β 为参数， u 为随机项。

又如投入产出模型中的生产量 X_j 与消耗量 x_{ij} 间的关系

$$x_{ij} = a_{ij} X_j \quad (1.6)$$

根据生产技术关系建立的模型称为技术方程式。

3、制度关系

制度关系指经济现象之间由政府政策和规定的制度所决定的关系。例如销售税的数量决定于销售额和税率，其中税率是由政府规定的。这样建立的模型称为制度方程式。

4、定义关系

这是指根据定义而表达的恒等式。这类关系是由经济理论或客观存在的经济关系决定的恒等关系，例如：

$$\text{国内生产总值} = \text{消费} + \text{投资} + \text{净出口}$$

国民经济中许多平衡关系都可以建立恒等关系，这样的模型称为定义方程式。

以上几种方程式中，最重要、最常用的是行为方程式和技术方程式，这两种方程中都有未知参数需要估计，且每个方程说明了经济结构的某一方面，所以这些方程称为结构方程式。

第一章小结

1、计量经济学是以经济理论和经济数据的事实为依据，运用数学、统计学的方法，通过建立数学模型来研究经济数量关系和规律的一门经济学科。计量经济学与理论经济学、数理经济学、经济统计学、数理统计学既有区别又有联系。

2、计量经济研究分为模型设定、参数估计、模型检验、模型运用等四个步骤。

3、模型的设定主要是选择变量和确定变量间联系的数学形式。适于对实际经济活动作计量分析的计量经济模型应包含经济变量、待确定的参数和随机误差项。行为方程、技术方程、制度方程和定义方程可作为建立模型时参考。

4、计量经济模型中的变量分为被解释变量（应变量）和解释变量、内生变量和外生变量。

5、参数是计量经济模型中表现经济变量相互依存程度的因素，通常具有相对稳定性。参数无法直接观测和计算，只能用适当的方法根据变量的样本观测值去估计。参数估计的方法应符合“尽可能地接近总体参数真实值”的准则。

6、计量经济研究中应用的数据包括时间序列数据、截面数据、面板数据、虚拟变量数据等。

7、对模型检验包括经济意义检验、统计推断检验、计量经济学检验和模型预测检验。

8、计量经济模型主要可应用于经济结构分析、政策评价和经济预测。

思考题

1.1 怎样理解产生于西方国家的计量经济学能够在中国的经济理论研究和现代化建设中发挥重要作用？

1.2 理论计量经济学和应用计量经济学的区别和联系是什么？

1.3 怎样理解计量经济学与理论经济学、经济统计学的关系？

1.4 在计量经济模型中应变量和解释变量的作用有什么不同？

1.5 一个完整的计量经济模型应包括哪些基本要素？你能举一个例子说明吗？

1.6 假如你是中央银行货币政策的研究者，需要你对增加货币供应量促进经济增长提出建议，你将考虑哪些因素？你认为可以怎样运用计量经济学的研究方法？

1.7 计量经济学模型的主要应用领域有那些？

1.8 如果要根据历史经验预测明年中国的粮食产量，你认为应当考虑那些因素？应当怎样来设定计量经济模型？

1.9 参数和变量的区别是什么？为什么对计量经济模型中的参数通常只能用样本观测值去估计？

1.10 你能分别举出三个时间序列数据、截面数据、面板数据、虚拟变量数据的实际例子，并分别说明这些数据的来源吗？

1.11 为什么对已经估计出参数的模型还要进行检验？你能举一个例子说明各种检验的必要性吗？

1.12 为什么计量经济模型可以用于政策评价？其前提条件与什么？

第二章 简单线性回归模型

引子:

中国旅游业总收入将超过 3000 亿美元吗?

根据国家旅游局统计, 2004 年中国旅游业快速增长, 全年入境人数 1.08 亿人次, 分别比 2003 年和 2002 年增长 18% 和 10%; 旅游外汇收入可达 250 亿美元, 分别比 2003 年和 2002 年增长 43.7% 和 22.6%; 国内旅游人数 9.3 亿人次以上, 国内旅游收入 4000 亿元以上, 比 2002 年分别增长 5.9% 和 3.1% 以上; 出境 2800 万人次, 分别比 2003 年和 2002 年增长 38.5% 和 68.7%。目前中国旅游业利用外资的规模达到 500 亿美元, 占国内各行业吸收外资总额的 11%。中国公民可以组团前往的旅游目的地国家和地区将达到 63 个。现在, 中国人均收入已经跨越 1000 美元关口, 按照国际经验, 将触发国内社会消费结构的升级。居民消费将由实物消费为主走上实物消费与服务消费并重的轨道。在消费结构升级的推动下, 中国旅游业将迎来新一轮的长期增长周期。据《中国旅游业发展“九五”计划和 2010 年远景目标纲要》, 到 2010 年, 中国旅游入境人数将达 6400 万-7100 万人次, 国际旅游外汇收入 380 亿-410 亿美元; 国内旅游人数将达到 20 亿-25 亿人次, 国内旅游收入 10000 亿-10500 亿元人民币; 两项合计总产出将达 13000 亿~14000 亿元人民币, 旅游总收入占 GDP 的比例将达 8%。旅游业已经成为中国第三产业中最具活力与潜力的新兴产业和国民经济中新的增长点。另据世界旅游及旅行理事会 (WTTC) 预测, 未来 10 年间, 中国旅游业将保持年均 10.4% 的增长速度, 其中个人旅游消费将以年均 9.8% 的速度增长, 企业/政府旅游的增长速度将达到 10.9%。到 2020 年, 中国将成为第一大旅游目的地国和第四大客源输出国。从 2004 中国国际旅游交易会上获悉, 到 2020 年, 中国旅游业总收入将超过 3000 亿美元, 相当于国内生产总值的 8% 至 11%。(资料来源: 国际金融报 2004 年 11 月 25 日第二版)《

推动中国旅游业快速发展的原因是多方面的, 是什么决定性的因素能使中国旅游业总收入到 2020 年达到 3000 亿美元? 旅游业的发展与这种决定性因素的数量关系究竟是什么? 显然, 需要寻求一些方法研究相互联系的经济变量之间的数量关系, 对这类问题的研究应当考虑以下几个方面:

- (1) 确定作为研究对象的经济变量 (如中国旅游业总收入)
- (2) 分析影响研究对象变动的主要因素 (如中国居民收入的增长)
- (3) 分析各种影响因素与所研究经济现象的相互关系 (决定相互联系的数学关系式)
- (4) 确定所研究的经济问题与影响因素间具体的数量关系 (需要特定的方法)
- (5) 分析并检验所得数量结论的可靠性 (需要统计检验)
- (6) 运用数量研究结果作经济分析和预测 (对数量分析的实际应用)

对经济变量相互关系的计量，最基本的方法是回归分析。回归分析是计量经济学的主要工具，也是计量经济学理论和方法的主要内容。只有一个解释变量的线性回归模型是最简单的，称为简单线性回归模型或一元线性回归模型。本章从最简单的一元线性回归模型入手，讨论在基本假定满足的条件下，对经济变量关系计量的基本理论和方法，这也是以后各章的重要基础。

第一节 回归分析与回归函数

一、相关分析与回归分析

（一）经济变量间的相互关系

许多社会与经济现象，除了自身的变动以外，它们相互之间很可能有一定的依存关系。各种经济变量相互之间的依存关系有两种不同的类型：一种是确定性的函数关系，另一类是不确定性的统计关系，也称为相关关系。

当一个或若干个变量 X 取一定数值时，某一个变量 Y 有确定的值与之相对应，我们称变量间的这种关系为确定性的函数关系。例如当销售价格 P 不变的情况下，某种商品销售量 X 与销售额 Y 之间的关系可表示为 $Y = P X$ 。一般情况下，确定性的函数关系可表示为 $Y = f(X)$ 。

当一个或若干个变量 X 取一定值时，与之相对应的另一个变量 Y 的值虽然不确定，但却按某种规律在一定范围内变化，我们称变量之间的这种关系为不确定的统计关系或相关关系，一般可表示为 $Y = f(X, u)$ ，其中 u 为随机变量。例如居民的可支配收入 X 与居民的消费支出 Y 之间的关系，通常具有相同收入水平居民的消费支出并不完全相同，这时居民可支配收入 X 与消费支出 Y 会呈现为不确定的相关关系。居民消费支出 X 之所以与居民可支配收入 Y 不呈现为确定的函数关系，是因为除了居民可支配收入 X 以外，还存在许多其他的因素也会影响居民消费支出 Y 。

变量之间的相关关系可用坐标图又称散点图去描述，例如，变量 X 和 Y 之间关系的散点图可描述为图 2.1：

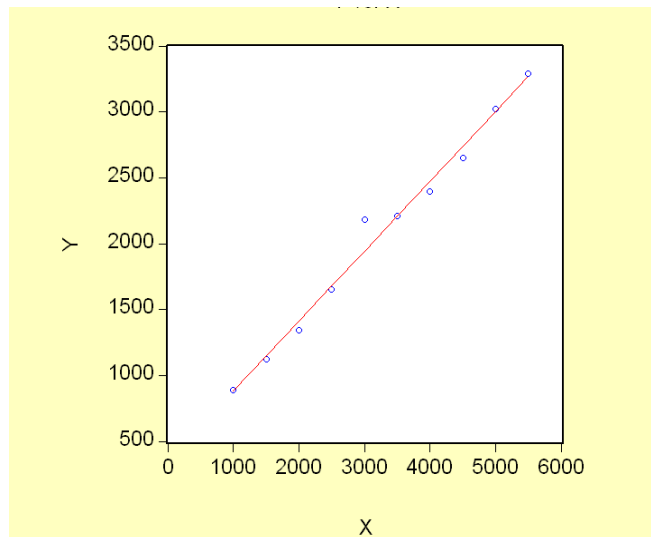


图 2.1 变量相关关系的散点图

由于涉及的变量数量、相关性质及相关程度的不同，变量之间的相关关系可以分为若干类型：

1、从相关关系涉及的变量数量看

只有两个变量的相关关系，称为简单相关关系。例如，人的身高与体重之间的相关关系。三个或三个以上变量的相关关系，称为多重相关或复相关。例如，某种商品的需求量与商品的价格及居民的收入水平之间的相关关系。

2、从变量相关关系的表现形式看

当变量之间相关关系的散点图中的点接近一条直线时，称为线性相关，当变量之间相关关系散点图中的点接近于一条曲线时，称为非线性相关。

3、从变量相关关系变化的方向看

两个变量趋于在同一个方向变化时，即同增或同减，称为变量之间存在正相关。当两个变量趋于在相反方向变化时，即当一个变量增加时，另一个变量却减少，称为变量之间存在负相关。

4、从变量相关的程度看

当一个变量的变化完全由另一个变量的变化所确定时，称为变量之间完全相关。例如前面所举的当价格不变的条件下，某种商品销售额与销售量之间的关系，在这种情况下，相关关系实际成为了函数关系，所以可以把函数关系视为相关关系的特例。

当两个变量的变化相互完全没有关系，即彼此互不影响时，称为二者不相关。两个现象的关系如果介于完全相关和不相关之间时，称为不完全相关，我们研究的相关关系通常都是指的这种不完全的相关关系。

(二) 简单线性相关关系的度量与检验

1、简单线性相关系数

在各种类型的相关分析中, 只有两个变量的线性相关关系的分析是最简单的。两个变量之间线性相关程度可以用简单线性相关系数去度量, 这种相关系数是最常用的, 也简称为相关系数。对于我们所研究的总体, 两个相互联系的变量的相关系数称为总体相关系数, 通常用 ρ 表示, 总体相关系数 ρ 可用式 (2. 1) 计算:

$$\rho = \frac{Cov(X,Y)}{Var(X)Var(Y)} \quad (2.1)$$

其中: $Var(X)$ 是变量 X 的方差 ; $Var(Y)$ 是变量 Y 的方差;

$Cov(X, Y)$ 是变量 X 和 Y 的协方差

总体相关系数 ρ 反映了总体两个变量 X 和 Y 的线性相关程度, 对于特定的总体来说, X 和 Y 的数值是既定的, 总体相关系数 ρ 是客观存在的特定数值。然而, 当总体较大时, 变量 X 和 Y 的全部数值一般不可能去直接观测, 所以总体相关系数一般是不能直接计算的未知量。通常可能做到的是从总体中随机抽取一定数量的样本, 通过 X 和 Y 的样本观测值 X_i 和 Y_i 去估计样本相关系数, 变量 X 和 Y 的样本相关系数通常用 r_{XY} 表示, 或简记为 r , 可用式 (2. 2) 或式 (2. 3) 去估计:

$$r_{XY} = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} \quad (2.2)$$

或

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (2.3)$$

其中: X_i 和 Y_i 分别是变量 X 和 Y 的样本观测值

\bar{X} 和 \bar{Y} 分别是变量 X 和 Y 样本观测值的平均值

n 是样本的个数, 也称样本容量

样本相关系数 r 是根据从总体中抽取的随机样本的观测值 X_i 和 Y_i 计算出来的, 它是对总体相关系数 ρ 的估计。可以证明, 这样计算的样本相关系数是总体相关系数的一致估计。

多个变量之间的线性相关程度, 则需要用复相关系数和偏相关系数去度量。

2、相关系数的特点:

由 (2.1) 和 (2.2) 式可看出, 相关系数有以下特点:

(1) 相关系数的取值在 -1 与 1 之间。

(2) 当 $r=0$ 时, 表明 X 与 Y 没有线性相关关系。

(3) 当 $0 < |r| < 1$ 时, 表明 X 与 Y 存在一定的线性相关关系, 若 $r > 0$ 表明 X 与 Y 为正相关, 若 $r < 0$ 表明 X 与 Y 为负相关。

(4) 当 $|r|=1$ 时, 表明 X 与 Y 完全线性相关, 若 $r=1$, 称 X 与 Y 完全正相关; 若 $r=-1$, 称 X 与 Y 完全负相关。

使用相关系数分析相关关系时应当注意:

(1) X 和 Y 都是相互对称的随机变量, 所以 $\gamma_{XY} = \gamma_{YX}$ 。

(2) 相关系数只反映变量间的线性相关程度, 不能说明非线性相关关系。

(3) 相关系数只能反映变量间线性相关的程度, 并不能确定变量的因果关系, 也不能说明相关关系具体接近于哪条直线。

(4) 样本相关系数是根据从总体中抽取的随机样本的观测值 X 和 Y 计算出来的, 它只是对总体相关系数 ρ 的估计。由于从总体中每抽取一个样本, 都可以根据其观测值估计出一个样本相关系数, 因此样本相关系数不是确定的值, 而是随抽样而变动的随机变量。对相关系数的统计显著性还有待进行检验¹。

(三) 回归分析

研究变量相互之间的相关关系时, 首先需要分析它们是否存在相关关系, 然后要明确其相关关系的类型, 而且还应计量其相关关系的密切程度, 在统计学中这种研究称为相关分析。相关分析主要是用一个指标 (相关系数) 去表明现象间相互依存关系的性质和密切程度。不过相关分析并不能说明变量间相关关系的具体形式, 也还不能从一个变量的变化去推测另一个变量的具体变化。如果要具体测定变量之间相关关系的数量形式, 还需要运用回归分析的方法。

“回归”这个词是由英国生物学家高尔顿在遗传学研究中首先提出来的。高尔顿发现相对于一定身高的父母, 子女的平均身高有朝向人类平均身高移动或回归的趋势。这就是“回归”的古典意义。

现在我们沿用“回归”这个词, 但其意义与回归的古典意义已有很大区别。现代意义的回归是关于一个变量 (被解释变量或应变量) 对另一个或多个变量 (解释变量) 依存关系的

¹ 对相关系数显著性检验的方法可参考有关《统计学》著作。

研究，用适当的数学模型去近似地表达或估计变量之间的平均变化关系，其目的是要根据已知的或固定的解释变量的数值，去估计所研究的被解释变量的总体平均值。

例如，研究个人消费支出与个人可支配收入的依存关系，对应于一定的个人可支配收入水平，个人消费支出并不确定，但总是在一定的范围内变动。对于每一个个人可支配收入水平，个人消费支出呈现出一定的分布，但平均来说，个人消费支出总是随着个人可支配收入的增加而增加的，其关系可见图 2。

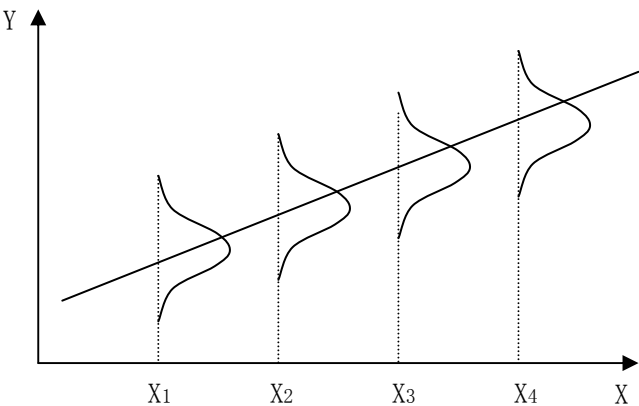


图 2.2 个人消费支出随个人可支配收入的变化

如果根据个人消费支出和个人可支配收入的观测数据，要去确定二者变动的统计规律性，也就要研究当解释变量个人可支配收入变动时，被解释变量个人消费支出的平均水平变动的规律，解决这样一类问题的方法是回归分析。

在理解回归分析时，应当注意回归所要揭示的是被解释变量与解释变量之间的平均关系。在这里，被解释变量是随机变量，解释变量在本质上可以是随机变量，但在回归分析中解释变量作为被解释变量变动的原因，我们总是假定在重复抽样中是取某些固定的值，所以在一般情况下解释变量是作为非随机变量来处理的。

显然，相关分析与回归分析有密切的联系，他们都是对变量间相关关系的研究，二者可以相互补充。相关分析可以表明变量间相关关系的性质和程度，只有当变量间存在一定程度的相关关系时，进行回归分析去寻求相关的具体数学形式才有实际的意义。同时，在进行相关分析时如果要具体确定变量间相关的具体数学形式，又要依赖于回归分析，而且相关分析中相关系数的确定也是建立在回归分析基础上的。

同时应当明确，相关分析与回归分析的研究目的和研究方法是有明显区别的。从研究目的上看，相关分析是用一定的数量指标（相关系数）度量变量间相互联系的方向和程度；回归分析却是要寻求变量间联系的具体数学形式，是要根据自变量的固定值去估计和预测被解释变量的平均值。从对变量的处理看，相关分析对称地对待相互联系的变量，不考虑二者的

因果关系，也就是不区分解释变量和被解释变量，相关的变量不一定具有因果关系，均视为随机变量；回归分析是建立在变量因果关系分析的基础上，研究其中解释变量的变动对被解释变量的具体影响，回归分析中必须明确划分被解释变量和解释变量，对变量的处理是不对称的。在回归分析中通常假定解释变量在重复抽样中是取固定值的非随机变量，只有被解释变量是具有一定概率分布的随机变量。

还应当强调，相关分析和回归分析只是从数据出发定量地分析经济变量间相互联系的手段，并不能决定经济现象相互之间的本质联系。经济现象间内在的本质联系，决定于它们的客观规律性，需要结合实际经验去分析，并要由经济学理论去加以说明。如果对本来没有内在联系的经济现象，仅凭数据进行相关分析和回归分析，有可能会是一种“伪相关”或“伪回归”，这样不仅没有实际的意义，而且会导致荒谬的结论。所以在对经济问题开展相关分析和回归分析时，要注意与定性的经济分析相结合，才能得到有实际意义的结果。

二、总体回归函数（PRF）

1、回归线与回归函数

回归分析研究的是总体中解释变量与被解释变量之间客观存在的协变规律性，在经济现象的研究中，这种协变规律是所研究的经济总体的特征。怎样去认识经济总体中相关变量的协变关系呢？由于实际的经济总体通常难以直接观测，这里以一个简化的例子去说明。

【例 2.1】假如有一个由 100 个家庭构成的总体，我们要研究的是每月家庭消费支出 Y 与每月家庭可支配收入 X 之间的关系，并要根据已知的家庭可支配收入水平去预测该总体每月家庭消费支出的平均水平。为了研究的方便，把总体 100 个家庭按收入水平分为 10 个组，分别考察各组中每个家庭的消费支出（见表 2.1）

表 2.1 家庭消费支出与家庭可支配收入 单位：元

	每 月 家 庭 可 支 配 收 入 X									
	1000	1500	2000	2500	3000	3500	4000	4500	5000	5500
每 月 家 庭 消 费 支	820	962	1108	1329	1632	1842	2037	2275	2464	2824
	888	1024	1201	1365	1726	1874	2110	2388	2589	3038
	932	1121	1264	1410	1786	1906	2225	2426	2790	3150
	960	1210	1310	1432	1835	1068	2319	2488	2856	3201
		1259	1340	1520	1885	2066	2321	2587	2900	3288
		1324	1400	1615	1943	2185	2365	2650	3021	3399
			1448	1650	2037	2210	2398	2789	3064	
			1489	1712	2078	2289	2487	2853	3142	
			1538	1778	2179	2313	2513	2934	3274	
			1600	1841	2298	2398	2538	3110		

出 Y			1702	1886	2316	2423	2567			
				1900	2387	2453	2610			
				2012	2498	2487	2710			
					2589	2586				
$E(Y X_i)$	900	1150	1400	1650	1900	2150	2400	2650	2900	3150

由表 2.1 可以看出, 由于解释变量可支配收入 X 与被解释变量消费支出 Y 之间不是确定性的函数关系而是不确定性的相关关系, 对于可支配收入 X 的每一个固定水平, 家庭消费支出 Y 并不确定。在给定家庭可支配收入 X 的条件下, 家庭消费支出 Y 形成一定的分布, 这种分布称为在 X 取某一特定值时 Y 的条件分布。当 X 取某一特定值时, Y 取各种值的概率, 称为 Y 的条件概率。例如当家庭可支配收入为 2500 元时, 家庭消费支出为 1712 元的条件概率为: $P(Y=1712 | X=2500) = 1/13$, 等等。对于 X 的每一个取值 X_i , 根据 Y 的条件分布和条件概率, 可以计算出 Y 的条件期望或称条件均值 $E(Y|X_i)$, 所计算的条件均值列于表 2.1 的最后一行。

对于 X 的每一个取值 X_i , 都有 Y 的条件期望 $E(Y|X_i)$ 与之对应, 根据表 2.1 的数据, 可作家庭可支配收入 X 与家庭消费支出 Y 的散点图, 如图 2.3 所示:

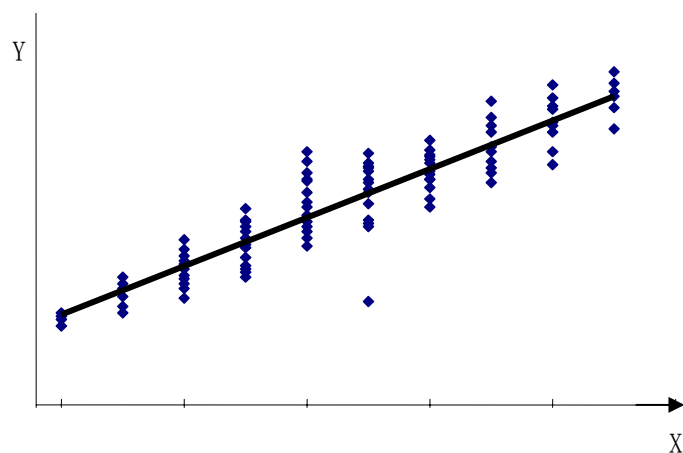


图 2.3

由表 2.1 和图 2.3 可以看出, 虽然每个家庭的消费支出存在差异, 但平均来说, 家庭消费支出是随家庭可支配收入的递增而递增的。还可以看出当 X_i 取各种值时, Y 的条件均值 (图 2.3 中的黑点) 的轨迹接近一条直线, 该直线称为 Y 对 X 的回归直线。当然, 若是 Y 的各个条件均值是位于一条曲线上, 则这条曲线就称为 Y 对 X 的回归曲线。

从上述 100 个家庭构成的总体的例子可以看出, 所研究的总体被解释变量家庭消费支出

Y 的条件均值 $E(Y|X_i)$ ，是随解释变量 X 的变化而有规律的变化，如果把 Y 的条件均值表示为 X 的某种函数，可写为：

$$E(Y|X_i) = f(X_i) \quad (2.5)$$

如 (2.5) 式那样，将总体被解释变量 Y 的条件均值表现为解释变量 X 的函数，这个函数称为总体回归函数 (Population Regression Function, 简记为 **PRF**)。这是总体回归函数的条件期望表示方式。

假如 Y 的总体条件均值 $E(Y|X_i)$ 是解释变量 X 的线性函数，可表示为：

$$E(Y|X_i) = f(X_i) = \beta_1 + \beta_2 X_i \quad (2.6)$$

其中 β_1 和 β_2 是未知的参数， β_1 称为截距系数， β_2 称为斜率系数。

需要指出，在实际的经济问题中，总体包含的单位数通常相当多，一般情况下不大可能像假定的 100 个家庭那样去取得总体所有的观测值，也不大可能直接计算 Y 的条件期望值。所以事实上总体回归函数的具体形式，只能根据经济理论对所研究经济问题的深刻认识以及实践经验去设定，也就是说需要对总体回归函数作出合理的假设。

在计量经济学中经常把总体回归函数设定为线性函数，这是因为线性函数是最简单的函数形式，而且线性回归函数中参数的估计与检验相对容易，用线性模型去近似地描述总体回归函数，常能获得较好的效果。

在计量经济学中线性模型的“线性”有两种解释：一是模型就变量而言是线性的，即 Y 的条件均值 $E(Y|X_i)$ 是解释变量 X_i 的线性函数，这时回归线是一条直线。按这一原则，

$E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ 或 $E(Y|X_i) = \beta_1 + \beta_2 (1/X_i)$ 都不是线性回归函数。二是模型就参数而言是线性的，即 Y 的条件均值 $E(Y|X_i)$ 是参数 β 的线性函数，而对于解释变量 X_i 则可以是线性的，也可以是非线性的。按这一原则， $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ 或 $E(Y|X_i) = \beta_1 + \beta_2 (1/X_i)$ 都是线性回归模型，而 $E(Y|X_i) = \beta_1 + \sqrt{\beta_2} X_i$ 或 $E(Y|X_i) = \beta_1 + (1/\beta_2) X_i$ 则不是线性回归模型。在计量经济学中，从回归理论和参数的估计方法考虑，通常是就参数而言判断是否线性回归模型，而对解释变量 X_i 则或者是线性的或者不是线性回归模型。

三、随机扰动项 u

以条件均值表现的总体回归函数 $E(Y|X_i)$ 描述的是随着解释变量的变化被解释变量的平均变动。但是相对于一定的 X_i ， Y 的个别值 Y_i 并不全在代表平均值轨迹的回归线上，而是围绕回归线上下波动，也就是说个别值 Y_i 总是分布在条件均值 $E(Y|X_i)$ 的周围。若令各个 Y_i 值与条件均值 $E(Y|X_i)$ 的偏差为 u_i ，显然 u_i 是个可正可负的随机变量，称为随机扰动项或随机误差项。即

$$u_i = Y_i - E(Y|X_i) \quad (2.7)$$

或
$$Y_i = E(Y|X_i) + u_i \quad (2.8)$$

(2.8) 式是总体回归函数的个别值表示方式，或称随机设定形式。

如果总体回归函数是只有一个解释变量的线性函数，则有

$$u_i = Y_i - \beta_1 - \beta_2 X_i$$

或
$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (2.9)$$

(2.8) 式表明，除了已列入模型的解释变量 X 以外，还有影响被解释变量 Y 变动的其他因素，这里的随机扰动项 u_i 就代表着那些对 Y 有影响但又未纳入模型的诸多因素的综合影响。

从 (2.5) 式和 (2.8) 式可以看出，条件期望表示方式与个别值表示方式是等价的。因为若对 (2.8) 式两边取其对于 X_i 的条件期望，则有

$$\begin{aligned} E(Y|X_i) &= E\{E(Y|X_i)\} + E(u_i|X_i) \\ &= E(Y|X_i) + E(u_i|X_i) \end{aligned}$$

显然，这里暗含着 $E(u_i|X_i) = 0$ 的假定条件，也就是回归线是通过 Y 的条件期望或条件均值的。

在总体回归函数中引进随机扰动项，主要有以下几方面的原因：

(1) 作为未知影响因素的代表。由于对所研究的经济现象的变动规律的认识并不完备，除了一些已知的主要因素以外，还有一些未被认识或尚不能肯定的因素影响被解释变量，因此只得用随机扰动项作为被模型省略掉的未知因素的代表。

(2) 作为无法取得数据的已知因素的代表。有一些因素已经知道对被解释变量有相当的

影响，但可能无法获得这些变量的定量数据。例如，在研究家庭消费支出时，根据有关经济理论的分析，认为家庭财产的数量对家庭消费支出也有影响，可是在一般情况下取得家庭财产的数据有困难，在计量经济模型中不得不把家庭财产略去，而这类变量的影响被归入到随机扰动项。

(3) 作为众多细小影响因素的综合代表。某些影响因素已经被认识到，其数据也可能获得，例如影响家庭消费支出的还可能有子女人数、性别构成、民族习惯、受教育程度，等等，但是这些因素或许对被解释变量家庭消费支出的影响比较小，或许其影响不很规则、有的可能不易数量化，从经济计量的成本考虑，通常不把它们列入模型，而将它们的联合影响处理为随机扰动项。

(4) 模型的设定误差。在设定经济计量模型时，总是力图使模型更为简单明了，当用较少的解释变量就能说明被解释变量的实质变化时，就不应把更多的解释变量列入模型；当用较简洁的函数形式就能说明变量之间的本质联系时，就尽量不采用更为复杂的函数形式。这样，变量和函数形式的设定可能会引起设定误差，这种设定误差也要由随机扰动项来表示。

(5) 变量的观测误差。对社会经济现象观测所得到的统计数据，由于主客观的原因，可能地会有一定的观测误差，这种观测误差只有归入随机扰动项。

(6) 经济现象的内在随机性。即使把所有相关的影响因素全部纳入模型，即使不存在观测误差，但是人所从事的一些经济行为还是可能具有不可重复性和随机性。例如，某些涉及人们思想行为的变量，很难完全控制，而是具有内在的随机性，这种内在的随机性也可能影响人们的经济行为。这类变量变内在的随机性的影响只能归入随机扰动项。

由此可见，随机扰动项有十分丰富的内容，在计量经济研究中起着重要的作用。一定程度上，随机扰动项的性质决定着计量经济方法的选择和使用。

四、样本回归函数（SRF）

对于实际的经济问题，通常总体包含的单位数很多，无法掌握所有单位的数值，总体回归函数实际上是未知的。我们可能做到的只是对应于解释变量 X 的选定水平，对被解释变量 Y 的某些样本进行观测，然后通过对样本观测获得的信息去估计总体回归函数。

仍然以【例 2.1】中 100 个家庭的可支配收入与消费支出为例，假设从 100 个家庭的总体中各随机抽取 10 个家庭进行观测，形成了两个随机样本，如表 2.2 和表 2.3 所示：

表 2.2		随机样本（一）								单位：元	
可支配收入 X	1000	1500	2000	2500	3000	3500	4000	4500	5000	5500	

消费支出 Y	888	1121	1340	1650	2179	2210	2398	2650	3021	3288
--------	-----	------	------	------	------	------	------	------	------	------

表 2.3

随机样本（二）

单位：元

可支配收入 X	1000	1500	2000	2500	3000	3500	4000	44500	5000	5500
消费支出 Y	932	1259	1448	1651	2298	2289	2365	2488	2856	3150

可将两个随机样本的数据绘制成散点图，其示意图见图 2.4。

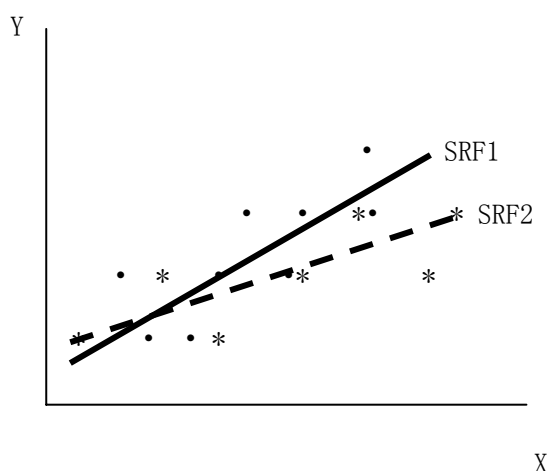


图 2.4 两个随机样本的示意图

从总体中抽取一定样本数据进行观测，对于解释变量 X 的一定值，取得的被解释变量 Y 的样本观测值也可计算其条件均值， Y 的样本观测值的条件均值随解释变量 X 而变动的轨迹，称为样本回归线。图 2.4 中所示的两条直线就是分别通过两个样本观测值拟合的最适合的样本回归直线。当然如果所拟合的是曲线，就称为样本回归曲线。

从图 2.4 可以看出，被解释变量（消费支出）的样本条件均值也是随解释变量（可支配收入）的变化而有规律的变化。如果把被解释变量 Y 的样本条件均值表示为解释变量 X 的某种函数，这个函数称为样本回归函数（Sample Regression Function，简记为 **SRF**）。显然，样本回归函数的函数形式应与设定的总体回归函数的函数形式一致。样本回归函数如为线性函数，可表示为：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad (2.10)$$

其中的 \hat{Y}_i 是回归线上与 X_i 相对应的 Y 的样本条件均值，可视为对总体条件期望 $E(Y|X_i)$ 的估计值； $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 分别是样本回归函数的截距系数和斜率系数，可视为对总体回归函数中系数 β_1 和 β_2 的估计。

与总体回归函数相类似，实际观测的被解释变量值 Y_i 并不完全等于样本条件均值 \hat{Y}_i ，二者之差可用 e_i 表示，那么

$$Y_i - \hat{Y}_i = e_i \quad (2.11)$$

或者
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad (2.12)$$

(2.12) 式是样本回归函数的另一种表达形式，与 (2.10) 式是等价的，这里的 e_i 称为剩余项，或称为残差， e_i 在概念上类似于总体扰动项 u_i 。在样本回归函数中引入 e_i 的原因，与将 u_i 引入总体回归函数的理由是相同的。

必须明确，样本回归函数与总体回归函数是有区别的。首先，总体回归函数虽然未知，但它是确定的；而由于从总体中每次抽样都能获得一个样本，就都可以拟合一条样本回归线，所以样本回归线却是随抽样波动而变化的，可以有許多条。所以，样本回归线还不是总体回归线，至多只是未知的总体回归线的近似反映。其次，总体回归函数的参数 β_1 和 β_2 是确定的常数；而样本回归函数的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是随抽样而变化的随机变量。此外，总体回归函数中的 u_i 是不可直接观测的；而样本回归函数中的 e_i 是只要估计出样本回归的参数就可以计算的数值。

表示总体经济活动规律的总体回归函数是未知的，在计量经济学中进行回归分析的目的，就是要根据有可能获得的样本回归函数去对总体回归函数作出合理的估计。然而，样本毕竟不等于总体，样本回归函数 SRF 几乎总是与总体回归 PRF 存在着差异，二者的关系可以从图 2.5 中看出：

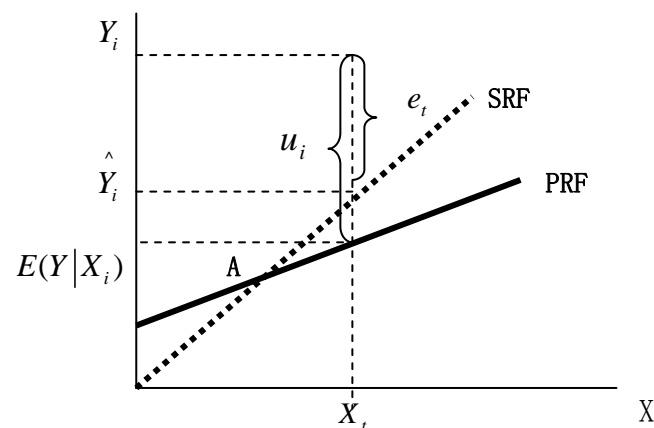


图 2.5 总体回归函数与样本回归函数的关系

回归分析的目的是要用样本回归函数去尽可能准确的估计总体回归函数。由于样本对总体存在代表性误差，SRF 又总会过高或过低估计 PRF，例如图 2.5 中 A 点左边部分就过高估计了 PRF，A 点右边部分又过低估计了 PRF。显然，需要寻求一种规则和方法，使得到的样本回归函数的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 能够“尽可能地接近”总体回归函数中的参数 β_1 和 β_2 。这就是计量经济学应当解决的参数估计的基本问题。

第二节 简单线性回归模型参数的估计

估计线性回归模型中参数的方法有若干种，这些估计方法都是以对模型的某些假定条件为前提的。因为只有具备这些假定条件，所作出的估计才具有良好的统计性质。所以，这些假设与所采用的估计方法是紧密相关的。

一、简单线性回归的基本假定

对简单线性回归模型的基本假定有两个方面：一是对变量和模型的假定；二是对随机扰动项 u_i 统计分布的假定。

在简单线性回归模型中对变量和模型的假定，首先是假定解释变量 X_i 是确定性变量，是非随机的，这是因为在重复抽样中 X_i 是取一组固定的值。或者 X_i 虽然是随机的，但与随机扰动项 u_i 也是不相关的。其次，是假定模型中的变量没有测量误差。此外，还要假设模型对变量和函数形式的设定是正确的，即不存在设定误差。

为了使对模型的估计具有良好的统计性质，在计量经济研究中对无法直接观测的随机扰动项 u_i 的分布，需要作如下一些基本假定：

假定 1：零均值假定 即在给定解释变量 X_i 的条件下，随机扰动项 u_i 的条件均值为零，即

$$E(u_i | X_i) = 0 \quad (2.13)$$

假定 2：同方差假定 即对于给定的每一个 X_i ，随机扰动项 u_i 的条件方差都等于某一个常数 σ^2 ，即

$$Var(u_i | X_i) = E[u_i - E(u_i | X_i)]^2 = E(u_i^2) = \sigma^2 \quad (2.14)$$

式中的 Var 表示方差。

假定 3：无自相关假定 即随机扰动项 u_i 的逐次值互不相关，或者说对于所有的 i 和 j ($i \neq j$)， u_i 和 u_j 的协方差为零，即

$$\begin{aligned} Cov(u_i, u_j) &= E[u_i - E(u_i)][u_j - E(u_j)] \\ &= E(u_i u_j) = 0 \end{aligned} \quad (2.15)$$

式中 Cov 表示协方差。

假定 4：随机扰动项 u_i 与解释变量 X_i 不相关 可表示为

$$Cov(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)] = 0 \quad (2.16)$$

这一假定表明模型中的 X_i 和 u_i 是各自独立影响 Y_i 的，这样才能分清楚解释变量 X_i 与随机扰动项 u_i 分别对 Y_i 的影响各为多少。

假定 5：正态性假定 即假定随机扰动项 u_i 服从期望为零，方差为 σ^2 的正态分布，表示为

$$u_i \sim N(0, \sigma^2) \quad (2.17)$$

以上这些对随机扰动项 u_i 分布的假定是德国数学家高斯最早提出的，也称为高斯假定或古典假定。满足以上古典假定的线性回归模型，也称为古典线性回归模型 (Classical Linear Regression Model, 简称 CLRM)。

顺便指出，由于 $Y_i = \beta_1 + \beta_2 X_i + u_i$ ， Y_i 的分布性质决定于 u_i ，为此，对 u_i 的零均值、同方差、无自相关及正态性假定也可以用对 Y_i 的假定来表示：

$$\text{假定 1: } E(Y_i | X_i) = \beta_1 + \beta_2 X_i \quad (2.18)$$

$$\text{假定 2: } Var(Y_i | X_i) = \sigma^2 \quad (2.19)$$

$$\text{假定 3: } Cov(Y_i, Y_j) = 0 \quad (i \neq j) \quad (2.20)$$

$$\text{假定 4: } Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2) \quad (2.21)$$

容易证明，以上对 Y_i 分布性质的假定与对随机扰动项 u_i 分布的古典假定是等价的。

二、普通最小二乘法

计量经济研究的直接目的是确定总体回归函数 $Y_i = \beta_1 + \beta_2 X_i + u_i$ ，然而能够得到的只是来自总体的若干样本的观测值，要用样本信息建立的样本回归函数尽可能“接近”地去估计总体回归函数。为此，可以从不同的角度去确定建立样本回归函数的准则，也就有了估计回归模型参数的多种方法。例如用产生该样本概率最大的原则去确定样本回归函数，称为极大似然准则；用使估计的剩余平方和最小的原则确定样本回归函数，称为最小二乘准则。本章只介绍在古典假定下的最小二乘法，也称为普通最小二乘估计（Ordinary Least Squares Estimators，简记为 OLS 或者 OLSE）。

为了使样本回归函数尽可能“接近”总体回归函数，就是要由样本回归函数 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ 估计的 \hat{Y}_i 与实际的 Y_i 的误差尽量小，即是要使剩余项 e_i 越小越好。可是作为误差 e_i 有正有负，其简单代数和 $\sum e_i$ 会相互抵消而趋于零。为使在数学上便于处理，可采用剩余平方和 $\sum e_i^2$ 最小的准则，这就是最小二乘准则，即

$$\min \sum e_i^2 = \min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

根据微积分中求极值的原理，要使 $\sum e_i^2$ 达到最小，待定系数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 应满足以下条件

$$\frac{\partial(\sum e_i^2)}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \quad (2.22)$$

$$\frac{\partial(\sum e_i^2)}{\partial \hat{\beta}_2} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = 0 \quad (2.23)$$

从而得如下方程组

$$\sum Y_i = n \hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad (2.24)$$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \quad (2.25)$$

其中的 n 为样本容量。(2.24)和(2.25)的方程组称为最小二乘的正规方程组。根据克莱姆法则解正规方程组，得

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \quad (2.26)$$

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \quad (2.27)$$

(2.26)和(2.27)式即是用样本观测值 X_i 和 Y_i 表现的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的最小二乘估计式。

如果令 $x_i = X_i - \bar{X}$ ， $y_i = Y_i - \bar{Y}$ ， x_i 和 y_i 分别称为 X_i 和 Y_i 的离差形式²。容易证明，(2.26)和(2.27)式可用离差形式表示为

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{2.28}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \tag{2.29}$$

（2.26）和（2.27）式或者(2.28)和(2.29)式是根据最小二乘准则推导出来的，称为简单线性回归模型参数的最小二乘估计式，由这些估计式可直接用样本观测值求得参数的估计值。

【例 2.2】：为了估计【例 2.1】的样本回归函数，通过表 2.2 样本(一)的数据可计算相关的数据，如表 2.4 所示。

表 2.4 样本(一)相关数据 单位:元

序 号	可支配 收入 X_i	消费 支出 Y_i	$x_i =$ $X_i - \bar{X}$	$y_i =$ $Y_i - \bar{Y}$	$x_i y_i$	x_i^2	y_i^2	\hat{Y}_i	$e_i =$ $Y_i - \hat{Y}_i$	e_i^2
1	1000	888	-2250	-1186.5	2669625	5062500	1407782.25	882	6	36
2	1500	1121	-1750	-953.5	1668625	3062500	909162.25	1147	-26	676
3	2000	1340	-1250	-734.5	918125	1562500	539490.25	1412	-72	5184
4	2500	1650	-750	-424.5	318375	562500	180200.25	1677	-27	729
5	3000	2179	-250	104.5	-26125	62500	10920.25	1942	237	56169
6	3500	2210	250	135.5	33875	62500	18360.25	2207	3	9
7	4000	2398	750	323.5	242625	562500	104652.25	2472	-74	5476
8	4500	2650	1250	575.5	719375	1562500	331200.25	2737	-87	7569
9	5000	3021	1750	946.5	1656375	3062500	895862.25	3002	19	361
10	5500	3288	2250	1213.5	2730375	5062500	1472582.25	3267	21	441
合计	32500	20745			10931250	20625000	5870212.5			76650
平均	3250	2074.5								

将有关数据代入(2.28)和(2.29)式，得

² 为了使表现形式尽量简洁，本书中一律用大写字母 X_i 、 Y_i 等表示观测值，用小写字母 x_i 、 y_i 等表示观测值的离差。

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{10931250}{20625000} = 0.5300$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 2074.5 - 0.53 \times 3250 = 352$$

即样本回归函数为 $\hat{Y}_i = 352 + 0.53X_i$

三、OLS 回归线的性质

用普通最小二乘法拟合的样本回归线有以下性质：

1、回归线通过样本均值

由 (2.29) 式可得 $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$ ，所以样本回归线必然通过 (\bar{X}, \bar{Y}) ，如图 2.6 所示：

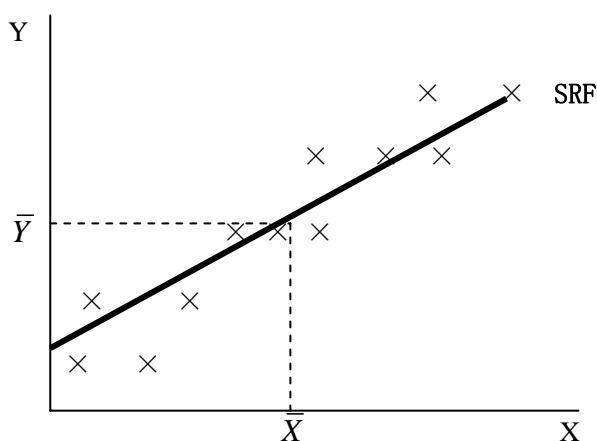


图 2.6 样本回归线通过样本均值

2、估计值 \hat{Y}_i 的均值等于实际值 Y_i 的均值

因为

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

将上式两边加总，再除以样本容量 n

$$\begin{aligned}\frac{\sum \hat{Y}_i}{n} &= \frac{\sum [\bar{Y} + \hat{\beta}_2 (X_i - \bar{X})]}{n} \\ &= \bar{Y} + \frac{\hat{\beta}_2}{n} \sum (X_i - \bar{X}) \quad [\text{其中 } \sum (X_i - \bar{X}) = 0]\end{aligned}$$

则
$$\frac{\sum \hat{Y}_i}{n} = \bar{Y} \quad (2.30)$$

3、剩余项 e_i 的均值为零

由最小二乘准则已知

$$\frac{\partial(\sum e_i^2)}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \quad (\text{见 2.22})$$

而且

$$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

所以

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

可知

$$\bar{e} = \frac{\sum e_i}{n} = 0 \quad (2.31)$$

4、被解释变量估计值 \hat{Y}_i 与剩余项 e_i 不相关，即 $Cov(\hat{Y}_i, e_i) = 0$

这是因为

$$\begin{aligned} Cov(\hat{Y}_i, e_i) &= E\{[\hat{Y}_i - E(\hat{Y}_i)][e_i - E(e_i)]\} \\ &= E[(\hat{Y}_i - \bar{Y})e_i] \\ &= E(y_i e_i) \\ &= \frac{\sum (y_i e_i)}{n} \end{aligned}$$

式中

$$\begin{aligned} \sum (y_i e_i) &= \sum y_i (y_i - \hat{\beta}_2 x_i) \\ &= \sum (\hat{\beta}_2 x_i)(y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \end{aligned}$$

则

$$\sum (y_i e_i) = 0 \quad (2.32)$$

5、解释变量 X_i 与剩余项 e_i 不相关

因为

$$Cov(X_i, e_i) = \frac{1}{n} \sum (e_i - \bar{e})(X_i - \bar{X})$$

$$= \frac{1}{n} \sum e_i X_i$$

$$\text{由 OLS 正规方程} \quad \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = \sum e_i X_i = 0$$

$$\text{所以} \quad \text{Cov}(X_i, e_i) = 0 \quad (2.33)$$

普通最小二乘法估计的回归线所具有的以上性质，在计量经济方法的估计、检验以及一些结论的证明中都会用到，具有重要的意义。

四、最小二乘估计式的统计性质

1、参数估计式的评价标准

计量经济模型中的参数一般是未知的，需要根据样本信息去加以估计。所估计的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 等都是样本数据的函数。由于取得的样本不同，样本数据也不同，即存在“抽样波动”，因此参数的估计值是随抽样而变化的随机变量，加之估计方法和假设前提不一定那么完备，用样本估计的参数数值不一定等于总体参数的真实值。那么，在比较不同估计方法的估计结果时，需要有一定的评价标准，这个标准就是应使参数估计值“尽可能地接近”总体参数的真实值。作为一个随机变量，参数估计值怎样才算“尽可能地接近”总体参数的真实值呢？理论计量经济学一般并不直接分析参数估计值本身，而是研究所运用的参数估计式是否符合一定的标准。通常选择参数的估计式时主要应考虑以下一些标准：

(1) 无偏性

如果已经确定了模型参数 β 的估计式 $\hat{\beta}$ ，在重复抽样中可取得变量一系列的观测值，将这些样本观测值代入参数估计式 $\hat{\beta}$ ，可得到参数的一系列估计值，这些参数估计值的分布称为 $\hat{\beta}$ 的抽样分布，其密度函数记为 $f(\hat{\beta})$ 。假如用另外一种方式确定了模型参数 β 的另一个估计式 β^* ，其抽样分布的密度函数记为 $f(\beta^*)$ 。

如果参数的估计式 $\hat{\beta}$ 的期望等于参数的真实值 β ，即 $E(\hat{\beta}) = \beta$ ，则称 $\hat{\beta}$ 是参数 β 的无偏估计式。如果参数估计式 β^* 的期望值不等于参数 β 的真实值，则称 β^* 是有偏的，其偏差为 $E(\beta^*) - \beta$ 。其关系如图 2.7 所示：

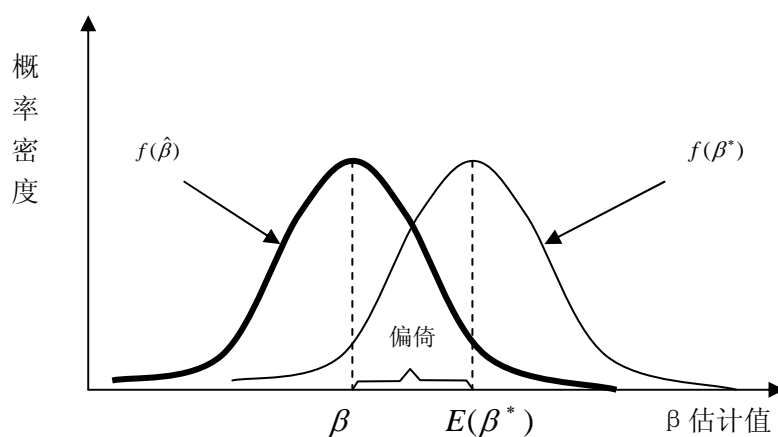


图 2.7 无偏估计与有偏估计

显然，在计量经济研究中应尽可能寻求符合无偏性要求的参数估计式。

(2)最小方差性

在计量经济研究中，通常用不同的方法可以获得若干不同的参数估计式，这些估计式抽样分布的方差也可能不同。如果对于参数 β 的任意一个估计式 $\beta^\#$ ，都有 $Var(\hat{\beta}) \leq Var(\beta^\#)$ ，则称 $\hat{\beta}$ 是参数 β 的最小方差估计式，或最佳估计式。显然，应当尽可能选择其抽样分布具有最小方差的参数估计式，这一原则即最小方差性，或称最佳性。如图 2.8 所示， $\hat{\beta}$ 和 β^* 的期望值相同，都是 β ，但若 $\hat{\beta}$ 的方差最小，则 $\hat{\beta}$ 是最佳估计式。

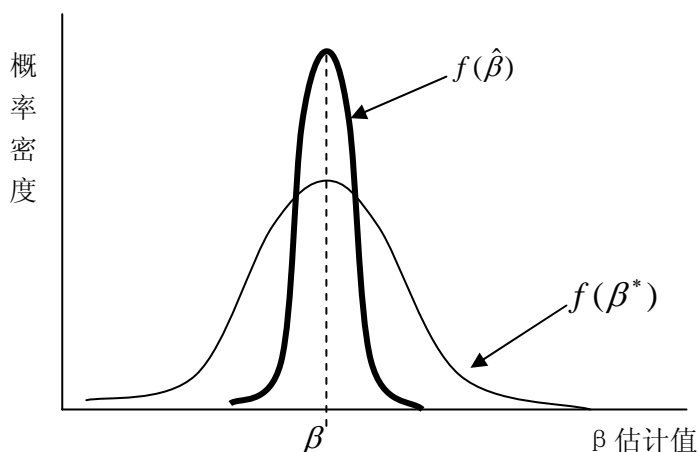


图 2.8 最佳估计式

(3)有效性

一个估计式若不仅具有无偏性而且具有最小方差性时，称这个估计式为有效估计式。无偏估计式可能有多，但在所有无偏估计式中，只有方差最小的最佳无偏估计式才是有效估计式。

(4)一致性

在样本容量较小的情况下,有时很难找到最佳无偏估计式,这时还需要考虑当样本容量充分大或趋于无穷大时估计式的渐近性质。

当样本容量趋于无穷大时,估计式 $\hat{\beta}$ 的抽样分布依概率收敛于总体参数的真实值 β ,即

$$P \lim_{n \rightarrow \infty} \hat{\beta} = \beta \quad \text{或} \quad \lim_{n \rightarrow \infty} P(|\hat{\beta} - \beta| < \varepsilon) = 1 \quad (2.34)$$

也就是说当 $n \rightarrow \infty$ 时,估计式 $\hat{\beta}$ 与总体参数真实值 β 的距离 $\hat{\beta} - \beta$ 的绝对值小于任意给定的正数 ε 的概率等于1,则称估计式 $\hat{\beta}$ 为一致估计式。

2、OLS 估计式的统计特性

可以证明,在古典假定完全满足的条件下,回归模型参数的最小二乘估计式具有以下统计性质:

(1) 无偏性

对于简单线性回归模型的最小二乘估计,因为

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum x_i^2} \\ &= \sum \frac{x_i}{\sum x_i^2} Y_i \end{aligned}$$

若令 $\frac{x_i}{\sum x_i^2} = k_i$,在重复抽样中, X_i 取一组固定的值, k_i 是一组常数。且 k_i 具有

$\sum k_i = 0, \sum k_i X_i = 1$ 的性质。则有

$$\hat{\beta}_2 = \sum k_i Y_i \quad (2.35)$$

由(2.35)式

$$\begin{aligned} \hat{\beta}_2 &= \sum k_i Y_i \\ &= \sum k_i (\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i u_i \\ &= \beta_2 + \sum k_i u_i \end{aligned} \quad (2.36)$$

则有

$$\begin{aligned}
E(\hat{\beta}_2) &= E(\beta_2) + E(\sum k_i u_i) \\
&= \beta_2 + \sum k_i E(u_i)
\end{aligned} \tag{2.37}$$

由古典假定 $E(u_i) = 0$ ，所以

$$E(\hat{\beta}_2) = \beta_2 \tag{2.38}$$

此外

$$\begin{aligned}
E(\hat{\beta}_1) &= E(\bar{Y} - \hat{\beta}_2 \bar{X}) \\
&= E(\bar{Y}) - \bar{X} E(\hat{\beta}_2) \\
&= (\beta_1 + \beta_2 \bar{X}) - \bar{X} \beta_2
\end{aligned}$$

则有

$$E(\hat{\beta}_1) = \beta_1 \tag{2.39}$$

这表明普通最小二乘法估计的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的期望值等于总体回归函数参数真实值 β_1 和 β_2 ，所以 OLS 估计式是无偏估计式。

(2) 最小方差性

为了说明 OLS 估计式的方差特性，必须导出 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 方差的公式：

$$\begin{aligned}
Var(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2 \\
&= E(\hat{\beta}_2 - \beta_2)^2
\end{aligned}$$

由 (2.36) 式

$$\begin{aligned}
Var(\hat{\beta}_2) &= E(\beta_2 + \sum k_i u_i - \beta_2)^2 \\
&= E(\sum k_i u_i)^2 \\
&= \sigma^2 \sum k_i^2 \\
&= \frac{\sigma^2}{\sum x_i^2}
\end{aligned} \tag{2.40}$$

类似地可以导出

$$Var(\hat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{n \sum x_i^2} \tag{2.41}$$

可以证明，在总体回归函数参数 β_1 和 β_2 的所有无偏估计量中，普通最小二乘估计 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 具有最小的方差，其证明过程较为繁琐，可见本章附录 2.1。

在计量经济学中还常用标准误差去度量估计量的精确性，标准误差是方差的平方根，可

用 SE (Standard error) 表示, 所以

$$SE(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \quad (2.42)$$

$$SE(\hat{\beta}_1) = \sigma \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}} \quad (2.43)$$

在 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差和标准误差的计算公式中, 除了样本观测值以外, 都包含了随机扰动项 u_i 的方差 σ^2 或 σ , 然而 σ^2 作为总体随机扰动项 u_i 的方差是未知的, 也需要通过样本去估计。可以证明, 在简单线性回归模型中, 用 (2.44) 式计算的 σ^2 的估计值 $\hat{\sigma}^2$ 是对 σ^2 的无偏估计 (证明过程见本章附录 2.2) :

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \quad (2.44)$$

式中: $\sum e_i^2$ 是剩余平方和; $n-2$ 是自由度。

例如, 在【例 2.2】中运用表 2.2 样本(一)的数据已估计出参数为 $\hat{\beta}_1 = 352$, $\hat{\beta}_2 = 0.53$ 。

由于总体方差 σ^2 未知, 可用表 2.4 中已算出的 $\sum e_i^2 = 76650$ 通过(2.44)式去估计, 即

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum e_i^2}{n-2} = \frac{76650}{10-2} = 9581.25 \\ \hat{\sigma} &= \sqrt{\hat{\sigma}^2} = \sqrt{9581.25} = 97.88386 \end{aligned}$$

(3) 线性特性

由(2.36)式直接看出, $\hat{\beta}_2 = \sum k_i Y_i$, 其中 k_i 是一组常数, 所以 $\hat{\beta}_2$ 是 Y_i 的线性函数。类似地,

$$\begin{aligned} \hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2 \bar{X} \\ &= \bar{Y} - \bar{X} \sum k_i Y_i \\ &= \sum \left(\frac{1}{n} - \bar{X} k_i \right) Y_i \end{aligned} \quad (2.45)$$

其中: n 、 \bar{X} 、 k_i 均为固定常数, 所以 $\hat{\beta}_1$ 也是 Y_i 的线性函数。

由以上的分析可以看出, 在古典假定条件下, OLS 估计式 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是总体参数 β_1 和 β_2

的最佳线性无偏估计式 (Best Linear Unbiased Estimator, 简记为 BLUE), 这一结论称为高斯—马尔可夫定理。OLS 估计式的这些特性具有很重要的意义。因为在其他条件不变的情况下, 线性估计量不仅比非线性估计量更为简单, 计算更为方便, 而且线性估计量比较容易确定其概率分布性质。此外 OLS 估计的最小方差特性和无偏特性结合起来, 使得按同样的置信度, OLS 估计量的置信区间最小, 最集中于真实值周围。所以寻求最佳线性无偏估计式是计量经济学努力的目标, 这也是 OLS 估计法能得到广泛应用的重要原因。

第三节 拟合优度的度量

样本回归线是对样本数据的一种拟合, 对于同一组样本数据来说, 用不同的方法估计回归函数的参数, 可拟合出不同的回归线。从散点图上看, 样本回归线对样本观测值总是存在或正或负的偏离。所估计的样本回归线对样本观测数据拟合的优劣程度, 称为样本回归线的拟合优度。为了评价所建立的样本回归函数对样本观测值的拟合程度, 需要对模型的拟合优度加以度量。在计量经济学中, 度量模型拟合优度的可决系数建立在对被解释变量总变差分解的基础之上。

一、总变差的分解

回顾已经估计的样本回归函数

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i = \hat{Y}_i + e_i \quad (\text{见 2.12})$$

如果以被解释变量平均值 \bar{Y} 为基准, 说明被解释变量观测值 Y_i 和估计值 \hat{Y}_i 对 \bar{Y} 的偏离程度, 上式可用离差表示为

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + e_i \quad (2.46)$$

(2.62) 式中各变量的关系如图 2.9 所示:

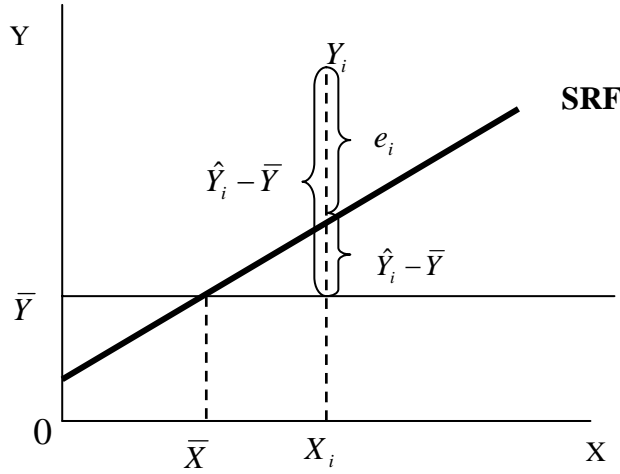


图 2.9 总变差的分解

将式 (2.46) 两边平方并对所有观测值加总，得到

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad (2.47)$$

或

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \quad (2.48)$$

在 (2.47) 或 (2.48) 式中：

(1) 被解释变量 Y 的样本观测值与其平均值的离差平方和 $\sum (Y_i - \bar{Y})^2 = \sum y_i^2$ ，称为总变差或总离差平方和，用 TSS 表示。

(2) 被解释变量 Y 的样本估计值与其平均值的离差平方和 $\sum (\hat{Y}_i - \bar{Y})^2 = \sum \hat{y}_i^2$ ，称为回归平方和，是由模型回归线作出解释的变差，用 ESS 表示。

(3) 被解释变量观测值与估计值之差的平方和 $\sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$ ，称为残差平方和，是回归线未作出解释的变差，用 RSS 表示。

这样，(2.47) 或 (2.48) 式也可写为

$$TSS = ESS + RSS \quad (2.49)$$

二、可决系数

将 (2.47) 式两边同除以 $TSS = \sum (Y_i - \bar{Y})^2$ ，得

$$1 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \quad (2.50)$$

或

$$1 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} + \frac{\sum e_i^2}{\sum y_i^2} \quad (2.51)$$

(2.50)或(2.51)式中 $\sum (\hat{Y}_i - \bar{Y})^2 / \sum (Y_i - \bar{Y})^2 = \sum \hat{y}_i^2 / \sum y_i^2$ 是由样本回归作出解释的离差平方和在总离差平方和中占的比重； $\sum (Y_i - \hat{Y}_i)^2 / \sum (Y_i - \bar{Y})^2 = \sum e_i^2 / \sum y_i^2$ 是回归线没有作出解释的离差平方和在总离差平方和中占的比重。

显然，如果样本回归线对样本观测值拟合程度越好，各样本观测点与回归线靠得越近，由样本回归作出解释的离差平方和在总离差平方和中占的比重也将越大，反之拟合程度越差，这部分所占比重就越小。所以 $\sum (\hat{Y}_i - \bar{Y})^2 / \sum (Y_i - \bar{Y})^2 = \sum \hat{y}_i^2 / \sum y_i^2$ 可以作为综合度量回归模型对样本观测值拟合优度的指标，这一比例称为可决系数（或称判定系数），在简单线性回归中一般用 r^2 表示，即

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \quad (2.52)$$

$$\text{或} \quad r^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2} \quad (2.53)$$

例如【例 2.2】用样本(一)数据估计的样本线性回归模型，在表 2.4 的相关数据中已经计算出 $\sum e_i^2 = 76650$ 和 $\sum y_i^2 = 5870212.5$ ，则可计算出可决系数为：

$$r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} = 1 - \frac{76650}{5870212.5} = 1 - 0.0131 = 0.9869$$

这说明，在样本(一)被解释变量(消费支出)观测值的总变差中，有 98.69% 由所估计的样本回归模型作出了解释。

可决系数 r^2 有如下特点：

- (1) 可决系数是非负的统计量；
- (2) 可决系数取值范围： $0 \leq r^2 \leq 1$ ；
- (3) 可决系数是样本观测值的函数，可决系数 r^2 是随抽样而变动的随机变量；

三、可决系数与相关系数的关系

可决系数与相关系数既有联系又有区别。在一元线性回归中，可决系数 r^2 在数值上是简单线性相关系数 r 的平方，即 $r = \pm \sqrt{r^2}$ 。

由 (2.2) 式，样本相关系数为

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (\text{见 2.2})$$

容易证明, 可决系数 r^2 也可表示为

$$r^2 = \frac{[\sum (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2} \quad (2.54)$$

虽然可决系数在数值上等于简单线性相关系数的平方, 但是应注意二者在概念上是有明显区别的。首先, 从意义上讲, 可决系数 r^2 是就估计的回归模型而言, 度量回归模型对本观测值的拟合程度, 也就是模型中解释变量对被解释变量变差的解释程度; 相关系数是就两个变量而言, 说明两个变量的线性依存程度。其次, 可决系数度量的是解释变量与被解释变量不对称的因果关系, 是在回归分析的基础上说明 X 对 Y 的变差的解释比例, 并不说明 Y 对 X 的解释; 而相关系数 r 度量的是 X 与 Y 对称的相关关系, 不涉及 X 与 Y 具体的因果关系。而且, 可决系数具有非负性, 取值范围为 $0 \leq r^2 \leq 1$; 而相关系数可正可负, 取值范围为 $-1 \leq r \leq 1$ 。

在计量经济学中, 主要研究回归模型的估计、检验和应用, 所以从实际应用看, 可决系数比相关系数更有意义。

第三节 回归系数的区间估计和假设检验

一、OLS 估计的分布性质

用样本数据通过 OLS 法估计的参数 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 都是随抽样而变化的随机变量, 由于存在抽样波动, 用样本得出的估计值并不一定等于参数的真实值。为此, 还需要通过估计的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 去对总体参数 β_1 和 β_2 的真实值作进一步的说明和推断, 这要求首先明确 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的具体分布性质。

前面已经指出, 在古典假定条件下, 假定随机扰动项 u_i 服从正态分布, 为此 Y_i 也服从正态分布。上一节又证明了 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 都是 Y_i 的线性函数, 所以即使在小样本情况下, $\hat{\beta}_1$ 和 $\hat{\beta}_2$

也服从正态分布。在大样本情况下，即使 Y_i 不服从正态分布， $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的分布也会趋于正态分布。前面还证明了 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是无偏估计，并确定了其方差的计算公式，所以 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的具体分布性质可以表示为

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 \frac{\sum X_i^2}{n \sum x_i^2}) \quad (2.55)$$

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum x_i^2}) \quad (2.56)$$

若将正态随机变量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 作标准化变换

$$z_1 = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 \frac{\sum X_i^2}{n \sum x_i^2}}} \sim N(0,1) \quad (2.57)$$

$$z_2 = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0,1) \quad (2.58)$$

即经标准化变换的 z_1 和 z_2 均服从标准正态分布。

可是， $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差以及标准正态变量 z_1 和 z_2 的确定，都要涉及随机扰动项 u_i 的方差 σ^2 ，而总体随机扰动项 u_i 是随机变量，其方差是未知的，只能通过 (2.44) 式计算 σ^2 的无偏估计值 $\hat{\sigma}^2$ 。

在大样本情况下，用无偏估计 $\hat{\sigma}^2$ 去代替 σ^2 ，可计算参数估计值的标准误差，这时用估计的标准误差作 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的标准化变换得到的 z_1 和 z_2 ，仍可视作标准正态分布变量。

在小样本情况下，若用无偏估计 $\hat{\sigma}^2$ 代替 σ^2 去估计标准误差，可得 $SE(\hat{\beta}_1)$ 和 $SE(\hat{\beta}_2)$ ，可以证明，用估计的标准误差作 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的标准化变换，所得的 $(\hat{\beta}_1 - \beta_1) / SE(\hat{\beta}_1)$ 和 $(\hat{\beta}_2 - \beta_2) / SE(\hat{\beta}_2)$ 已不再服从正态分布，而是服从自由度为 $n-2$ 的 t 分布，设这时的变换值为 t ，则

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 \frac{\sum X_i^2}{n \sum x_i^2}}} \sim t(n-2) \quad (2.59)$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} \sim t(n-2) \quad (2.60)$$

二、回归系数的区间估计

用 OLS 法得到的参数估计值只是对总体回归函数参数的点估计值，尽管在重复抽样中可以预期它的期望会等于参数的真实值，但还不能说明所得参数点估计值的可靠性。虽然前面已经确定了参数估计值的标准误差 $SE(\hat{\beta}_1)$ 和 $SE(\hat{\beta}_2)$ ，可是标准误差只是说明估计值与其均值的离散程度，还不能说明参数真实值的可能范围。参数估计值 $\hat{\beta}_2$ 可能比 β_2 小，也可能比 β_2 大，只是在 $\hat{\beta}_2$ 左右的一定区间范围内可能包含了 β_2 。我们可以设法找到可能包含参数真实值 β_2 的一定范围，并确定这样的范围包含参数真实值的可靠程度，这就是参数的区间估计。

为了确定 $\hat{\beta}_2$ 对真实值 β_2 “靠近”的程度，可设法找出两个正数 δ 和 α ，（其中 $0 < \alpha < 1$ ），以使得 $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ 这样的区间包含真实 β_2 的概率为 $1 - \alpha$ ，可表示为：

$$P(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta) = 1 - \alpha \quad (2.61)$$

这样的区间如果存在，就称为 β_2 的置信区间，其中 α 称为显著性水平， $1 - \alpha$ 称为置信系数或置信概率， $\hat{\beta}_2 - \delta$ 和 $\hat{\beta}_2 + \delta$ 分别称为下置信限和上置信限。

(2.61)式是回归模型中斜率系数 β_2 的区间估计式，类似地也可导出截距系数 β_1 的区间估计式。区间估计式的意义是在重复抽样之下，像这样的区间构造很多次，平均说来，这样的区间将有 $(1 - \alpha)\%$ 是包含参数真实值的。显然，构造参数的置信区间需要先确定参数估计式的抽样分布。前面对 u_i 的分布已作了正态性假定，并据此确定了 Y_i 、 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 和 $\hat{\sigma}^2$ 的分布，这就为构造置信区间创造了条件。

对回归系数的区间估计，可分为三种情况：

1、当总体方差 σ^2 已知时，在 u_i 的正态性假定下

$$z = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \sim N(0,1) \quad (\text{见 2.58})$$

由于 σ^2 已知， $SE(\hat{\beta}_2) = \sigma / \sqrt{\sum x_i^2}$ 可以确定，取定显著性水平 α 可得 Z 的临界值，例如取 $\alpha = 0.05$ ，即 $1 - \alpha = 0.95$ ，查正态分布表可知

$$P[-1.96 < z = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} < 1.96] = 0.95$$

$$P[\hat{\beta}_2 - 1.96SE(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + 1.96SE(\hat{\beta}_2)] = 0.95$$

所以，回归系数 β_2 的 95% 置信区间为 $[\hat{\beta}_2 - 1.96SE(\hat{\beta}_2), \hat{\beta}_2 + 1.96SE(\hat{\beta}_2)]$ 。

(2) 当总体方差 σ^2 未知，且样本容量充分大时，可用无偏估计 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 去代替 σ^2 。此时由于样本容量充分大，仍可认为

$$z = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma} / \sqrt{\sum x_i^2}} \sim N(0,1) \quad (2.62)$$

同样可以利用正态分布去确定 β_2 的置信区间。

(3) 当总体方差 σ^2 未知，且样本容量较小时，若用无偏估计 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 去代替 σ^2 ，此时

$$t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \sim t(n-2) \quad (\text{见 2.60})$$

这里的统计量 t 不再服从正态分布，而服从自由度为 $n-2$ 的 t 分布，可利用 t 分布去建立置信区间。

如果选取置信度为 $1 - \alpha$ ，查 t 分布表得显著性水平为 $\alpha/2$ ，自由度为 $n-2$ 的临界值 $t_{\alpha/2}$ ，可建立置信区间

$$P[-t_{\alpha/2} \leq t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \leq t_{\alpha/2}] = 1 - \alpha \quad (2.63)$$

即

$$P[\hat{\beta}_2 - t_{\alpha/2} \hat{SE}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \hat{SE}(\hat{\beta}_2)] = 1 - \alpha \quad (2.64)$$

同理， β_1 的置信区间也可按以上三种情况分别用类似的方法导出。

例如，【例 2.2】中运用表 2.2 样本(一)的数据已估计出参数为 $\hat{\beta}_1 = 352$ ， $\hat{\beta}_2 = 0.53$ 。

由于总体方差 σ^2 未知，且样本容量为 10 较小，可用估计的 $\hat{\sigma}^2 = 9581.25$ 代替 σ^2 ，表 2.4 中已算出 $\sum x_i^2 = 20625000$ ，则可计算出

$$\hat{SE}(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2 / \sum x_i^2} = \sqrt{9581.25 / 20625000} = 0.02155$$

以对参数 β_2 的区间估计为例，可取 $\alpha = 0.05$ ，查 t 分布表得自由度为 n-2 的临界值

$t_{\alpha/2} = t_{0.025} = 2.306$ ，将有关数据代入 (2.54) 式，即可得到 β_2 的置信度为 95% 的置信区间：

$$\begin{aligned} & P[\hat{\beta}_2 - t_{\alpha/2} \hat{SE}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \hat{SE}(\hat{\beta}_2)] \\ &= P(0.53 - 2.306 \times 0.02155 \leq \beta_2 \leq 0.53 + 2.306 \times 0.02155) \\ &= P(0.4803 \leq \beta_2 \leq 0.5797) \\ &= 95\% \end{aligned}$$

应当强调，由于 $\hat{\beta}_2$ 是随机变量，(2.64) 式那样的区间是随机区间，只是说明在重复抽样中，在 $1 - \alpha$ 的概率基础上，像这样的区间构造许多次，从长远看平均地说，这些区间中将有 95% 包含着 β_2 的真实值。

三、回归系数的假设检验

前面已经指出，简单线性回归模型的系数 β_1 、 β_2 和方差 σ^2 都不能直接观测或准确计算，只能通过样本观测值去估计，所得到的样本回归系数的估计量是随抽样而变动的随机变量。那末，像这样估计的回归系数和方差是否可靠？是否仅仅为抽样的偶然结果呢？还需要进行统计推断检验。

参数的区间估计与假设检验既有联系又有区别。参数的区间估计主要回答什么样的区间包含总体参数真实值的可靠程度问题；而假设检验是要根据已知的样本观测值，判断它是否与对总体参数作的某一个假设相一致。

对回归系数假设检验的基本思想，是所估计样本回归系数概率分布性质已确定的基础上，在对总体回归系数某种原假设成立的条件下，利用适当的有明确概率分布的统计量和给定的显著性水平 α ，构造一个小概率事件，判断原假设结果合理与否，是基于“小概率事件

不易发生”的原理，可以认为小概率事件在一次观察中基本不会发生，如果该小概率事件竟然发生了，就认为原假设不真，从而拒绝原假设，不拒绝备择假设。

对总体参数的假设检验可能有不同的要求，可以检验总体参数是否等于、大于或小于某特定的数值，这时原假设分别为 $H_0: \beta_2 = \beta_2^*$ 、 $H_0: \beta_2 \geq \beta_2^*$ 、 $H_0: \beta_2 \leq \beta_2^*$ ；也可以检验总体参数是否等于零。而且，原假设和备择假设的设定方式不同，判断是否拒绝区域的方式也不同，例如设定 $H_0: \beta_2 = \beta_2^*$ ， $H_1: \beta_2 \neq \beta_2^*$ ，进行的是双侧检验；而设定 $H_0: \beta_2 \geq \beta_2^*$ 或 $H_1: \beta_2 < \beta_2^*$ 时，进行的是单侧检验。针对不同的要求和条件，对总体参数的假设检验可分为多种情况。

1、Z 检验

当 σ^2 已知，或样本容量充分大时，按 (2.58) 式根据样本计算的 z^* 有

$$z^* = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} \sim N(0,1) \quad (2.65)$$

针对原假设 $H_0: \beta_2 = \beta_2^*$ ，备择假设 $H_1: \beta_2 \neq \beta_2^*$ ，可利用服从正态分布的 z^* 统计量作假设检验。给定显著性水平 α （例如 $\alpha = 0.05$ ），由正态分布表查出 Z 的临界值，例如为 1.96。把根据样本计算的 z^* 与 Z 的临界值作比较，如果 $-1.96 \leq z^* \leq 1.96$ ，就不拒绝原假设 $H_0: \beta_2 = \beta_2^*$ ；如果 $z^* < -1.96$ 或 $z^* > 1.96$ ，就拒绝 $H_0: \beta_2 = \beta_2^*$ ，而不拒绝 $H_1: \beta_2 \neq \beta_2^*$ ，即认为 β_2 显著不等于 β_2^* 。这种利用正态分布进行的显著性检验，也称为 Z 检验。

2、t 检验

当 σ^2 未知，且样本容量较小时，只能用无偏估计 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 去代替 σ^2 ，由 (2.52) 式已知

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} \sim t(n-2) \quad (2.66)$$

针对原假设 $H_0: \beta_2 = \beta_2^*$ ，备择假设 $H_1: \beta_2 \neq \beta_2^*$ ，给定显著性水平 α ，由 t 分布表可查出自由度为 n-2，对应概率为 $\alpha/2$ 的临界值 $t_{\alpha/2}(n-2)$ 。将根据样本计算的 t^* 统计量与临界值 $t_{\alpha/2}(n-2)$ 比较，如果 $-t_{\alpha/2} \leq t^* \leq t_{\alpha/2}$ ，就不拒绝 $H_0: \beta_2 = \beta_2^*$ ；如果 $t^* < -t_{\alpha/2}$ 或 $t^* > t_{\alpha/2}$ ，

就拒绝 $H_0: \beta_2 = \beta_2^*$ ，而不拒绝 $H_1: \beta_2 \neq \beta_2^*$ ，也就是认为 β_2 显著不等于 β_2^* 。像这样利用 t 分布进行的显著性检验，称为 t 检验。

为了检验所建立的回归模型中解释变量对被解释变量是否有显著影响，在计量经济学中经常把回归系数 $\beta_2 = 0$ 作为原假设。当原假设为 $H_0: \beta_2 = 0$ ，备择假设为 $H_1: \beta_2 \neq 0$ 时

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\hat{SE}(\hat{\beta}_2)} \sim t(n-2) \quad (2.67)$$

给定显著性水平 α ，由 t 分布表可查出自由度为 $n-2$ ，对应概率为 $\alpha/2$ 的临界值 $t_{\alpha/2}(n-2)$ 。如果 $-t_{\alpha/2} \leq t^* \leq t_{\alpha/2}$ ，就不拒绝 $H_0: \beta_2 = 0$ 即认为对应的解释变量对被解释变量没有显著影响；反之，如果 $t^* < -t_{\alpha/2}$ 或 $t^* > t_{\alpha/2}$ ，就拒绝 $H_0: \beta_2 = 0$ ，而不拒绝 $H_1: \beta_2 \neq 0$ ，即认为对应解释变量对被解释变量有显著影响。

对回归系数的假设检验是在给定的显著性水平下作出的，因此当给定的显著性水平不同时，对检验所得的结论很可能不同，甚至会得出相反的结论。例如，对于 $H_0: \beta_2 = 0$ ，在 $n=20$ ，由样本值计算的 t 统计量为 2.5000 时，若显著性水平取 0.05，临界值为 $t_{\alpha/2}(n-2) = t_{0.025}(18) = 2.1009$ ，应当拒绝 $H_0: \beta_2 = 0$ ；但若显著性水平取 0.01，临界值为 $t_{\alpha/2}(n-2) = t_{0.005}(18) = 2.8784$ ，则应当不拒绝 $H_0: \beta_2 = 0$ 。由此可以看出，在原假设既定，t 统计量已确定的情况下，对参数假设检验的结论与显著性水平息息相关，那末就可以从显著性水平出发去判断检验结果。在既定原假设下计算出回归系数的 t 统计量 t^* 以后，由 t 分布的性质可求得统计量大于 t^* 的概率：

$$P(t \geq t^* | H_0) = \alpha^* \quad (2.68)$$

这里的 α^* 是 t 统计量大于 t^* 值的概率，也就是尚不能拒绝原假设 $H_0: \beta_2 = 0$ 的最大显著水平，称为所估计的回归系数的 P 值。显然，我们所取的显著性水平 α （例如取 0.05）只要比 P 值 α^* 更大，就可在显著性水平 α 下拒绝 $H_0: \beta_2 = 0$ 。反之，所取的 α 还小于 P 值 α^* ，就应在显著性水平 α 下不拒绝 $H_0: \beta_2 = 0$ 。这就是回归系数显著性的 P 值检验方法。在 Eviews 及各种回归分析的软件中，在给出回归分析的结果时通常都会同时给出了原假设

$H_0: \beta_2 = 0$ 下所估计参数的 P 值，这给判断参数显著性检验的结论带来很大方便。

例如【例 2.2】中已估计出 $\hat{\beta}_2 = 0.53$ ，为了进一步检验家庭可支配收入是否对消费支出有显著影响，针对原假设 $H_0: \beta_2 = 0$ ，备择假设 $H_1: \beta_2 \neq 0$ ，可以计算 t 统计量

$$t^* = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0.53}{0.02155} = 25.5220$$

如果给定显著性水平 $\alpha = 0.05$ ，由 t 分布表可查出自由度为 $n-2=8$ 的临界值 $t_{0.025}(8) = 2.306$ 。因为 $t^* = 25.5220 > t_{\alpha/2} = 2.306$ ，所以可以拒绝 $H_0: \beta_2 = 0$ ，而不拒绝 $H_1: \beta_2 \neq 0$ ，即认为解释变量可支配收入对被解释变量消费支出确实有显著影响。

在回归分析的计算机软件（例如 Eviews）中，通常在给出回归系数的估计结果的同时，也给出了对应解释变量回归参数估计值的 P 值，例如在本例中与 $\hat{\beta}_2 = 0.53$ 对应的 P 值为 0.0000，远小于给定的 $\alpha = 0.05$ ，通过 P 值也可以判定应当拒绝 $H_0: \beta_2 = 0$ 的假设。

第五节 回归模型预测

一、回归分析结果的报告

回归模型经过估计检验以后，得到了一系列的数据，在以后的计量经济分析中还会得出更多的说明模型特性的有意义的数据。为了更清晰、更简明地表现这些数据，通常将这些数据加以整理，并用一定规范的格式去报告回归分析的结果。

例如对于【例 2.2】用 OLS 法所作的回归分析的结果得到以下数据：

$\hat{\beta}_1 = 352$ 、 $SE(\hat{\beta}_1) = 76.5826$ 、 $\hat{\beta}_2 = 0.53$ 、 $SE(\hat{\beta}_2) = 0.0216$ 、 $r^2 = 0.9869$ 、 $df=8$ 等。在 $H_0: \beta_1 = 0$ 的假设下，t 统计量为 4.5963；在 $H_0: \beta_2 = 0$ 的假设下，t 统计量为 24.5902。

在计量经济研究中，通常按以下规范格式表述以上各项数据：

$$\begin{aligned} \hat{Y}_i &= 352 + 0.5300 X_i & (2.69) \\ &(76.5826) \quad (0.0216) \\ t &= (4.5963) \quad (24.5902) \end{aligned}$$

$$r^2 = 0.9869 \quad df=8$$

其中列在回归方程下方第一排圆括号内的数据是对应参数估计值的标准误差；第二排圆括号内的数据分别是对应参数等于零的原假设下，所计算的 t 统计量（有时也可不列出标准误差，而只列出 t 统计量）。

按这种规范格式报告回归分析计算的结果，一是比较规范简洁，便于交流；二是可以较方便地看出所估计回归模型的特性，如回归系数是否显著。因此这种表现方式已被广泛采用。

二、被解释变量平均值预测

回归分析的目的之一是对被解释变量作合理的预测。所谓预测是指由已知的或预先测定的解释变量，去预测被解释变量在所观测的样本数据以外的数值。预测可以是对被解释变量未来时期的动态预测，也可以是对被解释变量在不同空间状况的空间预测。如果所建立的回归方程通过了各项统计检验，并且在经济上也是有实际意义的，我们估计出参数的回归模型就可以用于对被解释变量的预测。预测的基本方法是将解释变量预测期的数值 X_f 代入估计的模型，对被解释变量的预测期或样本以外的数值 \hat{Y}_f 作出定量的估计。

利用计量经济模型对被解释变量所作的预测，是在一定前提条件下进行的条件预测。首先要对模型在预测期的适应性作出判断，要在经济理论分析和研究的基础上，判定所研究的经济总体的经济结构在样本期和预测期并无明显变化。具体来说，这种预测是假定模型所设定的 Y 与 X 的关系式保持不变条件下的预测。此外，是认为用样本所估计的参数在预测期保持不变条件下的预测。当然，还必须是对解释变量在预测期的取值已经确定或已作出预测的条件下的预测。

对被解释变量的预测可分为对被解释变量 Y 的平均值预测和个别值预测。对 Y 的平均值预测又分为对平均值的点预测和区间预测。这几种预测的关系如图 2.10 所示：

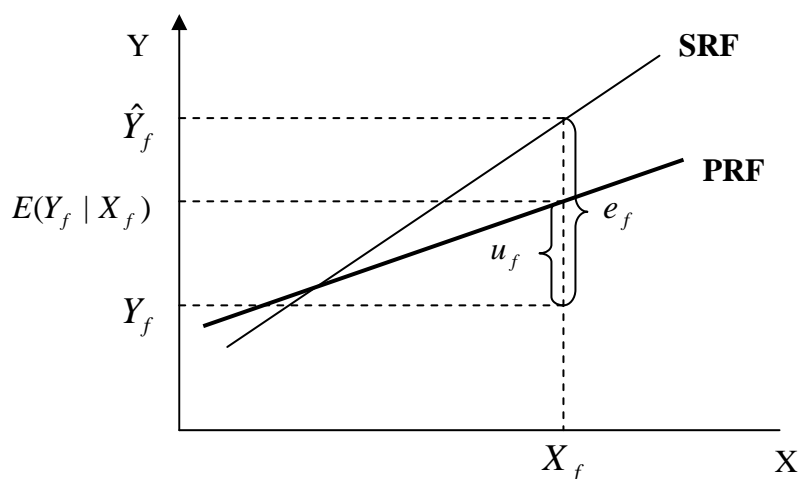


图 2.10 平均值预测与个别值预测的关系

对被解释变量预测的基本依据，是用样本估计的回归方程 $\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f$ ，由于存在抽样波动，估计的参数与总体真实参数有误差，因此用样本回归函数预测的被解释变量平均值 \hat{Y}_f 与预测期总体的真实平均值 $E(Y_f | X_f)$ 亦会有误差。此外，由于存在随机扰动，被解释变量在预测期的个别值 Y_f 与平均值 $E(Y_f | X_f)$ 也有误差。因此对 Y 的平均值预测和个别值预测要分别进行讨论。

1、对 Y 平均值的点预测

把解释变量的预测值 X_f 直接代入所估计的样本回归函数，就可以计算出被解释变量平均值的预测值

$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f \quad (2.70)$$

例如对于【例 2.2】，若是要预测当家庭可支配收入达到 6000 元时消费支出的平均水平，可将 $X_f = 6000$ 代入经过估计和检验的回归模型，得

$$\hat{Y}_f = 352 + 0.53 \times 6000 = 3632 \text{ (元)}$$

由样本回归函数的意义不难理解，用 (2.70) 式计算的 \hat{Y}_f 只是对 Y_f 的平均值作的点估计。 \hat{Y}_f 是由样本回归方程计算的，因为 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是随样本而变化的随机变量， \hat{Y}_f 也是一个随机变量。点预测值 \hat{Y}_f 不一定等于被解释变量预测期的真实平均值 $E(Y_f | X_f)$ ，我们还需要对 $E(Y_f | X_f)$ 可能的置信区间作出预测，也就是说要对 $E(Y_f | X_f)$ 进行区间预测。

2、对 Y 平均值的区间预测

为了由预测值 \hat{Y}_f 去对真实平均值 $E(Y_f | X_f)$ 作区间预测，应考虑预测值 \hat{Y}_f 的抽样分布，并寻找与 \hat{Y}_f 和 $E(Y_f | X_f)$ 都有关的统计量。

由前面的分析已知

$$E(\hat{Y}_f) = E(Y_f | X_f) = \beta_1 + \beta_2 X_f \quad (2.71)$$

还可以证明³

³ 此式证明过程较繁琐，本书对证明从略。

$$\text{Var}(\hat{Y}_f) = \sigma^2 \left[\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2} \right] \quad (2.72)$$

$$\text{SE}(\hat{Y}_f) = \sigma \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}} \quad (2.73)$$

一般情况下 σ^2 未知，可用无偏估计 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 去代替，此时

$$t = \frac{\hat{Y}_f - E(Y_f)}{\hat{\text{SE}}(\hat{Y}_f)} = \frac{\hat{Y}_f - E(Y_f | X_f)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}} \sim t(n-2) \quad (2.74)$$

显然，这里的 t 统计量与 \hat{Y}_f 和 $E(Y_f | X_f)$ 都有关，且服从自由度为 $n-2$ 的 t 分布。给定显著性水平 α ，查 t 分布表可得临界值 $t_{\alpha/2}(n-2)$ ，因此

$$P\{[\hat{Y}_f - t_{\alpha/2} \hat{\text{SE}}(\hat{Y}_f)] \leq E(Y_f | X_f) \leq [\hat{Y}_f + t_{\alpha/2} \hat{\text{SE}}(\hat{Y}_f)]\} = 1 - \alpha \quad (2.75)$$

即预测期平均值 $E(Y_f | X_f)$ 的置信度为 $1 - \alpha$ 的预测区间为：

$$[(\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}), (\hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}})] \quad (2.76)$$

例如对于【例 2.2】，要对可支配收入 $X_f = 6000$ 元时平均消费支出点预测值 $\hat{Y}_f = 3632$ 元的基础上，进一步对家庭平均消费支出作区间预测。给定显著性水平 $\alpha = 0.05$ ，查 t 分布表可得临界值 $t_{\alpha/2}(n-2) = t_{0.025}(10-2) = 2.306$ ，。前面利用 (2.44) 式已经计算出 $\hat{\sigma} = 97.8839$ ，在表 2.4 中已由样本数据计算出 $\sum x_i^2 = 20625000$ ， $\bar{X} = 3250$ 。由(2.76)式可计算消费支出平均值 $E(Y_f | X_f)$ 预测区间的上下限：

$$\begin{aligned} E(Y_f | X_f) &= \hat{Y}_f \mp t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}} \\ &= 3632 \mp 2.306 \times 97.8839 \times \sqrt{\frac{1}{10} + \frac{(6000 - 3250)^2}{20625000}} \\ &= 3632 \mp 154.19 (\text{元}) \end{aligned}$$

即是说，当家庭每月可支配收入达到 6000 元时，每月消费支出平均水平置信度为 95% 的区间预测值为 (3477.81, 3786.19) 元。

3、对 Y 个别值的预测

对应于给定的预测期解释变量的数值 X_f ，要在平均值预测的基础上，进一步确定 Y 个别值的预测区间，必须明确与预测值 \hat{Y}_f 和个别值 Y_f 都有关的统计量的概率分布。

由前面的分析已知，与预测期解释变量对应的残差项 $e_f = Y_f - \hat{Y}_f$ ，正是与 \hat{Y}_f 和 Y_f 都有关的随机变量，而且在 u_i 正态性假定下， e_f 也服从正态分布，在简单线性回归时可证明⁴

$$E(e_f) = E(Y_f - \hat{Y}_f) = 0 \quad (2.77)$$

$$Var(e_f) = E(Y_f - \hat{Y}_f)^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2} \right] \quad (2.78)$$

$$SE(e_f) = \sigma \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}} \quad (2.79)$$

当 σ^2 未知，用 $\hat{\sigma}^2 = \sum e_i^2 / (n-2)$ 代替时，可以证明对 e_f 标准化的变量服从自由度 n-2 的 t 分布，即

$$t = \frac{e_f - E(e_f)}{SE(e_f)} = \frac{Y_f - \hat{Y}_f}{SE(e_f)} \sim t(n-2) \quad (2.80)$$

这里的 t 统计量可用于关于个别值 Y_f 的区间预测。给定显著性水平 α ，查 t 分布表得自由度为 n-2 的临界值 $t_{\alpha/2}(n-2)$ ，则有

$$P\{[\hat{Y}_f - t_{\alpha/2} \hat{SE}(e_f)] \leq Y_f \leq [\hat{Y}_f + t_{\alpha/2} \hat{SE}(e_f)]\} = 1 - \alpha \quad (2.81)$$

因此，简单线性回归时 Y 的真实值 Y_f 的置信度为 $1 - \alpha$ 的预测区间为

$$\{[\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}], [\hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}]\} \quad (2.82)$$

例如，对于【例 2.2】，要对当可支配收入 $X_f = 6000$ 元时平均消费支出点预测值 $\hat{Y}_f = 3632$ 元的基础上，进一步对家庭消费支出个别值作区间预测。给定显著性水平

⁴ 证明过程较繁琐，本书对证明从略。

$\alpha=0.05$ ，由(2.87)式可计算消费支出个别值 Y_f 预测区间的上下限：

$$\begin{aligned} & \hat{Y}_f \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}} \\ &= 3632 \mp 2.306 \times 97.8839 \times \sqrt{1 + \frac{1}{10} + \frac{(6000 - 3250)^2}{20625000}} \\ &= 3632 \mp 273.36 \text{ (元)} \end{aligned}$$

即是说，当家庭每月可支配收入达到 6000 元时，每月消费支出个别值置信度为 95% 的区间预测值为 (3358.64, 3905.36) 元。

从对被解释变量的平均值预测和个别值预测，可以看出有以下特点：

(1) 由于抽样误差的存在，用样本估计的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 去预测的被解释变量平均值 \hat{Y}_f 与总体真实平均值 $E(Y_f | X_f)$ 存在误差，这主要决定于抽样波动。而用 \hat{Y}_f 对个别值 Y_f 的预测，不仅存在由于抽样波动引起的误差，而且要受随机扰动项 u_i 的影响。对比 (2.72) 和 (2.78) 式可以看出，由 \hat{Y}_f 对个别值预测的方差要大于对平均值预测的方差。从图 2.11 也可看出对个别值的预测区间比对平均值的预测区间更宽。

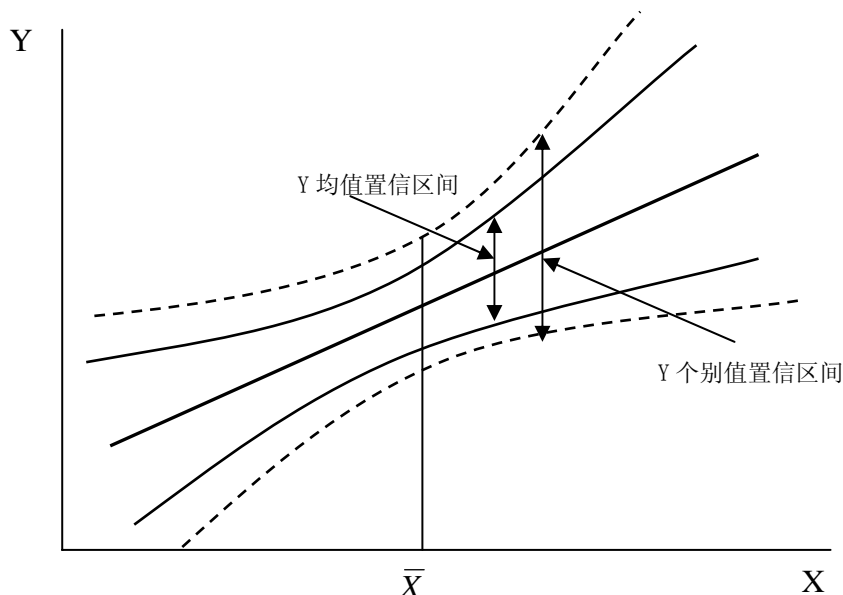


图 2.11 平均值和个别值的预测区间

(2) 对 Y_f 的平均值预测区间和个别值预测区间都不是常数，它们是随解释变量预测值 X_f 的变化而变化的，当 $X_f = \bar{X}$ 时， $(X_f - \bar{X})^2 = 0$ ，此时预测区间最窄， X_f 越是远离 \bar{X} ，

$(X_f - \bar{X})^2$ 越大, 预测区间也越宽 (见图 2.11)。所以用回归模型作预测时, X_f 的取值不宜偏离 \bar{X} 过远, 否则预测的精度会大大降低。

(3) 预测区间与样本容量 n 有关。由式 (2.76) 和 (2.82) 可看出, 样本容量越大, 不仅 n 越大, 而且 $\sum x_i^2$ 越大, 预测误差的方差将越小, 预测区间也将越窄。为此, 随着样本容量的增加, 预测的精度会提高, 如果样本容量过小, 预测的精度也将较差。当样本容量趋于无穷大 (即 $n \rightarrow \infty$) 时, 抽样误差趋于 0, 此时对平均值的预测误差亦趋于 0, 而对个别值的预测误差则只决定于随机扰动 u_i 的方差 σ^2 。

第六节 案例分析

一、研究的目的要求

居民消费在社会经济的持续发展中有重要的作用。居民合理的消费模式和居民适度的消费规模有利于经济持续健康的增长, 而且这也是人民生活水平的具体体现。改革开放以来随着中国经济的快速发展, 人民生活水平不断提高, 居民的消费水平也不断增长。但是在看到这个整体趋势的同时, 还应看到全国各地区经济发展速度不同, 居民消费水平也有明显差异。例如, 2002 年全国城市居民家庭平均每人每年消费支出为 6029.88 元, 最低的黑龙省仅为人均 4462.08 元, 最高的上海市达人均 10464 元, 上海是黑龙省的 2.35 倍。为了研究全国居民消费水平及其变动的原因, 需要作具体的分析。影响各地区居民消费支出有明显差异的因素可能很多, 例如, 居民的收入水平、就业状况、零售物价指数、利率、居民财产、购物环境等等都可能对居民消费有影响。为了分析什么是影响各地区居民消费支出有明显差异的最主要因素, 并分析影响因素与消费水平的数量关系, 可以建立相应的计量经济模型去研究。

二、模型设定

我们研究的对象是各地区居民消费的差异。居民消费可分为城市居民消费和农村居民消费, 由于各地区的城市与农村人口比例及经济结构有较大差异, 最具有直接对比可比性的是城市居民消费。而且, 由于各地区人口和经济总量不同, 只能用“城市居民每人每年的平均消费支出”来比较, 而这正是可从统计年鉴中获得数据的变量。所以模型的被解释变量 Y 选定为“城市居民每人每年的平均消费支出”。

因为研究的目的是各地区城市居民消费的差异, 并不是城市居民消费在不同时间的变

动，所以应选择同一时期各地区城市居民的消费支出来建立模型。因此建立的是 2002 年截面数据模型。

影响各地区城市居民人均消费支出有明显差异的因素有多种，但从理论和经验分析，最主要的影响因素应是居民收入，其他因素虽然对居民消费也有影响，但有的不易取得数据，如“居民财产”和“购物环境”；有的与居民收入可能高度相关，如“就业状况”、“居民财产”；还有的因素在运用截面数据时在地区间的差异并不大，如“零售物价指数”、“利率”。因此这些其他因素可以不列入模型，即便它们对居民消费有某些影响也可归入随即扰动项中。为了与“城市居民人均消费支出”相对应，选择在统计年鉴中可以获得的“城市居民每人每年可支配收入”作为解释变量 X。

从 2002 年《中国统计年鉴》中得到表 2.5 的数据：

表 2.5 2002 年中国各地区城市居民人均年消费支出和可支配收入

地 区	城市居民家庭平均每人每年消费支出(元) Y	城市居民人均年可支配收入(元) X
北京	10284.60	12463.92
天津	7191.96	9337.56
河北	5069.28	6679.68
山西	4710.96	5234.35
内蒙古	4859.88	6051.06
辽宁	5342.64	6524.52
吉林	4973.88	6260.16
黑龙江	4462.08	6100.56
上海	10464.00	13249.80
江苏	6042.60	8177.64
浙江	8713.08	11715.60
安徽	4736.52	6032.40
福建	6631.68	9189.36
江西	4549.32	6334.64
山东	5596.32	7614.36
河南	4504.68	6245.40

湖北	5608.92	6788.52
湖南	5574.72	6958.56
广东	8988.48	11137.20
广西	5413.44	7315.32
海南	5459.64	6822.72
重庆	6360.24	7238.04
四川	5413.08	6610.80
贵州	4598.28	5944.08
云南	5827.92	7240.56
西藏	6952.44	8079.12
陕西	5278.04	6330.84
甘肃	5064.24	6151.44
青海	5042.52	6170.52
宁夏	6104.92	6067.44
新疆	5636.40	6899.64

作城市居民家庭平均每人每年消费支出(Y)和城市居民人均年可支配收入(X)的散点图，如图 2.12：

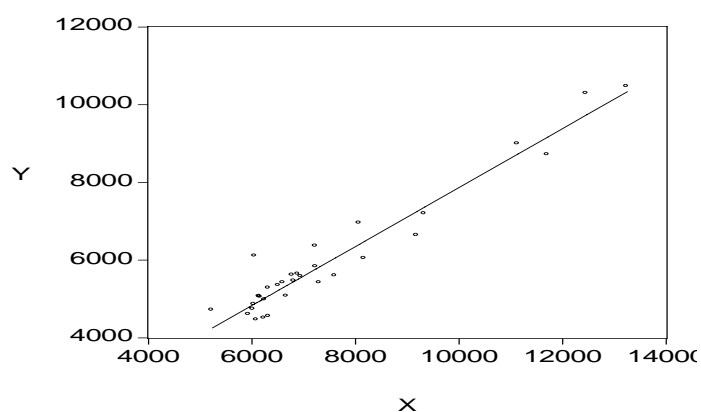


图 2.12

从散点图可以看出居民家庭平均每人每年消费支出(Y)和城市居民人均年可支配收入(X)大体呈现为线性关系，所以建立的计量经济模型为如下线性模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

三、估计参数

假定所建模型及随机扰动项 u_i 满足古典假定，可以用 OLS 法估计其参数。运用计算机软件 EViews 作计量经济分析十分方便。

利用 EViews 作简单线性回归分析的步骤如下：

1、建立工作文件

首先，双击 EViews 图标，进入 EViews 主页。在菜单一次点击 File\New\Workfile，出现对话框 “Workfile Range”。在 “Workfile frequency” 中选择数据频率：

Annual (年度)	Weekly (周数据)
Quarterly (季度)	Daily (5 day week) (每周 5 天日数据)
Semi Annual (半年)	Daily (7 day week) (每周 7 天日数据)
Monthly (月度)	Undated or irregular (未注明日期或不规则的)

在本例中是截面数据，选择 “Undated or irregular”。并在 “Start date” 中输入开始时间或顺序号，如 “1” 在 “end date” 中输入最后时间或顺序号，如 “31” 点击 “ok” 出现 “Workfile UNTITLED” 工作框。其中已有变量：“c” — 截距项 “resid” — 剩余项。

在 “Objects” 菜单中点击 “New Objects”，在 “New Objects” 对话框中选 “Group”，并在 “Name for Objects” 上定义文件名，点击 “OK” 出现数据编辑窗口。

若要将工作文件存盘，点击窗口上方 “Save”，在 “SaveAs” 对话框中给定路径和文件名，再点击 “ok”，文件即被保存。

2、输入数据

在数据编辑窗口中，首先按上行键 “↑”，这时对应的 “obs” 字样的空格会自动上跳，在对应列的第二个 “obs” 有边框的空格键入变量名，如 “Y”，再按下行键 “↓”，对因变量名下的列出现 “NA” 字样，即可依顺序输入响应的数据。其他变量的数据也可用类似方法输入。

也可以在 EViews 命令框直接键入 “data X Y”(一元时) 或 “data Y X_1 X_2 ...”(多元时)，回车出现 “Group” 窗口数据编辑框，在对应的 Y、X 下输入数据。

若要对数据存盘，点击 “fire/Save As”，出现 “Save As” 对话框，在 “Drives” 点所要存的盘，在 “Directories” 点存入的路径（文件名），在 “Fire Name” 对所存文件命名，或点已存的文件名，再点 “ok”。

若要读取已存盘数据，点击 “fire/Open”，在对话框的 “Drives” 点所存的磁盘名，在

“Directories”点文件路径，在“File Name”点文件名，点击“ok”即可。

3、估计参数

方法一：在 EViews 主页界面点击“Quick”菜单，点击“Estimate Equation”，出现“Equation specification”对话框，选 OLS 估计，即选击“Least Squares”，键入“Y C X”，点“ok”或按回车，即出现如表 2.6 那样的回归结果。

表 2.6

Method: Least Squares					
Date: 02/25/05 Time: 03:15					
Sample: 1 31					
Included observations: 31					
Variable	Coefficient	Std. Error	t-Statistic	Prob.	
C	282.2434	287.2649	0.982520	0.3340	
X	0.758511	0.036928	20.54026	0.0000	
R-squared	0.935685	Mean dependent var		5982.476	
Adjusted R-squared	0.933467	S.D. dependent var		1601.762	
S.E. of regression	413.1593	Akaike info criterion		14.94788	
Sum squared resid	4950317.	Schwarz criterion		15.04040	
Log likelihood	-229.6922	F-statistic		421.9023	
Durbin-Watson stat	1.481439	Prob(F-statistic)		0.000000	

在本例中，参数估计的结果为：

$$\hat{Y}_i = 282.2434 + 0.758511X_i$$

$$(287.2649) \quad (0.036928)$$

$$t=(0.982520) \quad (20.54026)$$

$$r^2 = 0.935685 \quad F=421.9023 \quad df=29$$

方法二：在 EViews 命令框中直接键入“LS Y C X”，按回车，即出现回归结果。

若要显示回归结果的图形，在“Equation”框中，点击“Resids”，即出现剩余项(Residual)、实际值(Actual)、拟合值(Fitted)的图形，如图 2.13 所示。

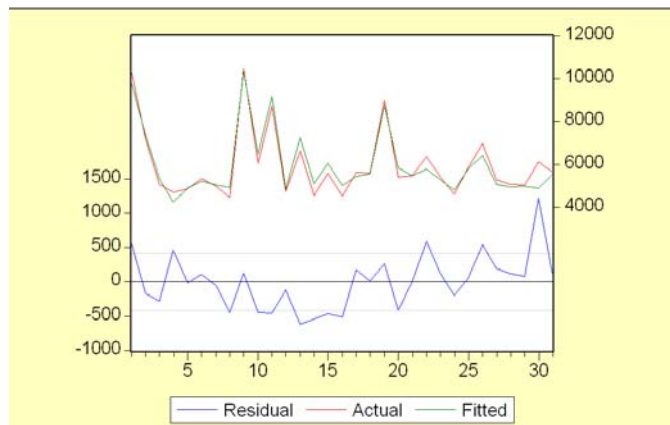


图 2.13

四、模型检验

1、经济意义检验

所估计的参数 $\hat{\beta}_2 = 0.758511$ ，说明城市居民人均年可支配收入每相差 1 元，可导致居民消费支出相差 0.758511 元。这与经济学中边际消费倾向的意义相符。

2、拟合优度和统计检验

用 EViews 得出回归模型参数估计结果的同时，已经给出了用于模型检验的相关数据。

拟合优度的度量：由表 2.6 中可以看出，本例中可决系数为 0.935685，说明所建模型整体上对样本数据拟合较好，即解释变量“城市居民人均年可支配收入”对被解释变量“城市居民人均年消费支出”的绝大部分差异作出了解释。

对回归系数的 t 检验：针对 $H_0: \beta_1 = 0$ 和 $H_0: \beta_2 = 0$ ，由表 2.6 中还可以看出，估计的回归系数 $\hat{\beta}_1$ 的标准误差和 t 值分别为： $SE(\hat{\beta}_1) = 287.2649$ ， $t(\hat{\beta}_1) = 0.982520$ ； $\hat{\beta}_2$ 的标准误差和 t 值分别为： $SE(\hat{\beta}_2) = 0.036928$ ， $t(\hat{\beta}_2) = 20.54026$ 。取 $\alpha = 0.05$ ，查 t 分布表得自由度为 $n - 2 = 31 - 2 = 29$ 的临界值 $t_{0.025}(29) = 2.045$ 。因为 $t(\hat{\beta}_1) = 0.982520 < t_{0.025}(29) = 2.045$ ，所以不能拒绝 $H_0: \beta_1 = 0$ ；因为 $t(\hat{\beta}_2) = 20.54026 > t_{0.025}(29) = 2.045$ ，所以应拒绝 $H_0: \beta_2 = 0$ 。这表明，城市人均年可支配收入对人均年消费支出有显著影响。

五、回归预测

由表 2.5 中可看出，2002 年中国西部地区城市居民人均年可支配收入除了西藏外均在 8000 以下，人均消费支出也都在 7000 元以下。在西部大开发的推动下，如果西部地区的城

市居民人均年可支配收入第一步争取达到 1000 美元(按现有汇率即人民币 8270 元), 第二步再争取达到 1500 美元(即人民币 12405 元), 利用所估计的模型可预测这时城市居民可能达到的人均年消费支出水平。可以注意到, 这里的预测是利用截面数据模型对被解释变量在不同空间状况的空间预测。

用 EViews 作回归预测, 首先在 “Workfile” 窗口点击 “Range”, 出现 “Change Workfile Range” 窗口, 将 “End data” 由 “31” 改为 “33”, 点 “OK”, 将 “Workfile” 中的 “Range” 扩展为 1—33。在 “Workfile” 窗口点击 “sampl”, 将 “sampl” 窗口中的 “1 31” 改为 “1 33”, 点 “OK”, 将样本区也改为 1—33。

为了输入 $X_{f1} = 8270$, $X_{f2} = 12405$ 在 EViews 命令框键入 data x /回车, 在 X 数据表中的 “32” 位置输入 “8270”, 在 “33” 的位置输入 “12405”, 将数据表最小化。

然后在 “Equation” 框中, 点击 “Forecast”, 得对话框。在对话框中的 “Forecast name” (预测值序列名) 键入 “ Y_f ”, 回车即得到模型估计值及标准误差的图形。双击 “Workfile” 窗口中出现的 “ Y_f ”, 在 “ Y_f ” 数据表中的 “32” 位置出现预测值 $Y_{f1} = 6555.132$, 在 “33” 位置出现 $Y_{f2} = 9691.577$ 。这是当 $X_{f1} = 8270$ 和 $X_{f2} = 12405$ 时人均消费支出的点预测值。

为了作区间预测, 在 X 和 Y 的数据表中, 点击 “View” 选 “Descriptive Stats\Common Sample”, 则得到 X 和 Y 的描述统计结果, 见表 2.7:

表 2.7

	X	Y	
Mean	7515.026	5982.476	
Median	6788.520	5459.640	
Maximum	13249.80	10464.00	
Minimum	5234.350	4462.080	
Std. Dev.	2042.682	1601.762	
Skewness	1.585893	1.629968	
Kurtosis	4.458645	4.787999	
Jarque-Bera	15.74267	17.85617	
Probability	0.000382	0.000133	
Observations	31	31	

根据表 2.7 的数据可计算:

$$\sum x_i^2 = \sigma_x^2(n-1) = 2042.682^2 \times (31-1) = 125176492.59$$

$$(X_{f1} - \bar{X})^2 = (8270 - 7515.026)^2 = 569985.74$$

$$(X_{f2} - \bar{X})^2 = (12405 - 7515.026)^2 = 23911845.72$$

取 $\alpha = 0.05$, Y_f 平均值置信度 95% 的预测区间为:

$$\hat{Y}_f \mp t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}$$

$$\begin{aligned} X_{f1} = 8270 \text{ 时} \quad & 6555.13 \mp 2.045 \times 413.1593 \times \sqrt{\frac{1}{31} + \frac{569985.74}{125176492.59}} \\ & = 6555.13 \mp 162.10 \end{aligned}$$

$$\begin{aligned} X_{f2} = 12405 \text{ 时} \quad & 9691.58 \mp 2.045 \times 413.1593 \times \sqrt{\frac{1}{31} + \frac{23911845.72}{125176492.59}} \\ & = 9691.58 \mp 499.25 \end{aligned}$$

即是说, 当 $X_{f1} = 8270$ 元时, Y_{f1} 平均值置信度 95% 的预测区间为 (6393.03, 6717.23)

元。当 $X_{f2} = 12405$ 元时, Y_{f2} 平均值置信度 95% 的预测区间为 (9292.33, 10090.83) 元。

Y_f 个别值置信度 95% 的预测区间为:

$$\hat{Y}_f \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum x_i^2}}$$

$$\begin{aligned} X_{f1} = 8270 \text{ 时} \quad & 6555.13 \mp 2.045 \times 413.1593 \times \sqrt{1 + \frac{1}{31} + \frac{569985.74}{125176492.59}} \\ & = 6555.13 \mp 860.32 \end{aligned}$$

$$\begin{aligned} X_{f2} = 12405 \text{ 时} \quad & 9691.58 \mp 2.045 \times 413.1593 \times \sqrt{1 + \frac{1}{31} + \frac{23911845.72}{125176492.59}} \\ & = 9691.58 \mp 934.49 \end{aligned}$$

即是说, 当第一步 $X_{f1} = 8270$ 时, Y_{f1} 个别值置信度 95% 的预测区间为 (5694.81, 7415.45) 元。当第二步 $X_{f2} = 12405$ 时, Y_{f2} 个别值置信度 95% 的预测区间为 (8757.09, 10626.07) 元。

在 “Equation” 框中, 点击 “Forecast” 可得预测值及标准误差的图形如图 2.14:

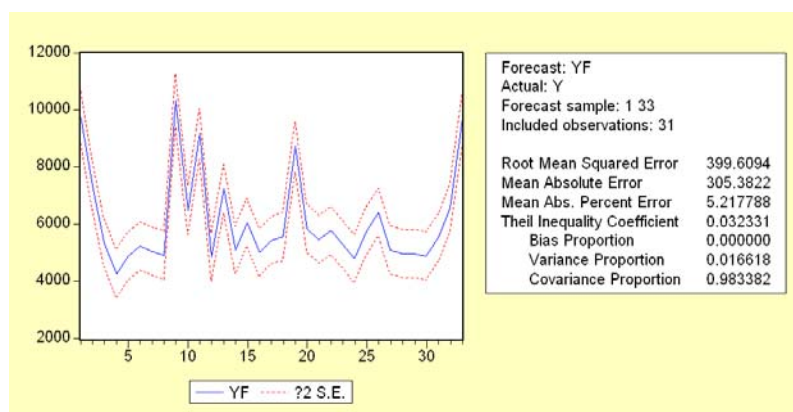


图 2.14

第二章小结

- 1、变量间的关系分为函数关系与相关关系。相关系数是对变量间线性相关程度的度量。
- 2、现代意义的回归是一个被解释变量对若干个解释变量依存关系的研究，回归的实质是由固定的解释变量去估计被解释变量的平均值。简单线性回归模型是只有一个解释变量的线性回归模型。
- 3、总体回归函数（PRF）是将总体被解释变量 Y 的条件均值 $E(Y_i|X_i)$ 表现为解释变量 X 的某种函数。样本回归函数（SRF）是将被解释变量 Y 的样本条件均值 \hat{Y}_i 表示为解释变量 X 的某种函数。总体回归函数与样本回归函数的区别与联系。
- 4、随机扰动项 u_i 是被解释变量实际值 Y_i 与条件均值 $E(Y_i|X_i)$ 的偏差，代表排除在模型以外的所有因素对 Y 的影响。
- 5、简单线性回归的基本假定：对模型和变量的假定、对随机扰动项 u 的假定（零均值假定、同方差假定、无自相关假定、随机扰动与解释变量不相关假定、正态性假定）
- 6、普通最小二乘法（OLS）估计参数的基本思想及估计式；OLS 估计式的分布性质及期望、方差和标准误差；OLS 估计式是最佳线性无偏估计式。
- 7、对回归系数区间估计的思想和方法。
- 8、拟合优度是样本回归线对样本观测数据拟合的优劣程度，可决系数是在总变差分解基础上确定的。可决系数的计算方法、特点与作用。
- 9、对回归系数假设检验的基本思想。对回归系数 t 检验的思想和方法；用 P 值判断参数的显著性。

10、被解释变量平均值预测与个别值预测的关系，被解释变量平均值的点预测和区间预测的方法，被解释变量个别值区间预测的方法。

11、运用 EViews 软件实现对简单线性回归模型的估计和检验。

第二章主要公式表

1、总体回归函数	$Y_i = \beta_1 + \beta_2 X_i + u_i$	$E(Y_i X_i) = \beta_1 + \beta_2 X_i$
2、样本回归函数	$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$	$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
3、基本假定	$E(u_i) = 0$ $Var(u_i) = Var(Y_i) = \sigma^2$ $Cov(u_i, X_i) = 0$	$E(Y_i) = \beta_1 + \beta_2 X_i$ $Cov(u_i, u_j) = E(u_i u_j) = 0$ $u_i \sim N(0, \sigma^2)$
4、最小二乘估计	$\hat{\beta}_2 = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2} = \frac{\sum x_i y_i}{\sum x_i^2}$ $\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N \sum X_i^2 - (\sum X_i)^2} \quad \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$	
5、参数 OLS 估计式的期望	$E(\hat{\beta}_k) = \beta_k$	
6、参数 OLS 估计式的方差	$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$	$Var(\hat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2}$
7、参数估计式的标准误差	$SE(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$	$SE(\hat{\beta}_1) = \sigma \sqrt{\frac{\sum X_i^2}{N \sum x_i^2}}$
8、 σ^2 的无偏估计	$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$	
9、t 检验统计量	$t^* = \frac{\hat{\beta}_2 - \beta_2}{\frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t(n-2)$	
8、样本可决系数	$1 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} + \frac{\sum e_i^2}{\sum y_i^2} \quad r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \quad r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$	
9、参数估计的置信区间	$P[\hat{\beta}_2 - t_{\alpha/2} SE(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} SE(\hat{\beta}_2)] = 1 - \alpha$	

10、平均值预测区间	$[\hat{Y}_F - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}, \hat{Y}_F + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}]$
11、个别值预测区间	$Y_F = \hat{Y}_F \mp t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_F - \bar{X})^2}{\sum x_i^2}}$

思考题与练习题

思考题

2.1 相关分析与回归分析的关系是什么？

2.2 什么是总体回归函数和样本回归函数？它们之间的区别是什么？

2.3 什么是随机扰动项和剩余项(残差)?它们之间的区别是什么？

2.4 为什么在对参数作最小二乘估计之前,要对模型提出古典假定？

2.5 总体方差和参数估计方差的区别是什么？

2.6 为什么可决系数可以度量模型的拟合优度？在简单线性回归中它与对参数的 t 检验的关系是什么？

2.7 有人说：“得到参数区间估计的上下限后，说明参数的真实值落入这个区间的概率为 $1-\alpha$ ”，如何评论这种说法？

2.8 对参数假设检验的基本思想是什么？

2.9 为什么对被解释变量个别值的预测区间会比对被解释变量平均值的预测区间更宽？

2.10 如果有人利用中国 1978 年—2000 年的样本估计的计量经济模型直接预测：“中国综合经济水平将在 2050 年达到美国 2002 年的水平”，你如何评论这种预测？

2.11 对本章开始提出的“中国旅游业总收入将超过 3000 亿美元？” ,你认为可以建立什么样的简单线性回归模型去分析？

练习题

2.1 为了研究深圳市地方预算内财政收入与国内生产总值的关系，得到以下数据：

年 份	地方预算内财政收入 Y (亿元)	国内生产总值(GDP)X (亿元)
1990	21.7037	171.6665
1991	27.3291	236.6630
1992	42.9599	317.3194
1993	67.2507	449.2889
1994	74.3992	615.1933
1995	88.0174	795.6950
1996	131.7490	950.0446
1997	144.7709	1130.0133
1998	164.9067	1289.0190
1999	184.7908	1436.0267
2000	225.0212	1665.4652
2001	265.6532	1954.6539

资料来源:《深圳统计年鉴 2002》, 中国统计出版社

- (1) 建立深圳地方预算内财政收入对 GDP 的回归模型;
- (2) 估计所建立模型的参数, 解释斜率系数的经济意
- (3) 对回归结果进行检验
- (4) 若是 2005 年年的国内生产总值为 3600 亿元, 确定 2005 年财政收入的预测值和预测区间 ($\alpha = 0.05$)。

2.2 某企业研究与发展经费与利润的数据(单位:万元)列于下表:

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
研究与发展经费	10	10	8	8	8	12	12	12	11	11
利 润 额	100	150	200	180	250	300	280	310	320	300

分析企业”研究与发展经费与利润额的相关关系, 并作回归分析。

2.3 为研究中国的货币供应量(以货币与准货币 M2 表示)与国内生产总值(GDP)的相互依存关系,分析表中 1990 年—2001 年中国货币供应量(M2)和国内生产总值(GDP)的有关数据:

年份	货币供应量(亿元)	国内生产总值(亿元)
	M2	GDP
1990	1529.3	18598.4
1991	19349.9	21662.5
1992	25402.2	26651.9
1993	34879.8	34560.5
1994	46923.5	46670.0
1995	60750.5	57494.9
1996	76094.9	66850.5
1997	90995.3	73142.7
1998	104498.5	76967.2
1999	119897.9	80579.4
2000	134610.3	88228.1
2001	158301.9	94346.4

资料来源:《中国统计年鉴 2002》,第 51 页、第 662 页,中国统计出版社

对货币供应量与国内生产总值作相关分析,并说明分析结果的经济意义。

2.4 表中是 16 支公益股票某年的每股帐面价值和当年红利:

公司序号	帐面价值(元)	红利(元)	公司序号	帐面价值(元)	红利(元)
1	22.44	2.4	9	12.14	0.80
2	20.89	2.98	10	23.31	1.94
3	22.09	2.06	11	16.23	3.00
4	14.48	1.09	12	0.56	0.28
5	20.73	1.96	13	0.84	0.84
6	19.25	1.55	14	18.05	1.80
7	20.37	2.16	15	12.45	1.21

8	26.43	1.60	16	11.33	1.07
---	-------	------	----	-------	------

根据上表资料：

- (1) 建立每股帐面价值和当年红利的回归方程；
- (2) 解释回归系数的经济意义；
- (3) 若序号为 6 的公司的股票每股帐面价值增加 1 元，估计当年红利可能为多少？

2.5 美国各航空公司业绩的统计数据公布在《华尔街日报 1999 年年鉴》(The Wall Street Journal Almanac 1999) 上。航班正点到达的比率和每 10 万名乘客投诉的次数的数据如下⁵。

航空公司名称	航班正点率 (%)	投诉率 (次/10 万名乘客)
西南(Southwest)航空公司	81.8	0.21
大陆(Continental)航空公司	76.6	0.58
西北(Northwest)航空公司	76.6	0.85
美国(US Airways)航空公司	75.7	0.68
联合(United)航空公司	73.8	0.74
美洲(American)航空公司	72.2	0.93
德尔塔 (Delta) 航空公司	71.2	0.72
美国西部(Americawest)航空公司	70.8	1.22
环球(TWA)航空公司	68.5	1.25

- (1) 画出这些数据的散点图
- (2) 根据散点图。表明二变量之间存在什么关系？
- (3) 求出描述投诉率是如何依赖航班按时到达正点率的估计的回归方程。
- (4) 对估计的回归方程的斜率作出解释。
- (5) 如果航班按时到达的正点率为 80%，估计每 10 万名乘客投诉的次数是多少？

2.6 研究青春发育与远视率（对数视力）的变化关系，测得结果如下表：

年龄 (岁) x	远视率 (%) y	对数视力 $Y=\ln y$
6	63.64	4.153

⁵资料来源：(美)David R. Anderson 等《商务与经济统计》，第 405 页，机械工业出版社

7	61.06	4.112
8	38.84	3.659
9	13.75	2.621
10	14.50	2.674
11	8.07	2.088
12	4.41	1.484
13	2.27	0.82
14	2.09	0.737
15	1.02	0.02
16	2.51	0.92
17	3.12	1.138
18	2.98	1.092

试建立曲线回归方程 $\hat{y} = a e^{bx}$ ($\hat{Y} = \ln a + b x$) 并进行计量分析。

2.7 为研究美国软饮料公司的广告费用 X 与销售数量 Y 的关系, 分析七种主要品牌软饮料公司的有关数据⁶

美国软饮料公司广告费用与销售数量

品牌名称	广告费用 X(百万美元)	销售数量 Y(百万箱)
Coca-Cola Classic	131.3	1929.2
Pepsi-Cola	92.4	1384.6
Diet-Coke	60.4	811.4
Sprite	55.7	541.5
Dr. Pepper	40.2	546.9
Moutain Dew	29.0	535.6
7-Up	11.6	219.5

分析广告费用对美国软饮料公司销售影响的数量关系。

⁶资料来源:(美)David R.Anderson 等《商务与经济统计》，第 405 页，机械工业出版社

2.8 从某公司分布在 11 个地区的销售点的销售量 (Y) 和销售价格 (X) 观测值得出以下结果:

$$\bar{X} = 519.8 \quad \bar{Y} = 217.82 \quad \sum X_i^2 = 3134543 \quad \sum X_i Y_i = 1296836$$

$$\sum Y_i^2 = 539512$$

(1) 作销售额对价格的回归分析, 并解释其结果。

(2) 回归直线未解释的销售变差部分是多少?

2.9 表中是中国 1978 年-1997 年的财政收入 Y 和国内生产总值 X 的数据:

中国国内生产总值及财政收入		单位: 亿元
年 份	国内生产总值 X	财政收入 Y
1978	3624. 1	1132. 26
1979	4038. 2	1146. 38
1980	4517. 8	1159. 93
1981	4860. 3	1175. 79
1982	5301. 8	1212. 33
1983	5957. 4	1366. 95
1984	7206. 7	1642. 86
1985	8989. 1	2004. 82
1986	10201. 4	2122. 01
1987	11954. 5	2199. 35
1988	14992. 3	2357. 24
1989	16917. 8	2664. 90
1990	18598. 4	2937. 10
1991	21662. 5	3149. 48
1992	26651. 9	3483. 37
1993	34560. 5	4348. 95
1994	46670. 0	5218. 10
1995	57494. 9	6242. 20

1006	66850.5	7407.99
1997	73452.5	8651.14

数据来源：《中国统计年鉴》

试根据这些数据完成下列问题：

- (1) 建立财政收入对国内生产总值的简单线性回归模型，并解释斜率系数的经济意义；
- (2) 估计所建立模型的参数，并对回归结果进行检验；
- (3) 若是 1998 年的国内生产总值为 78017.8 亿元，确定 1998 年财政收入的预测值和预测区间 ($\alpha = 0.05$)。

第二章附录

附录 2.1 简单线性回归最小二乘估计最小方差性质的证明

对于 OLS 估计式 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ ，已知其方差为

$$Var(\hat{\beta}_1) = \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2}$$

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

这里只证明 $Var(\hat{\beta}_2)$ 最小， $Var(\hat{\beta}_1)$ 最小的证明可以类似得出。

设 β_2 的另一个线性无偏估计为 β_2^* ，即

$$\beta_2^* = \sum w_i Y_i$$

其中

$$w_i \neq k_i, k_i = \frac{x_i}{\sum x_i^2}$$

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

因为 β_2^* 也是 β_2 的无偏估计，即 $E(\beta_2^*) = \beta_2$ ，必须有

$$\sum w_i = 0, \quad \sum w_i X_i = 1$$

同时

$$\begin{aligned} \text{Var}(\beta_2^*) &= \text{Var}(\sum w_i Y_i) \\ &= \sum w_i^2 \text{Var}(Y_i) \\ &= \sigma^2 \sum w_i^2 \quad [\text{因为 } \text{Var}(Y_i) = \sigma^2] \\ &= \sigma^2 \sum (w_i - k_i + k_i)^2 \\ &= \sigma^2 \sum (w_i - k_i)^2 + \sigma^2 \sum k_i^2 + 2\sigma^2 \sum (w_i - k_i)k_i \\ &= \sigma^2 \sum (w_i - k_i)^2 + \sigma^2 \sum k_i^2 + 2\sigma^2 (\sum w_i k_i - \sum k_i^2) \end{aligned}$$

上式最后一项中

$$\begin{aligned} \sum w_i k_i - \sum k_i^2 &= \frac{\sum w_i x_i}{\sum x_i^2} - \frac{\sum x_i^2}{(\sum x_i^2)^2} \\ &= \frac{\sum w_i (X_i - \bar{X})}{\sum x_i^2} - \frac{1}{\sum x_i^2} \\ &= \frac{\sum w_i X_i - \bar{X} \sum w_i}{\sum x_i^2} - \frac{1}{\sum x_i^2} \\ &= 0 \quad (\text{因为 } \sum w_i = 0, \quad \sum w_i X_i = 1) \end{aligned}$$

所以

$$\begin{aligned} \text{Var}(\beta_2^*) &= \sigma^2 \sum (w_i - k_i)^2 + \sigma^2 \sum \left[\frac{x_i^2}{(\sum x_i^2)^2} \right] \\ &= \sigma^2 \sum (w_i - k_i)^2 + \frac{\sigma^2}{\sum x_i^2} \\ &= \sigma^2 \sum (w_i - k_i)^2 + \text{Var}(\hat{\beta}_2) \end{aligned}$$

而 $\sigma^2 \geq 0$ ，因为 $w_i \neq k_i$ ，则有 $(w_i - k_i)^2 \geq 0$ ，为此

$$\text{Var}(\beta_2^*) \geq \text{Var}(\hat{\beta}_2)$$

只有 $w_i = k_i$ 时， $\text{Var}(\beta_2^*) = \text{Var}(\hat{\beta}_2)$ ，由于 β_2^* 是任意设定的 β_2 的线性无偏估计式，这表明 $\hat{\beta}_2$ 的 OLS 估计式具有最小方差性。

附录 2.2 σ^2 最小二乘估计的证明

用离差形式表示模型时

$$\begin{aligned} y_i &= Y_i - \bar{Y} \\ &= (\beta_1 + \beta_2 X_i + u_i) - (\beta_1 + \beta_2 \bar{X} + \bar{u}) \\ &= (u_i - \bar{u}) + \beta_2 x_i \end{aligned}$$

而且

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

因此

$$e_i = y_i - \hat{y}_i = (u_i - \bar{u}) - (\hat{\beta}_2 - \beta_2)x_i$$

则有

$$\begin{aligned} \sum e_i^2 &= \sum [(u_i - \bar{u}) - (\hat{\beta}_2 - \beta_2)x_i]^2 \\ &= \sum (u_i - \bar{u})^2 + (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 - 2(\hat{\beta}_2 - \beta_2) \sum (u_i - \bar{u})x_i \end{aligned}$$

取 $\sum e_i^2$ 的期望

$$E(\sum e_i^2) = E[\sum (u_i - \bar{u})^2] + \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 - 2E[(\hat{\beta}_2 - \beta_2) \sum (u_i - \bar{u})x_i]$$

式中

$$(1) \quad E[\sum (u_i - \bar{u})^2] = E[\sum u_i^2 - n(\bar{u})^2]$$

$$= \sum E(u_i^2) - \frac{1}{n} E(\sum u_i)^2$$

$$= \sum \sigma^2 - \frac{1}{n} E(u_1^2 + u_2^2 + \cdots + u_n^2 + 2u_1u_2 + \cdots + 2u_{n-1}u_n)$$

$$= \sum \sigma^2 - \frac{1}{n} E(u_1^2 + u_2^2 + \cdots + u_n^2)$$

$$= \sum \sigma^2 - \frac{1}{n} n\sigma^2 = (n-1)\sigma^2$$

$$(2) \quad \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 = \sum x_i^2 \frac{\sigma^2}{\sum x_i^2} = \sigma^2$$

$$(3) \quad -2E[(\hat{\beta}_2 - \beta_2) \sum (u_i - \bar{u})x_i] = -2E[\frac{\sum x_i u_i}{\sum x_i^2} (\sum x_i u_i - \bar{u} \sum x_i)]$$

$$= -2E[\frac{(\sum x_i u_i)^2}{\sum x_i^2}]$$

$$\begin{aligned}
 &= -2E[(\hat{\beta}_2 - \beta_2)^2 \sum x_i^2] \\
 &= -2 \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 = -2\sigma^2
 \end{aligned}$$

所以 $E(\sum e_i^2) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$

如果定义 $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$

其期望值为 $E(\hat{\sigma}^2) = E[\frac{\sum e_i^2}{n-2}] = \sigma^2$

这说明 $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 是 σ^2 的无偏估计。

第三章 多元线性回归模型

引子:

中国汽车的保有量会将达到 1.4 亿辆吗？

20 世纪 90 年代以来，随着中国经济的快速发展，居民收入不断增加，数以百万计的中国人开始得以实现拥有汽车的梦想，中国也成为世界上成长最快的汽车市场。与此同时中国的汽车工业也得到飞速发展，2004 年汽车产量达 507.05 万辆，同比增长 14.2%，比 1990 年增长了 886.5%。中国汽车产量的增长速度在世界各国是前所未有。汽车行业已成为拉动中国经济持续、快速、健康增长的“火车头”。“2020 年，中国的民用汽车保有量将比 2003 年的数字增长 6 倍，达到 1.4 亿辆左右”，这是中国交通部副部长在中国交通可持续发展论坛上做出的预测。据交通部规划司推算，中国未来汽车拥有水平的饱和度大约是每千人 150 辆左右，与目前新加坡的水平相似。照此推算，中国终极汽车保有量将达到 2.4 亿到 2.5 亿辆。

由国务院发展研究中心产业经济研究部提供的一份《中国汽车市场需求预测》显示，中国已经超过法国，在美国、日本、德国之后成为世界第四大汽车生产国。德国在 2004 年的汽车产量为 550 万辆，是仅次于美国和日本的世界第三大汽车生产国。预计中国汽车生产将在 2005 年增长 20%，产量将至 600 万辆，有可能超过德国而成为世界第三大汽车生产国。

但是也有人认为，尽管中国汽车产量的增长速度在世界各国前所未有，但仍然不要盲目乐观。因为中国进口汽车配额许可证管理将全部取消，其他商品的非关税措施，包括银行、保险、证券、分销等各大行业享有的保护性政策也将随之“失效”，“如果不认真对待或及早准备后过渡期的挑战，必将会对国内的相关产业带来严重冲击”。

目前，汽车行业投资增速过快，产能高速增加，有可能造成供需失衡加剧，从而导致汽车价格下降。而且随着几大集团投资的大项目，都要在 2005 年和 2006 年先后投入生产，轿车产能和需求的矛盾会更加突出。而且煤、电、油、运、原材料价格上涨的势头仍然难以控制。中国人民银行宣布的加息举措，虽然对启动汽车消费市场的影响有限，但对制造商、经销商来说却是一场“劫难”，因为这将意味着一个加息周期的轮回，汽车业的制造商和经销商们将难以应付；

此外，由于全球能源价格的暴涨，“未来中国可能将持续面临一段高油价的消费环境”，加之纷纷扬扬讨论的将要出台燃油税政策，将影响对汽车的有效需求。

也有人认为，外部的客观环境不会宽松，包括信贷、利率、保险、城市的交通条件等，都将影响汽车行业大势走向。未来道路增长将远远赶不上汽车增长的需要。还有人把新交通安全法中增加

的“机动车全赔”新规定，人保车险中新出台的“500元以下免赔条款”都归入到“对汽车消费不是什么好消息”之列。(资料来源：人民网、新华网、中新网等)

显然，影响中国汽车行业发展的因素并不是单一的，经济增长、消费趋势、市场行情、业界心态，内外环境，都会使中国汽车行业面临机遇和挑战。要分析中国汽车行业未来的趋势究竟会怎样？应当具体分析这样一些问题：

中国汽车市场发展的状况如何？（用销售量观测）

影响中国汽车销量的主要因素是什么？（如收入、价格、费用、道路状况、政策环境等）

各种因素对汽车销量影响的性质怎样？（正、负）

各种因素影响汽车销量的具体数量关系是什么？

所得到的数量结论是否可靠？

中国汽车行业今后的发展前景怎样？应当如何制定汽车的产业政策？

很明显，简单线性回归模型不能解决这类多因素问题的分析，还需要进一步寻求有多个解释变量情况的回归分析方法。

简单线性回归模型主要讨论一个被解释变量和一个解释变量之间的线性关系，但是，由于实际经济问题的复杂性，一个经济变量可能会同多个变量相联系。例如，消费者对某种商品的需求量不仅受该种商品价格的影响，而且还可能受消费者的收入水平、其他代用商品的价格等因素的影响；又如，影响一个国家货币需求量的不仅有经济总量 GDP，而且还有利率、物价水平、外汇储备等多种因素。因此，有必要将只有一个解释变量的一元回归模型推广到有多个解释变量的情况。本章将把上一章讨论的结论推广到包含多个解释变量的多元回归模型。

第一节 多元线性回归模型及古典假定

一、多元线性回归模型

社会经济现象是复杂的，通常一种社会经济现象总是和许多种现象相联系。一种社会经济现象与多种现象相联系的最简单形式，是一个被解释变量与多个解释变量的线性关系。例如，在生产理论中，著名的 Cobb-Douglas 生产函数描述了产出量与投入要素之间的关系，其形式为

$$Y = AK^{\alpha}L^{\beta}u \quad (3.1)$$

其中 Y 表示产出量，K、L 分别表示资本和劳动投入， α 、 β 为参数，u 为随机误差项。只是这里的被解释变量 Y 与解释变量 K、L 之间的关系是非线性的，但通过对数变换后可转化为如下形式

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L + \ln u \quad (3.2)$$

如果将 $\ln Y$ 视为被解释变量，将 $\ln K$ 和 $\ln L$ 视为解释变量， $\ln u$ 是随机误差项，该式关于参数 $\ln A$ 、 α 、 β 是线性的。

又如，为了对西部大开发中的电力供应作好安排，研究西部地区各省区电力消费的变化与各地区国内生产总值（GDP）及电力价格水平变动等因素的关系，这时的解释变量已不止一个，可建立如下计量经济模型：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (3.3)$$

其中： Y_i 为西部地区各省区电力消费量； X_2 为西部地区各省区国内生产总值（GDP）； X_3 为各地区电力价格变动， u_i 为随机误差项。

在计量经济学中，如果总体回归函数描述了一个被解释变量与多个解释变量之间的线性关系，由此而设定的总体回归函数就是多元线性回归模型。与一元线性回归模型类似，所谓多元线性回归模型是指对各个回归参数而言是线性的，而对于变量则可以是线性的，也可以不是线性的。

一般地，包含被解释变量 Y 与 $k-1$ 个解释变量 X_2, X_3, \dots, X_k 的多元总体线性回归函数的形式为：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i \quad (3.4)$$

其中 $\beta_j (j=1, 2, \dots, k)$ 为模型的参数； u_i 为随机误差项； $k-1$ 为解释变量的个数。

如果对被解释变量 Y 及解释变量 X_2, X_3, \dots, X_k 作了 n 次观测，所得的 n 组观测值 $(Y_i, X_{2i}, X_{3i}, \dots, X_{ki}) (i=1, 2, \dots, n)$ 将都满足如下线性关系

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i \quad (i=1, 2, \dots, n) \quad (3.5)$$

显然，多元总体线性回归函数的条件均值形式为：

$$E(Y|X_{2i}, X_{3i}, \dots, X_{ki}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} \quad (3.6)$$

多元线性回归模型与简单线性回归模型基本类似，只不过解释变量由一个增加到了多个。但是对于 Y_i 和 $X_{2i}, X_{3i}, \dots, X_{ki}$ 的观测值，已不能再像简单线性回归那样用二维平面坐标的散布图来表现了。由于多个解释变量会同时对被解释变量 Y 的变动发挥作用，因此，如果要考察其中某个解释变量对 Y 的影响，就必须使其它解释变量保持不变。在多元线性回归模型中，回归系数 $\beta_j (j=1, 2, \dots, k)$ 表示的是当控制其它解释变量不变的条件下，第 j 个解释变量的单位变动对被解释变量平均值的影

响，这样的回归系数称为偏回归系数。

在总体线性回归函数中，各个回归系数是未知的，只能利用样本观测值对之进行估计。如果将被解释变量的样本条件均值 \hat{Y}_i 表示为各个解释变量的线性函数，即得多元样本线性回归函数：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} \quad (3.7)$$

其中 $\hat{\beta}_j (j=1,2,\cdots,k)$ 是对总体回归参数 β_j 的估计。

与简单线性回归类似，多元回归中，由样本估计的被解释变量样本条件均值 \hat{Y}_i 与实际观测值 Y_i 之间通常也存在偏差，即剩余项或残差 e_i ，所以多元样本线性回归函数也可表示为：

$$Y_i = \hat{Y}_i + e_i$$

如果有 n 次样本观测值则

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + e_i \quad (i=1,2,\cdots,n) \quad (3.8)$$

多元线性回归分析要解决的主要问题，仍然是如何根据变量的样本观测值去估计回归模型中的各个参数，即要用样本回归函数去估计总体回归函数，并且对估计的参数及回归方程进行统计检验，最后利用回归模型进行预测和经济分析。只不过多元线性回归模型包含了多个解释变量，相应的分析过程及计算更为复杂。为了表达和分析的简便，对多元线性回归模型需要利用矩阵去表示和运算。在实际运用中，借助于 EViews 等计量经济软件进行运算也十分方便。

二、多元线性回归模型的矩阵形式

对被解释变量 Y 及多个解释变量作 n 次观测，所得的 n 组观测值 $(Y_i, X_{2i}, X_{3i}, \cdots X_{ki})(i=1,2,\cdots,n)$ 的线性关系，实际可写成方程组的形式

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\ &\cdots \cdots \cdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n \end{aligned} \quad (3.9)$$

这样的方程组可表示成矩阵形式

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad (3.10)$$

可以记

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix}$$

这里的 \mathbf{X} 是由解释变量 X_{ij} 的数据构成的矩阵，其中截距项可视为解释变量总是取值为 1。 \mathbf{X} 一般是非随机变量构成的，有时也称为 \mathbf{X} 的数据矩阵或设计矩阵。

这样，多元总体线性回归函数的矩阵形式可表示为

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{U} \quad (3.11)$$

或

$$E(\mathbf{Y}) = \mathbf{X} \boldsymbol{\beta} \quad (3.12)$$

类似地，多元样本线性回归函数的矩阵表示为

$$\mathbf{Y} = \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{e} \quad (3.13)$$

或

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\boldsymbol{\beta}} \quad (3.14)$$

其中，

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad \hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

分别为回归系数估计值向量、残差向量和 Y 的样本估计值向量。

三、多元线性回归模型的古典假定

在多元回归分析中，为了寻找有效的参数估计方法及对模型进行统计检验，也需要对模型中的随机扰动项和解释变量作一些假定。多元线性回归模型的基本假定条件有

1、零均值假定

假定随机扰动项的期望或均值为零

$$E(u_i) = 0 \quad i = 1, 2, \dots, n \quad (3.15)$$

用矩阵形式可表示为

$$E(\mathbf{U}) = E \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} Eu_1 \\ Eu_2 \\ \vdots \\ Eu_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.16)$$

2、同方差和无自相关假定

假定随机扰动项互不相关且方差相同

$$\begin{aligned} \text{Cov}(u_i, u_k) &= E[(u_i - Eu_i)(u_k - Eu_k)] \\ &= E(u_i u_k) = \begin{cases} \sigma^2, & i = k \\ 0, & i \neq k \end{cases} \quad (i, k = 1, 2, \dots, n) \end{aligned} \quad (3.17)$$

也就是说，随机扰动项的方差—协方差矩阵为

$$\begin{aligned} \text{Var}(\mathbf{U}) &= E[(\mathbf{U} - E\mathbf{U})(\mathbf{U} - E\mathbf{U})'] = E(\mathbf{U}\mathbf{U}') \\ &= \begin{bmatrix} E(u_1 u_1) & E(u_1 u_2) & \cdots & E(u_1 u_n) \\ E(u_2 u_1) & E(u_2 u_2) & \cdots & E(u_2 u_n) \\ \vdots & \vdots & \cdots & \vdots \\ E(u_n u_1) & E(u_n u_2) & \cdots & E(u_n u_n) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \end{aligned}$$

$$\text{即} \quad \text{Var}(\mathbf{U}) = \sigma^2 \mathbf{I}_n \quad (3.18)$$

其中 \mathbf{I}_n 为 n 阶单位阵。

3、随机扰动项与解释变量不相关假定

即假定

$$\text{Cov}(X_{ji}, u_i) = 0 \quad (j = 2, 3, \dots, k; i = 1, 2, \dots, n) \quad (3.19)$$

4、无多重共线性假定

假定各解释变量之间不存在线性关系，或者说各解释变量的观测值之间线性无关；在此条件下，解释变量观测值矩阵 \mathbf{X} 列满秩

$$\text{Rank}(\mathbf{X}) = k \quad (3.20)$$

此时，方阵 $\mathbf{X}'\mathbf{X}$ 满秩

$$\text{Rank}(\mathbf{X}'\mathbf{X}) = k \quad (3.21)$$

从而 $\mathbf{X}'\mathbf{X}$ 可逆， $(\mathbf{X}'\mathbf{X})^{-1}$ 存在。

5、正态性假定

假定随机扰动项 u_i 服从正态分布，即

$$u_i \sim N(0, \sigma^2) \quad (3.22)$$

上述这些假定条件称为多元线性回归模型的古典假定。在实际经济问题中，这些假定条件有时可能并不成立。如何识别这些假定条件是否满足，以及假定条件不成立时如何进行参数估计和检验，将在后面几章中讨论。

第二节 多元线性回归模型的估计

在对模型作出古典假定的基础上，即可对多元线性回归模型的参数加以估计，并分析参数估计式的统计性质。

一、多元线性回归性参数的最小二乘估计

与简单线性回归模型参数的估计类似，多元线性回归模型也需要用样本信息建立的样本回归函数尽可能“接近”地去估计总体回归函数。按最小二乘准则，采用使估计的剩余平方和最小的原则去确定样本回归函数。

设 $(Y_i, X_{2i}, X_{3i}, \dots, X_{ki})$ 为第 i 次观测样本 $(i = 1, 2, \dots, n)$ ，由 (3.8) 式，残差为

$$e_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki}) \quad (3.23)$$

要使残差平方和

$$\sum e_i^2 = \sum [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki})]^2$$

达到最小，其必要条件是

$$\frac{\partial(\sum e_i^2)}{\partial \hat{\beta}_j} = 0 \quad (j=1, 2, \dots, k) \quad (3.24)$$

即

$$\begin{aligned}
-2 \sum [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki})] &= 0 \\
-2 \sum X_{2i} [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki})] &= 0 \\
&\vdots \\
-2 \sum X_{ki} [Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki})] &= 0
\end{aligned}$$

注意上述各式中方括号内的各项恰好为残差 e_i ，从而上述 k 个方程可写成如下形式

$$\begin{bmatrix} \sum e_i \\ \sum X_{2i} e_i \\ \vdots \\ \sum X_{ki} e_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

或者表示为

$$\begin{bmatrix} \sum e_i \\ \sum X_{2i}e_i \\ \vdots \\ \sum X_{ki}e_i \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = X'e = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.23)$$

对样本回归函数 (3.14) 式两边同乘以样本观测值矩阵 \mathbf{X} 的转置矩阵 \mathbf{X}' , 有

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} \quad (3.24)$$

由极值条件 (3.23) 式, 可得正规方程组

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \quad (3.25)$$

由古典假定条件中的无多重共线性假定，可知 $(\mathbf{X}'\mathbf{X})^{-1}$ 存在，用 $(\mathbf{X}'\mathbf{X})^{-1}$ 左乘上述方程两端，得多元线性回归模型参数向量 β 最小二乘估计式的矩阵表达式为

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.26)$$

对于只有两个解释变量的线性回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, 其参数最小二乘估计式的代数表达式为

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (3.27)$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (3.28)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 \quad (3.29)$$

其中, $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$ 。

【例 3.1】从《中国统计年鉴》中取得西部各地区 2002 年“电力消费量”、“国内生产总值 (GDP)、电力价格变动（以“水电燃料价格指数”代表）等数据作为样本，列于表 3.1 中：

表 3.1 2002 年西部各地区电力消费等数据

地 区	电力消费量 (亿千瓦小时) Y	国内生产总值 (亿元) X_2	水电燃料价格指数 (%) X_3
内蒙古	320.43	1734.31	104.7
广西	356.95	2455.36	101.7
重庆	248.01	1971.30	109.0
四川	660.51	4875.12	103.4
贵州	366.63	1185.04	99.3
云南	353.20	2232.32	102.9
陕西	355.97	2035.96	103.2
甘肃	339.66	1161.43	102.6
青海	125.51	341.11	107.3
宁夏	178.76	329.28	105.2
新疆	214.60	1598.28	109.6

注：由于某些数据缺失，未列入西藏，但不影响样本的代表性。

根据所建模型（3.3）式，被解释变量观测值向量和解释变量数据矩阵分别为

$$\mathbf{Y} = \begin{bmatrix} 320.43 \\ 356.95 \\ \vdots \\ 214.60 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 1734.31 & 104.7 \\ 1 & 2455.36 & 101.7 \\ \vdots & \vdots & \vdots \\ 1 & 1598.28 & 109.6 \end{bmatrix}$$

将有关数据代入（3.26）式，得到

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1941.837 \\ 0.0936 \\ -17.1507 \end{bmatrix}$$

所估计的样本回归模型为

$$\hat{Y}_i = 1941.738 + 0.0936X_{2i} - 17.1507X_{3i}$$

二、参数最小二乘估计的性质

从（3.26）式可以看出，参数的最小二乘估计是样本观测值的函数，因此，参数估计量是随抽样而变化的随机变量，当我们将具体的样本观测值代入时，就可得到参数的估计值。

类似于简单线性回归，在模型古典假定成立的情况下，多元线性回归模型参数的最小二乘估计也具有线性性、无偏性与最小方差性等优良性质。

1、线性性质

最小二乘估计的参数估计量是被解释变量观测值 Y_i 的线性组合。由（3.26）式可以看出， $\hat{\beta}$ 等于取固定值的解释变量构成的 $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 与被解释变量观测值列向量 \mathbf{Y} 的乘积，从而 $\hat{\beta}_j (j=1,2,\dots,k)$ 为 Y_i 的线性函数。

2、无偏性

尽管参数估计量会随抽样波动而取不同的值，但其均值等于总体参数。即

$$E(\hat{\beta}) = \begin{bmatrix} E\hat{\beta}_1 \\ E\hat{\beta}_2 \\ \vdots \\ E\hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \beta \quad (3.30)$$

无偏性的证明较为繁琐，可参阅本章附录 3.1。

3、最小方差性

参数向量 β 的最小二乘估计 $\hat{\beta}$ 是 β 的所有线性无偏估计量中方差最小的估计量。最小方差性的证明较为繁琐，可参阅本章附录 3.2。

这就是说,在古典假定都满足的条件下，多元线性回归模型的最小二乘估计式也是最佳线性无偏估计式(BLUE)。

三、OLS 估计的分布性质

在多元线性回归中，各个参数的估计式 $\hat{\beta}$ 是随样本观测值而变动的随机变量，必须要确定其分布性质，才可能进行区间估计和假设检验。

根据正态性假定， u_i 是服从正态分布的，这决定了 \mathbf{Y} 也是服从正态分布的随机变量。由于最小

二乘估计的线性性质， $\hat{\beta}_j (j=1,2,\dots,k)$ 是 Y_i 的线性函数，这决定了 $\hat{\beta}$ 也是服从正态分布的随机变量。

由最小二乘估计的无偏性性质，已知 $E(\hat{\beta}) = \beta$

可以证明， $\hat{\beta}$ 的方差—协方差矩阵为

$$Var - Cov(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \quad (3.31)$$

$$Var(\hat{\beta}_j) = \sigma^2 c_{jj} \quad (3.32)$$

$$SE(\hat{\beta}_j) = \sigma \sqrt{c_{jj}} \quad (3.33)$$

其中 C_{jj} 是矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 中第 j 行第 j 列位置上的元素。

也就是说，在古典假定下， $\hat{\beta}_j (j=1,2,\dots,k)$ 服从正态分布。即

$$\hat{\beta}_j \sim N[\beta_j, Var(\hat{\beta}_j)] \quad (3.34)$$

三、随机扰动项方差的估计

参数估计量的方差或标准差是衡量参数估计量接近真实参数的重要指标，据此可以判断参数估计量的可靠性。但在参数估计量方差的表达式 (3.29) 中，随机扰动项的方差 σ^2 是未知的，参数估计量方差实际上无法直接计算。为此，需要对 σ^2 进行估计。

根据 (3.13) 式，当得到参数估计值以后，即可计算残差向量

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$$

据此可得残差平方和

$$\sum e_i^2 = \mathbf{e}'\mathbf{e}$$

可以证明，残差平方和具有如下性质（证明见本章附录 3.3）

$$E(\sum e_i^2) = E(\mathbf{e}'\mathbf{e}) = [n-k]\sigma^2$$

即

$$E\left(\frac{\sum e_i^2}{n-k}\right) = \sigma^2 \quad (3.35)$$

若记

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \quad (3.36)$$

则 $\hat{\sigma}^2$ 就是随机扰动项方差 σ^2 的无偏估计。一般地，称 $\hat{\sigma}^2$ 为残差的方差， $\hat{\sigma}$ 为残差的标准差。于是，参数估计量 $\hat{\beta}_j (j=1,2,\dots,k)$ 的方差 $Var(\hat{\beta}_j)$ 就可借助 $\hat{\sigma}$ 来估计，从而有如下估计式

$$Var(\hat{\beta}_j) = \hat{\sigma}^2 C_{jj} = \left(\frac{\sum e_i^2}{n-k} \right) C_{jj} \quad (3.37)$$

例如,对于【例 3.1】可计算出

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum e_i^2}{n-k} = \frac{10428.85}{11-3} = 1303.6063 \\ Var(\hat{\beta}_2) &= \left(\frac{\sum e_i^2}{n-k} \right) C_{22} = \frac{10428.85}{11-3} \times 0.000000067767 = 0.00008834 \end{aligned}$$

同理,可计算出 $Var(\hat{\beta}_1) = 152271.1801$, $Var(\hat{\beta}_3) = 12.6977$ 。

四、多元线性回归模型参数的区间估计

为了说明参数真实值的可能范围和可靠性，还需要在对参数点估计的基础上对多元线性回归模型参数作区间估计。

当用 (3.33) 式对随机扰动项的方差 σ^2 作出估计以后，用 $\hat{\sigma}^2$ 替代 σ^2 ，可以证明

$$t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{c_{jj}}} \sim t(n-k) \quad (3.38)$$

给定 α ，查 t 分布表的自由度为 $n-k$ 的临界值 $t_{\alpha/2}(n-k)$ ，则有

$$P[-t_{\alpha/2}(n-k) \leq t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \leq t_{\alpha/2}(n-k)] = 1 - \alpha \quad (j=1 \dots k)$$

$$\text{即} \quad P[\hat{\beta}_j - t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}}] = 1 - \alpha \quad (3.39)$$

$$\text{或} \quad P[\hat{\beta}_j - t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}}] = 1 - \alpha$$

这就是多元线性回归模型参数的置信度为 $1-\alpha$ 的置信区间。

例如，对于【例 3.1】的参数,前面已估计出： $\hat{\beta}_2 = 0.0936$ ， $\hat{\beta}_3 = -17.15069$ ，而且可计算得 $\sqrt{C_{22}} = \sqrt{0.000000067768} = 0.00026$ ， $\sqrt{C_{33}} = \sqrt{0.00974} = 0.09869$ ，并且可得

$$P[\hat{\beta}_2 - t_{\alpha/2} \hat{\sigma} \sqrt{c_{22}} \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \hat{\sigma} \sqrt{c_{22}}] = 1 - \alpha$$

$\hat{\sigma} = \sqrt{1303.6063} = 36.1055$ 。若取 $\alpha = 0.05$ ，查 t 分布表得 $t_{0.025}(8) = 2.306$ ，则有

$$P[0.0936 - 2.306 \times 36.1055 \times 0.00026 \leq \beta_j \leq 0.0936 + 2.306 \times 36.1055 \times 0.00026] = 1 - 0.05$$

即 $P(0.07196 \leq \beta_2 \leq 0.11765) = 0.95$

同理得

$$P[\hat{\beta}_3 - t_{\alpha/2} \hat{\sigma} \sqrt{c_{33}} \leq \beta_3 \leq \hat{\beta}_3 + t_{\alpha/2} \hat{\sigma} \sqrt{c_{33}}] = 1 - \alpha$$

$$P[(-17.1507) - 2.306 \times 36.1055 \times 0.09869 \leq \beta_3 \leq (-17.1507) + 2.306 \times 36.1055 \times 0.09869] = 1 - 0.05$$

即 $P(-25.3676 \leq \beta_3 \leq -8.9338) = 0.95$

第三节 多元线性回归模型的检验

对已经估计出参数的多元线性回归模型的检验，除了对假定条件是否满足的检验以外，主要是所估计的模型拟合优度的检验、模型中各个参数显著性的检验以及整个回归方程显著性的检验。

一、拟合优度检验

在简单线性回归模型中，我们用可决系数 r^2 来衡量估计的模型对观测值的拟合程度。在多元线性回归模型中，我们也需要讨论所估计的模型对观测值的拟合程度。

(一) 多重可决系数

与简单线性回归类似，为了说明多元线性回归线对样本观测值的拟合情况，也可以考察在 Y 的总变差中由多个解释变量作出了解释的那部分变差的比重，即“回归平方和”与“总离差平方和”的比值。在多元回归中这一比值称为多重可决系数，用 R^2 表示。

多元线性回归中 Y 的变差分解式为

$$\text{变差} \quad \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \quad (3.40)$$

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

$$(\text{总离差平方和}) = (\text{残差平方和}) + (\text{回归平方和})$$

$$\text{自由度} \quad n-1 = n-k + k-1$$

其中，总离差平方和 TSS 反映了被解释变量观测值总变差的大小；回归平方和 ESS 反映了被解释变量回归估计值总变差的大小，它是被解释变量观测值总变差中由多个解释变量作出解释的那部分变差；残差平方和 RSS 反映了被解释变量观测值与估计值之间的总变差，是被解释变量观测值总变差中未被列入模型的解释变量解释的那部分变差。显然，回归平方和 ESS 越大，残差平方和 RSS 就越小，从而被解释变量观测值总变差中能由解释变量解释的那部分变差就越大，模型对观测数据的拟合程度就越高。因此我们定义多重可决系数为

$$R^2 = \frac{ESS}{TSS} \quad (3.41)$$

或者表示为

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} \quad (3.42)$$

多重可决系数是介于 0 和 1 之间的一个数， R^2 越接近 1，模型对数据的拟合程度就越好。

多重可决系数可用矩阵去表示，因为

$$TSS = \mathbf{Y}'\mathbf{Y} - N\bar{Y}^2 \quad (3.43)$$

$$ESS = \hat{\beta}'\mathbf{X}'\mathbf{Y} - N\bar{Y}^2 \quad (3.44)$$

所以

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}'\mathbf{X}'\mathbf{Y} - N\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - N\bar{Y}^2} \quad (3.45)$$

（二）修正的可决系数

由(3.43)式容易证明，多重可决系数还可表示为

$$R^2 = \frac{\hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i + \cdots + \hat{\beta}_k \sum x_{ki} y_i}{\sum y_i^2} \quad (3.46)$$

(3.46) 式表明，多重可决系数是模型中解释变量个数的不减函数，也就是说，随着模型中解释变量的增加，多重可决系数 R^2 的值会变大。当被解释变量相同而解释变量个数不同时，这给运用多重可决系数去比较两个模型的拟合程度会带来缺陷。这时模型的解释变量个数不同，不能简单地

直接对比多重可决系数。可决系数只涉及到变差，没有考虑自由度¹。显然，如果用自由度去校正所计算的变差，可以纠正解释变量个数不同引起的对比困难。因为在样本容量一定的情况下，增加解释变量必定使得待估参数的个数增加，从而会损失自由度。为此，可以用自由度去修正多重可决系数 R^2 中的残差平方和与回归平方和，从而引入修正的可决系数 \bar{R}^2 (adjusted coefficient of determination)，其计算公式为

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n-k)}{\sum (Y_i - \bar{Y})^2 / (n-1)} = 1 - \frac{n-1}{n-k} \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} \quad (3.47)$$

修正可决系数与未经修正的多重可决系数之间有如下关系：

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k} \quad (3.48)$$

由（3.48）式可以看出，当 $k > 1$ 时， $\bar{R}^2 < R^2$ ，这意味着随着解释变量的增加， \bar{R}^2 将小于 R^2 。需要注意，可决系数 R^2 必定非负，但按（3.48）式计算的修正的可决系数 \bar{R}^2 可能为负值，这时规定 $\bar{R}^2 = 0$ 。

例如，对于【例 3.1】已知 $RSS = \sum e_i^2 = 10428.85$ ，可计算得 $TSS = \sum (Y_i - \bar{Y})^2 = 196400.19$ ，所以可决系数为

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{10428.85}{196400.19} = 0.9469$$

修正的可决系数为

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k} = 1 - (1 - 0.9469) \times \frac{11-1}{11-3} = 0.9336$$

需要强调，对用样本估计的回归模型计算的可决系数和修正的可决系数，也是随抽样而变动的随机边量，这样度量的拟合优度的显著性，还需要进行检验。

在实际的计量经济分析中，往往希望所建立模型的 R^2 或 \bar{R}^2 越大越好。但应明确，可决系数只是对模型拟合优度的度量， R^2 和 \bar{R}^2 越大，只是说明列入模型中的解释变量对被解释变量的联合影响程度越大，并非说明模型中各个解释变量对被解释变量的影响程度也大。在回归分析中，不仅要模型的拟合程度高，而且还要得到总体回归系数的可靠估计量。因此，在选择模型时，不能单纯地凭可决系数的高低断定模型的优劣，有时为了通盘考虑模型的可靠度及其经济意义，可以适当降低对可决系数的要求。

¹统计量的自由度指可自由变化的样本观测值个数，它等于所用样本观测值的个数减去对观测值的约束个数。

二、回归方程的显著性检验（F-检验）

由于多元线性回归模型包含多个解释变量，它们同被解释变量之间是否存在显著的线性关系呢？还需进一步作出判断。也就是要对模型中被解释变量与所有解释变量之间的线性关系在总体上是否显著成立作出推断。

对回归模型整体显著性的检验，所检验假设的形式为

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_1: \beta_j (j = 2, 3, \cdots, k) \text{ 不全为零}$$

这种检验是在方差分析的基础上利用 F 检验进行的。如前所述，被解释变量 Y 观测值的总变差有（3.40）式的分解形式，将自由度考虑进去进行方差分析，可得如下方差分析表：

表 3.2 方差分析表

变差来源	平方和	自由度	方差
源于回归	$ESS = \sum (\hat{Y}_i - \bar{Y})^2$	k-1	ESS/(k-1)
源于残差	$RSS = \sum (Y_i - \hat{Y}_i)^2$	n-k	RSS/(n-k)
总变差	$TSS = \sum (Y_i - \bar{Y})^2$	n-1	

可以证明，在 H_0 成立的条件下，统计量

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F(k-1, n-k) \quad (3.49)$$

即统计量 F 服从自由度为 p 和 n-k 的 F 分布。

给定显著性水平 α ，在 F 分布表中查出自由度为 k-1 和 n-k 的临界值 $F_\alpha(k-1, n-k)$ ，将样本观测值代入（3.50）式计算 F 值，然后将 F 值与临界值 $F_\alpha(k-1, n-k)$ 比较。若 $F > F_\alpha(k-1, n-k)$ ，则拒绝原假设 $H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$ ，说明回归方程显著，即列入模型的各个解释变量联合起来对被解释变量有显著影响；若 $F < F_\alpha(k-1, n-k)$ ，则不能拒绝原假设 $H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$ ，说明回归方程不显著，即列入模型的各个解释变量联合起来对被解释变量的影响不显著。

例如，【例 3.1】已计算得 $TSS = \sum (Y_i - \bar{Y})^2 = 196400.19$ ， $RSS = \sum e_i^2 = 10428.85$ ，则有 $ESS = 185971.34$ 。对于 $H_0: \beta_2 = \beta_3 = 0$ ，给定显著性水平 $\alpha = 0.05$ ，在 F 分布表中查出自由度

为 $k-1=3-1$ 和 $n-k=11-3$ 的临界值 $F_{0.05}(2,8)=4.46$ ，计算 F 统计量为

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{185971.34/(3-1)}{10428.85/(11-3)} = 71.32$$

由于 $F=71.32 > F_{0.05}(2,8)=4.46$ ，说明回归方程是显著的，即列入模型的解释变量“国内生产总值”和“水电燃料价格指数”联合起来对被解释变量“电力消费量”有显著影响。

需要指出的是，在一元线性回归中，由于解释变量只有一个，不存在解释变量联合影响的整体检验问题，也就用不着进行 F 检验。事实上，在一元回归情形下，F 检验与 t 检验是一致的，它们之间存在如下关系：

$$\begin{aligned} F &= \frac{ESS/(2-1)}{RSS/(n-2)} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum e_i^2/(n-2)} = \frac{\sum \hat{\beta}_1^2 (X_i - \bar{X})^2}{\sum e_i^2/(n-2)} \\ &= \frac{\hat{\beta}_1^2 \sum x_i^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2 / \sum x_i^2} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = t^2 \end{aligned} \quad (3.50)$$

即 F 统计量等于 t 统计量的平方。给定显著性水平 α ，查 $F_\alpha(1, n-2)$ 与 $t_{\alpha/2}(n-2)$ ，临界值之间也存在这种平方关系。也就是说，在一元回归情形下，对参数 β_1 的显著性检验（t 检验）与对回归总体线性关系的显著性检验（F 检验）是等价的。

由方差分析可以看出，F 检验与可决系数有密切联系。事实上，F 检验与拟合优度检验都是在把总变差 TSS 分解为回归平方和 ESS 与残差平方和 RSS 的基础上构造统计量进行的检验，区别在于前者考虑了自由度，后者未考虑自由度。一般来说，模型对观测值的拟合程度越高，模型总体线性关系的显著性就越强。F 统计量与可决系数 R^2 之间有如下关系：

$$F = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2} \quad (3.51)$$

可以看出，伴随着可决系数 R^2 和修正可决系数 \bar{R}^2 的增加，F 统计量的值将不断增加。当 $R^2=0$ 时， $F=0$ ；当 R^2 越大时，F 值也越大；当 $R^2=1$ 时， $F \rightarrow \infty$ 。这说明两者之间具有一致性，对 $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ 的检验，实际等价于对 $R^2 = 0$ 的检验。也就是说，对方程联合显著性检验的 F 检验，实际上也是对 R^2 的显著性检验。可决系数和修正可决系数只能提供对拟合优度的度量，它们的值究竟要达到多大才算模型通过了检验呢？并没有给出确定的界限。而 F 检验则不同，

它可以在给定显著性水平下，给出统计意义上严格的结论。

三、回归参数的显著性检验（t-检验）

多元线性回归分析的目的，不仅是获得较高拟合优度的模型，也不仅是要寻求方程整体的显著性，而是要对各个总体回归参数作出有意义的估计。因为方程的整体线性关系显著并不一定表示每个解释变量对被解释变量的影响都是显著的。因此，还必须分别对每个解释变量进行显著性检验。多元回归分析中对各个回归系数的显著性检验，目的在于分别检验当其它解释变量不变时，该回归系数对应的解释变量是否对被解释变量有显著影响。检验方法与简单线性回归的检验基本相同。

由参数估计量的性质（3.31）式已知，回归系数的估计量服从如下正态分布

$$\hat{\beta}_j \sim N[\beta_j, \text{Var}(\hat{\beta}_j)] \quad (\text{见 3.34})$$

因此其标准化随机变量服从标准正态分布

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \sim N(0,1) \quad (3.52)$$

由（3.33）式已知 $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$ ，而 σ^2 未知，故 $\text{Var}(\hat{\beta}_j)$ 也未知。但正如前面已经讨论过的，

可用 $\hat{\sigma}^2$ 代替 σ^2 对 $\hat{\beta}_j$ 作标准化变换，可以证明所构造的统计量服从自由度为 $n-k$ 的 t 分布，即

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{C_{jj}}} \sim t(n-k) \quad (3.53)$$

这样，就可以用 t 统计量对各个回归参数的显著性检验。具体过程如下：

1、提出检验假设

$$H_0: \beta_j = 0 \quad (j=1,2,\dots,k)$$

$$H_1: \beta_j \neq 0 \quad (j=1,2,\dots,k)$$

2、计算统计量

在 H_0 成立的条件下，（3.48）式变为

$$t = \frac{\hat{\beta}_j - 0}{\hat{\sigma} \sqrt{c_{jj}}} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{c_{jj}}} \sim t(n-k) \quad (3.54)$$

根据样本观测值计算 t 统计量的值

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{C_{jj}}} \quad (3.55)$$

3、检验

给定显著性水平 α ，查自由度为 $n-k$ 的 t 分布表，得临界值 $t_{\alpha/2}(n-k)$ 。

若 $|t| \geq t_{\alpha/2}(n-k)$ ，即 $-t_{\alpha/2}(n-k) \leq t^* \leq t_{\alpha/2}(n-k)$ ，就拒绝 H_0 ，而不拒绝 H_1 ，说明在其它解释变量不变的情况下，解释变量 X_j 对被解释变量 Y 的影响是显著的。

若 $|t| < t_{\alpha/2}(n-k)$ ，即 $t^* < -t_{\alpha/2}(n-k)$ 或 $t^* > t_{\alpha/2}(n-k)$ ，就不能拒绝 H_0 ，说明在其它解释变量不变的情况下，解释变量 X_j 对被解释变量 Y 的影响不显著。

从 t 分布表可以看出，在给定显著性水平 $\alpha = 0.05$ 的情况下，当自由度大于 10 时，临界值 $t_{\alpha/2}$ 基本上都接近 2。因此，当系数估计的 t 统计量超过 2 时，我们可以粗略作出判断，在显著性水平 0.05 下可拒绝原假设 H_0 ，认为相应解释变量对被解释变量的影响是显著的，此时犯错误的概率不超过 0.05。如果系数估计的 t 统计值远大于 2，则犯错误的概率更小。

例如，对于【例 3.1】，分别对于 $H_0: \beta_1 = 0$ 、 $H_0: \beta_2 = 0$ 、 $H_0: \beta_3 = 0$ ，给定显著性水平 $\alpha = 0.05$ ，查自由度为 $n-k=11-3=8$ 的 t 分布表，得临界值 $t_{0.025}(8) = 2.306$ 。已知 $\hat{\sigma} = 36.1055$ ，可分别计算

$$\begin{aligned} t_{\beta_1} &= \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{C_{11}}} = \frac{1941.837}{36.1055 \times 10.8075} = 4.9763 \\ t_{\beta_2} &= \frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{C_{22}}} = \frac{0.0936}{36.1055 \times 0.00026} = 9.9708 \\ t_{\beta_3} &= \frac{\hat{\beta}_3}{\hat{\sigma}\sqrt{C_{33}}} = \frac{-17.1507}{36.1055 \times 0.09869} = -4.8132 \end{aligned}$$

可以看出，由样本数据计算的三个回归系数的 t 统计量的绝对值均大于临界值 $t_{0.025}(8) = 2.306$ ，可以分别拒绝各个 H_0 ，而接受 H_1 ，说明在其它解释变量不变的情况下，解释变量“国内生产总值”、“水电燃料价格指数”分别对被解释变量 Y 的影响都是显著的。

一般来说，多元线性回归模型在经过参数估计和模型检验后，应对回归分析结果作出分析判断。倘若某个解释变量对被解释变量的影响不显著，则应在模型中剔除该解释变量，此时多元线性

回归模型应重新建立，且寻求新模型参数的估计及对新模型进行假设检验，直到获得较为满意的模型为止。

第四节 多元线性回归模型的预测

多元线性回归模型用于经济预测，是指对各个解释变量给定样本以外数值 $\mathbf{X}_f = (1, X_{2f}, X_{3f}, \dots, X_{kf})$ 的条件下，对预测期被解释变量 Y 的平均值 $E(Y_f)$ 及个别值 Y_f 进行估计，这种预测也分为点预测与区间预测。

一、点预测

设多元线性回归模型为

$$Y = \mathbf{X}\boldsymbol{\beta} + U$$

若根据观测样本已经估计出参数 $\boldsymbol{\beta}$ ，得到样本回归方程且模型通过检验，即

$$\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

把样本以外各个解释变量的值表示为行向量 $\mathbf{X}_f = (1, X_{2f}, X_{3f}, \dots, X_{kf})$ ，直接代入所估计的多元样本回归函数，就可以计算出被解释变量的点预测值 \hat{Y}_f ：

$$\begin{aligned}\hat{Y}_f &= \mathbf{X}_f \hat{\boldsymbol{\beta}} \\ &= \hat{\beta}_1 + \hat{\beta}_2 X_{2f} + \hat{\beta}_3 X_{3f} + \dots + \hat{\beta}_k X_{kf}\end{aligned}\quad (3.56)$$

对 (3.56) 式两边取期望得

$$\begin{aligned}E(\hat{Y}_f) &= E(\hat{\beta}_1 + \hat{\beta}_2 X_{2f} + \hat{\beta}_3 X_{3f} + \dots + \hat{\beta}_k X_{kf}) \\ &= \beta_1 + \beta_2 X_{2f} + \beta_3 X_{3f} + \dots + \beta_k X_{kf} \\ &= E(Y_f)\end{aligned}\quad (3.57)$$

说明 \hat{Y}_f 是 $E(Y_f)$ 的无偏估计，从而可以用 \hat{Y}_f 作为 $E(Y_f)$ 和 Y_f 的点预测值。

例如，对于【例 3.1】，当西部地区某省区的 GDP 达到 3500 亿元，而且水电燃料价格不变（指数为 100%）时，可预测其电力消费量为：

$$\hat{Y}_f = 1941.738 + 0.0936 \times 3500 - 17.1507 \times 100 = 554.268 \text{ (亿千瓦小时)}$$

二、平均值 $E(Y_f)$ 的区间预测

为了对预测期平均值 $E(Y_f)$ 作区间预测，必须明确得到的点预测值 \hat{Y}_f 与预测期平均值 $E(Y_f)$ 的关系，并分析其概率分布性质。如果记 \hat{Y}_f 和 $E(Y_f)$ 的偏差为 w_f ，即

$$w_f = \hat{Y}_f - E(Y_f) \quad (3.58)$$

因为 \hat{Y}_f 服从正态分布， w_f 也服从正态分布，而且

$$E(w_f) = E[\hat{Y}_f - E(Y_f)] = E(\hat{Y}_f) - E(Y_f) = 0 \quad (3.59)$$

可以证明， w_f 的方差为 $\sigma^2 \mathbf{X}_f'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f$ ，即

$$w_f \sim N[0, \sigma^2 \mathbf{X}_f'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f] \quad (3.60)$$

若用 $\hat{\sigma}^2$ 代替未知的 σ^2 ，构造如下统计量

$$t = \frac{\hat{Y}_f - E(Y_f)}{SE(\hat{Y}_f)} = \frac{\hat{Y}_f - E(Y_f)}{\hat{\sigma} \sqrt{\mathbf{X}_f'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f}} \quad (3.61)$$

该统计量 t 服从自由度为 $n-k$ 的 t 分布。

给定显著性水平 α ，查自由度为 $n-k$ 的 t 分布表，可得临界值 $t_{\alpha/2}(n-k)$ 。则 Y_f 平均值 $E(Y_f)$ 的置信度为 $1-\alpha$ 的预测区间为

$$\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{\mathbf{X}_f'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f} \leq E(Y_f) \leq \hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{\mathbf{X}_f'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f} \quad (3.62)$$

三、个别值 Y_f 的区间预测

要对预测期个别值 Y_f 作区间预测，除了已经得到的点预测值 \hat{Y}_f 以外，还需要分析已知的点预测值 \hat{Y}_f 和预测期个别值 Y_f 的联系，并明确其概率分布性质。显然，与点预测值 \hat{Y}_f 和预测期个别值 Y_f 有关的是残差 e_f ：

$$e_f = Y_f - \hat{Y}_f \quad (3.63)$$

因为 Y_f 和 \hat{Y}_f 均服从正态分布， e_f 也服从正态分布，而且

$$E(e_f) = E(Y_f - \hat{Y}_f)$$

$$\begin{aligned}
&= E(\mathbf{X}_f \boldsymbol{\beta} + u_f - \mathbf{X}_f \hat{\boldsymbol{\beta}}) \\
&= \mathbf{X}_f E(\boldsymbol{\beta}) + E(u_f) - \mathbf{X}_f E(\hat{\boldsymbol{\beta}}) \\
&= 0
\end{aligned} \tag{3.64}$$

还可以证明， e_f 的方差为 $\sigma^2[1 + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f']$ ，即

$$e_f \sim N\{0, \sigma^2[1 + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f']\} \tag{3.65}$$

若用 $\hat{\sigma}^2$ 代替未知的 σ^2 ，构造如下统计量

$$t = \frac{e_f - E(e_f)}{SE(e_f)} = \frac{Y_f - \hat{Y}_f}{\hat{\sigma} \sqrt{1 + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'}} \tag{3.66}$$

则该统计量 t 服从自由度为 $n-k$ 的 t 分布。

给定显著性水平 α ，查自由度为 $n-k$ 的 t 分布表，可得临界值 $t_{\alpha/2}(n-k)$ 。则 Y_f 的置信度为 $1-\alpha$ 的预测区间为

$$\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'} \leq Y_f \leq \hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f'} \tag{3.67}$$

第五节 案例分析

【例3.2】中国税收增长的分析

一、研究的目的要求

改革开放以来，随着经济体制改革的深化和经济的快速增长，中国的财政收支状况发生很大变化，中央和地方的税收收入1978年为519.28亿元，到2002年已增长到17636.45亿元，25年间增长了33倍，平均每年增长 %。为了研究影响中国税收收入增长的主要原因，分析中央和地方税收收入的增长规律，预测中国税收未来的增长趋势，需要建立计量经济模型。

影响中国税收收入增长的因素很多，但据分析主要的因素可能有：（1）从宏观经济看，经济整体增长是税收增长的基本源泉。（2）公共财政的需求，税收收入是财政收入的主体，社会经济的发展和社会保障的完善等都对公共财政提出要求，因此对预算支出所表现的公共财政的需求对当年的税收收入可能会有一定的影响。（3）物价水平。我国的税制结构以流转税为主，以现行价格计算的GDP等指标和经营者的收入水平都与物价水平有关。（4）税收政策因素。我国自1978年以来经历了两次大的税制改革，一次是1984-1985年的国有企业利改税，另一次是1994年的全国范围内的新

税制改革。税制改革对税收会产生影响，特别是1985年税收陡增215.42%。但是第二次税制改革对税收增长速度的影响不是非常大。因此，可以从以上几个方面，分析各种因素对中国税收增长的具体影响。

二、模型设定

为了全面反映中国税收增长的全貌，选择包括中央和地方税收的“国家财政收入”中的“各项税收”（简称“税收收入”）作为被解释变量，以反映国家税收的增长；选择“国内生产总值（GDP）”作为经济整体增长水平的代表；选择中央和地方“财政支出”作为公共财政需求的代表；选择“商品零售物价指数”作为物价水平的代表。由于财税体制的改革难以量化，而且1985年以后财税体制改革对税收增长影响不是很大，可暂不考虑税制改革对税收增长的影响。所以解释变量设定为可观测的“国内生产总值”、“财政支出”、“商品零售物价指数”等变量。

从《中国统计年鉴》收集到以下数据（见表3.3）：

年份	税收收入（亿元） (Y)	国内生产总值（亿元） (X ₂)	财政支出（亿元） (X ₃)	商品零售价格指数（%） (X ₄)
1978	519.28	3624.1	1122.09	100.7
1979	537.82	4038.2	1281.79	102.0
1980	571.70	4517.8	1228.83	106.0
1981	629.89	4862.4	1138.41	102.4
1982	700.02	5294.7	1229.98	101.9
1983	775.59	5934.5	1409.52	101.5
1984	947.35	7171.0	1701.02	102.8
1985	2040.79	8964.4	2004.25	108.8
1986	2090.73	10202.2	2204.91	106.0
1987	2140.36	11962.5	2262.18	107.3
1988	2390.47	14928.3	2491.21	118.5
1989	2727.40	16909.2	2823.78	117.8
1990	2821.86	18547.9	3083.59	102.1
1991	2990.17	21617.8	3386.62	102.9
1992	3296.91	26638.1	3742.20	105.4
1993	4255.30	34634.4	4642.30	113.2
1994	5126.88	46759.4	5792.62	121.7
1995	6038.04	58478.1	6823.72	114.8
1996	6909.82	67884.6	7937.55	106.1
1997	8234.04	74462.6	9233.56	100.8
1998	9262.80	78345.2	10798.18	97.4
1999	10682.58	82067.5	13187.67	97.0
2000	12581.51	89468.1	15886.50	98.5
2001	15301.38	97314.8	18902.58	99.2
2002	17636.45	104790.6	22053.15	98.7

表 3.3 中国税收收入及相关数据

设定的线性回归模型为：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

三、估计参数

利用EViews估计模型的参数，方法是：

1、建立工作文件：启动 EViews，点击 File\New\Workfile，在对话框“Workfile Range”。在“Workfile frequency”中选择“Annual”（年度），并在“Start date”中输入开始时间“1978”，在“end date”中输入最后时间“2002”，点击“ok”，出现“Workfile UNTITLED”工作框。其中已有变量：“c”——

截距项 “resid” 一剩余项。在 “Objects” 菜单中点击 “New Objects”，在 “New Objects”对话框中选 “Group”，并在 “Name for Objects”上定义文件名，点击 “OK” 出现数据编辑窗口。

2、输入数据：点击 “Quik” 下拉菜单中的 “Empty Group”，出现 “Group”窗口数据编辑框，点第一列与 “obs” 对应的格，在命令栏输入 “Y”，点下行键 “↓”，即将该序列命名为 Y，并依此输入 Y 的数据。用同样方法在对应的列命名 X₂、X₃、X₄，并输入相应的数据。或者在 EViews 命令框直接键入“data Y X₂ X₃ X₄ …”，回车出现 “Group”窗口数据编辑框，在对应的 Y、X₂、X₃、X₄下输入响应的数据。

3、估计参数：点击 “Procs “下拉菜单中的 “Make Equation”，在出现的对话框的 “Equation Specification” 栏中键入 “Y C X₂ X₃ X₄”，在 “Estimation Settings” 栏中选择 “Least Squares” (最小二乘法)，点 “ok”，即出现回归结果：

表3.4

Dependent Variable: Y
Method: Least Squares
Date: 07/05/05 Time: 16:54
Sample: 1978 2002
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2582.791	940.6128	-2.745860	0.0121
X2	0.022067	0.005577	3.956605	0.0007
X3	0.702104	0.033236	21.12466	0.0000
X4	23.98541	8.738302	2.744859	0.0121
R-squared	0.997430	Mean dependent var	4848.366	
Adjusted R-squared	0.997063	S.D. dependent var	4870.971	
S.E. of regression	263.9599	Akaike info criterion	14.13512	
Sum squared resid	1463172.	Schwarz criterion	14.33014	
Log likelihood	-172.6890	F-statistic	2717.238	
Durbin-Watson stat	0.948542	Prob(F-statistic)	0.000000	

根据表3.4中数据，模型估计的结果为：

$$\hat{Y}_i = -2582.791 + 0.022067 X_2 + 0.702104 X_3 + 23.98541 X_4$$

(940.6128)

(0.0056)

(0.0332)

(8.7363)

t=(-2.7459)

(3.9566)

(21.1247)

(2.7449)

$R^2 = 0.9974$

$\bar{R}^2 = 0.9971$

F=2717.238

df=21

四、模型检验

1、经济意义检验

模型估计结果说明，在假定其它变量不变的情况下，当年 GDP 每增长 1 亿元，税收收入就会增长 0.02207 亿元；在假定其它变量不变的情况下，当年财政支出每增长 1 亿元，税收收入会增长 0.7021 亿元；在假定其它变量不变的情况下，当年零售商品物价指数上涨一个百分点，税收收入就会增长 23.9854 亿元。这与理论分析和经验判断相一致。

2、统计检验

(1) 拟合优度：由表 3.4 中数据可以得到： $R^2 = 0.9974$ ，修正的可决系数为 $\bar{R}^2 = 0.9971$ ，这说明模型对样本的拟合很好。

(2) F 检验：针对 $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ ，给定显著性水平 $\alpha = 0.05$ ，在 F 分布表中查出自由度为 $k-1=3$ 和 $n-k=21$ 的临界值 $F_\alpha(3,21)=3.075$ 。由表 3.4 中得到 $F=2717.238$ ，由于 $F=2717.238 > F_\alpha(3,21)=3.075$ ，应拒绝原假设 $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ ，说明回归方程显著，即“国内生产总值”、“财政支出”、“商品零售物价指数”等变量联合起来确实对“税收收入”有显著影响。

(3) t 检验：分别针对 $H_0: \beta_j = 0$ ($j=1,2,3,4$)，给定显著性水平 $\alpha = 0.05$ ，查 t 分布表得自由度为 $n-k=21$ 临界值 $t_{\alpha/2}(n-k) = 2.080$ 。由表 3.4 中数据可得，与 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 、 $\hat{\beta}_3$ 、 $\hat{\beta}_4$ 对应的 t 统计量分别为 -2.7459、3.9566、21.1247、2.7449，其绝对值均大于 $t_{\alpha/2}(n-k) = 2.080$ ，这说明分别都应当拒绝 $H_0: \beta_j = 0$ ($j=1,2,3,4$)，也就是说，当在其它解释变量不变的情况下，解释变量“国内生产总值”(X_2)、“财政支出”(X_3)、“商品零售物价指数”(X_4) 分别对被解释变量“税收收入”Y 都有显著的影响。

第三章小结

1、多元线性回归模型是将总体回归函数描述为一个被解释变量与多个解释变量之间线性关系的模型。通常多元线性回归模型可以用矩阵形式表示。

2、多元线性回归模型中对随机扰动项 u 的假定，除了零均值假定、同方差假定、无自相关假定、随机扰动与解释变量不相关假定、正态性假定以外，还要求满足无多重共线性假定。

3、多元线性回归模型参数的最小二乘估计式；参数估计式的分布性质及期望、方差和标准误差；在基本假定满足的条件下，多元线性回归模型最小二乘估计式是最佳线性无偏估计式。

- 4、多元线性回归模型中参数区间估计的方法。
- 5、多重可决系数的意义和计算方法，修正可决系数的作用和方法。
- 6、F 检验是对多元线性回归模型中所有解释变量联合显著性的检验，F 检验是在方差分析基础上进行的。
- 7、多元回归分析中，为了分别检验当其它解释变量不变时，各个解释变量是否对被解释变量有显著影响，需要分别对所估计的各个回归系数作 t 检验。
- 8、利用多元线性回归模型作被解释变量平均值预测与个别值预测的方法。

第三章主要公式表

1、多元线性回归模型	$E(Y_i X_1, X_2, \dots, X_k) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki}$ $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} \quad E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$
2、样本回归函数	$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki}$ $Y_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{X}_{2i} + \hat{\beta}_3 \hat{X}_{3i} + \dots + \hat{\beta}_k \hat{X}_{ki} + e_i$ $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} \quad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
3、基本假定	$E(\mathbf{U}) = \mathbf{0}$ $Cov(u_i, u_k) = E(u_i u_k) = \begin{cases} \sigma^2, & i = k \\ 0, & i \neq k \end{cases} \quad \text{Rank}(\mathbf{X}) = k$ $Cov(X_{ji}, u_i) = 0 \quad (j = 1, 2, \dots, k) \quad u_i \sim N(0, \sigma^2)$
4、最小二乘估计	$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$ $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$
5、参数 OLS 估计的期望	$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
6、参数 OLS 估计的方差	$Var(\hat{\beta}_j) = \hat{\sigma}^2 C_{jj} = \left(\frac{\sum e_i^2}{n-k} \right) C_{jj}$
7、参数估计的标准误差	$SE(\hat{\beta}_j) = \sigma \sqrt{C_{jj}}$
8、 σ^2 的无偏估计	$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$

9、参数估计的置信区间	$P[\hat{\beta}_j - t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} \hat{\sigma} \sqrt{c_{jj}}] = 1 - \alpha$
10、多重可决系数	$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$
11、修正的可决系数	$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n - k)}{\sum (Y_i - \bar{Y})^2 / (n - 1)} = 1 - \frac{n - 1}{n - k} \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$
12、F 检验统计量	$F = \frac{ESS / (k - 1)}{RSS / (n - k)} \sim F(k - 1, n - k)$
13、t 检验统计量	$t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{c_{jj}}} \sim t(n - k)$
14、点预测值	$\hat{Y}_f = \mathbf{X}_f \hat{\boldsymbol{\beta}}$
15、平均值预测区间	$\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{\mathbf{X}_f (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f'} \leq E(Y_f) \leq \hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{\mathbf{X}_f (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f'}$
16、个别值预测区间	$\hat{Y}_f - t_{\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{X}_f (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f'} \leq Y_f \leq \hat{Y}_f + t_{\alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{X}_f (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_f'}$

思考题与练习题

思考题

3.1 若要将一个被解释变量对两个解释变量作线性回归分析：

- (1) 写出总体回归函数和样本回归函数；
- (2) 写出回归模型的矩阵表示；
- (3) 说明对此模型的古典假定；
- (4) 写出回归系数及随机扰动项方差的最小二乘估计量，并说明参数估计量的性质；

3.2 什么是偏回归系数？它与简单线性回归的回归系数有什么不同？

3.3 多元线性回归中的古典假定与简单线性回归时有什么不同？

3.4 多元线性回归分析中，为什么要对可决系数加以修正？修正可决系数与 F 检验之间有何区别与联系？

3.5 什么是方差分析？对被解释变量的方差分析与对模型拟合优度的度量有什么联系和区别？

3.6 多元线性回归分析中，F 检验与 t 检验的关系是什么？为什么在作了 F 检验以后还要作 t 检验？

3.7 试证明：在二元线性回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 中，当 X_2 和 X_3 相互独立时，对斜率系数 β_2 和 β_3 的 OLS 估计值，等于 Y_i 分别对 X_2 和 X_3 作简单线性回归时斜率系数的 OLS 估计值。

3.8 对于本章开始提出的“中国汽车保有量会超过一亿辆吗？”，你认为可建立什么样的计量经济模型去分析？

3.9 说明用 EViews 完成多元线性回归分析的具体操作步骤。

练习题

3.1 为研究中国各地区入境旅游状况，建立了各省市旅游外汇收入（Y，百万美元）、旅行社职工人数（X1，人）、国际旅游人数（X2，万人次）的模型，用某年 31 个省市的截面数据估计结果如下：

$$\hat{Y}_i = -151.0263 + 0.1179X_{1i} + 1.5452X_{2i}$$

$$t=(-3.066806) \quad (6.652983) \quad (3.378064)$$

$$R^2=0.934331 \quad \bar{R}^2 = 0.92964 \quad F=191.1894 \quad n=31$$

- (1) 从经济意义上考察估计模型的合理性。
- (2) 在 5% 显著性水平上，分别检验参数 β_1, β_2 的显著性。
- (3) 在 5% 显著性水平上，检验模型的整体显著性。

3.2 根据下列数据试估计偏回归系数、标准误差，以及可决系数与修正的可决系数：

$$\bar{Y} = 367.693, \quad \bar{X}_1 = 402.760, \quad \bar{X}_2 = 8.0, \quad n = 15,$$

$$\sum (Y_i - \bar{Y})^2 = 66042.269, \quad \sum (X_{1i} - \bar{X}_1)^2 = 84855.096,$$

$$\sum (X_{2i} - \bar{X}_2)^2 = 280.000, \quad \sum (Y_i - \bar{Y})(X_{1i} - \bar{X}_1) = 74778.346,$$

$$\sum (Y_i - \bar{Y})(X_{2i} - \bar{X}_2) = 4250.900, \quad \sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 4796.000$$

3.3 经研究发现，家庭书刊消费受家庭收入几户主受教育年数的影响，表中为对某地区部分家庭抽样调查得到样本数据：

家庭书刊年 消费支出 (元) Y	家庭月平均 收入 (元) X	户主受教育 年数 (年) T	家庭书刊年 消费支出 (元) Y	家庭月平均 收入 (元) X	户主受教育 年数 (年) T
450	1027.2	8	793.2	1998.6	14
507.7	1045.2	9	660.8	2196	10
613.9	1225.8	12	792.7	2105.4	12
563.4	1312.2	9	580.8	2147.4	8
501.5	1316.4	7	612.7	2154	10
781.5	1442.4	15	890.8	2231.4	14
541.8	1641	9	1121	2611.8	18
611.1	1768.8	10	1094.2	3143.4	16
1222.1	1981.2	18	1253	3624.6	20

- (1) 建立家庭书刊消费的计量经济模型；
- (2) 利用样本数据估计模型的参数；
- (3) 检验户主受教育年数对家庭书刊消费是否有显著影响；
- (4) 分析所估计模型的经济意义和作用

3.4 考虑以下“期望扩充菲利普斯曲线 (Expectations-augmented Phillips curve)”模型：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

其中： Y_t =实际通货膨胀率 (%)； X_{2t} =失业率 (%)； X_{3t} =预期的通货膨胀率 (%)

下表为某国的有关数据，

表 1. 1970-1982 年某国实际通货膨胀率 $Y(\%)$ ，
失业率 $X_2(\%)$ 和预期通货膨胀率 $X_3(\%)$

年份	实际通货膨胀率 Y (%)	失业率 X2 (%)	预期的通货膨胀率 X3 (%)
1970	5.92	4.90	4.78
1971	4.30	5.90	3.84
1972	3.30	5.60	3.31

1973	6.23	4.90	3.44
1974	10.97	5.60	6.84
1975	9.14	8.50	9.47
1976	5.77	7.70	6.51
1977	6.45	7.10	5.92
1978	7.60	6.10	6.08
1979	11.47	5.80	8.09
1980	13.46	7.10	10.01
1981	10.24	7.60	10.81
1982	5.99	9.70	8.00

(1) 对此模型作估计,并作出经济学和计量经济学的说明。

(2) 根据此模型所估计结果,作计量经济学的检验。

(3) 计算修正的可决系数(写出详细计算过程)。

3.5 某地区城镇居民人均全年耐用消费品支出、人均年可支配收入及耐用消费品价格指数的统计资料如表所示:

年份	人均耐用消费品支出 Y (元)	人均年可支配收入 X1 (元)	耐用消费品价格指数 X2 (1990 年=100)
1991	137.16	1181.4	115.96
1992	124.56	1375.7	133.35
1993	107.91	1501.2	128.21
1994	102.96	1700.6	124.85
1995	125.24	2026.6	122.49
1996	162.45	2577.4	129.86
1997	217.43	3496.2	139.52
1998	253.42	4283.0	140.44
1999	251.07	4838.9	139.12
2000	285.85	5160.3	133.35
2001	327.26	5425.1	126.39

利用表中数据,建立该地区城镇居民人均全年耐用消费品支出关于人均年可支配收入和耐用消费品价格指数的回归模型,进行回归分析,并检验人均年可支配收入及耐用消费品价格指数对城镇

居民人均全年耐用消费品支出是否有显著影响。

3.6 下表给出的是 1960—1982 年间 7 个 OECD 国家的能源需求指数(Y)、实际 GDP 指数(X1)、能源价格指数(X2)的数据,所有指数均以 1970 年为基准(1970=100)

年份	能源需求 指数 Y	实际 GDP 指数 X1	能源价格 指数 X2	年份	能源需求 指数 Y	实际 GDP 指数 X1	能源价格 指数 X2
1960	54.1	54.1	111.9	1972	97.2	94.3	98.6
1961	55.4	56.4	112.4	1973	100.0	100.0	100.0
1962	58.5	59.4	111.1	1974	97.3	101.4	120.1
1963	61.7	62.1	110.2	1975	93.5	100.5	131.0
1964	63.6	65.9	109.0	1976	99.1	105.3	129.6
1965	66.8	69.5	108.3	1977	100.9	109.9	137.7
1966	70.3	73.2	105.3	1978	103.9	114.4	133.7
1967	73.5	75.7	105.4	1979	106.9	118.3	144.5
1968	78.3	79.9	104.3	1980	101.2	119.6	179.0
1969	83.3	83.8	101.7	1981	98.1	121.1	189.4
1970	88.9	86.2	97.7	1982	95.6	120.6	190.9
1971	91.8	89.8	100.3				

(1)建立能源需求与收入和价格之间的对数需求函数 $\ln Y_t = \beta_0 + \beta_1 \ln X1_t + \beta_2 \ln X2_t + u_t$, 解释各回归系数的意义, 用 P 值检验所估计回归系数是否显著。

(2) 再建立能源需求与收入和价格之间的线性回归模型 $Y_t = \beta_0 + \beta_1 X1_t + \beta_2 X2_t + u$, 解释各回归系数的意义, 用 P 值检验所估计回归系数是否显著。

(3)比较所建立的两个模型, 如果两个模型结论不同, 你将选择哪个模型, 为什么?

第三章附录

附录 3.1 多元线性回归最小二乘估计无偏性的证明

因为

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + U) \\ &= (X'X)^{-1} (X'X)\beta + (X'X)^{-1} X'U \\ &= \beta + (X'X)^{-1} X'U\end{aligned}$$

对两边取期望, $E(\hat{\beta}) = \beta + (X'X)^{-1} X'[E(U)]$

$$= \beta \quad [\text{由假定 1: } E(\mathbf{U}) = \mathbf{0}]$$

即 $\hat{\beta}$ 是 β 的无偏估计。

附录 3.2 多元线性回归最小二乘估计最小方差性的证明

设 β^* 为 β 的另一个关于 \mathbf{Y} 的线性无偏估计式，可知

$$\beta^* = \mathbf{A}\mathbf{Y} \quad (\mathbf{A} \text{ 为常数矩阵})$$

由无偏性可得 $E(\beta^*) = E(\mathbf{A}\mathbf{Y}) = E[\mathbf{A}(\mathbf{X}\beta + \mathbf{U})]$

$$= E(\mathbf{A}\mathbf{X}\beta) + \mathbf{A}E(\mathbf{U})$$

$$= \mathbf{A}\mathbf{X}E(\beta) = \beta$$

所以必须有 $\mathbf{A}\mathbf{X} = \mathbf{I}$

要证明最小二乘法估计式的方差 $\text{Var}(\hat{\beta})$ 小于其他线性无偏估计式的方差 $\text{Var}(\beta^*)$ ，只要证明协方差矩阵之差

$$E[(\beta^* - \beta)(\beta^* - \beta)'] - E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

为半正定矩阵，则称最小二乘估计 $\hat{\beta}$ 是 β 的最小方差线性无偏估计式。

因为 $\beta^* - \beta = \mathbf{A}\mathbf{Y} - \beta = \mathbf{A}(\mathbf{X}\beta + \mathbf{U}) - \beta$

$$= \mathbf{A}\mathbf{X}\beta + \mathbf{A}\mathbf{U} - \beta$$

$$= \beta + \mathbf{A}\mathbf{U} - \beta = \mathbf{A}\mathbf{U}$$

所以 $E[(\beta^* - \beta)(\beta^* - \beta)'] = E[(\mathbf{A}\mathbf{U})(\mathbf{A}\mathbf{U})'] = E(\mathbf{A}\mathbf{U}\mathbf{U}'\mathbf{A}')$

$$= \mathbf{A}E(\mathbf{U}\mathbf{U}')\mathbf{A}' = \mathbf{A}\mathbf{A}'\sigma^2$$

由于 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}$

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}]'$$

$$= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}][\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}\mathbf{U}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

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

所以

$$\begin{aligned} E[(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'] - E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] &= \mathbf{A}\mathbf{A}'\sigma^2 - (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \\ &= [\mathbf{A}\mathbf{A}' - (\mathbf{X}'\mathbf{X})^{-1}]\sigma^2 \end{aligned}$$

由于

$$\begin{aligned} [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']' &= [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] [\mathbf{A}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \mathbf{A}\mathbf{A}' - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{A}' - \mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{A}\mathbf{A}' - (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

由线性代数知，对任一非奇异矩阵 \mathbf{C} ， $\mathbf{C}\mathbf{C}'$ 为半正定矩阵。如果令 $[\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] = \mathbf{C}$

则

$$\mathbf{C}\mathbf{C}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']' = \mathbf{A}\mathbf{A}' - (\mathbf{X}'\mathbf{X})^{-1}$$

由于半正定矩阵对角线元素非负，因此有 $\mathbf{A}\mathbf{A}' - (\mathbf{X}'\mathbf{X})^{-1} \geq \mathbf{0}$

即

$$E(\beta_j^* - \beta_j)^2 - E(\hat{\beta}_j - \beta_j)^2 \geq 0 \quad (j = 1, 2, \dots, k)$$

这证明了 β_j 的最小二乘估计 $\hat{\beta}_j$ 在 β_j 的所有无偏估计中是方差最小的估计式。

附录 3.3 残差平方和 $\sum e_i^2$ 的均值为 $(n-k)\sigma^2$ 的证明

由残差向量的定义及参数的最小二乘估计式，有

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']\mathbf{Y} \end{aligned}$$

可以记 $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ ，则

$$\begin{aligned} \mathbf{e} &= \mathbf{P}\mathbf{Y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] [\mathbf{X}\boldsymbol{\beta} + \mathbf{U}] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{U} \\ &= \mathbf{P}\mathbf{U} \end{aligned}$$

容易验证， \mathbf{P} 为对称等幂矩阵，即

$$\mathbf{P} = \mathbf{P}'$$

$$\mathbf{P}^2 = \mathbf{P}\mathbf{P} = \mathbf{P}$$

残差向量的协方差矩阵为

$$\begin{aligned} \text{Var}(\mathbf{e}) &= E(\mathbf{e}\mathbf{e}') = E[\mathbf{P}\mathbf{U}(\mathbf{P}\mathbf{U})'] \\ &= E[\mathbf{P}(\mathbf{U}\mathbf{U}')\mathbf{P}'] \\ &= \mathbf{P}[E(\mathbf{U}\mathbf{U}')] \mathbf{P}' \\ &= \mathbf{P}(\sigma^2 \mathbf{I}) \mathbf{P}' \\ &= \mathbf{P}\mathbf{P}' \sigma^2 = \mathbf{P} \sigma^2 \end{aligned}$$

利用矩阵迹的性质，有

$$\sum e_i^2 = \mathbf{e}'\mathbf{e} = \text{tr}(\mathbf{e}\mathbf{e}')$$

两边取期望得

$$\begin{aligned} E(\sum e_i^2) &= E(\mathbf{e}'\mathbf{e}) = E[\text{tr}(\mathbf{e}\mathbf{e}')] \\ &= \text{tr}[E(\mathbf{e}'\mathbf{e})] = \text{tr}[\mathbf{P} \sigma^2] \\ &= \sigma^2 \text{tr}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'] \\ &= \sigma^2 \{ \text{tr}(\mathbf{I}) - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}] \} \\ &= \sigma^2 [n - \text{tr}(\mathbf{I})] \\ &= (n - k) \sigma^2 \end{aligned}$$

第四章 多重共线性

引子:

古典假定总是能够满足吗?

——对古典假定的再讨论

在第二、三章,已经讨论了在古典假定完全满足的条件下线性回归模型的估计、检验及预测问题。然而,在现实的经济问题中古典假定的条件真的总是能够满足吗?

回顾对模型中随机扰动项和解释变量作的古典假定:

(1) 零均值假定:假定随机扰动项的期望或均值为零,即 $E(u_i) = 0$;

(2) 同方差假定:对于给定的每一个 X_i ,随机扰动项 u_i 的条件方差都等于某一个常数 σ^2 , 即 $Var(u_i | X_i) = E[u_i - E(u_i | X_i)]^2 = E(u_i^2) = \sigma^2$;

(3) 无自相关假定:即随机扰动项 u_i 的逐次值互不相关,或者说对于所有的 i 和 j ($i \neq j$), u_i 和 u_j 的协方差为零,即 $Cov(u_i, u_j) = E[u_i - E(u_i)][u_j - E(u_j)] = E(u_i u_j) = 0$

(4) 解释变量非随机或随机扰动项与解释变量不相关假定:即 $Cov(X_{ji}, u_i) = 0$;

(5) 无多重共线性假定:假定各解释变量之间不存在线性关系;

(6) 正态性假定:假定随机扰动项 u_i 服从正态分布,即 $u_i \sim N(0, \sigma^2)$ 。

正是有了这些古典假定,回归系数的OLS估计量才是最佳线性无偏估计量。然而实际的经济活动异常复杂,不一定总是能满足这些假定,从而可能给计量经济分析带来一系列的麻烦和问题。

假定(1)零均值假定的违反主要会对截距项的估计产生影响,并不影响更受关注的斜率系数的估计;违反假定(4)解释变量非随机或随机扰动项与解释变量不相关的影响,将在有关的章节中再讨论;假定(6)正态性假定的违反并不影响OLS估计是最佳线性无偏估计,加之在大样本情况下 u_i 会渐近服从正态分布,可以不再讨论。除此之外需要首先作深入讨论的,是假定(5)无多重共线性、假定(2)同方差性、假定(3)无自相关。这正是第四、五、六章将讨论的主题。

农业和建筑业的发展会减少财政收入吗？

国家财政收入主要来自各项税收收入，经济增长是其重要的影响因素。为了分析各主要因素对国家财政收入的影响，建立财政收入(亿元) (CS)为被解释变量，农业增加值(亿元)(NZ)、工业增加值(亿元)(GZ)、建筑业增加值(亿元)(JZZ)、总人口(万人)(TPOP)、最终消费(亿元)(CUM)、受灾面积(万公顷)(SZM)等为解释变量的计量模型。数据样本时期为1978年-2003年共26个年份的统计数据（资料来源：《中国统计年鉴2004》，中国统计出版社2004年版）

设定的理论模型为：

$$CS_i = \beta_0 + \beta_1 NZ_i + \beta_2 GZ_i + \beta_3 JZZ_i + \beta_4 TPOP_i + \beta_5 CUM_i + \beta_6 SZM_i + u_i$$

采用普通最小二乘法得到以下估计结果

关于财政收入的多元回归结果

Variable	Coefficient	Std. Error	t-Statistic	Prob.
农业增加值	-1.535090	0.129778	-11.82861	0.0000
工业增加值	0.898788	0.245466	3.661558	0.0017
建筑业增加值	-1.527089	1.206242	-1.265989	0.2208
总人口	0.151160	0.033759	4.477646	0.0003
最终消费	0.101514	0.105329	0.963783	0.3473
受灾面积	-0.036836	0.018460	-1.995382	0.0605
截距项	-11793.34	3191.096	-3.695704	0.0015
R-squared	0.995015	Mean dependent var	5897.824	
Adjusted R-squared	0.993441	S.D. dependent var	5945.854	
S.E. of regression	481.5380	Akaike info criterion	15.41665	
Sum squared resid	4405699.	Schwarz criterion	15.75537	
Log likelihood	-193.4165	F-statistic	632.0999	
Durbin-Watson stat	1.873809	Prob(F-statistic)	0.000000	

从主要指标分析可见，可决系数为0.995，校正的可决系数为0.993，模型拟合很好。F统计量为632.10，说明在 $\alpha = 0.05$ 水平下回归方程整体上是显著的，模型对财政收入的解释程度高达99.5%。但是t检验结果表明，除了工业增加值和总人口以外，其他因素对财政收入的影响均不显著。更难理解的是农业增加值和建筑业增加值的回归系数竟然是负数，这就是说农业和建筑业的发展反而会使财政收入减少，这显然与理论分析和实践经验不相符。为什么会出现这样的异常结果？如果设定的模型和数据的真实性没有问题，问题可能会出在哪里呢？

第四章专门讨论古典假定中无多重共线性假定被违反的情况，主要内容包括多重共线性的实质和产生的原因、多重共线性产生的后果、多重共线性的检测方法及无多重共线性假定违反后的处置方法。

第一节 什么是多重共线性

一、多重共线性的含义

第三章讨论多元线性回归模型的估计时，强调了假定无多重共线性，即假定各解释变量之间不存在线性关系，或者说各解释变量的观测值之间线性无关。在计量经济学中所谓的多重共线性(Multi-Collinearity)，不仅包括解释变量之间精确的线性关系，还包括解释变量之间近似的线性关系。

从数学意义上去说明多重共线性，就是对于解释变量 X_2, X_3, \dots, X_k ，如果存在不全为0的数 $\lambda_1, \lambda_2, \dots, \lambda_k$ ，能使得

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} = 0 \quad i = 1, 2, \dots, n \quad (4.1)$$

则称解释变量 X_2, X_3, \dots, X_k 之间存在着完全的多重共线性。

用矩阵表示，解释变量的数据矩阵为

$$\mathbf{X} = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{bmatrix} \quad (4.2)$$

当 $\text{Rank}(\mathbf{X}) < k$ 时，表明在数据矩阵 \mathbf{X} 中，至少有一个列向量可以用其余的列向量线性表示，则说明存在完全的多重共线性。

在实际经济问题中，完全的多重共线性并不多见。常见的情形是解释变量 X_2, X_3, \dots, X_k 之间存在不完全的多重共线性。所谓不完全的多重共线性，是指对于解释变量 X_2, X_3, \dots, X_k ，存在不全为0的数 $\lambda_1, \lambda_2, \dots, \lambda_k$ ，使得

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_k X_{ki} + u_i = 0 \quad i = 1, 2, \dots, n \quad (4.3)$$

其中， u_i 为随机变量。这表明解释变量 X_2, X_3, \dots, X_k 只是一种近似的线性关系。

如果 k 个解释变量之间不存在完全或不完全的线性关系，则称无多重共线性。若用矩阵

表示，这时 \mathbf{X} 为满秩矩阵，即 $\text{Rank}(\mathbf{X})=k$ 。

需要强调，解释变量之间不存在线性关系，并非不存在非线性关系，当解释变量存在非线性关系时，并不违反无多重共线性假定。

回归模型中解释变量的关系可能表现为三种情形：

(1) $r_{x_i x_j} = 0$ ，解释变量间毫无线性关系，变量间相互正交。事实上这时已不需要作多元回归，每个参数 β_j 都可以通过 Y 对 X_j 的一元回归来估计。

(2) $r_{x_i x_j} = 1$ ，解释变量间完全共线性。此时模型参数将无法确定。直观地看，当两变量按同一方式变化时，要区别每个解释变量对被解释变量的影响程度非常困难。

(3) $0 < r_{x_i x_j} < 1$ ，解释变量间存在一定程度的线性关系。实际中常遇到的是这种情形。

随着共线性程度的加强，会对参数估计值的准确性、稳定性带来影响。因此不完全的多重共线性事实上是有严重程度的问题。

二、产生多重共线性的背景

由于经济现象的变化涉及多个影响因素，而影响因素之间常常存在一定的相关性。多重共线性产生的经济背景主要有几种情形：

1、经济变量之间具有共同变化趋势。例如，对于时间序列数据收入、消费、就业率等，在经济上升时期均呈现增长的趋势，而当经济收缩期，又都呈现下降趋势。当这些变量同时作为解释变量进入模型时就可能带来多重共线性问题。

2、模型中包含滞后变量。当建立的模型中引入了解释变量的滞后变量时，如， X_t, X_{t-1}, X_{t-2} ，等等，而 X 变量与其滞后期变量常常呈现高度相关性，于是导致出现多重共线性。

3、利用截面数据建立模型也可能出现多重共线性。利用截面数据建模时，许多变量变化与发展规模相关，会呈现出共同增长的趋势，例如资本、劳动力、科技、能源等投入与产出的规模相关，这时容易出现多重共线性。有时如果出现部分因素的变化与另一部分因素的变化相关程度较高时，也容易出现共线性。如粮食产量与化肥用量、水浇地面积、农业投入资金建立回归模型，发现回归效果较差，原因是农业资金的影响已经通过化肥用量、水浇地面积两个因素体现出来。

4、样本数据自身的原因。例如，抽样仅仅限于总体中解释变量取值的一个有限范围，

使得变量变异不大；或由于总体受限，多个解释变量的样本数据之间存在相关，这时都可能出现多重共线性。

第二节 多重共线性产生的后果

一、完全多重共线性产生的后果

1、参数的估计值不确定

完全共线性时， $\mathbf{X}\mathbf{X}$ 矩阵的秩小于 k ，此时 $|\mathbf{X}\mathbf{X}| = 0$ ，正规方程组的解不惟一， $(\mathbf{X}\mathbf{X})^{-1}$ 不存在，回归参数的最小二乘估计表达式不成立。

这里以两个解释变量的回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 为例，说明完全多重共线性的影响。采用离差形式表示的两个解释变量的回归模型为：

$$\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \quad (4.4)$$

由第三章(3.27)、(3.28)式得到其OLSE估计式为：

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (4.5)$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (4.6)$$

假定 $X_{2i} = \lambda X_{3i}$ ，这里 λ 是一非零常数，将其分别带入(4.5)和(4.6)式，可得：

$$\hat{\beta}_2 = \frac{(\lambda \sum y_i x_{3i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\lambda \sum x_{3i} x_{3i})}{(\lambda^2 \sum x_{3i}^2)(\sum x_{3i}^2) - \lambda^2 (\sum x_{3i} x_{3i})^2} = \frac{0}{0} \quad (4.7)$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\lambda^2 \sum x_{3i}^2) - (\lambda \sum y_i x_{3i})(\lambda \sum x_{3i}^2)}{(\lambda^2 \sum x_{3i}^2)(\sum x_{3i}^2) - \lambda^2 (\sum x_{3i}^2)^2} = \frac{0}{0} \quad (4.8)$$

(4.7)和(4.8)式都是不定式，这说明当 $X_{2i} = \lambda X_{3i}$ 时，参数的估计值是不确定的。

从回归模型的建模思想看，在回归模型中回归系数 $\hat{\beta}_2$ 的含义是指在保持 X_3 不变的情况下，当 X_2 每变动一个单位时 Y 的平均变化；回归系数 $\hat{\beta}_3$ 的含义是指在保持 X_2 不变的情况下，当 X_3 每改变一个单位时 Y 的平均变化。如果 X_2 与 X_3 完全共线性，就没有办法能在

保持 X_2 不变的情况下, 分析 X_3 对 Y 的影响。或者说, 没有办法能从所给的样本中把 X_2 和 X_3 各自的影响分解开来。

2、参数估计值的方差无限大

还是以只有两个解释变量的回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 为例, 由第三章的 (3.28) 和 (3.29) 式, 已知 OLS 估计参数的方差为 $\text{Var-Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, 容易导出这时参数估计式 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差为

$$\text{Var}(\hat{\beta}_2) = \frac{\sum x_3^2}{(\sum x_2^2)(\sum x_3^2) - (\sum x_2 x_3)^2} \sigma^2 \quad (4.9)$$

$$\text{Var}(\hat{\beta}_3) = \frac{\sum x_2^2}{(\sum x_2^2)(\sum x_3^2) - (\sum x_2 x_3)^2} \sigma^2 \quad (4.10)$$

在完全共线性情况下 $X_{2i} = \lambda X_{3i}$, 代入 (4.9) 和 (4.10) 式中得

$$\text{Var}(\hat{\beta}_2) = \frac{\sum x_3^2}{(\lambda^2 \sum x_3^2)(\sum x_3^2) - (\lambda \sum x_3 x_3)^2} \sigma^2 = \frac{\sum x_3^2}{0} \sigma^2 = \infty$$

同理
$$\text{Var}(\hat{\beta}_3) = \frac{\lambda^2 \sum x_3^2}{(\lambda^2 \sum x_3^2)(\sum x_3^2) - (\lambda \sum x_3 x_3)^2} \sigma^2 = \frac{\sum x_2^2}{0} \sigma^2 = \infty$$

这表明, 在解释变量之间存在完全的共线性时, 参数估计值的方差将变成无穷大。

二、不完全多重共线性下产生的后果

完全多重共线性的情形只不过是一种极端。通常, 解释变量之间并不一定是完全的线性关系。如果模型中存在不完全的多重共线性, 这种情况下 $(\mathbf{X}'\mathbf{X})^{-1}$ 也存在, 可以得到参数的估计值, 但是对计量经济分析可能会产生一系列的影响。

1、参数估计值的方差增大

仍然以只有两个解释变量的回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 为例, X_2 与 X_3 不完全的共线性的关系为

$$x_{2i} = \lambda x_{3i} + v_i \quad (4.11)$$

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。

这种情况下，还是可以用OLS法估计回归系数 β_2 和 β_3 。若将(4.11)代入(4.6)式得：

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\lambda^2 \sum x_{3i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{3i} + \sum y_i v_i)(\lambda \sum x_{3i}^2)}{(\lambda^2 \sum x_{3i}^2 + \sum v_i^2)(\sum x_{3i}^2) - \lambda^2 (\sum x_{3i}^2)^2} \quad (4.12)$$

其中利用了关系式 $\sum x_{2i} v_i = 0$ 。因此在 X_2 与 X_3 近似共线性时， $\hat{\beta}_3$ 还是可以估计的。但是，如果 X_2 与 X_3 共线程度越高， v_i 会充分地小，以至于非常接近于零，此时 $\hat{\beta}_3$ 会愈加趋于不确定。对于 $\hat{\beta}_2$ 也可推出类似的表达式，并得到类似的结论。

在 X_2 与 X_3 为不完全的共线性时， X_2 与 X_3 的相关系数的平方用离差形式可表示为

$$r_{23}^2 = \frac{(\sum x_2 x_3)^2}{\sum x_2^2 \sum x_3^2} \quad (4.13)$$

将其代入到(4.9)和(4.10)式中，容易证明

$$\begin{aligned} \text{Var}(\hat{\beta}_2) &= \frac{\sum x_3^2}{(\sum x_2^2)(\sum x_3^2) - (\sum x_2 x_3)^2} \sigma^2 \\ &= \sigma^2 \frac{1}{\sum x_2^2 [1 - \frac{(\sum x_2 x_3)^2}{\sum x_2^2 \sum x_3^2}]} \\ &= \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \end{aligned} \quad (4.14)$$

同样地，可以得到：

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \quad (4.15)$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23} \sigma^2}{(1 - r_{23}^2) \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \quad (4.16)$$

从(4.14)和(4.15)式可以看出，随着共线性增加， r_{23} 趋于1，两个参数估计量的方差也将增大。同样地，其协方差在绝对值上也增大。由(4.14)、(4.15)、(4.16)式可以看出，方差和协方差增大的速度决定于方差扩大因子(简称VIF)。VIF定义为：

$$VIF = \frac{1}{(1 - r_{23}^2)} \quad (4.17)$$

VIF表明，参数估计量的方差是由于多重共线性的出现而膨胀起来的。随着共线性的增加，参数估计量的方差也增大，当 r_{23}^2 趋于1时，甚至可以变至无穷大。而当没有共线性时，VIF将是1。利用VIF的定义，（4.14）和（4.15）式可表达为：

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \cdot \text{VIF} \quad (4.18)$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} \cdot \text{VIF} \quad (4.19)$$

这表明了 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 的方差同VIF成正比关系。

2、对参数区间估计时，置信区间趋于变大

存在多重共线性时，参数估计值的方差增大，其标准误差也增大，导致总体参数的置信区间也随之变大。同上例，假设方差已知，正态分布下95%置信度下的分位点为1.96，当 $r_{23}=0.99$ 时， β_3 的置信区间要比相关系数为零时大 $\sqrt{100}$ 或约10倍(见表4.1)。

表4.1 增加共线性对 β_3 的95%置信区间 $\hat{\beta}_3 \pm 1.96 \text{se}(\hat{\beta}_3)$ 的影响

r_{23}	β_3 的95%置信区间
0.00	$\hat{\beta}_3 \pm 1.96 \sqrt{\sigma^2 / \sum x_{3i}^2}$
0.50	$\hat{\beta}_3 \pm 1.96 \sqrt{(1.33)} \sqrt{\sigma^2 / \sum x_{3i}^2}$
0.99	$\hat{\beta}_3 \pm 1.96 \sqrt{(100)} \sqrt{\sigma^2 / \sum x_{3i}^2}$
0.999	$\hat{\beta}_3 \pm 1.96 \sqrt{(500)} \sqrt{\sigma^2 / \sum x_{3i}^2}$

3、严重多重共线时，假设检验容易作出错误的判断

存在严重多重共线时，首先是参数的置信区间扩大，会使得接受一个本应拒绝的假设的概率增大；此外，在对回归系数的原假设(例如 $\beta_3=0$)的检验中，使用了t比率 $t = \hat{\beta}_3 / \sqrt{\text{Var}(\hat{\beta}_3)}$ ，在高度共线性时，参数估计值的方差增加较快，会使得t值变小，而使本应否定的“系数为0”的原假设被错误的接受。

4、当多重共线性严重时，可能造成可决系数 R^2 较高，经F检验的参数联合显著性也很高，但对各个参数单独的t检验却可能不显著，甚至可能使估计的回归系数符号相反，得

出完全错误的结论。出现这种情况，很可能正是存在严重多重共线性的表现。例如，在本章开始的“引子”里提出的“农业和建筑业的发展会减少财政收入吗？”的例子中，财政收入与农业和建筑业的增加值就表现出异常关系（变量的样本数据见练习题4.7）。

综上所述，严重的多重共线性常常会导致下列情形出现：使得用普通最小二乘得到的回归参数估计值很不稳定，回归系数的方差随着多重共线性强度的增加而加速增长，对参数难以作出精确的估计；造成回归方程高度显著的情况下，有些回归系数通不过显著性检验；甚至可能出现回归系数的正负号得不到合理的经济解释。但是应注意，如果研究目的仅在于预测 \mathbf{Y} ，而各个解释变量 \mathbf{X} 之间的多重共线性关系的性质在未来将继续保持，这时虽然无法精确估计个别的回归系数，但可估计这些系数的某些线性组合，因此多重共线性可能并不是严重问题。

第三节 多重共线性的检验

如何检验回归模型中变量之间存在多重共线性呢？下面介绍几种常用的多重共线性的检验方法。

一、简单相关系数检验法

简单相关系数检验法是利用解释变量之间的线性相关程度去判断是否存在严重多重共线性的一种简便方法。一般而言，如果每两个解释变量的简单相关系数(零阶相关系数)比较高，例如大于0.8，则可认为存在着较严重的多重共线性。但要注意，较高的简单相关系数只是多重共线性存在的充分条件，而不是必要条件。特别是在多于两个解释变量的回归模型中，有时较低的简单相关系数也可能存在多重共线性。因此并不能简单地依据相关系数进行多重共线性的准确判断。

二、方差扩大（膨胀）因子法

对于多元线性回归模型来说，如果分别以每个解释变量为被解释变量，作与其他解释变量的回归，这称为辅助回归。以 X_j 为被解释变量作对其他解释变量辅助线性回归的可决系数用 R_j^2 表示，则可以证明¹，解释变量 X_j 参数估计值 $\hat{\beta}_j$ 的方差可表示为

¹ 证明过程从略。

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \cdot \frac{1}{1 - R_j^2} = \frac{\sigma^2}{\sum x_j^2} \cdot \text{VIF}_j \quad (4.20)$$

其中的 VIF_j 是变量 X_j 的方差扩大因子(Variance Inflation Factor)，即

$$\text{VIF}_j = \frac{1}{(1 - R_j^2)} \quad (4.21)$$

注意这里的 VIF_j 是多个解释变量辅助回归确定多重可决系数 R_j^2 的基础上计算的方差扩大因子，是（4.17）式只有两个解释变量情况的拓展。

由于 R_j^2 度量了 X_j 与其他解释变量的线性相关程度，这种相关程度越强，说明变量间多重共线性越严重， VIF_j 也就越大。反之， X_j 与其他解释变量的线性相关程度越弱，说明变量间的多重共线性越弱， VIF_j 也就越接近于1。由此可见， VIF_j 的大小反映了解释变量之间是否存在多重共线性，可用它来度量多重共线性的严重程度。经验表明， $\text{VIF}_j \geq 10$ 时，说明解释变量与其余解释变量之间有严重的多重共线性，且这种多重共线性可能会过度地影响最小二乘估计。

在Eviews中，当解释变量之间存在完全或高度共线性时，将不能给出回归模型的参数估计结果，在其“Equation Specification”窗口中会显示错误提示的信息“nearly singular matrix”。在Eviews中，不能直接计算解释变量的方差扩大因子，而是需要根据VIF的定义式计算得到。

例如，计算某解释变量 X_j 的 VIF_j 值的过程如下：

首先建立一个辅助回归方程，命名为“eqjzz”。它是以 X_j 为被解释变量，其余解释变量为解释变量的回归方程。

然后在主窗口命令行输入：`scalar vifjzz=1/(1-eqjzz.@R2)`，该命令的含义是建立一个取值为此式的变量vifjzz，其中R2是 R^2 。执行后在主窗口的左下角状态栏上会出现“vifjzz successfully created”的字样，同时工作表中产生一个叫做vifjzz的新变量，双击此变量，主窗口左下角的状态栏上便会显示它的值。例如vifjzz=1047.4615，这个数值很大，根据前面的准则可以认为解释变量之间存在严重的多重共线性。

三、直观判断法

根据经验，通常以下情况的出现可能提示存在多重共线性的影响：

(1)当增加或删除一个解释变量，或者改变一个观测值时，回归参数的估计值发生较大变化，回归方程可能存在严重的多重共线性。

(2)从定性分析认为，一些重要的解释变量的回归系数的标准误差较大，在回归方程中没有通过显著性检验时，可初步判断可能存在严重的多重共线性。

(3)有些解释变量的回归系数所带正负号与定性分析结果违背时，很可能存在多重共线性。

(4)解释变量的相关矩阵中，自变量之间的相关系数较大时，可能会存在多重共线性问题。

四、逐步回归检测法

逐步回归的基本思想是将变量逐个的引入模型，每引入一个解释变量后，都要进行F检验，并对已经选入的解释变量逐个进行t检验，当原来引入的解释变量由于后面解释变量的引入而变得不再显著时，则将其剔除。以确保每次引入新的变量之前回归方程中只包含显著的变量。这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止。以保证最后所得到的解释变量集是最优的。

在逐步回归中，如果解释变量之间是高度相关的，则先前引入的解释变量可能会因为后来引入与之相关的解释变量而被剔除。逐步回归用这种有进有出的结果说明解释变量之间是否具有较高的相关性。如果解释变量之间是完全不相关的，那么引入的解释变量就不会再被剔除，而剔除的解释变量也就不会再被引入。解释变量之间具有怎样程度的相关性才会被剔除呢？则取决于研究目的与要求。在有的统计软件中（例如SPSS）可通过调整容忍度来进行。当出现多个解释变量之间高度相关的时候，逐步回归方法也是一种检测多重共线性的有效方法。

*五、特征值与病态指数²

(1)特征根分析

根据矩阵行列式的性质，矩阵的行列式等于其特征根的连乘积。因而当行列式 $|\mathbf{X}'\mathbf{X}| \approx 0$ 时，矩阵 $\mathbf{X}'\mathbf{X}$ 至少有一个特征根近似于零。反之，可以证明，当矩阵 $\mathbf{X}'\mathbf{X}$ 至少有

²这部分内容本科教学中供选择使用

一个特征根近似为零时， \mathbf{X} 的列向量之间必存在多重共线性。

设 λ 是矩阵 $\mathbf{X}'\mathbf{X}$ 的一个近似为零的特征根， \mathbf{c} 是对应于特征根的单位特征向量，则

$$\mathbf{X}'\mathbf{X}\mathbf{c} = \lambda \mathbf{c} \approx 0$$

$$\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} \approx 0$$

$$\mathbf{X}\mathbf{c} \approx 0$$

$$c_0 X_0 + c_1 X_1 + \cdots + c_k X_k \approx 0$$

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_k x_{ik} \approx 0 \quad i = 0, 1, 2, \dots, n$$

这就是前面定义的多重共线性关系。

矩阵 $\mathbf{X}'\mathbf{X}$ 有多少个特征根近似为零，设计矩阵就会有多少个多重共线性关系，并且这些多重共线性关系系数向量就等于接近于零的那些特征根对应的特征向量。

特征根近似为零的标准可以用下面的病态指数来确定。记 $\mathbf{X}'\mathbf{X}$ 的最大特征根为 λ_m ，称

$$CI_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, \quad i = 0, 1, 2, \dots, k \quad (4.22)$$

为特征根的病态指数（condition index）。注意特征根个数与病态指数都包含了常数项在内。

病态指数度量了矩阵 $\mathbf{X}'\mathbf{X}$ 的特征根散布程度，可以用来判断多重共线性是否存在以及多重共线性的严重程度。

一般认为， $0 < CI < 10$ 时，设计矩阵没有多重共线性； $10 \leq CI < 100$ 时，认为 \mathbf{X} 存在较强的多重共线性；当 $CI \geq 100$ 时，则认为存在严重多重共线性。

利用病态指数进行判断时，软件一般给出按大小顺序排列的特征根和病态指数，病态指数大，说明存在多重共线性，但不能判断哪个变量之间有多重共线性，这时还需要结合方差比例表进行分析，如果某几个解释变量的方差比例值在某一行同时较大时，则这几个解释变量之间就存在多重共线性。

第四节 多重共线性的补救措施

诊断出多重共线性还只完成了任务的一半，还需要在存在多重共线性的基础上采取措施进行补救，以便尽量降低回归模型中存在的多重共线性。这里介绍常用的几种降低多重共线

性的方法。

一、修正多重共线性的经验方法

1、剔除变量法。

这是最简单的一种方法。当回归方程中存在严重的多重共线性，可以删除引起多重共线性的不重要的解释变量。有几个变量方差扩大因子大于10时，可试探把方差扩大因子最大者所对应的解释变量首先剔除，再重新建立回归方程。如果仍然存在严重的多重共线性，则再继续剔除方差扩大因子最大者所对应的解释变量，直至回归方程中不再存在严重的多重共线性。

一般而言，在选择回归模型时，可以将回归系数的显著性检验、方差扩大因子VIF的多重共线性检验与解释变量经济含义(通过经济分析确定变量的相对重要性)结合起来考虑，以引进或剔除不重要的变量。不过，采用该方法时千万要注意，如果剔除的变量是重要变量，又可能引起模型的设定误差（见第九章的讨论）。

2、增大样本容量

由于多重共线性是一个样本特性，所以可能在同样变量的另一样本中共线性又没有那样严重。这时，如果可能，可以通过增大样本容量来减轻共线性的问题。建立一个实际经济问题的回归模型，如果所收集的样本数据太少，是容易产生多重共线性的，从前面(4.18)和(4.19)式可知，如果样本容量增加，则 $\sum x_{ji}^2$ 也会增加，结果会减小回归参数的方差，标准误差也同样会减小。因此尽可能地收集足够多的数据可以改进模型参数的估计。所以在运用回归分析研究经济问题时，要尽量使样本容量远大于解释变量的个数。

3、变换模型形式

将原设定的模型的形式作适当的变换，有可能消除或减弱原模型中解释变量之间的相关关系。例如，可采用差分法，这是指将原模型变形为差分模型形式进而减低多重共线性的一个方法。

将原模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \quad (4.23)$$

变形为

$$\Delta Y_i = \beta_1 \Delta X_{1i} + \beta_2 \Delta X_{2i} + \cdots + \beta_k \Delta X_{ki} + \Delta u_i \quad (4.24)$$

一般而言，差分后变量之间的相关性要比差分前要弱得多，所以差分后的模型可以有

效地降低出现共线性的可能性，此时可直接估计差分方程。在Eviews中运用差分法时，可以在方程定义栏中输入：“ $Y - Y(-1) \quad X1 - X1(-1) \quad X2 - X2(-1) \cdots \quad Xk - Xk(-1)$ ”，从而得到 $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ 的估计值。

因为差分常常会丢失一些信息，差分模型的误差项可能是序列相关的，可能会违背经典线性回归模型的相关假设，在具体运用时要慎重。

4、利用非样本先验信息

如果通过经济理论分析能够得到某些参数之间的线性关系，可以将这种线性关系作为约束条件，将此约束条件和样本信息结合起来进行约束最小二乘估计。

例如，考虑以下模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

如果依据长期的经验分析可以认为 $\beta_3 = 0.2\beta_2$ ，这样，我们就可以化简为下列模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i = \beta_1 + \beta_2 X_i + u_i \quad (4.25)$$

其中， $X_i = X_{2i} + 0.2X_{3i}$ 。如果估计出 $\hat{\beta}_2$ ，则也估计出了 $\hat{\beta}_3$ 。

5、横截面数据与时间序列数据并用

这是先验信息法的一个变种，将横截面数据与时间序列数据组合，称为数据并用 (pooling the data)。其基本做法是：首先利用横截面数据估计出部分参数，再利用时序数据估计出另外的部分参数，最后得到整个方程参数的估计。

假设我们要研究中国家用轿车需求，并收集到了关于家用轿车的销售数据(Y_t)、平均价格(P_t)和消费者收入(I_t)的时间序列数据。依据研究目的，设定的模型为

$$\ln Y_t = \beta_1 + \beta_2 \ln P_t + \beta_3 \ln I_t + u_t \quad (4.26)$$

目的是要估计价格弹性 β_2 和收入弹性 β_3 。

在时间序列分析中，价格和收入变量一般都有高度共线性的趋势。托宾提出了解决此问题的一种方法。即如果拥有关于消费者定点追踪的横截面数据，如城镇或农村居民住户调查数据，就可能可靠地估计收入弹性 β_3 。令收入弹性的横截面估计为 $\hat{\beta}_3^*$ ，就可以将前述时间序列回归写成

$$Y_t^* = \beta_1 + \beta_2 \ln P_t + u_t \quad (4.27)$$

其中, $Y_t^* = \ln Y - \hat{\beta}_3^* \ln I_t$ 。

这样就可以得到价格弹性的估计值。值得注意的是这里包含着假设: 收入弹性的横截面估计和从纯粹时间序列分析中得到的估计是一样的。当横截面估计在不同截面之间没有大的变化时这是一个值得考虑的方法。

6、变量变换

有时通过对模型中变量进行变换能够实现降低共线性的目的。例如, 常用的变量变换方式有:

(1) 计算相对指标。如原来的是总量指标, 可计算人均指标或结构相对数(比重)指标等。经过这样处理的数据有时可以降低共线性。

(2) 将名义数据转换为实际数据。将名义数据剔除价格影响后反映的信息在统计上常常是指纯的物量的变化, 不包含价格变动的影响, 有助于描述现象之间真实的数量变化关系。因此在多数经济分析中采用“实际”数据而不是名义数据, 有时名义数据转换为实际数据后可降低多重共线性。

(3) 将小类指标合并成大类指标。如例4.1中, 工业增加值、建筑业增加值之间呈现高度线性相关, 可将其合并成第二产业增加值。这一合并有助于消除多重共线性。

需要指出, 变量数据的变换只是有时可得到较好的结果, 但谁也无法保证一定可以得到很好的结果。

二、逐步回归法

依据前述逐步回归的思想, 可通过逐步回归筛选并剔除引起多重共线性的变量。其具体步骤如下: (1) 用被解释变量对每一个所考虑的解释变量做简单回归。(2) 以对被解释变量贡献最大的解释变量所对应的回归方程为基础, 按对被解释变量贡献大小的顺序逐个引入其余的解释变量。这个过程会出现3种情形。①若新变量的引入改进了 R^2 和F检验, 且回归参数的t检验在统计上也是显著的, 则在模型中保留该变量。②若新变量的引入未能明显改进 R^2 和F检验, 且对其他回归参数估计值的t检验也未带来什么影响, 则认为该变量是多余的, 应该舍弃。③若新变量的引入未能明显改进 R^2 和F检验, 且显著地影响了其他回归参数估计值的数值或符号, 同时本身的回归参数也通不过t检验, 则说明出现了严重的多重共线性, 应剔除该变量。

例如，在本章开始的“引子”里提出的“农业和建筑业的发展会减少财政收入吗？”的例子中，包含所有解释变量的模型存在多重共线性。为降低多重共线性，可采用逐步回归法筛选解释变量（样本数据见练习题4.7）。

(1) 用被解释变量分别对每个解释变量做简单回归，以可决系数为标准确定解释变量的重要程度，将解释变量排序。以下模型中括号内为参数估计的t检验统计量。

$$CS = -1032.163 + 0.9358NZ$$

$$t = (1.2722) \quad (10.7563) \quad R^2 = 0.8283, \quad F = 115.74, \quad N = 26$$

$$CS = 93.081 + 0.3486GZ$$

$$t = (0.2256) \quad (19.3667) \quad R^2 = 0.9415, \quad F = 385.997, \quad N = 26$$

$$CS = 369.657 + 2.2248JZZ$$

$$t = (0.8687) \quad (18.5554) \quad R^2 = 0.9348, \quad F = 344.307, \quad N = 26$$

$$CS = -47378.78 + 0.4674TPOP$$

$$t = (6.6850) \quad (7.5509) \quad R^2 = 0.7036, \quad F = 56.984, \quad N = 26$$

$$CS = -201.534 + 0.2609CUM$$

$$t = (0.4391) \quad (18.1181) \quad R^2 = 0.9316, \quad F = 326.917, \quad N = 26$$

$$CS = -14855.98 + 0.4461SZM$$

$$t = (2.0848) \quad (2.9426) \quad R^2 = 0.2652, \quad F = 8.6599, \quad N = 26$$

解释变量的重要程度依次为GZ、JZZ、CUM、NZ、TPOP、SZM。

(2) 以 $CS = 93.081 + 0.3486GZ$ 为基础，依次引入JZZ、CUM、NZ、TPOP、SZM。首先将JZZ引入模型，得

$$CS = -570.7853 + 1.2602GZ - 5.8415JZZ$$

$$(538.6945) \quad (0.5043) \quad (3.2295) \quad R^2 = 0.9488, \quad F = 212.9026, \quad N = 26$$

因JZZ引入，对可决系数改进不多，且回归系数未通过t检验，而剔除JZZ。接着引入CUM，得到模型：

$$CS = 513.788 + 0.785GZ - 0.329CUM$$

$$(508.125) \quad (0.3186) \quad (0.2398) \quad R^2 = 0.9459, \quad F = 201.036, \quad N = 26$$

同理剔除CUM，引入NZ，得

$$CS = 1984.283 + 0.7298GZ - 1.1126NZ$$

$$(298.414) \quad (0.0448) \quad (0.1282) \quad R^2 = 0.9863, \quad F = 828.17, \quad N = 26$$

模型的可决系数提高很多，且t统计量显著，保留NZ，继续引入TPOP。得模型：

$$CS = -13051.93 + 0.7883GZ - 1.5406NZ + 0.1512TPOP$$

$$(3243.474) \quad (0.03489) \quad (0.131) \quad (0.0325) \quad R^2 = 0.9931, \quad F = 1053.505, \quad N = 26$$

模型的可决系数提高很多，且t统计量显著，保留TPOP，继续引入SZM，得模型：

$$CS = -13480.15 + 0.7923GZ - 1.5599NZ + 0.1704TPOP - 0.036319SZM$$

$$(3069.964) \quad (0.0330) \quad (0.124) \quad (0.0323) \quad (0.0189) \quad R^2 = 0.9941, \quad F = 887.56, \quad N = 26$$

SZM的回归系数在0.05水平下不显著，剔除SZM。最后消除了多重共线性的模型为：

$$CS = -13051.93 + 0.7883GZ - 1.5406NZ + 0.1512TPOP$$

$$(3243.474) \quad (0.03489) \quad (0.131) \quad (0.0325) \quad R^2 = 0.9931, \quad F = 1053.505, \quad N = 26$$

逐步回归法的好处是将统计上不显著的解释变量一一剔除，最后得到的最优解释变量之间不仅相关系数不高，而且对被解释变量有较好的解释贡献。

*三、岭回归法简介³

为了降低多重共线性对回归模型的影响，计量经济学家们还致力于改进古典的最小二乘法，提出以采用有偏的估计为代价来提高估计量的稳定性的方法，如岭回归法、主成分法、偏最小二乘法等。下面简要介绍岭回归法的思想和方法。

(一) 岭回归的含义

岭回归（Ridge Regression）是A.E.Hoerl(霍尔)提出的一种改进最小二乘估计的方法，也叫岭估计（Ridge Estimate）。

当解释变量之间存在多重共线性时， $|\mathbf{X}'\mathbf{X}| \approx 0$ ，则 $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ 会增大，原因在于 $\mathbf{X}'\mathbf{X}$ 接近奇异。如果将 $\mathbf{X}'\mathbf{X}$ 加上一个正常数对角矩阵 $k\mathbf{I}$ ($k > 0$, \mathbf{I} 为单位矩阵)，即 $\mathbf{X}'\mathbf{X} + k\mathbf{I}$ ，使得 $|\mathbf{X}'\mathbf{X} + k\mathbf{I}| \approx 0$ 的可能性比 $|\mathbf{X}'\mathbf{X}| \approx 0$ 的可能性更小，那么 $\mathbf{X}'\mathbf{X} + k\mathbf{I}$ 接近奇异的程度就会比 $\mathbf{X}'\mathbf{X}$ 小得多。

这样可以得到 $\boldsymbol{\beta}$ 的岭回归估计为：

$$\tilde{\boldsymbol{\beta}}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{Y} \quad (4.28)$$

称 $\tilde{\boldsymbol{\beta}}(k)$ 为 $\boldsymbol{\beta}$ 的岭回归估计量， k 为岭回归参数。当解释变量之间存在多重共线性时，

³这部分内容本科教学中供选择使用

以 $\tilde{\beta}(k)$ 作为 β 的估计应比普通最小二乘估计稳定。当 k 较小时，回归系数很不稳定，而当 k 逐渐增大时，回归系数可能呈现稳定状态。因此要选择合适的 k 值时，岭回归参数才会优于普通最小二乘估计参数。当 $k=0$ 时，岭回归估计 $\tilde{\beta}(k)=\hat{\beta}$ ，实际就是普通最小二乘估计。

(二)岭回归估计的性质

性质1：岭回归的参数估计是回归参数的有偏估计

$$\begin{aligned} E(\tilde{\beta}(k)) &= E(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{X}\beta \end{aligned} \quad (4.29)$$

显然，只有当 $k=0$ 时，才有 $E(\tilde{\beta}(0))=\beta$ ，当 $k \neq 0$ 时， $\tilde{\beta}(k)$ 是 β 的有偏估计。有偏性是岭回归估计的一个重要性质。

性质2：从(4.28)式容易看出，在岭回归参数 k 与 \mathbf{Y} 无关的情形下， $\tilde{\beta}(k)$ 是最小二乘估计的一个线性变换，也是理论值 \mathbf{Y} 的线性函数。

性质3：可以证明岭估计量 $\tilde{\beta}(k)$ 方差比普通最小二乘估计 $\hat{\beta}^2$ 的方差要小。

岭回归估计的方差和偏倚与岭回归参数 k 有关，岭回归参数 k 的值越大， $\tilde{\beta}(k)$ 的偏倚越大，其方差就越小。要得到方差较小的估计结果，又不得不牺牲无偏性。为此可以用兼顾方差和偏倚的最小均方误差MSE原则[具体见第九章（9.16）式]，去分析岭回归的效果。

(三)岭回归参数k的选择

原则上是要选择使均方误差 $MSE[\hat{\beta}(k)]$ 达到最小的 k ，而最优 k 值依赖于未知参数 β 和 σ^2 ，因而在实际应用中必须通过样本来确定。目前还没有形成公认的选择岭回归参数的最优方法，常用的方法主要有岭迹法、方差扩大因子法、残差平方和方法。在实际应用中，可考虑使用逐步搜索的方法，即开始给定较小的 k 值，然后逐渐增加 k 的取值进行试验，直至岭估计量 $\tilde{\beta}(k)$ 的值趋于稳定为止。

显然，用逐步搜索的方法确定的 k 值，仍缺乏令人信服的理论依据，具有一定主观性，是一种将定性分析与定量分析相结合的方法。

第五节 案例分析

一、研究的目的要求

近年来，中国旅游业一直保持高速发展，旅游业作为国民经济新的增长点，在整个社会经济发展中的作用日益显现。中国的旅游业分为国内旅游和入境旅游两大市场，入境旅游外汇收入年均增长 22.6%，与此同时国内旅游也迅速增长。改革开放 20 多年来，特别是进入 90 年代后，中国的国内旅游收入年均增长 14.4%，远高于同期 GDP 9.76% 的增长率。为了规划中国未来旅游产业的发展，需要定量地分析影响中国旅游市场发展的主要因素。

二、模型设定及其估计

经分析，影响国内旅游市场收入的主要因素，除了国内旅游人数和旅游支出以外，还可能与相关基础设施有关。为此，考虑的影响因素主要有国内旅游人数 X_2 ，城镇居民人均旅游支出 X_3 ，农村居民人均旅游支出 X_4 ，并以公路里程 X_5 和铁路里程 X_6 作为相关基础设施的代表。为此设定了如下对数形式的计量经济模型：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + u_t$$

其中： Y_t ——第 t 年全国旅游收入

X_2 ——国内旅游人数（万人）

X_3 ——城镇居民人均旅游支出（元）

X_4 ——农村居民人均旅游支出（元）

X_5 ——公路里程（万公里）

X_6 ——铁路里程（万公里）

为估计模型参数，收集旅游事业发展最快的 1994—2003 年的统计数据，如表 4.2 所示：

表 4.2 1994 年—2003 年中国旅游收入及相关数据

年 份	国内旅游 收入 Y (亿元)	国内旅游 人数 X_2 (万人次)	城镇居民人均 旅游支出 X_3 (元)	农村居民人均 旅游支出 X_4 (元)	公路里 程 X_5 (万公里)	铁路里 程 X_6 (万公里)
1994	1023.5	52400	414.7	54.9	111.78	5.90
1995	1375.7	62900	464.0	61.5	115.70	5.97
1996	1638.4	63900	534.1	70.5	118.58	6.49
1997	2112.7	64400	599.8	145.7	122.64	6.60
1998	2391.2	69450	607.0	197.0	127.85	6.64
1999	2831.9	71900	614.8	249.5	135.17	6.74

2000	3175.5	74400	678.6	226.6	140.27	6.87
2001	3522.4	78400	708.3	212.7	169.80	7.01
2002	3878.4	87800	739.7	209.1	176.52	7.19
2003	3442.3	87000	684.9	200.0	180.98	7.30

数据来源：《中国统计年鉴2004》

利用Eviews软件，输入Y、X2、X3、X4、X5、X6等数据，采用这些数据对模型进行OLS回归，结果如表4.3：

表4.3

Dependent Variable: Y					
Method: Least Squares					
Date: 07/18/05 Time: 18:16					
Sample: 1994 2003					
Included observations: 10					
Variable	Coefficient	Std. Error	t-Statistic	Prob.	
C	-274.3773	1316.690	-0.208384	0.8451	
X2	0.013088	0.012692	1.031172	0.3607	
X3	5.438193	1.380395	3.939591	0.0170	
X4	3.271773	0.944215	3.465073	0.0257	
X5	12.98624	4.177929	3.108296	0.0359	
X6	-563.1077	321.2830	-1.752685	0.1545	
R-squared	0.995406	Mean dependent var		2539.200	
Adjusted R-squared	0.989664	S.D. dependent var		985.0327	
S.E. of regression	100.1433	Akaike info criterion		12.33479	
Sum squared resid	40114.74	Schwarz criterion		12.51634	
Log likelihood	-55.67396	F-statistic		173.3525	
Durbin-Watson stat	2.311565	Prob(F-statistic)		0.000092	

由此可见，该模型 $R^2 = 0.9954$ ， $\bar{R}^2 = 0.9897$ 可决系数很高，F 检验值 173.3525，明显显著。但是当 $\alpha = 0.05$ 时 $t_{\alpha/2}(n-k) = t_{0.025}(10-6) = 2.776$ ，不仅 X_2 、 X_6 系数的 t 检验不显著，而且 X_6 系数的符号与预期的相反，这表明很可能存在严重的多重共线性。

计算各解释变量的相关系数，选择 X2、X3、X4、X5、X6 数据，点” view/correlations” 得相关系数矩阵（如表 4.4）：

表 4.4

	X2	X3	X4	X5	X6
X2	1.000000	0.918851	0.751960	0.947977	0.941681
X3	0.918851	1.000000	0.865145	0.859191	0.963313
X4	0.751960	0.865145	1.000000	0.664946	0.818137
X5	0.947977	0.859191	0.664946	1.000000	0.897708
X6	0.941681	0.963313	0.818137	0.897708	1.000000

由相关系数矩阵可以看出：各解释变量相互之间的相关系数较高，证实确实存在严重多重共线性。

三、消除多重共线性

采用逐步回归的办法，去检验和解决多重共线性问题。分别作 Y 对 X2、X3、X4、X5、X6 的一元回归，结果如表 4.5 所示：

表 4.5

变量	X2	X3	X4	X5	X6
参数估计值	0.0842	9.0523	11.6673	34.3324	2014.146
t 统计量	8.6659	13.1598	5.1967	6.4675	8.7487
R^2	0.9037	0.9558	0.7715	0.8394	0.9054

按 R^2 的大小排序为：X3、X6、X2、X5、X4。

以 X3 为基础，顺次加入其他变量逐步回归。首先加入 X6 回归结果为：

$$\hat{Y}_t = -4109.639 + 7.850632X_3 + 285.1784X_6$$

$$t=(2.9086) \quad (0.46214) \quad R^2 = 0.957152$$

当取 $\alpha = 0.05$ 时， $t_{\alpha/2}(n-k) = t_{0.025}(10-3) = 2.365$ ，X6 参数的 t 检验不显著，予以剔除，加入 X2 回归得

$$\hat{Y}_t = -3326.393 + 6.194241X_3 + 0.029761X_2$$

$$t=(4.2839) \quad (2.1512) \quad R^2 = 0.973418$$

X2 参数的 t 检验不显著，予以剔除，加入 X5 回归得

$$\hat{Y}_t = -3059.972 + 6.736535X_3 + 10.90789X_5$$

$$t=(6.6446) \quad (2.6584) \quad R^2 = 0.978028$$

X3、X5 参数的 t 检验显著，保留 X5，再加入 X4 回归得

$$\hat{Y}_t = -2441.161 + 4.215884X_3 + 13.62909X_5 + 3.221965X_4$$

$$t=(3.944983) \quad (4.692961) \quad (3.06767)$$

$$R^2 = 0.991445 \quad \bar{R}^2 = 0.987186 \quad F=231.7935 \quad DW=1.952587$$

当取 $\alpha = 0.05$ 时， $t_{\alpha/2}(n-k) = t_{0.025}(10-4) = 2.447$ ，X3、X4、X5 系数的 t 检验都显著，

这是最后消除多重共线性的结果。

这说明，在其他因素不变的情况下，当城镇居民人均旅游支出 X_3 和农村居民人均旅游支出 X_4 分别增长1元时，国内旅游收入 Y_t 将分别增长4.21亿元和3.22亿元。在其他因素不变的情况下，作为旅游设施的代表，公路里程 X_5 每增加1万公里时，国内旅游收入 Y_t 将增长13.63亿元。

第四章小节

1、经典线性回归模型的假定之一是各个解释变量X之间不存在多重共线性。一般说来，多重共线性是指各个解释变量X之间有准确或近似准确的线性关系。

2、多重共线性的后果是：如果各个解释变量X之间有完全的共线性，则它们的回归系数是不确定的，并且它们的方差会无穷大。如果共线性是高度的但不完全的，则回归系数的估计是可能的，但有较大的标准误差的趋势。结果回归系数不能准确地加以估计。不过，如果目的是估计这些系数的线性组合用于预测，多重共线性不是严重问题。

3、诊断共线性的经验方法主要有：(1)多重共线性的明显表现是可决系数 R^2 异常高而回归系数在通常的t检验中在统计上不显著。(2)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，一般说来高的相关系数通常可认为有多重共线性。(3)当模

型中涉及多于两个解释变量的情形时，较低的零阶相关也可能出现多重共线性，这时需要检查偏相关系数。(4)如果 R^2 高而偏相关系数低，则多重共线性是可能的，这时会存在一个或多个解释变量是多余的。如果 R^2 高而偏相关系数也高，则多重共线性难以识别。(5)在建模时，首先可以将每一个解释变量 X_i 对其余所有解释变量进行辅助回归，并计算出相应的可决系数 R_i^2 。较高的 R_i^2 可能表明 X_i 和其余的解释变量高度相关，在不会引起严重的设定偏误的前提下，可考虑把 X_i 从模型中剔除。

4、降低多重共线性的经验方法有：(1)利用外部或先验信息；(2)横截面与时间序列数据并用；(3)剔除高度共线性的变量；(4)数据转换；(5)获取补充数据或新数据；(6)选择有偏估计量（如岭回归）。经验方法的效果取决于数据的性质和共线性的严重程度。

第四章主要公式表

方差—膨胀因子(简称VIF)	$VIF = \frac{1}{(1 - r_{23}^2)}$
多重共线性下参数估计式的方差	$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \cdot VIF$ $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \cdot \frac{1}{1 - R_j^2} = \frac{\sigma^2}{\sum x_j^2} \cdot VIF_j$
特征根的病态指数	$CI_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, \quad i = 0, 1, 2, \dots, k$
β 的岭回归估计	$\tilde{\beta}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{Y}$

思考题与练习题

思考题

- 4.1 多重共线性的实质是什么？为什么会出现多重共线性？
- 4.2 多重共线性对回归参数的估计有何影响？
- 4.3 多重共线性的典型表现是什么？判断是否存在多重共线性的方法有哪些？

4.4 针对出现多重共线性的不同情形，能采取的补救措施有哪些？

4.5 在涉及相关的宏观经济总量指标如GDP、货币供应量、物价总水平、国民总收入、就业人数等时间序列的数据中一般都会怀疑有多重共线性，为什么？

4.6 多重共线性的产生与样本容量的个数 n 、解释变量的个数 k 有无关系？

4.7 具有严重多重共线性的回归方程能否用来进行预测？

4.8 岭回归法的基本思想是什么，它对降低共线性有何作用？

4.9 以下陈述是否正确？请判断并说明理由

(1)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单个显著性是不可能的。

(2)尽管有完全的多重共线性，OLS估计量仍然是BLUE。

(3)如果有某一辅助回归显示出高的 R_j^2 值，则高度共线性的存在是肯定无疑的。

(4)变量的两两高度相关并不表示高度多重共线性。

(5)如果其他条件不变，VIF越高，OLS估计量的方差越大。

(6)如果在多元回归中，根据通常的t检验，全部偏回归系数分别都是统计上不显著的，你就不会得到一个高的 R^2 值。

(7)在Y对 X_2 和 X_3 的回归中，假如 X_3 的值很少变化，这就会使 $\text{var}(\hat{\beta}_3)$ 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。

(8)如果分析的目的仅仅是预测，则多重共线性是无害的。

练习题

4.1 假设在模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ 中， X_2 与 X_3 之间的相关系数为零，于是有人建议你进行如下回归：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_{1i}$$

$$Y_i = \gamma_1 + \gamma_3 X_{3i} + u_{2i}$$

(1)是否存在 $\hat{\alpha}_2 = \hat{\beta}_2$ 且 $\hat{\gamma}_3 = \hat{\beta}_3$ ？为什么？

(2) $\hat{\beta}_1$ 会等于 $\hat{\alpha}_1$ 或 $\hat{\gamma}_1$ 或两者的某个线性组合吗？

(3)是否有 $\text{var}(\hat{\beta}_2) = \text{var}(\hat{\alpha}_2)$ 且 $\text{var}(\hat{\beta}_3) = \text{var}(\hat{\gamma}_3)$ ？

4.2在决定一个回归模型的“最优”解释变量集时人们常用逐步回归的方法。不我待在逐步回归中既可采取每次引进一个解释变量的程序(逐步向前回归),也可以先把所有可能的解释变量都放在一个多元回归中,然后逐一地将它们剔除(逐步向后回归)。加进或剔除一个变量,通常是根据F检验看其对ESS的贡献而作出决定的。根据你现在对多重共线性的认识,你赞成任何一种逐步回归的程序吗?为什么?

4.3 下表给出了中国商品进口额Y、国内生产总值GDP、消费者价格指数CPI。

年份	商品进口额 (亿元)	国内生产总值 (亿元)	居民消费价格指数 (1985=100)
1985	1257.8	8964.4	100
1986	1498.3	10202.2	106.5
1987	1614.2	11962.5	114.3
1988	2055.1	14928.3	135.8
1989	2199.9	16909.2	160.2
1990	2574.3	18547.9	165.2
1991	3398.7	21617.8	170.8
1992	4443.3	26638.1	181.7
1993	5986.2	34634.4	208.4
1994	9960.1	46759.4	258.6
1995	11048.1	58478.1	302.8
1996	11557.4	67884.6	327.9
1997	11806.5	74462.6	337.1
1998	11626.1	78345.2	334.4
1999	13736.4	82067.5	329.7
2000	18638.8	89468.1	331.0
2001	20159.2	97314.8	333.3
2002	24430.3	105172.3	330.6
2003	34195.6	117251.9	334.6

资料来源：《中国统计年鉴》，中国统计出版社 2000 年、2004 年。

请考虑下列模型： $\ln Y_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + u_t$

(1)利用表中数据估计此模型的参数。

(2)你认为数据中有多重共线性吗？

(3)进行以下回归：

$$\ln Y_t = A_1 + A_2 \ln GDP_t + v_{1i}$$

$$\ln Y_t = B_1 + B_2 \ln CPI_t + v_{2i}$$

$$\ln GDP_t = C_1 + C_2 \ln CPI_t + v_{3i}$$

根据这些回归你能对数据中多重共线性的性质说些什么？

(4)假设数据有多重共线性，但 $\hat{\beta}_2$ 和 $\hat{\beta}_3$ 在 5% 水平上个别地显著，并且总的 F 检验也是显著的。对这样的情形，我们是否应考虑共线性的问题？

4.4 自己找一个经济问题来建立多元线性回归模型，怎样选择变量和构造解释变量数据矩阵 X 才可能避免多重共线性的出现？

4.5 克莱因与戈德伯格曾用 1921-1950 年 (1942-1944 年战争期间略去) 美国国内消费 Y 和工资收入 X_1 、非工资—非农业收入 X_2 、农业收入 X_3 的时间序列资料，利用 OLSE 估计得出了下列回归方程：

$$\hat{Y} = 8.133 + 1.059X_1 + 0.452X_2 + 0.121X_3$$

$$(8.92) \quad (0.17) \quad (0.66) \quad (1.09)$$

$$R^2 = 0.95 \quad F = 107.37$$

(括号中的数据为相应参数估计量的标准误)。

试对上述模型进行评析，指出其中存在的问题。

4.6 理论上认为影响能源消费需求总量的因素主要有经济发展水平、收入水平、产业发展、人民生活水平提高、能源转换技术等因素。为此，收集了中国能源消费总量 Y (万吨标准煤)、国内生产总值(亿元) X_1 (代表经济发展水平)、国民总收入(亿元) X_2 (代表收入水平)、工业增加值(亿元) X_3 、建筑业增加值(亿元) X_4 、交通运输邮电业增加值(亿元) X_5 (代表产业发展水平及产业结构)、人均生活电力消费 (千瓦时) X_6 (代表人民生活水平提高)、能源加工转换效率(%) X_7 (代表能源转换技术)等在 1985-2002 年期间的统计数据，具体如下：

年份	能源消费	国民总收入	G D P	工业	建筑业	交通运输邮电	人均生活电力消费	能源加工转换效率
	y	X1	X2	X3	X4	X5	X6	X7
1985	76682	8989.1	8964.4	3448.7	417.9	406.9	21.3	68.29
1986	80850	10201.4	10202.2	3967.0	525.7	475.6	23.2	68.32
1987	86632	11954.5	11962.5	4585.8	665.8	544.9	26.4	67.48
1988	92997	14922.3	14928.3	5777.2	810.0	661.0	31.2	66.54
1989	96934	16917.8	16909.2	6484.0	794.0	786.0	35.3	66.51
1990	98703	18598.4	18547.9	6858.0	859.4	1147.5	42.4	67.2
1991	103783	21662.5	21617.8	8087.1	1015.1	1409.7	46.9	65.9
1992	109170	26651.9	26638.1	10284.5	1415.0	1681.8	54.6	66
1993	115993	34560.5	34634.4	14143.8	2284.7	2123.2	61.2	67.32
1994	122737	46670.0	46759.4	19359.6	3012.6	2685.9	72.7	65.2
1995	131176	57494.9	58478.1	24718.3	3819.6	3054.7	83.5	71.05
1996	138948	66850.5	67884.6	29082.6	4530.5	3494.0	93.1	71.5
1997	137798	73142.7	74462.6	32412.1	4810.6	3797.2	101.8	69.23
1998	132214	76967.2	78345.2	33387.9	5231.4	4121.3	106.6	69.44
1999	130119	80579.4	82067.5	35087.2	5470.6	4460.3	118.1	70.45
2000	130297	88254.0	89468.1	39047.3	5888.0	5408.6	132.4	70.96
2001	134914	95727.9	97314.8	42374.6	6375.4	5968.3	144.6	70.41
2002	148222	103935.3	105172.3	45975.2	7005.0	6420.3	156.3	69.78

资料来源：《中国统计年鉴》2004、2000年版，中国统计出版社。

要求：

(1)建立对数线性多元回归模型

(2)如果决定用表中全部变量作为解释变量，你预料会遇到多重共线性的问题吗？为什么？

(3)如果有多重共线性，你准备怎样解决这个问题？明确你的假设并说明全部计算。

4.7 在本章开始的“引子”提出的“农业和建筑业的发展会减少财政收入吗？”的例

子中，如果所采用的数据如下表所示

1978-2003年财政收入及其影响因素数据

年份	财政收入 (亿元)CS	农业增加值 (亿元)NZ	工业增加值 (亿元)GZ	建筑业增加值(亿元)JZZ	总人口(万人)TP0P	最终消费 (亿元)CUM	受灾面积 (万公顷)SZM
1978	1132.3	1018.4	1607.0	138.2	96259	2239.1	50760
1979	1146.4	1258.9	1769.7	143.8	97542	2619.4	39370
1980	1159.9	1359.4	1996.5	195.5	98705	2976.1	44530
1981	1175.8	1545.6	2048.4	207.1	100072	3309.1	39790
1982	1212.3	1761.6	2162.3	220.7	101654	3637.9	33130
1983	1367.0	1960.8	2375.6	270.6	103008	4020.5	34710
1984	1642.9	2295.5	2789.0	316.7	104357	4694.5	31890
1985	2004.8	2541.6	3448.7	417.9	105851	5773.0	44370
1986	2122.0	2763.9	3967.0	525.7	107507	6542.0	47140
1987	2199.4	3204.3	4585.8	665.8	109300	7451.2	42090
1988	2357.2	3831.0	5777.2	810.0	111026	9360.1	50870
1989	2664.90	4228.0	6484.0	794.0	112704	10556.5	46991
1990	2937.10	5017.0	6858.0	859.4	114333	11365.2	38474
1991	3149.48	5288.6	8087.1	1015.1	115823	13145.9	55472
1992	3483.37	5800.0	10284.5	1415.0	117171	15952.1	51333
1993	4348.95	6882.1	14143.8	2284.7	118517	20182.1	48829
1994	5218.10	9457.2	19359.6	3012.6	119850	26796.0	55043
1995	6242.20	11993.0	24718.3	3819.6	121121	33635.0	45821
1996	7407.99	13844.2	29082.6	4530.5	122389	40003.9	46989
1997	8651.14	14211.2	32412.1	4810.6	123626	43579.4	53429
1998	9875.95	14552.4	33387.9	5231.4	124761	46405.9	50145
1999	11444.08	14472.0	35087.2	5470.6	125786	49722.7	49981
2000	13395.23	14628.2	39047.3	5888.0	126743	54600.9	54688
2001	16386.04	15411.8	42374.6	6375.4	127627	58927.4	52215
2002	18903.64	16117.3	45975.2	7005.0	128453	62798.5	47119
2003	21715.25	17092.1	53092.9	8181.3	129227	67442.5	54506

(资料来源：《中国统计年鉴2004》，中国统计出版社2004年版)

试分析：为什么会出现本章开始时所得到的异常结果？怎样解决所出现的问题？

第五章 异方差性

引子:

更为接近真实的结论是什么?

改革开放以来,各地区的医疗机构都有了较快发展,不仅政府建立了一批医疗机构,还建立了不少民营医疗机构。各地医疗机构的发展状况,除了其他因素外主要决定于对医疗服务的需求量,而医疗服务需求与人口数量有关。为了给制定医疗机构的规划提供依据,分析比较医疗机构与人口数量的关系,建立卫生医疗机构数与人口数的回归模型。根据四川省2000年21个地市州医疗机构数与人口数资料对模型估计的结果如下:

$$\hat{Y}_i = -563.0548 + 5.3735X_i$$

$$(291.5778) \quad (0.644284)$$

$$t = (-1.931062) \quad (8.340265)$$

$$R^2 = 0.785456, \bar{R}^2 = 0.774146, F = 69.56003$$

式中 Y 表示卫生医疗机构数(个), X 表示人口数量(万人)。从回归模型估计的结果看,人口数量对应参数的标准误差较小, t 统计量远大于临界值,说明人口数量对医疗机构确有显著影响,可决系数和修正的可决系数还可以, F 检验结果也明显显著。表明该模型的估计效果还不错,可以认为人口数量每增加1万人,平均说来医疗机构将增加5.3735个。

然而,这里得出的结论可能是不可靠的,平均说来每增加1万人口可能并不需要增加这样多的医疗机构,所得结论并不符合真实情况。那末,有什么充分的理由说明这一回归结果不可靠呢?更为接近真实的结论又是什么呢?

在现实经济活动中,最小二乘法的基本假定并非都能满足,上一章介绍的多重共线性只是其中一个方面,本章将讨论违背基本假定的另一个方面——异方差性。虽然它们都是违背了基本假定,但前者属于解释变量之间存在的问题,后者是随机误差项出现的问题。本章将讨论异方差性的实质、异方差出现的原因、异方差的后果,并介绍检验和修正异方差的若干方法。

第一节 异方差性的概念

一、异方差性的实质

第二章提出的基本假定中，要求对所有的 i ($i=1, 2, \dots, n$) 都有

$$Var(u_i) = \sigma^2 \quad (5.1)$$

也就是说 u_i 具有同方差性。这里的方差 σ^2 度量的是随机误差项围绕其均值的分散程度。由于 $E(u_i) = 0$ ，所以等价地说，方差 σ^2 度量的是被解释变量 Y 的观测值围绕回归线 $E(Y_i) = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ 的分散程度，同方差性实际指的是相对于回归线被解释变量所有观测值的分散程度相同。

设模型为

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i \quad i = 1, 2, \dots, n \quad (5.2)$$

如果其它假定均不变，但模型中随机误差项 u_i 的方差为

$$Var(u_i^2) = \sigma_i^2, \quad (i = 1, 2, 3, \dots, n). \quad (5.3)$$

则称 u_i 具有异方差性。

由于异方差性指的是被解释变量观测值的分散程度是随解释变量的变化而变化的，如图 5.1 所示，所以进一步可以把异方差看成是由于某个解释变量的变化而引起的，则

$$Var(u_i^2) = \sigma_i^2 = \sigma^2 f(X_i) \quad (5.4)$$

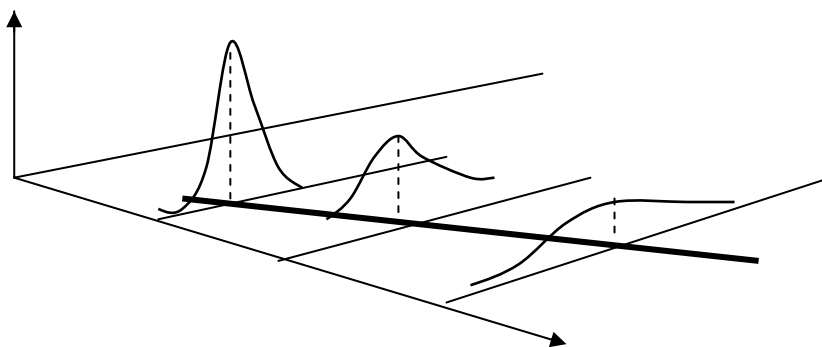


图 5.1

二、产生异方差的原因

由于现实经济活动的错综复杂性，一些经济现象的变动与同方差性的假定经常是相悖的。所以在计量经济分析中，往往会出现某些因素随其观测值的变化而对被解释变量产生不同的影响，导致随机误差项的方差相异。通常产生异方差有以下主要原因：

1、模型中省略了某些重要的解释变量

异方差性表现在随机误差上，但它的产生却与解释变量的变化有紧密的关系。如果计量模型本来应当为 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ ，假如被略去了 X_{3i} ，而采用了

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i^* \quad (5.5)$$

当被略去的 X_{3i} 与 X_{2i} 有呈同方向或反方向变化的趋势时， X_{3i} 随 X_{2i} 的有规律变化会体现在 (5.5) 式的 u_i^* 中。如果将某些未在模型中出现的重要影响因素归入随机误差项，而且这些影响因素的变化具有差异性，则会对被解释变量产生不同的影响，从而导致误差项的方差随之变化，即产生异方差性。在第四章已经讨论过，可以通过剔除变量的方法去避免多重共线性的影响，但是如果删除了重要的变量又有可能引起异方差性。这是在建模过程中应当引起注意的问题。

2、模型设定误差

模型的设定主要包括变量的选择和模型数学形式的确定。模型中略去了重要解释变量常常导致异方差，实际就是模型设定问题。除此而外，模型的函数形式不正确，如把变量间本来为非线性的关系设定为线性，也可能导致异方差。

3、测量误差的变化

样本数据的观测误差有可能随研究范围的扩大而增加，或随时间的推移逐步积累，也可能随着观测技术的提高而逐步减小。例如生产函数模型，由于生产要素投入的增加与生产规模相联系，在其他条件不变的情况下，测量误差可能会随生产规模的扩大而增加，随机误差项的方差会随资本和劳动力投入的增加而变化。另一方面当用时间序列数据估计生产函数时，由于抽样技术和数据收集处理方法的改进，观测误差有可能会随着时间的推移而降低。

4、截面数据中总体各单位的差异

通常认为，截面数据较时间序列数据更容易产生异方差。例如，运用截面数据研究消费和收入之间的关系时，如果采取不同家庭收入组的数据，低收入组的家庭用于购买生活必需品的比例相对较大，消费的分散程度不大，组内各家庭消费的差异也较小。高收入组的家庭有更多自由支配的收入，家庭消费有更广泛的选择范围，消费的分散程度较大，组内各家庭

消费的差异也较大。这种不同收入组家庭的消费偏离均值程度的差异，最终反映为随机误差项偏离其均值的程度有变化，而出现异方差。异方差性在截面数据中比在时间序列数据中可能更常出现，这是因为同一时点不同对象的差异，一般说来会大于同一对象不同时间的差异。不过，在时间序列数据发生较大变化的情况下，也可能出现比截面数据更严重的异方差。

以上只是对产生异方差的经验总结，在建立计量经济学模型的过程中，具体是什么原因产生异方差，应对变量的经济意义和数据所表现出的特征进行认真地分析。

第二节 异方差性的后果

在计量经济分析中，如果模型里存在异方差，则对模型会产生以下后果。

一、对参数估计式统计特性的影响

1、参数的 OLS 估计仍然具有无偏性

由第二章参数估计的统计特性可知，参数 OLS 估计的无偏性仅依赖于基本假定中随机误差项的零均值假定（即 $E(u_i) = 0$ ），以及解释变量的非随机性，异方差的存在并不影响参数估计式的无偏性。

2、参数 OLS 估计式的方差不再是最小的

在模型参数的所有线性估计式中，OLS 估计方差最小的重要前提条件之一是随机误差项为同方差，如果随机误差项是异方差的，将不能再保证最小二乘估计的方差最小。事实上可以证明，能够找到比 OLS 估计的方差更小的估计方法，本章第四节将会介绍这类估计方法。也就是说，在异方差存在时，虽然 OLS 估计仍保持线性无偏性和一致性，但已失去了有效性，即参数的 OLS 估计量不再具有最小方差。（证明见本章附录 5.1）。

二、对参数显著性检验的影响

在 u_i 存在异方差时，OLS 估计式不再具有最小方差，如果仍然用不存在异方差性时的 OLS 方式估计其方差，例如在一元回归时仍用 $Var(\hat{\beta}_2) = \sigma^2 / \sum x_i^2$ 去估计参数估计式的方差，将会低估存在异方差时的真实方差，从而低估 $SE(\hat{\beta}_2)$ ，这将导致夸大用于参数显著性检验的 t 统计量。如果仍用夸大的 t 统计量进行参数的显著性检验，可能造成本应接受的原假设被错误的拒绝，从而夸大所估计参数的统计显著性。

三、对预测的影响

尽管参数的 OLS 估计量仍然无偏，并且基于此的预测也是无偏的，但是由于参数估计量不是有效的，从而对 Y 的预测也将不是有效的。在 u_i 存在异方差时， σ_i^2 与 X_i 的变化有关，参数 OLS 估计的方差 $Var(\hat{\beta}_k)$ 不能唯一确定， Y 预测区间的建立将发生困难。而且 $Var(\hat{\beta}_k)$ 会增大， Y 预测值的精确度也将会下降。

异方差性的存在，会对回归模型的正确建立和统计推断带来严重后果，因此在计量经济分析中，有必要检验模型是否存在异方差。

第三节 异方差性的检验

要检验模型中是否有异方差，需要了解随机误差项 u_i 的概率分布。由于随机误差很难直接观测，只能对随机误差的分布特征进行某种推测，因此对异方差性的检验还没有完全可靠的准则，只能针对产生异方差不同原因的假设，提出一些检验异方差的经验办法。本节只介绍一些最常用的方法。

一、图示检验法

1、相关图形分析

方差描述的是随机变量相对其均值的离散程度，而被解释变量 Y 与随机误差项 u 有相同的方差，所以分析 Y 与 X 的相关图形，可以粗略地看到 Y 的离散程度及与 X 之间是否有相关关系。如果随着 X 的增加， Y 的离散程度有逐渐增大（或减小）的变化趋势，则认为存在递增型（或递减型）的异方差。通常在建立回归模型时，为了判断模型的函数形式，需要观测 Y 与 X 的相关图形，同时也可利用相关图形大致判断模型是否存在异方差性。例如，用 1998 年四川省各地市州农村居民家庭消费支出与家庭纯收入的数据（表 5.2），绘制出消费支出对纯收入的散点图（图 5.2），其中用 $y1$ 表示农村家庭消费支出， $x1$ 表示家庭纯收入。

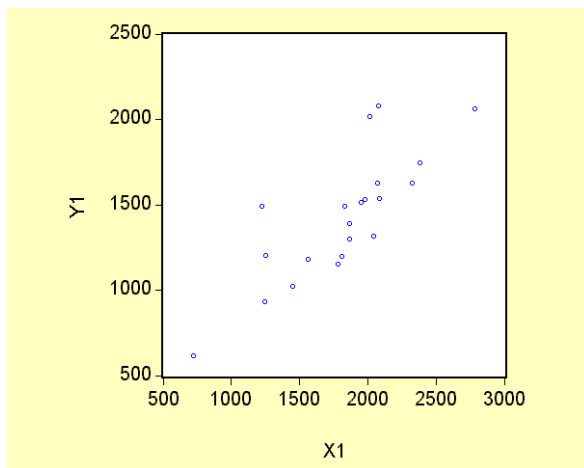


图 5.2

2、残差图形分析

虽然随机误差项无法观测，但样本回归的残差一定程度上反映了随机误差的某些分布特征，可通过残差的图形对异方差性作观察。例如，一元线性回归模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$ ，在 OLS 估计基础上得到残差的平方 e_i^2 ，然后绘制出 e_i^2 对 X_i 的散点图，如果 e_i^2 不随 X_i 而变化，如图 5.3a 所示，则表明 u_i 不存在异方差；如果 e_i^2 随 X_i 而变化，如图 5.3b、c、d 所示，则表明 u_i 存在异方差。

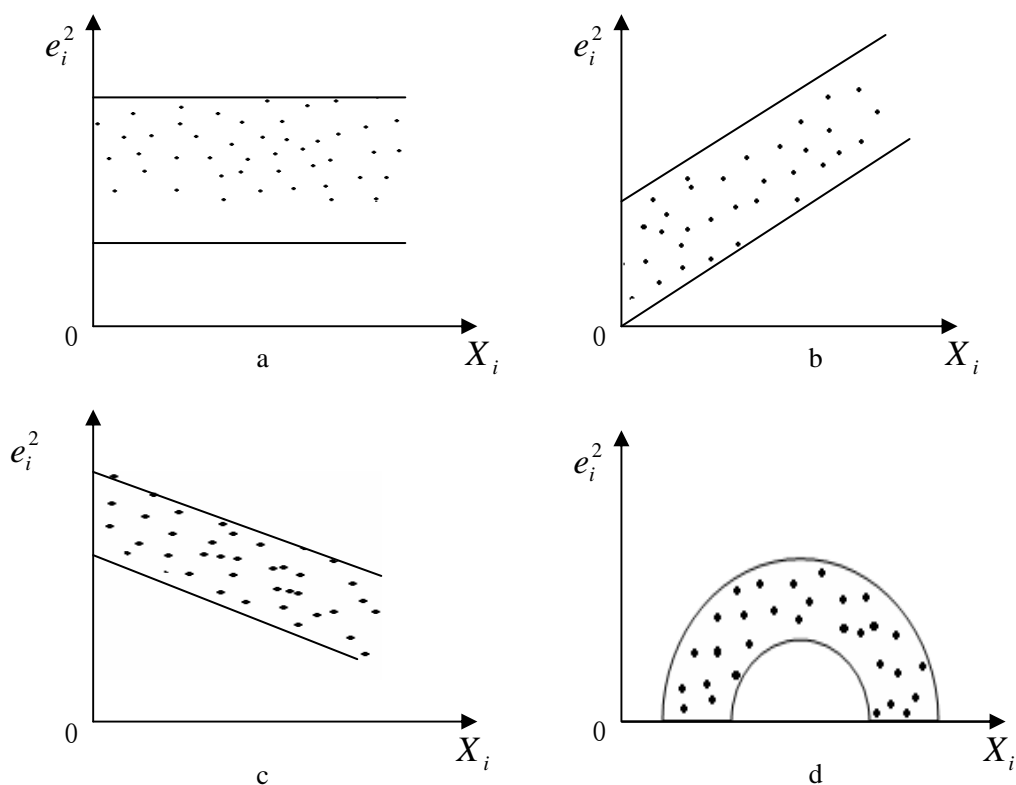


图 5.3

图形法的特点是简单易操作，不足是对异方差性的判断比较粗糙，由于引起异方差性的原因错综复杂，仅靠图形法有时很难准确对是否存在异方差下结论，还需要采用其他统计检验方法。

二、戈德菲尔德-夸特（Goldfeld-Quanadt）检验

该检验方法是戈德菲尔德和夸特于 1965 年提出的，可用于检验递增性或递减性异方差。此检验的基本思想是将样本分为两部分，然后分别对两个样本进行回归，并计算比较两个回归的剩余平方和是否有明显差异，以此判断是否存在异方差。

1、检验的前提条件

- (1) 此检验只适用于大样本。
- (2) 除了同方差假定不成立外，其它假定均满足。

2、检验的具体做法

- (1) 将观测值按解释变量 X_i 的大小顺序排序。
- (2) 将排列在中间的 C 个（约 $1/4$ ）的观察值删除掉，再将剩余的观测值分为两个部分，每部分观察值的个数为 $(n-c)/2$ 。
- (3) 提出假设。即 H_0 : 两部分数据的方差相等； H_1 : 两部分数据的方差不相等。
- (4) 构造 F 统计量。分别对上述两个部分的观察值作回归，由此得到的两个部分的残差平方和，以 $\sum e_{li}^2$ 表示前一部分样本回归产生的残差平方和，以 $\sum e_{2i}^2$ 表示后一部分样本回归产生的残差平方和，它们的自由度均为 $[(n-c)/2]-k$ ， k 为参数的个数。在原假设成立的条件下，因 $\sum e_{li}^2$ 和 $\sum e_{2i}^2$ 分别服从自由度均为 $[(n-c)/2]-k$ 的 χ^2 分布¹，可导出

$$F^* = \frac{\sum e_{2i}^2 / [\frac{n-c}{2} - k]}{\sum e_{li}^2 / [\frac{n-c}{2} - k]} = \frac{\sum e_{2i}^2}{\sum e_{li}^2} \sim F(\frac{n-c}{2} - k, \frac{n-c}{2} - k) \quad (5.7)$$

- (5) 判断。给定显著性水平 α ，查 F 分布表，得临界值 $F_{(\alpha)} = F_{(\alpha)}(\frac{n-c}{2} - k, \frac{n-c}{2} - k)$ 。计算统计量 F^* ，如果 $F^* > F_{(\alpha)}$ ，则拒绝原假设，不拒绝备择假设，即认为模型中的随机误差存在异方差。反之，如果 $F^* < F_{(\alpha)}$ ，则不拒绝原假设，认为模型中随机误差项不存在异方差。

戈德菲尔德-夸特检验的功效, 一是与对观测值的正确排序有关; 二是与删除数据的个数 c 的大小有关。经验认为, 当 $n=30$ 时, 可以取 $c=4$; 当 $n=60$ 时, 可以取 $c=10$ 为宜。该方法得到的只是异方差是否存在的判断, 在多个解释变量的情况下, 对判断是哪一个变量引起异方差还存在局限。

三、White 检验

White 检验的基本思想是, 如果存在异方差, 其方差 σ_t^2 与解释变量有关系, 分析 σ_t^2 是否与解释变量的某些形式有联系可判断异方差性。但是 σ_t^2 一般是未知的, 可用 OLS 估计的残差平方 e_t^2 作为其估计值。在大样本的情况下, 作 e_t^2 对常数项、解释变量、解释变量的平方及其交叉乘积等所构成辅助回归, 利用辅助回归相应的检验统计量, 即可判断是否存在异方差性。

例如, 二元线性回归模型为

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (5.8)$$

并且, 设异方差与 X_{2t}, X_{3t} 的一般关系为

$$\sigma_t^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t \quad (5.9)$$

其中 v_t 为随机误差项。White 检验的基本步骤如下

- 1、用 OLS 法估计 (5.12) 式, 计算残差 $e_t = Y_t - \hat{Y}_t$, 并求残差的平方 e_t^2 。
- 2、用残差平方 e_t^2 作为异方差 σ_t^2 的估计, 并作 e_t^2 对 $X_{2t}, X_{3t}, X_{2t}^2, X_{3t}^2, X_{2t} X_{3t}$ 的辅助回归, 即

$$\hat{e}_t^2 = \hat{\alpha}_1 + \hat{\alpha}_2 x_{2t} + \hat{\alpha}_3 x_{3t} + \hat{\alpha}_4 x_{2t}^2 + \hat{\alpha}_5 x_{3t}^2 + \hat{\alpha}_6 x_{2t} x_{3t} \quad (5.10)$$

式中 \hat{e}_t^2 表示 e_t^2 的估计。

- 3、计算统计量 nR^2 , 其中 n 为样本容量, R^2 为辅助回归的可决系数。
- 4、在 $H_0: \alpha_2 = \dots = \alpha_6 = 0, H_1: \alpha_j \ (j = 2, 3, \dots, 6)$ 中至少有一个不为零的原假设下, 可证明, nR^2 渐近地服从自由度为 5 的 χ^2 分布。给定显著性水平 α , 查 χ^2 分布表得临界值 $\chi_\alpha^2(5)$, 如果 $nR^2 > \chi_\alpha^2(5)$, 则拒绝原假设, 表明模型中随机误差存在异方差。

¹ 可参阅 [美] J.M.伍德里奇著《计量经济学导论》, 中国人民大学出版社, 2003, 第 240 页。

White 检验的特点是，不仅能够检验异方差的存在性，同时，在多变量的情况下，还能判断出是哪一变量引起的异方差。此方法不需要异方差的先验信息，但要求观测值为大样本。

四、ARCH 检验

通常，人们在做计量经济分析时对截面数据产生异方差给予足够的关注，而放松了对时间序列数据产生异方差的警惕。恩格尔（Engel）于 1982 年提出了在时间序列背景下也有可能出现异方差性，并从理论上提出了一种观测时间序列方差变动的方法，这就是所谓的 ARCH（AutoRegressive Conditional Heteroscedasticity）检验方法。ARCH 检验的思想是，在时间序列数据中，可认为存在的异方差性为 ARCH（自回归条件异方差）过程，并通过检验这一过程是否成立去判断时间序列是否存在异方差。

1、ARCH 过程

设 ARCH 过程为

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \cdots + \alpha_p \sigma_{t-p}^2 + v_t \quad (5.11)$$

式中 p 为 ARCH 过程的阶数，并且 $\alpha_0 > 0, \alpha_i \geq 0, (i = 1, 2, \cdots, p)$ ； v_t 为随机误差。

2、ARCH 检验的基本步骤

(1) 提出原假设：

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0; H_1 : \alpha_j \quad (j = 1, 2, \cdots, p) \text{ 中至少有一个不为零。}$$

(2) 对原模型作 OLS 估计，求出残差 e_t ，并计算残差平方序列 $e_t^2, e_{t-1}^2, \cdots, e_{t-p}^2$ ，以分别作为对 $\sigma_t^2, \sigma_{t-1}^2, \cdots, \sigma_{t-p}^2$ 的估计。

(3) 作辅助回归

$$\hat{e}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 e_{t-1}^2 + \cdots + \hat{\alpha}_p e_{t-p}^2 \quad (5.12)$$

式中 \hat{e}_t^2 表示 e_t^2 的估计。

(4) 计算式 (5.12) 辅助回归的可决系数 R^2 ，可以证明在 H_0 成立下，基于大样本，有 $(n-p)R^2$ 渐近服从 $\chi^2(p)$ ， p 为自由度，亦即式 (5.11) 中变量的滞后期数；给定显著性水平 α ，查 χ^2 分布表得临界值 $\chi_{\alpha}^2(p)$ ，如果 $(n-p)R^2 > \chi_{\alpha}^2(p)$ ，则拒绝原假设，表明模型中的随机误差项存在异方差³。

³ 陆懋祖，高等时间序列经济计量学，上海人民出版社，1999 年，第 300 页。

ARCH 检验的特点是，要求变量的观测值为大样本，并且是时间序列数据；它只能判断模型中是否存在异方差，而不能诊断出是哪一个变量引起的异方差。

五、Glejser 检验

Glejser 检验的基本思想是，由 OLS 法得到残差 e_i ，取 e_i 的绝对值 $|e_i|$ ，然后将 $|e_i|$ 对某个解释变量 X_i 回归，根据回归模型的显著性和拟合优度来判断是否存在异方差。该检验的特点是不仅能对异方差的存在进行判断，而且还能对异方差随某个解释变量变化的函数形式进行诊断。该检验要求变量的观测值为大样本。

Glejser 检验的具体步骤：

(1) 根据样本数据建立回归模型，并求残差序列 $e_i = Y_i - \hat{Y}_i$ 。

(2) 用残差绝对值 $|e_i|$ 对 X_i 的进行回归，由于 $|e_i|$ 与 X 的真实函数形式并不知道，可用各种函数形式去试验，从中选择最佳形式。Glejser 曾提出如下一些假设的函数形式：

$$|e_i| = \beta X_i + v_i; |e_i| = \alpha + \beta X_i + v_i; |e_i| = \beta \sqrt{X_i} + v_i; |e_i| = \beta \frac{1}{X_i} + v_i; |e_i| = \beta \frac{1}{\sqrt{X_i}} + v_i,$$

其中 v 为随机误差项。

(3) 根据选择的函数形式作 $|e_i|$ 对 X_i 的回归，用回归所得到的 R^2 、 t 、 F 等信息判断，若表明参数 β 显著不为零，即认为存在异方差性。

上述各种检验方法，很难说哪一种方法最为有效。这些检验方法的共同思想是，基于不同的假定，分析随机误差项的方差与解释变量之间的相关性，以判断随机误差项的方差是否随解释变量而变化。其中有的检验方法还能提供随机误差项的方差与解释变量之间关系的某些信息，这些信息对补救异方差性可能是有价值的。

第四节 异方差性的补救措施

通过检验如果证实存在异方差，则需要采取措施对异方差性进行修正，基本思想是采用适当的估计方法，消除或减小异方差对模型的影响。

一、对模型变换

当可以确定异方差的具体形式时，将模型作适当变换有可能消除或减轻异方差的影响。

以一元线性回归模型为例

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (5.13)$$

经检验 u_i 存在异方差，并已知 $\text{var}(u_i) = \sigma_i^2 = \sigma^2 f(X_i)$ ，其中 σ^2 为常数， $f(X_i)$ 为 X_i 的某种函数。显然，当 $f(X_i)$ 是常数时， u_i 为同方差，当 $f(X_i)$ 不是常数时， u_i 为异方差。

为变换模型，用 $\sqrt{f(X_i)}$ 去除 (5.20) 式的两端，得

$$\frac{Y_i}{\sqrt{f(X_i)}} = \frac{\beta_1}{\sqrt{f(X_i)}} + \beta_2 \frac{X_i}{\sqrt{f(X_i)}} + \frac{u_i}{\sqrt{f(X_i)}} \quad (5.14)$$

记 $Y_i^* = \frac{Y_i}{\sqrt{f(X_i)}}$ ； $X_i^* = \frac{X_i}{\sqrt{f(X_i)}}$ ； $\beta_1^* = \frac{\beta_1}{\sqrt{f(X_i)}}$ ； $v_i = \frac{u_i}{\sqrt{f(X_i)}}$ ，则有

$$Y_i^* = \beta_1^* + \beta_2 X_i^* + v_i \quad (5.15)$$

(5.15) 式的随机误差项为 v_i 的方差为

$$\text{var}(v_i) = \text{var}\left(\frac{u_i}{\sqrt{f(X_i)}}\right) = \frac{1}{f(X_i)} \text{var}(u_i) = \sigma^2 \quad (5.16)$$

可见，经变换后的 (5.15) 式的随机误差项 $v_i = \frac{u_i}{\sqrt{f(X_i)}}$ 已是同方差。

根据图示法或 Glejser 检验所得到的相应信息，可以对 $f(X_i)$ 的函数形式作出各种假定，

常见的 $f(X_i)$ 形式有以下几种：

(1) 设 $f(X_i) = X_i$ ，即 $\text{var}(u_i) = \sigma^2 X_i$ ，这时对式 (5.13) 两端同除 $\sqrt{X_i}$ ，得

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}} \quad (5.17)$$

令 $v_i = \frac{u_i}{\sqrt{X_i}}$ ，则 $\text{var}(v_i)$ 为同方差。因为

$$\text{var}(v_i) = \text{var}\left(\frac{u_i}{\sqrt{X_i}}\right) = \frac{1}{X_i} \text{var}(u_i) = \sigma^2 \quad (5.18)$$

(2) 设 $f(X_i) = X_i^2$ ，则 $\text{var}(u_i) = \sigma^2 X_i^2$ ，同理，得

$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 \frac{X_i}{X_i} + \frac{u_i}{X_i} \quad (5.19)$$

令 $v_i = \frac{u_i}{X_i}$ ，则 $\text{var}(v_i)$ 为同方差。因为

$$\text{var}(v_i) = \text{var}\left(\frac{u_i}{X_i}\right) = \frac{1}{X_i^2} \text{var}(u_i) = \sigma^2 \quad (5.20)$$

(3) 设 $f(X_i) = (a_0 + a_1 X_i)^2$ ，则 $\text{var}(u_i) = \sigma^2 (a_0 + a_1 X_i)^2$ 。同理有

$$\frac{Y_i}{a_0 + a_1 X_i} = \beta_1 \frac{1}{a_0 + a_1 X_i} + \beta_2 \frac{X_i}{a_0 + a_1 X_i} + \frac{u_i}{a_0 + a_1 X_i} \quad (5.21)$$

令 $v_i = \frac{u_i}{a_0 + a_1 X_i}$ ，则 $\text{var}(v_i)$ 为同方差。因为

$$\text{var}(v_i) = \text{var}\left(\frac{u_i}{a_0 + a_1 X_i}\right) = \frac{1}{(a_0 + a_1 X_i)^2} \text{var}(u_i) = \sigma^2 \quad (5.22)$$

二、加权最小二乘法

为了便于说明问题，以一元线性回归模型为例

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (5.23)$$

且存在异方差的形式为 $\text{var}(u_i) = \sigma_i^2 = \sigma^2 f(X_i)$ ，其中 σ^2 为常数， $f(X_i)$ 为 X_i 的某种函数。对 (5.20) 式按照最小二乘法的基本原则，是使残差平方和 $\sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$ 为最小。在同方差性假定下，普通最小二乘法是把每个残差平方 $e_i^2 (i=1,2,\dots,n)$ 都同等看待，都赋予相同的权数 1。但是，当存在异方差性时，方差 σ_i^2 越小，其样本值偏离均值的程度越小，其观测值越应受到重视。即方差越小，在确定回归线时的作用应当越大；反之方差 σ_i^2 越大，其样本值偏离均值的程度越大，其观测值所起的作用应当越小。也就是说，在拟合存在异方差的模型的回归线时，对不同的 σ_i^2 应该区别对待。从样本的角度，对较小的 e_i^2 给予较大的权数，对较大的 e_i^2 给予较小的权数，从而使 $\sum e_i^2$ 更好地反映 σ_i^2 对残差平方和的影响。通常可将权数取为 $w_i = 1/\sigma_i^2 (i=1,2,\dots,n)$ ，由此，当 σ_i^2 越小时， w_i 越大，当 σ_i^2 越大时， w_i 就越小。将权数与残差平方相乘以后再求和，得

$$\sum w_i e_i^2 = \sum w_i (Y_i - \beta_1^* - \beta_2^* X_i)^2 \quad (5.24)$$

(5.24) 式称为加权的残差平方和。根据最小二乘原理，若使得加权的残差平方和最小，即

$$\min : \sum w_i e_i^2 = \sum w_i (Y_i - \beta_1^* - \beta_2^* X_i)^2 \quad (5.25)$$

可得

$$\begin{aligned} \hat{\beta}_1^* &= \bar{Y}^* - \hat{\beta}_2^* \bar{X}^* \\ \hat{\beta}_2^* &= \frac{\sum w_i (X_i - \bar{X}^*)(Y_i - \bar{Y}^*)}{\sum w_i (X_i - \bar{X}^*)^2} \end{aligned} \quad (5.26)$$

其中 $\bar{X}^* = \frac{\sum w_i X_i}{\sum w_i}$, $\bar{Y}^* = \frac{\sum w_i Y_i}{\sum w_i}$ 。这样估计的参数 β_1^* 和 β_2^* 称为加权最小二乘估计。这种

求解参数估计式的方法为加权最小二乘法(Weighted Least Square, 简称 WLS)。

容易证明，对原模型变换的方法与加权最小二乘法实际上是等价的。例如以 (5.23) 式的一元线性模型为例，如果已知存在异方差，且 $\text{var}(u_i) = \sigma_i^2 = \sigma^2 f(X_i)$ ，变换后的模型为

$$\frac{Y_i}{\sqrt{f(X_i)}} = \frac{\beta_1}{\sqrt{f(X_i)}} + \beta_2 \frac{X_i}{\sqrt{f(X_i)}} + \frac{u_i}{\sqrt{f(X_i)}} \quad (5.27)$$

由前面的讨论知，(5.27)式的随机误差项 $u_i/\sqrt{f(X_i)}$ 已是同方差的。用 OLS 法估计(5.27)式的参数，其剩余平方和为

$$\sum e_i^2 = \sum \left(\frac{Y_i}{\sqrt{f(X_i)}} - \frac{\hat{\beta}_1}{\sqrt{f(X_i)}} - \hat{\beta}_2 \frac{X_i}{\sqrt{f(X_i)}} \right)^2 = \sum \frac{1}{f(X_i)} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad (5.28)$$

当对(5.23)式采用加权最小二乘法时，其权数为 $w_i = 1/\sigma_i^2 = 1/\sigma^2 f(X_i)$ ($i=1,2,\dots,n$)，其残差平方和为

$$\sum \left(\frac{e_i}{\sigma_i} \right)^2 = \sum \frac{1}{\sigma_i^2} (Y_i - \beta_1^* - \beta_2^* X_i)^2 = \sum \frac{1}{\sigma^2 f(X_i)} (Y_i - \beta_1^* - \beta_2^* X_i)^2 \quad (5.29)$$

将(5.28)式模型变换的残差平方和与(5.29)式加权最小二乘的残差平方和加以对比，可以看出二者的剩余平方和只相差常数因子 σ^2 ，能使其中一个最小时必能使另一个最小。对模型变换后用 OLS 估计其参数，实际与应用加权最小二乘法估计的参数是一致的。这也间接证明了加权最小二乘法可以消除异方差。只是对原模型变换后的模型拟合优度有可能变小，这是由于对样本观测值加权的结果。

三、模型的对数变换

在经济意义成立的情况下，如果对（5.13）式的模型作对数变换，其变量 Y_i 和 X_i 分别用 $\ln Y_i$ 和 $\ln X_i$ 代替，即

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad (5.30)$$

对数变换后的模型通常可以降低异方差性的影响。

首先，运用对数变换能使测定变量值的尺度缩小。它可以将两个数值之间原来 10 倍的差异缩小到只有 2 倍的差异。例如，100 是 10 的 10 倍，但在常用对数情况下， $\lg 100=2$ 是 $\lg 10=1$ 的两倍；再例如，80 是 8 的 10 倍，但在自然对数情况下， $\ln 80=4.3820$ 是 $\ln 8=2.0794$ 的两倍多。

其次，经过对数变换后的线性模型，其残差 e 表示相对误差（证明见附录 5.2），而相对误差往往比绝对误差有较小的差异。

但是特别要注意的是，对变量取对数虽然能够减少异方差对模型的影响，但应注意取对数后变量的经济意义。如果变量之间在经济意义上并非呈对数线性关系，则不能简单地对变量取对数，这时只能用其它方法对异方差进行修正。

第五节 案例分析

一、问题的提出和模型设定

根据本章引子提出的问题，为了给制定医疗机构的规划提供依据，分析比较医疗机构与人口数量的关系，建立卫生医疗机构数与人口数的回归模型。假定医疗机构数与人口数之间满足线性约束，则理论模型设定为

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (5.31)$$

其中 Y_i 表示卫生医疗机构数， X_i 表示人口数。由 2001 年《四川统计年鉴》得到如下数据。

表 5.1 四川省 2000 年各地区医疗机构数与人口数

地区	人口数（万人）	医疗机构数（个）	地区	人口数（万人）	医疗机构数（个）
	X	Y		X	Y
成都	1013.3	6304	眉山	339.9	827
自贡	315	911	宜宾	508.5	1530

攀枝花	103	934	广安	438.6	1589
泸州	463.7	1297	达州	620.1	2403
德阳	379.3	1085	雅安	149.8	866
绵阳	518.4	1616	巴中	346.7	1223
广元	302.6	1021	资阳	488.4	1361
遂宁	371	1375	阿坝	82.9	536
内江	419.9	1212	甘孜	88.9	594
乐山	345.9	1132	凉山	402.4	1471
南充	709.2	4064			

二、参数估计

进入 EViews 软件包，确定时间范围；编辑输入数据；选择估计方程菜单，估计样本回归函数如下

表 5.2

Dependent Variable: Y Method: Least Squares Date: 07/09/05 Time: 11:11 Sample: 1 21 Included observations: 21				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-563.0548	291.5778	-1.931062	0.0685
X	5.373498	0.644284	8.340265	0.0000
R-squared	0.785456	Mean dependent var	1588.238	
Adjusted R-squared	0.774164	S.D. dependent var	1311.037	
S.E. of regression	623.0330	Akaike info criterion	15.79747	
Sum squared resid	7375233.	Schwarz criterion	15.89695	
Log likelihood	-163.8734	F-statistic	69.56003	
Durbin-Watson stat	0.429831	Prob(F-statistic)	0.000000	

估计结果为

$$\hat{Y}_i = -563.0548 + 5.3735X_i$$

$$(-1.9311) \quad (8.3403) \quad (5.32)$$

$$R^2 = 0.7855, s.e. = 508.2665, F = 69.56$$

括号内为 t 统计量值。

三、检验模型的异方差

本例用的是四川省 2000 年各地市州的医疗机构数和人口数，由于地区之间存在的不同人口数，因此，对各种医疗机构的设置数量会存在不同的需求，这种差异使得模型很容易产生异方差，从而影响模型的估计和运用。为此，必须对该模型是否存在异方差进行检验。

（一）图形法

1、EViews 软件操作。

由路径：Quick/Qstimate Equation，进入 Equation Specification 窗口，键入 “y c x”，确认并 “ok”，得样本回归估计结果，见表 5.2。

(1) 生成残差平方序列。在得到表 5.2 估计结果后，立即用生成命令建立序列 e_i^2 ，记为 e2。生成过程如下，先按路径：Procs/Generate Series，进入 Generate Series by Equation 对话框，即

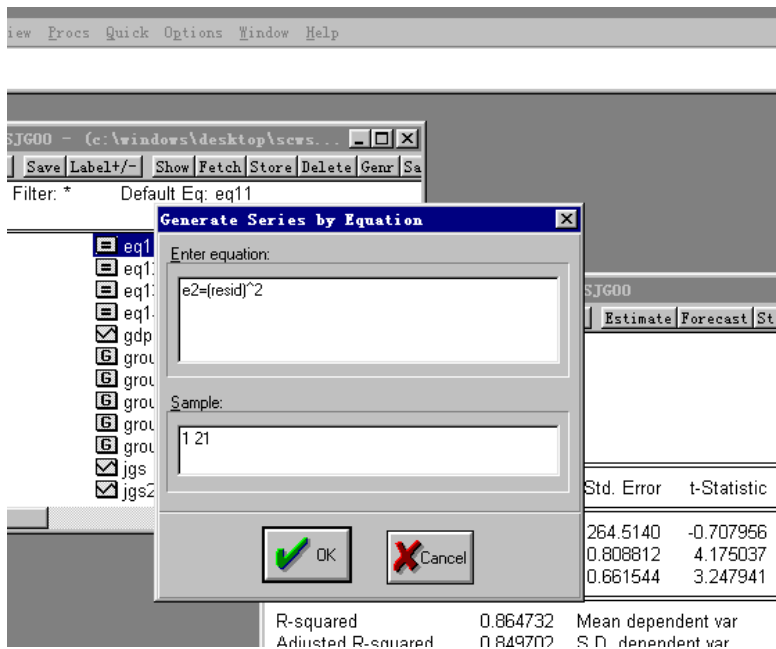


图 5.4

然后，在 Generate Series by Equation 对话框中（如图 5.4），键入 “e2= (resid) ^2”，则生成序列 e_i^2 。

(2) 绘制 e_i^2 对 X_i 的散点图。选择变量名 X 与 e2（注意选择变量的顺序，先选的变量将在图形中表示横轴，后选的变量表示纵轴），进入数据列表，再按路径 view/graph/scatter，可得散点图，见图 5.5。

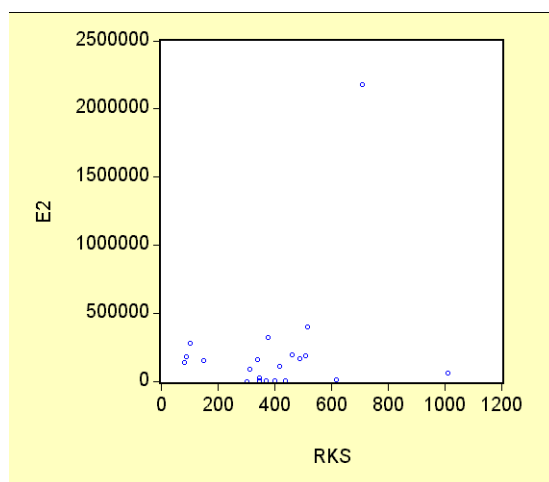


图 5.5

2、判断。由图 5.5 可以看出，残差平方 e_i^2 对解释变量 X 的散点图主要分布在图形中的下三角部分，大致看出残差平方 e_i^2 随 X_i 的变动呈增大的趋势，因此，模型很可能存在异方差。但是否确实存在异方差还应通过更进一步的检验。

(二) Goldfeld-Quanadt 检验

1、EViews 软件操作。

(1) 对变量取值排序（按递增或递减）。在 Procs 菜单里选 Sort Series 命令，出现排序对话框，如果以递增型排序，选 Ascending，如果以递减型排序，则应选 Descending，键入 X，点 ok。本例选递增型排序，这时变量 Y 与 X 将以 X 按递增型排序。

(2) 构造子样本区间，建立回归模型。在本例中，样本容量 $n=21$ ，删除中间 1/4 的观测值，即大约 5 个观测值，余下部分平分得两个样本区间：1—8 和 14—21，它们的样本个数均是 8 个，即 $n_1 = n_2 = 8$ 。

在 Sample 菜单里，将区间定义为 1—8，然后用 OLS 方法求得如下结果

表 5.3

Dependent Variable: Y				
Method: Least Squares				
Date: 07/09/05 Time: 11:14				
Sample: 1 8				
Included observations: 8				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	598.2525	119.2922	5.015018	0.0024
X	1.177650	0.490187	2.402452	0.0531
R-squared	0.490306	Mean dependent var	852.6250	
Adjusted R-squared	0.405357	S.D. dependent var	201.5667	
S.E. of regression	155.4343	Akaike info criterion	13.14264	
Sum squared resid	144958.9	Schwarz criterion	13.16250	
Log likelihood	-50.57056	F-statistic	5.771775	
Durbin-Watson stat	1.656269	Prob(F-statistic)	0.053117	

在 Sample 菜单里, 将区间定义为 14—21, 再用 OLS 方法求得如下结果

表 5.4

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: Y									
Method: Least Squares									
Date: 07/09/05 Time: 11:16									
Sample: 14 21									
Included observations: 8									
Variable	Coefficient	Std. Error	t-Statistic	Prob.					
C	-2941.087	430.3991	-6.833395	0.0005					
X	9.179365	0.692831	13.24907	0.0000					
R-squared	0.966949	Mean dependent var	2520.750						
Adjusted R-squared	0.961441	S.D. dependent var	1781.608						
S.E. of regression	349.8466	Akaike info criterion	14.76518						
Sum squared resid	734355.8	Schwarz criterion	14.78504						
Log likelihood	-57.06074	F-statistic	175.5379						
Durbin-Watson stat	1.812612	Prob(F-statistic)	0.000011						

(3) 求 F 统计量值。基于表 5.3 和表 5.4 中残差平方和的数据, 即 Sum squared resid 的值。由表 5.3 计算得到的残差平方和为 $\sum e_{1i}^2 = 144958.9$, 由表 5.4 计算得到的残差平方和为 $\sum e_{2i}^2 = 734355.8$, 根据 Goldfeld-Quanadt 检验, F 统计量为

$$F = \frac{\sum e_{2i}^2}{\sum e_{1i}^2} = \frac{734355.8}{144958.9} = 5.066 \quad (5.33)$$

(4) 判断。在 $\alpha = 0.05$ 下, 式 (5.33) 中分子、分母的自由度均为 6, 查 F 分布表得临界值为 $F_{0.05}(6,6) = 4.28$, 因为 $F = 5.066 > F_{0.05}(6,6) = 4.28$, 所以拒绝原假设, 表明模型确实存在异方差。

(三) White 检验

由表 5.2 估计结果, 按路径 view/residual tests/white heteroskedasticity(no cross terms or cross terms), 进入 White 检验。根据 White 检验中辅助函数的构造, 最后一项

为变量的交叉乘积项，因为本例为一元函数，故无交叉乘积项，因此应选 no cross terms，则辅助函数为

$$\sigma_t^2 = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + v_t \quad (5.34)$$

经估计出现 White 检验结果，见表 5.5。

从表 5.5 可以看出， $nR^2 = 18.0694$ ，由 White 检验知，在 $\alpha = 0.05$ 下，查 χ^2 分布表，得临界值 $\chi_{0.05}^2(2) = 5.9915$ （在（5.34）式中只有两项含有解释变量，故自由度为 2），比较计算的 χ^2 统计量与临界值，因为 $nR^2 = 18.0694 > \chi_{0.05}^2(2) = 5.9915$ ，所以拒绝原假设，不拒绝备择假设，表明模型存在异方差。

表 5.5

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
White Heteroskedasticity Test:									
F-statistic		55.49105	Probability		0.000000				
Obs*R-squared		18.06936	Probability		0.000119				
Test Equation:									
Dependent Variable: RESID^2									
Method: Least Squares									
Date: 07/09/05 Time: 11:18									
Sample: 1 21									
Included observations: 21									
Variable		Coefficient	Std. Error	t-Statistic	Prob.				
C		823726.3	130406.0	6.316626	0.0000				
X		-3607.112	554.1908	-6.508791	0.0000				
X^2		4.743829	0.532983	8.900521	0.0000				
R-squared		0.860446	Mean dependent var		351201.6				
Adjusted R-squared		0.844940	S.D. dependent var		454283.3				
S.E. of regression		178886.3	Akaike info criterion		27.15845				
Sum squared resid		5.76E+11	Schwarz criterion		27.30767				
Log likelihood		-282.1637	F-statistic		55.49105				
Durbin-Watson stat		1.688003	Prob(F-statistic)		0.000000				

四、异方差性的修正

（一）加权最小二乘法（WLS）

在运用 WLS 法估计过程中，我们分别选用了权数 $w_{1t} = \frac{1}{X_t}$, $w_{2t} = \frac{1}{X_t^2}$, $w_{3t} = \frac{1}{\sqrt{X_t}}$ 。权

数的生成过程如下，由图 5.4，在对话框中的 Enter Equation 处，按如下格式分别键入：

$w1 = 1/X$ ； $w2 = 1/X^2$ ； $w3 = 1/\text{sqr}(X)$ ，经估计检验发现用权数 w_{2t} 的效果最好。下

面仅给出用权数 w_{2t} 的结果。

表 5.7

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: Y									
Method: Least Squares									
Date: 07/09/05 Time: 11:24									
Sample: 1 21									
Included observations: 21									
Weighting series: W2									
Variable	Coefficient	Std. Error	t-Statistic	Prob.					
C	368.6090	84.16870	4.379407	0.0003					
X	2.952958	0.822688	3.589402	0.0020					
Weighted Statistics									
R-squared	0.938665	Mean dependent var	808.6991						
Adjusted R-squared	0.935437	S.D. dependent var	1086.410						
S.E. of regression	276.0493	Akaike info criterion	14.16943						
Sum squared resid	1447861.	Schwarz criterion	14.26891						
Log likelihood	-146.7790	F-statistic	12.88381						
Durbin-Watson stat	1.705980	Prob(F-statistic)	0.001955						
Unweighted Statistics									
R-squared	0.625222	Mean dependent var	1588.238						
Adjusted R-squared	0.605497	S.D. dependent var	1311.037						
S.E. of regression	823.4555	Sum squared resid	12883501						
Durbin-Watson stat	0.380523								

表 5.7 的估计结果如下

$$\hat{Y}_i = 368.6090 + 2.9530X_i$$

(4.3794) (3.5894) (5.36)

$$R^2 = 0.9387, D.W. = 1.7060, s.e. = 276.0493, F = 12.8838$$

括号中数据为 t 统计量值。

可以看出运用加权小二乘法消除了异方差性后，参数的 t 检验均显著，可决系数大幅提高，F 检验也显著，并说明人口数量每增加 1 万人，平均说来将增加 2.953 个卫生医疗机构，而不是引子中得出的增加 5.3735 个医疗机构。虽然这个模型可能还存在某些其他需要进一步解决的问题，但这一估计结果或许比引子中的结论更为接近真实情况。

第五章小结

1、异方差性是指模型中随机误差项的方差不是常量，而且它的变化与解释变量的变动有关。

2、产生异方差性的主要原因有：模型中略去的变量随解释变量的变化而呈规律性的变化、变量的设定问题、截面数据的使用，利用平均数作为样本数据等。

3、存在异方差性时对模型的 OLS 估计仍然具有无偏性，但最小方差性不成立，从而导致参数的显著性检验失效和预测的精度降低。

4、检验异方差性的方法有多种，常用的有图形法、Goldfeld-Quandt 检验、White 检

验、ARCH 检验以及 Glejser 检验，运用这些检验方法时要注意它们的假设条件。

5、修正异方差性的主要方法是加权最小二乘法，也可以用变量变换法和对数变换法。变量变换法与加权最小二乘法实际是等价的。

第五章主要公式表

异方差性	$Var(u_i) = \sigma_i^2$
Goldfeld-Quandt 检验 的 F 统计量	$F^* = \frac{\sum e_{2i}^2 / [\frac{n-c}{2} - k]}{\sum e_{1i}^2 / [\frac{n-c}{2} - k]} = \frac{\sum e_{2i}^2}{\sum e_{1i}^2}$
White 检验中的辅助函数 (原模型只有两个解释变量)	$\hat{e}_i^2 = \hat{\alpha}_1 + \hat{\alpha}_2 x_{2i} + \hat{\alpha}_3 x_{3i} + \hat{\alpha}_4 x_{2i}^2 + \hat{\alpha}_5 x_{3i}^2 + \hat{\alpha}_6 x_{2i} x_{3i}$
ARCH 检验中的辅助函数	$\hat{e}_i^2 = \hat{\alpha}_0 + \hat{\alpha}_1 e_{i-1}^2 + \cdots + \hat{\alpha}_p e_{i-p}^2$
Glejser 检验中常用的辅助函数	$ e = \beta X + v; e = \beta \sqrt{X} + v; e = \beta \frac{1}{X} + v;$ $ e = \beta \frac{1}{\sqrt{X}} + v; e = \alpha + \beta X + v$
一元函数下的加权最小二乘估计	$\hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}^*$ $\hat{\beta}_2^* = \frac{\sum w_i (X_i - \bar{X}^*)(Y_i - \bar{Y}^*)}{\sum w_i (X_i - \bar{X}^*)^2}$
一元函数下的对原模型的变换	设 $Y_i = \beta_1 + \beta_2 X_i + u_i$ 并且 $\text{var}(u_i) = \sigma_i^2 = \sigma^2 f(X_i)$ 则 $\frac{Y_i}{\sqrt{f(X_i)}} = \frac{\beta_1}{\sqrt{f(X_i)}} + \beta_2 \frac{X_i}{\sqrt{f(X_i)}} + \frac{u_i}{\sqrt{f(X_i)}}$
对数变换的模型	$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$

思考题与练习题

思考题

- 5.1 简述什么是异方差?为什么异方差的出现总是与模型中某个解释变量的变化有关?
- 5.2 试归纳检验异方差方法的基本思想,并指出这些方法的异同。
- 5.3 什么是加权最小二乘法,它的基本思想是什么?

5.4 产生异方差的原因是什么？试举例说明经济现象中的异方差性。

5.5 如果模型中存在异方差性，对模型有什么影响？这时候模型还能进行应用分析吗？

5.6 对数变化的作用是什么？进行对数变化应注意什么？对数变换后模型的经济意义有什么变化？

5.7 怎样确定加权最小二乘法中的权数？

练习题

5.1 设消费函数为

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

式中， Y_i 为消费支出； X_{2i} 为个人可支配收入； X_{3i} 为个人的流动资产； u_i 为随机误差

项，并且 $E(u_i) = 0, \text{Var}(u_i) = \sigma^2 X_{2i}^2$ （其中 σ^2 为常数）。试回答以下问题：

(1) 选用适当的变换修正异方差，要求写出变换过程；

(2) 写出修正异方差后的参数估计量的表达式。

5.2 根据本章第四节的对数变换，我们知道对变量取对数通常能降低异方差性，但须对这种模型的随机误差项的性质给予足够的关注。例如，设模型为 $Y = \beta_1 X^{\beta_2} u$ ，对该模型中的变量取对数后得如下形式

$$\ln Y = \ln \beta_1 + \beta_2 \ln X + \ln u$$

(1) 如果 $\ln u$ 要有零期望值， u 的分布应该是什么？

(2) 如果 $E(u) = 1$ ，会不会 $E(\ln u) = 0$ ？为什么？

(3) 如果 $E(\ln u)$ 不为零，怎样才能使它等于零？

5.3 由表中给出消费 Y 与收入 X 的数据，试根据所给数据资料完成以下问题：

(1) 估计回归模型 $Y = \beta_1 + \beta_2 X + u$ 中的未知参数 β_1 和 β_2 ，并写出样本回归模型的书写格式；

(2) 试用 Goldfeld-Quandt 法和 White 法检验模型的异方差性；

(3) 选用合适的方法修正异方差。

Y	X	Y	X	Y	X
55	80	152	220	95	140
65	100	144	210	108	145
70	85	175	245	113	150
80	110	180	260	110	160
79	120	135	190	125	165
84	115	140	205	115	180
98	130	178	265	130	185
95	140	191	270	135	190
90	125	137	230	120	200
75	90	189	250	140	205
74	105	55	80	140	210
110	160	70	85	152	220
113	150	75	90	140	225
125	165	65	100	137	230
108	145	74	105	145	240
115	180	80	110	175	245
140	225	84	115	189	250
120	200	79	120	180	260
145	240	90	125	178	265
130	185	98	130	191	270

5.4 由表中给出 1985 年我国北方几个省市农业总产值，农用化肥量、农用水利、农业劳动力、每日生产性固定生产原值以及农机动力数据，要求：

- (1) 试建立我国北方地区农业产出线性模型；
- (2) 选用适当的方法检验模型中是否存在异方差；
- (3) 如果存在异方差，采用适当的方法加以修正。

	农业总产值	农业劳动力	灌溉面积	化肥用量	户均固定	农机动力
地区	(亿元)	(万人)	(万公顷)	(万吨)	资产(元)	(万马力)
北京	19.64	90.1	33.84	7.5	394.3	435.3
天津	14.4	95.2	34.95	3.9	567.5	450.7
河北	149.9	1639.0	357.26	92.4	706.89	2712.6
山西	55.07	562.6	107.9	31.4	856.37	1118.5
内蒙古	60.85	462.9	96.49	15.4	1282.81	641.7
辽宁	87.48	588.9	72.4	61.6	844.74	1129.6
吉林	73.81	399.7	69.63	36.9	2576.81	647.6
黑龙江	104.51	425.3	67.95	25.8	1237.16	1305.8
山东	276.55	2365.6	456.55	152.3	5812.02	3127.9
河南	200.02	2557.5	318.99	127.9	754.78	2134.5
陕西	68.18	884.2	117.9	36.1	607.41	764
新疆	49.12	256.1	260.46	15.1	1143.67	523.3

5.5 表中的数据是美国 1988 研究与开发 (R&D) 支出费用 (Y) 与不同部门产品销售量 (X)。试根据资料建立一个回归模型, 运用 Glejser 方法和 White 方法检验异方差, 由此决定异方差的表现形式并选用适当方法加以修正。

工业群体	单位: 百万美元		
	销售量 X	R&D 费用 Y	利润 Z
1.容器与包装	6375.3	62.5	185.1
2.非银行业金融	11626.4	92.9	1569.5
3.服务行业	14655.1	178.3	276.8
4.金属与采矿	21869.2	258.4	2828.1
5.住房与建筑	26408.3	494.7	225.9
6.一般制造业	32405.6	1083	3751.9
7.休闲娱乐	35107.7	1620.6	2884.1
8.纸张与林木产品	40295.4	421.7	4645.7
9.食品	70761.6	509.2	5036.4
10.卫生保健	80552.8	6620.1	13869.9
11.宇航	95294	3918.6	4487.8
12.消费者用品	101314.3	1595.3	10278.9
13.电器与电子产品	116141.3	6107.5	8787.3
14.化工产品	122315.7	4454.1	16438.8
15.五金	141649.9	3163.9	9761.4
16.办公设备与电算机	175025.8	13210.7	19774.5
17.燃料	230614.5	1703.8	22626.6
18.汽车	293543	9528.2	18415.4

5.6 由表中给出的收入和住房支出样本数据, 建立住房支出模型。

住房支出	收入
1.8	5
2	5
2	5
2	5
2.1	5
3	10
3.2	10
3.5	10
3.5	10
3.6	10
4.2	15
4.2	15
4.5	15

4.8	15
5	15
4.8	20
5	20
5.7	20
6	20
6.2	20

假设模型为 $Y_i = \beta_1 + \beta_2 X_i + u_i$, 其中 Y 为住房支出, X 为收入。试求解下列问题:

(1) 用 OLS 求参数的估计值、标准差、拟合优度

(2) 用 Goldfeld-Quandt 方法检验异方差 (假设分组时不去掉任何样本值)

(3) 如果模型存在异方差, 假设异方差的形式是 $\sigma_i^2 = \sigma^2 X_i^2$, 试用加权最小二乘法重新

估计 β_1 和 β_2 的估计值、标准差、拟合优度。

5.7 表中给出 1969 年 20 个国家的股票价格 (Y) 和消费者价格年百分率变化 (X) 的一个横截面数据。

国家	股票价格变化率%Y	消费者价格变化率%X
1.澳大利亚	5	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
11.爱尔兰	6.4	4
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	4
19.英国	7.5	3.9
20.美国	9	2.1

试根据资料完成以下问题:

(1) 将 Y 对 X 回归并分析回归中的残差;

(2) 因智利的数据出现了异常, 去掉智利数据后, 重新作回归并再次分析回归中的残差;

(3) 如果根据第 1 条的结果你将得到有异方差性的结论, 而根据第 2 条的结论你又得到相反的结论, 对此你能得出什么样的结论?

5.8 表中给出的是 1998 年我国重要制造业销售收入与销售利润的数据资料

行业名称	销售收入	销售利润	行业名称	销售收入	销售利润
食品加工业	187.25	3180.44	医药制造业	238.71	1264.10
食品制造业	111.42	1119.88	化学纤维制造	81.57	779.46
饮料制造业	205.42	1489.89	橡胶制品业	77.84	692.08
烟草加工业	183.87	1328.59	塑料制品业	144.34	1345.00
纺织业	316.79	3862.90	非金属矿制品	339.26	2866.14
服装制造业	157.70	1779.10	黑色金属冶炼	367.47	3868.28
皮革羽绒制品	81.73	1081.77	有色金属冶炼	144.29	1535.16
木材加工业	35.67	443.74	金属制品业	201.42	1948.12
家具制造业	31.06	226.78	普通机械制造	354.69	2351.68
造纸及纸制品	134.40	1124.94	专用设备制造	238.16	1714.73
印刷业	90.12	499.83	交通运输设备	511.94	4011.53
文教体育用品	54.40	504.44	电子机械制造	409.83	3286.15
石油加工业	194.45	2363.80	电子通讯设备	508.15	4499.19
化学原料制品	502.61	4195.22	仪器仪表设备	72.46	663.68

试完成以下问题:

(1) 求销售利润对销售收入的样本回归函数, 并对模型进行经济意义检验和统计检验;

(2) 分别用图形法、Glejser 方法、White 方法检验模型是否存在异方差;

(3) 如果模型存在异方差, 选用适当的方法对异方差性进行修正。

5.9 下表所给资料为 1978 年至 2000 年四川省农村人均纯收入 X_t 和人均生活费支出 Y_t 的数据。

四川省农村人均纯收入和人均生活费支出			单位: 元/人		
时间	农村人均纯收入 X	农村人均生活费 支出Y	时间	农村人均纯收入 X	农村人均生活费 支出Y
1978	127.1	120.3	1990	557.76	509.16
1979	155.9	142.1	1991	590.21	552.39
1980	187.9	159.5	1992	634.31	569.46
1981	220.98	184.0	1993	698.27	647.43

1982	255.96	208.23	1994	946.33	904.28
1983	258.39	231.12	1995	1158.29	1092.91
1984	286.76	251.83	1996	1459.09	1358.03
1985	315.07	276.25	1997	1680.69	1440.48
1986	337.94	310.92	1998	1789.17	1440.77
1987	369.46	348.32	1999	1843.47	1426.06
1988	448.85	426.47	2000	1903.60	1485.34
1989	494.07	473.59			

数据来源：《四川统计年鉴》2001 年。

(1) 求农村人均生活费支出对人均纯收入的样本回归函数，并对模型进行经济意义检验和统计检验；

(2) 选用适当的方法检验模型中是否存在异方差；

(3) 如果模型存在异方差，选用适当的方法对异方差性进行修正。

5.10 在题 5.9 中用的是时间序列数据，而且没有剔除物价上涨因素。试分析如果剔除物价上涨因素，即用实际可支配收入和实际消费支出，异方差的问题是否会有所改善？由于缺乏四川省从 1978 年起的农村居民消费价格定基指数的数据，以 1978 年—2000 年全国商品零售价格定基指数（以 1978 年为 100）代替，数据如下表所示：

年份	商品零售价格指数	年份	商品零售消费价格指数	年份	商品零售消费价格指数
1978	100	1986	135.8	1994	310.2
1979	102	1987	145.7	1995	356.1
1980	108.1	1988	172.7	1996	377.8
1981	110.7	1989	203.4	1997	380.8
1982	112.8	1990	207.7	1998	370.9
1983	114.5	1991	213.7	1999	359.8
1984	117.7	1992	225.2	2000	354.4
1985	128.1	1993	254.9		

数据来源：《中国统计年鉴 2001》

第五章附录

附录 5.1 在异方差性条件下参数估计统计性质的证明

1、参数估计的无偏性仍然成立

设模型为
$$Y_i = \beta_1 + \beta_2 X_i + v_i, \quad i = 1, 2, \dots, n \quad (1)$$

用离差形式表示 $y_i = \beta_2 x_i + u_i$ (其中 $u_i = v_i - \bar{v}$) (2)

参数 β_2 的估计量 $\hat{\beta}_2$ 为

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (\beta_2 x_i + u_i)}{\sum x_i^2} = \frac{\beta_2 \sum x_i^2 + \sum x_i u_i}{\sum x_i^2} \\ &= \beta_2 + \frac{\sum x_i u_i}{\sum x_i^2}\end{aligned}\quad (3)$$

$$E(\hat{\beta}_2) = \beta_2 + E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right) = \beta_2 + \frac{\sum E(x_i u_i)}{\sum x_i^2} = \beta_2 \quad (4)$$

在证明中仅用到了假定 $E(x_i u_i) = 0$ 。

2、参数估计的有效性不成立

假设 (1) 式存在异方差，且 $\text{var}(u_i) = \sigma_i^2 = \sigma^2 X_i^2$ ，则参数 β_2 的估计 $\hat{\beta}_2$ 的方差为

$$\begin{aligned}\text{Var}(\hat{\beta}_2^*) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2 = E(\hat{\beta}_2 - \beta_2)^2 = E\left(\beta_2 + \frac{\sum x_i u_i}{\sum x_i^2} - \beta_2\right)^2 \\ &= E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right)^2 = E\left(\frac{\sum_{i=j} x_i^2 u_i^2 + 2 \sum_{i \neq j} x_i x_j u_i u_j}{(\sum x_i^2)^2}\right) = \frac{\sum_{i=j} x_i^2 E(u_i^2) + 2 \sum_{i \neq j} x_i x_j E(u_i u_j)}{(\sum x_i^2)^2} \\ &= \frac{\sum_{i=j} x_i^2 E(u_i^2)}{(\sum x_i^2)^2} = \frac{\sum_{i=j} x_i^2 \sigma_i^2}{(\sum x_i^2)^2} = \frac{\sigma^2 \sum x_i^2 X_i^2}{(\sum x_i^2)^2} = \frac{\sigma}{\sum x_i^2} \cdot \frac{\sum x_i^2 X_i^2}{\sum x_i^2}\end{aligned}\quad (5)$$

在上述推导中用了假定 $E(u_i u_j) = 0, i \neq j$ 。

下面对 (2) 式运用加权最小二乘法 (WLS)。设权数为 $w_i = \frac{1}{z_i}$ ，对 (2) 式变换为

$$\frac{y_i}{z_i} = \beta_2 \frac{x_i}{z_i} + \frac{u_i}{z_i} \quad (6)$$

可求得参数的估计 $\hat{\beta}_2$ ，根据本章第四节变量变换法的讨论，这时新的随机误差项 $\frac{u_i}{z_i}$ 为同方

差，即 $\text{var}\left(\frac{u_i}{z_i}\right) = \sigma^2$ ，而 $\hat{\beta}_2$ 的方差为

$$\text{var}(\hat{\beta}_2)_{wls} = \frac{\sigma^2}{\sum \left(\frac{x_i}{z_i} \right)^2} \quad (7)$$

为了便于区别, 用 $(\hat{\beta}_2)_{wls}$ 表示加权最小二乘法估计的 β_2 , 用 $(\hat{\beta}_2)_{ols}$ 表示 OLS 法估计的 β_2 。

比较 (5) 式与 (7) 式, 即在异方差下用 OLS 法得到参数估计的方差与用 WLS 法得到参数估计的方差相比较为

$$\frac{\text{var}(\hat{\beta}_2)_{wls}}{\text{var}(\hat{\beta}_2)_{ols}} = \frac{\frac{\sigma^2}{\sum \left(\frac{x_i}{z_i} \right)^2}}{\frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}} = \frac{\frac{\sigma^2}{\sum \left(\frac{x_i}{z_i} \right)^2}}{\frac{\sum x_i^2 \sigma^2 z_i^2}{(\sum x_i^2)^2}} = \frac{(\sum x_i^2)^2}{\sum \left(\frac{x_i}{z_i} \right)^2 (\sum x_i^2 z_i^2)} \quad (8)$$

令 $\frac{x_i}{z_i} = a_i, z_i x_i = b_i$, 由初等数学知识有 $\frac{(\sum ab)^2}{\sum a^2 \sum b^2} \leq 1$, 因此 (10) 式右端有

$$\frac{(\sum x_i^2)^2}{\sum \left(\frac{x_i}{z_i} \right)^2 (\sum x_i^2 z_i^2)} \leq 1 \quad (9)$$

从而, 有

$$\text{var}(\hat{\beta}_2)_{wls} \leq \text{var}(\hat{\beta}_2)_{ols}$$

这就证明了在异方差下, 仍然用普通最小二乘法所得到的参数估计值的方差不再最小。

附录 5.2 对数变换后残差为相对误差的证明

事实上, 设样本回归函数为

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad (10)$$

其中 $e_i = Y_i - \hat{Y}$ 为残差, 取对数后的样本回归函数为

$$\ln Y = \hat{\alpha}_1 + \hat{\alpha}_2 \ln X + e^* \quad (11)$$

其中残差为 $e^* = \ln Y - \ln \hat{Y}$, 因此

$$e^* = \ln Y - \ln \hat{Y} = \ln \left(\frac{Y}{\hat{Y}} \right) = \ln \left(\frac{\hat{Y} + Y - \hat{Y}}{\hat{Y}} \right) = \ln \left(1 + \frac{Y - \hat{Y}}{\hat{Y}} \right) \quad (12)$$

对 (12) 式的右端, 依据泰勒展式

$$\ln(1+X) = X - \frac{X^2}{2} + \frac{X^3}{3} - \frac{X^4}{4} + \cdots + (-1)^{n-1} \frac{X^n}{n} + \cdots \quad (13)$$

将 (13) 式中的 X 用 $\frac{Y-\hat{Y}}{\hat{Y}}$ 替换, 则 e^* 可近似地表示为

$$e^* \approx \frac{Y-\hat{Y}}{\hat{Y}} \quad (14)$$

即表明 (11) 式中的误差项为相对误差。

第六章 自相关

引子:

T 检验和 F 检验一定就可靠吗?

为了研究居民储蓄存款 Y 与居民收入 X 的关系, 设定模型为:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

取某地区 1985 年—2000 年居民储蓄存款 Y 和居民收入 X 的数据为样本数据, 用普通最小二乘法估计其参数, 结果为

$$\hat{Y}_i = 27.9123 + 0.3524 X_i$$

$$(1.8690) \quad (0.0055)$$

$$t = (14.9343) \quad (64.2069)$$

$$R^2 = 0.9966 \quad F = 4122.531$$

由估计检验结果可以看出, 回归系数的标准误差非常小, t 统计量较大, 说明居民收入 X 对居民储蓄存款 Y 的影响非常显著。同时可决系数也非常高, F 统计量=4122.531, 也表明模型异常的显著。

可是某些信息提示: 以上结论是不可靠的! 尽管所用样本数据都是真实的, 但这样的估计结果却可能是虚假的, 所计算的 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差及标准误差都严重地被低估了, t 统计量和 F 统计量都远远被虚假地夸大了, 因此所得结果是不可信的。

有什么充分的理由提出这样的质疑呢?

前面几章的讨论中, 我们假定随机误差项前后期之间是不相关的。但在经济系统中, 经济变量前后期之间很可能有关联, 使得随机误差项不能满足无自相关的假定。本章将探讨随机误差项不满足无自相关的古典假定时的参数估计问题。

第一节 什么是自相关

一、自相关的概念

自相关 (auto correlation) 又称序列相关 (serial correlation), 是指总体回归模型的随机误差项 u_i 之间存在相关关系。前面几章中强调, 在回归模型的古典假定中是假设随机误差

项是无自相关的，即 u_i 在不同观测点之间是不相关的，即

$$\text{Cov}(u_i, u_j) = E(u_i, u_j) = 0 \quad (i \neq j) \quad (\text{见 2.15})$$

如果该假定不能满足，就称 u_i 与 u_j 存在自相关，即不同观测点上的误差项彼此相关。

自相关的程度可用自相关系数去表示，随机误差项 u_t 与滞后一期的 u_{t-1} 的自相关系数为

$$\rho = \frac{\sum_{t=2}^n u_t u_{t-1}}{\sqrt{\sum_{t=2}^n u_t^2} \sqrt{\sum_{t=2}^n u_{t-1}^2}} \quad (6.1)$$

(6.1) 式定义的自相关系数 ρ 与普通相关系数的公式形式相同， ρ 的取值范围为 $-1 \leq \rho \leq 1$ 。(6.1) 式中 u_{t-1} 是 u_t 滞后一期的随机误差项，因此，将 (6.1) 式计算的自相关系数 ρ 称为一阶自相关系数。

根据自相关系数 ρ 的符号可以判断自相关的状态，如果 $\rho < 0$ ，则 u_t 与 u_{t-1} 为负相关；如果 $\rho > 0$ ，则 u_t 与 u_{t-1} 为正关；如果 $\rho = 0$ ，则 u_t 与 u_{t-1} 不相关。

二、自相关产生的原因

1、经济系统的惯性

自相关现象大多出现在时间序列数据中，而经济系统的经济行为都具有时间上的惯性。例如 GDP、价格、就业等经济数据，都会随经济系统的周期而波动。又如，在经济高涨时期，较高的经济增长率会持续一段时间，而在经济衰退期，较高的失业率也会持续一段时间，这种情况下经济数据很可能表现为自相关。

2、经济活动的滞后效应。

滞后效应是指某一变量对另一变量的影响不仅限于当期，而是延续若干期。由此带来变量的自相关。例如，居民当期可支配收入的增加，不会使居民的消费水平在当期就达到应有水平，而是要经过若干期才能达到。因为人的消费观念的改变存在一定的适应期。

3、数据处理造成的相关。

因为某些原因对数据进行了修正和内插处理，在这样的数据序列中可能产生自相关。例如，将月度数据调整为季度数据，由于采用了加合处理，修匀了月度数据的波动，使季度数据具有平滑性，这种平滑性可能产生自相关。对缺失的历史资料，采用特定统计方法进行内插处理，也可能使得数据前后期相关，而产生自相关。

4、蛛网现象。

蛛网现象是微观经济学中的一个概念。它表示某种商品的供给量 Y_t 受前一期价格 P_{t-1} 影响而表现出来的某种规律性，即呈蛛网状收敛或发散于供需的均衡点。许多农产品的供给呈现为蛛网现象，供给对价格的反应要滞后一段时间，因为供给的调整需要经过一定的时间才能实现。如果时期 t 的价格 P_t 低于上一期的价格 P_{t-1} ，农民就会减少时期 $t+1$ 的生产量。如此则形成蛛网现象，此时的供给模型为

$$Y_t = \beta_1 + \beta_2 P_{t-1} + u_t \quad (6.2)$$

这时式中的随机误差项 u_t 可能产生自相关。

5、模型设定偏误。

如果模型中省略了某些重要的解释变量或者模型函数形式不正确，都会产生系统误差，这种误差存在于随机误差项中，从而带来了自相关。由于设定误差造成的自相关，在经济计量分析中经常可能发生。例如，本来应该用两个解释变量去解释 Y ，即

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (6.3)$$

而建立模型时，模型设定为

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t \quad (6.4)$$

这样， X_{3t} 对 Y_t 的影响在 (6.4) 式中便归入到随机误差项 u_t 中，由于 X_{3t} 在不同观测点上是相关的，就造成了 u_t 是自相关的。

模型函数形式的设定误差也会导致自相关现象。如将本来应是 U 形成本曲线的模型设定为线性成本曲线的模型，则会导致自相关。由设定误差产生的自相关，可通过改变模型设定予以消除。

自相关关系主要存在于时间序列数据中，但是在横截面数据中也可能会出现自相关，通常称其为空间自相关 (Spatial auto correlation)。例如，一个家庭或一个地区的消费行为可能会影响另外一些家庭或另外一些地区，就是说不同观测点的随机误差项可能是相关的。多数经济时间序列在较长时间内都表现为上升或下降的超势，因此大多表现为正自相关。但就自相关本身而言，可以为正相关也可以为负相关。

三、自相关的表现形式

自相关的性质可以用自相关系数 ρ 的符号判断，即 $\rho < 0$ 为负相关， $\rho > 0$ 为正相关。当 $|\rho|$ 接近 1 时，表示相关的程度很高。自相关是 u_1, u_2, \dots, u_n 序列自身的相关，因 n 个随机误差项的关联形式不同而可能具有不同的自相关形式。自相关大多出现在时间序列数据

中，下面以时间序列为例说明自相关的不同表现形式。

对于样本观测期为 n 的时间序列数据，可得到总体回归模型（PRF）的随机误差项为 u_1, u_2, \dots, u_n ，如果自相关形式为

$$u_t = \rho u_{t-1} + v_t \quad (-1 < \rho < 1) \quad (6.5)$$

其中， ρ 为自相关系数， v_t 为满足古典假定的误差项，即 $E(v_t) = 0$ ， $Var(v_t) = \sigma^2$ ， $Cov(v_t, v_{t+s}) = 0$ ， $s \neq 0$ 。因为模型（6.5）中 u_{t-1} 是 u_t 滞后一期的值，则（6.5）式称为一阶自回归形式，记为 AR(1)。（6.5）式中的 ρ 也称为一阶自相关系数。

如果（6.5）式中的随机误差项 v_t 是不满足古典假定的误差项，即 v_t 中包含有 u_t 的成份，例如包含有 u_{t-2} 的影响，则需将 u_{t-2} 包含在回归模型中，即

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + v'_t \quad (6.6)$$

其中， ρ_1 为一阶自相关系数， ρ_2 为二阶自相关系数， v'_t 是满足古典假定的误差项。（6.6）式称为二阶自回归形式，记为 AR(2)。一般地，如果 u_1, u_2, \dots, u_t 之间的关系为

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_m u_{t-m} + v_t \quad (6.7)$$

其中， v_t 为满足古典假定的误差项。则称（6.7）式为 m 阶自回归形式，记为 AR(m)。

此外，自相关的形式可能为移动平均形式，记为 MA(n)，还可能为更复杂的移动平均自回归形式，记为 ARMA(m, n)，这些是时间序列分析的专题内容，本书不作讨论。

在经济计量分析中，通常采用（6.5）式的一阶自回归形式，即假定自回归形式为一阶自回归 AR(1)。这种假定简化了自回归形式，在实际分析中常能取得较好的效果，在本章中只讨论假定自相关为 AR(1) 的形式。

第二节 自相关的后果

当一个线性回归模型的随机误差项存在自相关时，就违背了线性回归方程的古典假定，如果仍然用普通最小二乘法（OLS）估计参数，将会产生严重后果。自相关产生的后果与异方差情形类似。

一、一阶自回归形式的性质

以一元线性回归模型为例，对于

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (6.8)$$

假定随机误差项 u 存在一阶自相关

$$u_t = \rho u_{t-1} + v_t \quad (6.9)$$

其中, u_t 为现期随机误差, u_{t-1} 为前期随机误差。 v_t 是满足古典假定的误差项, 即 v_t 满足零均值 $E(v_t) = 0$, 同方差 $Var(v_t) = \sigma_v^2$, 无自相关 $E(v_t v_s) = 0$ ($t \neq s$) 的假定。

在大样本情况下, ρ 的 OLS 估计式为

$$\hat{\rho} = \frac{\sum u_t u_{t-1}}{\sum u_{t-1}^2} \quad (6.10)$$

u_t 与 u_{t-1} 的相关系数为 (当样本较大时, $\sum u_t^2 \approx \sum u_{t-1}^2$)

$$\rho = \frac{\sum u_t u_{t-1}}{\sqrt{\sum u_t^2} \sqrt{\sum u_{t-1}^2}} \approx \frac{\sum u_t u_{t-1}}{\sum u_{t-1}^2} = \hat{\rho} \quad (6.11)$$

如果将随机误差项 u_t 的各期滞后值 $u_{t-1} = \rho u_{t-2} + v_{t-1}$, $u_{t-2} = \rho u_{t-3} + v_{t-2}$, \dots , 逐次代入

(6.9) 式可得

$$u_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \dots = \sum_{r=0}^{\infty} \rho^r v_{t-r} \quad (6.12)$$

(6.12) 式表明随机误差项 u_t 可表示为独立同分布的随机误差序列 $v_t, v_{t-1}, v_{t-2}, \dots$ 的加权和, 权数分别为 $1, \rho, \rho^2, \dots$, 当 $0 < \rho < 1$ 时, 这些权数随时间推移而呈几何衰减; 而当 $-1 < \rho < 0$ 时, 这些权数是随时间推移而交错振荡衰减的。

在随机误差项 u_t 存在一阶自回归形式的自相关时, 由 (6.12) 式可推得

$$E(u_t) = \sum_{r=0}^{\infty} \rho^r E(v_{t-r}) = 0 \quad (6.13)$$

$$Var(u_t) = \sum_{r=0}^{\infty} \rho^{2r} Var(v_{t-r}) = \frac{\sigma_v^2}{1 - \rho^2} = \sigma_u^2 \quad (6.14)$$

(6.13) 和 (6.14) 式表明, 在 u_t 为一阶自回归形式的自相关时, 随机误差项 u_t 依然满足零均值、同方差的假定。

由于现期的随机误差项 v_t 并不影响回归模型中随机误差项 u_t 的以前各期值 u_{t-k} ($k > 0$), 所以 v_t 与 u_{t-k} 不相关, 即有 $E(v_t u_{t-k}) = 0$ 。因此, 由 (6.9) 式可得随机误差项 u_t 与其以前各期 u_{t-k} 的协方差分别为

$$Cov(u_t, u_{t-1}) = E(u_t u_{t-1})$$

$$\begin{aligned}
&= E[(\rho u_{t-1} + v_t)u_{t-1}] \\
&= \rho E(u_{t-1}^2) + E(v_t u_{t-1}) \\
&= \rho \sigma_u^2 \\
&= \frac{\rho \sigma_v^2}{1 - \rho^2} \tag{6.15}
\end{aligned}$$

$$\begin{aligned}
Cov(u_t, u_{t-2}) &= E(u_t u_{t-2}) \\
&= E[(\rho u_{t-1} + v_t)u_{t-2}] \\
&= E[(\rho^2 u_{t-2} + \rho v_{t-1} + v_t)u_{t-2}] \\
&= \rho^2 E(u_{t-2}^2) \\
&= \frac{\rho^2 \sigma_v^2}{1 - \rho^2} \tag{6.16}
\end{aligned}$$

以此类推，可得

$$Cov(u_t, u_{t-k}) = \rho^k Var(u_{t-k}) = \frac{\rho^k \sigma_v^2}{1 - \rho^2} \tag{6.17}$$

这些协方差分别称为随机误差项 u_t 的一阶自协方差、二阶自协方差和 k 阶自协方差，这些自协方差均不为零，这正是存在自相关的含义。

二、自相关对参数估计的影响

以一元线性回归模型（6.8）为例，当 u_t 满足各项古典假定时，普通最小二乘估计 $\hat{\beta}_2$ 的方差为

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2} \quad (\text{见 2.40})$$

首先，当随机误差项 u_t 存在自相关时， $\hat{\beta}_2$ 依然是无偏的，即 $E(\hat{\beta}_2) = \beta_2$ ，因为在普通最小二乘法无偏性的证明中并不需要 u_t 满足无自相关的假定。

再看 OLS 估计式的方差，（6.8）式中 $\hat{\beta}_2$ 的最小二乘估计式为

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} = \beta_2 + \frac{\sum x_t u_t}{\sum x_t^2} \tag{6.18}$$

可以证明(见本章附录 6.1)

$$Var(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum_{t=1}^n x_t^2} \left(1 + 2\rho \frac{\sum_{t=1}^{n-1} x_t x_{t+1}}{\sum_{t=1}^n x_t^2} + 2\rho^2 \frac{\sum_{t=1}^{n-2} x_t x_{t+2}}{\sum_{t=1}^n x_t^2} + \cdots + 2\rho^{n-1} \frac{x_1 x_n}{\sum_{t=1}^n x_t^2} \right) \quad (6.19)$$

当随机误差项无自相关时， $\rho = 0$ ，此时（6.19）式等价于（2.40）式。当存在正的自相关时，（6.19）式的方差将大于（2.40）式的方差。可以证明，当存在自相关时，普通最小二乘估计量不再是最佳线性无偏估计量，即它在线性无偏估计量中不是方差最小的。

由式(6.19)可以看出，当 $\rho > 0$ ，即有正相关时，对所有的 j 都有 $\rho^j > 0$ 。另外回归模型中的解释变量在不同时期通常也呈正相关，即对于 x_t 和 x_{t+j} 来说 $\sum x_t x_{t+j}$ 是大于 0 的。因此（6.19）式右边括号内的值通常大于 0，如果仍用 OLS 法 $Var(\hat{\beta}_2) = \sigma^2 / \sum x_t^2$ 去计算 $\hat{\beta}_2$ 的方差，将会低估存在自相关时参数估计值的真实方差。

此外，当随机误差项 u_t 不存在自相关时，已知 $\hat{\sigma}^2 = \sum e_t^2 / (n-2)$ 是 u_t 的方差 σ^2 的无偏估计，即 $E(\hat{\sigma}^2) = E[\sum e_t^2 / (n-2)] = \sigma^2$ 。但是，如果随机误差项 u_t 存在一阶自相关，可以证明

$$E(\sum e_t^2) = \sigma^2[(n-2) - (2\rho \frac{\sum X_t X_{t+1}}{\sum X_t^2} + 2\rho^2 \frac{\sum X_t X_{t+2}}{\sum X_t^2} + \cdots + 2\rho^{n-1} \frac{\sum X_t X_n}{\sum X_t^2})] \quad (6.20)$$

当 u_t 及 X_t 都是正自相关时，(6.20)式中圆括号内的值为正值，将使 $\sum e_t^2$ 的值降低。这说明如果仍用 $\hat{\sigma}^2 = \sum e_t^2 / (n-2)$ 去估计 u_t 的方差 σ^2 ，则会导致低估真实的 σ^2 。显然，这将使得参数估计值的方差被进一步低估。

三、自相关对模型检验的影响

通过前面的讨论已知，当存在自相关时，如果忽视自相关问题，依然用满足古典假定的 OLS 法去估计参数及其方差，会低估真实的 σ^2 ，更会低估参数估计值的方差。由于对参数显著性检验的 t 统计量为 $t = (\hat{\beta}_2 - \beta_2) / SE(\hat{\beta}_2) \sim t(n-2)$ ，当参数估计值的方差被低估时，其标准误差 $SE(\hat{\beta}_2)$ 也将被低估，从而过高估计 t 统计量的值，会夸大所估计参数的显著性，对本来不重要的解释变量可能误认为重要而被保留。这时通常的回归系统显著性的 t 检验将

失去意义。

类似地，由于自相关的存在，参数的最小二乘估计量是无效的，使得 F 检验和 R^2 检验也是不可靠的。

四、自相关对模型预测的影响

模型预测的精度决定于抽样误差和总体误差项的方差 σ^2 。抽样误差来自于对 $\hat{\beta}_j$ 的估计，在自相关情形下， $\hat{\beta}_j$ 的方差的最小二乘估计变得不可靠，由此必定加大抽样误差。同时，在自相关情形下，对 σ^2 的估计 $\hat{\sigma}^2 = \sum e_i^2 / (n - k)$ 也会不可靠。由此可看出，影响预测精度的两大因素都因自相关的存在而加大不确定性，使预测的置信区间不可靠，从而降低了预测的精度。

第三节 自相关的检验

随机误差项存在自相关给普通最小二乘法的应用带来的后果是严重的。因此，必须设法诊断是否存在自相关。检测回归模型是否存在自相关的常用方法有以下几种。

一、图示检验法

图示法是一种直观的诊断方法，它是对给定的回归模型直接用普通最小二乘法估计其参数，求出残差项 e_t ，以 e_t 作为随机项 u_t 的估计值，再描绘 e_t 的散点图，根据散点图来判断 e_t 的相关性。残差 e_t 的散点图通常有两种绘制方式。

1. 绘制 e_{t-1} 和 e_t 的散点图。用 (e_{t-1}, e_t) ($t = 1, 2, \dots, n$) 作为散布点绘图，如果大部分点落在第 I、III 象限，表明随机误差项 u_t 存在着正自相关，如图 6.1 所示。如果大部分点落在第 II、IV 象限，那么随机误差项 u_t 存在着负自相关，如图 6.2 所示。

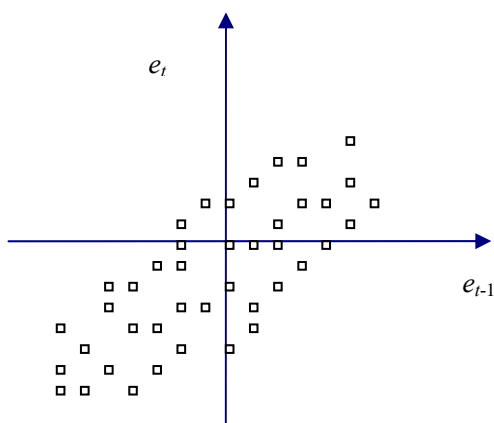


图 6.1 e_t 与 e_{t-1} 的关系

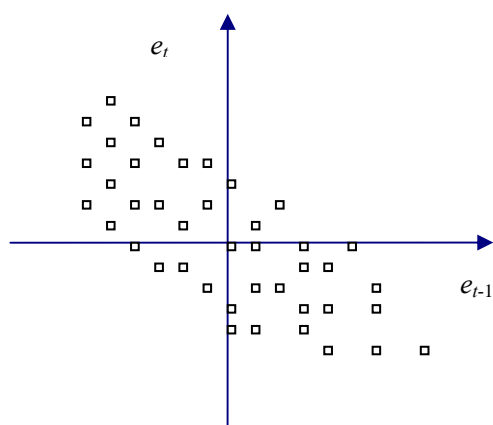


图 6.2 e_t 与 e_{t-1} 的关系

2. 按照时间顺序绘制回归残差项 e_t 的图形。如果 e_t ($t=1,2,\dots,n$) 随着 t 的变化逐次有规律地变化, 呈现锯齿形或循环形状的变化, 就可判断 e_t 存在相关, 表明 u_t 存在着自相关; 如果 e_t 随着 t 的变化逐次变化并不断地改变符号, 那么随机误差项 u_t 存在负自相关; 如图 6.3 所示。如果 e_t 随着 t 的变化逐次变化并不频繁地改变符号, 而是几个正的 e_t 后面跟着几个负的, 则表明随机误差项 u_t 存在正自相关, 如图 6.4 所示。

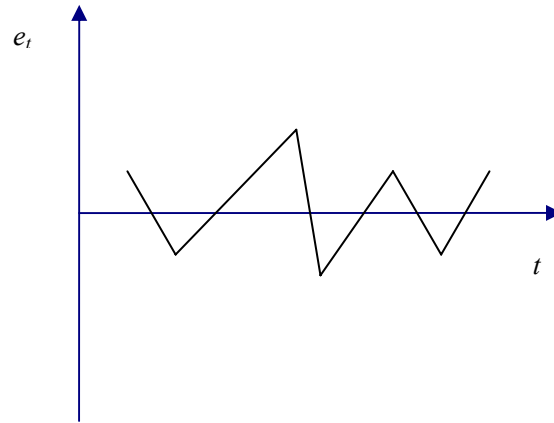


图 6.3 e_t 的分布

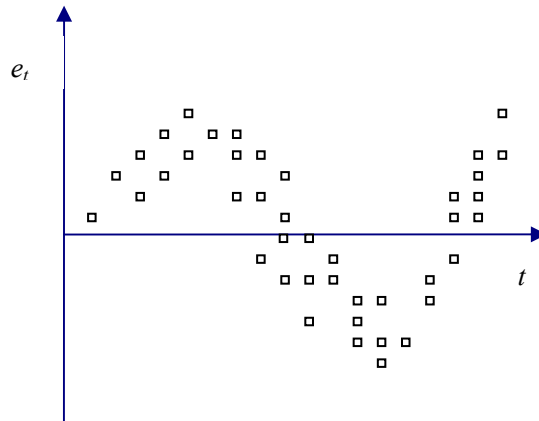


图 6.4 e_t 的分布

二、DW 检验法

DW 检验是 J.Durbin(杜宾)和 G.S.Watson(沃特森)于 1951 年提出的一种适用于小样本的检验方法。 DW 检验方法是检验自相关的常用方法，许多计量经济学和统计软件都提供 DW 值。

DW 检验法的前提条件是：

- (1) 解释变量 X 为非随机的；
- (2) 随机误差项为一阶自回归形式，即

$$u_t = \rho u_{t-1} + v_t \quad (v_t \text{ 满足古典假定}) \quad (6.21)$$

- (3) 线性模型的解释变量中不包含滞后的被解释变量，例如不应出现下列形式：

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$$

(4) 截距项不为零，即只适用于有常数项的回归模型；

(5) 数据序列无缺失项。

为了检验序列的相关性，构造的原假设是 $H_0: \rho = 0$ 。为了检验这一假设，构造 DW 统计量，首先要计算回归估计式的残差 e_t ，定义 DW 统计量为

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (6.22)$$

其中， $e_t = Y_t - \hat{Y}_t$ ， $t = 1, 2, \dots, n$ 。

由 (6.22) 式可得

$$DW = \frac{\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \quad (6.23)$$

如果认为 $\sum_{t=2}^n e_t^2 \approx \sum_{t=2}^n e_{t-1}^2 \approx \sum_{t=1}^n e_t^2$ ，则由 (6.23) 式得

$$DW \approx 2 \left[1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \right] \quad (6.24)$$

同理，在认为 $\sum_{t=2}^n e_t^2 \approx \sum_{t=2}^n e_{t-1}^2 \approx \sum_{t=1}^n e_t^2$ 时

$$\hat{\rho} \approx \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \quad (6.25)$$

$$\text{因此，} \quad DW \approx 2(1 - \hat{\rho}) \quad (6.26)$$

所以， DW 值与 $\hat{\rho}$ 的对应关系如表 6.1 所示。

表 6.1 DW 值与 $\hat{\rho}$ 的值的对应关系

$\hat{\rho}$	DW
-1	4
(-1,0)	(2,4)
0	2
(0,1)	(0,2)
1	0

由上述讨论可知 DW 的取值范围为 $0 \leq DW \leq 4$ 。

根据样本容量 n 和解释变量的数目 k' (不包括常数项), 查 DW 分布表, 可得临界值 d_L 和 d_U , 然后依下列准则考察计算的 DW 值, 以决定模型的自相关状态 (见表 6.2)。

表 6.2 DW 检验决策规则

$0 \leq DW \leq d_L$	误差项 u_1, u_2, \dots, u_n 间存在正相关
$d_L < DW \leq d_U$	不能判定是否有自相关
$d_U < DW < 4 - d_U$	误差项 u_1, u_2, \dots, u_n 间无自相关
$4 - d_U \leq DW < 4 - d_L$	不能判定是否有自相关
$4 - d_L \leq DW \leq 4$	误差项 u_1, u_2, \dots, u_n 间存在负相关

表 6.2 可以用坐标图更加直观地表示出来, 如图 6.5 所示。

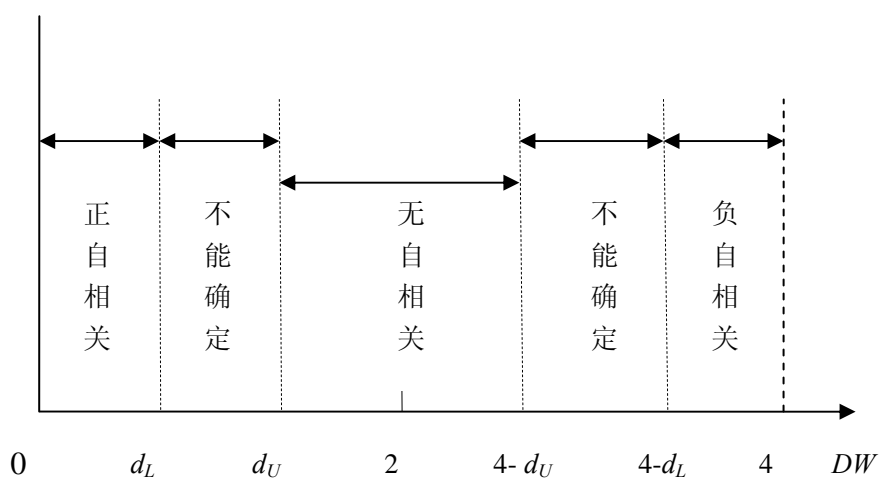


图 6.5 DW 检验示意图

需要注意的是， DW 检验尽管有着广泛的应用，但也有明显的缺点和局限性。

(1) DW 检验有运用的前提条件，只有符合这些条件 DW 检验才是有效的。

(2) DW 统计量的上、下界表一般要求 $n \geq 15$ ，这是因为样本如果再小，当 $n < 15$ 时， DW 检验上下界表的数据不完整，利用残差很难对自相关的存在性做出比较正确的诊断；

(3) DW 检验不适应随机误差项具有高阶序列相关的检验；

(4) DW 检验有两个不能确定的区域，一旦 DW 值落在这两个区域，就无法判断。这时，只有增大样本容量或选取其他方法。或者采用修正的 DW 检验法进行检验，即扩大拒绝区域，当 $d_U < d < 4 - d_U$ 时，不拒绝 $H_0: \rho = 0$ ，认为不存在自相关；而当 $d < d_U$ 或 $d > 4 - d_U$ 时，就拒绝 $H_0: \rho = 0$ ，认为存在自相关。这种修正实际是将拒绝区域扩大到不确定区域，由于自相关后果严重，在自相关不确定时，宁可拒绝 $H_0: \rho = 0$ ，而不宜轻易接受无自相关。

第四节 自相关的补救

如果自相关是由于模型设定误差造成的，可以通过改变模型的设定去消除。对于设定正确的模型，如果随机误差项有自相关，则可采用如下方法予以消除。

一、广义差分法

由于随机误差项 u_t 是不可观测的，通常我们假定 u_t 为一阶自回归形式，即

$$u_t = \rho u_{t-1} + v_t \quad (\text{见 6.21})$$

其中， $|\rho| < 1$ ， v_t 为满足古典假定的误差项。

当自相关系数 ρ 为已知时，可使用广义差分法解决自相关问题。以一元线性回归模型为例

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (6.27)$$

将模型 (6.27) 滞后一期可得

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \quad (6.28)$$

用 ρ 乘 (6.28) 式两边，得

$$\rho Y_{t-1} = \rho \beta_1 + \rho \beta_2 X_{t-1} + \rho u_{t-1} \quad (6.29)$$

用 (6.27) 式减去 (6.29) 式得

$$Y_t - \rho Y_{t-1} = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + u_t - \rho u_{t-1} \quad (6.30)$$

由 (6.21) 式, (6.30) 式中的 $u_t - \rho u_{t-1} = v_t$ 是满足古典假定的误差项。因此, 模型 (6.30) 满足古典假定, 随机误差项 $u_t - \rho u_{t-1} = v_t$ 无自相关。

令 $Y_t^* = Y_t - \rho Y_{t-1}$, $X_t^* = X_t - \rho X_{t-1}$, $\beta_1^* = \beta_1(1 - \rho)$, $\beta_2 = \beta_2^*$, 则式 (6.30) 可表示为

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + v_t \quad (6.31)$$

对模型 (6.31) 使用普通最小二乘估计, 可得到参数的最佳线性无偏估计量。因为 (6.30) 式中被解释变量与解释变量均为现期值减去前期值的一部分, 所以称为广义差分方程。在进行广义差分时, 解释变量 X 与被解释变量 Y 均以差分形式出现, 因而样本容量由 n 减少为 $n - 1$, 即丢失了第一个观测值。如果样本容量较大, 减少一个观测值对估计结果影响不大。但是, 如果样本容量较小, 则会对估计精度产生较大影响。此时, 可采用普莱斯—温斯滕 (Prais-Winsten) 变换, 将第一个观测值分别变换为 $Y_1\sqrt{1 - \rho^2}$ 和 $X_1\sqrt{1 - \rho^2}$, 补充到差分序列 Y_t^*, X_t^* 中, 再使用普通最小二乘法估计参数。

二、科克伦—奥克特 (Cochrane—Orcutt) 迭代法

在实际应用中, 自相关系数 ρ 往往是未知的, 必须通过一定的方法去估计 ρ 。最简单的方法是依据 DW 统计量去估计 ρ 。由 (6.26) 式 DW 与 ρ 的关系可知

$$\hat{\rho} \approx 1 - \frac{DW}{2} \quad (6.32)$$

但是, (6.31) 式得到的只是一个粗略的结果, 这样得到的 $\hat{\rho}$ 只是对 ρ 精度不高的估计, 根本原因在于对有自相关的回归模型使用了普通最小二乘法。为了得到 ρ 的更精确的估计值, 可采用科克伦—奥克特 (Cochrane—Orcutt) 迭代法。

科克伦—奥克特 (Cochrane—Orcutt) 迭代法的基本思想, 是通过逐次迭代去寻求更为满意的 ρ 的估计值, 然后再采用广义差分法。具体来说, 该方法是利用残差 e_t 去估计未知的 ρ 。

对于一元线性回归模型 $Y_t = \beta_1 + \beta_2 X_t + u_t$, 假定 u_t 为一阶自回归形式, 即

$$u_t = \rho u_{t-1} + v_t \quad (6.33)$$

科克伦—奥克特迭代法估计 ρ 的步骤如下：

第一步，使用 OLS 法估计模型 $Y_t = \beta_1 + \beta_2 X_t + u_t$ ，并计算残差 $e_t^{(1)}$

$$e_{tx}^{(1)} = Y_t - \hat{Y}_t = Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_t)$$

第二步，利用残差 $e_t^{(1)}$ 作如下的回归

$$e_t^{(1)} = \hat{\rho}^{(1)} e_{t-1}^{(1)} + v_t \quad (6.34)$$

第三步，用 OLS 法估计 (6.34) 式中的 $\hat{\rho}^{(1)}$ ，对模型 (6.32) 进行广义差分，即

$$Y_t - \hat{\rho}^{(1)} Y_{t-1} = \beta_1 (1 - \hat{\rho}^{(1)}) + \beta_2 (X_t - \hat{\rho}^{(1)} X_{t-1}) + u_t - \hat{\rho}^{(1)} u_{t-1} \quad (6.35)$$

令 $Y_t^* = Y_t - \hat{\rho}^{(1)} Y_{t-1}$ ， $X_t^* = X_t - \hat{\rho}^{(1)} X_{t-1}$ ， $\beta_1^* = \beta_1 (1 - \hat{\rho}^{(1)})$ ，对式 (6.35) 使用 OLS 法，可得样本回归函数为

$$\hat{Y}_t^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_t^* + e_t^{(2)} \quad (6.36)$$

第四步，由前一步估计的结果有 $\hat{\beta}_1 = \hat{\beta}_1^* / (1 - \hat{\rho}^{(1)})$ 和 $\hat{\beta}_2 = \hat{\beta}_2^*$ ，将 $\hat{\beta}_1, \hat{\beta}_2$ 代入原回归方程 (6.32)，求得新的残差 $e_t^{(3)}$

$$e_t^{(3)} = Y_t - \beta_1 - \beta_2 X_t \quad (6.37)$$

第五步，利用残差 $e_t^{(3)}$ 作回归

$$e_t^{(3)} = \rho^{(2)} e_{t-1}^{(3)} + v_t \quad (6.38)$$

用 OLS 法估计的 $\hat{\rho}^{(2)}$ 是对 ρ 的第二轮估计值。

当不能确认 $\hat{\rho}^{(2)}$ 是否是 ρ 的最佳估计值时，继续迭代估计 ρ 的第三轮估计值 $\hat{\rho}^{(3)}$ 。直到估计的 $\hat{\rho}^{(k)}$ 与 $\hat{\rho}^{(k-1)}$ 相差很小时，收敛并满足精度要求，或回归所得 DW 统计量说明已不存在自相关时为止。通常，经过迭代很快就能得到有较高精度的 $\hat{\rho}$ ，用作广义差分对自相关的修正效果也较好。

三、其它方法简介

(一) 一阶差分法

一阶差分法是模型存在完全一阶正自相关时消除自相关的一种简单有效方法。仍以一元线性回归模型为例

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (6.39)$$

其中 u_t 为一阶自回归 AR(1) $u_t = \rho u_{t-1} + v_t$ 。如果原模型存在完全一阶正自相关，即 $\rho=1$ ，由 (6.30) 式，可将模型 (6.39) 变换为

$$\Delta Y_t = \beta_2 \Delta X_t + u_t - u_{t-1} \quad (6.40)$$

其中， $\Delta Y_t = Y_t - Y_{t-1}$, $\Delta X_t = X_t - X_{t-1}$ 。这时的随机误差为

$$u_t = u_{t-1} + v_t \quad (6.41)$$

其中， v_t 为满足古典假定的误差项，则 (6.40) 式的随机误差项 $u_t - u_{t-1} = v_t$ 为满足古典假定的误差项，无自相关问题。对 (6.40) 式使用普通最小二乘法估计参数，可得到最佳线性无偏估计量。虽然实际经济问题中完全一阶正自相关并不多见。但是只要 ρ 是正的且比较大，一阶差分法往往是有效的。但应注意，一阶差分法得到的回归方程 (6.40) 中没有常数项，如回归分析要求有常数项，该方法就不一定适用。

(二) 德宾两步法

当自相关系数 ρ 未知时，也可采用德宾提出的两步法去消除自相关。

将广义差分方程 (6.30) 表示为

$$Y_t = \beta_1(1 - \rho) + \beta_2 X_t - \rho \beta_2 X_{t-1} + \rho Y_{t-1} + v_t \quad (6.42)$$

采用如下的两个步骤消除自相关。

第一步，将 (6.42) 式作为一个多元回归模型，使用普通最小二乘法估计其参数。把 Y_{t-1} 的回归系数 $\hat{\rho}$ 看作 ρ 的一个估计值，它是 ρ 的一个有偏、一致估计。

第二步，利用估计的 $\hat{\rho}$ 进行广义差分。求得序列 $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$ 和 $X_t^* = X_t - \hat{\rho} X_{t-1}$ ，然后使用 OLS 法对广义差分方程估计参数，求得最佳线性无偏估计量。

第五节 案例分析

一、研究目的

2003 年中国农村人口占 59.47%，而消费总量却只占 41.4%，农村居民的收入和消费是一个值得研究的问题。消费模型是研究居民消费行为的常用工具。通过中国农村居民消费模型的分析可判断农村居民的边际消费倾向，这是宏观经济分析的重要参数。同时，农村居民

消费模型也能用于农村居民消费水平的预测。

二、模型设定

正如第二章所讲述的，影响居民消费的因素很多，但由于受各种条件的限制，通常只引入居民收入一个变量做解释变量，即消费模型设定为

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad (6.43)$$

式中， Y_t 为农村居民人均消费支出， X_t 为农村人均居民纯收入， u_t 为随机误差项。表 6.3 是从《中国统计年鉴》收集的中国农村居民 1985-2003 年的收入与消费数据。

表 6.3 1985-2003 年农村居民人均收入和消费 单位：元

年份	全年人均纯收入 (现价)	全年人均消费性支出 (现价)	消费价格指数 (1985=100)	人均实际纯收入 (1985 可比价)	人均实际消费性支出 (1985 可比价)
1985	397.60	317.42	100.0	397.60	317.40
1986	423.80	357.00	106.1	399.43	336.48
1987	462.60	398.30	112.7	410.47	353.42
1988	544.90	476.70	132.4	411.56	360.05
1989	601.50	535.40	157.9	380.94	339.08
1990	686.30	584.63	165.1	415.69	354.11
1991	708.60	619.80	168.9	419.54	366.96
1992	784.00	659.80	176.8	443.44	373.19
1993	921.60	769.70	201.0	458.51	382.94
1994	1221.00	1016.81	248.0	492.34	410.00
1995	1577.70	1310.36	291.4	541.42	449.69
1996	1923.10	1572.10	314.4	611.67	500.03
1997	2090.10	1617.15	322.3	648.50	501.77
1998	2162.00	1590.33	319.1	677.53	498.28
1999	2214.30	1577.42	314.3	704.52	501.75
2000	2253.40	1670.00	314.0	717.64	531.85
2001	2366.40	1741.00	316.5	747.68	550.08
2002	2475.60	1834.00	315.2	785.41	581.85
2003	2622.24	1943.30	320.2	818.86	606.81

注：资料来源于《中国统计年鉴》1986-2004。

为了消除价格变动因素对农村居民收入和消费支出的影响，不宜直接采用现价人均纯收入和现价人均消费支出的数据，而需要用经消费价格指数进行调整后的 1985 年可比价格计的人均纯收入和人均消费支出的数据作回归分析。

根据表 6.3 中调整后的 1985 年可比价格计的人均纯收入和人均消费支出的数据，使用普通最小二乘法估计消费模型得

$$\hat{Y}_t = 106.7528 + 0.5998X_t \quad (6.44)$$

$$Se = (12.2238) \quad (0.0214)$$

$$t = (8.7332) \quad (28.3067)$$

$$R^2 = 0.9788, F = 786.0548, df = 17, DW = 0.7706$$

该回归方程可决系数较高，回归系数均显著。对样本量为 19、一个解释变量的模型、5%显著水平，查 DW 统计表可知， $d_L = 1.18$ ， $d_U = 1.40$ ，模型中 $DW < d_L$ ，显然消费模型中有自相关。这一点残差图中也可从看出，点击 EViews 方程输出窗口的按钮 Resids 可得到残差图，如图 6.6 所示。

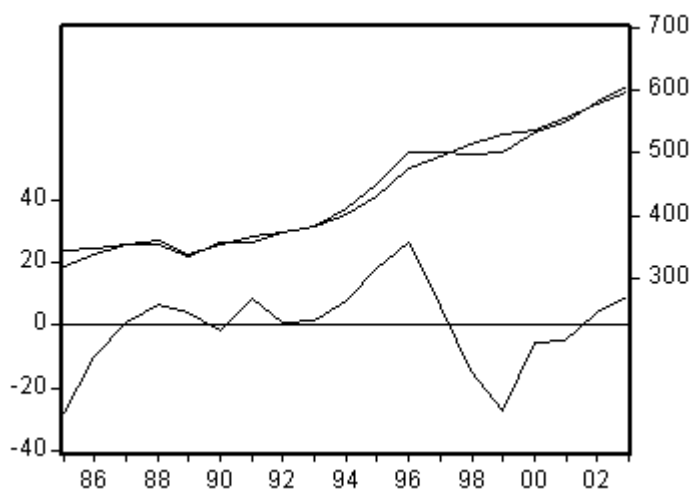


图 6.6 残差图

图 6.6 残差图中，残差的变动有系统模式，连续为正和连续为负，表明残差项存在一阶正自相关，模型中 t 统计量和 F 统计量的结论不可信，需采取补救措施。

三、自相关问题的处理

为解决自相关问题，选用科克伦—奥克特迭代法。由模型 (6.44) 可得残差序列 e_t ，在 EViews 中，每次回归的残差存放在 resid 序列中，为了对残差进行回归分析，需生成命名为 e 的残差序列。在主菜单选择 Quick/Generate Series 或点击工作文件窗口工具栏中的 Procs/

Generate Series, 在弹出的对话框中输入 $e = \text{resid}$, 点击 OK 得到残差序列 e_t 。使用 e_t 进行滞后一期的自回归, 在 EViews 命令栏中输入 $\text{ls } e \ e(-1)$ 可得回归方程

$$e_t = 0.4960 \ e_{t-1} \quad (6.45)$$

由式 (6.45) 可知 $\hat{\rho} = 0.4960$, 对原模型进行广义差分, 得到广义差分方程

$$Y_t - 0.4960Y_{t-1} = \beta_1(1 - 0.4960) + \beta_2(X_t - 0.4960X_{t-1}) + u_t \quad (6.46)$$

对式 (6.46) 的广义差分方程进行回归, 在 EViews 命令栏中输入 $\text{ls } Y-0.4960*Y(-1) \ c \ X-0.4960*X(-1)$, 回车后可得方程输出结果如表 6.4。

表6.4 广义差分方程输出结果

Dependent Variable: Y-0.496014*Y(-1)				
Method: Least Squares				
Date: 03/26/05 Time: 12:32				
Sample(adjusted): 1986 2003				
Included observations: 18 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	60.44431	8.964957	6.742287	0.0000
X-0.496014*X(-1)	0.583287	0.029410	19.83325	0.0000
R-squared	0.960914	Mean dependent var	231.9218	
Adjusted R-squared	0.958472	S.D. dependent var	49.34525	
S.E. of regression	10.05584	Akaike info criterion	7.558623	
Sum squared resid	1617.919	Schwarz criterion	7.657554	
Log likelihood	-66.02761	F-statistic	393.3577	
Durbin-Watson stat	1.397928	Prob(F-statistic)	0.000000	

由表 6.4 可得回归方程为

$$\hat{Y}_t^* = 60.4443 + 0.5833X_t^* \quad (6.47)$$

$$Se = (8.9650) \quad (0.0294)$$

$$t = (6.7423) \quad (19.8333)$$

$$R^2 = 0.9609 \quad F = 393.3577 \quad df = 16 \quad DW = 1.3979$$

式中, $\hat{Y}_t^* = Y_t - 0.4960Y_{t-1}$, $X_t^* = X_t - 0.4960X_{t-1}$ 。

由于使用了广义差分数据, 样本容量减少了 1 个, 为 18 个。查 5% 显著水平的 DW 统计表可知 $d_L = 1.16$, $d_U = 1.39$, 模型中 $DW = 1.3979 > d_U$, 说明广义差分模型中已无自相关, 不必再进行迭代。同时可见, 可决系数 R^2 、 t 、 F 统计量也均达到理想水平。

对比模型 (6.44) 和 (6.47), 很明显普通最小二乘法低估了回归系数 $\hat{\beta}_2$ 的标准误差。[原模型中 $Se(\hat{\beta}_2) = 0.0214$, 广义差分模型中为 $Se(\hat{\beta}_2) = 0.0294$ 。

经广义差分后样本容量会减少 1 个, 为了保证样本数不减少, 可以使用普莱斯—温斯腾变换补充第一个观测值, 方法是 $X_1^* = X_1\sqrt{1-\rho^2}$ 和 $Y_1^* = Y_1\sqrt{1-\rho^2}$ 。在本例中即为 $X_1\sqrt{1-0.4960^2}$ 和 $Y_1\sqrt{1-0.4960^2}$ 。由于要补充因差分而损失的第一个观测值, 所以在 EViews 中就不能采用前述方法直接在命令栏输入 Y 和 X 的广义差分函数表达式, 而是要生成 X 和 Y 的差分序列 X^* 和 Y^* 。在主菜单选择 Quick/Generate Series 或点击工作文件窗口工具栏中的 Procs/Generate Series, 在弹出的对话框中输入 $Y^* = Y - 0.4960 * Y(-1)$, 点击 OK 得到广义差分序列 Y^* , 同样的方法得到广义差分序列 X^* 。此时的 X^* 和 Y^* 都缺少第一个观测值, 需计算后补充进去, 计算得 $X_1^* = 345.236$, $Y_1^* = 275.598$, 双击工作文件窗口的 X^* 打开序列显示窗口, 点击 Edit+/- 按钮, 将 $X_1^* = 345.236$ 补充到 1985 年对应的栏目中, 得到 X^* 的 19 个观测值的序列。同样的方法可得到 Y^* 的 19 个观测值序列。在命令栏中输入 $Ls Y^* c X^*$ 得到普莱斯—温斯腾变换的广义差分模型为

$$Y_t^* = 60.4443 + 0.5833X_t^* \quad (6.48)$$

$$Se = (9.1298) \quad (0.0297)$$

$$t = (6.5178) \quad (19.8079)$$

$$R^2 = 0.9585 \quad F = 392.3519 \quad df = 19 \quad DW = 1.3459$$

对比模型 (6.47) 和 (6.48) 可发现, 两者的参数估计值和各检验统计量的差别很微小, 说明在本例中使用普莱斯—温斯腾变换与直接使用科克伦—奥克特两步法的估计结果无显著差异, 这是因为本例中的样本还不算太小。如果实际应用中样本较小, 则两者的差异会较大。通常对于小样本, 应采用普莱斯—温斯腾变换补充第一个观测值。

由差分方程 (6.46) 有

$$\hat{\beta}_1 = \frac{60.4443}{1 - 0.4960} = 119.9292 \quad (6.49)$$

由此，我们得到最终的中国农村居民消费模型为

$$Y_t = 119.9292 + 0.5833 X_t \quad (6.50)$$

由 (6.50) 的中国农村居民消费模型可知，中国农村居民的边际消费倾向为 0.5833，即中国农民每增加收入 1 元，将增加消费支出 0.5833 元。

第六章小结

- 1、当总体回归模型的随机误差项在不同观测点上彼此相关时就产生了自相关问题。
- 2、时间序列的惯性、经济活动的滞后效应、模型设定错误、数据的处理等多种原因都可能导致出现自相关。

- 3、在出现自相关时，普通最小二乘估计量依然是无偏、一致的，但不再是有效的。

如

果仍用 OLS 法计算参数估计值的方差，将会低估存在自相关时参数估计值的真实方差。而且会因低估真实的 σ^2 ，导致参数估计值的方差被进一步低估。由于真实 σ^2 的低估和参数估计值方差的低估，通常的 t 检验和 F 检验都不能有效地使用，也使预测的置信区间不可靠，降低了预测的精度。

- 4、随机误差项的自相关形式决定于其关联形式，可以为 m 阶自回归形式 $m = 1, 2, \dots, m$ ，即 $AR(m)$ 。为了研究问题的方便和考虑实际问题的代表意义，通常将自相关设定为一阶自相关即 $AR(1)$ 模式。用一阶自相关系数 ρ 表示自相关的程度与方向。

- 5、由于 u_t 不可观测，通常使用 u_t 的估计量 e_t 判断 u_t 的特性。绘制 e_{t-1} ， e_t 的散点图或按照时间顺序绘制回归残差项 e_t 的图形，可以判断自相关的存在。判断自相关的存在最常用的方法是依据 e_t 计算的 DW 统计量，但要注意 DW 检验法的前提条件和局限性。

- 6、如果自相关系数 ρ 是已知的，我们可以使用广义差分法消除序列相关。

- 7、如果自相关系数 ρ 是未知的，我们可采用科克伦—奥克特迭代法或德宾两步法求得 ρ 的估计值，然后用广义差分法消除序列相关。

主要公式表

1、自相关系数	$\rho = \frac{\sum_{t=2}^n u_t u_{t-1}}{\left(\sqrt{\sum_{t=2}^n u_t^2} \sqrt{\sum_{t=2}^n u_{t-1}^2} \right)}$
2、一阶自回归 形式 AR(1)	$u_t = \rho u_{t-1} + v_t$
3、 m 阶自回归 形式 AR(m)	$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_m u_{t-m} + v_t$
4、自相关时参数估计式的方差	$Var(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum_{t=1}^n x_t^2} \left(1 + 2\rho \frac{\sum_{t=1}^{n-1} x_t x_{t+1}}{\sum_{t=1}^n x_t^2} + 2\rho^2 \frac{\sum_{t=1}^{n-2} x_t x_{t+2}}{\sum_{t=1}^n x_t^2} + \cdots + 2\rho^{n-1} \frac{x_1 x_n}{\sum_{t=1}^n x_t^2} \right)$
5、 DW 统计量	$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$
6、 DW 值与 $\hat{\rho}$ 的关系	$DW \approx 2(1 - \hat{\rho})$
7、广义差分	$Y_t - \rho Y_{t-1} = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + u_t - \rho u_{t-1}$

思考题与练习题

思考题

6.1 如何使用 DW 统计量来进行自相关检验？该检验方法的前提条件和局限性有哪些？

6.2 当回归模型中的随机误差项为 AR(1)自相关时，为什么仍用 OLS 法会低估 $\hat{\beta}_j$ 的标准误差？

6.3 判断以下陈述的真伪，并给出合理的解释。

(1) 当回归模型随机误差项有自相关时，普通最小二乘估计量是有偏误的和非有效的。

(2) DW 检验假定随机误差项 u_i 的方差是同方差。

(3) 用一阶差分法消除自相关是假定自相关系数 ρ 为-1。

(4) 当回归模型随机误差项有自相关时，普通最小二乘估计的预测值的方差和标准误差不再是有效的。

6.4 对于四个解释变量的回归模型

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

如果样本量 $n=50$, 当 DW 统计量为如下数值时, 请判断模型中的自相关状况。

(1) DW=1.05 (2) DW=1.40

(3) DW=2.50 (4) DW=3.97

6.5 如何判别回归模型中的虚假自相关?

6.6 在回归模型

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

中, u_t 无自相关。如果我们错误地判定模型中有一阶自相关, 即 $u_t = \rho u_{t-1} + v_t$, 并使用了广义差分模型

$$Y_t - Y_{t-1} = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + v_t$$

将会产生什么问题?

练习题

6.1 下表给出了美国 1960-1995 年 36 年间个人实际可支配收入 X 和个人实际消费支出 Y 的数据。

美国个人实际可支配收入和个人实际消费支出

单位: 100 亿美元

年份	个人实际可支配收入 X	个人实际消费支出 Y	年份	个人实际可支配收入 X	个人实际消费支出 Y
1960	157	143	1978	326	295
1961	162	146	1979	335	302
1962	169	153	1980	337	301
1963	176	160	1981	345	305
1964	188	169	1982	348	308
1965	200	180	1983	358	324
1966	211	190	1984	384	341

1967	220	196	1985	396	357
1968	230	207	1986	409	371
1969	237	215	1987	415	382
1970	247	220	1988	432	397
1971	256	228	1989	440	406
1972	268	242	1990	448	413
1973	287	253	1991	449	411
1974	285	251	1992	461	422
1975	290	257	1993	467	434
1976	301	271	1994	478	447
1977	311	283	1995	493	458

注：资料来源于 Economic Report of the President，数据为 1992 年价格。

要求：（1）用普通最小二乘法估计收入—消费模型；

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

（2）检验收入—消费模型的自相关状况（5%显著水平）；

（3）用适当的方法消除模型中存在的问题。

6.2 在研究生产中劳动所占份额的问题时，古扎拉蒂采用如下模型

$$\text{模型 1} \quad Y_t = \alpha_0 + \alpha_1 t + u_t$$

$$\text{模型 2} \quad Y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + u_t$$

其中， Y 为劳动投入， t 为时间。据 1949-1964 年数据，对初级金属工业得到如下结果：

$$\text{模型 1} \quad \hat{Y}_t = 0.4529 - 0.0041t$$

$$t = (-3.9608)$$

$$R^2 = 0.5284 \quad DW = 0.8252$$

$$\text{模型 2} \quad \hat{Y}_t = 0.4786 - 0.0127t + 0.0005t^2$$

$$t = (-3.2724) (2.7777)$$

$$R^2 = 0.6629 \quad DW = 1.82$$

其中，括号内的数字为 t 统计量。

- 问：（1）模型 1 和模型 2 中是否有自相关；
- （2）如何判定自相关的存在？
- （3）怎样区分虚假自相关和真正的自相关。

6.3 下表是北京市连续 19 年城镇居民家庭人均收入与人均支出的数据。

北京市 19 年来城镇居民家庭收入与支出数据表（单位：元）

年份 顺序	人均收入 (元)	人均生活消 费支出(元)	商品零售 物价指数(%)	人均实 际收入(元)	人均实际 支出(元)
1	450.18	359.86	100.00	450.18	359.86
2	491.54	408.66	101.50	484.28	402.62
3	599.40	490.44	108.60	551.93	451.60
4	619.57	511.43	110.20	562.22	464.09
5	668.06	534.82	112.30	594.89	476.24
6	716.60	574.06	113.00	634.16	508.02
7	837.65	666.75	115.40	725.87	577.77
8	1158.84	923.32	136.80	847.11	674.94
9	1317.33	1067.38	145.90	902.90	731.58
10	1413.24	1147.60	158.60	891.07	723.58
11	1767.67	1455.55	193.30	914.47	753.00
12	1899.57	1520.41	229.10	829.14	663.64
13	2067.33	1646.05	238.50	866.81	690.17
14	2359.88	1860.17	258.80	911.85	718.77
15	2813.10	2134.65	280.30	1003.60	761.56
16	3935.39	2939.60	327.70	1200.91	897.04
17	5585.88	4134.12	386.40	1445.62	1069.91
18	6748.68	5019.76	435.10	1551.06	1153.70
19	7945.78	5729.45	466.90	1701.82	1227.13

- 要求：（1）建立居民收入—消费函数；
- （2）检验模型中存在的问题，并采取适当的补救措施予以处理；

(3) 对模型结果进行经济解释。

6.4 下表给出了日本工薪家庭实际消费支出与可支配收入数据

日本工薪家庭实际消费支出与实际可支配收入			单位: 1000 日元		
年份	个人实际可支配收入 X	个人实际消费支出 Y	年份	个人实际可支配收入 X	个人实际消费支出 Y
1970	239	300	1983	304	384
1971	248	311	1984	308	392
1972	258	329	1985	310	400
1973	272	351	1986	312	403
1974	268	354	1987	314	411
1975	280	364	1988	324	428
1976	279	360	1989	326	434
1977	282	366	1990	332	441
1978	285	370	1991	334	449
1979	293	378	1992	336	451
1980	291	374	1993	334	449
1981	294	371	1994	330	449
1982	302	381			

注: 资料来源于日本银行《经济统计年报》数据为 1990 年价格。

要求: (1) 建立日本工薪家庭的收入—消费函数;

(2) 检验模型中存在的问题, 并采取适当的补救措施予以处理;

(3) 对模型结果进行经济解释。

6.5 下表给出了中国进口需求(Y)与国内生产总值(X)的数据。

1985~2003 年中国实际 GDP、进口需求		单位: 亿元
年份	实际 GDP (X , 亿元)	实际进口额 (Y , 亿元)
1985	8964.40	2543.2

1986	9753.27	2983.4
1987	10884.65	3450.1
1988	12114.62	3571.6
1989	12611.32	3045.9
1990	13090.55	2950.4
1991	14294.88	3338.0
1992	16324.75	4182.2
1993	18528.59	5244.4
1994	20863.19	6311.9
1995	23053.83	7002.2
1996	25267.00	7707.2
1997	27490.49	8305.4
1998	29634.75	9301.3
1999	31738.82	9794.8
2000	34277.92	10842.5
2001	36848.76	12125.6
2002	39907.21	14118.8
2003	43618.58	17612.2

注：表中数据来源于《中国统计年鉴 2004》光盘。实际 GDP 和实际进口额均为 1985 年可比价指标。

要求：（1）检测进口需求模型 $Y_t = \beta_1 + \beta_2 X_t + u_t$ 的自相关性；

（2）采用科克伦—奥克特迭代法处理模型中的自相关问题。

6.6 下表给出了某地区 1980-2000 年的地区生产总值(Y)与固定资产投资额(X)的数据。

地区生产总值(Y)与固定资产投资额(X)			单位：亿元		
年份	地区生产 总值(Y)	固定资 产投资额(X)	年份	地区生产 总值(Y)	固定资 产投资额(X)
1980	1402	216	1990	3124	544
1981	1624	254	1991	3158	523
1982	1382	187	1992	3578	548

1983	1285	151	1993	4067	668
1984	1665	246	1994	4483	699
1985	2080	368	1995	4897	745
1986	2375	417	1996	5120	667
1987	2517	412	1997	5506	845
1988	2741	438	1998	6088	951
1989	2730	436	1999	7042	1185
			2000	8756	1180

要求：(1) 使用对数线性模型 $\ln Y_t = \beta_1 + \beta_2 \ln X_t + u_t$ 进行回归，并检验回归模型的自相关性；

(2) 采用广义差分法处理模型中的自相关问题。

(3) 令 $X_t^* = X_t / X_{t-1}$ (固定资产投资指数)， $Y_t^* = Y_t / Y_{t-1}$ (地区生产总值增长指数)，使用模型 $\ln Y_t^* = \beta_1 + \beta_2 \ln X_t^* + v_t$ ，该模型中是否有自相关？

第六章附录：

附录 6.1：存在自相关时参数估计值方差的证明

$$\begin{aligned}
\text{Var}(\hat{\beta}_2) &= E(\hat{\beta}_2 - \beta_2)^2 \\
&= E\left(\frac{\sum x_t u_t}{\sum x_t^2}\right)^2 \\
&= \left(\frac{1}{\sum x_t^2}\right)^2 E(x_1 u_1 + x_2 u_2 + \cdots + x_n u_n)^2 \\
&= \left(\frac{1}{\sum x_t^2}\right)^2 E[(x_1^2 u_1^2 + x_2^2 u_2^2 + \cdots + x_n^2 u_n^2) \\
&\quad + 2(x_1 x_2 u_1 u_2 + x_1 x_3 u_1 u_3 + \cdots + x_{n-1} x_n u_{n-1} u_n)] \\
&= \left(\frac{1}{\sum x_t^2}\right)^2 [(x_1^2 E(u_1^2) + x_2^2 E(u_2^2) + \cdots + x_n^2 E(u_n^2)) \\
&\quad + 2[x_1 x_2 E(u_1 u_2) + x_1 x_3 E(u_1 u_3) + \cdots + x_{n-1} x_n E(u_{n-1} u_n)]]
\end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma_u^2}{\sum x_t^2} + \frac{2}{(\sum x_t^2)^2} [x_1 x_2 \rho \sigma_u^2 + x_1 x_3 \rho^2 \sigma_u^2 + \cdots + x_{n-1} x_n \rho \sigma_u^2] \\
&= \frac{\sigma_u^2}{\sum_{t=1}^n x_t^2} (1 + 2\rho \frac{\sum_{t=1}^{n-1} x_t x_{t+1}}{\sum_{t=1}^n x_t^2} + 2\rho^2 \frac{\sum_{t=1}^{n-2} x_t x_{t+2}}{\sum_{t=1}^n x_t^2} + \cdots + 2\rho^{n-1} \frac{x_1 x_n}{\sum_{t=1}^n x_t^2})
\end{aligned}$$

第七章 分布滞后模型与自回归模型

引子:

货币政策效应的时滞

货币供给的变化对经济的影响很大，因此货币政策总是备受关注。当出现通货膨胀时，总是会要求控制货币供给量；当经济有衰退迹象时，又要求货币当局采用积极的货币政策来刺激经济复苏。可是人们发现，货币政策的传导总有个过程，当年的货币政策的效应总是难以立即显现出来，好象对当年 GDP 没有多少直接影响，货币供应量的增减也不会立即导致物价的变动。也就是说，货币政策的影响效应存在着时间上的滞后。在货币政策的传导过程中，货币扩张首先促使利率降低，或者一般价格水平的上升，这本身就需要一段时间。而这些因素对以 GDP 为代表的经济增长的影响，更是需要一段时间才能显示出来。只有经过一段时间以后，支出对利率的反应增强，投资、进出口和消费才会不断上升，货币政策才最终促使 GDP 增加。通常，货币扩张对 GDP 影响的最高点可能是在政策实施以后的一到两年间达到，货币当局在制定货币政策时，必须考虑未来一两年可能的经济情况。

在现实经济活动中，滞后现象是普遍存在的，这就要求我们在做经济分析时应该考虑时滞的影响。怎样才能把这类时间上滞后的经济关系纳入计量经济模型呢？

前面各章所讨论的回归模型属于静态模型，即认为被解释变量的变化仅仅依赖于解释变量的当期影响，没有考虑变量之间的前后联系。事实上，在现实经济活动中，由于经济活动主体的决策与行动都需要一个过程，加之人们生活习惯的延续、制度或技术条件的限制以及预期效应等因素的影响，经济变量的变化往往存在时滞现象。因此，为了探索受时滞因素影响的经济变量的变化规律，需要在回归模型中引入滞后变量进行分析。本章主要介绍经济分析中较为常用的分布滞后模型与自回归模型，讨论它们的产生背景、特点及估计。

第一节 滞后效应与滞后变量模型

一、经济活动中的滞后现象

一般来说，解释变量对被解释变量的影响不可能在短时间内完成，在这一过程中通常都存在时间滞后，也就是说解释变量需要通过一段时间才能完全作用于被解释变量。此外，由于经济活动的惯性，一个经济指标以前的变化态势往往会延续到本期，从而形成被解释变量

的当期变化同自身过去取值水平相关的情形。这种被解释变量受自身或其它经济变量过去值影响的现象称为滞后效应。

下面我们看两个涉及滞后效应的例子。

【例 7.1】 消费滞后

消费者的消费水平，不仅依赖于当年的收入，还同以前的收入水平有关。一般来说，消费者不会把当年的收入全部花光。假定消费者将每一年收入的 40% 用于当年花费，30% 用于第二年花费，20% 用于第三年花费，其余的作为长期储蓄。这样，该消费者的消费函数就可以表示成：

$$Y_t = \alpha + 0.4X_t + 0.3X_{t-1} + 0.2X_{t-2} + u_t$$

其中， Y_t 、 X_t 分别为第 t 年的消费和收入， α 为常数。

【例 7.2】 通胀滞后

通货膨胀与货币供应量的变化有着较为密切的联系。物价上涨最直接的原因是相对于流通中商品和服务的价值量来说货币供应过多，货币的超量供应通常是通货膨胀产生的必要条件。但是，货币供应量的变化对通货膨胀的影响并不是即期的，总存在一定时滞。美国一学者在研究通胀滞后效应时，就采用了如下模型：

$$P_t = \alpha + \beta_0 M_t + \beta_1 M_{t-1} + \beta_2 M_{t-2} + \cdots + \beta_s M_{t-s} + u_t$$

其中， P_t 、 M_t 分别为第 t 季度的物价指数和广义货币的增长率， s 是滞后（时滞）期。通过对实际数据的分析发现，西方发达国家的通货膨胀时滞期 s 大约为 2—3 个季度。

二、滞后效应产生的原因

为什么经济变量会存在滞后现象呢？原因众多，但主要有以下几方面：

1、心理预期因素

经济社会是一个复杂的有机体系，经济活动离不开人的参与，在这个系统中，人的心理因素对经济变量的变化有很大影响。由于人们的心理定势及社会习惯的作用，适应新经济条件和经济环境需要一个过程，从而表现为决策滞后。而且，经济主体的大多数行动，都会受到预期心理的影响。以消费为例，人们对某种商品的消费量不仅受商品当前价格影响，而且还受预期价格影响，当人们预计价格上涨时，就会加快当期的购买，而当人们预期价格要下降时，则会持币观望，减少当期的购买。由于对将来的预期要依据过去的经验，因此在一定条件下，这种“预期”因素的影响可转化为滞后效应。

2、技术因素

在国民经济运行中，从生产到流通再到使用，每一个环节都需要一段时间，从而形成时滞。例如，农产品产量对价格信息的反应总是滞后的，其原因就在于农产品的生产需要一个较长的时间过程；又例如，在工业生产中，当年的产出量会在某种程度上依赖于过去若干期内投资形成的固定资产规模；再例如，货币投放量的增减对物价水平会产生影响，但这种影响并不会全部在当期内反映，总会滞后一段时期。这些滞后效应都是因为经济活动的技术因素所致。

3、制度因素

契约、管理制度等因素也会形成一定程度的滞后。例如，企业要改变它的产品结构或产量，会受到过去签订的供货合同的制约；拥有一定数量定期存款的消费者，要调整自己的消费水平，会受到银行契约制度的限制；此外，管理层次过多、管理的低效率也会造成滞后效应。这些情况说明，当一种变量发生变化时，另一变量由于制度方面的原因，需经过一定时期才能作出相应的变动，从而形成滞后现象。

三、滞后变量模型

所谓滞后变量，是指过去时期的、对当前被解释变量产生影响的变量。滞后变量可分为滞后解释变量与滞后被解释变量两类。把滞后变量引入回归模型，这种回归模型称为滞后变量模型。在经济分析中，运用滞后变量模型可以使不同时期的经济现象彼此联系起来，同时也将经济活动的静态分析转化为动态分析，使模型更加切合实际经济的运行状况。

滞后变量模型的一般形式为

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_s X_{t-s} + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \cdots + \gamma_q Y_{t-q} + u_t \quad (7.1)$$

其中 s 、 q 分别为滞后解释变量和滞后被解释变量的滞后期长度。若滞后期长度为有限，称模型为有限滞后变量模型；若滞后期长度为无限，称模型为无限滞后变量模型。

1、分布滞后模型

如果滞后变量模型中没有滞后被解释变量，被解释变量只受解释变量的影响，且这种影响分布在解释变量不同时期的滞后值上，即模型形如

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_s X_{t-s} + u_t \quad (7.2)$$

具有这种滞后分布结构的模型称为**分布滞后模型**，其中 s 为滞后长度。根据滞后长度 s 取值的有限和无限，我们将模型分别称为有限分布滞后模型和无限分布滞后模型。前面两个例子

中所设定的回归模型就属于有限分布滞后模型。

在分布滞后模型中，各系数体现了解释变量的各个滞后值对被解释变量的不同影响程度，即通常所说的乘数效应：

β_0 ：称为短期乘数或即期乘数，表示本期 X 变动一个单位对 Y 值的影响大小；

β_i ：称为延迟乘数或动态乘数($i = 1, 2, \dots, s$)，表示过去各时期 X 变动一个单位对 Y 值的影响大小；

$\sum_{i=0}^s \beta_i$ ：称为长期乘数或总分布乘数，表示 X 变动一个单位时，包括滞后效应而形成的对 Y 总的影

响。

2、自回归模型

如果滞后变量模型的解释变量仅包括自变量 X 的当期值和被解释变量的若干期滞后值，即模型形如

$$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \dots + \gamma_q Y_{t-q} + u_t \quad (7.3)$$

则称这类模型为**自回归模型**，其中 q 称为自回归模型的阶数。

第二节 分布滞后模型的估计

一、分布滞后模型估计的困难

如前所述，分布滞后模型可分为有限分布滞后模型与无限分布滞后模型两类。对于无限分布滞后模型，由于滞后项无限多而样本观测总是有限的，因此不能直接对其进行估计。对于有限分布滞后模型，如果随机扰动项满足古典假定，可以考虑用最小二乘法对模型进行估计。阿尔特（Alt）和丁伯根（Tinbergen）曾建议使用最小二乘法递推地估计模型，其基本思路是，首先做被解释变量 Y_t 关于解释变量 X_t 的回归，然后做 Y_t 关于 X_t 和 X_{t-1} 的回归，再做 Y_t 关于 X_t 、 X_{t-1} 和 X_{t-2} 的回归，依次添加解释变量 X_t 的滞后项，直到滞后变量的回归系数开始变成统计上不显著或至少有一个变量的系数改变符号时结束。例如，为了获得燃油消耗量 Y 与订货量 X 之间的关系，阿尔特（Alt）曾利用十年的季度数据递推估计回归方程，得到如下结果：

$$\hat{Y}_t = 8.37 + 0.171X_t$$

$$\hat{Y}_t = 8.27 + 0.111X_t + 0.064X_{t-1}$$

$$\hat{Y}_t = 8.27 + 0.109X_t + 0.071X_{t-1} - 0.055X_{t-2}$$

$$\hat{Y}_t = 8.32 + 0.108X_t + 0.063X_{t-1} + 0.022X_{t-2} - 0.020X_{t-3}$$

根据回归结果，由于 X_{t-2} 的符号不稳定，并且 X_{t-2} 、 X_{t-3} 的符号为负，其经济意义难于解释，所以阿尔特（Alt）最后选择第二个回归模型作为最佳模型。

上述估计法表面上看似可行，但事实上还存在一些缺陷：

1、自由度问题

假设有限分布滞后模型的滞后长度为 s ，如果样本观测值个数 n 较小，随着滞后长度 s 的增大，有效样本容量 $n-s$ 变小，会出现自由度不足的问题。由于自由度的过分损失，致使估计偏差增大，统计显著性检验失效。

2、多重共线性问题

由于经济活动的前后继起性，经济变量的滞后值之间通常存在较强的联系，因此，分布滞后模型中滞后解释变量观测值之间往往会存在严重的多重共线性问题。如果直接使用最小二乘法进行估计，则至少有些参数的估计会有较大偏差，可能导致一些重要的滞后变量被剔除。

3、滞后长度难于确定

在实际经济分析中用分布滞后模型来处理滞后现象时，模型中滞后长度的确定较为困难，没有充分的先验信息可供使用。

针对分布滞后模型直接估计存在一些缺陷，人们进行了广泛的研究，提出了一系列修正估计方法。对于有限分布滞后模型，其基本思想是对滞后模型中系数施加某种约束，设法有目的地减少需要直接估计的模型参数的个数，以缓解多重共线性，保证自由度。对于无限分布滞后模型，主要是通过适当的模型变换，使其转化为只需估计有限个参数的自回归模型。有限分布滞后模型的常用估计方法主要有经验加权法、阿尔蒙法等。

二、经验加权估计法

所谓**经验加权估计法**，是根据实际经济问题的特点及经验判断，形成相应的约束，对解释变量的系数赋予一定的权数，利用这些权数构成各滞后变量的线性组合，以形成新的变量，再应用最小二乘法进行估计。权数分布的确定取决于模型滞后结构的不同类型，常见的滞后结构类型有：

（1）递减滞后结构。这类滞后结构假定权数是递减的，认为滞后解释变量对被解释变量的影响随着时间的推移越来越小，即遵循远小近大的原则（如图 7.1(a)）。这种滞后结构

在现实经济活动中较为常见，比较典型的例子是消费函数，显然，现期收入对消费的影响较大，越滞后，影响越小。

(2) 不变滞后结构。这类滞后结构假定权数不变，即认为滞后解释变量对被解释变量的影响不随时间而变化(如图 7.1(b))。

(3) Λ 型滞后结构。即两头小中间大，权数先递增后递减呈 Λ 型(如图 7.1(c))。这类滞后结构适合于前后期滞后解释变量对被解释变量的影响不大，而中期滞后解释变量对被解释变量的影响较大的分布滞后模型。如投资对产出的影响，就是以周期期中的投资对本期产出贡献最大，因此可选择 Λ 型滞后结构。

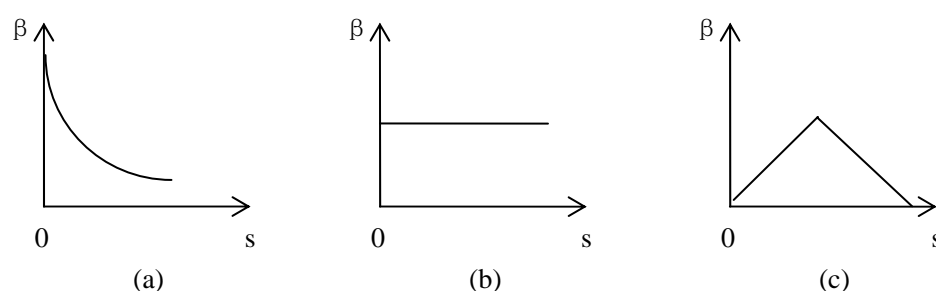


图 7.1 常见的滞后结构类型

例如，假设某经济变量服从一个滞后 3 期的分布滞后模型

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + u_t$$

如果根据经验判断滞后解释变量对被解释变量的影响递减，权数取某种形式，比如为

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}$$

即有：

$$Y_t = \alpha + \beta_0 \left(\frac{1}{2} X_t \right) + \beta_0 \left(\frac{1}{4} X_{t-1} \right) + \beta_0 \left(\frac{1}{6} X_{t-2} \right) + \beta_0 \left(\frac{1}{8} X_{t-3} \right) + u_t$$

$$= \alpha + \beta_0 \left(\frac{1}{2} X_t + \frac{1}{4} X_{t-1} + \frac{1}{6} X_{t-2} + \frac{1}{8} X_{t-3} \right) + u_t$$

则新的线性组合变量为

$$Z_t = \frac{1}{2} X_t + \frac{1}{4} X_{t-1} + \frac{1}{6} X_{t-2} + \frac{1}{8} X_{t-3}$$

原模型就变为经验加权模型

$$Y_t = \alpha + \beta_0 Z_t + u_t$$

若随机扰动项与解释变量不相关，从而与滞后解释变量的线性组合变量也不相关，因此可直

接应用最小二乘法对该模型进行估计。

经验加权法具有简单易行、不损失自由度、避免多重共线性干扰及参数估计具有一致性等特点。缺点是设置权数的主观随意性较大，要求分析者对实际问题的特征有比较透彻的了解。通常的做法是，依据先验信息，多选几组权数分别估计多个模型，然后根据可决系数、F-检验值、t-检验值、估计标准误以及 D-W 值，从中选出最佳估计方程。

【例 7.3】 已知 1955—1974 年期间美国制造业库存量 Y 和销售额 X 的统计资料如表 7.1（金额单位：亿美元）。设定有限分布滞后模型为：

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + u_t$$

运用经验加权法，选择下列三组权数（1）1，1/2，1/4，1/8；（2）1/4，1/2，2/3，1/4；（3）1/4，1/4，1/4，1/4；分别估计上述模型，并从中选择最佳的方程。

表 7.1

年份	Y	X	年份	Y	X
1955	450.69	264.80	1965	682.21	410.03
1956	506.42	277.40	1966	779.65	448.69
1957	518.70	287.36	1967	846.55	464.49
1958	500.70	272.80	1968	908.75	502.82
1959	527.07	302.19	1969	970.74	535.55
1960	538.14	307.96	1970	1016.45	528.59
1961	549.39	308.96	1971	1024.45	559.17
1962	582.13	331.13	1972	1077.19	620.17
1963	600.43	350.32	1973	1208.70	713.98
1964	633.83	373.35	1974	1471.35	820.98

数据来源：转摘自 D. N.Gujarati(古扎拉蒂),《计量经济学》(Basic Econometrics), 中译本, 中国人民大学出版社 2000, 第 611 页。

记新的线性组合变量分别为：

$$Z_1 = X_t + \frac{1}{2} X_{t-1} + \frac{1}{4} X_{t-2} + \frac{1}{8} X_{t-3}$$

$$Z_2 = \frac{1}{4} X_t + \frac{1}{2} X_{t-1} + \frac{2}{3} X_{t-2} + \frac{1}{4} X_{t-3}$$

$$Z_3 = \frac{1}{4} X_t + \frac{1}{4} X_{t-1} + \frac{1}{4} X_{t-2} + \frac{1}{4} X_{t-3}$$

在 Eviews 中，输入 X 和 Y 的数据，根据 X 的数据，由上述公式生成线性组合变量 Z₁、Z₂、Z₃ 的数据。然后分别估计如下经验加权模型

$$Y_t = \alpha + \beta Z_{kt} + u_t \quad k = 1, 2, 3$$

回归分析结果整理如下：

模型一：

$$\begin{aligned} \hat{Y}_t &= -66.60404 + 1.071502 Z_{1t} \\ &\quad (-3.6633) \quad (50.9191) \\ R^2 &= 0.994248 \quad DW = 1.440858 \\ F &= 2592 \end{aligned}$$

模型二：

$$\begin{aligned} \hat{Y}_t &= -133.1988 + 1.3667 Z_{2t} \\ &\quad (-5.029) \quad (37.35852) \\ R^2 &= 0.989367 \quad DW = 1.042935 \\ F &= 1396 \end{aligned}$$

模型三：

$$\begin{aligned} \hat{Y}_t &= -121.7394 + 2.23973 Z_{3t} \\ &\quad (-4.8131) \quad (38.68578) \\ R^2 &= 0.990077 \quad DW = 1.15853 \\ F &= 1496 \end{aligned}$$

从上述回归分析结果可以看出，模型一的扰动项无一阶自相关，模型二、模型三？扰动项存在一阶正自相关；再综合判断可决系数、F-检验值、t-检验值，可以认为：最佳的方程是模型一，即权数为（1，1/2，1/4，1/8）的分布滞后模型。

三、阿尔蒙法

为了消除多重共线性的影响，阿尔蒙（Almon）提出利用多项式来逼近滞后参数的变化结构，从而减少待估参数的数目。其基本原理是，在有限分布滞后模型滞后长度 s 已知的情况下，滞后项系数可以看成是相应滞后期 i 的函数。在以滞后期 i 为横轴、滞后系数取值为纵轴的坐标系中，如果这些滞后系数落在一条光滑曲线上，或近似落在一条光滑曲线上，则可以由一个关于 i 的次数较低的 m 次多项式很好地逼近，即

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \cdots + \alpha_m i^m \quad i = 0, 1, 2, \cdots, s; \quad m < s \quad (7.4)$$

此式称为阿尔蒙多项式变换（图 7.2）。

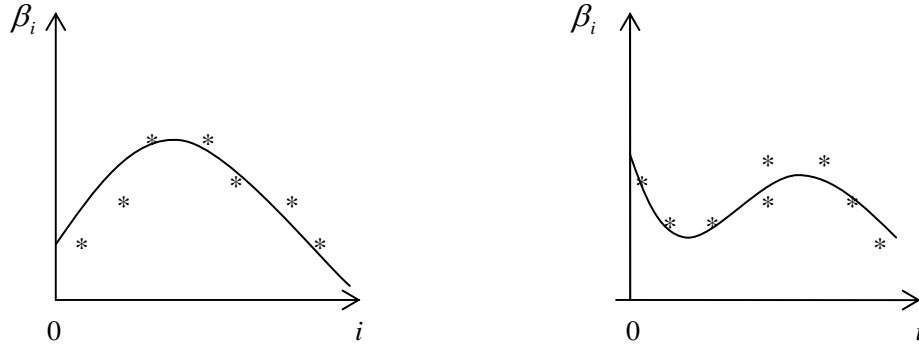


图 7.2 阿尔蒙多项式滞后结构

将阿尔蒙多项式变换具体列出来就是

$$i=0 \quad \beta_0 = \alpha_0 + \alpha_1 0 + \alpha_2 0^2 + \cdots + \alpha_m 0^m$$

$$i=1 \quad \beta_1 = \alpha_0 + \alpha_1 1 + \alpha_2 1^2 + \cdots + \alpha_m 1^m$$

$$i=2 \quad \beta_2 = \alpha_0 + \alpha_1 2 + \alpha_2 2^2 + \cdots + \alpha_m 2^m$$

.....

$$i=s \quad \beta_s = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \cdots + \alpha_m s^m$$

代入 (7.2) 式并整理各项, 模型变为如下形式

$$\begin{aligned} Y_t = & \alpha + \alpha_0 (X_t + X_{t-1} + X_{t-2} + \cdots + X_{t-s}) \\ & + \alpha_1 (X_{t-1} + 2X_{t-2} + 3X_{t-3} \cdots + sX_{t-s}) \\ & + \alpha_2 (X_{t-1} + 2^2 X_{t-2} + 3^2 X_{t-3} \cdots + s^2 X_{t-s}) \\ & \vdots \\ & + \alpha_m (X_{t-1} + 2^m X_{t-2} + 3^m X_{t-3} \cdots + s^m X_{t-s}) \\ & + u_t \end{aligned}$$

即
$$Y_t = \alpha + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \cdots + \alpha_m Z_{mt} + u_t \quad (7.5)$$

其中

$$\begin{aligned} Z_{0t} &= X_t + X_{t-1} + X_{t-2} + \cdots + X_{t-s} \\ Z_{1t} &= X_{t-1} + 2X_{t-2} + 3X_{t-3} \cdots + sX_{t-s} \\ Z_{2t} &= X_{t-1} + 2^2 X_{t-2} + 3^2 X_{t-3} \cdots + s^2 X_{t-s} \\ &\vdots \\ Z_{mt} &= X_{t-1} + 2^m X_{t-2} + 3^m X_{t-3} \cdots + s^m X_{t-s} \end{aligned}$$

为滞后变量的线性组合变量。

对于模型 (7.5)，在 u_t 满足古典假定的条件下，可用最小二乘法进行估计。将估计的参数 $\hat{\alpha}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_m$ 代入 (7.4) 式，就可求出原分布滞后模型参数的估计值。

在实际应用中，阿尔蒙多项式的次数 m 通常取得较低，一般取 2 或 3，很少超过 4。如果 m 取得过大则达不到通过阿尔蒙多项式变换减少变量个数的目的。

从上面的分析可以看出，通过阿尔蒙多项式变换，新模型中的变量个数少于原分布滞后模型中的变量个数，从而自由度得到保证，并在一定程度上缓解了多重共线性问题。

第三节 自回归模型的构建

在处理实际经济问题中，有时需要使用自回归模型进行分析。引入自回归模型主要有两条途径，一是对无限分布滞后模型的滞后结构作出某种假定，通过变换而形成；另一条途径是在模型中考虑了预期因素，然后基于经济原理对“期望模型”作出某种假定而导出。

一、库伊克 (Koyck) 模型

许多经济变量的滞后效应都在相当长的时期内存在。例如消费水平受收入的影响，可以追溯到较远的过去时期的收入水平；经济政策对经济效益的影响有一个逐步扩散的过程，目前的经济效益除了受不久前经济政策的影响外，还要受很久以前经济政策的影响，尽管这种影响可能很微弱。对于这种滞后现象，如果采用截尾的办法忽略某滞后期以前滞后解释变量对被解释变量的影响，建立有限分布滞后模型来进行分析，则存在滞后长度难于确定的问题。为了回避这一难点，可使用无限分布滞后模型来处理。

但是，正如前面所述，无限分布滞后模型中滞后项无限多，而样本观测总是有限的，因此不可能对其直接进行估计。显然，要使模型估计能够顺利进行，必须施加一些约束或假定条件，将模型的结构作某种转化。库伊克变换就是其中较具代表性的方法。

库伊克认为，对于如下无限分布滞后模型：

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + u_t \quad (7.6)$$

可以假定滞后解释变量 X_{t-i} 对被解释变量 Y 的影响随着滞后期 i ($i = 0, 1, 2, \dots$) 的增加而按几何级数衰减。即滞后系数的衰减服从某种公比小于 1 的几何级数：

$$\beta_i = \beta_0 \lambda^i, \quad 0 < \lambda < 1, \quad i = 0, 1, 2, \dots \quad (7.7)$$

其中 β_0 为常数，公比 λ 为待估参数。 λ 值的大小决定了滞后衰减的速度， λ 值越接近零，衰减速度越快（如图 7.3），通常称 λ 为分布滞后衰减率，称 $1 - \lambda$ 为调整速度。

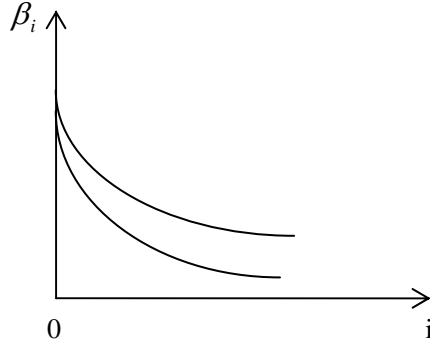


图 7.3 按几何级数衰减的滞后结构（库伊克）

将 (7.7) 式代入 (7.6) 式，得

$$\begin{aligned} Y_t &= \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \dots + u_t \\ &= \alpha + \beta_0 (X_t + \lambda X_{t-1} + \lambda^2 X_{t-2} + \dots) + u_t \\ &= \alpha + \beta_0 \sum_{i=0}^{\infty} \lambda^i X_{t-i} + u_t \end{aligned} \quad (7.8)$$

将 (7.8) 滞后一期，有

$$\begin{aligned} Y_{t-1} &= \alpha + \beta_0 \sum_{i=0}^{\infty} \lambda^i X_{t-1-i} + u_{t-1} \\ &= \alpha + \beta_0 \sum_{i=1}^{\infty} \lambda^{i-1} X_{t-i} + u_{t-1} \end{aligned} \quad (7.9)$$

对 (7.9) 式两边同乘 λ 并与 (7.8) 式相减，得

$$\begin{aligned} Y_t - \lambda Y_{t-1} &= (\alpha + \beta_0 \sum_{i=0}^{\infty} \lambda^i X_{t-i} + u_t) - (\lambda \alpha + \beta_0 \sum_{i=1}^{\infty} \lambda^i X_{t-i} + \lambda u_{t-1}) \\ &= \alpha(1 - \lambda) + \beta_0 X_t + (u_t - \lambda u_{t-1}) \end{aligned} \quad (7.10)$$

即

$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + (u_t - \lambda u_{t-1}) \quad (7.11)$$

这就是库伊克模型。上述变换过程也叫库伊克变换。

$$\text{令 } \alpha^* = (1 - \lambda)\alpha, \quad \beta_0^* = \beta_0, \quad \beta_1^* = \lambda, \quad u_t^* = u_t - \lambda u_{t-1}$$

则库伊克模型 (7.10) 式变为

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^* \quad (7.12)$$

这是一个一阶自回归模型。

由此可见，利用库伊克变换，可以将一个无限分布滞后模型变成只有一个本期解释变量 X_t 和滞后一期被解释变量 Y_{t-1} 的自回归模型。该模型以一个滞后被解释变量 Y_{t-1} 代替了大量的滞后解释变量 $X_{t-i} (i = 1, 2, \dots)$ ，使模型结构得到极大简化，而且最大限度地保证了自由度，解决了滞后长度难以确定的问题；同时，滞后一期的被解释变量 Y_{t-1} 与 X_t 的线性相关程度将低于 X 的各滞后值之间的相关程度，从而在很大程度上缓解了多重共线性。

当然，尽管库伊克变换具有上述优点，但也存在一些缺陷：

(1) 它假定无限滞后分布呈几何滞后结构，即滞后影响按某固定比例递减，解释变量当期值对被解释变量影响最大，滞后一期次之，并依此类推。这种假定对某些经济变量可能不适用，例如固定资产投资对总产出影响的滞后结构就不是这种类型。

(2) 库伊克模型的随机扰动项形如

$$u_t^* = u_t - \lambda u_{t-1} \quad (7.13)$$

说明新模型的随机扰动项 u_t^* 存在一阶自相关，且与解释变量 Y_{t-1} 相关。

(3) 将随机变量 Y_{t-1} 作为解释变量引入了模型，不一定符合基本假定。

(4) 库伊克变换是纯粹的数学运算结果，缺乏经济理论依据。

这些缺陷，特别是第二个缺陷，将给模型的参数估计带来一定困难。

二、自适应预期模型

在经济活动中，经济活动主体经常根据他们对某些经济变量未来走势的“预期”来改变自己的行为决策。例如，一家公司的价值在过去稳步增长，投资者就可能会预期这种情况会持续下去，并依据这种预期作出投资决策。又例如，企业会根据对产品未来价格走势的预期，决定现期的生产量以及是否对新设备进行投资。同样，为了确定种植哪种农作物最有利可图，农民往往要对各种农作物的未来价格进行预测。消费者在决定是否购买房屋、汽车或家用电器时，也需对这些消费品的未来价格进行预测。再例如，当期居民消费水平的高低，在一定程度上取决于对未来收入水平的预计，即取决于预期的收入水平。

这些例子表明，某些经济变量的变化会或多或少地受到另一些经济变量预期值的影响。为了处理这种经济现象，我们可以将解释变量预期值引入模型建立“期望模型”。例如，包

含一个预期解释变量的“期望模型”可以表现为如下形式：

$$Y_t = \alpha + \beta X_t^* + u_t \quad (7.14)$$

其中， Y_t 为被解释变量， X_t^* 为解释变量预期值， u_t 为随机扰动项。

在回归分析中，如何获取解释变量预期值，是上述模型的难点。预期是对未来的判断，在大多数情况下，预期值是不可观测的。因此，实际应用中需要对预期的形成机理作出某种假定。自适应预期假定就是其中之一，这种假定比较切合实际，具有一定代表性。

自适应预期假定认为，经济活动主体对某经济变量的预期，是通过一种简单的学习过程而行成的，其机理是，经济活动主体会根据自己过去在作预期时所犯错误的程度，来修正他们以后每一时期的预期，即按照过去预测偏差的某一比例对当前期望进行修正，使其适应新的经济环境。用数学式子表示就是

$$X_t^* = X_{t-1}^* + \gamma(X_t - X_{t-1}^*) \quad (7.15)$$

其中参数 γ 为调节系数，也称为适应系数。也就是说，本期预期值 X_t^* 等于前一期预期值 X_{t-1}^* 加上一修正量，该修正量 $\gamma(X_t - X_{t-1}^*)$ 是前一期预期误差 $(X_t - X_{t-1}^*)$ 的一部分。这一调整过程叫做自适应过程。

将 (7.15) 式改写为

$$X_t^* = \gamma X_t + (1 - \gamma) X_{t-1}^* \quad (7.16)$$

表明本期预期值是前一期预期值和本期实际值的加权平均，权数分别 $1 - \gamma$ 和 γ 。如果 γ 等于 0，说明本期实际值被忽略，预期没有进行修正。如果 γ 等于 1，则以本期实际值作为预期值，本期预期与前一期预期无关。在一般情况下， $0 \leq \gamma \leq 1$ 。

通常，将解释变量预期值满足自适应调整过程的期望模型，称为自适应预期模型 (Adaptive expectation model)。根据自适应预期假定，自适应预期模型可转化为自回归形式。

将 (7.16) 式代入 (7.14) 式得

$$Y_t = \alpha + \beta [\gamma X_t + (1 - \gamma) X_{t-1}^*] + u_t \quad (7.17)$$

同时，将 (7.14) 式滞后一期，并乘以 $1 - \gamma$ ，得

$$(1 - \gamma) Y_{t-1} = \alpha(1 - \gamma) + \beta(1 - \gamma) X_{t-1}^* + (1 - \gamma) u_{t-1} \quad (7.18)$$

(7.17) 式减去 (7.18) 式，整理得

$$Y_t = \gamma\alpha + \gamma\beta X_t + (1-\gamma)Y_{t-1} + [u_t - (1-\gamma)u_{t-1}] \quad (7.19)$$

令

$$\alpha^* = \gamma\alpha, \quad \beta_0^* = \gamma\beta, \quad \beta_1^* = 1-\gamma, \quad u_t^* = u_t - (1-\gamma)u_{t-1} \quad (7.20)$$

则 (7.19) 式变为

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^* \quad (7.21)$$

这是一个一阶自回归模型。如果能得到该模型参数 $\alpha^*, \beta_0^*, \beta_1^*$ 的估计值，代入 (7.20) 式即可求得自适应预期模型 (7.14) 的参数估计值。

三、局部调整模型

在经济活动中，会遇到为了适应解释变量的变化，被解释变量有一个预期的最佳值与之对应的现象。例如，企业为了确保生产或供应，必须保持一定的原材料储备，对应于一定的产量或销售量，存在着预期最佳库存量；为了确保一国经济健康发展，中央银行必须保持一定的货币供应，对应于一定的经济总量水平，应该有一个预期的最佳货币供应量。也就是说，解释变量的现值影响着被解释变量的预期值，即存在如下关系

$$Y_t^* = \alpha + \beta X_t + u_t \quad (7.22)$$

其中， Y_t^* 为被解释变量的预期最佳值， X_t 为解释变量的现值。

由于技术、制度、市场以及管理等各方面的限制，被解释变量的预期水平在单一周期内一般不会完全实现，而只能得到部分的调整。局部调整假设认为，被解释变量的实际变化仅仅是预期变化的一部分，即

$$Y_t - Y_{t-1} = \delta(Y_t^* - Y_{t-1}) \quad (7.23)$$

其中 δ 为调整系数，它代表调整速度。 δ 越接近 1，表明调整到预期最佳水平的速度越快。

若 $\delta = 1$ ，则 $Y_t = Y_t^*$ ，表明实际变动等于预期变动，调整在当期完全实现。若 $\delta = 0$ ，则

$Y_t = Y_{t-1}$ ，表明本期值与上期值一样，完全没有调整。一般情况下， $0 < \delta < 1$ 。

满足局部调整假设的模型 (7.22)，称为局部调整模型 (Partial adjustment model)。局部调整假设 (7.23) 式也可写成

$$Y_t = \delta Y_t^* + (1-\delta)Y_{t-1} \quad (7.24)$$

即被解释变量实际值是本期预期最佳值与前一期实际值的加权和，权数分别为 δ 和 $1-\delta$ 。

把（7.22）式代入（7.24）式，可得局部调整模型的转化形式

$$\begin{aligned} Y_t &= \delta(\alpha + \beta X_t + u_t) + (1 - \delta)Y_{t-1} \\ &= \delta\alpha + \delta\beta X_t + (1 - \delta)Y_{t-1} + \delta u_t \end{aligned} \quad (7.25)$$

令

$$\alpha^* = \delta\alpha, \quad \beta_0^* = \delta\beta, \quad \beta_1^* = 1 - \delta, \quad u_t^* = \delta u_t \quad (7.26)$$

则（9.25）式变为

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^* \quad (7.27)$$

这说明局部调整模型本质上是一个自回归模型。若能得到该模型的参数估计，代入（7.26）式就可求出原模型的参数估计。

从上述分析可以看出，库伊克模型、自适应预期模型与局部调整模型的最终形式，都是一阶自回归形式，这样，对这三类模型的估计就转化为对相应一阶自回归模型的估计。它们的区别在于两个方面：一是导出模型的经济背景与思想不同，库伊克模型是在无限分布滞后模型的基础上根据库伊克几何分布滞后假定而导出的；自适应预期模型是由解释变量的自适应过程而得到的；局部调整模型则是对被解释变量的局部调整而得到的。另一区别是，在这三个模型对应的自回归形式中，由于模型的形成机理不同而导致随机误差项的结构有所不同，这一区别将对模型的估计带来一定影响。

此外，有时需要将局部调整模型与自适应期望模型结合起来对某一经济问题进行研究，即建立局部调整—自适应期望综合模型。考虑如下模型：

$$Y_t^* = \alpha + \beta X_t^* + u_t$$

该模型反映了被解释变量的预期水平同解释变量预期值的关联性。对 Y_t^* 作局部调整假设，

对 X_t^* 作自适应假设下，局部调整—自适应期望综合模型可转化为如下形式的自回归模型

（读者不妨自己推导）：

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + \beta_2^* Y_{t-2} + u_t^* \quad (7.28)$$

第四节 自回归模型的估计

一、自回归模型估计的困难

上一节所讨论的库伊克模型、自适应预期模型与局部调整模型，模型结构上有一共性，

即最终都可表示为一阶自回归形式

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^* \quad (7.29)$$

因此，对这三个模型的估计就转化为对一阶自回归模型的估计。但是，上述一阶自回归模型的解释变量中含有滞后被解释变量 Y_{t-1} ， Y_{t-1} 是随机变量，它可能与随机扰动项相关；而且随机扰动项还可能自相关。也就是说，模型可能违背古典假定，从而给模型的估计带来一定困难。为了说明这一点，我们考察三个模型对应的一阶自回归模型中，随机扰动项的特征。

库伊克模型： $u_t^* = u_t - \lambda u_{t-1}$

自适应预期模型： $u_t^* = u_t - (1 - \gamma)u_{t-1}$

局部调整模型： $u_t^* = \delta u_t$

假定原模型中随机扰动项 u_t 满足古典假定，即

$$E(u_t) = 0$$

$$Var(u_t) = \sigma^2$$

$$Cov(u_i, u_j) = 0 \quad i \neq j$$

(1) 对于库伊克模型，有

$$\begin{aligned} Cov(u_t^*, u_{t-1}^*) &= E(u_t - \lambda u_{t-1} - E(u_t - \lambda u_{t-1}))(u_{t-1} - \lambda u_{t-2} - E(u_{t-1} - \lambda u_{t-2})) \\ &= E(u_t u_{t-1}) - \lambda E(u_{t-1}^2) - \lambda E(u_t u_{t-2}) + \lambda^2 E(u_{t-1} u_{t-2}) \\ &= -\lambda E(u_{t-1}^2) = -\lambda \sigma^2 \neq 0 \end{aligned}$$

$$\begin{aligned} Cov(Y_{t-1}, u_t^*) &= Cov(Y_{t-1}, u_t - \lambda u_{t-1}) \\ &= Cov(Y_{t-1}, u_t) - \lambda Cov(Y_{t-1}, u_{t-1}) \\ &= -\lambda Cov(Y_{t-1}, u_{t-1}) \neq 0 \end{aligned}$$

(2) 同理可证，自适应预期模型也有

$$Cov(u_t^*, u_{t-1}^*) \neq 0$$

$$Cov(Y_{t-1}, u_t^*) \neq 0$$

(1) 对于局部调整模型，有

$$Cov(u_t^*, u_{t-1}^*) = E(\delta u_t - E(\delta u_t))(\delta u_{t-1} - E(\delta u_{t-1})) = \delta^2 E(u_t u_{t-1}) = 0$$

$$Cov(Y_{t-1}, u_t^*) = Cov(Y_{t-1}, \delta u_t) = \delta Cov(Y_{t-1}, u_t) = 0$$

由此可见，对自回归模型的估计存在两个主要问题：一是出现了随机解释变量 Y_{t-1} ，而 Y_{t-1} 可能与 u_t 相关；二是随机扰动项 u_t 可能自相关，库伊克模型和自适应预期模型的随机扰动项都会导致自相关，只有局部调整模型的随机扰动项无自相关。如果用最小二乘法直接估计自回归模型，则估计可能是有偏的，而且不是一致估计。因此，估计自回归模型需要解决两个问题：一是设法消除 Y_{t-1} 与 u_t 的相关性；二是检验 u_t 是否存在自相关。

为了缓解解释变量 Y_{t-1} 与扰动项 u_t 存在相关带来的估计偏倚，可采用工具变量法；诊断一阶自回归模型扰动项是否存在自相关，可用德宾 h-检验法。而对于扰动项自相关的处理，问题比较复杂，涉及到动态回归模型的深入内容，在此从略。

二、工具变量法

所谓工具变量法，就是在进行参数估计的过程中选择适当的工具变量，代替回归模型中同随机扰动项存在相关性的解释变量。工具变量的选择应满足如下条件：

- (1) 与所代替的解释变量高度相关；
- (2) 与随机扰动项不相关；
- (3) 与其它解释变量不相关，以免出现多重共线性。

可以证明，利用工具变量法所得到的参数估计是一致估计。

在实际应用中，工具变量有多种选择方式，例如可选用 \hat{Y}_{t-1} 作工具变量，去代替滞后被解释变量 Y_{t-1} 进行估计，这样，一阶自回归模型就变为如下形式

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* \hat{Y}_{t-1} + u_t^* \quad (7.30)$$

其中 \hat{Y}_{t-1} 是 \hat{Y}_t 的滞后值。 \hat{Y}_t 是 Y 对 X 的滞后值的回归，即由如下回归方程得到

$$\hat{Y}_t = \hat{c}_0 + \hat{c}_1 X_{t-1} + \hat{c}_2 X_{t-2} + \cdots + \hat{c}_s X_{t-s} \quad (7.31)$$

滞后期 s 适当选取，一般取 2 或 3。由于 X_t 与 u_t^* 不相关， \hat{Y}_t 作为对 X 滞后值的回归，也与 u_t^* 不相关，进而 \hat{Y}_{t-1} 也与 u_t^* 不相关，因此，对模型 (7.30) 应用最小二乘法，可以得到参数的一致估计。

三、德宾 h-检验

关于随机扰动项是否存在自相关的诊断，前面我们曾介绍过 D—W 检验法，但这一检

验法不适合于方程含有滞后被解释变量的场合(见 D-W 检验的假设条件)。在自回归模型中, 滞后被解释变量是随机变量, 已有研究表明, 如果用 D—W 检验法, 则 d 统计量值总是趋近于 2。也就是说, 在一阶自回归中, 当随机扰动项存在自相关时, D—W 检验却倾向于得出非自相关的结论。为此, 德宾提出了检验一阶自相关的 h 统计量检验法。

h 统计量定义为

$$h = \hat{\rho} \sqrt{\frac{n}{1 - n\text{Var}(\hat{\beta}_1^*)}} = (1 - \frac{d}{2}) \sqrt{\frac{n}{1 - n\text{Var}(\hat{\beta}_1^*)}} \quad (7.32)$$

其中, $\hat{\rho}$ 为随机扰动项一阶自相关系数 ρ 的估计量, d 为 d 统计量, n 为样本容量, $\text{Var}(\hat{\beta}_1^*)$ 为滞后被解释变量 Y_{t-1} 的回归系数的估计方差。

德宾证明了在 $\rho = 0$ 的假定下, h 统计量的极限分布为标准正态分布。因此, 在大样本情况下, 可以用 h 统计量值判断随机扰动项是否存在一阶自相关。具体作法如下

(1) 对一阶自回归方程

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^*$$

直接进行最小二乘估计, 得到 $\text{Var}(\hat{\beta}_1^*)$ 及 d 统计量值。

(2) 将 $\text{Var}(\hat{\beta}_1^*)$ 、d 及样本容量 n 代入 (7.32) 式计算 h 统计量值。

(3) 给定显著性水平 α , 查标准正态分布表得临界值 h_α 。若 $|h| > h_\alpha$, 则拒绝原假设 $\rho = 0$, 说明自回归模型存在一阶自相关; 若 $|h| < h_\alpha$, 则接受原假设 $\rho = 0$, 说明自回归模型不存在一阶自相关。

例如, 假设对下列模型进行估计

$$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^*$$

得到 $d = 1.9$, $\text{Var}(\hat{\beta}_1^*) = 0.005$, 如果样本容量 $n=100$, 则有

$$\begin{aligned} h &= (1 - \frac{d}{2}) \sqrt{\frac{n}{1 - n\text{Var}(\hat{\beta}_1^*)}} \\ &= (1 - \frac{1}{2} \times 1.9) \sqrt{\frac{100}{1 - 100 \times 0.005}} \\ &= 0.7071 \end{aligned}$$

取显著性水平 $\alpha = 0.05$, 查标准正态分布表得临界值 $h_{\frac{\alpha}{2}} = 1.96$, 由于

$|h| = 0.7071 < h_{\frac{\alpha}{2}} = 1.96$ ，则接受原假设 $\rho = 0$ ，说明自回归模型不存在一阶自相关。

值得注意的是，该检验法可适用任意阶的自回归模型，对应的 h 统计量的计算式(7.32)仍然成立，即只用到 Y_{t-1} 回归系数的估计方差；此外，该检验法是针对大样本的，用于小样本效果较差。

第五节 案例分析

【案例 7.1】 为了研究 1955—1974 年期间美国制造业库存量 Y 和销售额 X 的关系，我们在例 7.3 中采用了经验加权法估计分布滞后模型。尽管经验加权法具有一些优点，但是设置权数的主观随意性较大，要求分析者对实际问题的特征有比较透彻的了解。下面用阿尔蒙法估计如下有限分布滞后模型：

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + u_t$$

将系数 $\beta_i (i=0,1,2,3)$ 用二次多项式近似，即

$$\beta_0 = \alpha_0$$

$$\beta_1 = \alpha_0 + \alpha_1 + \alpha_2$$

$$\beta_2 = \alpha_0 + 2\alpha_1 + 4\alpha_2$$

$$\beta_3 = \alpha_0 + 3\alpha_1 + 9\alpha_2$$

则原模型可变为

$$Y_t = \alpha + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + u_t$$

其中

$$Z_{0t} = X_t + X_{t-1} + X_{t-2} + X_{t-3}$$

$$Z_{1t} = X_{t-1} + 2X_{t-2} + 3X_{t-3}$$

$$Z_{2t} = X_{t-1} + 4X_{t-2} + 9X_{t-3}$$

在 Eviews 工作文件中输入 X 和 Y 的数据，在工作文件窗口中点击“Genr”工具栏，出现对话框，输入生成变量 Z_{0t} 的公式，点击“OK”；类似，可生成 Z_{1t} 、 Z_{2t} 变量的数据。进入 Equation Specification 对话框，键入回归方程形式

$$Y \ C \ Z0 \ Z1 \ Z2$$

点击“OK”，显示回归结果（见表 7.2）。

表 7.2

Dependent Variable: Y Method: Least Squares Date: 03/19/05 Time: 12:02 Sample(adjusted): 1958 1974 Included observations: 17 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-71.40754	19.92988	-3.582940	0.0033
Z0	0.661248	0.165480	3.995947	0.0015
Z1	0.902049	0.483131	1.867090	0.0846
Z2	-0.432155	0.166464	-2.596085	0.0222
R-squared	0.996797	Mean dependent var	818.6900	
Adjusted R-squared	0.996058	S.D. dependent var	279.9174	
S.E. of regression	17.57458	Akaike info criterion	8.773108	
Sum squared resid	4015.254	Schwarz criterion	8.969158	
Log likelihood	-70.57142	F-statistic	1348.639	
Durbin-Watson stat	1.848202	Prob(F-statistic)	0.000000	

表中 Z0、Z1、Z2 对应的系数分别为 α_0 、 α_1 、 α_2 的估计值 $\hat{\alpha}_0$ 、 $\hat{\alpha}_1$ 、 $\hat{\alpha}_2$ 。将它们代入分布滞后系数的阿尔蒙多项式中，可计算出 $\hat{\beta}_0$ 、 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 、 $\hat{\beta}_3$ 的估计值为：

$$\hat{\beta}_0 = \hat{\alpha}_0 = 0.661248$$

$$\hat{\beta}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 = 0.661248 + 0.902049 + (-0.432155) = 1.131142$$

$$\hat{\beta}_1 = \hat{\alpha}_0 + 2\hat{\alpha}_1 + 4\hat{\alpha}_2 = 0.661248 + 2 \times 0.902049 + 4 \times (-0.432155) = 0.736725$$

$$\hat{\beta}_1 = \hat{\alpha}_0 + 3\hat{\alpha}_1 + 9\hat{\alpha}_2 = 0.661248 + 3 \times 0.902049 + 9 \times (-0.432155) = -0.522$$

从而，分布滞后模型的最终估计式为：

$$Y_t = -6.419601 + 0.630281X_t + 1.15686X_{t-1} + 0.76178X_{t-2} - 0.55495X_{t-3}$$

在实际应用中，Eviews 提供了多项式分布滞后指令“PDL”用于估计分布滞后模型。下面结合本例给出操作过程：

在 Eviews 中输入 X 和 Y 的数据，进入 Equation Specification 对话框，键入方程形式

$$Y \quad C \quad PDL(X, 3, 2)$$

其中，“PDL 指令”表示进行多项式分布滞后（Polynomial Distributed Lags）模型的估计，括号中的 3 表示 X 的分布滞后长度，2 表示多项式的阶数。在 Estimation Settings 栏中选择 Least Squares(最小二乘法)，点击 OK，屏幕将显示回归分析结果（见表 7.3）。

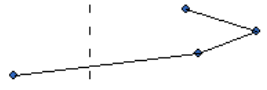
表 7.3

Dependent Variable: Y Method: Least Squares Date: 03/19/05 Time: 12:25 Sample(adjusted): 1958 1974				
---	--	--	--	--

Included observations: 17 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-71.40754	19.92988	-3.582940	0.0033
PDL01	1.131142	0.179991	6.284427	0.0000
PDL02	0.037739	0.162457	0.232299	0.8199
PDL03	-0.432155	0.166464	-2.596085	0.0222

R-squared	0.996797	Mean dependent var	818.6900
Adjusted R-squared	0.996058	S.D. dependent var	279.9174
S.E. of regression	17.57458	Akaike info criterion	8.773108
Sum squared resid	4015.254	Schwarz criterion	8.969158
Log likelihood	-70.57142	F-statistic	1348.639
Durbin-Watson stat	1.848202	Prob(F-statistic)	0.000000

Lag Distribution of X	i	Coefficient	Std. Error	T-Statistic
	0	0.66125	0.16548	3.99595
	1	1.13114	0.17999	6.28443
	2	0.73673	0.16428	4.48462
	3	-0.52200	0.23481	-2.22312
Sum of Lags		2.00711	0.06330	31.7065

需要指出的是，用“PDL”估计分布滞后模型时，Eviews所采用的滞后系数多项式变换不是形如（7.4）式的阿尔蒙多项式，而是阿尔蒙多项式的派生形式。因此，输出结果中 PDL01、PDL02、PDL03 对应的估计系数不是阿尔蒙多项式系数 α_0 、 α_1 、 α_2 的估计。但同前面分步计算的结果相比，最终的分布滞后估计系数 $\hat{\beta}_0$ 、 $\hat{\beta}_1$ 、 $\hat{\beta}_2$ 、 $\hat{\beta}_3$ 是相同的。

【案例 7.2】 货币主义学派认为，产生通货膨胀的必要条件是货币的超量供应。物价变动与货币供应量的变化有着较为密切的联系，但是二者之间的关系不是瞬时的，货币供应量的变化对物价的影响存在一定时滞。有研究表明，西方国家的通货膨胀时滞大约为 2—3 个季度。

在中国，大家普遍认同货币供给的变化对物价具有滞后影响，但滞后期究竟有多长，还存在不同的认识。下面采集 1996—2005 年全国广义货币供应量和物价指数的月度数据（见表 7.4）对这一问题进行研究。

表 7.4 1996—2005 年全国广义货币供应量及物价指数月度数据

月度	广义货币M2 (千亿元)	广义货币 增长量M2z (千亿元)	居民消费价 格同比指数 tbzs	月度	广义货币M2 (千亿元)	广义货币增 长量M2z (千亿元)	居民消费价 格同比指数 tbzs
Jan-96	58.401			Oct-00	129.522	-0.9518	100
Feb-96	63.778	5.377	109.3	Nov-00	130.9941	1.4721	101.3
Mar-96	64.511	0.733	109.8	Dec-00	134.6103	3.6162	101.5
Apr-96	65.723	1.212	109.7	Jan-01	137.5436	2.9333	101.2
May-96	66.88	1.157	108.9	Feb-01	136.2102	-1.3334	100

Jun-96	68.132	1.252	108.6		Mar-01	138.7445	2.5343	100.8
Jul-96	69.346	1.214	108.3		Apr-01	139.9499	1.2054	101.6
Aug-96	72.309	2.963	108.1		May-01	139.0158	-0.9341	101.7
Sep-96	69.643	-2.666	107.4		Jun-01	147.8097	8.7939	101.4
Oct-96	73.1522	3.5092	107		Jul-01	149.2287	1.419	101.5
Nov-96	74.142	0.9898	106.9		Aug-01	149.9418	0.7131	101
Dec-96	76.0949	1.9529	107		Sep-01	151.8226	1.8808	99.9
Jan-97	78.648	2.5531	105.9		Oct-01	151.4973	-0.3253	100.2
Feb-97	78.998	0.35	105.6		Nov-01	154.0883	2.591	99.7
Mar-97	79.889	0.891	104		Dec-01	158.3019	4.2136	99.7
Apr-97	80.818	0.929	103.2		Jan-02	159.6393	1.3374	99
May-97	81.151	0.333	102.8		Feb-02	160.9356	1.2963	100
Jun-97	82.789	1.638	102.8		Mar-02	164.0646	3.129	99.2
Jul-97	83.46	0.671	102.7		Apr-02	164.5706	0.506	98.7
Aug-97	84.746	1.286	101.9		May-02	166.061	1.4904	98.9
Sep-97	85.892	1.146	101.8		Jun-02	169.6012	3.5402	99.2
Oct-97	86.644	0.752	101.5		Jul-02	170.8511	1.2499	99.1
Nov-97	87.59	0.946	101.1		Aug-02	173.2509	2.3998	99.3
Dec-97	90.9953	3.4053	100.4		Sep-02	176.9824	3.7315	99.3
Jan-98	92.2114	1.2161	100.3		Oct-02	177.2942	0.3118	99.2
Feb-98	92.024	-0.1874	99.9		Nov-02	179.7363	2.4421	99.3
Mar-98	92.015	-0.009	100.7		Dec-02	185.0073	5.271	99.6
Apr-98	92.662	0.647	99.7		Jan-03	190.4883	5.481	100.4
May-98	93.936	1.274	99		Feb-03	190.1084	-0.3799	100.2
Jun-98	94.658	0.722	98.7		Mar-03	194.4873	4.3789	100.9
Jul-98	96.314	1.656	98.6		Apr-03	196.1301	1.6428	101
Aug-98	97.299	0.985	98.6		May-03	199.5052	3.3751	100.7
Sep-98	99.795	2.496	98.5		Jun-03	204.9314	5.4262	100.3
Oct-98	100.8752	1.0802	98.9		Jul-03	206.1931	1.2617	100.5
Nov-98	102.229	1.3538	98.8		Aug-03	210.5919	4.3988	100.9
Dec-98	104.4985	2.2695	99		Sep-03	213.5671	2.9752	101.1
Jan-99	105.5	1.0015	98.8		Oct-03	214.4694	0.9023	101.8
Feb-99	107.778	2.278	98.7		Nov-03	216.3517	1.8823	103
Mar-99	108.438	0.66	98.2		Dec-03	221.2228	4.8711	103.2
Apr-99	109.218	0.78	97.8		Jan-04	225.10193	3.87913	103.2
May-99	110.061	0.843	97.8		Feb-04	227.05072	1.94879	102.1
Jun-99	111.363	1.302	97.9		Mar-04	231.6546	4.60388	103
Jul-99	111.414	0.051	98.6		Apr-04	233.62786	1.97326	103.8

Aug-99	112.827	1.413	98.7		May-04	234.8424	1.21454	104.4
Sep-99	115.079	2.252	99.2		Jun-04	238.42749	3.58509	105
Oct-99	115.39	0.311	99.4		Jul-04	234.8424	-3.58509	105.3
Nov-99	116.559	1.169	99.1		Aug-04	239.72919	4.88679	105.3
Dec-99	119.898	3.339	99		Sep-04	243.757	4.02781	105.2
Jan-00	121.22	1.322	99.8		Oct-04	243.74	-0.017	104.3
Feb-00	121.5834	0.3634	100.7		Nov-04	247.13558	3.39558	102.8
Mar-00	122.5807	0.9973	99.8		Dec-04	253.2077	6.07212	102.4
Apr-00	124.1219	1.5412	99.7		Jan-05	257.75283	4.54513	101.9
May-00	124.0533	-0.0686	100.1		Feb-05	259.3561	1.60327	103.9
Jun-00	126.6053	2.552	100.5		Mar-05	264.5889	5.2328	102.7
Jul-00	126.3239	-0.2814	100.5		Apr-05	266.99266	2.40376	101.8
Aug-00	127.79	1.4661	100.3		May-05	269.2294	2.23674	101.8
Sep-00	130.4738	2.6838	100					

数据来源：中国经济统计数据库，<http://db.cei.gov.cn/>。

为了考察货币供应量的变化对物价的影响，我们用广义货币 M2 的月增长量 M2Z 作为解释变量，以居民消费价格月度同比指数 TBZS 为被解释变量进行研究。首先估计如下回归模型

$$TBZS_t = \alpha + \beta_0 M2Z_t + u_t$$

得如下回归结果（表 7.5）。

表7.5

Dependent Variable: TBZS				
Method: Least Squares				
Date: 07/03/05 Time: 17:10				
Sample(adjusted): 1996:02 2005:05				
Included observations: 112 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.4356	0.397419	255.2358	0.0000
M2Z	0.068371	0.151872	0.450190	0.6535
R-squared	0.001839	Mean dependent var	101.5643	
Adjusted R-squared	-0.007235	S.D. dependent var	2.911111	
S.E. of regression	2.921623	Akaike info criterion	4.999852	
Sum squared resid	938.9472	Schwarz criterion	5.048396	
Log likelihood	-277.9917	F-statistic	0.202671	
Durbin-Watson stat	0.047702	Prob(F-statistic)	0.653460	

从回归结果来看，M2Z 的 t 统计量值不显著，表明当期货币供应量的变化对当期物价水

平的影响在统计意义上不明显。为了分析货币供应量变化影响物价的滞后性，我们做滞后 6 个月的分布滞后模型的估计，在 Eviews 工作文档的方程设定窗口中，输入

TBZS C M2Z M2Z(-1) M2Z(-2) M2Z(-3) M2Z(-4) M2Z(-5) M2Z(-6)

结果见表 7.6。

表 7.6

Dependent Variable: TBZS				
Method: Least Squares				
Date: 07/03/05 Time: 17:09				
Sample(adjusted): 1996:08 2005:05				
Included observations: 106 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	100.0492	0.584318	171.2240	0.0000
M2Z	-0.011037	0.140613	-0.078493	0.9376
M2Z(-1)	0.016169	0.137998	0.117166	0.9070
M2Z(-2)	0.053044	0.136808	0.387723	0.6991
M2Z(-3)	0.028679	0.143155	0.200333	0.8416
M2Z(-4)	0.130825	0.139183	0.939951	0.3496
M2Z(-5)	0.137794	0.142502	0.966965	0.3359
M2Z(-6)	0.248778	0.143394	1.734924	0.0859
R-squared	0.055557	Mean dependent var	101.1377	
Adjusted R-squared	-0.011904	S.D. dependent var	2.347946	
S.E. of regression	2.361879	Akaike info criterion	4.629264	
Sum squared resid	546.6902	Schwarz criterion	4.830278	
Log likelihood	-237.3510	F-statistic	0.823546	
Durbin-Watson stat	0.094549	Prob(F-statistic)	0.570083	

从回归结果来看，M2Z 各滞后期的系数逐步增加，表明当期货币供应量的变化对物价水平的影响要经过一段时间才能逐步显现。但各滞后期的系数的 t 统计量值不显著，因此还不能据此判断滞后期究竟有多长。为此，我们做滞后 12 个月的分布滞后模型的估计，结果见表 7.7。

表 7.7

Dependent Variable: TBZS				
Method: Least Squares				
Date: 07/03/05 Time: 17:09				
Sample(adjusted): 1997:02 2005:05				
Included observations: 100 after adjusting endpoints				

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	98.35668	0.467897	210.2102	0.0000
M2Z	-0.167665	0.121743	-1.377203	0.1720
M2Z(-1)	-0.032065	0.111691	-0.287084	0.7747
M2Z(-2)	-0.000995	0.111464	-0.008925	0.9929
M2Z(-3)	0.004243	0.113815	0.037276	0.9704
M2Z(-4)	0.106581	0.112727	0.945480	0.3471
M2Z(-5)	0.043217	0.113161	0.381908	0.7035
M2Z(-6)	0.117581	0.118460	0.992575	0.3237
M2Z(-7)	0.140418	0.115571	1.214988	0.2277
M2Z(-8)	0.220875	0.114368	1.931271	0.0567
M2Z(-9)	0.140875	0.115354	1.221247	0.2253
M2Z(-10)	0.180497	0.115895	1.557410	0.1230
M2Z(-11)	0.246911	0.125543	1.966752	0.0524
M2Z(-12)	0.392359	0.130058	3.016798	0.0034
R-squared	0.317136	Mean dependent var	100.7830	
Adjusted R-squared	0.213913	S.D. dependent var	1.890863	
S.E. of regression	1.676469	Akaike info criterion	4.000434	
Sum squared resid	241.7072	Schwarz criterion	4.365158	
Log likelihood	-186.0217	F-statistic	3.072325	
Durbin-Watson stat	0.265335	Prob(F-statistic)	0.000906	

表 7.7 显示，从 M2Z 到 M2Z(-11)，回归系数都不显著异于零，而 M2Z (-12) 的回归系数 t 统计量值为 3.016798，在 5% 显著性水平下拒绝系数为零的原假设。这一结果表明，当期货币供应量变化对物价水平的影响在经过 12 个月（即一年）后明显地显现出来。为了考察货币供应量变化对物价水平影响的持续期，我们做滞后 18 个月的分布滞后模型的估计，结果见表 7.8。

表 7.8

Dependent Variable: TBZS				
Method: Least Squares				
Date: 07/03/05 Time: 17:08				
Sample(adjusted): 1997:08 2005:05				
Included observations: 94 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	97.41411	0.370000	263.2815	0.0000
M2Z	-0.083649	0.094529	-0.884900	0.3791
M2Z(-1)	-0.116744	0.093984	-1.242161	0.2181
M2Z(-2)	-0.119939	0.094428	-1.270156	0.2080

M2Z(-3)	-0.092993	0.095720	-0.971509	0.3345
M2Z(-4)	-0.032912	0.095823	-0.343468	0.7322
M2Z(-5)	-0.023891	0.097813	-0.244256	0.8077
M2Z(-6)	0.017290	0.100645	0.171794	0.8641
M2Z(-7)	0.028288	0.097570	0.289929	0.7727
M2Z(-8)	0.048708	0.095877	0.508021	0.6129
M2Z(-9)	0.025995	0.097569	0.266422	0.7907
M2Z(-10)	0.118247	0.096764	1.222011	0.2256
M2Z(-11)	0.157408	0.102558	1.534815	0.1291
M2Z(-12)	0.271281	0.112316	2.415326	0.0182
M2Z(-13)	0.325760	0.109217	2.982684	0.0039
M2Z(-14)	0.396242	0.107046	3.701601	0.0004
M2Z(-15)	0.335482	0.106776	3.141941	0.0024
M2Z(-16)	0.270811	0.107222	2.525697	0.0137
M2Z(-17)	0.200024	0.109278	1.830415	0.0712
M2Z(-18)	0.169696	0.101547	1.671114	0.0989
R-squared	0.610520	Mean dependent var	100.6085	
Adjusted R-squared	0.510519	S.D. dependent var	1.795733	
S.E. of regression	1.256348	Akaike info criterion	3.480597	
Sum squared resid	116.8024	Schwarz criterion	4.021724	
Log likelihood	-143.5881	F-statistic	6.105105	
Durbin-Watson stat	0.308938	Prob(F-statistic)	0.000000	

结果表明，从滞后 12 个月开始 t 统计量值显著，一直到滞后 16 个月为止，从滞后第 17 个月开始 t 值变得不显著；再从回归系数来看，从滞后 11 个月开始，货币供应量变化对物价水平的影响明显增加，再滞后 14 个月时达到最大，然后逐步下降。

通过上述一系列分析，我们可以做出这样的判断：在我国，货币供应量变化对物价水平的影响具有明显的滞后性，滞后期大约为一年，而且滞后影响具有持续性，持续的长度大约为半年，其影响力度先递增然后递减，滞后结构为 Λ 型。

当然，从上述回归结果也可以看出，回归方程的 R^2 不高，DW 值也偏低，表明除了货币供应量外，还有其他因素影响物价变化；同时，过多的滞后变量也可能引起多重共线性问题。如果我们分析的重点是货币供应量变化对物价影响的滞后性，上述结果已能说明问题。如果要提高模型的预测精度，则可以考虑对模型进行改进。根据前面的分析可知，分布滞后模型可以用子回归模型来代替，因此我们估计如下子自回归模型：

$$TBZS_t = \alpha + \beta TBZS_{t-1} + u_t$$

在 Eviews 工作文档的方程设定窗口中，输入

$$TBZS = C + TBZS(-1)$$

估计结果见表 7.9。

表 7.9

Dependent Variable: TBZS				
Method: Least Squares				
Date: 07/10/05 Time: 23:48				
Sample(adjusted): 1996:03 2005:05				
Included observations: 111 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.348792	1.938684	2.758982	0.0068
TBZS(-1)	0.946670	0.019081	49.61371	0.0000
R-squared	0.957596	Mean dependent var	101.4946	
Adjusted R-squared	0.957207	S.D. dependent var	2.828904	
S.E. of regression	0.585200	Akaike info criterion	1.784126	
Sum squared resid	37.32798	Schwarz criterion	1.832947	
Log likelihood	-97.01900	F-statistic	2461.520	
Durbin-Watson stat	1.779257	Prob(F-statistic)	0.000000	

第七章小结

- 1、由于心理、技术以及制度等原因，经济变量之间的影响往往具有滞后效应，滞后变量模型在经济分析中具有重要作用。分布滞后模型和自回归模型是两种常见的滞后变量模型。
- 2、分布滞后模型不能直接运用 OLS 方法进行估计，原因在于自由度损失、多重共线性和之后长度难于确定；克服这些困难的方法是采用变通估计方法，变通的估计方法有经验加权法、阿尔蒙法及库依克法。
- 3、自回归模型的产生背景主要在于两个方面：一是无限分布滞后模型不能直接估计，为了估计模型而对滞后结构作出某种假定（如库依克假定），然后通过变换形成自回归模型；二是在模型中引入了预期因素，由于变量的预期值无法观测，因此对“期望模型”中预期的形成作出某种假定，最后变换成自回归模型，例如自适应预期模型、局部调整模型。
- 4、库依克模型、自适应预期模型与局部调整模型的最终形式为自回归结构。在这三个模型中，只有局部调整模型满足扰动项无自相关、与解释变量 X_t 及 Y_{t-1} 不相关的古典假定，

从而可使用最小二乘法直接进行估计；而库伊克模型与自适应预期模型不满足古典假定，如果用最小二乘法直接进行估计，则估计是有偏的，且不是一致估计。

- 5、为了缓解扰动项与解释变量 Y_{t-1} 存在相关带来估计偏倚，克采用工具变量法；诊断一阶自回归模型扰动项是否存在自相关克采用德宾 h-检验法。

第七章主要公式表

滞后变量模型	一般形式	$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_s X_{t-s} + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \cdots + \gamma_q Y_{t-q} + u_t$
	分布滞后模型	$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_s X_{t-s} + u_t$
	自回归模型	$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \cdots + \gamma_q Y_{t-q} + u_t$
分布滞后模型的阿尔蒙估计法	基本模型	$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_s X_{t-s} + u_t$
	阿尔蒙变换	$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \cdots + \alpha_m i^m \quad i = 0, 1, 2, \dots, s; m < s$
	新模型	$Y_t = \alpha + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \cdots + \alpha_m Z_{mt} + u_t$ $Z_{it} = X_{t-1} + 2^i X_{t-2} + 3^i X_{t-3} + \cdots + s^i X_{t-s}$
库伊克模型	基本模型	$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + u_t$
	库伊克假定	$\beta_i = \beta_0 \lambda^i, \quad 0 < \lambda < 1, \quad i = 0, 1, 2, \dots$
	新模型	$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^*$ $\alpha^* = (1-\lambda)\alpha, \quad \beta_0^* = \beta_0, \quad \beta_1^* = \lambda, \quad u_t^* = u_t - \lambda u_{t-1}$
自适应预期模型	基本模型	$Y_t = \alpha + \beta X_t^* + u_t$
	自适应预期假定	$X_t^* = X_{t-1}^* + \gamma(X_t - X_{t-1}^*)$
	新模型	$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^*$ $\alpha^* = \gamma\alpha, \quad \beta_0^* = \gamma\beta, \quad \beta_1^* = 1-\gamma, \quad u_t^* = u_t - (1-\gamma)u_{t-1}$
局部调	基本模型	$Y_t^* = \alpha + \beta X_t + u_t$

整模型	局部调整假定	$Y_t - Y_{t-1} = \delta(Y_t^* - Y_{t-1})$
	新模型	$Y_t = \alpha^* + \beta_0^* X_t + \beta_1^* Y_{t-1} + u_t^*$ $\alpha^* = \delta\alpha, \quad \beta_0^* = \delta\beta, \quad \beta_1^* = 1 - \delta, \quad u_t^* = \delta u_t$
自回归模型自相关检验	德宾 h-检验 (h 统计量)	$h = \hat{\rho} \sqrt{\frac{n}{1 - n\text{Var}(\hat{\beta}_1^*)}} = (1 - \frac{d}{2}) \sqrt{\frac{n}{1 - n\text{Var}(\hat{\beta}_1^*)}}$

思考题与练习题

思考题

7.1 什么是滞后现象？产生滞后现象的原因主要有哪些？

7.2 对分布滞后模型进行估计存在哪些困难？实际应用中如何处理这些困难？

7.3 库伊克模型、自适应预期模型与局部调整模型有哪些共性和不同之处？模型估计会存在哪些困难？如何解决？

7.4 考虑如下模型：

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 Y_{t-1} + u_t$$

假定 Y_{t-1} 和 u_t 相关。为了消除相关，采用如下工具变量法：先求 Y_t 对 X_{1t} 和 X_{2t} 的回归，得到 Y_t 的估计值 \hat{Y}_t ，然后做如下回归：

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 \hat{Y}_{t-1} + u_t$$

其中 \hat{Y}_{t-1} 是第一步粗估计值 \hat{Y}_t 的滞后值。分析说明该方法为什么可以消除原模型中 \hat{Y}_{t-1} 和 u_t 之间的相关性。

7.5 检验一阶自回归模型随机扰动项是否存在自相关，为什么用德宾 h-检验而不用 D—W 检验？

练习题

7.1 表中给出了 1970~1987 年期间美国的个人消息支出(PCE)和个人可支配收入(PDI)数据，所有数字的单位都是 10 亿美元(1982 年的美元价)。

年份	PCE	PDI	年份	PCE	PDI	年份	PCE	PDI
1970	1492.0	1668.1	1976	1803.9	2001.0	1982	2050.7	2261.5
1971	1538.8	1728.4	1977	1883.8	2066.6	1983	2146.0	2331.9
1972	1961.9	1797.4	1978	1961.0	2167.4	1984	2249.3	2469.8
1973	1689.6	1916.3	1979	2004.4	2212.6	1985	2354.8	2542.8
1974	1674.0	1896.6	1980	2000.4	2214.3	1986	2455.2	2640.9
1975	1711.9	1931.7	1981	2042.2	2248.6	1987	2521.0	2686.3

估计下列模型：

$$PCE_t = A_1 + A_2 PDI_t + \mu_t$$

$$PCE_t = B_1 + B_2 PDI_t + B_3 PCE_{t-1} + v_t$$

- (1) 解释这两个回归模型的结果。
- (2) 短期和长期边际消费倾向（MPC）是多少？

7.2 表中给出了某地区 1980-2001 年固定资产投资 Y 与销售额 X 的资料（单位：亿元）。

年份	Y	X	年份	Y	X
1980	36.99	52.805	1991	128.68	168.129
1981	33.60	55.906	1992	123.97	163.351
1982	35.42	63.027	1993	117.35	172.547
1983	42.35	72.931	1994	139.61	190.682
1984	52.48	84.790	1995	152.88	194.538
1985	53.66	86.589	1996	137.95	194.657
1986	58.53	98.797	1997	141.06	206.326
1987	67.48	113.201	1998	163.45	223.541
1988	78.13	126.905	1999	183.80	232.724
1989	95.13	143.936	2000	192.61	239.459
1990	112.60	154.391	2001	182.81	235.142

试就下列模型，按照一定的处理方法估计模型参数，并解释模型的经济意义，探测模型扰动项的一阶自相关性。

- (1) 设定模型

$$Y_t^* = \alpha + \beta X_t + u_t$$

运用局部调整假定。

- (2) 设定模型

$$Y_t^* = \alpha X_t^\beta e^{u_t}$$

运用局部调整假定。

- (3) 设定模型

$$Y_t = \alpha + \beta X_t^* + u_t$$

运用自适应预期假定。

- (4) 运用阿尔蒙多项式变换法，估计分布滞后模型：

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_4 X_{t-4} + u_t$$

7.3 表中给出了某地区 1962-1995 年基本建设新增固定资产 Y (亿元) 和全省工业总产值 X (亿元) 按当年价格计算的历史资料。

年份	Y	X	年份	Y	X
1962	0.94	4.95	1979	2.06	42.69
1963	1.69	6.63	1980	7.93	51.61
1964	1.78	8.51	1981	8.01	61.5
1965	1.84	9.37	1982	6.64	60.73
1966	4.36	11.23	1983	16	64.64
1967	7.02	11.34	1984	8.81	66.67
1968	5.55	19.9	1985	10.38	73.78
1969	6.93	29.49	1986	6.2	69.52
1970	7.17	36.83	1987	7.97	79.64
1971	2.33	21.19	1988	27.33	92.45
1972	2.18	18.14	1989	12.58	102.94
1973	2.39	19.69	1990	12.47	105.62
1974	3.3	23.88	1991	10.88	104.88
1975	5.24	29.65	1992	17.7	113.3
1976	5.39	40.94	1993	14.72	127.13
1977	1.78	33.08	1994	13.76	141.44
1978	0.73	20.3	1995	14.42	173.75

- (1) 设定模型 $Y_t^* = \alpha + \beta X_t + \mu_t$ 作部分调整假定，估计参数，并作解释。
- (2) 设定模型 $Y_t = \alpha + \beta X_t^* + \mu_t$ 作自适应假定，估计参数，并作解释。
- (3) 比较上述两种模型的设定，哪一个模型拟合较好？

7.4 给出某地区各年末货币流通量 Y，社会商品零售额 X1、城乡居民储蓄余额 X2 的数据

单位：亿元

年份	Y	X1	X2	年份	Y	X1	X2
1953	10518	78676	4163	1970	38500	240332	26156
1954	14088	101433	4888	1971	47100	274534	30944
1955	13375	103989	5689	1972	57200	299197	35961
1956	18354	124525	7406	1973	60000	314006	39667
1957	16867	126467	9156	1974	62500	318954	43320
1958	18515	134446	10193	1975	64500	336015	46184
1959	22558	154961	13939	1976	68000	352924	48311
1960	29036	170370	15495	1977	63000	378115	53313
1961	41472	149182	12553	1978	66000	415830	61290
1962	34826	154564	10080	1979	76000	452032	70033
1963	30000	142548	11602	1980	85000	512543	92800
1964	24300	143415	15031	1981	90000	547956	109707
1965	29300	156998	17108	1982	101000	591088	133799
1966	33900	176387	19301	1983	100000	646427	164314
1967	36100	178162	20485	1984	160000	733162	201199
1968	39600	167074	22572	1985	192000	919045	277185

利用表中数据设定模型： $Y_t^* = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \mu_t$

$$Y_t^* = \alpha X_{1t}^{\beta_1} X_{2t}^{\beta_2} e^{u_t}$$

其中 Y_t^* 为长期(或所需求的)货币流通量。试根据总价调整假设，作模型变换，估计并检验参数，对参数经济意义作出解释，求出短期和长期货币流通需求同和需求弹性。

7.5 设 $M_t = \alpha + \beta_1 Y_t^* + \beta_2 R_t^* + \mu_t$

其中： M 为实际货币流通量， Y^* 为期望社会商品零售总额， R^* 为期望储蓄总额，对于期望值作如下假定：

$$Y_t^* = \gamma_1 Y_t + (1 - \gamma_1) Y_{t-1}^*$$

$$R_t = \gamma_2 R_t + (1 - \gamma_2) R_{t-1}^*$$

其中 γ_1, γ_2 为期望系数，均为小于 1 的正数。

- (1) 如何利用可观测的量来表示 M_t ?
- (2) 分析这样变换存在什么问题?
- (3) 利用 7.4 题的数据进行回归，估计模型，并作检验。

7.6 考虑如下回归模型：

$$\hat{y}_t = -3012 + 0.1408x_t + 0.2306x_{t-1}$$

$$t = (-6.27) \quad (2.6) \quad (4.26)$$

$$R^2 = 0.727$$

其中 y =通货膨胀率， x =生产设备使用率。

- (1) 生产设备使用率对通货膨胀率的短期影响和长期影响分别是多大？
- (2) 如果你手中无原始数据，并让你估计下列回归模型 $y_t = b_1 + b_2x_t + b_3y_{t-1} + \mu_t$ ，你怎样估计生产设备使用率对通货膨胀率的短期影响和长期影响。

7.7 表中给出了某地区消费总额 Y （亿元）和货币收入总额 X （亿元）的年度资料，

年份	X	Y	年份	X	Y
1975	103.169	91.158	1990	215.539	204.75
1976	115.07	109.1	1991	220.391	218.666
1977	132.21	119.187	1992	235.483	227.425
1978	156.574	143.908	1993	280.975	229.86
1979	166.091	155.192	1994	292.339	244.23
1980	155.099	148.673	1995	278.116	258.363
1981	138.175	151.288	1996	292.654	275.248
1982	146.936	148.1	1997	341.442	299.277
1983	157.7	156.777	1998	401.141	345.47
1984	179.797	168.475	1999	458.567	406.119
1985	195.779	174.737	2000	500.915	462.223
1986	194.858	182.802	2001	450.939	492.662
1987	189.179	180.13	2002	626.709	539.046
1988	199.963	190.444	2003	783.953	617.568
1989	205.717	196.9	2004	890.637	727.397

分析该地区消费同收入的关系

- (1) 做 Y_t 关于 X_t 的回归，对回归结果进行分析判断；
- (2) 建立分布滞后模型，用库伊克变换转换为库伊克模型后进行估计，并对估计结果进行分析判断；
- (3) 建立局部调整——自适应期望综合模型进行分析。

第八章 虚拟变量回归

引子

男女大学生的消费真的有差异吗?

在校大学生的消费行为越来越受到社会的关注,学生家长也很关心自己的子女上大学究竟要准备多少花费。由共青团中央、全国学联共同发布的《2004 中国大学生消费与生活形态研究报告》显示,当代大学生在消费结构方面呈现出多元化趋势。大学生除了日常生活费开支以外,还有人际交往消费、网络通讯消费、书报消费、衣着类消费、化妆品类消费、电脑类消费、旅游类消费、食品类消费、学习用品类消费、各种考证类等消费。大学生时尚化、个性化消费增多已成为趋势与潮流。不同性别大学生的消费结构有所不同,专科生、本科生、研究生的消费结构更有差异。有的记者调查发现,不同年级之间,男女同学之间,消费水平、消费结构、消费方式上都存在着差异。年级越高,消费水平也随之增长,随着阅历的增加,对自己形象的重视,精神享受的追求、学习的投入、配备手机电脑的需求也随之增长。同年级的男生的消费高于女生,虽然女生在化妆品、衣服饰品方面的投入明显高于男生。然而时代在变,对美的追求已不再限于女生,男生对于个人形象、装扮也已慢慢重视起来。此外男生在人际交往方面比女生投入了更多的"本钱"。请客吃饭、朋友聚会、节日送礼已不再罕见。所谓的"人情消费"已从社会向校园中扩张蔓延,而在乎"面子"的男同胞已成为追随这一潮流的"先驱"。高年级女生对于吃饭的投入相对较少,而在化妆品、服饰、零食方面的投入却增长不少。(注:来源于 Solic 教育网、网易教育频道、新华网等)

为了研究男女大学生、不同层次大学生、不同年级大学生的消费结构是否有差异,需要将这些定性的因素引入计量模型,怎样才能模型中有效地表示这些定性因素的作用呢?

第一节 虚拟变量

一、虚拟变量的基本概念

在前面的分析中,被解释变量主要受到一些可以直接度量的变量影响,如收入、产出、商品需求量、价格、成本、资金、人数等。但现实经济生活中,影响被解释变量变动的因素,除了这些可以直接获得实际观测数据的定量变量外,还包括一些本质上为定性因素(或称属性因素)的影响,例如性别、种族、肤色、职业、季节、文化程度、战争、自然灾害、政府

经济政策的变动等因素。在实际经济分析中，这些定性变量有时具有不可忽视的重要影响。例如，研究某个企业的销售水平，产业部门（制造业、零售业）、所有制（私营、非私营）、地理位置（东、中、西部）、管理者素质的高低等是值得经常考虑的影响因素，这些因素有共同的特征，即都是表示某种属性的，不能直接用数据精确描述的因素。因此，被解释变量的变动经常是定量因素和属性因素共同作用的结果。在计量经济模型中，应当同时包含定量和属性两种因素对被解释变量的影响作用。

定量因素是指那些可直接测度的数值型因素，如 GDP、 M_2 等。定性因素，或称为属性因素，是不能直接测度的、说明某种属性或状态存在与否的非数值型因素，如男性或女性、城市居民或非城市居民、气候条件正常或异常、政府经济政策不变与改革等。在计量经济学的建模中应当将定量因素和定性因素同时纳入模型之内。

为了在模型中反映定性因素，可以将定性因素转化为虚拟变量去表现。虚拟变量（或称为属性变量、双值变量、类型变量、定性变量、二元型变量等），是人工构造的取值为 0 和 1 的作为属性变量代表的变量，一般用字母 D（或 DUM，英文 dummy 的缩写）表示。属性因素通常具有若干类型或水平，通常虚拟变量的取值为 0 和 1，当虚拟变量取值为 0，即 $D=0$ 时，表示某种属性或状态不出现或不存在，即不是某种类型；当虚拟变量取值为 1，即 $D=1$ 时，表示某种属性或状态出现或存在，即是某种类型。例如，构造政府经济政策人工变量，当经济政策不变时，虚拟变量取值为 0，当经济政策改变时，虚拟变量取值为 1。这种做法实际上是一种变换或映射，将不能精确计量的定性因素的水平或状态变换为用 0 和 1 来定量描述。

二、虚拟变量的设置规则

在计量经济学模型中引入虚拟变量，可以使我们同时兼顾定量因素和定性因素的影响和作用。但是，在设置虚拟变量时应遵循一定的规则。

1、虚拟变量数量的设置规则

虚拟变量个数的设置规则是：若定性因素有 m 个相互排斥的类型（或属性、水平），在有截距项的模型中只能引入 $m-1$ 个虚拟变量，否则会陷入所谓“虚拟变量陷阱”，产生完全的多重共线性。在无截距项的模型中，定性因素有 m 个相互排斥的类型时，引入 m 个虚拟变量不会导致完全多重共线性，不过这时虚拟变量参数的估计结果，实际上是 $D=1$ 时的样本均值。

例如，城镇居民和农村居民住房消费支出的模型可设定为：

$$C_i = \alpha_1 + \beta Y_i + \alpha_2 D_i + u_i \quad (8.1)$$

其中， C_i 为居民的住房消费支出， Y_i 为居民的可支配收入， D_i 为虚拟变量，

$D_i = \begin{cases} 1 & \text{城镇居民, 即当 } D_i = 1 \text{ 时为城镇居民; 当 } D_i = 0 \text{ 时为其他 (农村居民)。} \\ 0 & \text{其他} \end{cases}$ 这里区分城

镇居民和农村居民的定性变量的类型有 $m=2$ 个，按虚拟变量的设置规则应引入 $m-1=2-1=1$ 个虚拟变量。

但是，如果引入了 $m=2$ 个虚拟变量： $D_{2i} = \begin{cases} 1 & \text{城镇居民} \\ 0 & \text{其他} \end{cases}$ ， $D_{3i} = \begin{cases} 1 & \text{农村居民} \\ 0 & \text{其他} \end{cases}$ ，

则有：

$$C_i = \alpha_1 + \beta Y_i + \alpha_2 D_{2i} + \alpha_3 D_{3i} + u_i \quad (8.2)$$

这时，当 $D_{2i}=1$ 时同时有 $D_{3i}=0$ ；反之，当 $D_{2i}=0$ 时有 $D_{3i}=1$ 。即对于任何被调查的居民家庭都有 $D_{2i} + D_{3i}=1$ ， D_2 和 D_3 存在完全的共线性，无法利用 OLS 估计其参数，从而陷入“虚拟变量陷阱”。由此，所谓的“虚拟变量陷阱”的实质是出现完全多重共线性。可见，虚拟变量有其积极作用的一面，也有不良影响的一面，引入的虚拟变量适当，则发挥了积极的作用，引入的虚拟变量过度，则会带来负面的影响。

2、虚拟变量的“0”和“1”的选取原则

虚拟变量取“1”或“0”的原则，应从分析问题的目的出发予以界定。从理论上讲，虚拟变量取“0”值通常代表为比较的基础类型；而虚拟变量取“1”值通常代表为被比较的类型。例如，引入政府经济政策的变动对被解释变量的影响时，由于此时的比较是在政府经济政策不变的基础上进行的，故虚拟变量确定为：

$$D_i = \begin{cases} 1 & \text{基础类型: 政府经济政策变动} \\ 0 & \text{比较类型: 政府经济政策不变} \end{cases}$$

三、虚拟变量的作用

在计量经济模型中，虚拟变量可以发挥多方面的作用：

- (1) 可以作为属性因素的代表，如性别、所有制等；
- (2) 作为某些非精确计量的数量因素的代表，如受教育程度、管理者素质等；
- (3) 作为某些偶然因素或政策因素的代表，如战争、灾害、改革前后等；
- (4) 还可以作为时间序列分析中季节（月份）的代表；

(5) 可以实现分段回归，研究斜率、截距的变动，或比较两个回归模型的结构差异。

在计量经济学中，把包含有虚拟变量的模型称为虚拟变量模型。常用的虚拟变量模型有三种类型：(1) 解释变量中只包含虚拟变量，作用是在假定其他因素都不变时，只研究定性变量是否使被解释变量表现出显著差异；(2) 解释变量中既含定量变量，又含虚拟变量，研究定量变量和虚拟变量同时对被解释变量的影响；(3) 被解释变量本身为虚拟变量的模型，是被解释变量本身取值为 0 或 1 的模型，适于对某社会经济现象进行“是”与“否”的判断研究。

特别要注意的是，定型或属性变量，通常由 1 个以上的虚拟变量描述。例如，分析考证区域这样一个定性因素的影响时，若将区域因素划分为东、中、西三种属性时，在有截距项的回归模型中，只能引入 2 个虚拟变量，而这两个虚拟变量只是描述了 1 个定性因素（区域因素），而不是 2 个定性因素。当然，当定性因素为性别因素时，1 个虚拟变量就描述了 1 个定性因素。

第二节 虚拟解释变量的回归

在计量经济模型中，加入虚拟解释变量的途径有两种基本类型：一是加法类型；二是乘法类型。不同的途径引入虚拟变量有不同的作用，加法方式引入虚拟变量改变的是截距；乘法方式引入虚拟变量改变的是斜率。

一、用虚拟变量表示不同截距的回归——加法类型

以加法类型引入虚拟解释变量的模型，如 (8.3) 式那样，

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 D + u_i \quad (8.3)$$

在(8.3)所设定的计量经济模型中，虚拟解释变量与其他解释变量是相加关系。以加法形式引入虚拟解释变量，从计量经济模型的意义看，其作用是改变了设定模型的截距水平。

以加法方式引入虚拟变量时，分为四种情形：(1) 解释变量只有一个分为两种相互排斥类型的定性变量而无定量变量；(2) 解释变量包含一个定量变量和一个分为两种类型的定性变量；(3) 解释变量包含一个定量变量和一个两种以上类型的定性变量；(4) 解释变量包含一个定量变量和两个定性变量。

1、解释变量只有一个分为两种相互排斥类型的定性变量而无定量变量的回归

这种情况的模型又被称为方差分析模型，例如 (8.4) 式

$$Y_i = \alpha + \beta D_i + u_i \quad (8.4)$$

其中， Y_i 为居民的年可支配收入， D_i 为虚拟解释变量， $D_i=1$ 代表城镇居民； $D_i=0$ 代表非城镇居民。

(8.4) 式的意义是，假设其他因素（包括文化程度、职业、性别等）保持不变的条件下，研究城镇居民和非城镇居民的收入是否存在差别。当 u_i 满足古典假设时，由式 (8.4) 有：

$$\text{非城镇居民的年平均收入： } E(Y_i | D_i = 0) = \alpha \quad (8.5)$$

$$\text{城镇居民的年平均收入： } E(Y_i | D_i = 1) = \alpha + \beta \quad (8.6)$$

即在 (8.4) 式中，截距项 α 给出了非城镇居民的年平均可支配收入水平，而另一系数 β 则表明城镇居民年平均可支配水平不同于非城镇居民年平均可支配收入的部分。由式 (8.5) 和 (8.6) 可知，虚拟解释变量的作用是改变设定模型的截距水平。

为了检验城镇居民和非城镇居民的年均可支配收入是否有显著差别，可构造假设 $H_0: \beta = 0$ ，即城镇与非城镇居民年均可支配收入无差别。对式 (8.4) 回归，依据 β 估计值的 t 检验是否显著，可作出接受或不能接受 H_0 假设的判断。

2、解释变量包含一个定量变量和一个分为两种类型定性变量的回归

$$\text{例如} \quad Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + \mu_i \quad (8.7)$$

其中： Y : 消费支出； X : 收入； $D_i = \begin{cases} 1 & \text{城镇居民} \\ 0 & \text{农村居民} \end{cases}$

模型 (8.7) 的意义在于描述收入和城乡差别对居民消费支出的影响。(8.7) 式由一个定量解释变量 X 和一个分为两种类型的虚拟解释变量组成。注意这里一个定性变量具有两种类型，只使用了一个虚拟变量。当 (8.7) 式中的 u_i 服从古典假定时，有：

$$\text{基础类型： 农村居民消费支出： } E(Y_i | X_i, D_i = 0) = \alpha_1 + \beta X_i \quad (8.8)$$

$$\text{比较类型： 城镇居民消费支出： } E(Y_i | X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta X_i \quad (8.9)$$

其中 α_1 为差异截距系数。

(8.7) 式可图示为 8.1，表明非城镇居民与城镇居民两种类型收入函数的斜率相同（均为 β ），而截距水平不同。这说明，城镇居民和非城镇居民在消费支出水平上，存在着规模

为 α_1 的差异，而由收入因素而产生的平均消费支出水平变化却是相同的。

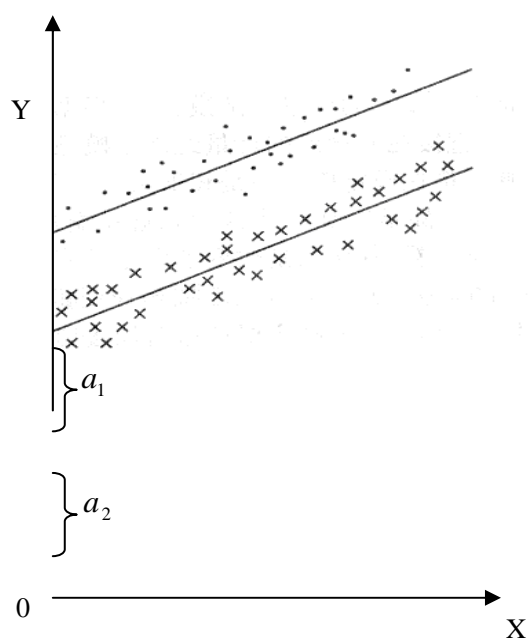


图 8.1 城镇农村居民消费支出水平的差异

在 $H_0: \alpha_1 = 0$ 的假设下，对参数 α_1 估计值的 t 检验，可以进行消费支出是否存在城乡差异的检验。

3、解释变量包含一个定量变量和一个两种以上类型的定性变量的回归

考虑以下模型：

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \quad (8.10)$$

其中： Y_i 为年医疗保健费用支出， X_i 为居民的年可支配收入，

$$D_2 = \begin{cases} 1 & \text{高中及高中教育以上} \\ 0 & \text{其他} \end{cases}, \quad D_3 = \begin{cases} 1 & \text{大专及大专以上} \\ 0 & \text{其他} \end{cases}$$

显然，模型（8.9）是描述居民的年医疗保健费用支出与居民可支配收入（定量变量）和受教育程度（定性变量）间的因果关系。这里，定性因素（受教育的程度）划分为三种类型：高中以下、高中、大专及大专以上。注意这里的定性变量有 3 种类型，依据虚拟变量设置规则引入了 $m-1=3-1=2$ 个虚拟变量，而且一个定性变量多种类型时，虚拟变量可同时取值为 0，但不能同时取值为 1，因为同一定性变量的各种类型间“非此即彼”。

当式（8.10）服从古典假定时，有：

$$\text{基础类型：高中以下教育： } E(Y_i | X_i, D_2 = 0, D_3 = 0) = \alpha_1 + \beta X_i \quad (8.11)$$

$$\text{比较类型：高中教育： } E(Y_i | X_i, D_2 = 1, D_3 = 0) = (\alpha_1 + \alpha_2) + \beta X_i \quad (8.12)$$

$$\text{大专及大专以上： } E(Y_i | X_i, D_2 = 0, D_3 = 1) = (\alpha_1 + \alpha_3) + \beta X_i \quad (8.13)$$

这表明，三种不同教育程度居民的医疗保健费用年均支出的起点水平（截距）不同，差异截距系数为 α_2 和 α_3 。对式（8.10）进行回归，检验 $H_0: \alpha_2 = 0$ 和 $H_0: \alpha_3 = 0$ 的 t 检验可以发现与比较基准组（高中以下教育水平）相比，另两种类型截距的差异在统计上是否存在显著差异。关于 $\alpha_2 = \alpha_3 = 0$ 的联合假设检验，也可由方差分析或 F 检验完成。

4、解释变量包含一个定量变量和两个定性变量的回归

以加法形式引入虚拟解释变量的作法，很容易扩展到处理一个以上定性变量的情形。例如依据某地区家庭调查资料所建立的卷烟需求模型：

$$Q_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta Y_i + u_i \quad (8.14)$$

其中， Q_i 为卷烟需求量， Y_i 为居民可支配收入， D_{2i} 和 D_{3i} 是虚拟解释变量，

$$D_{2i} = \begin{cases} 1 & \text{城镇居民} \\ 0 & \text{其他} \end{cases}, \quad D_{3i} = \begin{cases} 1 & \text{男性} \\ 0 & \text{女性} \end{cases}$$

一般认为，城镇居民的卷烟消费量高于非城镇居民，同时男性居民的吸烟量大于女性居民。为了分析城乡差别和性别差别对卷烟需求的影响，模型（8.14）以加法形式引入了两个虚拟解释变量。注意，这里有两个定性变量选用了两个虚拟变量去表示，这并不会出现“虚拟变量陷阱”，对比前面一个定性变量有三种类型时也用了两个虚拟变量，二者性质是不同的。而且注意这里的 D_{2i} 和 D_{3i} 是代表不同定性变量的虚拟变量，可以同时为 0，也可同时为 1，因为不同定性变量间并没有“非此即彼”的关系。

当式（8.14）满足古典假设时，有：

$$\text{基础类型：农村女性居民： } E(Q_i | Y_i, D_2 = 0, D_3 = 0) = \alpha_1 + \beta Y_i \quad (8.15)$$

$$\text{比较类型：农村男性居民： } E(Q_i | Y_i, D_2 = 0, D_3 = 1) = (\alpha_1 + \alpha_3) + \beta Y_i \quad (8.16)$$

$$\text{城镇女性居民： } E(Q_i | Y_i, D_2 = 1, D_3 = 0) = (\alpha_1 + \alpha_2) + \beta Y_i \quad (8.17)$$

$$\text{城镇男性居民： } E(Q_i | Y_i, D_2 = 1, D_3 = 1) = (\alpha_1 + \alpha_2 + \alpha_3) + \beta Y_i \quad (8.18)$$

显然，模型（8.14）是以农村女性居民为基础类型，并假设各种类型居民的卷烟需求函数只是有不同的截距，相对于收入的斜率系数 β 相同。用 t 检验分别检验 $\hat{\alpha}_2$ 和 $\hat{\alpha}_3$ 的统计显著性，可验证两个定性变量对截距是否有显著影响。

上述讨论的结果，可以推广到解释变量有多个定量变量和多个定性变量的情形。在推广过程中需要注意引入虚拟变量的个数应遵从前述的设置规则。例如，在考虑季节因素对冷饮销售量影响时，有春、夏、秋、冬四个类型的季节，依据设置规则，可引入 $m-1=4-1=3$ 个虚拟解释变量。

二、用虚拟变量表示不同斜率的回归——乘法类型

以乘法形式引入虚拟解释变量，是在所设定的计量经济模型中，将虚拟解释变量与其他解释变量相乘作为解释变量，以表示模型中斜率系数的差异。以乘法形式引入虚拟解释变量的主要作用在于：①关于两个回归模型的比较；②因素间的交互影响分析；③提高模型对现实经济现象的描述精度。

1、回归模型的比较——结构变化检验

以加法方式引入虚拟解释变量，属性因素仅影响不同类型模型的平均水平，而不会影响不同类型模型的相对变化。但是在现实经济生活中，属性因素也可能影响模型的斜率系数发生变化。例如，随着可支配收入水平的提高，城乡居民的消费结构将出现较大的差异，这种差异会表现在定性因素对斜率的影响上。又如，研究我国改革开放前后储蓄——收入总量间关系是否发生了变化时，也存在着经济结构变化而导致模型斜率发生变化的问题。这类问题可归结于两个回归模型的比较。例如，在研究改革开放前后储蓄——收入总量关系时，所设定的模型为：

$$\text{改革开放前： } Y_t = \lambda_1 + \lambda_2 X_t + u_{1t} \quad t=1950, 1951, \dots, 1977 \quad (8.19)$$

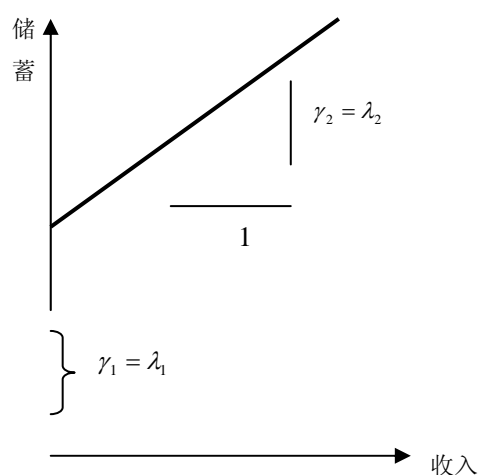
$$\text{改革开放后： } Y_t = \gamma_1 + \gamma_2 X_t + u_{2t} \quad t=1978, 1979, \dots, 2004 \quad (8.20)$$

其中：Y 为储蓄总额（亿元），X 为收入总额（亿元）， u_{1t} 、 u_{2t} 为随机扰动项。如果我

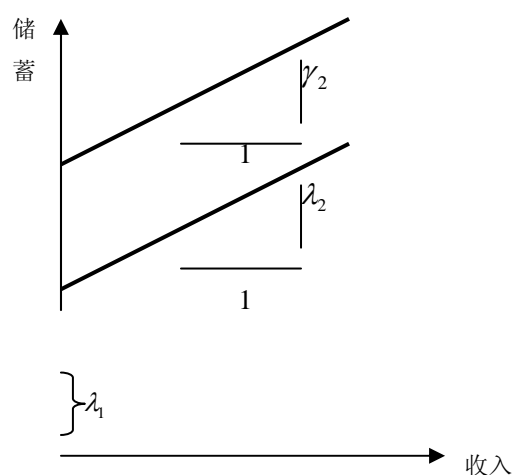
们分别对式 (8.19) 和式 (8.20) 在不同的时间区间内回归, 则可能得到以下四种结果:

- (1) $\lambda_1 = \gamma_1, \lambda_2 = \gamma_2$, 表明这两个回归模型是相同的, 或称为重合回归;
- (2) $\lambda_1 = \gamma_1, \lambda_2 \neq \gamma_2$, 表明这两个回归模型仅在位置水平上 (即截距水平上) 存在差异, 或称为平行回归;
- (3) $\lambda_1 \neq \gamma_1, \lambda_2 = \gamma_2$, 表明这两个回归模型具有相同的位置水平 (或起点相同) 而变化速率不等, 或称为共点回归;
- (4) $\lambda_1 \neq \gamma_1, \lambda_2 \neq \gamma_2$, 表明这两个回归模型完全不相同, 或称为不同的回归。

以上四种情形可用图示法描述 (见图 8.2):



(a) 重合回归



(b) 平行回归

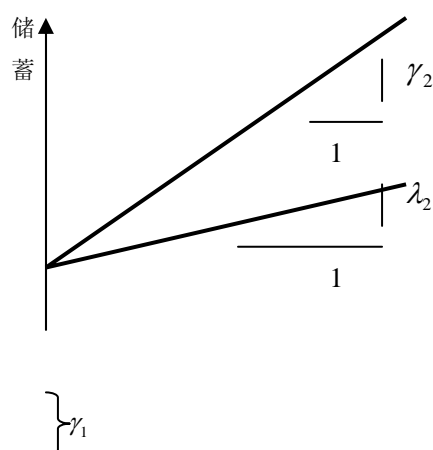
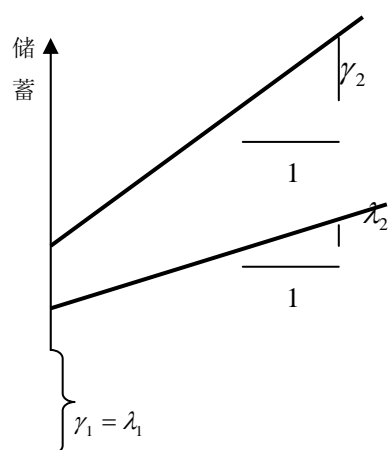




图 8.2 储蓄—收入回归模型

现在的问题是，当我们运用样本数据对式 (8.14) 和式 (8.15) 进行回归后，如何界定所得结果在统计意义上属于哪一种类型呢？这时可采用以乘法形式引入虚拟变量的方法。例如，对于改革开放前后储蓄——收入模型，可设定为：

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \quad (8.21)$$

其中，Y 为储蓄；X 为收入；D 为虚拟变量， $D_t = \begin{cases} 0 & \text{改革开放以前} \\ 1 & \text{改革开放以后(为什么?)} \end{cases}$

显然在式 (8.21) 中，以乘法形式引入了虚拟变量所形成的解释变量为 $D_t X_t$ ，以加法形式引入虚拟变量所形成的解释变量是 D_t 。

事实上，当式 (8.21) 满足古典假设时，有

$$\text{改革开放前: } E\{Y_t | D_t = 0, X_t\} = \alpha_1 + \beta_1 X_t \quad (8.22)$$

$$\text{改革开放后: } E\{Y_t | D_t = 1, X_t\} = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t \quad (8.23)$$

(8.22) 式和 (8.23) 式分别是改革开放后和改革开放前的平均储蓄函数。与 (8.19) 式及 (8.20) 式相比，有： $\lambda_1 = \alpha_1$ 、 $\lambda_2 = \beta_1$ ； $\gamma_1 = \alpha_1 + \alpha_2$ 、 $\gamma_2 = \beta_1 + \beta_2$ 。在 (8.21) 式中， α_2 称为截距差异系数， β_2 称为斜率差异系数，分别代表改革开放前后储蓄函数截距与斜率所存在的差异。当我们利用 1950—2000 年间的的数据估计式(8.21)时，等价于分别对 (8.19) 式和 (8.20) 式两个储蓄函数进行估计。

假如对 (8.21) 式用 OLS 法估计得

$$\begin{aligned} \hat{Y}_t &= -1.7502 + 1.4839 D_t + 0.1504 X_t - 0.1034 D_t X_t \\ &\quad (0.3319) \quad (0.4704) \quad (0.0163) \quad (0.0332) \\ t &= (-5.2733) \quad (3.1545) \quad (9.2270) \quad (-3.1144) \end{aligned}$$

结果表明，截距和斜率差异系数 α_2 、 β_2 在统计意义下均为显著的，说明改革开放前后的储蓄——收入行为确是不相同。即

$$\text{改革开放前} \quad \hat{Y}_i = -1.7502 + 0.1504X_i$$

$$\begin{aligned} \text{改革开放后} \quad \hat{Y}_i &= (-1.7502 + 1.4839) + (0.1504 - 0.1034)X_i \\ &= -0.2663 + 0.0470X_i \end{aligned}$$

以乘法形式引入虚拟变量作回归模型的比较和结构变化检验有一些优点：（1）用一个回归替代了多个回归，简化了分析过程；（2）可以方便地对模型结构的差异作各种假设检验；（3）合并了的回归增加了自由度，提高了参数估计的精确性。但是，也应注意合并后模型的 u_i 应服从基本假定，特别是所比较的方程的方差应相同，否则会出现异方差。

2、交互效应分析

当分析解释变量对变量的影响时，大多数情形只是分析了解释变量自身变动对被解释变量的影响作用，而没有深入分析解释变量间的相互作用对被解释变量的影响。前面讨论的分析两个定性变量对被解释变量影响的虚拟变量模型中，暗含着一个假定：两个定性变量是分别独立地影响被解释变量的。但是在实际经济活动中，两个定性变量对被解释变量的影响可能存在一定的交互作用，即一个解释变量的边际效应有时可能要依赖于另一个解释变量。为描述这种交互作用，可以把两个虚拟变量的乘积以加法形式引入模型。

考虑下列模型：

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \quad (8.24)$$

其中： Y_i 为农副产品生产总收益， X 为农副产品生产投入， D_{2i} 为代表油菜籽生产虚拟变量， D_{3i} 为代表养蜂生产虚拟变量：

$$D_{2i} = \begin{cases} 1 & \text{发展油菜籽生产} \\ 0 & \text{其他} \end{cases} ; D_{3i} = \begin{cases} 1 & \text{发展养蜂生产} \\ 0 & \text{其他} \end{cases}$$

显然（8.22）式描述了是否发展油菜籽生产与是否发展养蜂生产的差异对农副产品总收益的影响。虚拟解释变量 D_{2i} 和 D_{3i} 是以加法形式引入的，那么暗含着假设：油菜籽生产和养蜂生产是分别独立地影响农副产品生产总收益。但是，在发展油菜籽生产时，同时也发展养蜂生产，所取得的农副产品生产总收益，可能会高于不发展养蜂生产的情况。即在是否发展油菜籽生产与养蜂生产的虚拟变量 D_{2i} 和 D_{3i} 间，很可能存在着一定的交互作用，且这种交互影响对被解释变量农副产品生产收益会有影响。

为了描述交互作用对被解释变量的效应，在（8.24）式中以加法形式引入两个虚拟解释

变量的乘积，即

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i \quad (8.25)$$

(8.25) 式中各变量的含义与 (8.24) 式相同。

基础类型：为不发展油菜籽生产，也不发展养蜂生产时农副产品生产总收益的平均支出：

$$E(Y_i | D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i \quad (8.26)$$

对比类型：为同时发展油菜籽生产和养蜂生产时，农副产品生产总收益的平均支出

$$E(Y_i | D_2 = 1, D_3 = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i \quad (8.27)$$

这里的截距水平由四项组成，其中：

α_2 为是否发展油菜籽生产对农副产品生产总收益的截距差异系数；

α_3 为是否发展养蜂生产对农副产品生产总收益的截距差异系数；

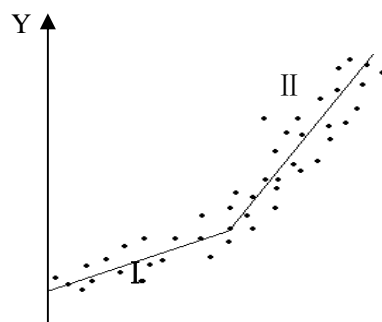
α_4 同时发展油菜籽生产和养蜂生产时对农副产品生产总收益的交互效应系数。

关于交互效应是否存在，可借助于交互效应虚拟解释变量系数的显著性检验来加以判断。如果 t 检验表明交互效应虚拟变量 $D_{2i} D_{3i}$ 在统计意义上是显著时，说明交互效应对 Y_i 存在显著影响。

3、分段线性回归

有的社会经济现象的变动，会在解释变量达到某个临界值时发生突变，为了区分不同阶段的截距和斜率可利用虚拟变量进行分段回归。

例如，某公司为了激励公司销售人员，按其销售额的一定比例计提奖励，但是销售额在某一目标水平 X^* 以下和以上时计提奖励的方法不同。当销售额高于 X^* 时，计提奖励额与销售额的比例要高于销售额低于 X^* 时的比例，也就是高于 X^* 时，奖励额与销售额的线性关系更为陡峭（如图 8.3 所示）。为了确切地描述奖励额度（Y）与销售额（X）间的关系，需要分两段进行回归。这种分段回归可以用虚拟变量来实现。



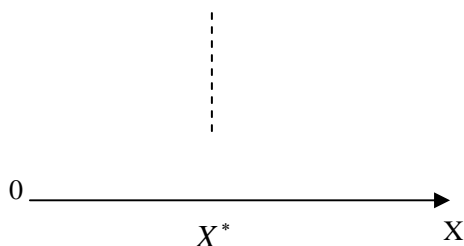


图 8.3 奖励额与销售额的关系

设虚拟变量 D 为：

$$D = \begin{cases} 1 & X \geq X^* \\ 0 & X < X^* \end{cases}$$

则奖励额度 (Y_t) 和销售额 (X_t) 间的关系式可以统一地表示为：

$$Y_t = \alpha_0 + \beta_1 X_t + \beta_2 (X_t - X^*) D_t + u_t \quad (8.28)$$

其中, Y_t 为奖励额, X_t 为销售额, X^* 为已知的销售目标临界水平。利用统计资料估计 (8.28)

式的参数, 就可以得到不同斜率和截距的回归方程：

$$\text{销售额低于 } X^* \text{ 时: } E(Y_t | X_t, D=0, X^*) = \hat{\alpha}_0 + \hat{\beta}_1 X_t \quad (8.29)$$

$$\text{销售额不低于 } X^* \text{ : } E(Y_t | X_t, D=1, X^*) = (\hat{\alpha}_0 + \hat{\beta}_1 X^*) + (\hat{\beta}_1 + \hat{\beta}_2)(X_t - X^*) \quad (8.30)$$

$$\text{整理得} \quad E(Y_t | X_t, D=1, X^*) = (\hat{\alpha}_0 - \hat{\beta}_2 X^*) + (\hat{\beta}_1 + \hat{\beta}_2) X_t \quad (8.31)$$

显然, β_1 是图 8.3 中第 I 段回归直线的斜率, 而 $\hat{\beta}_1 + \hat{\beta}_2$ 则是第 II 段回归直线的斜率。只要检验 $\hat{\beta}_2$ 的统计显著性, 就可以判断在所设定的临界水平 X^* 处是否存在 “突变”。

应当注意, 在分段回归中, 第一、二段回归不仅截距不同, 而且斜率也不同。在分为两段回归时, 使用了一个虚拟变量, 容易推广, 分为 K 段回归时, 可用 $K-1$ 个虚拟变量。

*第三节 虚拟被解释变量^①

在计量经济学模型中, 虚拟变量除了可以作为解释变量外, 还可以作为被解释变量。当虚拟变量作为被解释变量时, 其作用是对某一经济现象或活动进行 “是” 与 “否” 的判断或决策。例如, 研究是否购买商品住房、是否参加人寿或财产保险、是否能按期偿还贷款、新

^①本节内容本科教学中供选择使用。

产品在市场上是否畅销、对某一改革措施所持的态度等。这些问题的特征是被研究的对象(即被解释变量)在受到多种因素影响时,其取值只有两种状态:“是”与“否”。这在计量经济学中被称为“二元型响应”现象,这种现象常在市场研究或社会问题研究中遇到。如何处理二元型响应被解释变量模型的估计、推断问题,是本节要解决的问题。

一、线性概率模型(LPM)

1、什么是线性概率模型

假设住户是否购买商品房的决定主要依赖于其收入水平。那么考虑下列模型:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (8.32)$$

其中, X_i 为住户的收入; Y 为一虚拟变量表示的住户购买商品住房的情况:

$$Y = \begin{cases} 1 & \text{已购买商品住房} \\ 0 & \text{未购买商品住房} \end{cases}$$

现在的问题是:我们前面讨论的回归分析主要是研究 $E(Y_i | X_i) = \beta_1 + \beta_2 X_i$ 的问题,即研究条件均值的轨迹的问题,而在上述模型中,被解释变量是某种属性发生与否的状况,怎样把某种属性发生与否的问题同条件均值的轨迹研究联系起来?当然,在计量经济学中,研究被解释变量某种属性发生与否,通常是研究这种属性发生与否的概率。也就是说,上述问题可表述为:怎样把被解释变量某种属性发生与否的概率问题同条件均值的轨迹研究联系起来?另外,若概率问题与条件均值轨迹能够联系起来的话,那么,我们所讨论的线性回归分析会出现什么问题?

分析 (8.32) 式,其中, u_i 服从 $E(u_i)=0$, 有:

$$E(Y_i | X_i) = \beta_1 + \beta_2 X_i \quad (8.33)$$

另一方面, Y_i 是取值为 0 和 1 的随机变量,那么 Y_i 有下列分布 (p_i 为 $Y_i=1$ 的概率):

Y_i	0	1
概率	$1 - p_i$	p_i

根据数学期望的定义

$$E(Y_i) = 0 \times (1 - p_i) + 1 \times p_i = p_i \quad (8.34)$$

也就是说, $E(Y_i)$ 等于 Y_i 取值为 1 时的概率,即:

$$E(Y_i) = \text{Prob}(Y_i = 1 | X_i) = p_i \quad (8.35)$$

注意事件 $Y=1$ 发生是在给定收入 X 的条件下发生的, 因此 $E(Y_i) = E(Y_i | X_i)$ 于是, 比较 (8.33) 式和 (8.34) 式, 则有:

$$E(Y_i | X_i) = \beta_1 + \beta_2 X_i = p_i \quad (8.36)$$

表明购买商品用房的概率是收入的线性函数。像(8.32)式那样, 以虚拟变量作为被解释变量的模型的条件期望实际上等于随机变量 Y_i 取值为 1 的条件概率。即当住户的收入水平为 X 时, 其购买商品住房的概率可表示成 X 的线性函数, 故 (8.32) 式也被称为线性概率模型 (LPM)。显然, 只要得到 (8.32) 式中 β_1 和 β_2 和估计量后, 就可以估计出不同收入水平住户购买商品住房的概率。

由于购买商品住房的概率 p_i 必须在 0 和 1 之间, 故在估计式(8.32)式时必须满足约束条件

$$0 \leq E(Y_i = 1 | X_i) \leq 1 \quad (8.37)$$

2、线性概率模型的估计

从形式上看, (8.32) 式与普通的线性计量经济模型相似, 是否能够运用 OLS 法直接对其进行估计呢? 答案是否定的。因为直接采用 OLS 法对 (8.32) 式那样的模型进行估计, 将会遇到一些特殊的问题, 使得估计结果失去了合理的经济解释, 因而需要寻求相应的处理方法。

(1) 随机扰动项 u_i 的非正态性

在线性概率模型中, 关于 u_i 的正态性假设不再成立, 因为 (8.30) 式的随机误差项为:

$$u_i = Y_i - \beta_1 - \beta_2 X_i \quad (8.38)$$

此时, 当 $Y_i=1$ 时 $u_i = 1 - \beta_1 - \beta_2 X_i$

当 $Y_i=0$ 时 $u_i = -\beta_1 - \beta_2 X_i$

显然, 这里的 u_i 不遵从正态分布, 而是服从二项分布。

线性概率模型中的随机扰动项 u_i 不遵从正态分布, 对参数的估计并不产生影响, OLS 法本身并不要求随机扰动项 u_i 具备正态性, 此时参数的 OLS 估计仍是最佳无偏估计量。但

对参数的假设检验和区间估计要求随机扰动项 u_i 遵从正态分布。不过，随着样本容量的无限增大，根据中心极限定理，OLS 估计量的概率分布将会趋近于正态分布。因此，大样本条件下线性概率模型的统计推断，也可以按正态性假设条件下 OLS 的统计推断方式进行。这就是说，直接运用 OLS 法对线性概率模型进行估计，对参数的估计不会产生太大影响。

(2) 随机扰动项 u_i 的异方差性

根据 Y_i 的概率分布有： $Y_i=1$ 时， $u_i=1-(\beta_1+\beta_2X_i)$ 的概率为 p_i ； $Y_i=0$ 时， $u_i=-(\beta_1+\beta_2X_i)$ 的概率为 $1-p_i$ ，即

$$\begin{array}{ccc} u_i & -\beta_1 + \beta_2 X_i & 1 - \beta_1 + \beta_2 X_i \\ \text{概率} & 1 - p_i & p_i \end{array}$$

根据方差的定义

$$\begin{aligned} \text{Var}(u_i) &= E(u_i - E(u_i))^2 \\ &= E(u_i^2) \\ &= (-\beta_1 - \beta_2 X_i)^2 (1 - p_i) + (-\beta_1 + \beta_2 X_i)^2 p_i \\ &= (-\beta_1 - \beta_2 X_i)^2 (1 - \beta_1 - \beta_2 X_i) + (1 - \beta_1 - \beta_2 X_i)^2 (\beta_1 + \beta_2 X_i) \\ &= (-\beta_1 - \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i) \\ &= p_i(1 - p_i) \end{aligned} \quad (8.39)$$

这里利用了 $p_i = \beta_1 + \beta_2 X_i$ 。(8.39)式表示，当 u_i 满足 $E(u_i)=0$ 和 $E(u_i u_j)=0(i \neq j)$ 时， u_i 的方差却是 Y_i 条件期望的函数，即 $\text{Var}(u_i) = f(E(Y_i | X_i))$ ，这表明 u_i 是异方差的。这时利用 OLS 法所得的 LPM 的估计量不再具有最小方差的特性，且各参数估计量的标准差也不可信。也就是说，LPM 参数的 OLS 估计量虽仍为线性无偏估计量，但不是最佳估计量。

为了消除异方差性的影响，可利用第五章中有关修正异方差的方法，例如可用加权最小二乘法（WLS）修正异方差。

根据前面的讨论，已知 LPM 中 u_i 的方差是 Y_i 条件期望的函数，故选择权重的一种方法是：

$$\sqrt{w_i} = \sqrt{E(Y_i | X_i)[1 - (Y_i | X_i)]} = \sqrt{p_i(1 - p_i)} \quad (8.40)$$

其中， w_i 为权重。

对 (8.32) 式两边加权，有：

$$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_1}{\sqrt{w_i}} + \beta_2 \frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \quad (8.41)$$

(8.41)式中权重 w_i 是未知的，随机扰动项 $u_i / \sqrt{w_i}$ 也是未知的，在实践中为了估计 w_i 进而估计 LPM 模型，可采取以下步骤：

第一步，不考虑异方差，用 OLS 法估计原模型 (8.30)，计算 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ 作为 $E(Y_i | X_i) = P_i$ 的估计值 \hat{P}_i ，取 $\hat{w}_i = \hat{Y}_i(1 - \hat{Y}_i) = \hat{P}_i(1 - \hat{P}_i)$ 作为 w_i 的估计值

第二步，用 \hat{w}_i 按照(8.41)式对观察数据 Y_i 和 X_i 进行变换，再用 OLS 法估计变换后的模型参数，得 LPM 的参数，从而消除异方差。

(3)不满足 $0 \leq E(Y_i | X_i) \leq 1$ 的约束

在线性概率模型中， $E(Y_i | X_i)$ 表示在给定 X 的条件下，事件 Y 发生的概率，从理论上， $E(Y_i | X_i)$ 的取值范围必须在 0 和 1 之间，然而在实证分析中， $E(Y_i | X_i)$ 的估计量 \hat{Y}_i 并不一定介于 0 和 1 之间，也就是说， \hat{Y}_i 的值可能大于 1，也可能小于 0，这是 LPM 的 OLS 法估计存在的实际问题。解决这一问题的方法之一，是当 $\hat{Y}_i > 1$ 时，就认定 $\hat{Y}_i = 1$ ；当 $\hat{Y}_i < 0$ 时，就认定 $\hat{Y}_i = 0$ 。这是人为的把大概率事件当作必然事件，把小概率事件当作不可能事件。另一类方法，是选择 Logit 模型或 Probit 模型等能够保证满足 $0 \leq E(Y_i | X_i) \leq 1$ 约束的非线性模型。

3、非线性概率模型

应当指出的是，虽然我们可以采用 WLS 解决异方差性问题、增大样本容量减轻非正态性问题，通过约束迫使所估的事件 Y 发生的概率落入 0-1，但是，LPM 与经济意义的要求不符：随着 X 的变化， X 对 p_i 的“边际效应”保持不变。如在住户是否购买商品房的例子中，

当 $\hat{\beta}_2 = 0.1$ 时, 表明 X 每变化一个单位 (比如说 1000 元), 拥有商品住房的概率恒等地增加 0.1。这就是说, 无论住户的收入水平为 8000 元, 还是 22000 元, 拥有商品住房的概率都以相同的增量增加。在线性概率模型中, 不论 X 的变化是在什么水平上发生的, 参数都不发生变化, 显然这与现实经济所发生的情况是不符的。

因此, 表现概率平均变化比较理想的模型应当具有这样的特征:

(1) 概率 $p_i = E(Y_i = 1 | X_i)$ 随 X 的变化而变化, 但永远不超出 0—1 区间。

(2) 随着 $x_i \rightarrow -\infty$, $p_i \rightarrow 0$; 随着 $x_i \rightarrow \infty$, $p_i \rightarrow 1$; 即随着 x_i 变小, 概率 p_i 趋于零的速度越来越慢; 而随着 X_i 变大, 概率 p_i 趋于 1 的速度也越来越慢。 p_i 随 X_i 变化而变化, 且变化速率不是常数, p_i 和 X_i 之间是非线性关系。

符合这些特征的函数可用图 8.4 形象地刻画。

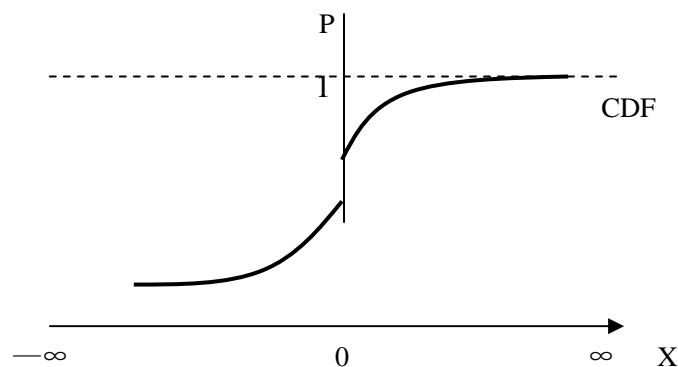


图 8.4 非线性概率函数的图形

从图中可知, 图 8.4 所示的模型满足 $0 \leq E(Y_i | X_i) \leq 1$, 以及 p_i 是 X_i 非线性函数的假设, 呈现出 S—型的曲线特征。因此可以设法找到符合这种 S—型曲线特征的函数形式来作为二元响应计量经济模型的设定形式。

原则上, 任何适当的、连续的、定义在实轴上的概率分布都将满足上述两个条件。对于连续随机变量来说, 密度函数的积分代表概率的大小, 也就是说, 连续随机变量的 (累积) 分布函数 (CDF) 可以满足上述两个要求。通常选择逻辑斯蒂分布函数和正态分布的累积分布函数去设定非线性概率模型。当选用逻辑斯蒂分布时, 就生成了 Logit 模型 (对数单位模型), 本书只介绍 Logit 模型。

二、对数单位模型 (Logit 模型)

1、Logit 模型的基本概念

如上所述，当选择用逻辑斯蒂分布函数（logistic distribution）去设定二元响应计量经济模型时，有

$$\text{Prob}(Y_i=1) = p_i = \frac{e^{\beta_1 + \beta_2 X_i}}{1 + e^{\beta_1 + \beta_2 X_i}} = \frac{1}{1 + e^{-\beta_1 - \beta_2 X_i}} = \frac{1}{1 + e^{-z_i}} \quad (8.42)$$

其中， $z_i = \beta_1 + \beta_2 X_i$ 。

(8.42) 式有以下特征：

(1) 随着 $z_i \rightarrow \infty$ ， $p_i \rightarrow 1$ （1 为 p_i 的饱和值）；反之， $z_i \rightarrow -\infty$ 时， $p_i \rightarrow 0$ ；即 $-\infty \leq z_i \leq \infty$ ， $0 \leq p_i \leq 1$ ； $z_i = 0$ 时， $p_i = 0.5$ 。

(2) (8.42) 式有一个拐点，在拐点之前，随 z_i 或 X_i 增大， p_i 的增长速度越来越快；在拐点之后，随 z_i 或 X_i 增大， p_i 的增长速度越来越慢，逐渐趋近于 1。

这些特征正好满足前面讨论的非线性概率模型的要求。

考虑在估计中便利，我们采用以下变换：

$$\begin{aligned} L_i &= \ln \left(\frac{\text{Prob}(Y=1)}{\text{Prob}(Y=0)} \right) = \ln \left(\frac{p_i}{1-p_i} \right) \\ &= \ln \left(\frac{e^{\beta_1 + \beta_2 X_i} / 1 + e^{\beta_1 + \beta_2 X_i}}{1 - (e^{\beta_1 + \beta_2 X_i} / 1 + e^{\beta_1 + \beta_2 X_i})} \right) = \ln(e^{\beta_1 + \beta_2 X_i}) = \beta_1 + \beta_2 X_i \end{aligned} \quad (8.43)$$

(8.43) 中，比率 $\frac{p_i}{1-p_i}$ 通常被称为机会比率，即所研究的事件（或属性）“发生”与“没有发生”的概率之比。机会比率在市场调查民意测验等社会学以及流行病学方面有着广泛的应用。“机会比率的对数” $L_i = \ln(\frac{p_i}{1-p_i})$ 被称为对数单位，这里的对数单位 L_i 不仅是 X_i 的线性函数，而且也是 β 的线性函数。所以，(8.43) 也称为对数单位模型（或 logit 模型）。

2、Logit 模型的估计

虽然 Logit 模型 (8.42) 或 (8.43) 式满足非线性概率模型的要求，但由于 p_i 不仅对 X_i 是非线性关系，而且对 β_1 和 β_2 也是非线性关系，不能直接运用 OLS 法估计参数。必须设法把非线性关系转换为可以运用 OLS 估计的线性形式。

若记 p_i 为事件发生的概率，那么有：

$$p_i = \frac{1}{1 + e^{-Z_i}}$$

$$1 - p_i = 1 - \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{Z_i}} \quad (8.44)$$

由(8.43)和(8.44)式有：

$$\frac{p_i}{1 - p_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \quad (8.45)$$

对 (8.45) 式两边取自然对数：

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right) = \ln(e^{Z_i}) = Z_i = \beta_1 + \beta_2 X_i \quad (8.46)$$

模型 (8.46) 表明， X_i 变动一个单位，机会比率的对数（注意不是概率 p_i ）平均变化 β_2 个单位。需要注意对数单位模型的以下特点：

(1) 随着 p_i 从 0 变化到 1，或 Z 从 $-\infty$ 变化到 ∞ ，对数单位 L_i 从 $-\infty$ 变化到 ∞ ，即概率 p_i 在 0 与 1 之间，但对数单位 L_i 并不一定在 0 与 1 之间。

(2) 虽然对数单位 L_i 对 X_i 是线性的，但概率 p_i 对 X_i 并不是线性的，这与线性概率模型不同。

(3) 注意对数单位模型中参数的意义： β_2 是 X_i 每变动一个单位时，对数单位 L_i （机会比率的对数）的平均变化，然而我们研究的目的并不是对数单位 L_i ，而是概率 p_i 。

(4) 如果设法估计出参数 β_1 和 β_2 ，给定某一水平 $X_i = X^*$ ，若欲估计 p_i ，当 β_1 和 β_2 估计量已知时，可从 (8.46) 式中直接得到 ($Z_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$)，就可能计算出要估计的概率 p_i 。

从经济计量的角度引入随机扰动项，将式 (8.46) 改记为：

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right) = Z_i = \beta_1 + \beta_2 X_i + u_i \quad (8.47)$$

现在的问题是如何得到 β_1 和 β_2 的估计量？对 (8.47) 式直接估计会遇到以下困难：

(1) 当事件发生时 $p_i = 1$ ， $L_i = \ln(1/0)$ ；当事件没有发生时 $p_i = 0$ ， $L_i = \ln(0/1)$ ，

机会比率 $p_i/(1-p_i)$ 的对数都无意义，不能直接用 OLS 法估计模型，而只能采用极大似然法（ML）估计参数。当样本容量 N 较大，可选用加权最小二乘法进行估计。

（2）估计参数需要的机会比率对数 L_i 的数据无法观测。解决办法是对应于每个 X_i ，样本观测值个数 N_i 较大时，可利用整理汇总的数据，用相对频率作为对 p_i 的估计，并估计机会比率对数 L_i 。例如购商品房的模型，对于收入水平 X_i ，家庭总数为 N_i ，其中购商品房家庭数为 n_i ，可计算相对频率 $\hat{P}_i = n_i/N_i$ 。样本容量 N_i 足够大时， \hat{P}_i 可视为对 p_i 的较好估计，并可用来估计机会比率对数 L_i ： $\hat{L}_i = \ln[\hat{P}_i/(1-\hat{P}_i)]$ 。

（3）（8.47）式模型的随机项 u_i 为异方差，可以证明， N 足够大时

$$u_i \sim N[0, \frac{1}{N_i \hat{P}_i (1-\hat{P}_i)}] \quad (8.48)$$

为了估计 u_i 的方差 σ_i^2 ，可通过用相对频率 \hat{P}_i 代替 p_i 去估计：

$$\hat{\sigma}_i^2 = \frac{1}{N_i \hat{P}_i (1-\hat{P}_i)} \quad (8.49)$$

估计出 u_i 的方差以后，可用加权最小二乘法去估计参数，权数 w_i 为：

$$w_i = \sqrt{\hat{\sigma}_i^2} = 1/\sqrt{N_i \hat{P}_i (1-\hat{P}_i)} \quad (8.50)$$

可以看出，对数单位模型参数的估计程序是较为繁琐的，但运用 Eviews 进行估计却较方便，具体的估计步骤在下一节的案例中介绍。

另外，关于二元选择模型的模型设定检验、异方差性检验、拟合优度分析等内容，已超出本书的讨论范围，在此不作讨论。

第四节 案例分析

改革开放以来，随着经济的发展中国城乡居民的收入快速增长，同时城乡居民的储蓄存款也迅速增长。经济学界的一种观点认为，20 世纪 90 年代以后由于经济体制、住房、医疗、养老等社会保障体制的变化，使居民的储蓄行为发生了明显改变。为了考察改革开放以来中国居民的储蓄存款与收入的关系是否已发生变化，以城乡居民人民币储蓄存款年底余额代表

居民储蓄 (Y)，以国民总收入 GNI 代表城乡居民收入，分析居民收入对储蓄存款影响的数量关系。

表 8.1 为 1978—2003 年中国的国民总收入和城乡居民人民币储蓄存款年底余额及增加额的数据。

表 8.1 国民总收入与居民储蓄存款 单位：亿元

年 份	国民总收 入 (GNI)	城乡居民人 民币储蓄存 款年底余额 (Y)	城乡居民人 民币储蓄存 款增加额 (YY)	年 份	国民总收 入 (GNI)	城乡居民人 民币储蓄存 款年底余额 (Y)	城乡居民人 民币储蓄存 款 增 加 额 (YY)
1978	3624.1	210.6	NA	1991	21662.5	9241.6	2121.800
1979	4038.2	281.0	70.4	1992	26651.9	11759.4	2517.800
1980	4517.8	399.5	118.5	1993	34560.5	15203.5	3444.100
1981	4860.3	532.7	124.2	1994	46670.0	21518.8	6315.300
1982	5301.8	675.4	151.7	1995	57494.9	29662.3	8143.500
1983	5957.4	892.5	217.1	1996	66850.5	38520.8	8858.500
1984	7206.7	1214.7	322.2	1997	73142.7	46279.8	7759.000
1985	8989.1	1622.6	407.9	1998	76967.2	53407.5	7615.400
1986	10201.4	2237.6	615.0	1999	80579.4	59621.8	6253.000
1987	11954.5	3073.3	835.7	2000	88254.0	64332.4	4976.700
1988	14922.3	3801.5	728.2	2001	95727.9	73762.4	9457.600
1989	16917.8	5146.9	1374.2	2002	103935.3	86910.6	13233.20
1990	18598.4	7119.8	1923.4	2003	116603.2	103617.7	16631.90

数据来源：《中国统计年鉴 2004》，中国统计出版社。表中“城乡居民人民币储蓄存款年增加额”为年鉴数值，与用年底余额计算的数值有差异。

为了研究 1978—2003 年期间城乡居民储蓄存款随收入的变化规律是否有变化，考证城乡居民储蓄存款、国民总收入随时间的变化情况，如下图所示：

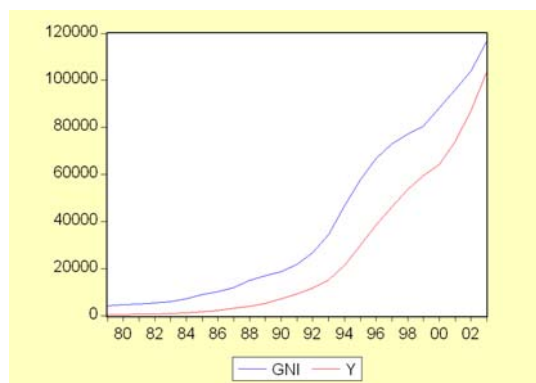


图 8.5

从图 8.5 中，尚无法得到居民的储蓄行为发生明显改变的详尽信息。若取居民储蓄的增量（YY），并作时序图（见图 8.6）

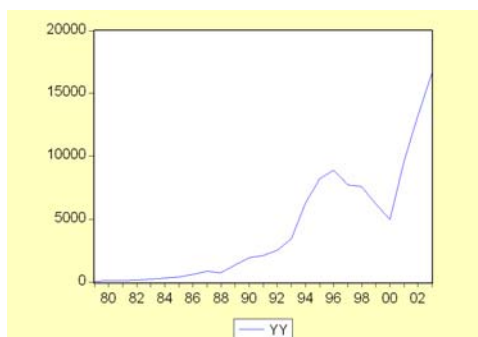


图 8.6

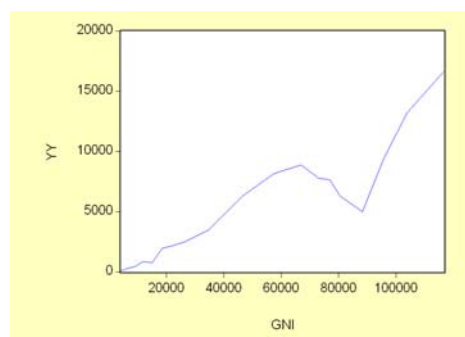


图 8.7

从居民储蓄增量图可以看出，城乡居民的储蓄行为表现出了明显的阶段特征：在 1996 年和 2000 年有两个明显的转折点。再从城乡居民储蓄存款增量与国民总收入之间关系的散布图看（见图 8.7），也呈现出了相同的阶段性特征。

为了分析居民储蓄行为在 1996 年前后和 2000 年前后三个阶段的数量关系，引入虚拟变量 D_1 和 D_2 。 D_1 和 D_2 的选择，是以 1996、2000 年两个转折点作为依据，1996 年的 GNI 为 66850.50 亿元，2000 年的 GNI 为 82254.00 亿元，并设定了如下以加法和乘法两种方式同时引入虚拟变量的模型：

$$YY_t = \beta_1 + \beta_2 GNI_t + \beta_3 (GNI_t - 66850.50) D_{1t} + \beta_4 (GNI_t - 82254.00) D_{2t} + u_t$$

$$\text{其中： } D_{1t} = \begin{cases} 1 & t = 1996 \text{ 年以后} \\ 0 & t = 1996 \text{ 年及以前} \end{cases} \quad D_{2t} = \begin{cases} 1 & t = 2000 \text{ 年以后} \\ 0 & t = 2000 \text{ 年及以前} \end{cases}$$

对上式进行回归后，有：

Dependent Variable: YY

Method: Least Squares

Date: 06/16/05 Time: 23:27

Sample (adjusted): 1979 2003

Included observations: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-830.4045	172.1626	-4.823374	0.0001
GNI	0.144486	0.005740	25.17001	0.0000
(GNI-66850.50)*DUM1	-0.291371	0.027182	-10.71920	0.0000
(GNI-88254.00)*DUM2	0.560219	0.040136	13.95810	0.0000
R-squared	0.989498	Mean dependent var		4168.652
Adjusted R-squared	0.987998	S.D. dependent var		4581.447
S.E. of regression	501.9182	Akaike info criterion		15.42040
Sum squared resid	5290359.	Schwarz criterion		15.61542
Log likelihood	-188.7550	F-statistic		659.5450
Durbin-Watson stat	1.677712	Prob(F-statistic)		0.000000

即有:

$$YY_t = -830.4045 + 0.1445GNI_t - 0.2914(GNI_t - 66850.50)D_{1t} + 0.5602(GNI_t - 88254.00)D_{2t}$$

$$se = (172.1626) (0.0057) (0.0272) (0.0401)$$

$$t = (-4.8234) (25.1700) (-10.7192) (13.9581)$$

$$R^2 = 0.9895 \quad \bar{R}^2 = 0.9880 \quad F = 659.5450 \quad DW = 1.6777$$

由于各个系数的 t 检验均大于 2，表明各解释变量的系数显著地不等于 0，居民人民币储蓄存款年增加额的回归模型分别为：

$$YY_t = \begin{cases} YY_t = -830.4045 + 0.1445GNI_t + \varepsilon_{1t} & t \leq 1996 \\ YY_t = 18649.8312 - 0.1469GNI_t + \varepsilon_{2t} & 1996 < t \leq 2000 \\ YY_t = -30790.0596 + 0.4133GNI_t + \varepsilon_{3t} & t > 2000 \end{cases}$$

这表明三个时期居民储蓄增加额的回归方程在统计意义上确实是不相同的。1996 年以前收入每增加 1 亿元，居民储蓄存款的增加额为 0.1445 亿元；在 2000 年以后，则为 0.4133

亿元，已发生了很大变化。上述模型与城乡居民储蓄存款与国民总收入之间的散布图是吻合的，与当时中国的实际经济运行状况也是相符的。

需要指出的是，在上述建模过程中，主要是从教学的目的出发运用虚拟变量法则，没有考虑通货膨胀因素。而在实证分析中，储蓄函数还应当考虑通货膨胀因素。

第八章小结

1、虚拟变量是人工构造的取值为 0 和 1 的作为属性变量代表的变量。

2、虚拟变量个数的设置有一定规则：在有截距项的模型中，若定性因素有 m 个相互排斥的类型，只能引入 $m-1$ 个虚拟变量，否则会陷入所谓“虚拟变量陷阱”，产生完全的多重共线性。

3、在计量经济模型中，加入虚拟解释变量的途径有两种基本类型：一是加法类型；二是乘法类型。以加法方式引入虚拟变量改变的是模型的截距；以乘法方式引入虚拟变量改变的是模型的斜率。

4、解释变量只有一个分为两种相互排斥类型的定性变量而无定量变量的回归，称为方差分析模型。

5、解释变量包含一个分为两种类型定性变量的回归时，只使用了一个虚拟变量；解释变量包含一个两种以上类型的定性变量的回归时，定性变量有 m 种类型，依据虚拟变量设置规则引入了 $m-1$ 个虚拟变量。

7、解释变量包含两个（或 K 个）定性变量的回归中，可选用了两个（或 K 个）虚拟变量去表示，这并不会出现“虚拟变量陷阱”。

8、以乘法形式引入虚拟解释变量的主要作用在于：对回归模型结构变化的检验；定性因素间交互作用的影响分析；分段线性回归等。

9、以虚拟变量作为被解释变量的模型中，被解释变量 Y_i 的条件期望实际上是 Y_i 取值为 1 的条件概率。线性概率模型（LPM）存在一定局限性，模型估计也面临某些困难。对数单位模型（Logit 模型）是以虚拟变量作为被解释变量的非线性模型之一。

第八章主要公式表

虚拟变量表示不同截距的回归——加法类型	$Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 D + u_t \quad D_i = \begin{cases} 1 \\ 0 \end{cases}$
虚拟变量表示不同斜率的回归——乘法类型	$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$
用虚拟变量作交互效应分析	$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$
分段线性回归	$Y_t = \alpha_0 + \beta_1 X_t + \beta_2 (X_t - X^*) D_t + u_t$ $D = \begin{cases} 1 & X \geq X^* \\ 0 & X < X^* \end{cases}$
线性概率模型	$E(Y_i X_i) = \beta_1 + \beta_2 X_i = p_i$
对数单位模型 (Logit 模型)	$p_i = E(Y_i = 1 X) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$ $p_i = E(Y_i = 1 X) = \frac{1}{1 + e^{-Z_i}}$ $L_i = \ln \frac{p_i}{1 - p_i} = Z_i = \beta_1 + \beta_2 X_i + u_i$
机会比率	$\frac{p_i}{1 - p_i}$
对数单位—机会比率的对数	$L_i = \ln \left(\frac{p_i}{1 - p_i} \right)$

思考题与练习题

思考题

8.1 什么是虚拟变量？它在模型中有什么作用？

8.2 虚拟变量为何只选 0、1，选 2、3、4 行吗？为什么？

8.3 对 (8.10) 式的模型，如果选择这样一个虚拟变量： $D = \begin{cases} 1 & \text{大专及大专以上} \\ 0 & \text{高中} \\ -1 & \text{高中以下} \end{cases}$ ，这样的设置方式隐含了什么假定？这一假定合理吗？

8.4 引入虚拟解释变量的两种基本方式是什么？它们各适用于什么情况？

8.5 四种加法方式引入虚拟变量会产生什么效应？

8.6 引入虚拟被解释变量的背景是什么？含有虚拟被解释变量模型的估计方法有哪

些？

8.7 设服装消费函数为

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

X_i —收入水平； Y_i — 年服装消费支出； $D_3 = \begin{cases} 1, \text{大专及大专以上} \\ 0, \text{其他} \end{cases}$ ； $D_2 = \begin{cases} 1, \text{女性} \\ 0, \text{男性} \end{cases}$ 。

试写出不同人群组的服装消费函数模型。

8.9 利用月度数据资料，为了检验下面的假设，应引入多少个虚拟解释变量？

- (1) 一年里的 12 个月全部表现出季节模式；
- (2) 只有 2 月、6 月、8 月、10 月和 12 月表现出季节模式。

练习题

8.1 Sen 和 Srivastava (1971) 在研究贫富国之间期望寿命的差异时，利用 101 个国家的数据，建立了如下的回归模型：

$$\hat{Y}_i = -2.40 + 9.39 \ln X_i - 3.36(D_i(\ln X_i - 7))$$

$$(4.37) \quad (0.857) \quad (2.42)$$

$$R^2 = 0.752$$

其中： X 是以美元计的人均收入；

Y 是以年计的期望寿命；

Sen 和 Srivastava 认为人均收入的临界值为 1097 美元 ($\ln 1097 = 7$)，若人均收入超过 1097 美元，则被认定为富国；若人均收入低于 1097 美元，被认定为贫穷国。

括号内的数值为对应参数估计值的 t -值。

- (1) 解释这些计算结果。
- (2) 回归方程中引入 $D_i(\ln X_i - 7)$ 的原因是什么？如何解释这个回归解释变量？
- (3) 如何对贫穷国进行回归？又如何对富国进行回归？
- (4) 从这个回归结果中可得到的一般结论是什么？

8.2 表中给出 1965—1970 年美国制造业利润和销售额的季度数据。假定利润不仅与销售额有关，而且和季度因素有关。要求：

- (1) 如果认为季度影响使利润平均值发生变异，应如何引入虚拟变量？
- (2) 如果认为季度影响使利润对销售额的变化率发生变异，应当如何引入虚拟变量？

(3) 如果认为上述两种情况都存在，又应当如何引入虚拟变量？

(4) 对上述三种情况分别估计利润模型，进行对比分析。

年份季度	利润 (Y)	销售额 (X)	年份季度	利润 (Y)	销售额 (X)
1965—1	10503	114862	1968—1	12539	148862
2	12092	123968	2	14849	153913
3	10834	123545	3	13203	155727
4	12201	131917	4	14947	168409
1966—1	12245	129911	1969—1	14151	162781
2	14001	140976	2	15949	176057
3	12213	137828	3	14024	172419
4	12820	145465	4	14315	183327
1967—1	11349	136989	1970—1	12381	170415
2	12615	145126	2	13991	181313
3	11014	141536	3	12174	176712
4	12730	151776	4	10985	180370

8.3 在统计学教材中，采用了方差分析方法分析了不同班次对劳动效率的影响，其样本数据为

早班	中班	晚班
34	49	39
37	47	40
35	51	42
33	48	39
33	50	41
35	51	42
36	51	40

试采用虚拟解释变量回归的方法对上述数据进行方差分析。

8.4 Joseph Cappelleri 基于 1961—1966 年的 200 只 Aa 级和 Baa 级债券的数据（截面数据和时间序列数据的合并数据），分别建立了 LPM 和 Logit 模型：

$$\text{LPM} \quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

$$\text{Logit} \quad Li = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

其中： $Y_i=1$ 债券信用等级为 Aa（穆迪信用等级）

$Y_i=0$ 债券信用等级为 Baa（穆迪信用等级）

X_2 =债券的资本化率，作为杠杆的测度（= $\frac{\text{长期债券的市值}}{\text{总资本的市值}} \times 100$ ）

X_3 = 利润率（= $\frac{\text{税后收入}}{\text{总资产净值}} \times 100$ ）

X_4 = 利润率的标准差，测度利润率的变异性

X_5 = 总资产净值，测度规模

上述模型中 β_2 和 β_4 事先期望为负值，而 β_3 和 β_5 期望为正值（为什么）。

对于 LPM，Cappelleri 经过异方差和一阶自相关校正，得到以下结果：

$$\hat{Y}_i = 0.6860 - 0.0179 X_{2i}^2 + 0.0486 X_{3i} + 0.0572 X_{4i} + 0.378 \times 10^{-7} X_{5i}$$

$$Se = (0.1775)(0.0024) \quad (0.0486) \quad (0.0178) \quad (0.039 \times 10^{-8})$$

$$R^2 = 0.6933$$

对于 Logit 模型，Cappelleri 在没有对异方差进行弥补的情形下用 ML 得以下结果：

$$\ln\left(\frac{p_i}{1-p_i}\right) = -1.6622 - 0.3185 X_{2i}^2 + 0.6248 X_{3i} - 0.9041 X_{4i} + 0.92 \times 10^{-6} X_{5i}$$

试解下列问题：

（1）为什么要事先期望 β_2 和 β_4 为负值？

（2）在 LPM 中，当 $\beta_4 > 0$ 是否合理？

（3）对 LPM 的估计结果应做什么样的解释？

（4）已知 $X_2^2 = 9.67\%$ ， $X_3 = 7.77\%$ ， $X_4 = 0.5933\%$ ， $X_5 = 3429$ （千元），问债券晋升 Aa 信用等级的概率有多大？

8.5 Greene 在分析讲授某门经济学课程采用新的教学方法效应时，搜集了如下表所示的

数据，其中，Grade 是学生在接受新教学方法（PSI， $PSI = \begin{cases} 1 & \text{接受新教学方法} \\ 0 & \text{没有采用新方法} \end{cases}$ ）后学习

成绩是否有所提高的虚拟变量， $GRADE = \begin{cases} 1 & \text{有所提高} \\ 0 & \text{没有提高} \end{cases}$ ，其他变量分别为平均级点 GPA，

非期末考试成绩分数 TUCE。试用 Logit 模型对此进行估计，并分析相应的边际效应。

obs	GRADE	GPA	TUCE	PSI	obs	GRADE	GPA	TUCE	PSI
1	0.000000	2.660000	20.00000	0.000000	17	0.000000	2.750000	25.00000	0.000000
2	0.000000	2.890000	22.00000	0.000000	18	0.000000	2.830000	19.00000	0.000000
3	0.000000	3.280000	24.00000	0.000000	19	0.000000	3.120000	23.00000	1.000000
4	0.000000	2.920000	12.00000	0.000000	20	1.000000	3.160000	25.00000	1.000000
5	1.000000	4.000000	21.00000	0.000000	21	0.000000	2.060000	22.00000	1.000000
6	0.000000	2.860000	17.00000	0.000000	22	1.000000	3.620000	28.00000	1.000000
7	0.000000	2.760000	17.00000	0.000000	23	0.000000	2.890000	14.00000	1.000000
8	0.000000	2.870000	21.00000	0.000000	24	0.000000	3.510000	26.00000	1.000000
9	0.000000	3.030000	25.00000	0.000000	25	1.000000	3.540000	24.00000	1.000000
10	1.000000	3.920000	29.00000	0.000000	26	1.000000	2.830000	27.00000	1.000000
11	0.000000	2.630000	20.00000	0.000000	27	1.000000	3.390000	17.00000	1.000000
12	0.000000	3.320000	23.00000	0.000000	28	0.000000	2.670000	24.00000	1.000000
13	0.000000	3.570000	23.00000	0.000000	29	1.000000	3.650000	21.00000	1.000000
14	1.000000	3.260000	25.00000	0.000000	30	1.000000	4.000000	23.00000	1.000000
15	0.000000	3.530000	26.00000	0.000000	31	0.000000	3.100000	21.00000	1.000000
16	0.000000	2.740000	19.00000	0.000000	32	1.000000	2.390000	19.00000	1.000000

8.6 依据下列大型超市的调查数据，分析股份制因素是否对销售规模产生影响。

销 售 规模	性质	销 售 规模	性质	销 售 规模	性质	销 售 规模	性质	销 售 规模	性质
1345	非股份制	1566	非股份制	2533	股份制	1144	非股份制	1461	非股份制
2435	股份制	1187	非股份制	1602	非股份制	1566	股份制	1433	股份制

1715	股份制	1345	非股份制	1839	非股份制	1496	股份制	2115	非股份制
1461	股份制	1345	非股份制	2218	股份制	1234	非股份制	1839	股份制
1639	股份制	2167	股份制	1529	非股份制	1345	非股份制	1288	股份制
1345	非股份制	1402	股份制	1461	股份制	1345	非股份制	1288	非股份制
1602	非股份制	2115	股份制	3307	股份制	3389	股份制	1345	非股份制
1839	股份制	2218	股份制	3833	股份制	981	股份制	1839	非股份制
2365	非股份制	3575	股份制	1839	股份制	1345	非股份制	2613	股份制
1234	非股份制	1972	股份制	1926	股份制	2165	非股份制		

*第九章 设定误差与测量误差¹

引子:

简单一定胜于复杂吗?

西方国家盛行“Occam's razor”原则²,意思是“简单优于复杂”的节约性原则。经济模型永远无法完全把握现实,在建立模型中一定的抽象和简化是不可避免的。

在研究进口数量时,分析进口(IM)与国内生产总值(GDP)、汇率(EX)的关系,建立并估计了以下模型

$$\begin{aligned} \hat{IM}_t = & -1159.179 + 1.142897GDP_t - 0.815842GDP_{t-1} - 0.022569EX_t^2 \\ t = & (-2.268276) \quad (7.71607) \quad (-5.66842) \quad (-6.857844) \end{aligned} \quad (1)$$

$$R^2 = 0.978378 \quad \bar{R}^2 = 0.974965 \quad DW=2.047965 \quad F=286.5846$$

如果根据“简单优于复杂”的原则,直接分析进口与国内生产总值的关系,得到回归结果

$$\begin{aligned} IM_t = & -1067.337 + 0.2307GDP_t + e_t \\ t = & (-2.0288) \quad (16.2378) \end{aligned} \quad (2)$$

$$R^2 = 0.9230 \quad \bar{R}^2 = 0.9195 \quad DW=0.5357 \quad F=263.6657$$

这两个方程的t检验和F检验结果显示都显著,方程(2)中GDP的t检验值还优于方程(1),而且方程(2)函数形式也更为简单。能否根据“Occam's razor”原则,判断简单的方程(2)比复杂的方程(1)更好呢?

对模型的设定是计量经济研究的重要环节。所设定的模型要求正确地描述被解释变量与解释变量之间的真实关系,在第二章提出线性回归模型的基本假定时,除了对随机扰动项 u_i 分布的假定以外,也强调了假定模型对变量和函数形式的设定是正确的,假定模型中的变量没有测量误差。但是在实际的建模实践中,对模型的设定不一定能够完全满足这样的要求,从而会使模型出现设定误差。本章以OLS估计为基础,分别讨论模型设定误差的后果以及检验方法。

¹ 本章内容本科教学供选择

² 见古扎拉蒂《计量经济学》下册第447页,中国人民大学出版社,2000

第一节 设定误差

一、设定误差的类型

计量经济模型是对变量间经济关系因果性的设想，若所设定的回归模型是“正确”的，主要任务是所选模型参数的估计和假设检验。若检验统计量 R^2 , t , F 和 DW 等在统计意义上是显著的，则模型的建模过程结束。反之，若这些统计量中的一个或多个不显著，我们就会去寻找其他的估计方法进行参数估计和检验，例如，在加权和广义差分的基础上用最小二乘法解决异方差性或自相关性问题。但是如果对计量模型的各种诊断或检验仍不能令人满意，这时就应把注意力集中到模型的设定方面，考虑所建模型是否遗漏了重要的变量？是否包含了多余的变量？所选模型的函数形式是否正确？随机扰动项的设定是否合理？关于被解释变量和解释变量的数据收集是否有误差？等等。所有这些，在计量经济学中被统称为设定误差。

从误差来源看，设定误差主要包括：（1）变量的设定误差，包括相关变量的遗漏（欠拟合）、无关变量的误选（过拟合）；（2）变量数据的测量误差；（3）模型函数形式的设定误差；（4）随机扰动项设定误差。本章主要讨论前两类设定误差。

出现设定误差的原因是多方面的。首先，数据来源渠道可能不畅。在建模过程中，尽管某个变量有着重要的经济意义和计量经济学解释作用，但这个变量的数据很难取得，而被迫将该变量排斥在模型之外，例如消费行为分析中消费者财富的变量就是例证。其次，虽然知道模型中应当包含哪些变量，但却不知道这些变量应当以什么确切的函数形式出现在回归模型中。也就是说，经济管理的基本理论并没有提示模型中变量的准确函数形式。例如，经济学理论不会肯定消费水平与有关变量的关系是线性的还是对数线性的，或者是两者的某种混合形式的。最后，更为重要的是，事实上我们事先并不知道所研究的实证数据中所隐含的真实模型究竟是什么。正是上述这些原因，设定误差在建模中是较容易出现。设定误差的存在可能会对模型形成不良的后果。

二、变量设定误差的后果

变量设定误差主要有两类：一类是相关变量的遗漏，也称为模型“欠拟合”；另一类是无关变量的误选，也称为模型“过拟合”。从实质上看，变量设定误差的主要后果，是一个或多个解释变量与随机扰动项之间存在着相关性，而影响参数估计的统计特性。

1、遗漏相关变量（欠拟合）的偏误

采用遗漏了重要解释变量的模型进行估计而带来的偏误，称为遗漏相关变量偏误。

如果正确的模型应当为：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (9.1)$$

其离差形式为 $y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u})$ (9.2)

但是由于某种原因，设定模型时将变量 X_{3i} 遗漏了，实际采用的回归模型为：

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \quad (9.3)$$

假定其他有关线性模型的古典假设都成立，则 (9.3) 式中 α_2 的 OLS 估计式为：

$$\hat{a}_2 = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2} \quad (9.4)$$

将正确模型的离差形式 (9.2) 式代入 (9.4) 式，得：

$$\begin{aligned} \hat{a}_2 &= \frac{\sum x_{2i} [\beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u})]}{\sum x_{2i}^2} \\ &= \frac{\sum \beta_2 x_{2i}^2 + \beta_3 \sum x_{2i} x_{3i} + \sum x_{2i} (u_i - \bar{u})}{\sum x_{2i}^2} \\ &= \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i} (u_i - \bar{u})}{\sum x_{2i}^2} \end{aligned} \quad (9.5)$$

对 (9.5) 式两边取期望，有：

$$E(\hat{\alpha}_2) = E\left(\beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i} (u_i - \bar{u})}{\sum x_{2i}^2}\right) \quad (9.6)$$

当样本容量无限增大时，观察 $\hat{\alpha}_2$ 的概率极限性质，对 (9.5) 式两边取概率极限：（证明见附录 9.1）

$$p \lim_{n \rightarrow \infty} \hat{\alpha}_2 = \beta_2 + \beta_3 \frac{Cov(X_{2i}, X_{3i})}{Var(X_{2i})} + \frac{Cov(X_{2i}, u_i)}{Var(X_{2i})} \quad (9.7)$$

由此可以看出， X_3 的遗漏将产生如下后果：

(1) 如果漏掉的 X_3 与 X_2 相关，则参数 $\hat{\alpha}_1$ 和 $\hat{\alpha}_2$ 将是有偏且不一致性的，即 $E(\hat{\alpha}_1) \neq \beta_1$ ，

$E(\hat{\alpha}_2) \neq \beta_2$ ，且 $p \lim_{n \rightarrow \infty}(\hat{\alpha}_1) \neq \beta_1$ ， $p \lim_{n \rightarrow \infty}(\hat{\alpha}_2) \neq \beta_2$ 。

这是由于(9.3)式中 $v_i = \beta_3 X_{3i} + u_i$ ，所以

$$\text{Cov}(v_i, X_{2i}) = \text{Cov}(\beta_3 X_{3i} + u_i, X_{2i}) = \text{Cov}(\beta_3 X_{3i}, X_{2i}) + \text{Cov}(u_i, X_{2i}) \quad (9.8)$$

(9.8)式中，虽然 $\text{Cov}(u_i, X_{2i}) = 0$ ，但 $\text{Cov}(\beta_3 X_{3i}, X_{2i}) = \beta_3 \text{Cov}(X_{3i}, X_{2i}) \neq 0$ 。在小样本下，

(9.6)式中的第二项求期望不会为零，表明 OLS 估计量在小样本下有偏。在大样本下，(9.7)

第二项中的 $\frac{1}{n} \sum x_{2i} x_{3i}$ 也不会随着样本的增大而趋于零，表明 OLS 估计量在大样本下非一

致，即有 $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_2 \neq \beta_2$ 。因此，如果漏掉的 X_3 与 X_2 相关，OLS 估计量在大样本下也是非

一致的。

(2) 若 X_3 与 X_2 不相关，即 $\sum x_{2i} x_{3i} = 0$ ， $\hat{\alpha}_2$ 满足无偏性和一致性，但可以证明这时截距项的估计 $\hat{\alpha}_1$ 却是有偏的（证明从略）。

(3) $\hat{\alpha}_2$ 的方差是 $\hat{\beta}_2$ 方差的有偏估计：

对于 (9.3) 式，已知

$$\text{Var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \quad (\text{见 2.40})$$

而对于(9.1)式，有（见 4.14）

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2})} = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad (9.9)$$

如第三章所讨论的， $\text{Var}(\hat{\beta}_2)$ 是 β_2 方差的无偏估计，而如果漏掉的 X_3 与 X_2 相关， $r_{23}^2 \neq 0$ ，

$\text{Var}(\hat{\alpha}_2) \neq \text{Var}(\hat{\beta}_2)$ ，故 $\text{Var}(\hat{\alpha}_2)$ 是有偏的。

(4) 漏掉 X_3 的 (9.3) 式中的随机扰动项 v_i 的方差估计量 $\hat{\sigma}_v^2 = \text{RSS}_v / (n - 2)$ 将是有偏的，即 $E(\hat{\sigma}_v^2) \neq \sigma_u^2$ ；

(5) 与方差相关的检验，包括假设检验、区间估计等，在关于参数的统计显著性方面，都容易导出错误的结论。

对从模型中遗漏变量时参数估计性质的认识，还有两点要特别注意：

(1) 若 X_3 与 X_2 相关， $r_{23}^2 \neq 0$ ，显然 $\text{Var}(\hat{\alpha}_2) \neq \text{Var}(\hat{\beta}_2)$ ，由(4.14)式可看出似乎有

$Var(\hat{\alpha}_2) < Var(\hat{\beta}_2)$ 。但实际情形并不完全如此。可以注意到, (9.1) 和 (9.3) 的剩余平方和 RSS 是不一样的, 其自由度也是不等的。在样本容量相同的条件下 $RSS_v/(n-2) \neq RSS_u/(n-3)$, 或 $\hat{\sigma}_v^2 \neq \hat{\sigma}_u^2$ 。因此, 有可能从 (9.3) 式回归得到的 $RSS_v/(n-2)$ 大于从 (9.1) 式回归得到的 $RSS_u/(n-3)$ 。

(2) 若 X_3 与 X_2 不相关, 有 $r_{23}^2 = 0$ 和 $\sum x_{2i}x_{3i}/\sum x_{2i}^2 = 0$, 由 (9.6) 和 (4.14), 似乎分别有 $E(\hat{\alpha}_2) = \beta_2$, $Var(\hat{\beta}_2) = Var(\hat{\alpha}_2)$ 。若这两个等式成立, 意味着尽管变量 X_3 在理论上分析是有关的变量, 但从所选模型中略去似乎也不会导致什么危害。这种认识实际也不正确。因为 $\widehat{Var}(\hat{\alpha}_2) = \frac{\hat{\sigma}_v^2}{\sum x_{2i}^2} = \frac{RSS_v/n-2}{\sum x_{2i}^2}$, 为 $\widehat{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}_u^2}{\sum x_{2i}^2} = \frac{RSS_u/n-3}{\sum x_{2i}^2}$ 的有偏估计, 即使 X_3 与 X_2 不相关, 也有 $\widehat{Var}(\hat{\beta}_2) \neq \widehat{Var}(\hat{\alpha}_2)$, 致使假设检验程序很有可能是可疑的。况且, 在大多数的实证经济研究中, X_3 与 X_2 通常都是相关的, 更可能会产生上述后果。因此必须清楚, 一旦根据相关理论把模型建立起来, 再从中遗漏变量需要充分地谨慎。

2、包含无关变量(过拟合)的偏误

模型中包括了不重要的解释变量, 即采用误选了无关解释变量的模型进行估计而带来的偏误, 称为包含无关变量偏误。

为讨论方程中包含了无关变量的情形, 假设正确的模型是:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad (9.10)$$

而回归模型加入了无关变量 X_3 , 被设定为:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i \quad (9.11)$$

可将(9.10)式视为以 $\alpha_3 = 0$ 为约束的(9.11)式的特殊形式。采用 OLS 法对式 (9.11) 进行参数估计, 由 (3.27)式有:

$$\hat{\alpha}_2 = \frac{\sum x_{2i}y_i \sum x_{3i}^2 - \sum x_{3i}y_i \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} \sum x_{3i})^2} \quad (9.12)$$

将 (9.10) 式的离差形式 $y_i = \beta_2 x_{2i} + (u_i - \bar{u})$ 代入 (9.12) 式, 并整理, 得:

$$\hat{\alpha}_2 = \beta_2 + \frac{(\sum x_{3i}^2)(\sum x_{2i}(u_i - \bar{u})) - (\sum x_{2i}x_{3i})(\sum x_{3i}(u_i - \bar{u}))}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} \sum x_{3i})^2} \quad (9.13)$$

当 X_2 与 X_3 为非随机时，对上式求数学期望，得

$$E(\hat{\alpha}_2) = \beta_2$$

其方差为

$$Var(\hat{\alpha}_2) = \frac{\sigma_v^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad (9.14)$$

由以上可知，无关变量的设定误差的后果为：

(1) 可以证明，(9.11) 式参数的 OLS 估计量是无偏的，且为一致性估计量。即：

$E(\hat{\alpha}_2) = \beta_2$ ， $\lim_{n \rightarrow \infty} \hat{\alpha}_2 = \beta_2$ 。同理，可证明 $E(\hat{\alpha}_1) = \beta_1$ ， $E(\hat{\alpha}_3) = \beta_3 = 0$ ； $\lim_{n \rightarrow \infty} \hat{\alpha}_1 = \beta_1$ 和 $\lim_{n \rightarrow \infty} \hat{\alpha}_3 = \beta_3 = 0$ 。其中，参数 $\hat{\alpha}_2$ 一致性的证明见本章附录 9.2。

(2) $\hat{\alpha}_2$ 不是有效估计量。

因为 β_2 的方差为 $\frac{\sigma^2}{\sum x_{2i}^2}$ ，那么：

$$\frac{Var(\hat{\alpha}_2)}{Var(\hat{\beta}_2)} = \frac{1}{(1 - r_{23}^2)} \quad (9.15)$$

虽然变量 X_3 对被解释变量 Y 是无关的，但解释变量 X_3 与 X_2 之间很可能一定程度相关，即 $0 \leq r_{23}^2 \leq 1$ ，则 $Var(\hat{\alpha}_2) \geq Var(\hat{\beta}_2)$ 。这表明，无关变量 X_3 的误选，会使得 $\hat{\alpha}_2$ 的方差增大，导致 $\hat{\alpha}_2$ 的估计精度下降，且偏离程度随着解释变量间相关程度的增加而增大。此结论对 $\hat{\alpha}_1$ 也成立；

(3) $E(\hat{\sigma}_v^2) = \sigma_v^2$ ，即随机误差项的方差的估计仍为无偏估计；

(4) 通常的区间估计和假设检验程序依然有效，但 $\hat{\alpha}_2$ 的方差增大，接受错误假设的概率会较高。

比较遗漏相关变量和误选无关变量两类设定误差可以看出，如果遗漏了相关变量，将导致参数估计量和假设检验是有偏的，且为不一致的；如果误选了无关变量，虽然参数估计量具有无偏性、一致性，又会损失参数估计量的有效性。由于事先并不可能清楚地知道隐含在数据中的真实数量关系，建模过程中将面临如何选择更为恰当变量的两难境地。若是主要注重检验的无偏性、一致性，那么可能会宁愿误选无关变量也不愿遗漏相关变量；若是主要注重估计量的有效性，一般的选择则是宁愿删除相关变量。通常误选无关变量不如遗漏相关变

量的后果严重。因此，一定程度上模型的设定实际是对偏误与有效进行权衡，偏爱哪一方取决于模型的研究目的。若建模目的只是为了进行预测，最小均方误差则可能是兼顾有效性和无偏性的良好准则。

均方误差（简记作 MSE）是参数估计值 β^* 与参数真实值 β 离差平方的期望

$$MSE(\beta^*) = E(\beta^* - \beta)^2 \quad (9.16)$$

容易证明，均方误差与方差有如下关系：

$$E(\beta^* - \beta)^2 = E[\beta^* - E(\beta^*)]^2 + [E(\beta^*) - \beta]^2 \quad (9.17)$$

均方误差 $E(\beta^* - \beta)^2$ 是方差 $E[\beta^* - E(\beta^*)]^2$ 与偏倚的平方 $[E(\beta^*) - \beta]^2$ 之和，包含了两个方面的因素。当在较小偏倚（或无偏性）和较小方差（或最小方差性）“二者不可得兼”时，需要进行“权衡与折衷”，可用均方误差准则。

第二节 设定误差的检验

相关变量的遗漏和无关变量的误选，在不同程度上给模型的设定形成了不良影响，有必要对变量设定误差进行检验。当然，这种假设检验必须在经济理论指导下进行，不可抛弃经济理论而进行假设检验。对于是否误选无关变量的检验，只要针对无关变量系数的期望值为零的假设，用 t 检验或 F 检验，对无关变量系数作显著性检验即可。对于遗漏变量设定误差的检验有多种方法，例如 DW 检验、拉格朗日乘数检验（Lagrange Multiplier, LM）、豪斯曼检验（Hausman-test）、RESET 一般性检验等。这里只讨论设定误差的一些最常用的检验方法。

一、DW 检验

用 DW 检验去检验是否遗漏相关变量，其基本思想是认为遗漏的相关变量应包含在随机扰动项中，那么回归所得的残差序列就会呈现单侧的正（负）相关性，因此可从自相关性的角度检验相关变量的遗漏。

从遗漏变量的模型看，可以认为遗漏变量模型是无遗漏变量模型的一个特例：被遗漏变量的系数为 0。例如，式（9.3）是式（9.1）中变量 X_{3i} 的系数为 0。我们称（9.1）为无约束回归模型，而（9.3）为受约束回归模型。

DW 检验的具体步骤如下：

1. 对回归模型运用 OLS 法得残差序列 e_i 。

2. 设定 H_0 :受约束回归模型, H_1 :无约束回归模型。按遗漏解释变量的递增次序对残差序列 e_i 进行排序, 对排序后的残差序列 e_i 计算 d 统计量

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (9.25)$$

3. 查 Durbin-Watson 表, 若 d 为显著, 则拒绝原假设, 受约束回归模型不成立, 存在模型设定误差, 否则接受原假设, 受约束回归模型成立, 模型无设定误差。

例如, 对表 7.1 的数据设定总生产成本函数, 准备使用如下的三个备选模型:

$$(1) Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$$

$$(2) Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2$$

$$(3) Y_i = \beta_1 + \beta_2 X_i$$

其中只有 (1) 为真实模型, 试用 DW 法检验模型设定误差。

表 9.2 总成本 (Y) 和产出(X)数据

	总成本 (Y)	产出 (X)
1	193	1
2	226	2
3	240	3
4	244	4
5	257	5
6	260	6
7	274	7
8	297	8
9	350	9
10	420	10

首先, 对上述三个模型分别代入数据回归得:

$$(1) \hat{Y}_i = 141.767 + 63.478 X_i - 12.962 X_i^2 + 0.939 X_i^3$$

$$\begin{array}{cccc}(6.375) & (4.778) & (0.9856) & (0.0592) \\ (22.238) & (13.285) & (-13.151) & (15.861) \\ R^2=0.9983 & \bar{R}^2=0.9975 & & DW=2.70\end{array}$$

$$(2) \hat{Y}_i = 222.383 - 8.0250 X_i + 2.542 X_i^2$$

$$\begin{array}{ccc}(23.488) & (9.809) & (0.869) \\ (9.468) & (-0.818) & (2.925) \\ R^2=0.9284 & \bar{R}^2=0.9079 & DW=1.038\end{array}$$

$$(3) \hat{Y}_i = 166.467 + 19.933 X_i$$

$$\begin{array}{ccc}(19.201) & (3.066) & \\ (8.752) & (6.502) & \\ R^2=0.8409 & \bar{R}^2=0.8210 & DW=0.716\end{array}$$

由于本例中，遗漏变量已经按递增次序排列，此时的 DW 值等于 d 值，无需重新计算 d 统计量。对上述模型的 DW 统计量的分析及查表情况如下：

对于模型（1）有 $DW=2.70$ ，当 $n=10$ 、 $k'=3$ 、 $\alpha=5\%$ 时， $d_L=0.525$ ， $d_U=2.016$ ，不能表明存在显著的正相关关系，接受 H_0 ，表示没有遗漏的变量。

对于模型（3）有 $DW=0.716$ ，当 $n=10$ 、 $k'=1$ 、 $\alpha=5\%$ 时， $d_L=0.879$ ， $d_U=1.320$ ，显然存在正的自相关，拒绝 H_0 ，表明存在遗漏变量；

对于模型（2），计算结果有 $n=10$ ， $DW=1.038$ ，那么，当 $n=10$ ， $k'=2$ ， $\alpha=5\%$ 时， $d_L=0.697$ ， $d_U=1.641$ ，显然有 $0.697 < 1.038 < 1.641$ ，属于无法确定的区域。这时，可采用修正的 DW 检验法进行检验，即扩大拒绝区域，可依据 $DW = 1.038 < d_U = 1.641$ ，宁可判别残差中存在正的自相关，认为也存在遗漏变量。

二、拉格朗日乘数（LM）检验

拉格朗日乘数检验的基本思想，是认为模型中遗漏的相关变量包含在随机扰动项中，因此随机扰动项或回归所得的残差序列应与遗漏的相关变量呈现出某种依存关系，可以进行残差序列与相关变量的回归，在一定显著水平下若相关变量具有统计显著性，则认为存在遗漏变量形成的设定偏误，若相关变量不具有统计显著性，则认为没有遗漏变量形成的设定误差。

拉格朗日乘数检验的具体步骤如下：

- 1、对存在遗漏变量设定偏误的模型（受约束回归模型）进行回归，得残差序列 e_i ；
- 2、用残差序列 e_i 对全部的解释变量（包括遗漏变量）进行回归，得可决系数 R^2 ；
- 3、设定 H_0 : 受约束回归模型， H_1 : 无约束回归模型。在大样本情况下，构造检验统计量 nR^2 ，恩格尔（Engle）曾经证明，

$$nR^2 \stackrel{asy}{\sim} \chi^2(\text{约束个数}) \quad (9.26)$$

其中：“asy”（asymptotically）表示“渐近地”；约束个数是 H_0 中设定的受约束个数。

- 4、进行显著性检验的判断：若 $nR^2 > \chi^2_\alpha(\text{约束个数})$ ，则拒绝 H_0 ，认为受约束模型不成立，存在遗漏变量；否则，接受 H_0 ，认为受约束模型成立，进而无遗漏变量。

*三、一般性检验（RESET）³

RESET 检验（regression error specification test）是拉姆齐(Ramsey)于 1969 年提出的一种检验方法。其检验的基本思想为：如果事先知道遗漏了哪个变量，只需将此变量引入模型，估计并检验其参数是否显著不为零即可，可是问题是并不知道遗漏了哪个变量，这时可寻找一个替代变量 Z 来进行上述检验。RESET 检验中，替代变量 Z 通常选用所设定模型被解释变量拟合值 \hat{Y} 若干次幂的线性组合。若模型估计所得的残差包含着遗漏的相关变量，那么这个残差可用被解释变量拟合值的线性组合近似表示；若这个线性组合是显著的，则认为原模型的设定有误。由于可引入若干个替代变量去判断是否有多个变量被遗漏，所以该方法被称为一般性设定偏误检验。

RESET 检验的基本步骤为：

第 1 步：对模型进行回归，用 OLS 法估计

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

分别得到 Y_i 的拟合值 \hat{Y}_i 和残差 e_i 。若残差 e_i 与拟合值 \hat{Y}_i 之间存在某种函数关系，则可用拟合值 \hat{Y}_i 若干次幂的线性组合充当工具变量；

第 2 步：用被解释变量 Y_i 的拟合值 \hat{Y}_i 的线性组合，测度残差中是否包含着遗漏的相关变量。具体做法为，在第 1 步的模型中增加一个包含拟合值 \hat{Y}_i 的函数。这个函数通常选择为

³ 这部分内容本科教学供选择

拟合值 \hat{Y}_i 的平方、立方和四次方的线性组合。例如：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \delta_1 \hat{Y}_i^2 + \delta_2 \hat{Y}_i^3 + \delta_3 \hat{Y}_i^4 + v_i \quad (9.36)$$

并对上述模型进行估计；

第 3 步：构造原假设： $H_0: \delta_j = 0, j=1,2,3$ 。然后用 F 统计量进行检验。F 检验统计量为

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/n - (k+J)} = \frac{(R_U^2 - R_R^2)/J}{(1 - R_U^2)/n - (k+J)} \quad (9.37)$$

即

$$F = \frac{(RSS_R - RSS_U)/3}{RSS_U/n - (k+3)} = \frac{(R_U^2 - R_R^2)/3}{(1 - R_U^2)/n - (k+3)} \quad (9.38)$$

其中， RSS_U 和 R_U^2 分别为对方程(9.36)进行回归得到的残差平方和与拟合优度， RSS_R 和 R_R^2 分别为当原假设： $H_0: \delta_j = 0, j=1,2,3$ 成立时，对方程(9.36)进行回归得到的残差平方和与拟合优度，J 为约束条件的个数。

若 F 统计值大于 F 临界值，则拒绝原假设，表明存在某种形式的设定误差问题。

第三节 测量误差

经济计量研究中需要运用大量的观测数据，在搜集相关的数据时，经常遇到所搜集的数据不能确实地反映变量间经济行为的情况。在计量经济模型中使用了经济变量不准确的数据时，则称模型中包含了测量误差。测量误差将会影响计量经济分析的结果。

一、模型变量的测量误差

测量误差指在收集数据过程中的登记误差、在数据加工整理过程中的整理误差以及其他统计误差。计量经济研究中运用的观测数据出现测量误差，原因是多方面的。首先，受人为因素和技术因素的影响，对经济现象和过程的调查登记本身就可能产生误差，例如虚报和误解指标含义而产生的统计误差；其次，数据的加工处理过程中也可能导致一定的误差，例如错误的汇总或分组导致的偏差，又如经过修匀加工的数据与实际情况的偏差；此外，数据的不当使用也会出现误差，例如错误地理解和运用了不同内涵、不同范围、不同计量单位的数据。可以把这些有关数据的误差统称为“测量误差”。测量误差可能是被解释变量的测量误差，也可能是解释变量的测量误差。

为了说明测量误差的后果，设正确的回归模型为

$$Y_i^* = \alpha + \beta^* X_i^* + u_i \quad (9.39)$$

其中： Y_i^* 为被解释变量的理论真实值； X_i^* 为解释变量的理论真实值，且 Y_i^* 和 X_i^* 都是不可直接测量的，而只能通过下列测量过程得到其样本数据：

$$Y_i = Y_i^* + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad (9.40)$$

$$X_i = X_i^* + \omega_i \quad \omega_i \sim N(0, \sigma_\omega^2) \quad (9.41)$$

$$\text{且：} \quad \text{Cov}(\varepsilon_i, \omega_i) = 0 \quad \text{Cov}(\varepsilon_i, X_i) = 0$$

其中： Y_i 是 Y_i^* 的测量结果； ε_i 是 Y_i^* 的测量误差； X_i 是 X_i^* 的测量结果； ω_i 是 X_i^* 的测量误差； ε_i 与 ω_i 相互无关， ε_i 与 X_i 也无关，且各误差都没有序列相关。

用观测到的样本数据进行回归时，等价于对下式回归：

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - \omega_i) + u_i + \varepsilon_i \\ &= \alpha + \beta X_i + u_i + \varepsilon_i - \beta \omega_i \end{aligned} \quad (9.42)$$

将式 (9.39)、(9.40)、(9.41) 分别以离差形式表示：

$$y_i^* = \beta x_i^* + (u_i - \bar{u}) \quad (9.43)$$

$$y_i = y_i^* + (\varepsilon_i - \bar{\varepsilon}) \quad (9.44)$$

$$x_i = x_i^* + (\omega_i - \bar{\omega}) \quad (9.45)$$

对 (9.42) 采用 OLS 法，有

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

将 (9.44) 和 (9.45) 代入，并在大样本下，取概率极限得（推导过程见附录 9.3）

$$P \lim \hat{\beta} = \frac{\beta \text{Var} X_i^*}{\text{Var} X_i^* + \sigma_\omega^2} = \frac{\beta}{1 + \frac{\sigma_\omega^2}{\sigma_{X^*}^2}} \quad (9.46)$$

其中 σ_ω^2 为 (9.41) 式中 ω 的方差； $\sigma_{X^*}^2$ 为 X_i^* 的方差。因为 (9.46) 式中 $\sigma_\omega^2 / \sigma_{X^*}^2 > 0$ ，这表明当测量误差存在时，OLS 法常常会低估真实的回归参数。

值得指出的是，回归变量中的测量误差是数据问题，目前计量经济学家们还提不出有效的解决方法。一般的做法往往是忽略测量误差问题，主观上希望测量误差足够小，从而不破坏计量的合理性。

二、测量误差的检验

测量误差的存在使得回归系数被低估，将直接影响计量经济分析的结果，因此有必要对是否存在测量误差进行检验。

关于测量误差存在与否的检验是豪斯曼（Hausman）1978 年提出的检验方法⁴，豪斯曼方法的具体步骤为：

- （1）对所研究的回归模型，无论是否存在测量误差，先采用 OLS 法得到参数估计量；
- （2）对可能存在测量误差的解释变量，选择与其相关的工具变量，将可能存在测量误差的解释变量对选择的工具变量进行回归，并获得回归残差 ω ；
- （3）将回归残差 ω 加入第（1）步中的回归表达式，再次进行 OLS 估计，得 ω 的参数估计值 $\hat{\beta}_\omega$ 及假设检验结果；
- （4）若 $\hat{\beta}_\omega$ 为显著时，则认为解释变量的确存在观测误差，反之，认为解释变量不存在测量误差。

现以一个例子说明上述检验步骤：

例 7.2 利用观测到的样本数据作回归，已得到以下结果：

$$EXP = -46.81 + 0.00324AID + 0.00019INC - 0.597POP \quad (9.47)$$

$$t = (-0.56) \quad (13.64) \quad (8.12) \quad (-5.17)$$

$$R^2 = 0.993 \quad F = 2190$$

其中：EXP 为某贫困地区地方政府的支出；AID 为中央政府的拨款量；INC 为贫困地区地方政府的财政收入；POP 为该地区的总人口数。现怀疑中央政府的拨款量 AID 存在测量误差。现选择工具变量 PS（PS 为贫困人口数），其原因为扶贫支出是该地区地方政府支出中比重最大的支出，其经费来源主要是依赖中央政府的拨款，因此 PS 与 AID 有较高相关性。将 AID 对 PS 进行回归，得到如下的残差变量 $\hat{\omega}$ ：

$$\hat{\omega}_i = AID - (77.95 + 0.845PS) \quad (9.48)$$

⁴ J.A.Hausman: "Specification Tests in Econometrics", Econometrics, vol.46, pp1251-1271, Nov.1978.

$$t = (-1.28) \quad (18.02) \quad R^2 = 0.87$$

将 $\hat{\omega}_i$ 项加入 (9.47)，再回归得到以下结果：

$$E\hat{X}P = -138.51 + 0.00174AID + 0.00018INC - 0.275POP + 1.372\hat{\omega}_i \quad (9.49)$$

$$t = (-1.41) \quad (1.94) \quad (7.55) \quad (-1.29) \quad (1.73)$$

从 (9.49) 看出，因为 $\hat{\omega}_i$ 系数的 t 值是 1.73(<1.96)，在 5% 的显著性水平下，双侧 t 检验接受原假设（不存在测量误差），但在 10% 的显著性水平上，双侧 t 检验则拒绝原假设而接受备择假设（存在测量误差）。

我们注意到，引进对测量误差可能性的修正，使 AID 变量的系数变小，这从另一个侧面说明，测量误差夸大了 AID 对 EXP 的影响。

第四节 案例分析

以引子中所提出的问题为例，分析影响中国进口量的主要因素（数据如表 9.3 所示）。

表 9.3

单位：人民币亿元、亿美元

年份	GDP	进口总额IM (人民币)	进口总额 IMdollar (美元)	汇率 EXCHANGE
1980	4517.8	298.8000	200.17	149.8400
1981	4862.4	375.3800	220.15	170.5100
1982	5294.7	364.9900	192.85	189.2600
1983	5934.5	422.6000	213.90	197.5700
1984	7171.0	637.8300	274.10	232.7000
1985	8964.4	1257.800	422.52	293.6600
1986	10202.20	1498.300	429.04	345.2800
1987	11962.50	1614.200	432.16	372.2100
1988	14928.30	2055.100	552.75	372.2100
1989	16909.20	2199.900	591.40	376.5100
1990	18547.90	2574.300	533.45	478.3200

1991	21617.80	3398.700	637.91	532.3300
1992	26638.10	4443.300	805.85	551.4600
1993	34634.40	5986.200	1039.59	576.2000
1994	46759.40	9960.100	1156.14	861.8700
1995	58478.10	11048.10	1320.84	835.1000
1996	67884.60	11557.40	1388.33	831.4200
1997	74462.60	11806.50	1423.70	828.9800
1998	78345.20	11626.10	1402.37	827.9100
1999	82067.50	13736.40	1656.99	827.8300
2000	89468.10	18638.80	2250.94	827.8400
2001	97314.80	20159.20	2435.53	827.7000
2002	105172.3	24430.30	2951.70	827.7000
2003	117251.9	34195.60	4127.60	827.7000

数据来源：《中国统计年鉴 2004》中国统计出版社

设定如下的模型。

$$IM_t = \alpha_1 + \alpha_2 GDP_t + u_t \quad (9.50)$$

其中， IM_t 是进口总额， GDP_t 是国内生产总值。

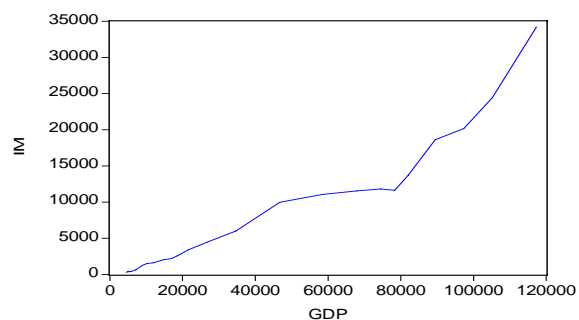
为了分析此模型是否有变量设定误差，进行变量设定误差检验。

有人认为，货物与服务的进口量受到一国的生产规模、货物与服务的进口价格、汇率等其他影响因素，而不能仅仅用 GDP 来解释商品进口的变化。因此，设定的回归模型应该为：

$$IM_t = \beta_1 + \beta_2 f(GDP_t) + \beta_3 g(Exchange_t) + u_t \quad (9.51)$$

其中：GDP 为国内生产总值， $f(GDP)$ 为 GDP 的线性函数，Exchange 为美元兑换人民币的汇率， $g(Exchange)$ 为 Exchange 的线性函数。如果是这样，显然设定的回归模型 (9.50) 式中可能遗漏了变量 GDP、Exchange 以及两者的线性组合。那么 GDP、Exchange 以及两者的线性组合是否被遗漏的重要变量呢？

依据表9.3的数据，录入到EViews响应的数据表中，考证 $IM=f(GDP)$ 基本关系图：



对 (9. 50) 进行回归, 有回归结果

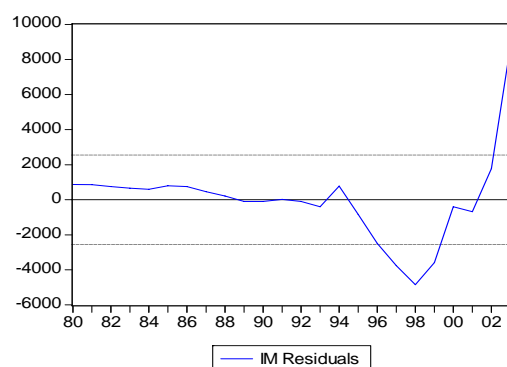
$$im_i = -1067.337 + 0.2307GDP_i + e_i$$

$$se = (792.2620) \quad (0.0142)$$

$$t = (-2.0288) \quad (16.2378)$$

$$R^2 = 0.9230 \quad \bar{R}^2 = 0.9195 \quad DW = 0.5357 \quad F = 263.6657$$

并作 (9. 50) 回归的残差图:



显然, 存在自相关现象, 其主要原因可能是建模时遗漏了重要的相关变量造成的。

1、DW 检验

模型 $im_i = -1067.337 + 0.2307GDP_i + e_i$ 的 DW 统计量表明, 存在正的自相关, 由于遗漏变量 **exchange** 或 **GDP** 已经按从小到大顺序排列, 因此, 无需重新计算 d 统计量。对 $n=24$ 和 $k'=1$, 5% 的德宾 - 沃森 d- 统计量的临界值为 $d_L = 1.273$ 和 $d_U = 1.446$, $0.5357 < d_L = 1.273$, 表明存在显著的遗漏变量现象。

为此, 进行如下的校正:

Dependent Variable: IM

Method: Least Squares

Date: 07/08/05 Time: 15:40

Sample (adjusted): 1981 2003

Included observations: 23 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-224.3632	1892.132	-0.118577	0.9069
GDP	1.148259	0.151433	7.582606	0.0000
GDP(-1)	-0.822444	0.147359	-5.581213	0.0000
EXCHANGE	-4.290746	8.348744	-0.513939	0.6135
EXCHANGE^2	-0.018637	0.008353	-2.231162	0.0386
R-squared	0.978691	Mean dependent var	8434.222	
Adjusted R-squared	0.973956	S.D. dependent var	9025.326	
S.E. of regression	1456.525	Akaike info criterion	17.59515	
Sum squared resid	38186370	Schwarz criterion	17.84200	
Log likelihood	-197.3443	F-statistic	206.6799	
Durbin-Watson stat	1.962659	Prob(F-statistic)	0.000000	

其中，exchange 的系数在统计意义上不显著，可以剔除，则有：

Dependent Variable: IM

Method: Least Squares

Date: 07/08/05 Time: 15:43

Sample (adjusted): 1981 2003

Included observations: 23 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1159.179	511.0396	-2.268276	0.0352
GDP	1.142897	0.148119	7.716070	0.0000
GDP(-1)	-0.815842	0.143928	-5.668420	0.0000

EXCHANGE^2	-0.022569	0.003291	-6.857844	0.0000
<hr/>				
R-squared	0.978378	Mean dependent var	8434.222	
Adjusted R-squared	0.974965	S.D. dependent var	9025.326	
S.E. of regression	1428.041	Akaike info criterion	17.52277	
Sum squared resid	38746720	Schwarz criterion	17.72024	
Log likelihood	-197.5118	F-statistic	286.5846	
Durbin-Watson stat	2.047965	Prob(F-statistic)	0.000000	

可以认为，这时模型设定无变量设定误差。

2、LM 检验

按照 LM 检验步骤，首先生成残差序列 e_i （用 EE 表示），用 EE 对全部解释变量（包括遗漏变量）进行回归，有：

Dependent Variable: EE

Method: Least Squares

Date: 07/08/05 Time: 15:45

Sample (adjusted): 1981 2003

Included observations: 23 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	448.1584	511.0396	0.876954	0.3915
GDP	0.912201	0.148119	6.158568	0.0000
GDP(-1)	-0.815842	0.143928	-5.668420	0.0000
EXCHANGE^2	-0.022569	0.003291	-6.857844	0.0000
<hr/>				
R-squared	0.727360	Mean dependent var	-37.56085	
Adjusted R-squared	0.684312	S.D. dependent var	2541.624	
S.E. of regression	1428.041	Akaike info criterion	17.52277	
Sum squared resid	38746720	Schwarz criterion	17.72024	

Log likelihood	-197.5118	F-statistic	16.89632
Durbin-Watson stat	2.047965	Prob(F-statistic)	0.000014

再计算 $nR^2 = 23 \times 0.72736 = 16.72928$ ，查表 $\chi_{0.025}^2(2) = 7.37776$ ，显然， $16.72928 > 7.37776$ ，拒绝 H_0 ：受约束回归模型，接受 H_1 ：无约束回归模型的假设，即确实存在遗漏变量。因此，在本章的引子中不能判断虽然简单但遗漏了重要变量的方程（1）比复杂的方程（2）更好。

第九章小结

1、计量经济学模型中的古典假设不是无条件的假设，而是有条件的假设。一是所设定的条件期望方程没有方程设定误差；二是所设定的回归模型没有模型设定误差。

2、方程设定误差主要指：（1）真实变量的遗漏；（2）无关变量的引入；（3）解释变量、被解释变量中存在观测误差。此外还有错误函数形式的误设和随机扰动项的非正确设定等。

3、当模型中遗漏了真实的变量时，模型的参数估计是有偏且不一致；参数估计的方差估计不正确，随机扰动项方差的估计也是不正确的，将使得假设检验、区间估计失效。

4、当模型包含无关变量，后果不如遗漏变量那么严重，模型的参数估计仍然是无偏且一致的，随机扰动项的方差将被正确估计，但所估计的方差将趋之于过大，从而使得参数估计的有效性降低，参数估计较为不准确，区间估计的精度下降。

5、检验方程设定误差的常用方法有：（1）DW 检验；（2）LM 检验；（3）Husman 检验；（4）RESET 检验。

6、测量误差分为被解释变量测量误差和解释变量测量误差。测量误差使参数的 OLS 估计有偏且不一致，常常低估真正的回归参数。

第九章主要公式表

均方误差（简记作 MSE）	$MSE(\beta^*) = E(\beta^* - \beta)^2$
均方误差与方差的关系	$MSE(\beta^*) = E\{\beta^* - E(\beta^*)\}^2 + \{E(\beta^*) - \beta\}^2$
DW 检验	$d = \sum_{i=2}^n (e_i - e_{i-1})^2 / \sum_{i=1}^n e_i^2$

拉格朗日乘数检验

$$nR_{\text{sys}}^2 \sim \chi^2(\text{约束个数})$$

思考题与练习题

思考题

- 9.1 什么是设定误差？设定误差有那些基本表现？
- 9.2 不同类型的设定误差对模型参数估计的影响有哪些相同之处？又有哪些区别？
- 9.2 检验变量设定误差有哪几种方法？它们的共性和差异是什么？
- 9.3 如何进行遗漏变量设定误差的后果分析？其检验有哪些方法？如何检验？
- 9.4 如何进行无关变量设定误差的后果分析？其检验有哪些方法？如何检验？
- 9.5 什么是测量误差？测量误差与变量设定误差有何区别？
- 9.6 如何对测量误差和设定误差的后果进行分析？其检验有哪些方法？如何检验？

练习题

- 9.1 设真实模型为无截距模型：

$$Y_i = \alpha_2 X_2 + u_i$$

回归分析中却要求截距项不能为零，于是，有人采用的实证分析回归模型为：

$$Y_i = \beta_1 + \beta_2 X_2 + \varepsilon_i$$

试分析这类设定误差的后果。

9.2 资本资产定价模型 现代投资理论中的资本资产定价模型（CAPM）设定，一定时期内的证券平均收益率与证券波动性（通常由贝塔系数 β 度量）有以下关系

$$\bar{R}_i = \alpha_1 + \alpha_2 (\beta_i) + u_i \quad (1)$$

其中， \bar{R}_i = 证券*i*的平均收益率， β_i = 证券*i*的真正 β 系数， u_i = 随机扰动项；

由于证券*i*的真正 β 系数不可直接观测，通常采用下式进行估算：

$$r_{it} = \alpha_1 + \beta^* r_{m_t} + e_t \quad (2)$$

其中， r_{it} = 时间 t 证券 i 的收益率， r_{m_t} = 时间 t 的市场收益率（通常是某个股票市场的综合指数的收益率）， e_t = 残差项； β^* 是真正 β 系数的一个估计值，且有 $\beta_i^* = \beta_1 + v_i$ ， v_i 是观测误差。

在实际的分析中，我们采用的估计式不是（1）而是：

$$\bar{R}_i = \alpha_1 + \alpha_2 (\beta_i^*) + u_i \quad (3)$$

（1）观测误差 v_i 对 α_2 的估计会有什么影响？

（2）从（3）估计的 α_2 会是真正 α_2 的一个无偏估计吗？若不是，会是真正 α_2 的一致性估计吗？

9.3 1978年-2003年的全国居民消费水平与国民收入的数据如下。

年 份	国民总收入 (GNI)	国内生产总值(GDP)	全国居民消费水平(CT)	农村居民消费水平(CN)	城镇居民消费水平(CC)
1978	3624.1	3624.1	184	138	405
1979	4038.2	4038.2	207	158	434
1980	4517.8	4517.8	236	178	496
1981	4860.3	4862.4	262	199	562
1982	5301.8	5294.7	284	221	576
1983	5957.4	5934.5	311	246	603
1984	7206.7	7171.0	327	283	662
1985	8989.1	8964.4	437	347	802
1986	10201.4	10202.2	485	376	805
1987	11954.5	11962.5	550	417	1089
1988	14922.3	14928.3	693	508	1431
1989	16917.8	16909.2	762	553	1568
1990	18598.4	18547.9	803	571	1686
1991	21662.5	21617.8	896	621	1925

1992	26651.9	26638.1	1070	718	2356
1993	34560.5	34634.4	1331	855	3027
1994	46670.0	46759.4	1746	1118	3891
1995	57494.9	58478.1	2236	1434	4874
1996	66850.5	67884.6	2641	1768	5430
1997	73142.7	74462.6	2834	1876	5796
1998	76967.2	78345.2	2972	1895	6217
1999	80579.4	82067.5	3138	1927	6796
2000	88254.0	89468.1	3397	2037	7402
2001	95727.9	97314.8	3609	2156	7761
2002	103935.3	105172.3	3818	2269	8047
2003	116603.2	117251.9	4089	2361	8471

若依据弗里德曼的持久收入假设，消费函数的真正模型应为

$$CC_i = \alpha + \beta GNI_i + u_i$$

(1) 试用 Eviews 软件，采用两种以上检验方法对实证分析模型

$$CC_i = \gamma_1 + \gamma_2 GDP_i + \mu_i$$

进行变量设定检验；

(2) 若 $GNI_i^* = GDP_i + \omega_i$ 。

试用 Eviews 软件，采用两种以上检验方法对实证分析模型

$$CC_i = \gamma_1 + \gamma_2 GDP_i + \varepsilon_i$$

进行测量误差检验。

9.4 考虑真正的 Cobb-Douglas 生产函数：

$$\ln Y_i = \alpha_1 + \alpha_2 \ln L_{1i} + \alpha_3 \ln L_{2i} + \alpha_4 \ln K_i + u_i$$

其中， Y = 产出， L_1 = 生产性劳力， L_2 = 非生产性劳力， K = 资本；

若在对横截面数据进行的实证分析中，采用的回归模型是：

$$\ln Y_i = \beta_1 + \beta_2 \ln L_{1i} + \beta_3 \ln K_i + u_i$$

试问：

(1) 表达式 $E(\hat{\beta}_2) = \alpha_2$ 和 $E(\hat{\beta}_3) = \alpha_4$ 成立吗？

(2) 若已经知道 L_2 是生产函数中的一个无关变量, (1) 中答案是否也成立?

9.5 假设制造业企业工人的平均劳动生产率 (Y) 与工人的平均培训时间(t)和平均能力 (X) 之间存在依存关系, 可建立如下的回归模型:

$$Y = \beta_0 + \beta_1 t + \beta_2 X + u$$

若政府给那些工人能力低的企业以政府培训补助, 则平均培训时间就和工人平均能力负相关。现在考虑这个因素, 采用如下模型进行回归:

$$Y = \alpha_0 + \alpha_1 t + \varepsilon$$

问由此获得的 $\hat{\alpha}_1$ 会有怎样的偏误。

第九章附录

附录 9.1 $\hat{\alpha}_2$ 概率极限性质的证明

$$\begin{aligned} p \lim_{n \rightarrow \infty} \hat{\alpha}_2 &= p \lim_{n \rightarrow \infty} \beta_2 + p \lim_{n \rightarrow \infty} \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + p \lim_{n \rightarrow \infty} \frac{\sum x_{2i} (u_i - \bar{u})}{\sum x_{2i}^2} \\ &= \beta_2 + \beta_3 \frac{p \lim_{n \rightarrow \infty} \frac{1}{n} \sum x_{2i} x_{3i}}{p \lim_{n \rightarrow \infty} \frac{1}{n} \sum x_{2i}^2} + \frac{p \lim_{n \rightarrow \infty} \frac{1}{n} \sum x_{2i} (u_i - \bar{u})}{p \lim_{n \rightarrow \infty} \frac{1}{n} \sum x_{2i}^2} \\ &= \beta_2 + \beta_3 \frac{Cov(X_{2i}, X_{3i})}{Var(X_{2i})} + \frac{Cov(X_{2i}, u_i)}{Var(X_{2i})} \end{aligned}$$

其中: $\frac{1}{n} \sum x_{2i}^2$ 为 X_2 的样本方差, $\frac{1}{n} \sum x_{2i} x_{3i}$ 为 X_2 和 X_3 的样本协方差, $\frac{1}{n} \sum x_{2i} (u_i - \bar{u})$ 为 X_2 和 u_i 的样本协方差。

附录 9.2 参数 $\hat{\alpha}_2$ 一致性的证明

$$\begin{aligned} p \lim_{n \rightarrow \infty} \hat{\alpha}_2 &= p \lim_{n \rightarrow \infty} \beta_2 + p \lim_{n \rightarrow \infty} \left(\frac{(\sum x_{3i}^2)(\sum x_{2i} (u_i - \bar{u})) - (\sum x_{2i} x_{3i})(\sum x_{3i} (u_i - \bar{u}))}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} \sum x_{3i})^2} \right) \\ &= \beta_2 + \left(\frac{p \lim_{n \rightarrow \infty} [(\sum x_{3i}^2)(\sum x_{2i} (u_i - \bar{u})) - (\sum x_{2i} x_{3i})(\sum x_{3i} (u_i - \bar{u}))]}{p \lim_{n \rightarrow \infty} [\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} \sum x_{3i})^2]} \right) \end{aligned}$$

$$\begin{aligned}
&= \beta_2 + \left(\frac{p \lim_{n \rightarrow \infty} (\sum x_{3i}^2) p \lim_{n \rightarrow \infty} (\sum x_{2i} (u_i - \bar{u})) - p \lim_{n \rightarrow \infty} (\sum x_{2i} x_{3i}) p \lim_{n \rightarrow \infty} (\sum x_{3i} (u_i - \bar{u}))}{p \lim_{n \rightarrow \infty} (\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} \sum x_{3i})^2)} \right) \\
&= \beta_2 + \left(\frac{Var(X_{3i}) Cov(X_{2i}, u_i) - Cov(X_{2i}, X_{3i}) Cov(X_{3i}, u_i)}{Var(X_{2i}) Var(X_{3i}) - p \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum x_{2i} \sum x_{3i} \right)^2} \right) \\
&\stackrel{Cov(X_{2i}, u_i)=0, Cov(X_{3i}, u_i)=0}{=} \beta_2 + \left(\frac{Var(X_{3i}) \times 0 - Cov(X_{2i}, X_{3i}) \times 0}{Var(X_{2i}) Var(X_{3i}) - p \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum x_{2i} \sum x_{3i} \right)^2} \right) = \beta_2
\end{aligned}$$

同理，可证 $E(\hat{\alpha}_1) = \beta_1$ ， $E(\hat{\alpha}_3) = \beta_3 = 0$ ； $p \lim_{n \rightarrow \infty} \hat{\alpha}_1 = \beta_1$ 和 $p \lim_{n \rightarrow \infty} \hat{\alpha}_3 = \beta_3 = 0$ 。

附录 9.3 有测量误差模型参数估计结果的推导

$$\begin{aligned}
\hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} \\
&= \frac{\sum [x_i^* + (\omega_i - \bar{\omega})][y_i^* + (\varepsilon_i - \bar{\varepsilon})]}{\sum [x_i^* + (\omega_i - \bar{\omega})]^2} \\
&= \frac{\sum x_i^* y_i^* + \sum (\omega_i - \bar{\omega}) y_i^* + \sum x_i^* (\varepsilon_i - \bar{\varepsilon}) + \sum (\omega_i - \bar{\omega}) (\varepsilon_i - \bar{\varepsilon})}{\sum x_i^{*2} + \sum (\omega_i - \bar{\omega})^2 + 2 \sum x_i^* (\omega_i - \bar{\omega})} \\
&= \frac{\sum x_i^* (\beta x_i^* + (u_i - \bar{u})) + \sum (\omega_i - \bar{\omega}) (\beta x_i^* + (u_i - \bar{u})) + \sum x_i^* (\varepsilon_i - \bar{\varepsilon}) + \sum (\omega_i - \bar{\omega}) (\varepsilon_i - \bar{\varepsilon})}{\sum x_i^{*2} + \sum (\omega_i - \bar{\omega})^2 + 2 \sum x_i^* (\omega_i - \bar{\omega})} \\
&= \frac{\beta \sum x_i^{*2} + \sum x_i^* (u_i - \bar{u}) + \beta \sum (\omega_i - \bar{\omega}) x_i^* + \sum (\omega_i - \bar{\omega}) (u_i - \bar{u}) + \sum x_i^* (\varepsilon_i - \bar{\varepsilon}) + \sum (\omega_i - \bar{\omega}) (\varepsilon_i - \bar{\varepsilon})}{\sum x_i^{*2} + \sum (\omega_i - \bar{\omega})^2 + 2 \sum x_i^* (\omega_i - \bar{\omega})}
\end{aligned}$$

因此，有测量误差模型参数估计的概率极限为

$$\begin{aligned}
p \lim \hat{\beta} &= \frac{p \lim [\beta \sum x_i^{*2} + \sum x_i^* (u_i - \bar{u}) + \beta \sum (\omega_i - \bar{\omega}) x_i^* + \sum (\omega_i - \bar{\omega}) (u_i - \bar{u}) + \sum x_i^* (\varepsilon_i - \bar{\varepsilon}) + \sum (\omega_i - \bar{\omega}) (\varepsilon_i - \bar{\varepsilon})]}{p \lim [\sum x_i^{*2} + \sum (\omega_i - \bar{\omega})^2 + 2 \sum x_i^* (\omega_i - \bar{\omega})]} \\
&= \frac{\beta Var(X_i^*) + Cov(X_i^*, u_i) + \beta Cov(X_i^*, \omega_i) + Cov(\omega_i, u_i) + Cov(X_i^*, \varepsilon_i) + Cov(\omega_i, \varepsilon_i)}{Var(X_i^*) + Var(\omega_i) + 2Cov(X_i^*, \omega_i)} \\
&= \frac{\beta Var X_i^*}{Var X_i^* + \sigma_\omega^2} = \frac{\beta}{1 + \frac{\sigma_\omega^2}{\sigma_{X^*}^2}}
\end{aligned}$$

*第十章 时间序列计量经济模型¹

引子:

是真回归还是伪回归?

在经典的回归分析中,通常的做法是:首先采用普通最小二乘法(OLS)对回归模型进行估计,然后根据可决系数 R^2 或 F 检验统计量值的大小来判定变量之间的相依程度,根据回归系数估计值的 t 统计量对系数的显著性进行判断,最后在回归系数显著不为零的基础上对回归系数估计值给予经济解释。

为了分析美国的个人可支配收入(I)与个人消费总支出(E)的关系,遵照以上作法,收集了 1970 年至 1991 年的季度时间序列数据,用 OLS 法作 E 关于 I 的线性回归,得到如下结果:

$$E_t = -171.4412 + 0.9672 I_t$$

$$t = (-7.4809) \quad (119.8711)$$

$$R^2 = 0.9940 \quad DW = 0.5316$$

从回归结果来看, R^2 非常高,个人可支配收入 I 的回归系数 t 统计量也非常大,边际消费倾向符合经济假设。

(资料来源:古扎拉蒂《计量经济学》下册,第 719 页,中国人民大学出版社)

凭借经验判断,这个模型的设定是好的,所用数据也是可靠的,样本容量很充分,这应是满意的结果。准备将这个计量结果用于经济结构分析和经济预测。

可是有人提出,这个回归结果可能是虚假的!可能只不过是一种“伪回归”!如果真是这样,将所估计的模型直接用于经济结构分析和预测,“就要千万小心!”。

这里用时间序列数据进行的回归,究竟是真回归还是伪回归呢?为什么模型、样本、数据、检验结果都很理想,却可能得到“伪回归”的结果呢?

时间序列数据被广泛地运用于计量经济研究。经典时间序列分析和回归分析有许多假定前提,如序列的平稳性、正态性等,如果直接将经济变量的时间序列数据用于建模分析,实际上隐含了这些假定。在这些假定成立的条件下,进行的 t、F、 χ^2 等检验才具有较高的

¹ 本章内容本科教学供选择

可靠度。但是，越来越多的经验证据表明，经济分析中所涉及的大多数时间序列是非平稳的。那末，如果直接将非平稳时间序列当作平稳时间序列来进行分析，会造成什么不良后果？如何判断一个时间序列是否为平稳序列？当我们在计量经济分析中涉及到非平稳时间序列时，应作如何处理呢？这就是本章要讨论的基本内容。

第一节 时间序列计量经济分析的基本概念

一、伪回归问题

经典计量经济学建模过程中，通常假定经济时间序列是平稳的，而且主要以某种经济理论或对某种经济行为的认识来确立计量经济模型的理论关系形式，借此形式进行数据收集、参数估计以及模型检验，这是 20 世纪 70 年代以前计量经济学的主导方法。然而，这种方法所构建的计量经济模型在 20 世纪 70 年代出现石油危机后引起的经济动荡面前却失灵了。这里的失灵不是指这些模型没能预见石油危机的出现，而是指这些模型无法预计石油危机的振荡对许多基本经济变量的动态影响。因此引起了计量经济学界对经典计量经济学方法论的反思，并将研究的注意力转向宏观经济变量非平稳性对建模的影响。人们发现，由于经济分析中所涉及的经济变量数据基本上是时间序列数据，而大多数经济时间序列是非平稳的，如果直接将非平稳时间序列当作平稳时间序列进行回归分析，则可能会带来不良后果，如伪回归问题。

所谓“伪回归”，是指变量间本来不存在有意义的关系，但回归结果却得出存在有意义关系的错误结论。经济学家早就发现经济变量之间可能会存在伪回归现象，但在什么条件下会产生伪回归现象，长期以来无统一认识。直到 20 世纪 70 年代，Grange、Newbold 研究发现，造成“伪回归”的根本原因在于时间序列变量的非平稳性。他们用 Monte Carlo 模拟方法研究表明，如果用传统回归分析方法对彼此不相关联的非平稳变量进行回归，t 检验值和 F 检验值往往会倾向于显著，从而得出“变量相依”的“伪回归结果”。

因此，在利用回归分析方法讨论经济变量有意义的经济关系之前，必须对经济变量时间序列的平稳性与非平稳性进行判断。如果经济变量时间序列是非平稳的，则需要寻找新的处理方法。20 世纪 80 年代发展起来的协整理论就是处理非平稳经济变量关系的行之有效的方法。该理论自从诞生以来，受到众多经济学家的重视，并广泛运用于对实际经济问题的研究。

二、随机过程的概念

在概率论和数理统计中，随机变量是分析随机现象的有力工具。对于一些简单的随机

现象，一个随机变量就足够了，如候车人数，某单位一天的总用水量等。对于一些复杂的随机现象，用一个随机变量来描述就不够了，而需要用若干个随机变量来加以刻画。例如平面上的随机点，某企业一天的工作情况（产量、次品率、耗电量、出勤人数等）都需要用多个随机变量来刻画。

还有些随机现象，要认识它必须研究其发展变化过程，这一类随机现象不能只用一个或多个随机变量来描述，而必须考察其动态变化过程，随机现象的这种动态变化过程就是随机过程。例如，某一天电话的呼叫次数 ξ ，它是一个随机变量。若考察它随时间 t 变动的情况，则需要考察依赖于时间 t 的随机变量 ξ_t ， $\{\xi_t\}$ 就是一个随机过程。又例如，某国某年的 GNP 总量，是一个随机变量，但若考查它随时间变化的情形，则 $\{GNP_t\}$ 就是一个随机过程。

一般地，若对于每一特定的 t ($t \in T$)， Y_t 为一随机变量，则称这一族随机变量 $\{Y_t\}$ 为一个**随机过程**。若 T 为一连续区间，则 $\{Y_t\}$ 为**连续型随机过程**。若 T 为离散集合，如 $T = (0, 1, 2, \dots)$ 或 $T = (\dots, -2, -1, 0, 1, 2, \dots)$ ，则 $\{Y_t\}$ 为**离散型随机过程**。随机过程的统计特征通常用其分布及数字特征来刻画。

离散型时间指标集的随机过程通常称为随机型时间序列，简称为时间序列。经济分析中常用的时间序列数据都是经济变量随机序列的一个实现。

三、时间序列的平稳性

所谓时间序列的平稳性，是指时间序列的统计规律不会随着时间的推移而发生变化。也就是说，生成变量时间序列数据的随机过程的特征不随时间变化而变化。以平稳时间序列数据作为计量经济模型变量的观测值时，其估计方法、检验过程才可能采用前面几章所介绍的方法。

直观上，一个平稳的时间序列可以看做作一条围绕其均值上下波动的曲线。从理论上，有两种意义的平稳性，一是严格平稳，另一是弱平稳。严格平稳是指随机过程 $\{Y_t\}$ 的联合分布函数与时间的位移无关。设 $\{Y_t\}$ 为一随机过程， n, h 为任意实数，若联合分布函数满足：

$$F_{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}}(y_1, \dots, y_n) = F_{Y_{t_1+h}, \dots, Y_{t_n+h}}(y_1, \dots, y_n) \quad (10.1)$$

则称 $\{Y_t\}$ 为严格平稳过程，它的分布结构不随时间推移而变化。

弱平稳是指随机过程 $\{Y_t\}$ 的期望、方差和协方差不随时间推移而变化。若 $\{Y_t\}$ 满足：

$$E(Y_t) = u$$

$$\text{Var}(Y_t) = r_0 = \sigma^2 \quad (10.2)$$

$$\text{Cov}(Y_t, Y_s) = \text{Cov}(Y_{t+h}, Y_{s+h}) = r(t-s, 0) = r_{t-s}$$

则称 $\{Y_t\}$ 为弱平稳随机过程。在以后的讨论中，关于平稳性的概念通常是指弱平稳。

所谓时间序列的非平稳性，是指时间序列的统计规律随着时间的位移而发生变化，即生成变量时间序列数据的随机过程的特征随时间而变化。当生成序列的随机过程是非平稳的时候，其均值函数，方差函数不再是常数，自协方差函数也不仅仅是时间间隔 $t-s$ 的函数，前面所介绍的高斯—马尔科夫定理不再成立，一个变量对其他变量的回归可能会导致伪回归结果，前面所介绍的计量经济技术也将遇到困难。

在经济领域中，我们所得到的许多时间序列观测值大都不是由平稳过程产生的。例如，国内生产总值 GDP 大多数情况下随时间的位移而持续增长；货币供给量 M2 在正常状态下会随时间的位移而扩大。也就是说，2000 年 GDP 或 M2 观测值的随机性质与 1996 年的 GDP 和 M2 的随机性质有相当的区别。

由于在实际中遇到的时间序列数据很可能是非平稳序列，而平稳性在计量经济建模中又具有重要地位，因此有必要对观测值的时间序列数据进行平稳性检验。

第二节 时间序列平稳性的单位根检验

时间序列平稳性的检验方法主要有传统方法和现代方法，前者以自相关函数检验为代表，后者以单位根检验为代表。本书只介绍目前最常用的单位根检验法。

一、单位根过程

一般来讲，由于经济系统惯性的作用，经济时间序列往往存在着前后依存关系，这种前后依存关系是时间序列预测的基础。假定 $\{Y_t\}$ 为一时间序列，最简单的一种前后依存关系就是变量当前的取值主要与其前一时期的取值状况有关，而与其前一时期以前的取值状况无直接关系，也就是说 Y_t 主要与 Y_{t-1} 相关，与 Y_{t-2} ， Y_{t-3} ，……无关。可用如下的一阶自回归模型来描述这种关系：

$$Y_t = \phi Y_{t-1} + \varepsilon_t \quad (10.3)$$

常记作 AR(1)。

如果 Y_t 不仅与前一期 Y_{t-1} 有关，而且与 Y_{t-2} 相关，显然，在这种情况下用 AR(1) 来刻画 Y_t 的动态依存关系就不恰当了，而需要在模型中引入 Y_{t-2} 。一般的，如果 Y_t 与过去时期直到 Y_{t-p} 的取值相关，则 $\{Y_t\}$ 的动态关系就需要使用包含 Y_{t-1} ，…… Y_{t-p} 在内的 p 阶自回归模型来加以刻画。P 阶自回归模型的一般形式为：

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \quad (10.4)$$

为了说明单位根过程的概念，这里侧重以 AR(1) 模型 $Y_t = \varphi Y_{t-1} + \varepsilon_t$ 进行分析。根据平稳时间序列分析的理论可知，当 $|\varphi| < 1$ 时，该序列 $\{Y_t\}$ 是平稳的，此模型是经典的 Box-Jenkins 时间序列 AR(1) 模型。但是，如果 $\varphi = 1$ ，则序列的生成过程变为随机游走过程(Random Walk Process)：

$$Y_t = Y_{t-1} + \varepsilon_t \quad (10.5)$$

其中， $\{\varepsilon_t\}$ 独立同分布且均值为零、方差恒定为 σ^2 。随机游走过程的方差为：

$$\begin{aligned} \text{Var}(Y_t) &= \text{Var}(Y_{t-1} + \varepsilon_t) \\ &= \text{Var}(Y_{t-2} + \varepsilon_{t-1} + \varepsilon_t) \\ &= \text{Var}(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-1} + \varepsilon_t) \\ &= t\sigma^2 \end{aligned}$$

当 $t \rightarrow \infty$ 时，序列的方差趋于无穷大，这说明随机游走过程是非平稳的。

如果一个序列是随机游动过程，则称这个序列是一个单位根过程²。较随机游动更一般的，是一般的单位根过程。若随机过程 $\{Y_t\}$ 遵从：

$$Y_t = \gamma Y_{t-1} + u_t \quad (10.6)$$

其中， $\gamma = 1$ ， $\{u_t\}$ 为一平稳过程，且 $E(u_t) = 0$ ， $\text{Cov}(u_t, u_{t-s}) = \mu_s < \infty$ ， $s = 0, 1, 2, \dots$ 。则称序列 $\{Y_t\}$ 为（不带漂移的）单位根过程。带漂移和时间趋势的单位根过程服从如下模型：

$$Y_t = \alpha + \beta t + Y_{t-1} + u_t \quad (10.7)$$

²将一阶自回归模型表示成如下形式： $Y_t - \varphi Y_{t-1} = \varepsilon_t$ 或 $(1 - \varphi L)Y_t = \varepsilon_t$ ，其中， L 是滞后算子，

即 $LY_t = Y_{t-1}$ 。根据模型的滞后多项式 $(1 - \varphi L)$ ，可以写出对应的线性方程： $1 - \varphi z = 0$ （通常称为特征方程），该方程的根为： $z = 1/\varphi$ 。当 $|\varphi| < 1$ 时序列是平稳的，特征方程的根满足条件 $|z| > 1$ ，我们称方程的根在单位园以外；当 $\varphi = 1$ 时，序列的生成过程变为随机游动过程，由前面可知，随机游动过程是非平稳的，由于此时方程的根 $z = 1$ ，所以通常称序列含有单位根，或者说序列的生成过程为“单位根过程”。由此可见，检验序列的非平稳性就变为检验特征方程是否有单位根，这就是单位根检验方法的由来。

显然，随机游动过程是一般单位根过程的一个特例。

从单位根过程的定义可以看出，含一个单位根的过程 $\{Y_t\}$ ，其一阶差分：

$$\Delta Y_t = Y_t - Y_{t-1} = u_t$$

是一平稳过程，像这种经过一次差分后变为平稳的序列称为一阶单整序列(Integrated Process)，记为 $\{Y_t\} \sim I(1)$ 。有时，一个序列经一次差分后可能还是非平稳的，如果序列经过二阶差分后才变成平稳过程，则称序列为二阶单整序列，记为 $\{Y_t\} \sim I(2)$ 。一般地，如果序列 $\{Y_t\}$ 经过 d 次差分后平稳，而 $d-1$ 次差分却不平稳，那么称 $\{Y_t\}$ 为 d 阶单整序列，记为 $\{Y_t\} \sim I(d)$ ， d 称为整形阶数。特别地，若序列 $\{Y_t\}$ 本身是平稳的，则称序列为零阶单整序列，记为 $\{Y_t\} \sim I(0)$ 。

二、Dickey-Fuller 检验 (DF 检验)

我们知道大多数的经济变量，如 GDP、总消费、价格水平以及货币供给量 M2 等都会呈现出强烈的趋势特征。这些具有趋势特征的经济变量，当发生经济振荡或冲击后，一般会出现两种情形，一是受到振荡或冲击后，经济变量逐渐又回到它们的长期趋势轨迹；二是这些经济变量没有回到原有轨迹，而呈现出随机游走的状态。若我们研究的经济变量遵从一个非平稳过程(比如随机游走过程)，当运用最小二乘法时，前面所介绍的高斯-马尔科夫定理不再成立，一个变量对其他变量的回归可能会导致伪回归结果。同时，如果我们所研究的经济变量(如 GDP)是非平稳的，则经济出现突发性振荡(如石油价格猛增，金融危机或政府开支骤减等)所造成的影响不会在短期内消失，其影响将是持久性的。这也是研究单位根检验的重要意义所在。

假设数据序列是由下列自回归模型生成的：

$$Y_t = \gamma Y_{t-1} + \varepsilon_t \quad (10.8)$$

其中， ε_t 独立同分布，期望为零，方差为 σ^2 ，我们要检验该序列是否含有单位根。检验的原假设为： $H_0 : \gamma = 1$ ，回归系数 γ 的 OLS 估计为：

$$\hat{\gamma} = \frac{\sum y_{t-1} y_t}{\sum y_{t-1}^2}$$

检验所用的统计量为：

$$t = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}_{\hat{\gamma}}}$$

在 $H_0: \gamma = 1$ 成立的条件下，t 统计量为：

$$t = \frac{\hat{\gamma} - 1}{\hat{\sigma}_{\hat{\rho}}} \quad (10.9)$$

但麻烦的是，Dickey、Fuller 通过研究发现，在原假设成立的情况下，该统计量不服从 t 分布。由于 t 检验统计量不再服从传统的 t 分布，所以传统的 t 检验方法失效。可以证明，上述统计量的极限分布存在，一般称其为 Dickey—Fuller 分布。根据这一分布所作的检验称为 DF 检验，为了区别，t 统计量的值有时也称为 τ 值。

Dickey、Fuller 得到 DF 检验的临界值，并编制了 DF 检验临界值表供查。在进行 DF 检验时，比较 t 统计量值与 DF 检验临界值，就可在某个显著性水平上拒绝或接受原假设。在实际应用中，可按如下检验步骤进行：

(1) 根据所观察的数据序列，用 OLS 法估计一阶自回归模型：

$$Y_t = \gamma Y_{t-1} + \varepsilon_t$$

得到回归系数 γ 的 OLS 估计

$$\hat{\gamma} = \frac{\sum y_{t-1} y_t}{\sum y_{t-1}^2}$$

(2) 提出假设： $H_0: \gamma = 1$ ，检验用统计量为常规 t 统计量，

$$t = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}_{\hat{\gamma}}}$$

(3) 计算在原假设成立的条件下 t 统计量值，查 DF 检验临界值表得临界值，然后将 t 统计量值与 DF 检验临界值进行比较：若 t 统计量值小于 DF 检验临界值，则拒绝原假设 $H_0: \gamma = 1$ ，说明序列不存在单位根；若 t 统计量值大于或等于 DF 检验临界值，则接受原假设 $H_0: \gamma = 1$ ，说明序列存在单位根。

此外，Dickey、Fuller 研究发现，DF 检验的临界值同序列的数据生成过程以及回归模型的类型有关，因此他们针对如下三种方程编制了临界值表，后来 Mackinnon 把临界值表加以扩充，形成了目前使用广泛的临界值表，在 Eviews 软件中使用的是 Mackinnon 临界值表。

$$\text{模型 I:} \quad Y_t = \gamma Y_{t-1} + \varepsilon_t$$

$$\text{模型 II:} \quad Y_t = \alpha + \gamma Y_{t-1} + \varepsilon_t$$

$$\text{模型 III:} \quad Y_t = \alpha + \beta t + \gamma Y_{t-1} + \varepsilon_t$$

三、Augmented Dickey-Fuller 检验 (ADF 检验)

上述 DF 检验存在的问题是, 在检验所设定的模型时, 假设随机扰动项 ε_t 不存在自相关。但大多数的经济数据序列是不能满足此项假设的, 当随机扰动项存在自相关时, 直接使用 DF 检验法会出现偏误, 为了保证单位根检验的有效性, 人们对 DF 检验进行拓展, 从而形成了扩展的 DF 检验(Augmented Dickey-Fuller Test), 简称为 ADF 检验。

假设基本模型为如下三种类型:

$$\text{模型 I:} \quad Y_t = \gamma Y_{t-1} + u_t$$

$$\text{模型 II:} \quad Y_t = \alpha + \gamma Y_{t-1} + u_t$$

$$\text{模型 III:} \quad Y_t = \alpha + \beta t + \gamma Y_{t-1} + u_t$$

其中 u_t 为随机扰动项, 它可以是一个一般的平稳过程。为了借用 DF 检验的方法, 将模型变为如下形式:

$$\text{模型 I:} \quad Y_t = \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$$

$$\text{模型 II:} \quad Y_t = \alpha + \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$$

$$\text{模型 III:} \quad Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$$

可以证明, 在上述模型中检验原假设 $H_0: \gamma = 1$ 的 t 统计量的极限分布, 同 DF 检验的极限分布相同, 从而可以使用相同的临界值表, 这种检验称为 ADF 检验。

【例 10.1】根据《中国统计年鉴 2004》, 得到我国 1978—2003 年的 GDP 序列, 检验其是否为平稳序列。在 Eviews 中录入数据, 其结果如表 10.1, 时间序列图见图 10.1。

表 10.1 中国 1978—2003 年度 GDP 序列

年度	GDP	年度	GDP	年度	GDP
----	-----	----	-----	----	-----

1978	3624.1	1987	11962.5	1996	67884.6
1979	4038.2	1988	14928.3	1997	74462.6
1980	4517.8	1989	16909.2	1998	79395.7
1981	4862.4	1990	18547.9	1999	82067.5
1982	5294.7	1991	21617.8	2000	89468.1
1983	5934.5	1992	26638.1	2001	97314.8
1984	7171	1993	34634.4	2002	105172.3
1985	8964.4	1994	46759.4	2003	116898.4
1986	10202.2	1995	58478.1		

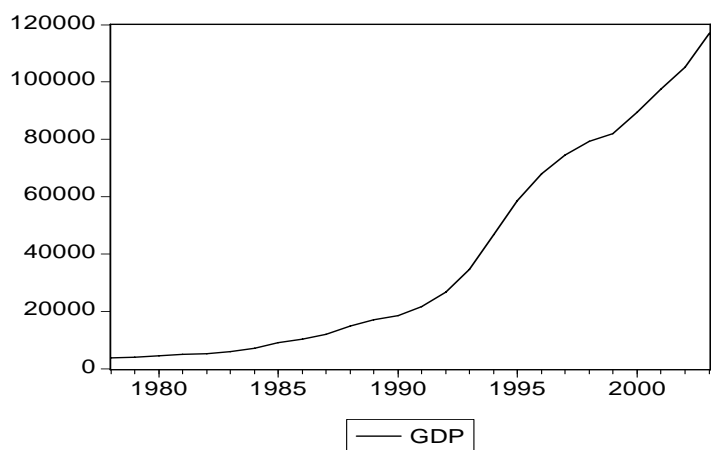


图 10.1 GDP 时间序列图

由 GDP 时间序列图可以看出，该序列可能存在趋势项，因此选择 ADF 检验的第三种模型进行检验。估计结果如下：

$$\hat{\Delta GDP}_t = -1565.141 + 355.62t - 0.02883GDP_{t-1} + 1.016\Delta GDP_{t-1} - 0.460382\Delta GDP_{t-2}$$

在原假设 $H_0: \gamma = 1$ 下，单位根的 t 检验统计量的值为：

$$t = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}_{\hat{\gamma}}} = \frac{-0.028830}{0.036679} = -0.786011$$

在 1%、5%、10% 三个显著性水平下，单位根检验的 Mackinnon 临界值分别为 -4.4167、-3.6219、-3.2474，显然，上述 t 检验统计量值大于相应临界值，从而不能拒绝 H_0 ，表明我国 1978——2003 年度 GDP 序列存在单位根，是非平稳序列。

第三节 协 整

一、协整的概念

在给出协整（Cointegration）概念之前，先看一个货币需求分析的例子。经典的理论分析告诉我们，一个国或地区的货币需求量主要取决于规模变量和机会成本变量，即实际收入、价格水平以及利率。如果以对数形式的计量经济模型将货币需求函数描述出来，其形式为：

$$\ln M_t = \beta_0 + \beta_1 \ln P_t + \beta_2 \ln Y_t + \beta_3 r_t + u_t$$

其中，M 为货币需求，P 为价格水平，Y 为实际收入总额，r 为利率，u 为扰动项， β_i 为模型参数。

人们关心的问题是：如何估计出上述回归模型，检验模型参数是否满足条件： $\beta_1 = 1, \beta_2 > 0, \beta_3 < 0$ ，并回答估计出来的货币需求函数是否揭示了货币需求的长期均衡关系。如果上述货币需求函数是适当的，那么货币需求对长期均衡关系的偏离将是暂时的，扰动项序列是平稳序列，估计出来的货币需求函数就揭示了货币需求的长期均衡关系。相反，如果扰动项序列有随机趋势而呈现非平稳现象，那么模型中的误差会逐步积聚，使得货币需求对长期均衡关系的偏离在长时期内不会消失。因此，上述货币需求模型是否具有实际价值，关键在于扰动项序列是否平稳。

但面临的问题是，货币供给量、实际收入、价格水平以及利率可能是非平稳的 $I(1)$ 序列。一般情况下，多个非平稳序列的线性组合也是非平稳序列。如果货币供给量、实际收入、价格水平以及利率的任何线性组合都是非平稳的，那么上述货币需求模型的扰动项序列就不可能是平稳的，从而模型并没有揭示出货币需求的长期稳定关系。反过来说，如果上述货币需求模型描述了货币需求的长期均衡关系，那么扰动项序列必定是平稳序列，也就是说，非平稳的货币供给量、实际收入、价格水平以及利率四变量之间存在平稳的线性组合。

上述例子揭示了这样一个事实：“包含非平稳变量的均衡系统，必然意味着这些非平稳变量的某种组合是平稳的”。这正是协整理论的思想。

所谓协整，是指多个非平稳经济变量的某种线性组合是平稳的。例如，收入与消费，工资与价格，政府支出与税收，出口与进口等，这些经济时间序列一般是非平稳序列，但它们之间却往往存在长期均衡关系。下面给出协整的严格定义：

对于两个序列 $\{x_t\}$ 与 $\{y_t\}$ ，如果 $y_t \sim I(1)$, $x_t \sim I(1)$ ，而且存在一组非零常数

α_1, α_2 , 使得 $\alpha_1 x_t + \alpha_2 y_t \sim I(0)$, 则称 x_t 和 y_t 之间是协整的。

一般的, 设有 $k (\geq 2)$ 个序列 $\{y_{1t}\}, \{y_{2t}\}, \dots, \{y_{kt}\}$, 用 $Y_t = (y_{1t}, y_{2t}, \dots, y_{kt})'$ 表示由此 k 个序列构成的 k 维向量序列, 如果:

(1) 每一个序列 $\{y_{1t}\}, \{y_{2t}\}, \dots, \{y_{kt}\}$ 都是 d 阶单整序列, 即 $y_{jt} \sim I(d)$;

(2) 存在非零向量 $\alpha = (a_1, a_2, \dots, a_k)'$, 使得 $\alpha' Y_t = a_1 y_{1t} + a_2 y_{2t} + \dots + a_k y_{kt}$ 为 $(d-b)$ 阶单整序列, 即 $\alpha' Y_t \sim I(d-b), 0 < b \leq d$ 。

则称向量序列 $Y_t = (y_{1t}, y_{2t}, \dots, y_{kt})'$ 的分量间是 d, b 阶协整的, 记为 $Y_t \sim CI(d, b)$, 向量 $\alpha = (a_1, a_2, \dots, a_k)'$ 称为协整向量。

特别地, 若 $d = b = 1$, 则 $Y_t \sim CI(1, 1)$, 说明尽管各个分量序列是非平稳的一阶单整序列, 但它们的某种线性组合却是平稳的。这种 $(1, 1)$ 阶协整关系在经济计量分析中较为常见。例如, 假设变量 y_{1t} 与变量 $y_{it} (i = 2, \dots, m)$ 之间存在 $(1, 1)$ 阶协整关系, 协整向量为 $\alpha = (1, -\beta_2, \dots, -\beta_m)'$, 则这种协整关系可表示为:

$$y_{1t} = \alpha + \beta_2 y_{2t} + \dots + \beta_m y_{mt} + u_t \quad (10.10)$$

组合变量 u_t 就为 $I(0)$ 过程。

协整概念的提出对于用非平稳变量建立经济计量模型, 以及检验这些变量之间的长期均衡关系非常重要。

(1) 如果多个非平稳变量具有协整性, 则这些变量可以合成一个平稳序列。这个平稳序列就可以用来描述原变量之间的均衡关系。

(2) 当且仅当多个非平稳变量之间具有协整性时, 由这些变量建立的回归模型才有意义。所以协整性检验也是区别真实回归与伪回归的有效方法。

(3) 具有协整关系的非平稳变量可以用来建立误差修正模型。由于误差修正模型把长期关系和短期动态特征结合在一个模型中, 因此既可以克服传统计量经济模型忽视伪回归的问题, 又可以克服建立差分模型忽视水平变量信息的弱点。

二、协整检验

协整性的检验有两种方法, 一种是基于回归残差的协整检验, 这种检验也称为单一方程的协整检验; 另一种是基于回归系数的完全信息协整检验。这里我们仅考虑单一方程的情

形，而且主要介绍两变量协整关系的 EG 两步法检验。

第一步，若 X_t 与 Y_t 是一阶单整 (I (1)) 序列，即 ΔX_t 和 ΔY_t 是平稳的，用 OLS 法对回归方程(也称为协整回归方程)

$$X_t = \alpha + \beta Y_t + u_t \quad (10.11)$$

进行估计，得到残差序列 $e_t = X_t - (\hat{\alpha} + \hat{\beta} Y_t)$ 。

第二步，检验 e_t 的平稳性。若 e_t 为平稳的，则 X_t 与 Y_t 是协整的，反之则不是协整的。因为若 X_t 与 Y_t 不是协整的，则它们的任一线性组合都是非平稳的。因此残差 e_t 将是非平稳。换言之，对残差序列 e_t 是否具有平稳性的检验，也就是对 X_t 与 Y_t 是否存在协整的检验。

检验 e_t 为非平稳的假设可用两种方法。一种方法是对残差序列进行 DF 检验，即对 e_t 进行单位根检验，其检验方法在前面已介绍，但要注意的是，DF 检验和 ADF 检验使用的临界值应该用 Engle-Granger 编制的专用临界值表。

另一种方法是协整回归 DW 检验。具体做法为，用协整回归所得的残差构造 DW 统计量：

$$CRDW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} \quad (10.12)$$

若 e_t 是随机游走的，则 $e_t - e_{t-1}$ 的数学期望为 0，故 DW 也应接近于 0。因此，只需检验： $H_0: DW = 0$ 是否成立，若 H_0 成立， e_t 为随机游走， X_t 与 Y_t 间不存在协整，反之则存在协整。Sargan 和 Bhargava 最早编制了用于检验协整的 DW 临界值表。表 10.2 是观察数为 100 时，该检验的临界值。例如，当 $DW=0.71$ 时，在 1% 的显著性水平上我们能拒绝 $H_0: DW = 0$ ，即拒绝非协整假设。

表 10.2 检验 $DW=0$ 的临界值

显著性水平%	DW 临界值
1	0.511
5	0.386
10	0.322

三、误差修正模型 (Error Correction Model ,ECM)

误差修正模型(ECM, 也称误差修正模型)是一种具有特定形式的计量经济模型。其基本思路是, 若变量间存在协整关系, 即表明这些变量间存在着长期稳定的关系, 而这种长期稳定的关系是在短期动态过程的不断调整下得以维持。产生这种结果的原因在于, 大多数的经济时间序列的一阶差分是平稳序列。同时, 存在着某种联系方式(如线性组合)把相互协整过程和长期稳定均衡状态结合起来。这时相互协整隐含的意义是: 即使所研究的水平变量各自都是一阶差分后平稳, 受支配于长期分量, 但这些变量的某些线性组合也可以是平稳的, 即所研究变量中的长期分量相互抵消, 产生了一个平稳的时间序列。之所以能够这样, 是因为一种调节过程(误差修正机制)在起作用, 防止了长期关系的偏差在规模或数量上的扩大。因此, 任何一组相互协整的时间序列变量都存在误差修正机制, 反映短期调节行为。

建立误差修正模型一般采用两步, 分别建立区分数据长期特征和短期待征的计量经济学模型。从理论上讲, 第一步, 建立长期关系模型。即通过水平变量和 OLS 法估计出时间序列变量间的关系。若估计结果形成平稳的残差序列时, 那么这些变量间就存在相互协整的关系。长期关系模型的变量选择是合理的, 回归系数具有经济意义。第二步, 建立短期动态关系。即误差修正方程。将长期关系模型中各变量以一阶差分形式重新加以构造, 并将长期关系模型所产生的残差序列作为解释变量引入, 在一个从一般到特殊的检验过程中, 对短期动态关系进行逐项检验, 不显著的项逐渐被剔除, 直到最适当的表示方法被找到为止。值得注意的是, 作为解释变量引入的长期关系模型的残差, 代表着在取得长期均衡的过程中各时点上出现“偏误”的程度, 使得第二步可以对这种偏误的短期调整或误差修正机制加以估计。

下面以建立我国货币需求函数为例, 说明误差修正模型的建模过程。

货币需求函数通常在局部调整的结构下加以设定。在这种模型中, 当前实际货币需求余额是关于实际货币需求余额滞后值、实际国民收入(通常用 GDP 表示)和机会成本等变量的回归。那么这种依据交易方程设定的模型可作为长期关系模型, 其一般形式为:

$$\left(\frac{M}{P}\right)_t = \beta_0 + \beta_1 Y_t + \beta_2 \pi_t + \beta_3 \left(\frac{M}{P}\right)_{t-1} + v_t \quad (10.13)$$

其中: M 为相应的名义货币余额, P 为物价指数(通常用 GDP 的平减指数表示), Y 为实际的国内生产总值(GDP), π 为季度通货膨胀率(根据综合物价指数衡量)。这里关于实际收入(产业规模)和机会成本变量的长期弹性分别由 $\beta_1/(1-\beta_3)$ 和 $\beta_2/(1-\beta_3)$ 给出。

第二阶段误差修正方程的一般形式是:

$$\Delta\left(\frac{M}{P}\right)_t = \alpha_0 + \sum_{i=0}^l \beta_i \Delta Y_{t-i} + \sum_{i=0}^l \gamma_i \Delta \pi_{t-i} + \sum_{i=0}^l \sigma_i \Delta\left(\frac{M}{P}\right)_{t-i-1} + \lambda EC_{t-1} + v_t \quad (10.14)$$

其中，EC 为长期关系模型中的残差。

在具体建模中，首先要对长期关系模型的设定是否合理进行单位根检验，以保证 EC 为平稳序列。其次，对短期动态关系中各变量的滞后项，进行从一般到特殊的检验，在这个检验过程中，不显著的滞后项逐渐被剔除，直到找出了最佳形式为止。通常滞后期在 $l=0,1,2,3$ 中进行试验。

第四节 案例分析

为了深入分析研究中国城镇居民的生活费支出与可支配收入的具体数量关系，收集了中国城镇居民月人均可支配收入（SR）和生活费支出（ZC）1992 年至 1998 年各月度数据序列（见表 10.3）。

表 10.3 城镇居民月人均生活费支出和可支配收入序列

序列	月份	1992	1993	1994	1995	1996	1997	1998
可支配收入 Sr	1	151.83	265.93	273.98	370.00	438.37	521.01	643.40
	2	159.86	196.96	318.81	385.21	561.29	721.01	778.62
	3	124.00	200.19	236.45	308.62	396.82	482.38	537.16
	4	124.88	199.48	248.00	320.33	405.27	492.96	545.79
	5	127.75	200.75	261.16	327.94	410.06	499.90	567.99
	6	134.48	208.50	273.45	338.53	415.38	508.81	555.79
	7	145.05	218.82	278.10	361.09	434.70	516.24	570.23
	8	138.31	209.07	277.45	356.30	418.21	509.98	564.38
	9	144.25	223.17	292.71	371.32	442.30	538.46	576.36
	10	143.86	226.51	289.36	378.72	440.81	537.09	599.40
	11	149.12	226.62	296.50	383.58	449.03	534.12	577.40
	12	139.93	210.32	277.60	427.78	449.17	511.22	606.14
生活费支出	1	139.47	221.74	234.28	307.10	373.58	419.39	585.70
	2	168.07	186.49	272.09	353.55	471.77	528.09	598.82
	3	110.47	185.92	202.88	263.37	350.36	390.04	417.27
	4	113.22	185.26	227.89	281.22	352.15	405.63	455.60
	5	115.82	187.62	235.70	299.73	369.57	426.81	466.20
	6	118.20	12.11	237.89	308.18	370.41	422.00	455.19
	7	118.03	186.75	239.71	315.87	376.90	428.70	458.57

Zc	8	124.45	187.07	252.52	331.88	387.44	459.29	475.40
	9	147.70	219.23	286.75	385.99	454.93	517.06	591.41
	10	135.14	212.80	270.00	355.92	403.77	463.98	494.57
	11	135.20	205.22	274.37	355.11	410.10	422.96	496.69
	12	128.03	192.64	250.01	386.08	400.48	460.92	516.16

数据来源：转摘自易丹辉《数据分析与 Eviews 的应用》，中国统计出版社 2002，P141。

由于所用数据为时间序列数据，需要检验其平稳性，并用 EG 两步法考察它们之间是否存在协整关系。

根据协整关系的检验方法，首先回答人均可支配收入（SR）和生活费支出（ZC）序列是否为非平稳序列，即考察其单整阶数。

在 Eviews 中具体操作过程如下：

在 Eviews 中建立文档，录入人均可支配收入（SR）和生活费支出（ZC）序列的数据。双击人均可支配收入（SR）序列，出现工作文件窗口，在其左上方点击 Eview 键出现下拉菜单，点击 Unit Root Test，出现对话框（图 10.2），选择带截距项（intercept），滞后差分项（Lagged differences）选 2 阶，点击 OK，得到估计结果，见表 10.4。

从检验结果看，在 1%、5%、10% 三个显著性水平下，单位根检验的 Mackinnon 临界值分别为 -3.5121、-2.8972、-2.5855，t 检验统计量值 -0.862611 大于相应临界值，从而不能拒绝 H_0 ，表明人均可支配收入（SR）序列存在单位根，是非平稳序列。

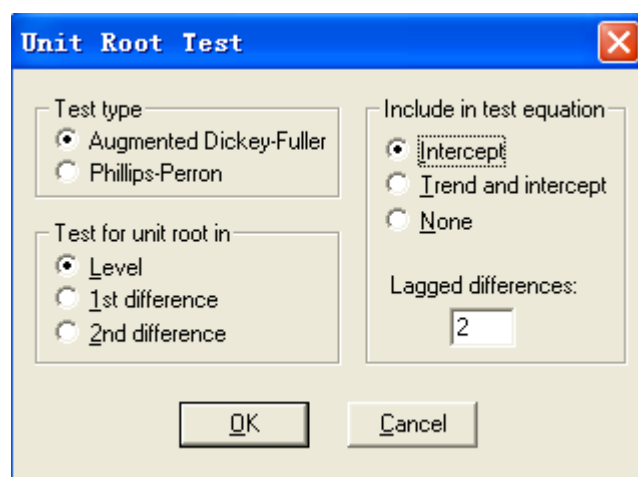


图 10.2 单位根检验回归方程设定（水平变量）

表 10.4 SR 序列的 ADF 检验结果为

ADF Test Statistic	-0.862611	1%	Critical Value*	-3.5121
		5%	Critical Value	-2.8972

10% Critical Value			-2.5855		
*MacKinnon critical values for rejection of hypothesis of a unit root.					
Augmented Dickey-Fuller Test Equation					
Dependent Variable: D(SR)					
Method: Least Squares					
Date: 06/08/05 Time: 10:31					
Sample(adjusted): 4 84					
Included observations: 81 after adjusting endpoints					
Variable	Coefficient	Std. Error	t-Statistic	Prob.	
SR(-1)	-0.034595	0.040105	-0.862611	0.3910	
D(SR(-1))	-0.409380	0.108905	-3.759060	0.0003	
D(SR(-2))	-0.336998	0.107273	-3.141502	0.0024	
C	22.63601	15.75919	1.436369	0.1549	
R-squared	0.221103	Mean dependent var		5.952346	
Adjusted R-squared	0.190756	S.D. dependent var		60.73081	
S.E. of regression	54.63220	Akaike info criterion		10.88725	
Sum squared resid	229820.1	Schwarz criterion		11.00549	
Log likelihood	-436.9334	F-statistic		7.285920	
Durbin-Watson stat	2.151282	Prob(F-statistic)		0.000230	

为了得到人均可支配收入（SR）序列的单整阶数，在单位根检验（Unit Root Test）对话框（图 10.3）中，指定对一阶差分序列作单位根检验，选择带截距项（intercept），滞后差分项（Lagged differences）选 2 阶，点击 OK，得到估计结果，见表 10.5。

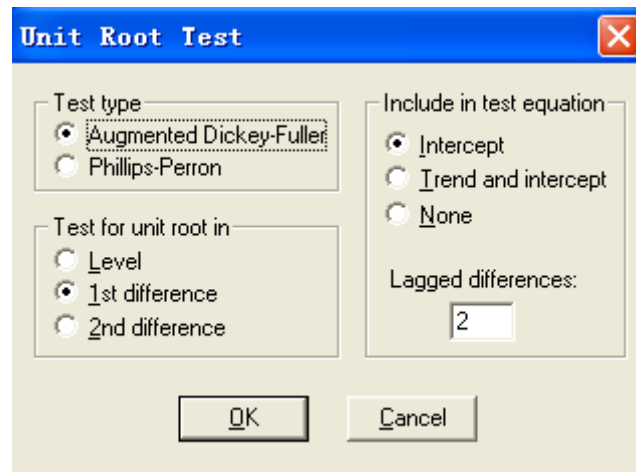


图 10.3 单位根检验回归方程设定（一阶差分序列）

表 10.5 SR 差分序列的 ADF 检验结果

ADF Test Statistic	-8.374339	1%	Critical Value*	-3.5132
		5%	Critical Value	-2.8976
		10%	Critical Value	-2.5858
*MacKinnon critical values for rejection of hypothesis of a unit root.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(SR,2)				
Method: Least Squares				
Date: 06/08/05 Time: 10:40				
Sample(adjusted): 5 84				
Included observations: 80 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(SR(-1))	-2.188331	0.261314	-8.374339	0.0000
D(SR(-1),2)	0.674099	0.190534	3.537949	0.0007
D(SR(-2),2)	0.225326	0.111513	2.020631	0.0468
C	12.59155	6.180708	2.037234	0.0451
R-squared	0.718058	Mean dependent var	0.348250	
Adjusted R-squared	0.706929	S.D. dependent var	99.32732	
S.E. of regression	53.77189	Akaike info criterion	10.85609	

Sum squared resid	219747.6	Schwarz criterion	10.97519
Log likelihood	-430.2434	F-statistic	64.51970
Durbin-Watson stat	2.095341	Prob(F-statistic)	0.000000

从检验结果看，在 1%、5%、10%三个显著性水平下，单位根检验的 Mackinnon 临界值分别为-3.5121、-2.8972、-2.5855，t 检验统计量值为-8.374339，小于相应临界值，从而拒绝 H_0 ，表明人均可支配收入（SR）的差分序列不存在单位根，是平稳序列。即 SR 序列是一阶单整的， $SR \sim I(1)$ 。

采用同样方法，可检验得到 ZC 序列也是一阶单整的，即 $ZC \sim I(1)$ 。

为了分析可支配收入（SR）和生活费支出（ZC）之间是否存在协整关系，我们先作两变量之间的回归，然后检验回归残差的平稳性。

以生活费支出（ZC）为被解释变量，可支配收入（SR）为解释变量，用 OLS 回归方法估计回归模型，结果见表 10.6。

表 10.6 ZC 对 SR 的 OLS 回归结果

Dependent Variable: ZC				
Method: Least Squares				
Date: 06/08/05 Time: 10:58				
Sample: 1 84				
Included observations: 84				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	18.98866	8.674160	2.189107	0.0314
SR	0.819677	0.021777	37.63950	0.0000
R-squared	0.945287	Mean dependent var	318.3649	
Adjusted R-squared	0.944620	S.D. dependent var	134.7917	
S.E. of regression	31.72051	Akaike info criterion	9.775326	
Sum squared resid	82507.66	Schwarz criterion	9.833202	
Log likelihood	-408.5637	F-statistic	1416.732	
Durbin-Watson stat	1.609062	Prob(F-statistic)	0.000000	

估计的回归模型为：

$$ZC_t = 18.98866 + 0.819677SR_t + \hat{u}_t \quad (10.15)$$

为了检验回归残差的平稳性，在工作文档窗口中，点击 **Genr** 功能键，命令 **ut=Resid**，将上述 OLS 回归得到的残差序列命名为新序列 **ut**，然后双击 **ut** 序列，对 **ut** 序列进行单位根检验。由于残差序列的均值为 0，所以选择无截距项、无趋势项的 DF 检验，模型设定见图 10.4，估计结果见表 10.7。

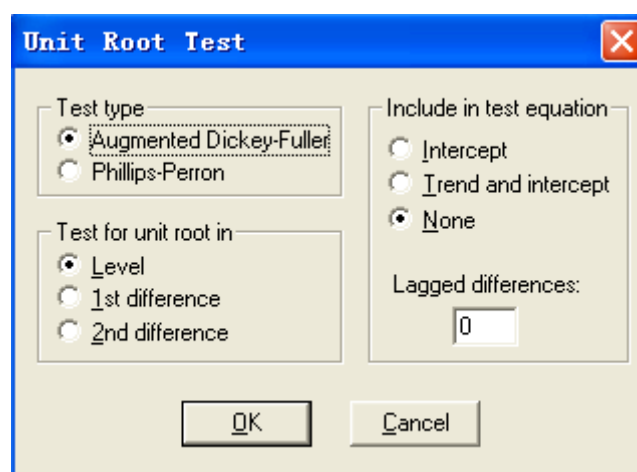


图 10.4 回归残差序列单位根检验的模型设定

表 10.7

ADF Test Statistic	-7.430111	1%	Critical Value*	-2.5909
		5%	Critical Value	-1.9441
		10%	Critical Value	-1.6178
*MacKinnon critical values for rejection of hypothesis of a unit root.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(UT)				
Method: Least Squares				
Date: 06/08/05 Time: 11:21				
Sample(adjusted): 2 84				
Included observations: 83 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
UT(-1)	-0.804627	0.108293	-7.430111	0.0000

R-squared	0.402360	Mean dependent var	0.051836
Adjusted R-squared	0.402360	S.D. dependent var	40.23706
S.E. of regression	31.10614	Akaike info criterion	9.724662
Sum squared resid	79342.53	Schwarz criterion	9.753805
Log likelihood	-402.5735	Durbin-Watson stat	1.973914

在 5% 的显著性水平下，t 检验统计量值为 -7.430111，大于相应临界值，从而拒绝 H_0 ，表明残差序列不存在单位根，是平稳序列，说明可支配收入（SR）和生活费支出（ZC）之间存在协整关系。

可支配收入（SR）和生活费支出（ZC）之间存在协整，表明两者之间有长期均衡关系。但从短期来看，可能会出现失衡，为了增强模型的精度，可以把协整回归（10.15）式中的误差项 \hat{u}_t 看作均衡误差，通过建立误差修正模型把生活费支出的短期行为与长期变化联系起来。误差修正模型的结构如下：

$$\Delta ZC_t = \alpha + \beta \Delta SR_t + \gamma \hat{u}_{t-1} + \varepsilon_t \quad (10.16)$$

在 Eviews 中，点击 Genr 功能键，生成可支配收入（SR）和生活费支出（ZC）的差分序列：

$$DZC_t = \Delta ZC_t = ZC_t - ZC_{t-1}$$

$$DSR_t = \Delta SR_t = SR_t - SR_{t-1}$$

然后以 DZC_t 作为被解释变量，以 DSR_t 和 \hat{u}_{t-1} 作为解释变量，估计回归模型（10.16），结果见表 10.8。

表 10.8

Dependent Variable: DZC				
Method: Least Squares				
Date: 07/03/05 Time: 21:30				
Sample(adjused): 2 84				
Included observations: 83 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.326424	3.456724	0.094432	0.9250

DSR	0.768942	0.059678	12.88490	0.0000
UT(-1)	-0.779148	0.113186	-6.883800	0.0000
R-squared	0.691102	Mean dependent var	4.538434	
Adjusted R-squared	0.683380	S.D. dependent var	55.71666	
S.E. of regression	31.35122	Akaike info criterion	9.763859	
Sum squared resid	78631.93	Schwarz criterion	9.851287	
Log likelihood	-402.2001	F-statistic	89.49261	
Durbin-Watson stat	1.996276	Prob(F-statistic)	0.000000	

最终得到误差修正模型的估计结果：

$$\begin{aligned}\hat{\Delta ZC}_t &= 0.3264 + 0.7689\Delta SR_t - 0.7791\hat{u}_{t-1} \\ t &= (0.094) \quad (12.88) \quad (-6.88) \\ R^2 &= 0.6911 \quad DW = 1.9963\end{aligned}$$

上述估计结果表明，城镇居民月人均生活费支出的变化不仅取决于可支配收入的变化，而且还取决于上一期生活费支出对均衡水平的偏离，误差项 u_t 的估计系数 -0.7791 体现了对偏离的修正，上一期偏离越远，本期修正的量就越大，即系统存在误差修正机制。

第十章小结

1、大多数经济时间序列是非平稳的，如果直接将非平稳时间序列当作平稳时间序列来进行回归分析，则可能造成“伪回归”，即变量间本来不存在相依关系，但回归结果却得出存在相依关系的错误结论。经济学家研究发现，造成“伪回归”的根本原因在于时序序列变量的非平稳性。

2、时间序列的平稳性，是指时间序列的统计规律不会随着时间的推移而发生变化。严格平稳是指随机过程 $\{Y_t\}$ 的联合分布函数与时间的位移无关。弱平稳是指随机过程 $\{Y_t\}$ 的一阶矩和二阶矩不随时间推移而变化。

3、单位根过程是最常见的非平稳过程。如果非平稳序列 $\{Y_t\}$ 经过 d 次差分后平稳，而 $d-1$ 次差分却不平稳，那么称 $\{Y_t\}$ 为 d 阶单整序列，记为 $\{Y_t\} \sim I(d)$ ， d 称为整形阶数。

4、时间序列平稳性的检验方法主要有两类：自相关函数检验法和单位根检验法。本书

只介绍最常用的单位根检验法——DF 检验法和 ADF 检验法。

5、协整是指多个非平稳经济变量的某种线性组合是平稳的。协整分析对于检验变量之间的长期均衡关系非常重要，而且也是区别真实回归与伪回归的有效方法。

6、任何一组相互协整的时间序列变量都存在误差修正机制。误差修正模型把长期关系和短期动态特征结合在一个模型中，既可以克服传统计量经济模型忽视伪回归的问题，又可以克服建立差分模型忽视水平变量信息的弱点。

第十章主要公式表

时间序列的平稳性	严格平稳	$F_{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}}(y_1, \dots, y_n) = F_{Y_{t_1+h}, \dots, Y_{t_n+h}}(y_1, \dots, y_n)$
	弱平稳	$E(Y_t) = \mu$, $Cov(Y_t, Y_s) = Cov(Y_{t+h}, Y_{s+h}) = r(t-s, 0) = r_{t-s}$
单位根过程	随机游动	$Y_t = \alpha + \beta t + Y_{t-1} + \varepsilon_t$ $\{\varepsilon_t\}$ 为白噪声序列。
	一般的单位根过程	$Y_t = \alpha + \beta t + Y_{t-1} + u_t$ $\{u_t\}$ 为一般平稳过程。
DF 检验	模型 I	$Y_t = \gamma Y_{t-1} + \varepsilon_t$
	模型 II	$Y_t = \alpha + \gamma Y_{t-1} + \varepsilon_t$
	模型 III	$Y_t = \alpha + \beta t + \gamma Y_{t-1} + \varepsilon_t$
ADF 检验	模型 I	$Y_t = \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$
	模型 II	$Y_t = \alpha + \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$
	模型 III	$Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \alpha_i \Delta Y_{t-i} + \varepsilon_t$
协整	$y_t \sim I(1), x_t \sim I(1)$	如果存在一组非零常数 α_1 、 α_2 ，使得 $\alpha_1 x_t + \alpha_2 y_t \sim I(0)$

思考题与练习题

思考题

- 10.1** 对时间序列进行分析，为什么提出平稳性问题？
- 10.2** 简述模型出现“伪回归”的含义。
- 10.3** 什么是非平稳？为什么随机游走过程是非平稳的？
- 10.4** 试述单位根检验的基本步骤。
- 10.5** 怎样判断变量之间是否存在协整关系。
- 10.6** 什么是误差修正机制？误差修正模型的特点是什么？

练习题

10.1 下表是某国的宏观经济数据（GDP——国内生产总值，单位：10 亿美元；PDI——一个人可支配收入，单位：10 亿美元；PCE——个人消费支出，单位：10 亿美元；利润——公司税后利润，单位：10 亿美元；红利——公司净红利支出，单位：10 亿美元）。

某国 1980 年到 2001 年宏观经济季度数据

季度	GDP	PDI	PCE	利润	红利	季度	GDP	PDI	PCE	利润	红利
Jan-80	2878.8	1990.6	1800.5	44.7	24.5	Jan-91	3860.5	2783.7	2475.5	159.35	64
Feb-80	2860.3	2020.1	1087.5	44.4	23.9	Feb-91	3844.4	2776.7	2476.1	143.7	68.4
Mar-80	2896.6	2045.3	1824.7	44.9	23.3	Mar-91	3864.5	2814.1	2487.4	147.6	71.9
Apr-80	2873.7	2045.2	1821.2	42.1	23.1	Apr-91	3803.1	2808.8	2468.8	140.3	72.4
Jan-81	2942.9	2073.9	1849.9	48.8	23.8	Jan-92	3756.1	2795	2484	114.4	70
Feb-81	2947.4	2098	1863.5	50.7	23.7	Feb-92	3771.1	2824.8	2488.9	114	68.4
Mar-81	2966	2106.6	1876.9	54.2	23.8	Mar-92	3754.4	2829	2502.5	114.6	69.2
Apr-81	2980.8	2121.1	1904.6	55.7	23.7	Apr-92	3759.6	2832.6	2539.3	109.9	72.5
Jan-82	3037.3	2129.7	1929.3	59.4	25	Jan-93	3783.3	2843.6	2556.5	113.6	77
Feb-82	3089.7	2149.1	1963.3	60.1	25.5	Feb-93	3886.5	2867	2604	133	80.5
Mar-82	3125.8	2193.9	1989.1	62.8	26.1	Mar-93	3944.4	2903	2639	145.7	83.1
Apr-82	3175.3	2272	2032.1	68.3	26.5	Apr-93	4012.1	2960.6	2678.2	141.6	84.2
Jan-83	3253.3	2300.7	2063.9	79.1	27	Jan-94	4221.8	3123.6	2824.3	125.2	87.2
Feb-83	3267.6	2315.2	2062	81.2	27.8	Feb-94	4144	3065.9	2741	152.6	82.2
Mar-83	3264.3	2337.9	2073.7	81.3	28.3	Mar-94	4166.4	3102.7	2754.6	141.8	81.7
Apr-83	3289.1	2382.7	2067.4	85	29.4	Apr-94	4194.2	3118.5	2784.8	136.3	83.4
Jan-84	3259.4	2334.7	2050.8	89	29.8	Jan-95	4221.8	3123.6	2824.9	125.2	87.2
Feb-84	3267.7	2304.5	2059	91.2	30.4	Feb-95	4254.8	3189.6	2849.7	124.8	90.8
Mar-84	3239.1	2315	2065.5	97.1	30.9	Mar-95	4309	3156.5	2893.3	129.8	94.1
Apr-84	3226.4	2313.7	2039.9	86.8	30.5	Apr-95	4333.5	3178.7	2895.3	134	97.4

Jan-85	3154	2282.5	2051.8	75.8	30	Jan-96	4390.5	3227.5	2922.4	109.2	105.1
Feb-85	3190.4	2390.3	2086.9	81	29.7	Feb-96	4387.7	3281.4	2947.9	106	110.7
Mar-85	3249.9	2354.4	2114.4	97.8	30.1	Mar-96	4412.6	3272.6	2993.4	111	112.3
Apr-85	3292.5	2389.4	2137	103.4	30.6	Apr-96	4427.1	3266.2	3012.5	119.2	111
Jan-86	3356.7	2424.5	2179.3	108.4	32.6	Jan-97	4460	3295.2	3011.5	140.2	108
Feb-86	3369.2	2434.9	2194.7	109.2	35	Feb-97	4515.3	3241.7	3045.8	157.9	105.5
Mar-86	3381	2444.7	2213	110	36.6	Mar-97	4559.3	3285.7	3075.8	169.1	105.1
Apr-86	3416.3	2459.5	2242	110.3	38.3	Apr-97	4625.5	3335.8	3074.6	176	106.3
Jan-87	2466.4	2463	2271.3	121.5	39.2	Jan-98	4655.3	3380.1	3128.2	195.5	109.6
Feb-87	3525	2490.3	2280.8	129.7	40	Feb-98	4704.8	3386.3	3147.8	207.2	113.3
Mar-87	3574.4	2541	2302.6	135.1	41.4	Mar-98	4779.7	3443.1	3170.6	213.4	117.5
Apr-87	3567.2	2556.2	2331.6	134.8	42.4	Apr-98	4779.7	3473.9	3202.9	226	121
Jan-88	3591.8	2587.3	2347.1	137.5	43.5	Jan-99	4809.8	3473.9	3200.9	221.3	124.6
Feb-88	3707.7	2631.9	2394	154	44.5	Feb-99	4832.4	3450.9	3208.6	206.2	127
Mar-88	3735.6	2653.2	2404.5	158	46.6	Mar-99	4845.6	3446.9	3241.1	195.7	129
Apr-88	3779.6	2680.9	2421.6	167.8	48.9	Apr-99	4859.7	3493	3241.6	203	130.7
Jan-89	3780.8	2699.2	2437.9	168.2	50.5	Jan-00	4880.8	3531.4	3258.8	199.1	132.3
Feb-89	3784.3	2697.6	2435.4	174.1	51.8	Feb-00	4832.4	3545.3	3258.6	193.7	132.5
Mar-89	3807.5	2715.3	2454.7	178.1	52.7	Mar-00	4903.3	3547	3281.2	196.3	133.8
Apr-89	3814.6	2728.1	2465.4	173.4	57.6	Apr-00	4855.1	3529.5	3251.8	199	136.2
Jan-90	3830.8	2742.9	2464.6	174.3	57.6	Jan-01	4824	3514.8	3241.1	189.7	137.8
Feb-90	3732.6	2692	2414.2	144.5	58.7	Feb-01	4840.7	3537.4	3252.4	182.7	136.7
Mar-90	3733.5	2722.5	2440.3	151	59.3	Mar-01	4862.7	3539.9	3271.2	189.6	138.1
Apr-90	3808.5	2777	2469.2	154.6	60.5	Apr-01	4868	3547.5	3271.1	190.3	138.5

(1) 画出利润和红利的散点图，并直观地考察这两个时间序列是否是平稳的。

(2) 应用单位根检验分别检验两个时间序列是否是平稳的。

10.2 下表数据是 1970-1991 年美国制造业固定厂房设备投资 Y 和销售量 X ，以 10 亿美元计价，且经过季节调整，根据该数据，判断厂房开支和销售量序列是否平稳？

年份	固定厂房设备 投资	销售量	年份	固定厂房设备 投资	销售量
1970	36.99	52.805	1981	128.68	168.129
1971	33.6	55.906	1982	123.97	163.351
1972	35.42	63.027	1983	117.35	172.547
1973	42.35	72.027	1984	139.61	190.682
1974	52.48	84.79	1985	182.88	194.538
1975	53.66	86.589	1986	137.95	194.657
1976	58.53	98.797	1987	141.06	206.326
1977	67.48	113.201	1988	163.45	223.541
1978	78.13	126.905	1989	183.8	232.724

1979	95.13	143.936	1990	192.61	239.459
1980	112.6	154.39	1991	182.81	235.142

10.3 根据习题 10.1 的数据，回答如下问题：

(1) 如果利润和红利时间序列并不是平稳的，而如果你以利润来回归红利，那么回归的结果会是虚假的吗？为什么？你是如何判定的，说明必要的计算。

(2) 取利润和红利两个时间序列的一阶差分，确定一阶差分时间序列是否是平稳的。

10.4 从《中国统计年鉴》中取得 1978 年—2005 年全国全社会固定资产投资额的时间序列数据，检验其是否平稳，并确定其单整阶数。

10.5 下表是 1978—2003 年中国财政收入 Y 和税收 X 的数据（单位：亿元），判断 $\ln Y$ 和 $\ln X$ 的平稳性，如果是同阶单整的，检验它们之间是否存在协整关系，如果协整，则建立相应的协整模型。

年度	财政收入 Y	税收 X	年度	财政收入 Y	税收 X
1978	1132.26	519.28	1995	6242.2	6038.04
1980	1159.93	571.7	1996	7407.99	6909.82
1985	2004.82	2040.79	1997	8651.14	8234.04
1989	2664.9	2727.4	1998	9875.95	9262.8
1990	2937.1	2821.86	1999	11444.08	10682.58
1991	3149.48	2990.17	2000	13395.23	12581.51
1992	3483.37	3296.91	2001	16386.04	15301.38
1993	4348.95	4255.3	2002	18903.64	17636.45
1994	5218.1	5126.88	2003	21715.25	20017.31

(1) **10.6** 下表是某地区消费模型建立所需的数据，对实际人均年消费支出 C 和人均年收入 Y（单位：元）

年份	人均消费 支出 C	人均年收 入 Y	年份	人均消费 支出 C	人均年收 入 Y
1950	92.28	151.20	1971	151.20	274.08
1951	97.92	165.60	1972	163.20	286.68
1952	105.00	182.40	1973	165.00	288.00
1953	118.08	198.48	1974	170.52	293.52
1954	121.92	203.64	1975	170.16	301.92
1955	132.96	211.68	1976	177.36	313.80
1956	123.84	206.28	1977	181.56	330.12

1957	137.88	255.48	1978	200.40	361.44
1958	138.00	226.20	1979	219.60	398.76
1959	145.08	236.88	1980	260.76	491.76
1960	143.04	245.40	1981	271.08	501.00
1961	155.40	240.00	1982	290.28	529.20
1962	144.24	234.84	1983	318.48	522.72
1963	132.72	232.68	1984	365.40	671.16
1964	136.20	238.56	1985	418.92	811.80
1965	141.12	239.88	1986	517.56	988.44
1966	132.84	239.04	1987	577.92	1094.64
1967	139.20	237.48	1988	655.76	1231.80
1968	140.76	239.40	1989	756.24	1374.60
1969	133.56	248.04	1990	833.76	1522.20
1970	144.60	261.48			

分别取对数，得到 lc 和 ly ：

(2) 对 lc 和 ly 进行平稳性检验。

(3) 用 EG 两步检验法对 lc 和 ly 进行协整性检验并建立误差修正模型。

分析该模型的经济意义。

第十一章 联立方程组模型

引子:

是先有鸡，还是先有蛋？

货币供应量及通货膨胀的关系倍受经济学家的关注。货币数量论认为：货币量增长是通货膨胀的主要原因。正如经济学家米尔顿·弗里德曼曾指出的：“通货膨胀永远而且处处是一种货币现象。”（曼昆著《经济学原理》246页）

也有经济学家认为：“人们持有货币是因为货币是交换媒介。与债券或股票这类其他资产不同，人们可以用货币购买他们购物单上的物品与劳务。他们为这种目的选择持有多少货币取决于这些物品与劳务的价格。价格越高，正常交易要求的货币越多。”“这就是说，物价水平上升（货币价值下降）增加了货币需求量。”（曼昆著《经济学原理》245页）

对货币供应量、经济增长及通货膨胀的关系也一直是各国政府和货币当局争论的问题。在出现通货膨胀时，政府强调是货币当局的货币供应量过多，使得总需求中的投资和消费过快增长，导致了通货膨胀；货币当局又争辩，是由于经济增长过快，投资和消费对货币需求增长，导致物价水平上升，而迫使货币供应量增加。究竟是物价上升导致货币供应量增加，还是货币供应量增加导致物价上涨？为了验证这种类似“是先有鸡，还是先有蛋？”的争论，在建立模型时，有人主张建立分析物价水平和经济增长影响货币供应量的方程；也有人主张建立分析货币供应量影响物价水平和经济增长的方程。这两个方程是什么关系？当经济增长、物价水平和货币供应量的样本数据都是既定的，两个方程可以同时估计吗？

迄今为止我们讨论的都是单一方程计量经济模型，但有的经济问题的计量需要运用联立方程模型。本章介绍联立方程模型的基础知识，包括联立方程模型的概念和类型、联立方程模型的识别问题及识别的方法、联立方程的估计方法等。

第一节 联立方程模型及其偏倚

一、联立方程模型的性质

单一方程模型中只有一个被解释变量，而有一个或多个解释变量，这类模型最主要的特征是被解释变量与解释变量间为一种单向的因果关系，通常解释变量是变化的原因，被解

释变量是变化的结果。单一方程模型中所研究的对象是单一的变量。但是，经济现象是错综复杂的，许多情况下所研究的问题不只是一个单一的变量，而是一个由多变量构成的经济系统。在经济系统中多个经济变量之间可能存在着双向的或多向的因果关系。例如，对某种商品的需求量 Q 的研究中，商品需求量 Q 受到商品价格 P 的影响，同时商品价格 P 又受到商品需求量 Q 的影响，这时需求量 Q 与价格 P 是相互影响，存在着双向的因果关系。在这种情况下，只用单一方程已经不能正确反映经济系统中诸多因素间的复杂关系了，而需要采用能够表现互为因果关系的联立方程模型。

所谓联立方程模型是指用若干个相互关联的单一方程，同时去表示一个经济系统中经济变量相互联立依存性的模型，即用一个联立方程组去表现多个变量间互为因果的联立关系。联立方程组中每一个单一方程中包含了一个或多个相互关联的内生变量，每一个方程的被解释变量都是内生变量，解释变量则可以是内生变量，也可以是外生变量。联立方程模型也称为联立方程组模型。

例如，商品需求与价格的模型，根据经济理论，商品的需求量 Q 受商品的价格 P 和消费者的收入 X 等因素的影响，可建立需求模型：

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 X_t + u_t \quad (11.1)$$

同时，该商品价格 P 也受商品需求量 Q 和其他代用商品价格 P^* 的影响，又可建立价格模型：

$$P_t = \beta_0 + \beta_1 Q_t + \beta_2 P_t^* + v_t \quad (11.2)$$

(11.1)和(11.2)式中的商品需求 Q 与商品价格 P ，事实上存在双向因果关系，不能只用单一方程模型去描述这种联立依存性，而需要把两个单一方程组成一个联立方程组，同时去研究商品的需求量 Q 和商品价格 P 的数量关系和变化规律，从而形成如下联立方程模型：

$$\begin{aligned} Q_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 X_t + u_t \\ P_t &= \beta_0 + \beta_1 Q_t + \beta_2 P_t^* + v_t \end{aligned} \quad (11.3)$$

又如，凯恩斯宏观经济模型，设变量有国民总收入 Y 、消费 C 、投资 I 、政府支出 G 。收入 Y 既是决定消费 C 和投资 I 的解释变量，同时又被消费 C 、投资 I 和政府支出 G 所决定。用联立方程组模型可清晰地描述它们之间的关系：

$$\begin{aligned} Y_t &= C_t + I_t + G_t \\ C_t &= \alpha_0 + \alpha_1 Y_t + u_{1t} \\ I_t &= \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_{2t} \end{aligned} \quad (11.4)$$

上式中 Y_{t-1} 表示收入 Y_t 的滞后一期数值，称为滞后内生变量。

与单一方程模型相比，联立方程模型有以下特点：

(1) 联立方程组模型是由若干个单一方程组成的。模型中不止一个被解释变量，通常建立 M 个方程就应有 M 个被解释变量

(2) 联立方程组模型里既有非确定性方程（即随机方程）又有确定性方程，但必须含有随机方程。

(3) 被解释变量和解释变量之间不仅是单向的因果关系，而可能是互为因果，有的变量在某个方程为解释变量，但同时在另一个方程中可能为被解释变量。因此解释变量有可能是随机的不可控变量。

(4) 解释变量可能与随机扰动项相关，违反 OLS 基本假定。如将(11.1)式代入(11.2)式

$$P_t = \beta_0 + \beta_1 P_t^* + \beta_2(\alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + v_t) + u_t \quad (11.5)$$

考虑 (11.2) 与 (11.5) 式，显然 P_t 不仅与 v_t 相关，而且与 u_t 相关。

二、联立方程模型中变量的类型

在单一方程模型中，被解释变量与解释变量的区分十分清晰，解释变量是变动的原因，被解释变量是变动的结果。在联立方程模型中，多个变量可能互为因果，同一变量可能作为被解释变量，同时又可能作为解释变量，显然将变量只是区分为解释变量与被解释变量的意义已经不大。而需要将变量区分为内生变量和外生变量。

在联立方程模型中，从变量的性质看，一些变量是由模型体现的经济系统本身所决定的，称为内生变量，内生变量的取值是模型求解的结果，由于受模型中随机扰动项的影响，内生变量是随机变量。例如 (11.3) 式模型中的需求量 Q 和商品价格 P ，它们的取值由模型所决定。同样，(11.4) 式模型中的收入 Y 、消费 C 和投资 I 也都是其取值由模型决定的内生变量。另一些变量是在模型体现的经济系统之外给定的，在模型中是非随机的，称为外生变量。例如 (11.3) 式模型中的消费者收入 X 和其他代用商品价格 P^* ，它们的取值是在模型之外因素决定的。同理，(11.4) 式模型中的政府支出 G 和收入的滞后值 Y_{t-1} 也都是由模型之外因素决定的外生变量。应当注意，一个变量在模型中是内生变量还是外生变量，是由经济理论和经济意义决定的，而不是从数学形式决定的。区分内生变量和外生变量对联立方程模型的估计和应用有重要意义。为了求解模型中的内生变量，一般说来联立方程中方程的个数应等于内生变量的个数。如果联立方程模型中内生变量的个数恰好等于方程组中方程的个数，则称该方程组为完备的。

在联立方程模型中，外生变量数值的变化能够影响内生变量的变化，而内生变量却不能反过来影响外生变量。对模型体系来讲，外生变量是由模型体系以外的因素所决定的，外生变量是可控制的变量，它与随机误差项不相关，所以是非随机变量。

在联立方程模型中，有一些变量本来是内生变量，但模型中可能出现了这些变量过去时期的滞后值或更大范围的数值。例如在（11.4）中，收入 Y 是内生变量，而模型中收入滞后值 Y_{t-1} 却不能由模型决定。像这样代表内生变量滞后值的变量称为滞后内生变量。在模型中滞后内生变量或更大范围的内生变量的作用视同于外生变量，并与外生变量一起称为前定变量。在单一方程模型中，前定变量一般作为解释变量，内生变量一般作为被解释变量；而在联立方程模型中，内生变量既可作为被解释变量，又可作为解释变量。

三、联立方程模型的偏倚性

在联立方程模型中，一个方程中的解释变量，在另一个方程中可以是被解释变量。因此联立方程模型很可能会违反古典假定。下面以（11.4）的宏观经济模型为例，来说明联立方程模型出现的问题，设宏观经济模型为

$$\begin{aligned} Y_t &= C_t + I_t + G_t \\ C_t &= \alpha_0 + \alpha_1 Y_t + u_{1t} \\ I_t &= \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_{2t} \end{aligned} \quad (11.6)$$

由第 1 个方程和第 2 个方程可以看出，因为变量 Y 与变量 C 有联系，并且变量 C 与随机误差项 u_1 相关，所以变量 Y 与 u_1 相关，而变量 Y 在第 2 个方程作解释变量，这就违背了解释变量与随机误差项不相关的假定。将第 2 个方程和第 3 个方程代入第 1 个方程，得

$$Y_t = \alpha_0 + \alpha_1 Y_t + u_{1t} + \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_{2t} + G_t \quad (11.7)$$

整理后得到如下结果

$$Y_t = \frac{\alpha_0 + \beta_0}{1 - \alpha_1 - \beta_1} + \frac{\beta_2}{1 - \alpha_1 - \beta_1} Y_{t-1} + \frac{1}{1 - \alpha_1 - \beta_1} G_t + \frac{1}{1 - \alpha_1 - \beta_1} (u_{1t} + u_{2t}) \quad (11.8)$$

由上式看出，变量 Y 与 $(u_{1t} + u_{2t})$ 相关，但在第 3 个方程里 Y 作为解释变量说明对投资的影响，这又违背了解释变量与随机误差项不相关的假定。当用普通最小二乘法去估计每一个方程时，如果解释变量与随机误差项相关，则参数的估计将是有偏的和不一致的（详细证明见附录 11.1）。这种由于联立方程模型中内生变量作为解释变量与随机误差项相关，而引起的 OLS 估计的参数有偏且不一致，称为联立方程偏倚性。联立方程偏倚性是联立方程固有的，所以一般情况下 OLS 法不适合于估计联立方程模型。

下面给出结构型模型的一个例子，设一个简化的凯恩斯宏观经济模型为

$$C_t = \beta_1 + \beta_2 Y_t + u_t \quad (11.11)$$

$$Y_t = C_t + I_t \quad (11.12)$$

其中 C 为消费， Y 为收入，它们是内生变量； I 是作为外生变量的投资； u 为随机误差项。

将上述结构型方程组表示成标准形式：

$$C_t - \beta_2 Y_t - \beta_1 + 0I_t = u_t \quad (11.13)$$

$$-C_t + Y_t + 0 - I_t = 0 \quad (11.14)$$

可用矩阵表示为

$$\begin{pmatrix} 1 & -\beta_2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} C_t \\ Y_t \end{pmatrix} + \begin{pmatrix} -\beta_1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ I_t \end{pmatrix} = \begin{pmatrix} u_t \\ 0 \end{pmatrix} \quad (11.15)$$

记矩阵为

$$\mathbf{B} = \begin{pmatrix} 1 & -\beta_2 \\ -1 & 1 \end{pmatrix} \quad \mathbf{\Gamma} = \begin{pmatrix} -\beta_1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} C_t \\ Y_t \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 \\ I_t \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_t \\ 0 \end{pmatrix}$$

结构型模型的矩阵形式可简记为：

$$\mathbf{BY} + \mathbf{\Gamma X} = \mathbf{U} \quad (11.16)$$

2、简化型模型

所谓简化型模型，是每个内生变量都只被表示为前定变量及随机扰动项函数的联立方程模型。

直观地看，在简化型模型中的每一个方程的右端不再出现内生变量。简化型模型的建立有两个实现的途径：一是直接写出模型的简化形式，在已知模型所包含的全部前定变量的条件下，将每个内生变量直接表示为前定变量和随机误差项的函数；二是通过结构型模型导出简化型模型，从结构型模型出发，经过代数运算，求解出内生变量，从而将每个内生变量用前定变量和随机误差项的函数来表示。

由（11.10）式的结构型模型，若 $|\mathbf{B}| \neq 0$ ，由矩阵知识知，内生变量结构型参数矩阵 \mathbf{B} 的逆矩阵 \mathbf{B}^{-1} 一定存在，对（11.10）式两端同时左乘 \mathbf{B}^{-1} ，得

$$\mathbf{Y} + \mathbf{B}^{-1}\mathbf{\Gamma X} = \mathbf{B}^{-1}\mathbf{u} \quad (11.17)$$

移项得

$$\mathbf{Y} = -\mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{X} + \mathbf{B}^{-1}\mathbf{u} \quad (11.18)$$

分别令

$$\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{\Gamma} \quad (11.19)$$

$$\mathbf{V} = \mathbf{B}^{-1}\mathbf{u} \quad (11.20)$$

则简化型模型的一般形式为

$$\mathbf{Y} = \mathbf{\Pi}\mathbf{X} + \mathbf{V} \quad (11.21)$$

在式(11.21)中， $\mathbf{\Pi}$ 表示简化型模型的参数矩阵， \mathbf{V} 表示简化型模型的随机误差项向量。

由(11.19)式可以看出，简化型模型的参数 $\mathbf{\Pi}$ 是结构型模型参数 \mathbf{B} 和 $\mathbf{\Gamma}$ 的函数。

事实上可以通过代数变换，将结构型模型转化为简化型模型。例如(11.11)和(11.12)式的结构型联立方程模型，将(11.12)代入(11.11)中得

$$C_t = \beta_1 + \beta_2(C_t + I_t) + u_t$$

即

$$C_t = \frac{\beta_1}{1-\beta_2} + \frac{\beta_2}{1-\beta_2}I_t + \frac{1}{1-\beta_2}u_t$$

由(11.12)，有 $C_t = Y_t - I_t$ ，代入((11.11)式得

$$Y_t - I_t = \beta_1 + \beta_2 Y_t + u_t$$

即

$$Y_t = \frac{\beta_1}{1-\beta_2} + \frac{1}{1-\beta_2}I_t + \frac{1}{1-\beta_2}u_t$$

因此，由(11.11)和(11.12)式的结构型联立方程模型导出的简化型模型为

$$C_t = \frac{\beta_1}{1-\beta_2} + \frac{\beta_2}{1-\beta_2}I_t + \frac{1}{1-\beta_2}u_t \quad (11.22)$$

$$Y_t = \frac{\beta_1}{1-\beta_2} + \frac{1}{1-\beta_2}I_t + \frac{1}{1-\beta_2}u_t \quad (11.23)$$

容易验证，用代数形式导出的简化型模型(11.22)和(11.23)式，与用(11.19)——(11.21)式矩阵导出的结果是一致的。

与结构型模型相比较，简化型模型有以下特点：

(1) 在简化型模型中每一个方程的右端不再出现内生变量，而只有前定变量作为解释变量。例如在式(11.22)和式(11.23)中，等式的右端只有前定变量 I 作为解释变量。

(2) 简化型模型中的前定变量与随机误差项不相关。事实上，因为简化型模型的随机误差项是结构型模型随机误差项的线性函数，而在结构型模型中的前定变量与随机误差项不相关，所以在简化型模型中同样有前定变量与随机误差项不相关。简化型模型中每个方程的解释变量全是前定变量，从而避免了联立方程偏倚。因此，从理论上讲可以对简化型模型的参数运用 OLS 法进行估计，只不过注意估计出的是简化型模型的参数估计值。但简化型模型中的参数是原结构型模型参数的函数，由估计的简化型模型参数，有可能求解出结构型参数。

(3) 简化型模型的参数综合反映了前定变量对内生变量的直接影响与间接影响，其参数表现了前定变量对内生变量的影响乘数。

(4) 在已知前定变量取值的条件下，可利用简化型模型参数的估计式直接对内生变量进行预测分析。

3、递归模型

所谓递归模型，是指在该模型中，第一个方程的内生变量 Y_1 仅由前定变量表示，而无其它内生变量；第二个方程内生变量 Y_2 表示成前定变量和一个内生变量 Y_1 的函数；第三个方程内生变量 Y_3 表示成前定变量和两个内生变量 Y_1 与 Y_2 的函数；按此规律下去，最后一个方程内生变量 Y_m 可表示成前定变量和 $m-1$ 个内生变量 Y_1, Y_2, \dots, Y_{m-1} 的函数。

例如，以三个内生变量 Y_1, Y_2, Y_3 和三个前定变量 X_1, X_2, X_3 为例，构造一个递归型联立方程组模型。

$$\begin{aligned} Y_1 &= \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + u_1 \\ Y_2 &= \alpha_{21}Y_1 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + u_2 \\ Y_3 &= \alpha_{31}Y_1 + \alpha_{32}Y_2 + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3 + u_3 \end{aligned} \quad (11.24)$$

将上式转化为标准形式

$$\begin{aligned} Y_1 - \beta_{11}X_1 - \beta_{12}X_2 - \beta_{13}X_3 &= u_1 \\ Y_2 - \alpha_{21}Y_1 - \beta_{21}X_1 - \beta_{22}X_2 - \beta_{23}X_3 &= u_2 \\ Y_3 - \alpha_{31}Y_1 - \alpha_{32}Y_2 - \beta_{31}X_1 - \beta_{32}X_2 - \beta_{33}X_3 &= u_3 \end{aligned} \quad (11.25)$$

(11.25) 式的矩阵形式为

$$\mathbf{BY} + \mathbf{\Gamma X} = \mathbf{u} \quad (11.26)$$

在式 (11.26) 中

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ -\alpha_{21} & 1 & 0 \\ -\alpha_{31} & -\alpha_{32} & 1 \end{pmatrix} \quad (11.27)$$

$$\mathbf{\Gamma} = - \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{pmatrix} \quad (11.28)$$

由此，我们看到内生变量参数矩阵 \mathbf{B} 是一个下三角阵，而前定变量的参数矩阵 $\mathbf{\Gamma}$ 只在原结构型模型中前定变量参数前多了一个负号。

递归模型是联立方程组模型中一种特殊的形式。它的特点是可以直接运用 OLS 方法对模型中的方程依次进行估计，而不会产生联立方程组的偏倚性问题。虽然满足内生变量递归特点的递归型模型确实存在，但在建模中并不多见。而且应指出，递归型模型中事实上没有变量间互为因果的特征，所以它并不是真正意义上的联立方程模型。

第二节 联立方程模型的识别

一、对模型识别的理解

由前面的讨论已知，简化型模型中的前定变量与随机扰动项不相关，避免了联立方程偏倚，因此对简化型模型一般可以运用 OLS 法估计其参数。然而通常的研究目的是要获得结构型模型的参数估计值，虽然已知结构型模型的参数是简化型模型参数的函数，但能否从简化型参数求解出结构型参数呢？这涉及到联立方程模型的识别问题。

联立方程模型的识别可以从多方面去理解，从根本上说识别是模型的设定问题。

例如，设农产品供需均衡模型为

$$Q_d = \alpha_0 + \alpha_1 p + u_1 \quad (11.29)$$

$$Q_s = \beta_0 + \beta_1 p + u_2 \quad (11.30)$$

$$Q_d = Q_s \quad (11.31)$$

对于方程 (11.29) 和 (11.30)，由于在均衡条件下，农产品的供给与需求是一致的，所以，这时用 OLS 法估计其参数，那么无法区分估计出的参数究竟是需求方程的还是供给方程的，这就是联立方程组模型的识别问题。

又如，设宏观经济模型为

$$Y_t = C_t + I_t \quad (11.32)$$

$$C_t = \alpha_0 + \alpha_1 Y_t + u_{1t} \quad (11.33)$$

$$I_t = \beta_0 + \beta_1 Y_t + u_{2t} \quad (11.34)$$

其中 Y 为国民总收入， C 为消费， I 为投资。

(11.33) 与 (11.34) 式分别是消费函数和投资函数的参数，在经济意义应该是惟一的，但经过一定的数学变换，可以发现事实并非如此。由 (11.32) 式移项得

$$I_t = Y_t - C_t \quad (11.35)$$

将 (11.35) 式代入投资函数 (11.34) 式得

$$Y_t - C_t = \beta_0 + \beta_1 Y_t + u_{2t}$$

$$\text{则有} \quad C_t = -\beta_0 + (1 - \beta_1)Y_t - u_{2t} \quad (11.36)$$

比较式消费函数 (11.33) 与投资函数 (11.36) 式，可以看出二者变量都是 C_t 和 Y_t 。现在的问题是，通过样本数据 C_t 和 I_t 所估计的参数究竟是消费函数的参数还是投资函数的参数呢？显然这时联立方程模型有无法识别的问题。

从上述两个例子，可以看到联立方程确实存在识别问题。联立方程模型的识别可以从多方面去理解。对联立方程识别最直观的理解，是看能否从简化型模型参数估计值中合理地求解出结构型模型参数的估计值。如果结构型模型参数的估计值能由简化型模型的参数求解出，则称这个结构方程是可识别的，否则是不可识别的。从理论上，也可从方程是否具有确定的统计形式去认识联立方程的识别。如果模型中一个结构方程与另一个结构方程含有相同的变量（包括解释变量与被解释变量），而且变量之间具有相同的统计关系，则这两个方程具有相同的统计形式，则它们都是不可识别的。此外，也可以从方程中是否排除了必要的变量去理解。如果一个结构方程包含了模型的所有变量，或者说该结构方程的变量系数均未实行零限制，则称该方程为不可识别。反过来，当模型中的结构方程有零限制出现时，即某些变量不出现在模型中某个结构方程里时，则该方程才有可能被识别。

关于模型识别的定义是针对结构型联立方程组来说的，在结构型模型中，除了定义方程、均衡方程（定义方程）不存在识别问题，而每一个需要估计参数的结构方程都有识别问题。如果结构型模型中的每一个结构方程都是可识别的，则称该联立方程模型是可识别的。在结构型模型中，只要有一个结构方程不可识别，则该联立方程模型就是不可识别的。

二、联立方程模型识别的类型

由于模型提供的信息有差异，联立方程模型的识别性质可分为三种类型：不可识别、恰好识别和过度识别。

1、不可识别

如果结构型模型中某个方程参数的估计值不能够由简化型模型参数估计值求解出，则称该方程是不可识别。

例如，商品需求与供给的结构型模型为

$$Q_t^d = \alpha_1 + \alpha_2 P_t + u_{1t} \quad (11.37)$$

$$Q_t^s = \beta_1 + \beta_2 P_t + u_{2t} \quad (11.38)$$

$$Q_t^d = Q_t^s \quad (11.39)$$

由均衡条件（11.39），可导出内生变量 P 与 Q 的简化型模型为

$$P_t = \pi_1 + v_{1t} \quad (11.40)$$

$$Q_t = \pi_2 + v_{2t} \quad (11.41)$$

$$\text{其中 } \pi_1 = \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2}; \quad \pi_2 = \frac{\alpha_2 \beta_1 - \alpha_1 \beta_2}{\alpha_2 - \beta_2}; \quad v_{1t} = \frac{u_{2t} - u_{1t}}{\alpha_2 - \beta_2}; \quad v_{2t} = \frac{\alpha_2 u_{2t} - \beta_2 u_{1t}}{\alpha_2 - \beta_2}$$

在上述简化型模型与结构型模型参数的关系式中，由估计的两个简化型参数 $\hat{\pi}_1$ 与 $\hat{\pi}_2$ ，无法求解出结构型模型的 4 个参数 α_1 、 α_2 、 β_1 、 β_2 。因此，方程式（11.37）和（11.38）为不可识别，从而该联立方程组模型是不可识别。直观的理解，这是因为供给方程和需求方程的结构形式一致，没有提供分别估计各个结构参数的足够信息，或者说对从模型的设定上方程没有施加足够的约束。

2、恰好识别

如果结构型模型中某个方程的参数能够由简化型模型参数估计值惟一地解出，则称该方程是恰好识别。

例如，对上述需求与供给结构型模型补充一些信息，在供给函数中引入前定变量价格的滞后值，即上一期的价格 P_{t-1} ，这时需求与供给模型为

$$\begin{aligned} Q_t^d &= \alpha_1 + \alpha_2 P_t + u_{1t} \\ Q_t^s &= \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + u_{2t} \\ Q_t^d &= Q_t^s \end{aligned} \quad (11.42)$$

这时，需求供给的简化型模型为

$$\begin{aligned} P_t &= \pi_{11} + \pi_{12}P_{t-1} + v_{1t} \\ Q_t &= \pi_{21} + \pi_{22}P_{t-1} + v_{2t} \end{aligned} \quad (11.43)$$

其中

$$\begin{aligned} \pi_{11} &= \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} & \pi_{12} &= \frac{\beta_3}{\alpha_2 - \beta_2} \\ \pi_{21} &= \frac{\alpha_2\beta_1 - \beta_2\alpha_1}{\alpha_2 - \beta_2} & \pi_{22} &= \frac{\alpha_2\beta_3}{\alpha_2 - \beta_2} \\ v_{1t} &= \frac{u_{2t} - u_{1t}}{\alpha_2 - \beta_2} & v_{2t} &= \frac{\alpha_2u_{2t} - \beta_2u_{1t}}{\alpha_2 - \beta_2} \end{aligned}$$

从简化型模型与结构型模型参数的关系可以看到，这时简化型模型的参数个数为 4 个，结构型模型的参数个数为 5 个，因此也不能在已知简化型模型参数估计值的条件下，惟一地解出结构型模型的所有参数。但可以看出，需求方程的参数 α_1 和 α_2 是可以被惟一求解出，即

$$\alpha_1 = \pi_{21} - \alpha_2\pi_{11} \quad \alpha_2 = \frac{\pi_{22}}{\pi_{12}}$$

即是说，此时需求方程是恰好识别的；而供给方程的参数估计值不能被唯一解出，故供给方程过度识别。

上述例子给出一个启示，模型中引进新的前定变量 P_{t-1} 后，能使不可识别的模型向可以识别转变，这为改进模型的识别状态提供了重要线索。很自然，如果继续对模型补充信息，再引进前定变量，模型的识别状况会进一步变好吗？

例如，在需求方程中再引进一个新的前定变量收入 I_t ，这时模型为

$$\begin{aligned} Q_t^d &= \alpha_1 + \alpha_2P_t + \alpha_3I_t + u_{1t} \\ Q_t^s &= \beta_1 + \beta_2P_t + \beta_3P_{t-1} + u_{2t} \\ Q_t^d &= Q_t^s \end{aligned} \quad (11.44)$$

由该结构型模型导出简化型模型

$$\begin{aligned} P_t &= \pi_{11} + \pi_{12}I_t + \pi_{13}P_{t-1} + u_{1t} \\ Q_t &= \pi_{21} + \pi_{22}I_t + \pi_{23}P_{t-1} + u_{2t} \end{aligned} \quad (11.45)$$

其中

$$\pi_{11} = \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} \quad \pi_{12} = \frac{-\alpha_3}{\alpha_2 - \beta_2} \quad \pi_{13} = \frac{\beta_3}{\alpha_2 - \beta_2}$$

$$\pi_{21} = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\alpha_2 - \beta_2} \quad \pi_{22} = \frac{\alpha_3\beta_1}{\alpha_2 - \beta_2} \quad \pi_{23} = \frac{\alpha_2\beta_3}{\alpha_2 - \beta_2};$$

$$v_{1t} = \frac{u_{2t} - u_{1t}}{\alpha_2 - \beta_2} \quad v_{2t} = \frac{\alpha_2 u_{2t} - \beta_2 u_{1t}}{\alpha_2 - \beta_2}$$

由上述简化型模型与结构型模型参数的关系可以看出，简化型模型的参数是 6 个，结构型模型的参数也是 6 个，所以由简化型模型的参数估计值可以惟一地求解出结构型模型的参数，即

$$\begin{aligned} \alpha_1 &= \pi_{21} - \alpha_2\pi_{11} & \alpha_2 &= \frac{\pi_{23}}{\pi_{13}} & \alpha_3 &= \alpha_2\pi_{12} - \pi_{22} \\ \beta_1 &= \pi_{21} - \beta_2\pi_{11} & \beta_2 &= \frac{\pi_{22}}{\pi_{12}} & \beta_3 &= \pi_{23} - \beta_2\pi_{13} \end{aligned}$$

这表明该联立方程模型中的每一个方程都是恰好识别，所以联立方程模型是恰好识别。

3、过度识别

如果结构型模型中某个方程的参数能够由简化型模型参数估计值解出，但求解出的值不惟一，则称该方程是过度识别。

例如，在需求方程中再继续引进一个前定变量消费者拥有的财富 R_t ，这时模型为

$$\begin{aligned} Q_t^d &= \alpha_1 + \alpha_2 P_t + \alpha_3 I_t + \alpha_4 R_t + u_{1t} \\ Q_t^s &= \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + u_{2t} \\ Q_t^d &= Q_t^s \end{aligned} \quad (11.46)$$

由该结构型模型求出简化型模型为

$$\begin{aligned} P_t &= \pi_{11} + \pi_{12} I_t + \pi_{13} R_t + \pi_{14} P_{t-1} + u_{1t} \\ Q_t &= \pi_{21} + \pi_{22} I_t + \pi_{23} R_t + \pi_{24} P_{t-1} + u_{2t} \end{aligned} \quad (11.47)$$

其中

$$\begin{aligned} \pi_{11} &= \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} & \pi_{12} &= \frac{-\alpha_3}{\alpha_2 - \beta_2} & \pi_{13} &= \frac{-\alpha_4}{\alpha_2 - \beta_2} & \pi_{14} &= \frac{\beta_3}{\alpha_2 - \beta_2} \\ \pi_{21} &= \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\alpha_2 - \beta_2} & \pi_{22} &= \frac{\alpha_3\beta_2}{\alpha_2 - \beta_2} & \pi_{23} &= \frac{-\alpha_4\beta_2}{\alpha_2 - \beta_2} & \pi_{24} &= \frac{\alpha_2\beta_3}{\alpha_2 - \beta_2} \\ v_{1t} &= \frac{u_{2t} - u_{1t}}{\alpha_2 - \beta_2} & v_{2t} &= \frac{\alpha_2 u_{2t} - \beta_2 u_{1t}}{\alpha_2 - \beta_2} \end{aligned}$$

从简化型模型与结构型模型参数的关系看出，简化型模型的参数为 8 个，而这时结构型模型的参数为 7 个。虽然可以从参数的关系式求解出结构型模型的参数，但解并不惟一。例如，

该结构型联立方程模型的供给方程中价格 P_i 的系数 β_2 ，就可由上述关系式导出两个表达式，即 $\beta_2 = \pi_{22}/\pi_{12}$ 和 $\beta_2 = \pi_{23}/\pi_{13}$ ，产生这样问题的原因，是为需求方程提供了过多的信息，或者说为供给方程施加了过多的约束，即供给方程不仅排除了收入变量，而且还排除了财产变量，因而供给方程是过度识别的。

三、联立方程模型识别的方法

由上述商品需求与供给联立方程模型的例子可以看出，从简化型模型与结构型模型参数的关系去判断模型的可识别性，实际上是非常麻烦的，特别是联立方程模型规模很大的时候。因此，需要寻求更为规范的方法对联立方程模型的识别性进行判断。这类规范的识别方法主要是模型识别的阶条件和秩条件。

1、模型识别的阶条件

模型识别阶条件的基本思想是，一个结构型方程的识别取决于不包含在这个方程中，而包含在模型其他方程中变量的个数，可从这类变量的个数去判断方程的识别性质。

如果模型中有 M 个方程，共有 M 个内生变量和 K 个前定变量；其中第 i 个方程包含 m_i 个内生变量和 k_i 个前定变量。模型识别的阶条件可以表述为：当模型的一个方程中不包含的变量（内生变量和前定变量）的总个数，大于或等于模型中内生变量总个数 M 减 1，则该方程能够识别。这就是说，被模型中第 i 个方程排除的变量个数为 $(M + K) - (m_i + k_i)$ ，当第 i 个方程是可识别时，必须有

$$(M + K) - (m_i + k_i) \geq M - 1 \quad (11.48)$$

整理后可得

$$K - k_i \geq m_i - 1 \quad (11.49)$$

即没有包含在第 i 个方程中的前定变量个数 $K - k_i$ ，大于或等于出现在该方程的内生变量个数 m_i 减 1。

由模型识别的阶条件可以判断：当 $K - k_i > m_i - 1$ 时，则第 i 方程是过度识别；当 $K - k_i = m_i - 1$ 时，则第 i 方程是恰好识别；当 $K - k_i < m_i - 1$ 时，则第 i 方程是不可识别。

例如，设定的联立方程组模型为

$$Y_t = C_t + I_t + G_t \quad (11.50)$$

$$C_t = \alpha_1 + \alpha_2 Y_t - \alpha_3 T_t + u_{1t} \quad (11.51)$$

$$I_t = \beta_1 + \beta_2 Y_t - \beta_3 Y_{t-1} + u_{2t} \quad (11.52)$$

$$T_t = \gamma_1 + \gamma_2 Y_t + u_{3t} \quad (11.53)$$

模型中有 Y_t 、 C_t 、 I_t 和 T_t 等 $M=4$ 个内生变量，有 G_t 和 Y_{t-1} 等 $K=2$ 个前定变量。下面分别对模型中的每一个方程用阶条件进行判断。

方程 (11.51) 有 $m_2 = 3, k_2 = 0$ ，这时 $K - k_2 = 2 - 0 = 2$ ，而 $m_2 - 1 = 3 - 1 = 2$ ，结果相等，所以，该方程可能是恰好识别。

方程 (11.52) 有 $m_3 = 2, k_3 = 1$ ，这时 $K - k_3 = 2 - 1 = 1$ ，而 $m_3 - 1 = 2 - 1 = 1$ ，结果相等，所以，该方程可能是恰好识别。

方程 (11.53)，有 $m_4 = 2, k_4 = 0$ ，这时 $K - k_4 = 2 - 0 = 2$ ，而 $m_4 - 1 = 2 - 1 = 1$ ，则 $K - k_4 > m_4 - 1$ ，所以，该方程可能是过度识别。

由于方程 (11.50) 为定义方程式，故不需判断其识别性。综合上述判断，该模型有可能是可识别。

应当指出，模型识别的阶条件只是联立方程模型中方程识别状态的必要条件，但非充分条件。还需要寻求用联立方程模型识别的充分必要条件去加以判断。

2、模型识别的秩条件

模型识别的阶条件还不是识别的充分条件，即是说，方程不满足识别的阶条件时，方程是不可识别的；但方程满足识别的阶条件时，并非一定是可识别的。例如(11.47)式的供给方程中，没有包含需求方程 (11.46) 式中的收入变量 I ，按照阶条件供给方程是可识别的。但是当需求方程 (11.46) 中收入 I 的系数 α_3 为 0 时，表明收入 I 仅是可能而实际并没有列入需求方程，这时还不能确保供给方程(11.47)是可识别的。此时需要运用联立方程模型识别的充分必要条件——秩条件。

联立方程模型识别的秩条件可以表述为：在有 M 个内生变量 M 个方程的完整联立方程模型中，当且仅当一个方程中不包含但在其他方程包含的变量（不论是内生变量还是外生变量）的结构参数，至少能够构成一个非零的 $M-1$ 阶行列式时，该方程是可以识别的。或者表述为，当且仅当一个方程所排斥（不包含）的变量的参数矩阵的秩等于 $M-1$ 时，该方程可以识别。

设结构型模型为

$$BY + \Gamma X = U$$

在上式中 B 为内生变量的系数矩阵, Γ 为前定变量的系数矩阵, 记矩阵 (B_0, Γ_0) 为该方程组中第 i 个方程中没有包含的内生变量和前定变量系数所构成的矩阵, 如果当 (B_0, Γ_0) 的秩为 $M-1$ 时, 即只有当至少有一个 $M-1$ 阶非零行列式时, 该方程才是可识别的。

类似阶条件有三种情况, 秩条件也有三种情况: 当只有一个 $M-1$ 阶非零行列式时, 该方程是恰好识别; 当不止一个 $M-1$ 阶非零行列式时, 该方程是过度识别; 当不存在 $M-1$ 阶非零行列式时, 该方程是不可识别。

运用秩条件判别模型的识别性, 步骤如下:

(1) 将结构型模型转变为结构型模型的标准形式, 并将全部参数列成完整的参数表(方程中不出现变量的参数以 0 表示);

(2) 考察第 i 个方程的识别问题: 划去该方程的那一行, 并划去该方程出现的变量的系数(该行中非 0 系数)所在列, 余下该方程不包含的变量在其他方程中的系数的矩阵 (B_0, Γ_0) ;

(3) 计算 $\text{Rank}(B_0, \Gamma_0)$, 检验所余系数矩阵 (B_0, Γ_0) 的秩, 看是否等于 $M-1$, 或检验所余系数是否能构成非零 $M-1$ 阶行列式。

(4) 判断: 如果 $\text{Rank}(B_0, \Gamma_0) = M-1$, 则该方程为可识别; 根据非零行列式个数判别是恰好识别, 还是过度识别。

例如, 设定的联立方程模型为

$$Y_t = C_t + I_t + G_t \quad (11.56)$$

$$C_t = \alpha_1 + \alpha_2 Y_t - \alpha_3 T_t + u_{1t} \quad (11.57)$$

$$I_t = \beta_1 + \beta_2 Y_t - \beta_3 Y_{t-1} + u_{2t} \quad (11.58)$$

$$T_t = \gamma_1 + \gamma_2 Y_t + u_{3t} \quad (11.59)$$

模型中, $M=4$ 个内生变量, Y_t 、 C_t 、 I_t 、 T_t 分别是收入、消费、投资、税收; 前定变量 G_t 和 Y_{t-1} 分别是政府支出和上年收入。

由给定方程组模型写出其结构性模型的标准形式

$$-\alpha_1 + C_t + 0I_t - \alpha_2 Y_t + \alpha_3 T_t + 0G_t + 0Y_{t-1} = u_{1t} \quad (11.60)$$

$$-\beta_1 + 0C_t + I_t - \beta_2 Y_t + 0T_t + 0G_t + \beta_3 Y_{t-1} = u_{2t} \quad (11.61)$$

$$-\gamma_1 + 0C_t + 0I_t - \gamma_2 Y_t + T_t + 0G_t + 0Y_{t-1} = u_{3t} \quad (11.62)$$

$$0 - C_t - I_t + Y_t + 0T_t - G_t - 0Y_{t-1} = 0 \quad (11.63)$$

由结构型的标准形式写出其系数矩阵 (B, Γ) ，即

$$(B, \Gamma) = \begin{pmatrix} -\alpha_1 & 1 & 0 & -\alpha_2 & \alpha_3 & 0 & 0 \\ -\beta_1 & 0 & 1 & -\beta_2 & 0 & 0 & \beta_3 \\ -\gamma_1 & 0 & 0 & -\gamma_2 & 1 & 0 & 0 \\ 0 & -1 & -1 & 1 & 0 & -1 & 0 \end{pmatrix}$$

或者将以上一般形式的结构参数列于表 11.1

表 11.1

		C	I	Y	T	G	Y_{t-1}
方程 1	$-\alpha_1$	1	0	$-\alpha_2$	α_3	0	0
方程 2	$-\beta_1$	0	1	$-\beta_2$	0	0	β_3
方程 3	$-\gamma_1$	0	0	$-\gamma_2$	1	0	0
方程 4	0	-1	-1	1	0	-1	0

下面利用秩条件判断该模型的识别性。

(1) 分析消费函数方程 1 的识别问题，划去方程 1 的那一行，并划去该行中非 0 系数所在列（即 C、Y、T 对应的列），余下方程 1 不包含的变量在其他方程中的系数，构成 (B_0, Γ_0) ，并列于表 11.2:

$$(B_0, \Gamma_0) = \begin{pmatrix} 1 & 0 & \beta_3 \\ 0 & 0 & 0 \\ -1 & -1 & 0 \end{pmatrix}$$

所余系数矩阵 (B_0, Γ_0) 能构成 $M-1=3$ 阶行列式:

$$\begin{vmatrix} 1 & 0 & \beta_3 \\ 0 & 0 & 0 \\ -1 & -1 & 0 \end{vmatrix} = 0$$

(B_0, Γ_0) 只能构成一个等于零的 $M-1$ 阶行列式，或者说 $\text{Rank}(B_0, \Gamma_0) < M-1$ ，这说明消费函数是不可识别的。值得注意的是，在阶条件的判断中该方程是有可能为恰好识别（见式 (11.51) 的阶条件判断），这一例子正好说明阶条件只是必要条件，而非充分条件，亦即满

足阶条件的未必一定满足秩条件。

(2) 分析投资函数方程 2 的识别问题, 同样道理可以划去方程 2 的那一行, 并划去该行中非 0 系数所在列 (即 I 、 Y 和 Y_{t-1} 对应的列), 余下方程 2 不包含的变量在其他方程中的

系数, 构成 (B_0, Γ_0) , 得到 $(B_0, \Gamma_0) = \begin{pmatrix} 1 & \alpha_3 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & -1 \end{pmatrix}$, 其行列式为

$$\begin{vmatrix} 1 & \alpha_3 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & -1 \end{vmatrix} \neq 0$$

只能构成一个不等于零的 $M-1$ 阶行列式, 则说明 $\text{Rank}(B_0, \Gamma_0) = M-1=3$, 即表明投资函数为恰好识别。

(3) 分析税收函数方程 3 的识别问题, 可以划去方程 3 的那一行, 并划去该行中非 0 系数所在列 (即 Y 和 T 对应的列), 余下方程 3 不包含的变量在其他方程中的系数, 构成 (B_0, Γ_0) , 得到 (B_0, Γ_0) 为

$$(B_0, \Gamma_0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \beta_3 \\ -1 & -1 & -1 & 0 \end{bmatrix}$$

这是一个三行四列的矩阵, 故可构成四个三阶行列式, 即

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & -1 \end{vmatrix} \quad \begin{vmatrix} 1 & 0 & 0 \\ 0 & 0 & \beta_3 \\ -1 & -1 & 0 \end{vmatrix} \quad \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & \beta_3 \\ -1 & -1 & 0 \end{vmatrix} \quad \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & \beta_3 \\ -1 & -1 & 0 \end{vmatrix}$$

很明显在这四个三阶行列式里只有

$$\begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & \beta_3 \\ -1 & -1 & 0 \end{vmatrix} = 0$$

其余三个均为非零行列式, 则表明税收函数是过度识别。

最后一个方程为恒定式, 可以不需判断其识别性。综上所述, 由于消费函数是不可识别, 所以, 整个方程组为不可识别。

3、模型识别的一般步骤和经验方法

从前面的介绍可以看出, 模型识别的秩条件是充分必要条件, 但识别程序过于繁琐; 模

型识别的阶条件比较简便,但又只是必要条件。在用联立方程模型作实际的计量经济研究时,为了简化识别的工作量,可以将两种方法结合运用。首先用阶条件判断方程是否可以识别,如果不可识别,说明不满足识别的必要条件,即可作出结论。如果阶条件显示可以识别,因还不是充分条件,再用秩条件分析其充分条件是否满足,若不满足即可作出不可识别的结论。若秩条件表明是可识别的,再用阶条件分析究竟是恰好识别,还是过度识别。模型识别的一般步骤如图 11.1 所示:

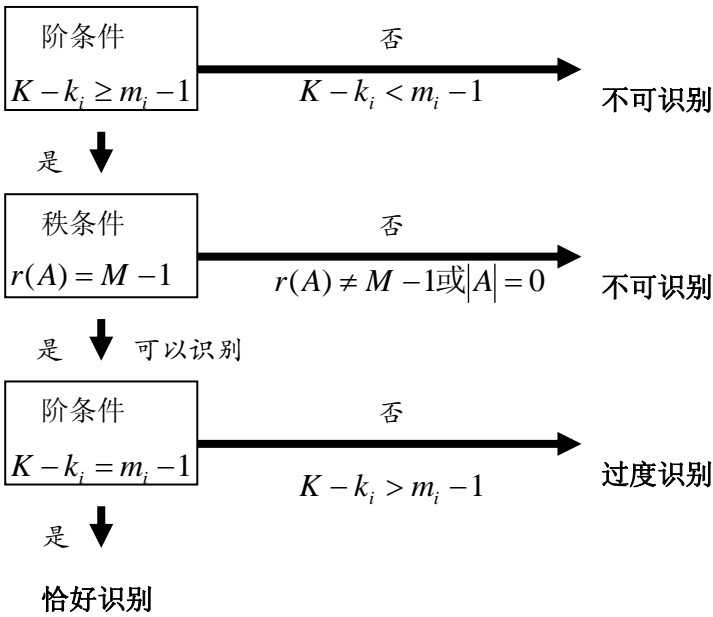


图 11.1 模型识别的一般步骤

模型的识别不是统计问题,而是模型的设定问题,因此在设定模型时就应设法尽量保证模型的可识别性。一般说来在设定联立方程模型时应遵循以下原则:“在建立联立方程结构模型时,要使新引入的方程中包含前面已引入的每一个方程都不包含的至少 1 个变量(内生变量或前定变量);同时,要使前面已引入的每一个方程都包含至少 1 个新引入方程未包含的变量,并要互不相同。”因为只有新引入的方程包含前面每一个方程都不包含的至少 1 个变量,才能保证不破坏前面已有方程的可识别性。而且,只有前面每一个方程都包含至少 1 个新引入方程所未包含的变量,才能保证新引入的方程是可识别的。

第三节 联立方程模型的估计

一、联立方程模型估计方法的选择

联立方程模型在模型型式上有结构型和简化型之分，从模型的识别条件上又有恰好识别、过度识别和不可识别之分。由于模型的类型不同，建立模型的目的不同，模型的估计方法也有多种选择。

从模型的研究目的来看，如果研究目的是为了作经济结构分析，验证某种经济理论，着重关注的是模型的结构参数，应当力争尽可能准确估计结构型参数。如果研究目的是为了评价政策或论证某些经济政策的效应，就应当力争准确估计简化型参数，因为简化型参数正好能够反映“政策乘数”和“效果乘数”。如果研究目的只是为了作经济预测，要用预测期的外生变量值预测内生变量，只要直接估计简化型参数即可，因为简化型模型已表现了外生变量对各内生变量的影响。

从模型的识别条件来看，对于恰好识别模型，需要用间接最小二乘法、工具变量法等估计参数；对于过度识别模型，需要用二段最小二乘法、三段最小二乘法等估计参数；对于不足识别模型，则不能估计其结构型参数。对于递归型模型可直接用 OLS 法估计参数。

此外，还应考虑数据的可用性和计算方法的复杂性，对联立方程组模型的估计，通常有两类方法，一类是单一方程估计法或称为有限信息估计法；另一类是系统估计法或称为完全信息估计法。

单一方程估计法是指对方程组模型中的每一个方程逐一进行估计，最后得到模型中全部方程的估计。单一方程估计法有普通最小二乘法（OLS）、间接最小二乘估计法（ILS）、二段最小二乘估计法（TSLS）、有限信息最大似然法（LIML）等。单一方程估计法的特点是估计方程的参数只考虑该方程本身所带来的（有限）信息，不考虑整个模型所提供的全部信息，所以也称有限信息法。

系统估计法是指在考虑整个模型所提供的全部信息的情况下，对模型中的全部方程同时进行估计的方法。系统估计法有三段最小二乘法（3SLS）、完全信息最大似然估计法（FIML）等。因为该方法在估计模型时，用到了模型全部信息，所以也称完全信息法。

从对参数估计的统计特性看，系统估计法要优于单一方程估计法；但从方法的复杂性和可操作性看，单一方程法又优于系统估计法。所以在实际中单一方程估计法仍然得到广泛运用。系统估计法已超出本书范围，本章只介绍常用的单一方程估计法。

二、递归模型的估计——OLS 法

在第一节联立方程模型里已介绍了递归型模型，由于该模型构造的特殊性，递归模型中

各内生变量之间的联系只是单向的，都满足 OLS 基本假定，实际上并没有联立方程偏倚问题。

例如，第一节（11.24）式给出的递归型模型为

$$Y_1 = \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + u_1 \quad (11.67)$$

$$Y_2 = \alpha_{21}Y_1 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + u_2 \quad (11.68)$$

$$Y_3 = \alpha_{31}Y_1 + \alpha_{32}Y_2 + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3 + u_3 \quad (11.69)$$

其中 Y_1, Y_2, Y_3 为内生变量， X_1, X_2, X_3 为前定变量， u_1, u_2, u_3 为随机误差项。

递归型模型的第一个方程（11.67）式，由于在等式的右端只有前定变量和随机误差项，无内生变量，并且前定变量与随机误差项不相关，所以满足基本假定，可以直接用 OLS 法估计参数。第二个方程（11.68）式，其右端除了前定变量和随机误差项以外，还有内生变量 Y_1 ，但 Y_1 与随机误差项 u_2 并不相关，所以该方程满足基本假定，可用 OLS 法估计参数。同理，第三个方程（11.69）式也能用 OLS 法估计参数。

尽管递归型模型的解释变量中包含了内生变量，但根据递归的特点，它们与随机误差项不相关，不会产生联立方程模型的偏倚性，因此，如果联立方程模型为递归型模型，则可直接运用 OLS 法估计其参数。

三、恰好识别模型的估计 ——间接最小二乘法（ILS）

将结构型模型转化为简化型模型，由于简化型模型中的每一个方程的右端只有前定变量，并且前定变量与随机误差项不相关，可以用最小二乘法估计其参数。如果模型为恰好识别的模型，通过模型的简化型参数可以唯一确定结构型参数的估计值，显然，这种情况下可以先用 OLS 法估计简化型参数，然后再求解出结构型参数。这就是间接最小二乘法（ILS）的基本思想。

应用间接最小二乘法的步骤为：

- 1、将结构型模型转化为简化型模型，并建立简化型模型与结构型模型之间参数的关系式；
- 2、对简化型模型中的每一个方程用 OLS 法估计其参数，得到简化型方程的参数估计量；
- 3、在恰好识别的条件下，利用简化型模型与结构型模型之间参数的关系式唯一地解出结构型方程的参数估计量。

例如，（11.44）式商品需求与供给的结构型模型为

$$\begin{aligned}
Q_t^d &= \alpha_1 + \alpha_2 P_t + \alpha_3 I_t + u_{1t} \\
Q_t^s &= \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + u_{2t} \\
Q_t^d &= Q_t^s
\end{aligned}
\tag{见 11.44}$$

由该结构型模型导出的简化型模型为

$$\begin{aligned}
P_t &= \pi_{11} + \pi_{12} I_t + \pi_{13} P_{t-1} + u_{1t} \\
Q_t &= \pi_{21} + \pi_{22} I_t + \pi_{23} P_{t-1} + u_{2t}
\end{aligned}
\tag{见 11.45}$$

运用阶条件和秩条件对方程（11.44）进行判断，可知整个模型为恰好识别。运用 OLS 法估计简化型模型（11.45）式中的参数，求得各个参数的估计值 $\hat{\pi}_{ij}$ ($i=1,2; j=1,2,3$)。将估计的 $\hat{\pi}_{ij}$ 带入参数关系式，即可通过简化型模型的参数估计求解出方程（11.44）的参数估计：

$$\begin{aligned}
\hat{\alpha}_1 &= \hat{\pi}_{21} - \hat{\alpha}_2 \hat{\pi}_{11} & \hat{\alpha}_2 &= \frac{\hat{\pi}_{23}}{\hat{\pi}_{13}} & \hat{\alpha}_3 &= \hat{\alpha}_2 \hat{\pi}_{12} - \hat{\pi}_{22} \\
\hat{\beta}_1 &= \hat{\pi}_{21} - \hat{\beta}_2 \hat{\pi}_{11} & \hat{\beta}_2 &= \frac{\hat{\pi}_{22}}{\hat{\pi}_{12}} & \hat{\beta}_3 &= \hat{\pi}_{23} - \hat{\beta}_2 \hat{\pi}_{13}
\end{aligned}$$

可以证明，间接最小二乘法参数估计有以下特性：简化型参数的估计是无偏的，并且在大量样本下是一致估计式；但因结构型参数与简化型参数是非线性关系，结构型参数的估计在小样本中是有偏的，不过在大样本中是一致估计量。还可以证明，间接最小二乘法估计的结构型参数不是完全有效的，即一般不具有最小方差。这些特性的证明已超出本书范围，故本书从略。

最后还应强调，间接最小二乘法的运用有一定的假定前提，首先，结构型模型应是恰好识别；其次，在简化型模型中的每一个方程都应满足基本假定；而且，在简化型模型中的前定变量不存在严重的多重共线性。

四、过度识别模型的估计——二段最小二乘法（TSLS）

在计量经济分析中，许多结构型模型是过度识别的，这种情况下间接最小二乘法不适用。联立方程模型中出现的联立方程偏倚，是因为内生变量作为了解释变量，而且与随机误差项相关，故造成参数的估计有偏和非一致。如果能够找到一种变量，它与作为解释变量的内生变量高度相关，但与同期的随机误差项不相关，问题便可得到解决。例如由简化型模型估计的 \hat{Y}_t 就可能是这样的变量，用这种变量替代内生变量去作为解释变量，就可能避免联立方程偏倚的出现。这就是二段最小二乘法（简称 TSLS）的基本思想。

例如，由结构型模型变换得到的简化型模型中的第 i 个方程为

$$Y_i = \pi_{i1}X_{1i} + \pi_{i2}X_{2i} + \cdots + \pi_{ik}X_{ki} + v_i \quad (11.70)$$

其中的 $(\pi_{i1}X_{1i} + \pi_{i2}X_{2i} + \cdots + \pi_{ik}X_{ki})$ 构成了由前定变量 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 决定的 Y_i 的精确分量部分, 随机误差 v_i 构成 Y_i 的随机分量部分。在简化型模型中, 前定变量与随机误差项不相关, 所以可以对 (11.70) 式用 OLS 法估计参数, 这样便可得到上述精确分量的估计 \hat{Y}_i 。

作为精确分量 \hat{Y}_i 与 Y_i 高度相关, 但是 \hat{Y}_i 与 v_i 不相关。如果用 \hat{Y}_i 替换作为结构型模型解释变量的 Y_i , 显然根据 OLS 原理 \hat{Y}_i 与结构型模型的同期随机误差项也不相关, 从而避免了联立方程偏倚, 因此, 这时对经过变量替代的新结构型方程, 可以用 OLS 法估计参数。可以看出, 二段最小二乘法是分为两个阶段使用最小二乘法进行参数估计的方法, 实际是用 \hat{Y}_i 作为 Y_i 的工具变量。由于恰好识别是过度识别的特殊情况, 所以二段最小二乘法既可以用于过度识别条件下的参数估计, 也可用于恰好识别的情况。

二段最小二乘法的具体步骤如下:

(1) 将结构型模型变换为简化型模型，将结构方程中内生变量直接对所有的前定变量回归。

[illegible]

(2)运用 OLS 法分别估计简化型方程的参数 $\hat{\pi}_{ij}$, 利用所估计的 $\hat{\pi}_{ij}$ 和前定变量 \mathbf{X} 求 \hat{Y}_i , 如

$$\hat{Y}_i = \hat{\pi}_{i1} X_1 + \hat{\pi}_{i2} X_2 + \cdots + \hat{\pi}_{ik} X_k \quad (11.72)$$

(3) 用估计的 \hat{Y}_i 去替代结构方程中作为解释变量的内生变量 Y_i , 得

[illegible]

再运用 OLS 法估计该结构方程的参数, 得到参数的 2SLS 估计值。

(4) 对结构型模型的每一个方程如此进行估计, 最终完成对整个模型的参数估计。

例如 (11.46) 式的结构型模型为

$$Q_t^d = \alpha_1 + \alpha_2 P_t + \alpha_3 I_t + \alpha_4 R_t + u_{1t} \quad (11.74)$$

$$Q_t^s = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + u_{2t} \quad (11.75)$$

$$Q_t^d = Q_t^s \quad (11.76)$$

前面已经验证了此结构型模型中的供给方程是过度识别。为了用 2SLS 法估计其参数，直接写出它的简化型模型为

$$P_t = \pi_{11} + \pi_{12} I_t + \pi_{13} R_t + \pi_{14} P_{t-1} + u_{1t} \quad (11.77)$$

$$Q_t = \pi_{21} + \pi_{22} I_t + \pi_{23} R_t + \pi_{24} P_{t-1} + u_{2t} \quad (11.78)$$

首先，估计简化型模型中关于价格 P_t 的方程，得到估计式

$$\hat{P}_t = \hat{\pi}_{21} + \hat{\pi}_{22} I_t + \hat{\pi}_{23} R_t + \hat{\pi}_{24} P_{t-1} \quad (11.79)$$

设残差为 $e_{1t} = P_t - \hat{P}_t$ ，则 $P_t = \hat{P}_t + e_{1t}$ ，其中 \hat{P}_t 与 e_{1t} 不相关。

其次，将 $P_t = \hat{P}_t + e_{1t}$ 代入结构型模型中的需求方程 (11.74)，得

$$\begin{aligned} Q_t^d &= \alpha_1 + \alpha_2 (\hat{P}_t + e_{1t}) + \alpha_3 I_t + \alpha_4 R_t + u_{1t} \\ &= \alpha_1 + \alpha_2 \hat{P}_t + \alpha_3 I_t + \alpha_4 R_t + u_t^* \end{aligned} \quad (11.80)$$

其中， $u_t^* = \alpha_2 e_{1t} + u_{1t}$ 。可以证明，这时 \hat{P}_t 与 u_t^* 渐进不相关，因此，对 (11.80) 式可直接用 OLS 法估计参数，而 (11.80) 式正是结构型模型中商品需求方程，即完成了对需求方程参数的估计。

类似地，也可对 (11.75) 式供给方程用 TSLS 法估计参数。这样便完成了对整个结构型模型的参数估计。

运用二段最小二乘法时要注意使用条件：

- (1) 结构方程必须可以识别。
- (2) 结构方程中的随机误差项要满足 OLS 的基本假定。
- (3) 结构方程中的所有前定变量不存在严重的多重共线性，而且与随机误差项不相关。
- (4) 样本容量要足够大。

(5) 运用二段最小二乘法时应关注简化型模型的可决系数 R^2 ，第一段回归时 R^2 高，说明 \hat{Y}_t 与 Y_t 很接近，若第一段简化型方程回归中 R^2 很低，说明 \hat{Y}_t 对 Y_t 的代表性不强， Y_t 很大程度上受随机分量决定，TSLS 估计事实上将无意义。

可以证明（同样本书省略了这些证明），两阶段最小二乘法参数估计有以下特性：

- (1) 小样本时，TSLS 法所得到的参数估计量是有偏的；

(2) 大样本时，TSLS 法所得到的参数估计量具有一致性；

(3) 尽管 TSLS 法是针对过度识别而提出的，但对于恰好识别情况仍然可以使用，并且估计的结果与间接最小二乘法 (ILS) 估计结果一致。但在过度识别条件下，用 TSLS 法只能提供每个参数的惟一估计值，而用 ILS 法则能提供多个估计值。

二段最小二乘法较为简便，易于操作，当模型中结构方程较多时尤其方便。而且二段最小二乘法具有一致性特征，对可以识别的模型都适用，只要样本足够大，是估计联立方程模型的常用方法。

第四节 案例分析

一、研究目的和模型设定

依据凯恩斯宏观经济调控原理，建立简化的中国宏观经济调控模型。经理论分析，采用基于三部门的凯恩斯总需求决定模型，在不考虑进出口的条件下，通过消费者、企业、政府的经济活动，分析总收入的变动对消费和投资的影响。设理论模型如下：

$$Y_t = C_t + I_t + G_t \quad (11.81)$$

$$C_t = \alpha_0 + \alpha_1 Y_t + u_{1t} \quad (11.82)$$

$$I_t = \beta_0 + \beta_1 Y_t + u_{2t} \quad (11.83)$$

其中， Y_t 为支出法 GDP， C_t 为消费， I_t 为投资， G_t 为政府支出；内生变量为 Y_t, C_t, I_t ；

前定变量为 G_t ，即 $M=3, K=1$ 。

二、模型的识别性

根据上述理论方程，其结构型的标准形式为

$$-C_t - I_t + Y_t - G_t = 0$$

$$-\alpha_0 + C_t - \alpha_1 Y_t = u_{1t}$$

$$-\beta_0 + I_t - \beta_1 Y_t = u_{2t}$$

标准形式的系数矩阵 (B, Γ) 为

$$(B, \Gamma) = \begin{pmatrix} 0 & -1 & -1 & 1 & -1 \\ -\alpha_0 & 1 & 0 & -\alpha_1 & 0 \\ -\beta_0 & 0 & 1 & -\beta_1 & 0 \end{pmatrix}$$

由于第一个方程为恒等式，所以不需要对其识别性进行判断。下面判断消费函数和投资函数的识别性。

1、消费函数的识别性

首先，用阶条件判断。这时 $m_2 = 2, k_2 = 0$ ，因为 $K - k_2 = 1 - 0 = 1$ ，并且 $m_2 - 1 = 2 - 1 = 1$ ，所以 $K - k_2 = m_2 - 1$ ，表明消费函数有可能为恰好识别。

其次，用秩条件判断。在 (B, Γ) 中划去消费函数所在的第二行和非零系数所在的第一、二、四列，得

$$(B_0, \Gamma_0) = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$$

显然， $Rank(B_0, \Gamma_0) = 2$ ，则由秩条件，表明消费函数是可识别。再根据阶条件，消费函数是恰好识别。

2、投资函数的识别性

由于投资函数与消费函数的结构相近，判断过程与消费函数完全一样，故投资函数的阶条件和秩条件的判断予以省略。结论是投资函数也为恰好识别。

综合上述各方程的判断结果，得出该模型为恰好识别。

三、宏观经济模型的估计

由于消费函数和投资函数均为恰好识别，因此，可用间接最小二乘估计法（ILS）估计参数。选取 GDP、消费、投资，并用财政支出作为政府支出的替代变量。这些变量取自 1978 年——2003 年中国宏观经济的历史数据，见表 11.1。

表 11.1

年份	支出法 GDP	消费	投资	政府支出
1978	3605.6	2239.1	1377.9	480.0
1979	4074.0	2619.4	1474.2	614.0
1980	4551.3	2976.1	1590.0	659.0
1981	4901.4	3309.1	1581.0	705.0
1982	5489.2	3637.9	1760.2	770.0
1983	6076.3	4020.5	2005.0	838.0
1984	7164.4	4694.5	2468.6	1020.0
1985	8792.1	5773.0	3386.0	1184.0
1986	10132.8	6542.0	3846.0	1367.0
1987	11784.7	7451.2	4322.0	1490.0
1988	14704.0	9360.1	5495.0	1727.0
1989	16466.0	10556.5	6095.0	2033.0

1990	18319.5	11365.2	6444.0	2252.0
1991	21280.4	13145.9	7517.0	2830.0
1992	25863.7	15952.1	9636.0	3492.3
1993	34500.7	20182.1	14998.0	4499.7
1994	46690.7	26796.0	19260.6	5986.2
1995	58510.5	33635.0	23877.0	6690.5
1996	68330.4	40003.9	26867.2	7851.6
1997	74894.2	43579.4	28457.6	8724.8
1998	79003.3	46405.9	29545.9	9484.8
1999	82673.1	49722.7	30701.6	10388.3
2000	89340.9	54600.9	32499.8	11705.3
2001	98592.9	58927.4	37460.8	13029.3
2002	107897.6	62798.5	42304.9	13916.9
2003	121511.4	67442.5	51382.7	14764.0

资料来源：《中国统计年鉴 2004》，中国统计出版社。

1、恰好识别模型的 ILS 估计。

根据 ILS 法，首先将结构型模型转变为简化型模型，则宏观经济模型的简化型为

$$Y = \pi_{00} + \pi_{01}G$$

$$C = \pi_{10} + \pi_{11}G$$

$$I = \pi_{20} + \pi_{21}G$$

其中结构型模型的系数与简化型模型系数的关系为

$$\pi_{00} = \frac{\alpha_0 + \beta_0}{1 - \alpha_1 - \beta_1}, \quad \pi_{01} = \frac{1}{1 - \alpha_0 - \beta_0}, \quad \pi_{10} = \alpha_0 + \alpha_1 \frac{\alpha_0 + \beta_0}{1 - \alpha_1 - \beta_1},$$

$$\pi_{11} = \frac{\alpha_1}{1 - \alpha_1 - \beta_1}, \quad \pi_{20} = \beta_0 + \beta_1 \frac{\alpha_0 + \beta_0}{1 - \alpha_1 - \beta_1}, \quad \pi_{21} = \frac{\beta_1}{1 - \alpha_1 - \beta_1}$$

其次，用 OLS 法估计简化型模型的参数。进入 EViews 软件，确定时间范围；编辑输入数据；选择估计方程菜单。则估计简化型样本回归函数的过程是：按路径：Quick/Estimate Equation/ Equation Spesfication，进入“Equation Spesfication”对话框。

在“Equation Spesfication”对话框里，分别键入：“GDP C GOV”、“COM C GOV”、“INV C GOV”，其中，GDP 表示 Y，COM 表示 C，INV 表示 I，GOV 表示 G。得到三个简化型方程的估计结果，写出简化型模型的估计式：

$$\hat{Y} = -205.4438 + 8.0192G$$

$$\hat{C} = 481.985 + 4.6319G$$

$$\hat{I} = -370.3287 + 3.1593G$$

即简化型系数的估计值分别为

$$\begin{aligned}\hat{\pi}_{00} &= -205.4438, & \hat{\pi}_{01} &= 8.0192, & \hat{\pi}_{10} &= 481.985, \\ \hat{\pi}_{11} &= 4.6319, & \hat{\pi}_{20} &= -370.3287, & \hat{\pi}_{21} &= 3.1593\end{aligned}$$

最后，因为模型是恰好识别，则由结构型模型系数与简化型模型系数之间的关系，可惟一地解出结构型模型系数的估计。解得的结构型模型的参数估计值为

$$\begin{aligned}\hat{\alpha}_0 &= 600.6493, & \hat{\alpha}_1 &= 0.5776 \\ \hat{\beta}_0 &= -289.3838, & \hat{\beta}_1 &= 0.3940\end{aligned}$$

从而结构型模型的估计式为

$$\begin{aligned}Y &= C + I + G \\ C &= 600.6493 + 0.5776Y + u_1 \\ I &= -289.3838 + 0.3940Y + u_2\end{aligned}$$

2、过度识别模型的 2SLS 估计。

考虑在宏观经济活动中，当期消费行为还要受到上一期消费的影响，当期的投资行为也要受到上一期投资的影响，因此，在上述宏观经济模型里再引入 C_t 和 I_t 的滞后一期变量 C_{t-1} 和 I_{t-1} 。这时宏观经济模型可写为

$$\begin{aligned}Y_t &= C_t + I_t + G_t \\ C_t &= \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + u_{1t} \\ I_t &= \beta_0 + \beta_1 Y_t + \beta_2 I_{t-1} + u_{2t}\end{aligned}$$

用阶条件和秩条件对上述模型进行识别判断(具体的判断过程从略)，结论是消费函数和投资函数均是过度识别。需要运用二段最小二乘法对方程组的参数进行估计。

首先，估计消费函数。进入 EViews 软件，确定时间范围；编辑输入数据。然后按路径：Qucik/Estimate equation/Equation specification/Method/TSLS，进入估计方程对话框，将 method 按钮点开，这时会出现估计方法选择的下拉菜单，从中选“TSLS”，即两阶段最小二乘法。

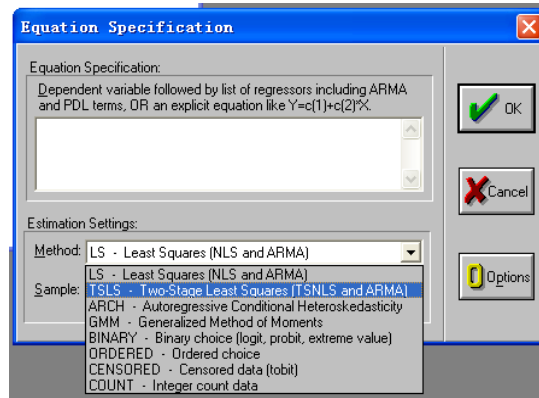


图 11.2

当 TSL 法选定后，便会出现“Equation Specification”对话框，见图 11.3。

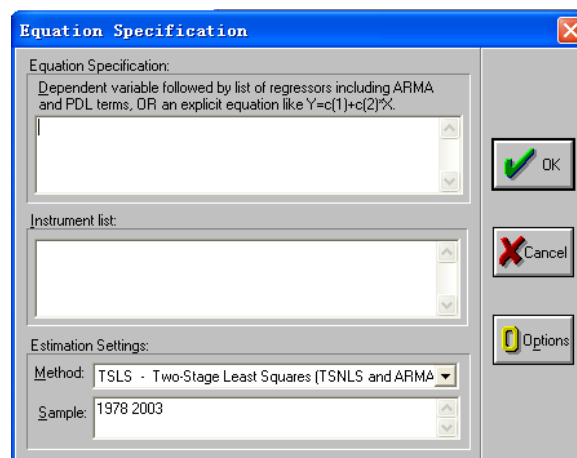


图 11.3

“Equation Specification”对话框有两个窗口，第一个窗口是用于写要估计的方程；第二个窗口是用于写该方程组中所有的前定变量，EViews 要求将截距项也看成前定变量。具体书写格式如下：第一个窗口写：“COM C GDP COM(-1)”；第二个窗口写：“C GOV COM(-1) INV(-1)”。其中，COM(-1)，INV(-1)分别表示消费变量 COM 和投资变量 INV 的滞后一期。然后按“OK”，便显示出估计结果，见表 11.5。

表 11.5

Dependent Variable: COM				
Method: Two-Stage Least Squares				
Date: 06/01/05 Time: 20:11				
Sample(adjusted): 1979 2003				
Included observations: 25 after adjusting endpoints				
Instrument list: C GOV COM(-1) INV(-1)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	760.1016	241.0503	3.153291	0.0046
GDP	0.393229	0.051167	7.685133	0.0000
COM(-1)	0.342025	0.095291	3.589265	0.0016
R-squared	0.998760	Mean dependent var	24219.91	
Adjusted R-squared	0.998647	S.D. dependent var	22011.34	
S.E. of regression	809.5354	Sum squared resid	14417648	
F-statistic	8856.594	Durbin-Watson stat	0.767087	
Prob(F-statistic)	0.000000			

根据表 11.5 写出消费函数的 2SLS 估计式为

$$C_t = 760.1016 + 0.3932Y_t + 0.3420C_{t-1} + u_{1t}$$

其次，估计投资函数。与估计消费函数过程一样，得到如下估计结果，见表 11.6。

表 11.6

Dependent Variable: INV				
Method: Two-Stage Least Squares				
Date: 06/01/05 Time: 20:21				
Sample(adjusted): 1979 2003				
Included observations: 25 after adjusting endpoints				
Instrument list: C GOV COM(-1) INV(-1)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-542.5631	397.8729	-1.363659	0.1865
GDP	0.524589	0.122685	4.275913	0.0003
INV(-1)	-0.369164	0.348573	-1.059074	0.3011
R-squared	0.994537	Mean dependent var	15799.04	
Adjusted R-squared	0.994040	S.D. dependent var	15119.00	
S.E. of regression	1167.206	Sum squared resid	29972123	
F-statistic	1999.299	Durbin-Watson stat	0.786536	
Prob(F-statistic)	0.000000			

由表 11.6 写出投资函数的估计式

$$I_t = -542.5631 + 0.5246Y_t - 0.3692I_{t-1} + u_{2t}$$

最后，写出该方程组模型的估计式为

$$\begin{aligned} Y_t &= C_t + I_t + G_t \\ C_t &= 760.1016 + 0.3932Y_t + 0.3420C_{t-1} + u_{1t} \\ I_t &= -542.5631 + 0.5246Y_t - 0.3692I_{t-1} + u_{2t} \end{aligned}$$

第十一章小结

1、联立方程模型是指用若干个相互关联的单一方程，同时表示一个经济系统中经济变

量相互联立依存性的模型，即用一个联立方程组去表现多个变量间互为因果的联立关系。联立方程组中每一个单一方程中包含一个或多个相互关联的内生变量，每一个方程的被解释变量都是内生变量，解释变量则可以是内生变量，也可以是外生变量。通常内生变量的个数应与模型中方程的个数一致。

2、联立方程模型中，从变量的性质看，一些变量是由模型体现的经济系统本身所决定的，称为内生变量，内生变量的取值是模型求解的结果，由于受模型中随机扰动项的影响，内生变量是随机变量。另一些变量是在模型体现的经济系统之外给定的，在模型中是非随机的，称为外生变量。外生变量数值的变化能够影响内生变量的变化，而内生变量却不能反过来影响外生变量。

3、联立方程模型中由于内生变量作为解释变量与随机误差项相关，用 OLS 法估计的参数有偏且不一致而引起的偏倚性，称为联立方程偏倚。

4、联立方程模型描述经济变量之间现实经济结构关系的模型，称为结构型模型。结构型模型表现变量间直接的经济联系，将某内生变量直接表示为内生变量和前定变量的函数。把每个内生变量都只表示为前定变量及随机扰动项函数的联立方程模型，称为简化型模型。简化型模型能直接用于对内生变量的预测。

5、联立方程模型的识别可以从多方面去理解，可从方程是否具有确定的统计形式去认识，也可以从方程中是否排除了必要的变量去理解。但对联立方程识别最直观的理解，是看能否从简化型模型参数估计值中合理求解出结构型模型参数的估计值。由简化型模型的参数求解结构型模型的参数时，能唯一求解，结构方程是恰好识别；能求解但解不惟一，结构方程过度识别；无法求解，则结构方程是不可识别。

6、判断模型可识别性的方法有模型识别的阶条件和秩条件。两种方法可结合运用。

7、联立方程模型的估计方法有多种。递归型联立方程模型 OLS 法估计。恰好识别的联立方程模型可用间接最小二乘法估计。过度识别和恰好识别的联立方程模型可用二段最小二乘法估计。不可识别的联立方程模型无法估计。

8、运用 EViews 软件实现对联立方程模型的估计和检验。

第十一章主要公式表

结构型模型的一般形式	$\beta_{11}Y_{1t} + \beta_{12}Y_{2t} + \cdots + \beta_{1M}Y_{Mt} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + \cdots + \gamma_{1k}X_{kt} = u_{1t}$ $\beta_{21}Y_{1t} + \beta_{22}Y_{2t} + \cdots + \beta_{2M}Y_{Mt} + \gamma_{21}X_{1t} + \gamma_{22}X_{2t} + \cdots + \gamma_{2k}X_{kt} = u_{2t}$ <p>.....</p> $\beta_{M1}Y_{1t} + \beta_{M2}Y_{2t} + \cdots + \beta_{MM}Y_{Mt} + \gamma_{M1}X_{1t} + \gamma_{M2}X_{2t} + \cdots + \gamma_{Mk}X_{kt} = u_{Mt}$
结构型模型的矩阵形式	$BY + \Gamma X = U$
简化型模型的矩阵形式	$Y = \Pi X + V$
模型识别的阶条件 (必要条件)	<p>当 $K - k_i = m_i - 1$ 时, 则第 i 方程是恰好识别;</p> <p>当 $K - k_i > m_i - 1$ 时, 则第 i 方程是过度识别;</p> <p>当 $K - k_i < m_i - 1$ 时, 则第 i 方程是不可识别。</p>
识别的秩条件 (充分必要条件)	<p>当且仅当一个方程所排斥 (不包含) 的变量的参数矩阵 (B_0, Γ_0) 的秩 $\text{Rank}(B_0, \Gamma_0) = M - 1$ 时, 方程可以识别, $\text{Rank}(B_0, \Gamma_0) \neq M - 1$, 方程不可识别;</p> <p>当只有一个 $M - 1$ 阶非零行列式时, 该方程是恰好识别;</p> <p>当不止一个 $M - 1$ 阶非零行列式时, 该方程是过度识别;</p> <p>当不存在 $M - 1$ 阶非零行列式时, 该方程是不可识别。</p> <p>若 $\text{Rank}(B_0, \Gamma_0) < M - 1$, 则该方程不可识别。</p>

思考题与练习题

思考题

- 11.1 除了单一方程模型以外, 为什么还要建立联立方程模型?
- 11.2 联立方程模型有那些种类? 各类联立方程模型的特点是什么?
- 11.3 什么是联立方程偏倚? 为什么会产生联立方程偏倚?
- 11.4 写出结构型模型的一般形式和结构参数矩阵。
- 11.5 为什么不能直接用普通最小二乘法对联立方程模型的参数进行估计?
- 11.6 识别的阶条件与秩条件的含义是什么? 为什么在识别的过程中要将阶条件与秩条件结合运用?
- 11.7 在哪种情况下, 可直接用最小二乘法估计联立方程模型的参数?
- 11.8 间接最小二乘法的条件、步骤、参数估计的特性是什么?
- 11.9 两阶段最小二乘法的条件、步骤、参数估计的特性是什么?

练习题

11.1 考虑以下凯恩斯收入决定模型：

$$\begin{aligned}C_t &= \beta_{10} + \beta_{11}Y_t + u_{1t} \\I_t &= \beta_{20} + \beta_{21}Y_t + \beta_{22}Y_{t-1} + u_{2t} \\Y_t &= C_t + I_t + G_t\end{aligned}$$

其中，C=消费支出，I=投资指出，Y=收入，G=政府支出； G_t 和 Y_{t-1} 是前定变量。

(1) 导出模型的简化型方程并判定上述方程中哪些是可识别的（恰好或过度）。

(2) 你将用什么方法估计过度可识别方程和恰好可识别方程中的参数。

11.2 考虑如下结果：

$$\text{OLS: } W_t = 0.276 + 0.258P_t + 0.046P_{t-1} + 4.959V_t \quad R^2=0.924$$

$$\text{OLS: } P_t = 2.693 + 0.232W_t - 0.544X_t + 0.247M_t + 0.064M_{t-1} \quad R^2=0.982$$

$$\text{2SLS: } W_t = 0.272 + 0.257P_t + 0.046P_{t-1} + 4.966V_t \quad R^2=0.920$$

$$\text{2SLS: } P_t = 2.686 + 0.233W_t - 0.544X_t + 0.246M_t + 0.064M_{t-1} \quad R^2=0.981$$

其中 W_t 、 P_t 、 M_t 和 X_t 分别是收益，价格，进口价格以及劳动生产力的百分率变化（所有百分率变化，均相对于上一年而言），而 V_t 代表未填补的职位空缺率（相对于职工总人数的百分率）。

试根据上述资料对“由于 OLS 和 2SLS 结果基本相同，故 2SLS 是无意义的。”这一说法加以评论。

11.3 考虑如下的货币供求模型：

$$\text{货币需求: } M_t^d = \beta_0 + \beta_1Y_t + \beta_2R_t + \beta_3P_t + u_{1t}$$

$$\text{货币供给: } M_t^s = \alpha_0 + \alpha_1Y_t + u_{2t}$$

其中，M=货币，Y=收入，R=利率，P=价格， u_{1t}, u_{2t} 为误差项；R和P是前定变量。

(1) 需求函数可识别吗？

(2) 供给函数可识别吗？

(3) 你会用什么方法去估计可识别的方程中的参数？为什么？

(4) 假设我们把供给函数加以修改，多加进两个解释变量 Y_{t-1} 和 M_{t-1} ，会出现什么识别问

题？你还会用你在（3）中用的方法吗？为什么？

11.4 考虑以下模型：

$$\begin{aligned} R_t &= \beta_0 + \beta_1 M_t + \beta_2 Y_t + u_{1t} \\ Y_t &= \alpha_0 + \alpha_1 R_t + u_{2t} \end{aligned}$$

其中 M_t （货币供给）是外生变量； R_t 为利率， Y_t 为 GDP，它们为内生变量。

（1）请说出此模型的合理性。

（2）这些方程可识别吗？

假使我们把上题的模型改变如下：

$$\begin{aligned} R_t &= \beta_0 + \beta_1 M_t + \beta_2 Y_t + \beta_3 Y_{t-1} + u_{1t} \\ Y_t &= \alpha_0 + \alpha_1 R_t + u_{2t} \end{aligned}$$

判断此方程组是否可识别，其中 Y_{t-1} 为滞后内生变量。

11.5 设我国的关于价格、消费、工资模型设定为

$$\begin{aligned} W_t &= \alpha_1 + \alpha_2 I_t + u_{1t} \\ C_t &= \beta_1 + \beta_2 I_t + \beta_3 W_t + u_{2t} \\ P_t &= \gamma_1 + \gamma_2 I_t + \gamma_3 W_t + \gamma_4 C_t + u_{3t} \end{aligned}$$

其中，I 为固定资产投资，W 为国有企业职工年平均工资，C 为居民消费水平指数，P 为价格指数，C、P 均以上一年为 100%，样本数据见表 11.6。试完成以下问题：

（1）该方程组是否可识别？

（2）选用适当的方法估计模型的未知参数？（要求：分别用 ILS 和 2SLS 两种方法估计参数）。

（3）比较所选方法估计的结果。

表 11.6

年份	固定资产投资 I（亿元）	职工年均工资 W（元）	消费水平指数 C（100%）	价格指数 P（100%）
1975	544.94	613	101.9	100.2
1976	523.94	605	101.8	100.3
1977	548.30	602	100.9	102.0
1978	668.72	644	105.1	100.7
1979	699.36	705	106.7	102.0
1980	745.90	803	109.5	106.0

1981	667.51	812	106.8	102.4
1982	945.31	831	105.4	101.9
1983	851.96	865	107.1	101.5
1984	1185.18	1034	11.4	102.8
1985	1680.51	1213	113.2	108.8
1986	1978.50	1414	104.9	106.0

11.6 表 11.6 给出了某国宏观经济统计资料，试判断模型的识别性，再用 2SLS 法估计如下宏观经济模型

$$\begin{aligned}
C_t &= \alpha_0 + \alpha_1 Y_t + u_{1t} \\
I_t &= \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_{2t} \\
Y_t &= C_t + I_t + G_t + X_t
\end{aligned}$$

其中， C_t, I_t, Y_t 分别表示消费，投资和收入； Y_{t-1}, G_t, X_t 分别表示收入的滞后一期，政府支出和净出口。

表 11.6

年份	C	I	Y	G	X
1978	1759	989	3036	869	-11
1979	1710	1026	3880	963	-19
1980	2129	1185	4083	881	-12
1981	2322	1169	4371	869	11
1982	2478	1279	4742	906	79
1983	2736	1432	5225	1013	44
1984	3070	1711	5985	1204	0
1985	3630	2356	6955	1259	-290
1986	3744	2453	7330	1319	-186
1987	4274	2742	8180	1424	-260
1988	4880	3237	9400	1380	-97
1989	5064	3403	9782	1425	-110
1990	5053	3355	10157	1467	282
1991	5376	3719	11091	1673	323
1992	6104	4550	12670	1881	135
1993	6536	6049	14379	2077	-283

1994	7300	6441	16200	2241	218
1995	8389	7008	17902	2204	301
1996	9335	7516	19620	2353	416
1997	10629	8006	21345	2684	-34

11.7 设联立方程组模型为

$$Q_t^d = \alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + u_{1t}$$

$$Q_t^s = \beta_0 + \beta_1 P_t + \beta_2 W_t + u_{2t}$$

$$Q_t^d = Q_t^s = Q_t$$

其中， Q_t^d, Q_t^s, P_t 分别为需求量，供给量和价格，它们为内生变量； Y_t, W_t 分别为收入和气候条件，它们为外生变量。试判断模型的识别性，并分别用 ILS 法和 2SLS 法求参数的估计，对所估计模型进行评价。样本数据见表 11.7。

表 11.7

时间 t	Q_t	P_t	Y_t	W_t
1	11	20	8.1	42
2	16	18	8.4	58
3	11	12	8.5	35
4	14	21	8.5	46
5	13	27	8.8	41
6	17	28	9.0	56
7	14	25	8.9	48
8	15	27	9.4	50
9	12	30	9.5	39
10	18	28	9.9	52

第十一章附录

附录 11.1 联立方程偏倚的证明

例如，设联立方程模型为

$$C_t = \beta_0 + \beta_1 Y_t + u_t \tag{1}$$

$$Y_t = C_t + I_t \tag{2}$$

对(1)式 β_1 的 OLS 估计为:

$$\hat{\beta}_1 = \frac{\sum c_t y_t}{\sum y_t^2} = \frac{\sum C_t y_t}{\sum y_t^2} = \frac{\sum (\beta_0 + \beta_1 Y_t + u_t) y_t}{\sum y_t^2} = \beta_1 + \frac{\sum u_t y_t}{\sum y_t^2} \quad (3)$$

其中利用了 $\sum y_t = 0$ 和 $\sum Y_t y_t / \sum y_t^2 = 1$ 。对上式两边取期望, 得

$$E(\hat{\beta}_1) = \beta_1 + E\left(\frac{\sum u_t y_t}{\sum y_t^2}\right)$$

这里的 $E(\sum u_t y_t / \sum y_t^2) \neq 0$, 则 $E(\hat{\beta}_1) \neq \beta_1$, $\hat{\beta}_1$ 是 β_1 的有偏估计。

对 (3) 式取概率极限, 得

$$p \lim(\hat{\beta}_1) = p \lim(\beta_1) + p \lim\left(\frac{\sum u_t y_t}{\sum y_t^2}\right) = p \lim(\beta_1) + \frac{p \lim(\frac{1}{n} \sum u_t y_t)}{p \lim(\sum y_t^2)} \quad (4)$$

其中: $(\sum u_t y_t)/n$ 是 Y 与 u 的样本协方差, 其总体协方差为

$$p \lim\left(\frac{1}{n} \sum u_t y_t\right) = Cov(Y_t, u_t) = \frac{\sigma^2}{1 - \beta_1}$$

$(\sum y_t^2)/n$ 是 Y 的样本方差, 其总体方差为

$$p \lim\left(\frac{1}{n} \sum y_t^2\right) = \sigma_Y^2$$

因此

$$p \lim(\hat{\beta}_1) = \beta_1 + \frac{1}{1 - \beta_1} \frac{\sigma^2}{\sigma_Y^2}$$

因为 $\frac{\sigma^2}{\sigma_Y^2} \neq 0$, 则 $p \lim(\hat{\beta}_1) \neq \beta_1$, 这说明 $\hat{\beta}_1$ 不是 β_1 的一致估计。

第十二章 实证项目的计量经济研究

——课程论文分析

前面各章分别从不同的角度讨论了计量经济学的基本理论和方法,可以看出这些理论方法有很强的针对性,都是从对实际经济问题的计量研究中提出来的。这些理论方法之所以具有生命力,也完全在于能够用于对实际经济问题的分析。运用计量经济学的基本理论和方法对实际经济管理问题作具体的计量研究,是学习计量经济学的根本目的,也是计量经济学重要的教学环节。计量经济学的应用领域十分广泛,研究的方法也多种多样,不可能在本书中去一一列举,但是如何运用计量经济方法去作实证项目研究,还是有某些规律可循的。

目前,一些学校要求学生在学习计量经济理论与方法的同时,以课程论文的形式对实证项目作一些具体的计量经济研究,并将其作为计量经济学教学的组成部分,这对于提高学生的素质和能力是非常有效的。对于计量经济学的初学者来说,以实证项目研究为内容的计量经济学课程论文,往往不知该从何处着手。本章将以完成一个学期的计量经济学课程论文为例,对实证项目的计量经济研究的构成要素、基本步骤等提出建议,包括对一般性原则和常用方法、有关选题、文献综述与评价、数据搜集、论文写作以及一些具体的计量经济建模分析技术等方面的内容展开讨论。在本章的附录中,给出一篇作为本科学生课程论文的实证项目计量研究的示例,供读者参考。应当强调的是,对实际经济问题计量研究的方式并不是唯一的,也不存在什么万能或统一的神奇模式,熟能生巧,实践才是学习实证项目计量经济研究的惟一方法。

第一节 实证项目研究的选题

一、问题的提出

计量经济实证研究首要的问题是选题,选题是确定“做什么和如何开始”的问题。当然,不同岗位或不同专业的读者关于选题可能有着不同的想法。选题应从实际需要出发,这取决于你的研究项目的要求,或者你所从事工作的需要,或者是上级对你的安排。计量经济实证研究要对所分析的经济问题得出数量上的结论,需要事先对所研究的目标和内在的经济联系有相当的认识,也就是说要有一定的理论准备和调查研究。作为计量经济学的初学者,可以结合已经学习过的经济管理课程,选择需要作实证分析的题目;或许你接触到了经济或管理中有值得从数量上加以实证估计和检验的问题;或者虽然别人已经作过理论上的研究,但缺

乏数量上的概念和界线，而你对这方面的数量结论感兴趣。这些都可能成为你选题的目标。

选题是一个不断探索、逐步深化认识的过程，一般而言，“做什么和如何开始”的问题可从两个层面去考虑：首先应确定自己感兴趣地研究领域，例如，“关于中国利率的研究”；然后是在所感兴趣的研究领域中选定感兴趣的具体题目，例如“关于提高住房贷款利率对北京房地产市场的影响分析”。这是两个不同层面的选题，前者只是在金融问题中的利率研究方面确定了一个总的领域，而后者则是具体化地明确一个真正的研究问题。

这里强调“感兴趣的领域”和“感兴趣的具体题目”，因为“兴趣是最好的老师”，在你真正对一个问题发生兴趣的时候，是你对它已经有了相当了解，有了从数量上深究愿望的时候。显然，研究领域要依据自身的专业，或者结合自己在经济学、管理学、社会学等方面的知识结构，去选择感兴趣的领域。一般界定自己感兴趣的领域，应当说不是一件太难的事情，而困惑的往往是如何从这些领域中具体地选择自己感兴趣的题目。表面上看

具体的研究题目从性质上通常分为两种类型，一类是关于理论验证方面的研究，另一类是关于实证应用分析方面的研究，或者是两者的结合。对理论的验证，主要是指对某些已有的观点、理论、命题等，采用定量分析的手段进行具体验证，看这些理论是否符合观测到的现实。这里强调的是定量分析的验证，而不是对理论本身的定性分析。例如，在经济学或金融学的相关课程中，曾学习过有关经济代理行为以及经济变量之间关系的理论、消费的收入决定理论、投资的决定理论；或者诸如“奥肯定律”以及 $MV = TP$ 的数量表达式等。选择题目时就要分析，在这些理论关系中是否存在需要用定量分析手段进行验证的问题？哪些理论是值得进行定量验证的，以及在定量验证过程中对哪些理论有进一步完善的可能。实证应用分析研究，主要指针对现实经济生活中已存在的一些看法和观点，运用计量经济分析方法来阐释自己的观点，或者去发现新的结论。例如，关于中国农村经济问题的讨论中，存在着不同的观点和不同的研究方法，可以运用某地区农村经济的相关数据进行计量经济研究，并将其结果与已有的观点和方法进行比较分析。

二、研究题目的选择

如上所述，课程论文的选题具有多样性和灵活性，尽管不存在万能或神奇的方法和公式，但如下的基本方面可在选题过程中供参考：

1. 要尽量选择在经济和社会领域中受到广泛关注的问题。所研究问题的题目要具体化，不宜空洞。题目应当体现出对所研究问题的了解程度，要明确究竟是要对理论作验证，还是要对现实经济活动作实证分析。这是进行计量经济学建模的前提。

2. 要明确研究的范围。研究的范围可以是宏观经济领域，例如国民经济的运行、经济

政策的传导评价等；研究的范围也可以是微观方面的，例如对某企业的管理、财务分析，或对一所大学的学生管理工作的研究。研究的范围也决定了收集数据的范围。

3. 所选题目的大小要适中。应当充分考虑研究的条件和现实可能性，包括理论把握的程度、数据获得的难易、计量分析方法的条件、完成项目研究的人力和时间的条件，等等。作为课程论文，特别是要考虑完成实证项目的时间约束。题目不能选得太大或过于综合，否则工作量太大，在半个学期的时间内是难以完成的。

4. 要充分考虑数据来源的可能性。没有变量数据来源的模型是不可能进行具体计量研究的。

经过上述各方面的工作后，对所研究的问题就会有大致地了解，对对计量经济实证项目的选题，就会有大致判断。这时需要整理自己的思路，对选题所研究的问题进行较为清晰地说明，从而确定具体的题目。

研究题目的选择是指确定研究的内容，为实证研究项目或课程论文具体定位。社会经济的计量选题不可能一一列举，作为举例，这里对某些实证研究的选题提出一些建议：

(1)宏观经济方面：例如可研究 GDP 的增长与固定资产投资增长之间的关系；研究税收对利率的影响；研究消费函数、投资函数、货币需求函数；研究财政金融政策的效应等。在宏观经济中选择题目作计量研究，其好处在于相关数据易于从各种年鉴中获得；其不足在于宏观经济的问题往往较为综合，影响因素众多，涉及诸多方面的知识，需要花费较多的时间和精力，作为本科课程论文要求在较短时间内完成有相当的难度，通常只能研究其中一个问题的某个方面。

(2)微观经济方面：例如估计公司的生产、财务成本、供给和需求函数；公司的原材料和产品市场的分析；投资决策分析；股票市场交易制度对股价的效应；商业银行绩效分析，等等。一般来说，微观方面的研究，题目比较具体化，针对性强，适合短时间内进行研究。其不足是微观经济的数据收集有相当的难度；

(3)城市和区域经济方面：例如估计中心城市、城镇、农村等对住房、交通和其他公共设施等的需求；产业结构、行业布局分析；不同区域的财政收支、教育发展、全要素生产率、能源价格、科技进步、人力资本的数量分析，等等；

(4)国际经济贸易方面：例如估计国家的进出口函数；研究汇率以及汇率决定因素之间的关系；研究国外直接投资（FDI）的效应；

(5)发展经济学方面：例如度量国家、省市自治区、不同经济发展区域（如东中西部）的入均 GNI、人均 GDP 的决定因素，投资、消费等对经济的拉动等；

(6)市场营销或产业组织方面：例如度量广告对销售额、利润或市场份额的影响；估计研究与开发(R&D)支出和人力资本生产力之间的关系；研究由于产权调整、兼并、合并与市场份额及利润率之间的关系。

(7)公共财政方面：例如估计中央或地方财政收入、财政支出与其特点；研究经济活动与财政政策变量之间的依存关系；研究农业税减免的效应；研究医疗卫生、道路、教育等与其决定因素之间的关系；

(8)人口、社会学方面：例如解释城市、农村在犯罪、贫困、离婚率、家庭人口、就业等方面的成因，研究人口出生率、居民生活质量及比较差异等。

三、文献资料的利用、综述与评价

在选题过程中除了自己作深入研究以外，选题时要充分借鉴他人的研究成果，包括图书、期刊等文献，也包括 Internet 网络资源。充分有效地利用各种文献和互联网提供的信息，可以避免重复作别人已经完成的工作，也可以从中发现自己可能的创新之处。

目前，可供利用的图书资料和文献很多，例如可利用 *Journal of Economic Literature(JEL)* 采用的分类系统和一些优秀期刊的信息。*Journal of Economic Literature* 是国内外大多数大学图书馆必定的季刊杂志，主要提供经过分类的上一季度出版的书籍和期刊文章清单，对每一篇论文确认一组编号并归类于经济学的某一子领域，按照题目进行编排，甚至包括文章的摘要。例如研究劳动力流动性问题，那么首先要在领域分类代码为 J 的“劳动和人口经济”下查找对应项，可以看到相关分类号码为 J6 的“流动性、失业和空位”，翻到“本期期刊文章主题索引”，该项内容为经过分类的最近发表文章的详细列表，依据此表就可以把自己感兴趣的内容进行摘录，帮助进行选题。另外，该期刊也列出了本期期刊的内容并给出了一些书籍和文章的摘要，仔细阅读这些摘要，有助于对题目有更多的了解。

国外的一些杂志，例如《应用经济学》(*Applied Economics*)、《应用计量经济学》(*Journal of Applied Econometrics*)、《牛津经济学和统计学评论》(*Oxford Bulletin of Economics and Statistics*)、《国际货币基金成员报告》(*International Monetary Fund Staff Papers*)等期刊，都具有极强的应用性导向，比较适合于作为我们选题的参考。如果在选题的初期只考虑了一个大概范围，那么这些专业杂志可能有助于缩小论文选题的范围。

大多数学生已非常熟悉利用 Internet 进行查询，通过键入关键词、主题等，就可以查询到与关键词或主题相关的内容。在选题过程中，主要注意搜索引擎的专业化和搜索查询的效率。一般使用较多的搜索引擎是 www.google.com。从专业化的角度看，网上经济文献分类体系是很好的关于经济学和其他社会科学分类的搜索引擎。目前常用的包括 www.econlit.org

和 www.isinet.com/products/citation/ssci/。前者简称为 *EconLit*，后者简称为 *SSCI*(Social Science Citation Index)。*EconLit* 在选题或寻找相关著作时，可以在线按照题目、作者或关键词的方式从大量的刊物文章中进行所需信息的搜索；*SSCI* 主要提供曾引用过某一研究内容的期刊、书籍等的清单(通常根据作者姓名查找)，在寻找与社会科学领域相关的论文时非常有用，其中包括那些引用率较高的论文。

对收集到的相关文献应注意进行整理。在相关文献的综述过程中，进一步明确别人的主要的观点和分歧，对那些与自己所选题目相似或密切相关的文献，应当特别关注建立计量经济学模型的基本思路，被解释变量和解释变量是如何确定的，采用的数据是哪些类型、数据来源以及测度方法、使用了哪些估计和假设检验方法等。

对相关文献应当有个总结性的文字材料，以利于梳理思路。拉姆·拉玛纳山（Ramu Ramanathan）建议至少写 4 篇总结性的书面文字材料，每篇文字材料应当有 3~5 页^①。文字材料通常由文献回顾性综述和文献评价两个部分组成。回顾性综述主要是交代所研究问题的理论与实证分析的发展沿革、回顾主要研究流派的观点、论点、命题以及支撑这些观点的理论与实证研究方法等。通常对文献的评价可从理论和方法论两个方面展开。从理论方面，主要是对理论的前提、理论命题或立论的准确性、论证推理的逻辑性等方面进行评价。从方法论方面，主要考证方法的假设条件、应用范围、应用对象以及实证衡量标准等。对文献的评价具有相当难度，需要综合运用所学知识和社会实践经验，对相关文献的现有研究成果给出自己的判定和评价，指出现有研究成果中存在的不足，发现其他尚未涉足的研究领域和内容。对相关文献的评价是一种基本的训练，也有利于发现实证研究项目可能的创新之处。

相关文献的回顾性综述是论文不可缺少的组成部分。从课程论文的写作看，文献综述要考虑研究目的、个人写作偏好、论文的长短等因素。有些人习惯于在专门的一章内，对与课程论文相关的文献进行综述，表明作者对所研究问题国内外发展现状的系统把握；也有人将文献综述作为某章（一般是引言或概论的章节）中的一部分，以保持整个课程论文在结构上的连贯性。不过关键不在于形式，而是要注意文献综述的内容与实质。

第二节 模型设定与数据处理

一、建模的基本思路

^① Ramu Ramanathan 著，薛菁睿译，应用计量经济学，机械工业出版社，北京，2003 年 9 月，第 393 页。

常用的建模思路，主要有结构模型方法和动态建模方法。

第一章中已介绍了一般的建模步骤，这是被称为“结构模型方法论”的传统计量经济学主导的建模思路。其基本要点是：从先验经济理论出发，在理论模型右边加上一个满足古典假设的误差项，然后采用某种统计方法，如普通最小二乘法，进行估计和检验。如果模型通过检验，则通过增加变量、删除变量、更换变量、改变函数形式等方式修改模型，重新进行估计和检验，直到模型通过各种检验为止。这种从少数方程和变量的简单模型入手，经过不断修改和补充，直至得到一个更为一般的模型，这种建模方法又被称为从“特殊到一般”的建模方法过程（simple-to-general approach）。从建模思路上看，这是一种以先验经济理论为建立模型的出发点、重点关注模型参数的估计、并将参数估计值与其理论预期值是否一致作为判断标准，以进行不同层次的检验和修正的思路，也称为“理论驱动型”建模思路，在建模中得到普遍应用。

动态建模方法是针对“特殊到一般”建模思路提出的一种建模方法论，将计量经济模型研究的重心从模型估计和检验，转向模型设定方法的探讨，从统计理论和经济理论两个方面，强调逻辑上的一致性。从统计理论角度，分析设定的计量经济模型是否与实证数据具有一致性；从经济理论角度，看设定的计量经济模型是否与经济理论保持了一致性。因此，模型设定过程中的一致性问题是计量经济学研究的中心环节。动态计量经济学的建模过程，是一个“从一般到特殊”的动态建模过程。首先，建立一个包含所有信息的最一般模型，以保证随机扰动项为独立同分布的正态变量；其次，对模型参数加以变化使其向经济理论靠近，并依据各种类型的检验结果，将模型简化为变量和参数都较少的节俭模型；然后，再对模型进行严格检验；最后，求出模型中内含的长期稳态解，用于检验经济理论、评价政策和预测未来等。这里需要强调的是，动态建模有着双重含义：一是对数据分布信息的动态挖掘，二是建模过程中有个不断反思与改进的过程。

二、模型设定的要求

模型的建立有理论问题，更为重要的是实践问题。关于建模的依据、变量的选择、模型形式的选择等在第一章和第七章已经讨论过，只有根据这些原则经过反复调整、反复检验才能得到较为理想的模型。

一般说来，判别模型优良程度总有一定的标准，可供参考的标准主要有以下几个方面：

（1）模型应当与数据所表现的现实相一致，这是一条建模的基本准则。

（2）模型应当与经济理论相一致，当存在若干相矛盾的理论时，一个模型至少应与一种理论相一致。

(3) 模型必须有外生变量构成其回归变量，并且模型中含有明确的因果关系。

(4) 模型中的参数应当具有相对稳定性，这是模型能用于预测和政策分析的必备条件，即估计出的模型参数必须可靠，并具有时不变性，即使将来新数据的协方差与原估算时用的样本数据的协方差不同了，参数的估计值也应不受影响。

(5) 模型必须具有对数据的代表性和优良的拟合性，即由模型算出的内生变量估计值与实际观测值之差，只是随机误差。所谓随机是指误差值无法由历史数据预测出来，否则就一定存在强于现有模型的设定形式。例如，随机误差出现序列相关，除了采用某些变通的估计方法处理序列相关问题以外，还应把序列相关视为模型设定有误的征兆，通常采用扩充滞后回归变量、重新设定模型的方法来解决。

(6) 模型应当具有尽可能大的包容性。当一个模型能够完全解释另一个模型的结论时就称前者包容后者，包容性是衡量模型优劣的一条重要标准。一个成功的模型，应当不仅仅能反映数据中所含的规律性，而且还应能解释其他运用同样数据的对立模型的长处与不足。包容性较强的计量经济模型一般能较好地揭示更普遍的经济规律。

(7) 模型的简洁性。从实践上考虑，模型越简洁其自由度也就越大；从认识论上考虑，模型越复杂人们全盘把握它的困难程度就越大，而且，复杂的设计常常能掩盖设计方案中的纰漏。简洁性准则迫使模型设计者采取科学的诚实态度。

能满足上述标准的模型，即可称为与理论和数据保持一致性的模型。总而言之，计量经济模型的设定过程，是一个综合考虑经济理论、样本数据、模型特征、使用要求等因素，依据前述标准进行科学性创作的过程。

三、模型变量与函数形式的设定

设定计量经济模型首先要确定模型中的变量。正如第一章已经讨论的，模型变量的选择，要根据模型的研究目的，要以经济理论为指导，通常不可能把所有的因素都列入模型，而只能抓住主要影响因素和主要特征，而不得不舍弃某些因素。因此根据研究的需要，对变量有取舍的问题，为避免出现对变量的设定误差，对模型中变量是否恰当需要加以检验。

回归模型的设定除了选择模型中的变量以外，另一重要方面是要使所设定的变量间函数形式能够体现变量间的基本关系。第二章已经说明了总体回归模型是对总体回归函数的描述，总体回归函数正是计量经济要去估计的目标。但其真实的函数形式事先并不知道，所谓模型函数形式的设定，是指根据对变量间相互关系的已有认识，把 Y 的条件期望设定为解释变量 X 的某种函数。总体条件期望函数 $E(Y|X_i)=f(X_i)$ ，可以设定为各种具体的函数

形式。在计量经济学的实践中，通常把总体回归函数的具体函数形式设定为初等函数，应当注意的是不同函数形式中参数的经济意义有较大差异。常用的函数形式如表 12.1 所示。

表 12.1 不同函数形式及参数的意义

设定	函数形式	边际效应 (dY/dX)	弹性系数 $\left[(dY/Y) / (dX/X) \right]$	参数 β_2 的意义
线性函数	$Y = \beta_1 + \beta_2 X$	β_2	$\beta_2 \frac{X}{Y}$	$\frac{dY}{dX}$
线性对数	$Y = \beta_1 + \beta_2 \ln X$	$\frac{\beta_2}{X}$	$\frac{\beta_2}{Y}$	$\frac{dY}{dX/X}$
倒数	$Y = \beta_1 + \beta_2 \left(\frac{1}{X} \right)$	$-\frac{\beta_2}{X^2}$	$-\frac{\beta_2}{XY}$	$-X^2 \frac{dY}{dX}$
多项式(二次函数)	$Y = \beta_1 + \beta_2 X + \beta_3 X^2$	$\beta_2 + 2\beta_3 X$	$(\beta_2 + \beta_3 X) \frac{X}{Y}$	$\frac{dY}{dX} - 2\beta_3 X$
交互作用	$Y = \beta_1 + \beta_2 X + \beta_3 XZ$	$\beta_2 + \beta_3 Z$	$(\beta_2 + \beta_3 Z) \frac{X}{Y}$	$\frac{dY}{dX} - \beta_3 Z$
对数线性	$\ln Y = \beta_1 + \beta_2 X$	$\beta_2 Y$	$\beta_2 X$	$\frac{dY/Y}{dX}$
对数倒数	$\ln Y = \beta_1 + \beta_2 \left(\frac{1}{X} \right)$	$-\frac{\beta_2 Y}{X^2}$	$-\frac{\beta_2}{X}$	$-X \frac{dY/Y}{dX/X}$
对数多项式(对数二次函数)	$\ln Y = \beta_1 + \beta_2 X + \beta_3 X^2$	$Y(\beta_2 + 2\beta_3 X)$	$X(\beta_2 + 2\beta_3 X)$	$\frac{dY/Y}{dX} - 2\beta_3 X$
双对数(对数对数)	$\ln Y = \beta_1 + \beta_2 \ln X$	$\beta_2 \frac{Y}{X}$	β_2	$\frac{dY/Y}{dX/X}$
对数曲线	$\ln \left(\frac{Y}{1-Y} \right) = \beta_1 + \beta_2 X$	$\beta_2 Y(1-Y)$	$\beta_2 X(1-Y)$	$\left(\frac{1}{1-Y} \right) \frac{dY/Y}{dX}$

表 12.1 中被解释变量与解释变量的关系许多都是非线性的，其中有的虽然变量间为非线性的，但对参数而言却是线性的，可直接按对于参数为线性的回归模型去估计与检验；有的通过初等函数变换就可得到对参数为线性的回归模型。例如：

(1) 双对数模型

如果设定的非线性模型为

$$Y_i = \beta_1 X^{\beta_2} e^{u_i} \quad (12.18)$$

可通过取自然对数得

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i \quad (12.19)$$

变换后的模型被解释变量和解释变量都是对数形式，斜率系数 β_2 衡量的是被解释变量 Y 关于解释变量 X 的弹性，即当 X 每变动百分之一时，Y 的均值变动的百分比。

(2)半对数模型

如果设定的非线性模型为

$$\ln Y_i = \alpha_1 + \alpha_2 X_i + u_i \quad (12.20)$$

或者

$$Y_i = \beta_1 + \beta_2 \ln X_i + v_i \quad (12.21)$$

这种模型也称不变百分率增长模型，其中斜率系数 α_2 衡量的是当变量 X 的绝对量每发生单位变动时，引起被解释变量 Y 平均值的相对变动比率。斜率系数 β_2 衡量的是当变量 X 变动百分之一时，Y 的均值变动的绝对量。

(3)倒数变换模型

如果设定的非线性模型为

$$Y_i = \beta_1 + \beta_2 (1/X_i) + u_i \quad (12.22)$$

这种模型表示随着 X 的递增 Y 将呈现非线性的递减，但最终以 β_1 为渐近线。

对于上述可变换为对参数线性的非线性模型，都可以方便地用线性回归的方法去估计和检验模型。除此以外，还有一些通过初等函数不能变换为对参数为线性的回归模型，这类模型参数的估计面临一些更为复杂的问题，需要探索专门的方法，例如在满足一定的条件下，可借助于泰勒级数展开来近似线性化，但这已经超出了本书的范围。

四、数据的收集与处理

1、数据来源

最基本的数据主要来自于各种统计年鉴、月报、季报等，如《中国统计年鉴》及各地区或各部门编制的年鉴、报告等。一些信息类的报刊也经常提供经济数据。

现在许多年鉴等数据报告已经通过网络对公众提供，这里仅列举出中国国内最常用的一

些网上数据来源：

1. 中国经济信息网（简称中经网，网址：<http://www.cei.gov.cn>）

2. 中国经济统计数据库（WEB 版）：<http://db.cei.gov.cn>）是一个综合、有序的庞大经济数据库群，内容涵盖宏观经济、产业经济、经济专题、区域经济、行业经济、以及世界经济等各个领域，是一个拥有 15 万指标量，面向社会各界用户提供权威、全面、及时的经济类数据信息的基础资料库。

3. 国家统计局统计数据网页：<http://www.stats.gov.cn/tjsj/>

4. 中国人民银行统计数据网页：<http://www.pbc.gov.cn/diaochatongji/tongjishuju/>

5. 中国证券监督管理委员会统计数据网页：<http://www.csrc.gov.cn/cn/statinfo/>

当所研究的问题无法从公众信息渠道获得时，则需要通过专门组织的调查去获取数据，这当然会面对很多特别的困难。

从各种渠道获取的数据，不能不加分析的拿来就用，应认真分析这些数据的内涵、数据包含的范围、数据的计算方法、数据所说明的问题、与其他数据的关系等。如果数据与模型中变量的要求不一致，则应当对数据作必要的加工或调整，或者应当重新寻求符合模型要求的数据。

2、数据类型

经济数据的类型有多种，不同的数据类型有其自身的性质，需要采用有针对性的估计方法。在实证项目计量研究中，常用的数据类型包括横截面数据；时间序列数据；混合横截面数据；面板数据；虚拟变量数据等。

（1）横截面数据

横截面数据在经济学和其他社会科学领域中得以广泛应用，特别是在对不同类型经济活动作比较研究时，更适于使用截面数据。在经济学中，横截面数据分析主要与应用经济领域密切相关，例如地方公共财政、产业组织理论、城市经济学、劳动经济学、人口与健康经济学等。在检验微观经济假设或评价微观经济政策时，有关个人、家庭、企业、城市等的数据都是至关重要的。

（2）时间序列数据

时间序列数据的重要特征是其与时间的相关性，很少假设经济数据的观测独立于时间。特别是在研究经济活动发展变化的规律性时，适于使用时间序列数据。应当注意，大多数的经济或其他社会科学的时间序列数据都存在着近期的相关性，例如，本期 GDP 与上期 GDP 之间在增长趋势上总是有着相关性，对上期 GDP 增长稳定性的一些了解，会提示本期 GDP

增长趋势的一定范围。考虑到时间序列数据的时间相关性，要设定时间序列数据模型事先有许多工作要做，包括平稳性检验、协整分析等，以便更好地解释和利用经济时间序列数据的相互依赖性。

（3）混合横截面数据与面板数据

混合横截面数据集（pooled cross section data set）是指既有横截面数据特点又有时间序列数据特征的数据集。面板数据集（panel data set）是不同指标在不同时间的表现形式，即由横截面数据集中每个数据的一个时间序列组成。

（4）虚拟变量数据

正如第八章所讨论过的，虚拟变量数据是由若干人工变量数据组成，包括两水平数据（0 和 1）、多水平数据（1，2，3，4）等。虚拟变量数据可为解释变量数据，也可作为被解释变量数据。正如第八章已讨论的，不同的虚拟变量数据有着不同的特征和不同的分析方法。

3、数据处理

在实际使用数据估计模型之前，需要对数据作预处理，对数据进行一些初步检查和分析，初步把握样本数据的一些统计特征，包括描述性统计、平稳性检验、协整分析和因果关系分析等。

1、数据的描述性统计

描述性统计主要分为图解、基本统计量和若干相关性的分析。

图解分析是指对数据的观测值绘制图形，从图中可以得到一些有价值的信息，例如识别数据非正常值；识别被解释变量和解释变量之间的依存关系等。若为时间序列图形，则可以了解到变量的时间路径和基本增长率；若为被解释变量和解释变量之间散布图，则可查看是否存在非线性关系，以初步选择条件期望方程的具体形式。应当指出的是，图解分析虽然直观，但也可能会产生误导，原因在于图解分析只是建立在样本数据的基础上，图中的形状并不能保证两个变量之间依存关系的真实性。例如，若图形显示两个变量为非线性关系，但这种非线性的关系可能并不真实，因为图形的非线性可能是由于第三个变量的变动而引起的。也就是说，没有在其他变量保持不变的条件下观测两个变量之间的图形。因此，图解法只是对变量进行的初步了解，而更重要的是根据相关理论去设定总体回归函数，并进行相应的模型设定检验。

数据的基本统计量，包括最小值、最大值、均值、标准差、峰度、偏态、变异系数、相关系数等。这些统计量在计量经济分析中常有其特定的作用，例如，变异系数描述了变量均值与标准差的比例，若将变异系数较小的变量作为解释变量，这些解释变量的变化不大，

可能会表现出非显著性的特征；又如，峰度、偏态等对分布函数的描述有着特殊的图示作用。再如，在灵敏度分析中通常会采用解释变量的一个标准差变化会引起被解释变量多少个标准差的变化。

相关系数矩阵通常被用以分析模型中相联系变量的相关程度。理想情形是被解释变量与某一解释变量之间的相关系数的数值较高，而两个解释变量之间的相关系数的数值较低。但也应注意，虽然解释变量之间相关系数较高会形成多重共线，而较小的相关系数并不意味着一定不产生多重共线性问题；若解释变量间的相关系数较高时，应事前予以注意。

2、时间序列数据的处理

如果经济变量采用的数据是时间序列数据，为了避免“伪回归”，应该按第十章的要求对时间序列变量的平稳性进行单位根检验。如果检验结果表明变量是平稳的，才可以用最小二乘法去估计模型。如果经检验表明时间序列变量为非平稳，则应进行协整分析，若它们之间存在协整关系，即两者的线性组合存在平稳关系，说明它们之间存在一个长期稳定的比例关系。需特别强调，对变量的平稳性检验和对变量之间的协整检验，是利用时间序列变量建立计量经济模型的先决条件。

在时间序列变量平稳性检验的基础上，还可以运用 Granger 因果检验等方法，对变量间的因果关系进行检验，进一步验证所建模型对变量间因果关系的设定是否合理。

第三节 计量经济分析

一、模型的估计

设定的模型确定以后，即可用收集的数据去估计模型中的参数。在本书讨论的范围内所用的估计方法主要是最小二乘法，事实上 OLS 不仅简便易用，而且在很多情形下都是既简便又适用的估计方法。模型中参数的估计与对模型的检验通常是个反复的过程，如果模型估计和检验的结果表明模型完全满足古典假定的要求，模型也通过各项检验，则参数的估计值就是计量的结果。如果经检验发现某些古典假定不能满足，则应按前几章所讨论的方法对模型加以适当调整，或采用其他估计方法，如加权最小二乘（WLS）、广义差分、工具变量等方法去估计模型中的参数。

二、模型的检验

如第一章已说明的，计量经济模型的检验主要包括经济意义的检验、统计推断检验、计量经济学检验、模型预测检验。此外还有模型的诊断性检验，主要包括对变量的检验、残差

检验和稳定性检验。对变量的检验包括参数约束、遗漏变量、包含多余变量的检验。残差检验包括正态性、ARCH、White 等检验。模型的诊断性检验中，除了第九章介绍的对设定误差的几种检验以外，有的检验已超出了本书的范围，这里只对与其他检验有交叉的内容，不加区分地进行讨论。

对计量经济模型的检验，首先要检验所估计的模型参数的数值和符号是否符合特定的经济意义。如果所估计的参数与经济理论或实际经验的结论不符合，应当分析模型设定是否有问题，分析是否违反了基本假定。在确认模型、数据、假定、估计方法均无问题的情况下，应当反思经济理论和经验是否不完备，或许你还会发现理论与经验有某些值得创新之处。

对模型的统计推断检验，主要是可决系数的分析、t 检验、F 检验，通过检验分析模型和各个变量是否显著。若模型或某些变量不显著，则应认真分析其原因，特别是要分析是否违反了某种基本假定条件。估计模型并分析 F 统计量、 \bar{R}^2 等可以捕获对被解释变量中变动百分比的解释信息，t 统计量可能表明所选变量显著性的信息，回归系数的符号可能提供估计值的经济背景合理与否的信息。

模型的计量经济学检验，主要是对多重共线性、异方差性、自相关，以及设定误差的检验。在对模型的检验中，除了例行的计量经济检验外，要特别注意解释变量与随机扰动项相关关系的检验，因为在选定了 OLS 作为估计方法后，解释变量与随机扰动项的相关关系将使参数的 OLS 估计具有不一致性。对解释变量与随机扰动项是否相关的检验，可从模型设定误差检验入手，因为如第七章所讨论的，各种模型设定误差的最终表现，均为随机扰动项与解释变量相关。此外，还可检验解释变量的测量误差和联立方程偏倚是否严重，因为解释变量的测量误差和联立方程偏倚也表现为解释变量与随机扰动项的相关性。若测量误差和联立方程偏倚问题不严重，则表明解释变量与随机扰动项的相关程度较弱，不会导致参数估计量较为严重的不一致性。

此外，模型的计量经济学检验中还应考虑对总体回归函数设定的检验，分析设定的条件期望方程的具体函数形式是否恰当，某些变量是否应该表述为对数形式？某些变量是否只取水平值还是需要取其平方值？定性因素的引入方式？虚拟变量的选择是否足够？交互效应的数量分析是否需要？等等。

在模型的计量经济学检验中，对于不同数据类型的模型，通常残差统计检验所关注的重点有所差异。一般说来，对横截面数据应当着重考证异方差性是否存在；对时间序列数据应当特别考证自相关性是否存在。但这并不意味着截面数据不会产生自相关，也不表明时间

序列数据不会生产异方差。就如第五章“引子”表明的，应当警惕在某些情况下时间序列数据的异方差性，甚至可能比截面数据更为严重。对时间序列数据的进一步应用需要格外小心。例如，关于方程的估计，所采用的样本数据是水平值还差分值？若为水平值，是否需要时间趋势变量？或者采用数据的差分形式是否更为合适？时间序列数据是否需要考虑季节性因素？从动态角度考虑，在分布滞后模型中滞后期应当如何选择？

显然，在模型检验中不存在什么“通行的模式”可以在模型检验过程中遵照执行，因为在模型检验的每个阶段都需要进行大量地判断，并且不同的人所使用的检验方法也不尽相同。鉴于此，某些一般性的建议和指导原则，可能对于完成实证分析是很有益的。首先，要避免在没有对模型进行更多分析之前就仓促地给出结论。建议先依据一些经济理论框架，或对一些基本经济行为的理解去分析模型；其次，运用本书各章介绍的方法对所建模型进行一连串的诊断性检验，以确保所得结论对模型设定的改变不是太敏感。例如，检验是否应当在模型中加入省略变量、检验是否非线性、检验是否存在滞后被解释变量等。最后，在若干类似的模型中，运用所选择统计量的数值来判断某个模型是否优于其它的模型。

三、模型的调整

模型的调整，是指对模型检验提出的问题如何予以解决，包括多重共线性、异方差性、自相关性以及模型设定误差、测量误差等。多重共线性、异方差性、自相关性等的弥补措施在本书相关章节中已进行了详尽地讨论，在实证分析中只是如何灵活运用的问题。例如，若是发现时间序列数据的模型中存在自相关，则或是重新建立模型，或是采用科克伦-奥克特等方法进行校正弥补；若是发现在横截面数据模型中存在异方差性，则可采用加权最小二乘法进行参数估计；若是发现解释变量与随机扰动项之间存在相关性，则应采用工具变量法或二段最小二乘法估计参数，以避免参数估计的不一致性。

值得注意的是，模型的检验与调整并不是截然分离的，这里只是为了论证方便，人为地将其分为了两个阶段去说明。模型的估计、检验和调整通常是一个统一体中的不同侧面，相辅相成；并且模型的估计、检验和调整要进行若干次的重复，即所谓的重新估计、重新检验和重新调整，至到模型满意为止。

四、模型计量结果的分析

经过检验证明所估计的模型是符合要求的，最后应对模型所提供的数量信息作具体的分析。根据建立模型的目的，可能是经济结构分析、经济预测、政策评价，其中经济预测和政策评价都要以所确立的经济结构为基础。所谓经济结构分析主要指模型中变量间的数量关系，这种数量关系是由所估计的参数体现的，所以应对所估计的参数数值与符号的经济意义

作具体研究和评价。对实证分析结果的解释力度作出说明，主要体现在对回归系数的符号、大小以及检验结果的解释方式上。除了一般性的解释外，例如回归系数的符号、大小和统计显著性等，还可从灵敏度分析、弹性分析等多个角度，对回归系数的经济意义进行说明，也可从不同估计方法差异性比较的角度进行解释。

五、研究结果的报告

实证项目的计量经济研究的结论得到以后，为了让别人了解研究的成果，应形成研究报告（或课程论文）。研究报告并没有固定的格式，这里对通常可以考虑包括的内容提供一些建议：

1、引言

引言包含对所研究问题或研究基本目标的陈述，并说明所研究项目的理论意义或应用价值所在。还可包括相关文献综述及评论，以及本项目研究中所得主要结论的简单描述。

2、理论分析与研究思路

主要对所要研究的问题作简要的理论描述，一般没有必要对经济理论进行完整陈述，只需说明理论上对所研究问题有什么结论，对所提出的有关概念、范畴给出明确的界定和解释。并且从理论上对计量分析的前提条件、基本思路和预计达到的目标作简要说明。

3、计量经济模型与估计方法

对自己所建的计量经济学模型进行系统全面地论述。主要包括两个方面的内容：一是关于模型的描述，对整个建模思路进行说明，特别是研究的主要对象（被解释变量）的确定、影响因素的分析及解释变量的选择、模型函数形式的假设，等等。二是关于推断方法的描述，主要是所选择的估计和假设检验的方法，指出所用方法与他人研究类似问题时所使用的方法有何差别等。

4、数据及处理

说明数据的来源及对数据所作的加工处理。如果采用了代用数据，应说明代用的理由和处理方式。另外，关于数据的初步挖掘分析也应说明，包括对数据进行的描述性统计分析、平稳性检验、协整分析等，这些关于数据的初步挖掘分析，有利于对所设定的模型形式进行改进和完善。

5、结果分析

结果分析包括按照第二章给出的规范格式报告回归分析计算的结果，以及估计和假设检验结果经济意义的解释，还可包括关于模型解释力度的讨论等。对估计和假设检验结果的解释，应当对估计和检验的每个阶段进行详细的说明。建议重点放在陈述所得结果的多种特征，

并与建模前对所研究问题的若干假设和期望值的认识进行比较分析,若出现了意料之外的结果,则应给出相应的解释。模型的估计(重新估计)、检验(重新检验)和调整(重新调整)的所有步骤和结果都应当在此部分有所记录,包括中间的失败经历。

6、结论

主要是对实证项目研究的结论和观点等进行总结,或者根据计量分析结果提出政策建议。另外,在结论部分应当包含本项研究的局限性和进一步应当做的工作。对研究的局限性还应提出相应的建议和打算。

在本章附录中,给出了一篇本科学计量经济学课程论文的示例,这是一篇对经济理论作实际验证性质的研究报告。应当指出,这里只是作为实证项目研究报告形式的举例,对于本报告的内容和方法是否恰当,读者可以充分展开讨论。

第十二章附录

附录 12.1 实证项目研究(课程论文)示例

库兹尼茨倒“U”理论的实证分析^①

一、问题的提出

改革开放以来,中国经济高速增长,1989年—2002年的13年间,经济年均增长9.3%,比世界平均增长速度快6.3个百分点。2001年底,中国国内生产总值达到95933亿元,比1988年增长近两倍,经济总量已跃居世界第六位。中国的改革开放的过程,同时也是向贫穷挑战的过程。据统计,中国城乡居民家庭人均纯收入分别从1978年的133.6元和343.4元上升到2001年的2366.4元和6859.6元。中国农村没有解决温饱问题的贫困人口,从1978年的2.5亿人减少到2000年的3000万人,人民生活水平总体上实现了由温饱到小康的历史性跨越。但是与此同时城乡居民收入差距却重新拉大,由1995年2.71:1扩大为2000年的2.79:1,城乡居民生活消费支出的差距也由2.70:1扩大到2.99:1;农村居民的消费额在居民消费额中的比重相应地由56.3%降到47%。全社会不同居民组织间的收入差距也在不断扩大。80年代中期,中国的基尼系数为0.28,1995年上升到0.38,1998年又升至0.415,比世界平均水平高出1.9个百分点。现在,5%的最富裕县与5%的最贫困县的人均GDP相差16.4倍。统计显示,中国最贫困的20%家庭仅占社会全部收入的4.27%,而富有的20%的家庭则占有全部收入的50.24%。在建设“社会更加和谐,人民生活更加富足”的全面小康社会的过程中,如何改革收入分配制度,缩小贫富差距,再一次成为人们关注的焦点。

对收入的差距,经济学中有著名的库兹涅茨“倒U假设”,这种假设是否能得到实际经济发展的证实?中国的贫富差距的发展趋势是否符合“倒U假设”的阶段特征?“倒U假设”对中国改善贫富差距状况能有什么启示呢?这是本项目研究的主要目的。

二、理论综述

库兹涅茨的“倒U假设”认为:在经济发展过程中,收入分配差别的长期变动轨迹是“先恶化,后改进”,或用他的话说:“收入分配不平等的长期趋势可以假设为:在前工业文明向工业文明过渡的经济增长早期阶段迅速扩大,而后是短暂的稳定,然后在增长的后阶段逐渐缩小。”对于出现倒U现象的原因他解释为,一方面,增长是储蓄和积累的函数,但储蓄和积累集中于少数富有阶层。另一方面,增长是同工

^① 本文原作者为西南财经大学经济学院2000级学生焦少飞,本书作了删改。

业化和城市化相伴随的，由于城市内部收入分配比农村更不平等，因而城市化水平的提高意味着经济中更不平等部分的增加。这使收入分配状况首先恶化。但是这种恶化会由于法律 and 政策的干预、人口变动、新兴行业的不断涌现而改变。

对“倒 U 假设”其他经济学家也作了类似的研究。索洛等人通过对英、荷、德等国二战后的收入差别的分析，证实了倒 U 假设的后半段，即这些国家的收入差别在战后随着经济的发展确实改进了。1970 年代，魏斯考夫通过对拉美一些发展中国家的分析，证明了库氏假设的前半段，即随着这些国家经济的发展收入不平等的状况恶化了。

阿德尔曼和毛瑞斯对此作了横截面分析，即利用同一时期不同发展水平的国家的资料进行分析。其实是假设处于不同发展水平的国家相当于处于不同的发展阶段，从而把倒 U 现象由动态的历史现象转化为静态的国别现象。阿德尔曼和毛瑞斯在 1970 年代初收集了 43 个国家的数据，第一次为相对收入不平等的研究提供了大量经验性证明，其结果支持了倒 U 理论。他们测算的回归方程为：

$$I=7.23+0.0258Y-0.000014Y^2 \quad (R^2=0.12) \\ (2.9) \quad (2.7) \quad (-2.8)$$

其中，I 为不平等指标，指 20% 最高收入者的收入与 20% 最低收入者的收入之比，Y 为人均国民收入，括号中的数值为 t 统计量。二次项前的负号证实了倒 U 现象的存在。

有关库兹涅茨假设的理论也存在着争论。对倒 U 理论的批判包括对模型假设的否定，及一些发展经济学家提出的公平增长理论。

三、模型设定

研究贫富差距与经济发展阶段的关系，需要考虑以下几个方面：

(1) 对贫富差距常用收入的差距去衡量，用什么数据表现收入差距呢？国外经济学家用的“收入不平等指标”，指 20% 最高收入者的收入与 20% 最低收入者的收入之比，但此指标只是间歇性测算，缺乏时间序列数据，也缺乏完整的截面数据。而基尼系数既能更准确的表现收入差距的程度，又可获得时间序列数据和截面数据。所以决定选用的“基尼系数”作为被解释变量去衡量收入差距。

(2) 数据性质的选择。由于世界上多数国家是发展中国家，不可能提供完整经济发展阶段的时间序列数据。但不同的国家经济发展所处的阶段有明显差异，可参考他人研究成果的假设：假设处于不同发展水平的国家相当于一国处于不同的发展阶段，从而把倒“U”理论由动态的历史现象变为静态的国别现象。并假设大多数国家的收入水平都是经历从低到高的发展。所以本项目选择 30 个国家的截面数据去建立模型。考虑到取得时序序列数据受到制约，从横截面数据角度的分析是合理的。

(3) 影响因素的分析

根据库兹涅茨假设的理论，收入差距决定于经济发展阶段，因此以人均国内生产总值表示的经济发展水平，是必须要考虑的主要影响因素。除此以外，根据经济理论，还有众多因素会影响收入的差距。

首先，生育率对收入分配的不均应存在显著的影响，人口—贫困周期理论说明，人口的过快发展将阻碍经济的增长，基尼系数将随之提高。

其次，政府通过向高收入者征税，并转而加大对贫困者的转移支付和补贴，可以缩小贫富差距。

另外，1991 年世界发展报告指出：“教育是影响收入不公平的最重要的单一变量”，通过政府对教育（主要是基础教育）的投入，可以使贫困阶层的教育负担减轻，减少辍学率，提高其素质，这是贫困者改变生存状况的必要条件。

因此，准备将“人均国内生产总值”、“生育率”、“政府的转移支付和补贴”、“教育投入”等作为模型的解释变量。

(4) 模型形式的设计

由于本文很大程度上是对库氏假设进行实证分析，所以首先对被解释变量(Y)与人均国内生产总值(X)进行回归分析，并将方程形式设定为二次型。

$$Y=C+C_1X+C_2X^2$$

然后，把影响基尼系数的其他因素“人口出生率 B”、“政府转移支付和补贴 G”、“教育投入 E”等变量以

某种方式引入模型。

四、数据的收集

本文获取了 30 个国家的数据如表所示：

国家	Y	X	Y1	X1	X2	E	B	B1	G	G1980	G1996	G1
1	0.31	1300	-1.171	7.17	51.41	2.8	2.1	0.742	0.435	47	40	-0.36151
2	0.35	1341	-1.05	7.2	51.86	2.5	1.5	0.405	0.28	28	28	-0.55284
3	0.33	1715	-1.109	7.45	55.46	1.7	1.8	0.588	0.3	24	36	-0.52288
4	0.47	3219	-0.755	8.08	65.24	3.4	1.7	0.531	0.105	17	4	-0.97881
5	0.4068	3227	-0.899	8.08	65.28	1.7	2.6	0.956	0.125	7	18	-0.90309
6	0.48	4320	-0.734	8.37	70.07	6	2.6	0.956	0.215	19	24	-0.66756
7	0.32	4437	-1.139	8.4	70.52	3.1	2.2	0.788	0.25	14	36	-0.60206
8	0.45	5949	-0.799	8.69	75.53	3.6	2	0.693	0.5	50	50	-0.30103
9	0.56	6145	-0.58	8.72	76.1	4.7	2.3	0.833	0.415	32	51	-0.38195
10	0.3364	7960	-1.089	8.98	80.68	3.7	1.2	0.182	0.365	24	49	-0.43771
11	0.48	8500	-0.734	9.05	81.86		2.2	0.788	0.355	31	40	-0.44977
12	0.47	8848	-0.755	9.09	82.59	2.7	1.5	0.405	0.505	43	58	-0.29671
13	0.35	10321	-1.05	9.24	85.41	3.8	0.1	-2.3	0.415	45	38	-0.38195
14	0.35	11431	-1.05	9.34	87.31	3.5	0.5	-0.69	0.335	35	32	-0.47496
15	0.41	11680	-0.892	9.37	87.72	2.8	1.7	0.531	0.07	6	8	-1.1549
16	0.315	11973	-1.155	9.39	88.18	4.7	0.4	-0.92	0.57	48	66	-0.24413
17	0.3658	14468	-1.006	9.58	91.77	5.8	0.8	-0.22	0.465	55	38	-0.33255
18	0.42	16110	-0.868	9.69	93.84	2	1.2	0.182	0.11	11	11	-0.95861
19	0.3	16389	-1.204	9.7	94.17	5.6	0.2	-1.61	0.555	53	58	-0.25571
20	0.3274	16651	-1.117	9.72	94.48	1.5	0.1	-2.3	0.605	63	58	-0.21824
21	0.262	16978	-1.339	9.74	94.86	5.3	0.4	-0.92	0.665	66	67	-0.17718
22	0.36	17118	-1.022	9.75	95.02	5.5	1.5	0.405	0.63	65	61	-0.20066
23	0.32	17167	-1.139	9.75	95.08	5	0.5	-0.69	0.635	62	65	-0.19723
24	0.25	17555	-1.386	9.77	95.51	4.2	0.1	-2.3	0.565	55	58	-0.24795
25	0.27	17568	-1.309	9.77	95.53	5.6	0.2	-1.61	0.61	60	62	-0.21467
26	0.26	17799	-1.347	9.79	95.78	9	0.3	-1.2	0.74	71	77	-0.13077
27	0.266	18484	-1.324	9.82	96.52	6.1	0.1	-2.3	0.595	59	60	-0.22548
28	0.37	18485	-0.994	9.82	96.53	5.8	0.6	-0.51	0.54	54	54	-0.26761
29	0.32	20322	-1.139	9.92	98.4	6.9	1.2	0.182	0.635	65	62	-0.19723
30	0.377	22609	-0.976	10	100.5	6.7	0.9	-0.11	0.57	54	60	-0.24413

资料来源：<http://www.worldbank.org/research/growth/dddeisqu.htm>;

http://ceinet.lib.swufe.edu.cn/index/Transform.asp?cedb=9&ThreeBlockCode=030901&Template=dbsjnj029&blockcode=DBsjnj_zh ,

《1999 年世界发展指标》 中国财政经济出版社 2000,《1998—1999 年世界发展报告》 中国财政经济出版社 2000

表中：Y 为 1988 年 30 个国家的基尼系数。

X 为 1988 年各国的人均国内生产总值。

Y_1 , X_1 分别为 Y 和 X 的自然对数； X_2 为 X_1^2 。

E 为 1980 年各国公共教育支出占国民生产总值的百分比，考虑到教育对收入分配影响的滞后性，所以取 1980 年的数据。

B 为各国的人口出生率，由于资料所限其实际取值为 1980-1990 年各国的人口年均增长率。 B_1

为 B 的自然对数。

G 为政府支出中的补贴和其它经常性转移支付占政府总支出的比重，同样由于资料所限，G 实际取 1980 年（ G_{1980} ）和 1996 年（ G_{1996} ）两年的平均值代替 1988 年的数值， G_1 为 G 的自然对数。

五、模型的估计与调整

1、基尼系数对人均国内生产总值的回归

由于本文很大程度上是对库氏假设进行实证分析，所以首先对被解释变量(Y)与人均国内生产总值(X)进行回归分析，并将方程形式设定为二次型。

$$Y=C+C_1X+C_2X^2$$

Eviews 的最小二乘计算结果为：

Dependent Variable: Y

Method: Least Squares

Date: 12/14/02 Time: 10:02

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.388851	0.041131	9.454050	0.0000
X	4.62E-06	8.84E-06	0.522862	0.6053
X2	-4.58E-10	3.91E-10	-1.170653	0.2520
R-squared	0.246671	Mean dependent var		0.361880
Adjusted R-squared	0.190869	S.D. dependent var		0.077480
S.E. of regression	0.069695	Akaike info criterion		-2.394747
Sum squared resid	0.131148	Schwarz criterion		-2.254627
Log likelihood	38.92121	F-statistic		4.420452
Durbin-Watson stat	1.861097	Prob(F-statistic)		0.021842

从中可以看出，解释变量系数检验的 t 值不显著，而且解释变量的系数太小。考虑到人均国内生产总值与基尼系数数据间的量纲级差距太大，改变模型设定的形式，对方程两边取对数形式可得：

$$\begin{aligned} \log Y &= -13.595 + 3.0163 \log X - 0.1783 (\log X)^2 \\ &\quad (0.0476) \quad (0.8266) \quad (3.5572) \quad (1) \\ t &= (-3.822) \quad (3.649) \quad (-3.747) \\ R^2 &= 0.4067 \quad df = 2.099 \end{aligned}$$

Eviews 的最小二乘计算结果为：

Dependent Variable: LOG(Y)

Method: Least Squares

Date: 06/07/05 Time: 21:42

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-13.59484	3.557195	-3.821788	0.0007
LOG(X)	3.016296	0.826650	3.648821	0.0011
(LOG(X))^2	-0.178324	0.047592	-3.746932	0.0009

R-squared	0.406672	Mean dependent var	-1.037703
Adjusted R-squared	0.362722	S.D. dependent var	0.208154
S.E. of regression	0.166168	Akaike info criterion	-0.656992
Sum squared resid	0.745522	Schwarz criterion	-0.516872
Log likelihood	12.85487	F-statistic	9.253020
Durbin-Watson stat	2.099527	Prob(F-statistic)	0.000870

(1) 经济意义检验。从回归结果可以看出，因为二次项前的系数为负值，所以此模型已经证实了库氏倒 U 理论。即随着经济的增长、人均国民生产总值的提高，基尼系数（也即贫富差距）会先增大后减小。模型采用双对数形式仍然可以说明这一现象。

(2) 统计推断检验。从回归的结果看，可决系数 $R^2=0.4067$ ，考虑到所采用的是截面数据，认为模型的拟合程度可以接受；系数显著性检验：给定 $\alpha=0.05$ ，查 t 分布表，在自由度为 $n-3=27$ 时得临界值 2.052，由于各解释变量系数的 t 值均大于临界值，所以人均国内生产总值对基尼系数有显著的影响。

(3) 计量经济学检验。给定显著性水平 0.05，查 D-W 表，当 $n=30, k=2$ 时，得下限临界值 $dL=1.284$ ，上限临界值 $dU=1.567$ ，因为 DW 统计量为 2.099 小于 $4-dL=2.433$ ，根据判定区域知不存在自相关。

作异方差的 White 检验如下表所示： 检验知 $Obs \cdot R\text{-squared}=4.424$ ，表明不存在异方差性。

White Heteroskedasticity Test:

F-statistic	1.499307	Probability	0.238064
Obs*R-squared	4.424486	Probability	0.219126

Test Equation:

Dependent Variable: RESID^2

Method: Least Squares

Date: 06/07/05 Time: 21:44

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.715980	5.148010	-1.693077	0.1024
LOG(X)	2.712955	1.614843	1.680011	0.1049
(LOG(X))^2	-0.235642	0.141554	-1.664675	0.1080
((LOG(X))^2)^2	0.000519	0.000317	1.640259	0.1130

R-squared	0.147483	Mean dependent var	0.024851
Adjusted R-squared	0.049116	S.D. dependent var	0.026246
S.E. of regression	0.025593	Akaike info criterion	-4.369401
Sum squared resid	0.017031	Schwarz criterion	-4.182575
Log likelihood	69.54101	F-statistic	1.499307
Durbin-Watson stat	2.379520	Prob(F-statistic)	0.238064

从 White 检验知 $Obs \cdot R\text{-squared}=30 \times 0.147483=4.424$ ，表明不存在异方差性。

本文至此已经从实际数据的回归分析的角度证明了库氏倒 U 理论的正确性。但影响基尼系数的因素还

包括诸如人口出生率 B、政府支出中的补贴和其它经常性转移支付占总支出的比重 G、教育投入 E 等。还需要对这些因素作分析。

2、基尼系数对其他因素的回归

(1) 对基尼系数 Y 和教育投入 E 进行回归得：

$$Y=0.407-0.011E$$

Eviews 的最小二乘计算结果为

Dependent Variable: Y

Method: Least Squares

Date: 12/25/02 Time: 10:39

Sample: 1 30

Included observations: 29

Excluded observations: 1

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.407126	0.035998	11.30982	0.0000
E	-0.011378	0.007677	-1.482148	0.1499
R-squared	0.075240	Mean dependent var		0.357807
Adjusted R-squared	0.040990	S.D. dependent var		0.075512
S.E. of regression	0.073948	Akaike info criterion		-2.304429
Sum squared resid	0.147645	Schwarz criterion		-2.210133
Log likelihood	35.41422	F-statistic		2.196763
Durbin-Watson stat	1.696502	Prob(F-statistic)		0.149879

从回归结果看，教育因素对收入分配的影响公然为负数，并且不显著。采用其他模型设定形式，t 值和可决系数的值仍不理想，原因可能是教育的严重滞后性。所以从计量经济学的角度本文还无法证明 E 对 Y 的显著性影响，下面的分析中将不得不舍弃这一在经济意义上合理的因素。

(2) 对基尼系数 Y 和人口出生率 B 回归得：

$$Y=0.2876+0.0646B$$

Eviews 的最小二乘计算结果为

Dependent Variable: Y

Method: Least Squares

Date: 12/14/02 Time: 10:39

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.287558	0.017825	16.13231	0.0000
B	0.064628	0.012622	5.120263	0.0000
R-squared	0.483558	Mean dependent var		0.361880
Adjusted R-squared	0.465113	S.D. dependent var		0.077480
S.E. of regression	0.056666	Akaike info criterion		-2.838952
Sum squared resid	0.089908	Schwarz criterion		-2.745539
Log likelihood	44.58429	F-statistic		26.21709
Durbin-Watson stat	1.828663	Prob(F-statistic)		0.000020

t 检验表明，人口出生率 B 对基尼系数的影响是显著的。

(3) 加入变量 G

考虑到人口出生率与政府转移支付之间可能出现多重共线，所以采用如下方程形式，回归可得：

$$\log(Y) = -0.097\log(G/B) - 1.1039$$

Dependent Variable: Y1

Method: Least Squares

Date: 12/25/02 Time: 11:10

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.103952	0.030211	-36.54127	0.0000
R1	-0.097219	0.018512	-5.251648	0.0000
R-squared	0.496220	Mean dependent var		-1.037703
Adjusted R-squared	0.478228	S.D. dependent var		0.208154
S.E. of regression	0.150357	Akaike info criterion		-0.887266
Sum squared resid	0.633004	Schwarz criterion		-0.793852
Log likelihood	15.30898	F-statistic		27.57981
Durbin-Watson stat	1.845702	Prob(F-statistic)		0.000014

这里可将 G/B 作为一个弹性的概念来理解，该弹性系数越大，即对出生率的每百分之一的增加，政府转移支付的增加量越大，基尼系数会减小。从回归结果看，G/B 每增加 1%，基尼系数将下降 0.097%，而且解释变量对被解释变量存在显著的影响。虽然同上面的模型相比，修正可决系数有微小下降，但从本文模型研究目的来看，这一模型拟合较好。

将上述分析的结果，加入方程（1）并回归，得：

$$\log(Y) = -0.1148(\log x)^2 + 2.013\log(x) - 0.089\log(G/B) - 9.829$$

$$(0.0417) \quad (0.715) \quad (0.023) \quad (3.028)$$

$$t = (-2.754) \quad (2.816) \quad (-3.925) \quad (-3.246)$$

$$R^2 = 0.627 (\text{修正值为 } 0.58) \quad DW = 2.342$$

Dependent Variable: LOG(Y)

Method: Least Squares

Date: 06/07/05 Time: 01:29

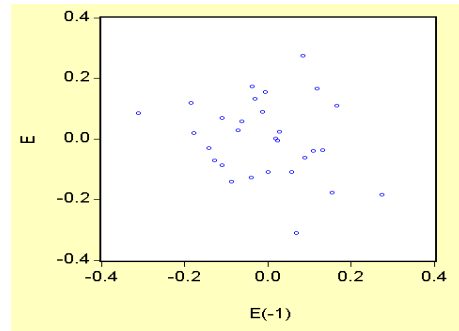
Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
(LOG(X))^2	-0.114823	0.041697	-2.753745	0.0106
LOG(X)	2.012955	0.714789	2.816154	0.0092
LOG(G/B)	-0.089069	0.022691	-3.925212	0.0006
C	-9.829011	3.028425	-3.245585	0.0032
R-squared	0.627444	Mean dependent var		-1.037703
Adjusted R-squared	0.584457	S.D. dependent var		0.208154
S.E. of regression	0.134181	Akaike info criterion		-1.055685
Sum squared resid	0.468119	Schwarz criterion		-0.868859
Log likelihood	19.83528	F-statistic		14.59607

Durbin-Watson stat 2.342498 Prob(F-statistic) 0.000009

检验：给定显著性水平 0.05，查 D-W 表,当 n=30,k =3 时，得 $d_L=1.214$ ， $d_U=1.65$ ， $4-d_U=4-1.65=2.35$ ，因为 DW 统计量为 2.342，可判断不存在自相关。由剩余项的图示也可以看出不存在自相关：



由 White 检验知 $\text{Obs} \cdot R\text{-squared}=5.909687$ ，明显小于自由度为 5， $\alpha=0.05$ 的 χ^2 值 11.0705，而且各项系数也不显著。Eviews 的计算结果为：

White Heteroskedasticity Test:

F-statistic	1.177506	Probability	0.349163
Obs*R-squared	5.909687	Probability	0.315106

Test Equation:

Dependent Variable: RESID^2

Method: Least Squares

Date: 06/07/05 Time: 21:50

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.293480	4.672055	-1.347048	0.1905
(LOG(X))^2	-0.160091	0.127797	-1.252701	0.2224
((LOG(X))^2)^2	0.000326	0.000284	1.150375	0.2613
LOG(X)	1.905291	1.462582	1.302690	0.2050
LOG(G/B)	-0.000183	0.004146	-0.044171	0.9651
(LOG(G/B))^2	-0.001515	0.001975	-0.766855	0.4506
R-squared	0.196990	Mean dependent var		0.015604
Adjusted R-squared	0.029696	S.D. dependent var		0.021930
S.E. of regression	0.021602	Akaike info criterion		-4.655222
Sum squared resid	0.011199	Schwarz criterion		-4.374982
Log likelihood	75.82833	F-statistic		1.177506
Durbin-Watson stat	2.592180	Prob(F-statistic)		0.349163

所以本文模型估计的最终结果为：

$$\log(Y) = -9.829 + 2.013 \log(x) - 0.1148 (\log x)^2 - 0.089 \log(G/B)$$

$$\begin{array}{cccc}
 (3.028) & (0.715) & (0.0417) & (0.023) \\
 t=(-3.246) & (2.816) & (-2.754) & (-3.925) \\
 R^2=0.627 & \bar{R}^2=0.58 & & DW=2.342
 \end{array}$$

六、本文的结论

(1) 人均国民生产总值对收入差距确实存在影响，原因是经济发展所产生的“扩散效应”。但这很大程度上取决于政府的政策取向。

(2) 生育率对收入分配的不均等存在显著的影响。从宏观层面上讲，人口—贫困周期理论说明，人口的高速增长造成了对食品供给的压力，而且强化了对储蓄、外汇储备及人力资源的发展的约束力。人口的过快增长，必将减少生产性资本的积累，有的国家还可能因进口粮食而耗费了本来可用于进口资本品的外汇。同时国家赡养率的提高使教育、医疗保健的供给不足，从而影响了人力资源的质量。这一切都将阻碍经济的增长，进而影响经济的发展。更重要的是上述问题所产生的负面效果将大部分由穷人承担，从而基尼系数将随之提高。从微观层面上讲，经济理论告诉我们，对于贫困阶层来说，儿童在某种程度上是一种经济投入品，其父母期待为其年老时提供经济支持的形式获得养育儿童的回报。再加上贫困阶层养育儿童的机会成本很低（原因在于诸如妇女的低就业率等）及较高的儿童死亡率使儿童的出生率大大高于富有阶层。这样穷人将可能越生越穷。

(3) 政府通过向高收入者征税，并转而加大对贫困者的转移支付和各项补贴可以在一定程度上缩小贫富差距。从图形上讲，可以降低“倒U”曲线的顶点。

所以随着经济的发展，收入分配不平等的程度将加深，并在经济达到一定规模后缩小。但在经济增长之初，如果贫富差距过大将影响经济的增长以及社会的稳定。本文分析表明政府完全可以通过一定的措施缓解这一状况。

本文未能从计量经济学的角度证明教育对基尼系数的影响，但“教育是影响收入不公平的最重要的单一变量”的结论应当是合理的，通过政府对教育（主要是基础教育）的投入，可以使贫困阶层的教育负担减轻，减少辍学率，提高其素质，这是贫困者改变生存状况的必要条件。本项研究在这方面的失败，原因可能是教育效果的严重滞后性。

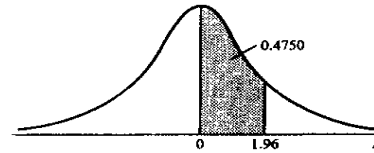
七、政策建议（略）

参考文献（略）

附 录

附表 1 标准化正态分布下的面积

例



$$p_r(0 \leq z \leq 1.96) = 0.4750$$

$$p_r(z \geq 1.96) = 0.5 - 0.4750 = 0.025$$

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4454	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981

2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

注：本表给出的分布的右侧（即 $z \geq 0$ ）面积。由于正态分布是围绕 $z=0$ 而对称分布的，故左侧面积与

相应的右侧面积相同。例如， $p(-1.96 \leq z \leq 0)=0.4750$ ，因此， $p(-1.96 \leq z \leq 1.96)=0.95$ 。

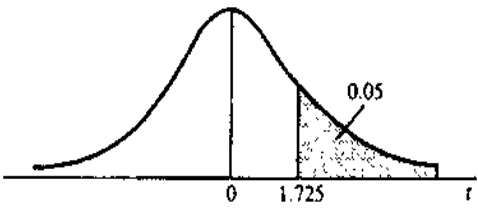
附表 2 t 分布的百分点

例

$p_r(t > 2.086) = 0.025$

$p_r(t > 1.725) = 0.05 \text{ for } df = 20$

$p_r(|t| > 1.725) = 0.10$



df \ pr	0.25	0.10	0.05	0.025	0.01	0.005	0.001
	0.50	0.20	0.10	0.05	0.02	0.010	0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505

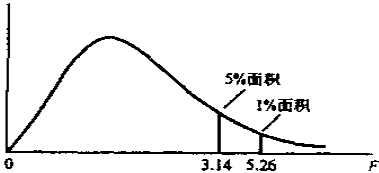
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

注：每列顶头的较小概率指单侧面积；较大的概率则指双侧面积。

附表 3 F 分布的上端百分点

例

$p_r(F>1.59)=0.25$
 $p_r(F>2.42)=0.10$ 对于自由度 $N_1=10$
 $p_r(F>3.14)=0.05$ 和 $N_2=9$
 $p_r(F>5.26)=0.01$



分母自由度 N_2	分子自由度 N_1												
	p_r	1	2	3	4	5	6	7	8	9	10	11	12
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	0.5	161	200	216	225	230	234	237	239	241	242	243	244
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.69	1.68
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.63	1.62
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07

| .01 10.6 8.02 6.99 6.42 6.06 5.80 5.61 5.47 5.35 5.26 5.18 5.11

续表

分子自由度 N_1													分母自由 度 N_2
15	20	24	30	40	50	60	100	120	200	500	∞	p_r	
9.49	9.58	9.63	9.67	9.71	9.74	9.76	9.78	9.80	9.82	9.84	9.85	.25	1
61.2	61.7	62.0	62.3	62.5	62.7	62.8	63.0	63.1	63.2	63.3	63.3	.10	
246	248	249	250	251	252	252	253	253	254	254	254	.05	
3.41	3.43	3.43	3.44	3.45	3.45	3.46	3.47	3.47	3.48	3.48	3.48	.25	2
9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.48	9.49	9.49	9.49	.10	
19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	.05	
99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	.01	3
2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	.25	
5.20	5.18	5.18	5.17	5.16	5.15	5.15	5.14	5.14	5.14	5.14	5.13	.10	
8.70	8.66	8.64	8.62	8.59	8.58	8.57	8.55	8.55	8.54	8.53	8.53	.05	4
26.9	26.7	26.6	26.5	26.4	26.4	26.3	26.2	26.2	26.2	26.1	26.1	.01	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	
3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.78	3.77	3.76	3.76	.10	5
5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.66	5.65	5.64	5.63	.05	
14.2	14.0	13.9	13.8	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5	.01	
1.89	1.88	1.88	1.88	1.88	1.88	1.87	1.87	1.87	1.87	1.87	1.87	.25	6
3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.13	3.12	3.12	3.11	3.10	.10	
4.62	4.56	4.53	4.50	4.46	4.44	4.43	4.41	4.40	4.39	4.37	4.36	.05	
9.72	9.55	9.47	9.38	9.29	9.24	9.20	9.13	9.11	9.08	9.04	9.02	.01	7
1.76	1.76	1.75	1.75	1.75	1.75	1.74	1.74	1.74	1.74	1.74	1.74	.25	
2.87	2.84	2.82	2.80	2.78	2.77	2.76	2.75	2.74	2.73	2.73	2.72	.10	
3.94	3.87	3.84	3.81	3.77	3.75	3.74	3.71	3.70	3.69	3.68	3.67	.05	8
7.56	7.40	7.31	7.23	7.14	7.09	7.06	6.99	6.97	6.93	6.90	6.88	.01	
1.68	1.67	1.67	1.66	1.66	1.66	1.65	1.65	1.65	1.65	1.65	1.65	.25	
2.63	2.59	2.58	2.56	2.54	2.52	2.51	2.50	2.49	2.48	2.48	2.47	.10	9
3.51	3.44	3.41	3.38	3.34	3.32	3.30	3.27	3.27	3.25	3.24	3.23	.05	
6.31	6.16	6.07	5.99	5.91	5.86	5.82	5.75	5.74	5.70	5.67	5.65	.01	
1.62	1.61	1.60	1.60	1.59	1.59	1.59	1.58	1.58	1.58	1.58	1.58	.25	8
2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.32	2.31	2.30	2.29	.10	
3.22	3.15	3.12	3.08	3.04	2.02	3.01	2.97	2.97	2.95	2.94	2.93	.05	
5.52	5.36	5.28	5.20	5.12	5.07	5.03	4.96	4.95	4.91	4.88	4.86	.01	9
1.57	1.56	1.56	1.55	1.55	1.54	1.54	1.53	1.53	1.53	1.53	1.53	.25	
2.34	2.30	2.28	2.25	2.23	2.22	2.21	2.19	2.18	2.17	2.17	2.16	.10	
3.01	2.94	2.90	2.86	2.83	2.80	2.79	2.76	2.75	2.73	2.72	2.71	.05	9
4.96	4.81	4.73	4.65	4.57	4.52	4.48	4.42	4.40	4.36	4.33	4.31	.01	

续表

分母自由度 N_2	分子自由度 N_1												
	p_r	1	2	3	4	5	6	7	8	9	10	11	12
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.55	1.54
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28
	0.5	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
	01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.52	1.51
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.50	1.49
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16
13	.25	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.08	2.05
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.44
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.44	1.43
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.41
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.96	1.93
	.05	4.41	3.55	3.26	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.40
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.94	1.91
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	.25	1.40	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.39
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.92	1.89
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23

续表

分子自由度 N_1													分母自由度 N_2
15	20	24	30	40	50	60	100	120	200	500	∞	p_r	
1.53	1.52	1.52	1.51	1.51	1.50	1.50	1.49	1.49	1.49	1.48	1.48	.25	10
2.24	2.20	2.18	2.16	2.13	2.12	2.11	2.09	2.08	2.07	2.06	2.06	.10	
2.85	2.77	2.74	2.70	2.66	2.64	2.62	2.59	2.58	2.56	2.55	2.54	.05	
4.56	4.41	4.33	4.25	4.17	4.12	4.08	4.01	4.00	3.96	3.93	3.91	.01	
1.50	1.49	1.49	1.48	1.47	1.47	1.47	1.46	1.46	1.46	1.45	1.45	.25	11
2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	2.00	1.99	1.98	1.97	.10	
2.72	2.65	2.61	2.57	2.53	2.51	2.49	2.46	2.45	2.43	2.42	2.40	.05	
4.25	4.10	4.02	3.94	3.86	3.81	3.78	3.71	3.69	3.66	3.62	3.60	.01	
1.48	1.47	1.46	1.45	1.45	1.44	1.44	1.43	1.43	1.43	1.42	1.42	.25	12
2.10	2.06	2.04	2.01	1.99	1.97	1.96	1.94	1.93	1.92	1.91	1.90	.10	
2.62	2.54	2.51	2.47	2.43	2.40	2.38	2.35	2.34	2.32	2.31	2.30	.05	
4.01	3.86	3.78	3.70	3.62	3.57	3.54	3.47	3.45	3.41	3.38	3.36	.01	
1.46	1.45	1.44	1.43	1.42	1.42	1.42	1.41	1.41	1.40	1.40	1.40	.25	13
2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.88	1.86	1.85	1.85	.10	
2.53	2.46	2.42	2.38	2.34	2.31	2.30	2.26	2.25	2.23	2.22	2.21	.05	
3.82	3.66	3.59	3.51	3.43	3.38	3.34	3.27	3.25	3.22	3.19	3.17	.01	
1.44	1.43	1.42	1.41	1.41	1.40	1.40	1.39	1.39	1.39	1.38	1.38	.25	14
2.01	1.96	1.94	1.91	1.89	1.87	1.86	1.83	1.83	1.82	1.80	1.80	.10	
2.46	2.39	2.35	2.31	2.27	2.24	2.22	2.19	2.18	2.16	2.14	2.13	.05	
3.66	3.51	3.43	3.35	3.27	3.22	3.18	3.11	3.09	3.06	3.03	3.00	.01	
1.43	1.41	1.41	1.40	1.39	1.39	1.38	1.38	1.37	1.37	1.36	1.36	.25	15
1.97	1.92	1.90	1.87	1.85	1.83	1.82	1.79	1.79	1.77	1.76	1.76	.10	
2.40	2.33	2.29	2.25	2.20	2.18	2.16	2.12	2.11	2.10	2.08	2.07	.05	
3.52	3.37	3.29	3.21	3.13	3.08	3.05	2.98	2.96	2.92	2.89	2.87	.01	
1.41	1.40	1.39	1.38	1.37	1.37	1.36	1.36	1.35	1.35	1.34	1.34	.25	16
1.94	1.89	1.87	1.84	1.81	1.79	1.78	1.76	1.75	1.74	1.73	1.72	.10	
2.35	2.28	2.24	2.19	2.15	2.12	2.11	2.07	2.06	2.04	2.02	2.01	.05	
3.41	3.26	3.18	3.10	3.02	2.97	2.93	2.86	2.84	2.81	2.78	2.75	.01	
1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.33	1.33	.25	17
1.91	1.86	1.84	1.81	1.78	1.76	1.75	1.73	1.72	1.71	1.69	1.69	.10	
2.31	2.23	2.19	2.15	2.10	2.08	2.06	2.02	2.01	1.99	1.97	1.96	.05	
3.31	3.16	3.08	3.00	2.92	2.87	2.83	2.76	2.75	2.71	2.68	2.65	.01	
1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.32	1.32	1.32	.25	18
1.89	1.84	1.81	1.78	1.75	1.74	1.72	1.70	1.69	1.68	1.67	1.66	.10	
2.27	2.19	2.15	2.11	2.06	2.04	2.02	1.98	1.97	1.95	1.93	1.92	.05	
3.23	3.08	3.00	2.92	2.84	2.78	2.75	2.68	2.66	2.62	2.59	2.57	.01	
1.38	1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.32	1.31	1.31	1.30	.25	19
1.86	1.81	1.79	1.76	1.73	1.71	1.70	1.67	1.67	1.65	1.64	1.63	.10	
2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.94	1.93	1.91	1.89	1.88	.05	
3.15	3.00	2.92	2.84	2.76	2.71	2.67	2.60	2.58	2.55	2.51	2.49	.01	
1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.31	1.31	1.30	1.30	1.29	.25	20
1.84	1.79	1.77	1.74	1.71	1.69	1.68	1.65	1.64	1.63	1.62	1.61	.10	
2.20	2.12	2.08	2.04	1.99	1.97	1.95	1.91	1.90	1.88	1.86	1.84	.05	
3.09	2.94	2.86	2.78	2.69	2.64	2.61	2.54	2.52	2.48	2.44	2.42	.01	

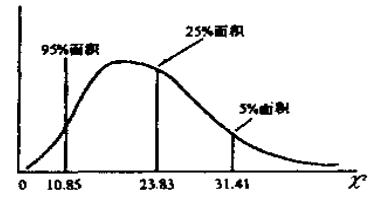
续表

分母自 由度 N_2	分子自由度 N_1												
	p_r	1	2	3	4	5	6	7	8	9	10	11	12
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.38	1.37
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86
	0.5	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
	01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.37	1.36
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.36	1.35
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.84	1.81
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.28	3.18	3.09	3.02	2.96
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.35	1.34
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.73	1.71
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.29
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.27	1.26
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.62	1.60
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34
200	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.31	1.29	1.28	1.27	1.26	1.25
	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60	1.57
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27
∞	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.24
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18

续表

分子自由度 N_1													分母自由 度 N_2
15	20	24	30	40	50	60	100	120	200	500	∞	p_r	
1.36	1.34	1.33	1.32	1.31	1.31	1.30	1.30	1.30	1.29	1.29	1.28	.25	22
1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.61	1.60	1.59	1.58	1.57	.10	
2.15	2.07	2.03	1.98	1.94	1.91	1.89	1.85	1.84	1.82	1.80	1.78	.05	
2.98	2.83	2.75	2.67	2.58	2.53	2.50	2.42	2.40	2.36	2.33	2.31	.01	
1.35	1.33	1.32	1.31	1.30	1.29	1.29	1.28	1.28	1.27	1.27	1.26	.25	24
1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.58	1.57	1.56	1.54	1.53	.10	
2.11	2.03	1.98	1.94	1.89	1.86	1.84	1.80	1.79	1.77	1.75	1.73	.05	
2.89	2.74	2.66	2.58	2.49	2.44	2.40	2.33	2.31	2.27	2.24	2.21	.01	
1.34	1.32	1.31	1.30	1.29	1.28	1.28	1.26	1.26	1.26	1.25	1.25	.25	26
1.76	1.71	1.68	1.65	1.61	1.59	1.58	1.55	1.54	1.53	1.51	1.50	.10	
2.07	1.99	1.95	1.90	1.85	1.82	1.80	1.76	1.75	1.73	1.71	1.69	.05	
2.81	2.66	2.58	2.50	2.42	2.36	2.33	2.25	2.23	2.19	2.16	2.13	.01	
1.33	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.25	1.25	1.24	1.24	.25	28
1.74	1.69	1.66	1.63	1.59	1.57	1.56	1.53	1.52	1.50	1.49	1.48	.10	
2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.73	1.71	1.69	1.67	1.65	.05	
2.75	2.60	2.52	2.44	2.35	2.30	2.26	2.19	2.17	2.13	2.09	2.06	.01	
1.32	1.30	1.29	1.28	1.27	1.26	1.26	1.25	1.24	1.24	1.23	1.23	.25	30
1.72	1.67	1.64	1.61	1.57	1.55	1.54	1.51	1.50	1.48	1.47	1.46	.10	
2.01	1.93	1.89	1.84	1.79	1.76	1.74	1.70	1.68	1.66	1.64	1.62	.05	
2.70	2.55	2.47	2.39	2.30	2.25	2.21	2.13	2.11	2.07	2.03	2.01	.01	
1.30	1.28	1.26	1.25	1.24	1.23	1.22	1.21	1.21	1.20	1.19	1.19	.25	40
1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.43	1.42	1.41	1.39	1.38	.10	
1.92	1.84	1.79	1.74	1.69	1.66	1.64	1.59	1.58	1.55	1.53	1.51	.05	
2.52	2.37	2.29	2.20	2.11	2.06	2.02	1.94	1.92	1.87	1.83	1.80	.01	
1.27	1.25	1.24	1.22	1.21	1.20	1.19	1.17	1.17	1.16	1.15	1.15	.25	60
1.60	1.54	1.51	1.48	1.44	1.41	1.40	1.36	1.35	1.33	1.31	1.29	.10	
1.84	1.75	1.70	1.65	1.59	1.56	1.53	1.48	1.47	1.44	1.41	1.39	.05	
2.35	2.20	2.12	2.03	1.94	1.88	1.84	1.75	1.73	1.68	1.63	1.60	.01	
1.24	1.22	1.21	1.19	1.18	1.17	1.16	1.14	1.13	1.12	1.11	1.10	.25	120
1.55	1.48	1.45	1.41	1.37	1.34	1.32	1.27	1.26	1.24	1.21	1.19	.10	
1.75	1.66	1.61	1.55	1.50	1.46	1.43	1.37	1.35	1.32	1.28	1.25	.05	
2.19	2.03	1.95	1.86	1.76	1.70	1.66	1.56	1.53	1.48	1.42	1.38	.01	
1.23	1.21	1.20	1.18	1.16	1.14	1.12	1.11	1.10	1.09	1.08	1.06	.25	200
1.52	1.46	1.42	1.38	1.34	1.31	1.28	1.24	1.22	1.20	1.17	1.14	.10	
1.72	1.62	1.57	1.52	1.46	1.41	1.39	1.32	1.29	1.26	1.22	1.19	.05	
2.13	1.97	1.89	1.79	1.69	1.63	1.58	1.48	1.44	1.39	1.33	1.28	.01	
1.22	1.19	1.18	1.16	1.14	1.13	1.12	1.09	1.08	1.07	1.04	1.00	.25	∞
1.49	1.42	1.38	1.34	1.30	1.26	1.24	1.18	1.17	1.13	1.08	1.00	.10	
1.67	1.57	1.52	1.46	1.39	1.35	1.32	1.24	1.22	1.17	1.11	1.00	.05	
2.04	1.88	1.79	1.70	1.59	1.52	1.47	1.36	1.32	1.25	1.15	1.00	.01	

附表 4 χ^2 分布的上端百分点



例

$p_r(\chi^2 > 23.83) = 0.25$ for $df=20$

$p_r(\chi^2 > 10.85) = 0.95$

$p_r(\chi^2 > 31.41) = 0.05$

p_r 自由度	.995	.990	.975	.950	.900
1	392704×10^{-10}	157088×10^{-9}	982068×10^{-9}	393214×10^{-8}	.0157908
2	.0100251	.0201007	.0506356	.102587	.210720
3	.0717212	.114832	.215795	.351846	.584375
4	.206990	.297110	.484419	.710721	1.063623
5	.411740	.554300	.831211	1.145476	1.61031
6	.675727	.872085	1.237347	1.63539	2.20413
7	.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	83.3581

.750	.500	.250	.100	.050	0.25	.010	0.005
.1015308	.454937	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
.575364	1.38629	2.77259	4.60517	5.99147	7.37776	9.21034	10.5966
1.212534	2.36597	4.10835	6.25139	7.81473	9.34840	11.3449	12.8381
1.92255	3.35670	5.38527	7.77944	9.48773	11.1433	13.2767	14.8602
2.67460	4.35146	6.62568	9.23635	11.0705	12.8325	15.0863	16.7496
3.45460	5.34812	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476
4.25485	6.34581	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777
5.07064	7.34412	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550
5.89883	8.34283	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893
6.73720	9.34182	12.5489	15.9871	18.3070	20.4831	23.2093	25.1882
7.58412	10.3410	13.7007	17.2750	19.6751	21.9200	24.7250	26.7569
8.43842	11.3403	14.8454	18.5494	21.0261	23.3367	26.2170	28.2995
9.29906	12.3398	15.9839	19.8119	22.3621	24.7356	27.6883	29.8194
10.1653	13.3393	17.1170	21.0642	23.6848	26.1190	29.1413	31.3193
11.0365	14.3389	18.2451	22.3072	24.9958	27.4884	30.5779	32.8013
11.9122	15.3385	19.3688	23.5418	26.2962	28.8454	31.9999	34.2672
12.7919	16.3381	20.4887	24.7690	27.5871	3.1910	33.4087	35.7185
13.6753	17.3379	21.6049	25.9894	28.8693	31.5264	34.8053	37.1564
14.5620	18.3376	22.7178	27.2036	30.1435	32.8523	36.1908	38.5822
15.4518	19.3374	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968
16.3444	20.3372	24.9348	29.6151	32.6705	34.4789	38.9321	41.4010
17.2396	21.3370	26.0393	30.8133	33.9244	36.7807	40.2894	42.7956
18.1373	22.3369	27.1413	32.0069	35.1725	38.0757	41.6384	44.1813
19.0372	23.3367	28.2412	33.1963	36.4151	39.3641	42.9798	45.5585
19.9393	24.3366	29.3389	34.3816	37.6525	40.6465	44.3141	46.9278
20.8434	25.3364	30.4345	35.5631	38.8852	41.9232	45.6417	48.2899
21.7494	26.3363	31.5284	36.7412	40.1133	43.1944	46.9630	49.6449
22.6572	27.3363	32.6205	37.9159	41.3372	44.4607	48.2782	50.9933
23.5666	28.3362	33.7109	39.0875	42.5569	45.7222	49.5879	52.3356
24.4776	29.3360	34.7998	40.2560	43.7729	46.9792	50.8922	53.6720
33.6603	39.3354	45.6160	51.8050	55.7585	59.3417	63.6907	66.7659
42.9421	49.3349	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900
52.2938	59.3347	66.9814	74.3970	79.0819	83.2976	88.3794	91.9517
61.6983	69.3344	77.5766	85.5271	90.5312	95.0231	100.425	104.215
71.1445	79.3343	88.1303	96.5782	101.879	106.629	112.329	116.321
80.6247	89.3342	98.6499	107.565	113.145	118.136	124.116	128.299
90.1332	99.3341	109.141	118.498	124.342	129.561	135.807	140.169

附表 5a 德宾—沃森 d 统计量： 在 0.05 显著性水平上 d_L 和 d_U 的显著点

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$		$k' = 10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	0.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	1.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.303	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.907	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010

60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

续表

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.378	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.078	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	—	—	—	—	—	—	—	—	—	—
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	—	—	—	—	—	—	—	—
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	—	—	—	—	—	—
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.1044	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.647	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.892	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.986	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.929	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

注：n=观测个数， k' =不包含常数项的解释变量个数。

例，若 $n=40$ 和 $k'=4$ ，则 $d_L=1.285$ 和 $d_U=1.721$ 。如果所计算的 d 值小于 1.285，即表明有正的一阶序列相关；如果它大于 1.721，则表明无正的一阶序列相关迹象，但若 d 落在这两个上下限之间，则表明尚无迹象足以判明是否有正的一阶序列相关。

附表 5b 德宾—沃森 d 统计量：在 0.01 显著性水平上 d_L 和 d_U 的显著点

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.390	1.142	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.435	1.036	0.294	1.676	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.497	1.003	0.345	1.489	0.229	2.102	—	—	—	—	—	—	—	—	—	—	—	—
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	—	—	—	—	—	—	—	—	—	—
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	—	—	—	—	—	—	—	—
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	20.30	0.193	2.453	0.124	2.892	—	—	—	—	—	—
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	—	—	—	—
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	—	—
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	1.183	2.667	1.122	2.981	0.078	3.287
15	0.811	1.070	0.700	1.252	0.591	1.464	0.488	1.704	0.391	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101
16	0.844	1.086	0.737	1.252	0.633	1.446	0.532	1.663	0.437	1.900	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944
17	0.874	1.102	0.772	1.255	0.672	1.432	0.574	1.630	0.480	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811
18	0.902	1.118	0.805	1.259	0.708	1.422	0.613	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697
19	0.928	1.132	0.835	1.265	0.742	1.415	0.650	1.584	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597
20	0.952	1.147	0.863	1.271	0.773	1.411	0.685	1.567	0.598	1.737	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510
21	0.975	1.161	0.890	1.277	0.803	1.408	0.718	1.554	0.633	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434
22	0.997	1.174	0.914	1.284	0.831	1.407	0.748	1.543	0.667	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367
23	1.018	1.187	0.938	1.291	0.858	1.407	0.777	1.534	0.698	1.673	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308
24	1.037	1.199	0.960	1.298	0.882	1.407	0.805	1.528	0.728	1.658	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255
25	1.055	1.211	0.981	1.305	0.906	1.409	0.831	1.523	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209
26	1.072	1.222	1.001	1.312	0.928	1.411	0.855	1.518	0.783	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168
27	1.089	1.233	1.019	1.319	0.949	1.413	0.878	1.515	0.808	1.626	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131
28	1.104	1.244	1.037	1.325	0.969	1.415	0.900	1.513	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098
29	1.119	1.254	1.054	1.331	0.988	1.418	0.921	1.512	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.9478	0.595	2.068
30	1.133	1.263	1.070	1.339	1.006	1.421	0.941	1.511	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041
31	1.147	1.273	1.085	1.345	1.023	1.425	0.960	1.510	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017
32	1.160	1.282	1.100	1.352	1.040	1.428	0.979	1.510	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995
33	1.172	1.291	1.114	1.358	1.055	1.432	0.996	1.510	0.936	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975
34	1.184	1.299	1.128	1.364	1.070	1.435	0.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957
35	1.195	1.307	1.140	1.370	1.085	1.439	0.028	1.512	0.971	1.598	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940
36	1.206	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.988	1.588	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925
37	1.217	1.323	1.165	1.382	1.112	1.446	1.058	1.514	1.004	1.586	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.585	0.566	1.658	0.913	1.735	0.860	1.816	0.807	1.899
39	1.237	1.337	1.187	1.393	1.137	1.453	1.085	1.517	1.034	1.584	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887
40	1.246	1.344	1.198	1.398	1.148	1.457	1.098	1.518	1.048	1.584	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876
45	1.288	1.376	1.245	1.423	1.201	1.474	1.156	1.528	1.111	1.584	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834
50	1.324	1.403	1.285	1.446	1.245	1.491	1.205	1.538	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805
55	1.356	1.427	1.320	1.466	1.284	1.506	1.247	1.548	1.209	1.592	1.072	1.638	1.134	1.685	1.095	1.734	1.057	1.785

60	1.383	1.449	1.350	1.484	1.317	1.520	1.283	1.558	1.249	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771
65	1.407	1.468	1.377	1.500	1.346	1.534	1.315	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761
70	1.429	1.485	1.400	1.515	1.372	1.546	1.343	1.578	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.587	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748
80	1.466	1.515	1.441	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745
85	1.482	1.528	1.458	1.553	1.435	1.578	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743
90	1.496	1.540	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.382	1.661	1.360	1.687	1.336	1.714	1.312	1.741
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.642	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741
100	1.522	1.562	1.503	1.583	1.482	1.604	1.462	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768

附表 6 协整检验临界值表

N	模型形式	α	ϕ_{∞}	s.e.	ϕ_1	ϕ_2
1	无常数项, 无趋势项	0.01	-2.5658	(0.0023)	-1.960	-10.04
		0.05	-1.9393	(0.0008)	-0.398	0.0
		0.10	1.6156	(0.0007)	-0.181	0.0
1	常数项, 无趋势项	0.01	-3.4336	(0.0024)	-5.999	-29.25
		0.05	-2.8621	(0.0011)	-2.738	-8.36
		0.10	-2.5671	(0.0009)	-1.438	-4.48
1	常数项, 趋势项	0.01	-3.9638	(0.0019)	-8.353	-47.44
		0.05	-3.4126	(0.0012)	-4.039	-17.83
		0.10	-3.1279	(0.0009)	-2.418	-7.58
2	常数项, 无趋势项	0.01	-3.9001	(0.0022)	-10.534	-30.03
		0.05	-3.3377	(0.0012)	-5.967	-8.98
		0.10	-3.0462	(0.0009)	-4.069	-5.73
2	常数项, 趋势项	0.01	-4.3266	(0.0022)	-15.531	-34.03
		0.05	-3.7809	(0.0013)	-9.421	15.06
		0.10	-3.4959	(0.0009)	-7.203	-4.01
3	常数项, 无趋势项	0.01	-4.2981	(0.0023)	-13.790	-46.37
		0.05	-3.7429	(0.0012)	-8.352	13.41
		0.10	-3.4518	(0.0010)	-6.241	-2.79
3	常数项, 无趋势项	0.01	-4.6676	(0.0022)	-18.492	-49.35
		0.05	-4.1193	(0.0011)	-12.024	-13.13
		0.10	-3.8344	(0.0009)	-9.188	-4.85
4	常数项, 无趋势项	0.01	-4.6493	(0.0023)	-17.188	-59.20
		0.05	-4.1000	(0.0012)	-10.745	-21.57
		0.10	-3.8110	(0.0009)	-8.317	-5.19
4	常数项, 趋势项	0.01	-4.9695	(0.0021)	-22.504	-50.22
		0.05	-4.4294	(0.0012)	-14.501	-19.54
		0.10	-4.1474	(0.0010)	-11.165	-9.88
5	常数项, 无趋势项	0.01	-4.9587	(0.0026)	-22.140	-37.29
		0.05	-4.4185	(0.0013)	-13.641	-21.16
		0.10	-4.1327	(0.0009)	-10.638	-5.48
5	常数项, 趋势项	0.01	-5.2497	(0.0024)	-26.606	-49.56
		0.05	-4.7154	(0.0013)	-17.432	-16.50
		0.10	-4.4345	(0.0010)	-13.654	-5.77
6	常数项, 无趋势项	0.01	-5.2400	(0.0029)	-26.278	-41.65

6	常数项，趋势项	0.05	-4.7048	(0.0018)	-17.120	-11.17
		0.10	-4.4242	(0.0010)	-13.347	0.0
		0.01	-5.5127	(0.0033)	-30.735	-52.50
		0.05	-4.9767	(0.0017)	20.883	-9.05
		0.10	-4.6999	(0.0011)	-16.445	0.0

注：1. 临界值计算公式是 $C(\alpha)=\phi_{\infty}+\phi_1T^{-1}+\phi_2T^{-2}$ ，其中 T 表示样本容量。

2. N 表示协整回归公式中所含变量个数， α 表示检验水平。

3. 摘自 Mackinnon(1991)。

续表

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16	0.060	3.446	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.084	3.286	0.053	3.506	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.113	3.146	0.075	3.358	0.047	3.357	—	—	—	—	—	—	—	—	—	—	—	—
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601	—	—	—	—	—	—	—	—	—	—
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639	—	—	—	—	—	—	—	—
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671	—	—	—	—	—	—
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.142	0.050	3.562	0.032	3.700	—	—	—	—
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725	—	—
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657
26	0.381	2.460	0.324	2.610	0.272	2.758	0.225	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490
28	0.444	2.363	0.387	2.499	0.333	2.9635	0.283	2.772	0.237	0.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	0.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.196	3.087	0.160	3.201
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.614	2.637	0.371	2.733	0.330	2.828
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2566	0.418	2.657	0.377	2.748
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456
55	0.979	1.891	0.940	1.945	0.902	0.002	0.863	2.059	0.825	1.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.934	1.150	1.963	1.126	1.993	1.102	2.023
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883

注： n =观测个数

k' = 不包含常数项的解释变量个数。

参考文献:

- 1、[美]古扎拉蒂著《计量经济学》(第三版),林少宫译,中国人民大学出版社,1999
- 2、[美]J.M.伍德里奇著《计量经济学导论》中国人民大学出版社 2003
- 3、[美]拉姆.拉玛纳山著《应用经济计量学》机械工业出版社 2003
- 4、Damodar N.Gujarrati Basic Econometrics.4th edition.McGraw-Hill Company,2001
- 5、Jeffrey M.Wooldridge Introductory Econometrics ,Second Edition Original language published by Thomson learning
- 6、庞皓主编《计量经济学》,西南财经大学出版社,2001
- 7、李子奈编著《计量经济学》(第二版),高等教育出版社,2005
- 8、张晓峒《计量经济学基础》,南开大学出版社,2001
- 9、张晓峒《计量经济学分析》(修订版),经济科学出版社,2000
- 10、王维国《计量经济学》东北财经大学出版社,2002
- 11、王文博编著《计量经济学》,西安交通大学出版社,2004
- 12、L. 克莱因《经济计量学教科书》谢嘉译,商务印数馆,1983
- 13、张晓峒《计量经济学软件 EViews 使用指南》(第二版)南开大学出版社 2004