专业课辅导+真题资料+微信公众号：432应统书院

普通高等教育"十一五"国家级规划教材

21世纪统计学系列教材

Statistics

Applied Regression Analysis

应用回归分析

（第三版）

何晓群　刘文卿　编著

中国人民大学出版社

专业课辅导+真题资料+微信公众号：432应统书院

统计学概论　　贾俊平

统计学（第四版）　　贾俊平　何晓群　金勇进

（普通高等教育“十一五”国家级规划教材；教育部推荐教材；国家统计局优秀统计教材）

《统计学》（第四版）学习指导书　　贾俊平

应用回归分析（第三版）　　何晓群　刘文卿

（普通高等教育“十一五”国家级规划教材）

统计分析与SPSS的应用（第三版）　　薛　薇

抽样技术（第二版）　　金勇进　杜子芳　蒋　妍

（普通高等教育“十一五”国家级规划教材）

应用时间序列分析（第二版）　　王　燕

非参数统计　　王　星

多元统计分析（第二版）　　何晓群

应用随机过程（第二版）　　张　波

国民经济核算原理与中国实践（第二版）　　高敏雪　李静萍　等

（普通高等教育“十一五”国家级规划教材）

《国民经济核算原理与中国实践》（第二版）学习指导书

高敏雪　等

市场调查方法与技术（第二版）　　简　明　金勇进　蒋　妍

经济社会统计（第二版）　　李静萍　高敏雪

21世纪统计学系列教材

策划编辑　陈永凤　黄　佳
责任编辑　魏　文
封面设计　三众工作室／耿中虎

人大经管图书在线　http://www.rdjg.com.cn
了解图书出版信息　下载教学辅助资料

ISBN 978-7-300-14276-0

ISBN 978-7-300-14276-0
9 787300 142760 >
定价：32.00元

专业课辅导+真题资料+微信公众号：432应统书院

普通高等教育“十一五”国家级规划教材

Statistics 21世纪统计学系列教材

Applied Regression Analysis

应用回归分析

（第三版）

何晓群　刘文卿　编著

中国人民大学出版社
· 北京 ·

专业课辅导+真题资料+微信公众号：432应统书院

图书在版编目（CIP）数据

应用回归分析/何晓群等编著. —3版. —北京：中国人民大学出版社，2011.8
21世纪统计学系列教材
ISBN 978-7-300-14276-0

Ⅰ.①应… Ⅱ.①何… Ⅲ.①回归分析-高等学校-教材 Ⅳ.①O212.1

中国版本图书馆CIP数据核字（2011）第178614号

普通高等教育"十一五"国家级规划教材
21世纪统计学系列教材
应用回归分析（第三版）
何晓群　刘文卿　编著
Yingyong Huigui Fenxi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511398（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	三河汇鑫印务有限公司		
规　　格	185 mm×260 mm　16开本	**版　　次**	2001年6月第1版
			2011年9月第3版
印　　张	18.75 插页1	**印　　次**	2011年9月第1次印刷
字　　数	401 000	**定　　价**	32.00元

版权所有　侵权必究　　印装差错　负责调换

专业课辅导+真题资料+微信公众号：432应统书院

总　序

改革开放以来，高等统计教育有了很大的发展。随着课程设置的不断调整，有不少教材出版，同时也翻译引进了一些国外优秀教材。作为培养我国统计专门人才的摇篮，中国人民大学统计学系自1952年创建以来，走过了风风雨雨，一直坚持着理论与应用相结合的办学方向，培养能够理论联系实际、解决实际问题的高层次人才。随着新知识经济和网络时代的到来，我们在教学科研的实践中，深切地感受到，无论是自然科学领域、社会科学领域的研究，还是国家宏观管理和企业生产经营管理，甚至在人们的日常生活中，信息需求量日益增多，信息处理技术更加复杂，作为信息技术支柱的统计方法，越来越广泛地应用于各个领域。

面对新的形势，我们一直在思索，课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要。在反复酝酿、不断尝试的基础上，我们决定与统计学界的同仁，共同编写、出版一套面向21世纪的统计学系列教材。

这套系列教材聘请了中国科学院院士、中国科技大学陈希孺教授，上海财经大学数量经济研究院张尧庭教授，中国科学院数学与系统科学研究所冯士雍研究员等作为编委。他们长期任中国人民大学的兼职教授，一直关心、支持着统计学的学科建设和应用统计的发展。中国人民大学应用统计科学研究中心2000年已成为国家级研究基地，这些专家是首批专职或兼职研究人员。这一开放性研究基地的运作，将有利于提升我国应用统计科学研究的水平，也必将进一步促进高等统计教育的发展。

这套教材是我们奉献给新世纪的，希望它能够为促进应用统计教育水平的提高增添一份力量。这套教材力求体现以下特点：

第一，在教材选择上，主要面向经济类统计学专业。选材既包括统计教材也包括风险管理与精算方面的教材。尽管名为统计学系列教材，但并不求大、求全，而是力求精选。对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材，并未列入本次出版计划。

第二，每部教材的内容和写作，注意广泛吸收国内外优秀教材的成果。教材力求简明

专业课辅导+真题资料+微信公众号：432应统书院

易懂、内容系统和实用，注重对统计方法思想的阐述，并结合大量实际数据和实例说明统计方法的特点及应用条件。

第三，强调与计算机的结合。为着力提高学生运用统计方法分析解决问题的能力，教材所涉及的统计计算，要求运用目前已有的统计软件。根据教材内容，选择使用 SAS，SPSS，TSP，STATISTICA，EViews，MINITAB，Excel 等。

感谢中国人民大学出版社的同志们，他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望，经过反复论证，使这套教材得以出版。感谢参与教材编写的同行专家、统计学系的教师。愿大家的辛勤劳动能够结出丰硕的果实。我们期待着与统计学界的同仁，共同创造应用统计辉煌的明天。

易丹辉

2000 年 8 月

于中国人民大学

专业课辅导+真题资料+微信公众号：432应统书院

前　言

回归分析是统计学中一个非常重要的分支，在自然科学、管理科学和社会经济等领域有着非常广泛的应用。本书是针对统计学专业和财经管理类专业教学的需要而编写的。

本书写作的指导思想是在不失严谨的前提下，明显不同于纯数理类教材，努力突出实际案例的应用和统计思想的渗透，结合统计软件全面系统地介绍回归分析的实用方法，尽量结合中国社会经济、自然科学等领域的研究实例，把回归分析的方法与实际应用结合起来，注意定性分析与定量分析的紧密结合，努力把同行以及我们在实践中应用回归分析的经验和体会融入其中。

本书2001年出版以来，得到了读者的广泛认可，第三版是继续作为普通高等教育“十一五”国家级规划教材出版的。在这期间，SPSS社会科学统计软件包（Statistical Package for the Social Science）已经从13.0版本提高到18.0版本，使我们可以在第三版中为读者提供更多的实用方法，同时书中增加了更多的例题和练习题。本书的案例主要运用目前在国内最流行的统计软件SPSS 18.0，部分内容用Excel和SAS软件完成，所讲述的方法都结合实例介绍软件的实施过程。每章后面给出本章小结与评注和思考与练习题。

全书共分10章。第1章对回归分析的研究内容和建模过程给出综述性介绍；第2章和第3章详细介绍了一元和多元线性回归的参数估计、显著性检验及其应用；第4章对违背回归模型基本假设的异方差、自相关和异常值等问题给出了诊断和处理方法，在这一章增加了BOX-COX变换；第5章介绍了回归变量选择与逐步回归方法；第6章就多重共线性的产生背景、诊断方法、处理方法等方面结合实际经济问题给予讨论；第7章岭回归估计是解决共线性问题的一种非常实用的方法；第8章介绍了主成分回归与偏最小二乘，这是第三版新增加的内容；第9章介绍了可化为线性回归的曲线回归、多项式回归，以及不能线性化的非线性回归模型的计算；第10章分别介绍了自变量中含定性变量和因变量是定性变量的回归问题，以及因变量是多类别和有序变量的回归问题。

本书作为回归分析的应用教材，讲述的重点是结合SPSS软件使用回归分析中的各种方法，比较各种方法的适用条件，并正确解释分析结果。为了保持教材的完整性，对一些

专业课辅导+真题资料+微信公众号：432应统书院

基本的公式和定理给出了推导和证明，对一些基本的理论性质也做了必要的说明。对统计学专业的本科生可以全面系统地讲述教材的内容；对非统计学专业的本科生应该舍弃其中理论性质的内容；对非统计学专业的研究生可以根据具体情况选择讲授其中的内容。根据我们的教学实践，本书讲授54课时较为合适，若有计算机和投影设备的配合，教学将会更为方便和有效。

在本书的写作过程中得到了中国人民大学21世纪统计学系列教材编审委员会和中国人民大学出版社的支持。编写大纲经过教材编写委员会的认真讨论，教材第一版得到张尧庭、吴喜之两位教授的认真审阅，他们提出了不少中肯的意见。我们的研究生刘达通、于田田、刘龙腾、韩剑雷、赵朋飞、王彦飞为本书第三版的编写作了大量上机计算。本书的大部分案例是我们多年教学和科研工作的积累，部分案例为体现其典型性引用他人著作。在此谨向对本书出版提供帮助的师长和朋友表示衷心的感谢。

本书的完成是我们两位作者多年友好合作的结果，我们期望它的出版能为回归分析在中国的应用做出积极的贡献。由于水平所限，书中难免有不足之处，尤其是在一些应用研究的体会性讨论中，恐有偏颇之处，恳切希望读者批评指正。

何晓群　刘文卿

2011年1月

于中国人民大学明德楼

专业课辅导+真题资料+微信公众号：432应统书院

教师教学服务说明

中国人民大学出版社工商管理分社以出版经典、高品质的工商管理、财务会计、统计、市场营销、人力资源管理、运营管理、物流管理、旅游管理等领域的各层次教材为宗旨。同时，为了更好地服务于一线教师教学，工商管理分社近年来着力建设数字化、立体化的网络教学资源。老师们可以通过以下方式获得免费下载教学资源的权限：

（1）在“人大经管图书在线”（www. rdjg. com. cn）注册并下载“教师服务登记表”，或者直接填写下面的“教师服务登记表”后，加盖院系公章，然后邮寄或者传真给我们。我们收到表格后将在一个工作日内为您开通相关资源的下载权限。

（2）如果您有“人大出版社教研服务网络”（http://www. ttrnet. com）会员卡，可以将卡号发到我们的公共邮箱，无须重复注册，我们将直接为您开通相关专业领域教学资源的下载权限。

如果您需要帮助，请随时联系我们：

联系人：钟馨（010－62515732）　　李文重（010－82501704）
传真：010－62514775　　邮箱：rdcbsjg@crup. com. cn
邮寄地址：北京市海淀区中关村大街甲 59 号文化大厦 1501 室
中国人民大学出版社工商管理分社　邮编：100872

教师服务登记表

姓　　名		□先生　□女士	职　　称		
座机/手机			电子邮箱		
通信地址			邮　　编		
任教学校			所在院系		
所授课程	课程名称	现用教材名称	出版社	对象（本科生/研究生/MBA/其他）	学生人数
需要哪本教材的配套资源					
人大经管图书在线用户名					
院/系领导（签字）： 院/系办公室盖章					

专业课辅导+真题资料+微信公众号：432应统书院

目　录

第 1 章　回归分析概述 …… 1
1.1　变量间的统计关系 …… 1
1.2　回归方程与回归名称的由来 …… 4
1.3　回归分析的主要内容及其一般模型 …… 5
1.4　建立实际问题回归模型的过程 …… 8
1.5　回归分析应用与发展述评 …… 13
思考与练习 …… 15
第 2 章　一元线性回归 …… 16
2.1　一元线性回归模型 …… 16
2.2　参数 β_0，β_1 的估计 …… 21
2.3　最小二乘估计的性质 …… 26
2.4　回归方程的显著性检验 …… 28
2.5　残差分析 …… 39
2.6　回归系数的区间估计 …… 42
2.7　预测和控制 …… 43
2.8　本章小结与评注 …… 47
思考与练习 …… 53
第 3 章　多元线性回归 …… 56
3.1　多元线性回归模型 …… 56
3.2　回归参数的估计 …… 60
3.3　参数估计量的性质 …… 66
3.4　回归方程的显著性检验 …… 69
3.5　中心化和标准化 …… 74
3.6　相关阵与偏相关系数 …… 77

专业课辅导+真题资料+微信公众号：432应统书院

3.7 本章小结与评注 …… 81
思考与练习 …… 87
第4章 违背基本假设的情况 …… 90
4.1 异方差性产生的背景和原因 …… 90
4.2 一元加权最小二乘估计 …… 92
4.3 多元加权最小二乘估计 …… 101
4.4 自相关性问题及其处理 …… 103
4.5 BOX-COX 变换 …… 115
4.6 异常值与强影响点 …… 119
4.7 本章小结与评注 …… 124
思考与练习 …… 126
第5章 自变量选择与逐步回归 …… 130
5.1 自变量选择对估计和预测的影响 …… 130
5.2 所有子集回归 …… 132
5.3 逐步回归 …… 141
5.4 本章小结与评注 …… 150
思考与练习 …… 155
第6章 多重共线性的情形及其处理 …… 158
6.1 多重共线性产生的背景和原因 …… 158
6.2 多重共线性对回归模型的影响 …… 159
6.3 多重共线性的诊断 …… 162
6.4 消除多重共线性的方法 …… 166
6.5 本章小结与评注 …… 168
思考与练习 …… 170
第7章 岭回归 …… 171
7.1 岭回归估计的定义 …… 171
7.2 岭回归估计的性质 …… 173
7.3 岭迹分析 …… 174
7.4 岭参数 k 的选择 …… 176
7.5 用岭回归选择变量 …… 177
7.6 本章小结与评注 …… 185
思考与练习 …… 186
第8章 主成分回归与偏最小二乘 …… 188
8.1 主成分回归 …… 188
8.2 偏最小二乘 …… 194
8.3 本章小结与评注 …… 201
思考与练习 …… 203

专业课辅导+真题资料+微信公众号：432应统书院

第 9 章 非线性回归 …… 204
9.1 可化为线性回归的曲线回归 …… 204
9.2 多项式回归 …… 211
9.3 非线性模型 …… 217
9.4 本章小结与评注 …… 230
思考与练习 …… 232
第 10 章 含定性变量的回归模型 …… 235
10.1 自变量含定性变量的回归模型 …… 235
10.2 自变量含定性变量的回归模型的应用 …… 238
10.3 因变量是定性变量的回归模型 …… 243
10.4 Logistic 回归模型 …… 245
10.5 多类别 Logistic 回归 …… 252
10.6 因变量顺序数据的回归 …… 257
10.7 本章小结与评注 …… 259
思考与练习 …… 261
部分练习题参考答案 …… 267
附录 …… 277
表 1 简单相关系数临界值表 …… 277
表 2 t 分布表 …… 278
表 3 F 分布表 …… 279
表 4 DW 检验上下界表 …… 285

参考文献 …… 287

专业课辅导+真题资料+微信公众号：432应统书院

第1章 Chapter 1 回归分析概述

为了在系统学习回归分析之前对该课程的思想方法、主要内容、发展现状等有个概括的了解，本章将由变量间的统计关系，引申出社会经济与自然科学等现象中的相关与回归问题，并扼要介绍“回归”名称的由来及近代回归分析的发展、回归分析研究的主要内容，以及建立回归模型的步骤与建模过程中应注意的问题。

1.1 变量间的统计关系

社会经济与自然科学等现象之间的相互联系和制约是一个普遍规律。例如社会经济的发展总是与一定的经济变量的数量变化紧密联系的。社会经济现象不仅同和它有关的现象构成一个普遍联系的整体，而且在它的内部存在着许多彼此关联的因素，在一定的社会环境、地理条件、政府决策影响下，一些因素推动或制约另外一些与之联系的因素发生变化。这种状况表明，在经济现象的内部和外部联系中存在着一定的相关性，人们往往利用这种相关关系来制定有关的经济政策，以指导、控制社会经济活动的发展。要认识和掌握客观经济规律就必须探求经济现象间经济变量的变化规律，变量间的统计关系是经济变量变化规律的重要特征。

互有联系的经济现象及经济变量间关系的紧密程度各不一样。一种极端的情况是一个变量的变化能完全决定另一个变量的变化。例如，一个保险公司承保汽车5万辆，每辆保费收入为1 000元，则该保险公司汽车承保总收入为5 000万元。如果把承保总收入记为y，承保汽车辆数记为x，则$y=1\,000x$。x与y两个变量间完全表现为一种确定性关系，即函数关系，如图1.1所示。

再如，银行的一年期存款利率为2.55%，存入的本金用x表示，到期的本息用y表示，则$y=x+2.55\%x$。这里y与x仍表现为一种线性函数关系。对于任意两个变量间的函数关系，可以表述为下面的数学形式

专业课辅导+真题资料+微信公众号：432应统书院

图 1.1 函数关系图

$$y=f(x)$$

再如，工业企业的原材料消耗总额用 y 表示，生产量用 x_1 表示，单位产量消耗用 x_2 表示，原材料价格用 x_3 表示，则

$$y=x_1x_2x_3$$

这里的 y 与 x_1，x_2，x_3 仍是一种确定性的函数关系，但它们显然不是线性函数关系。我们可以将变量 y 与 p 个变量 x_1，x_2，…，x_p 之间存在着的某种函数关系用下面的形式表示

$$y=f(x_1,x_2,\cdots,x_p)$$

经济问题中还有很多函数关系的例子。物理学中的自由落体距离公式、初等数学中的许多计算公式等表示的都是变量间的函数关系。

然而，现实世界中还有不少情况是两事物之间有着密切的联系，但它们密切的程度并没有到由一个可以完全确定另一个。下面举几个例子。

（1）我们都知道某种高档消费品的销售量与城镇居民的收入密切相关，居民收入高了这种消费品的销售量就大。但是由居民收入 x 并不能完全确定某种高档消费品的销售量 y，因为这种高档消费品的销售量还受人们的消费习惯、心理因素、其他商品的吸引程度及价格的高低等诸多因素的影响。这样变量 y 与变量 x 就是一种非确定的关系，见图 1.2。

图 1.2 y 与 x 非确定性关系图

专业课辅导+真题资料+微信公众号：432应统书院

(2) 粮食产量 y 与施肥量 x 之间有着密切的关系，在一定的范围内，施肥量越多，粮食产量就越高。但是，施肥量并不能完全确定粮食产量，因为粮食产量还与其他因素的影响有关，如降雨量、田间管理水平等。因此粮食产量 y 与施肥量 x 之间不存在确定的函数关系。

(3) 储蓄额与居民的收入密切相关，但是由居民收入并不能完全确定储蓄额。因为影响储蓄额的因素很多，如通货膨胀、股票价格指数、利率、消费观念、投资意识等。因此尽管储蓄额与居民收入有密切的关系，但它们之间并不存在一种确定性关系。

再如：广告费支出与商品销售额、保险利润与保费收入、工业产值与用电量等。这方面的例子不胜枚举。

以上变量间关系的一个共同特征是它们之间有密切关系，但这是一种非确定性关系。由于经济问题的复杂性，有许多因素因为我们的认识以及其他客观原因的局限，并没有包含在内，或者由于试验误差、测量误差以及其他种种偶然因素的影响，使得另外一个或一些变量的取值带有一定的随机性。因而当一个或一些变量取定值后，不能以确定值与之对应。

从图 1.1 看到确定性的函数关系，各对应点完全落在一条直线上。而由图 1.2 看到，各对应点并不完全落在一条直线上，即有的点在直线上，有的点在直线的两侧。这种对应点不能分布在一条直线上的变量间的关系，也就是变量 x 与 y 之间有一定的关系，但是又没有密切到可以通过 x 唯一确定 y 的程度，这种关系正是统计学中研究的重要内容。在推断统计中，我们把上述变量间具有密切关联而又不能由某一个或某一些变量唯一确定另外一个变量的关系，称为变量间的统计关系或相关关系。这种统计关系规律性的研究是统计学中研究的主要对象，现代统计学中关于统计关系的研究已形成两个重要的分支，它们叫相关分析和回归分析。

回归分析和相关分析都是研究变量间关系的统计学课题。在应用中，两种分析方法经常相互结合和渗透，但它们研究的侧重点和应用面不同。它们的差别主要有以下几点：一是在回归分析中，变量 y 称为因变量，处在被解释的特殊地位。在相关分析中，变量 y 与变量 x 处于平等的地位，即研究变量 y 与变量 x 的密切程度与研究变量 x 与变量 y 的密切程度是一回事。二是相关分析中所涉及的变量 y 与 x 全是随机变量。而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。通常的回归模型中，我们总是假定 x 是非随机的确定变量。三是相关分析的研究主要是为刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

由于回归分析与相关分析的研究侧重点不同，使得它们的研究方法也大不相同。回归分析已成为现代统计学中应用最广泛、研究最活跃的一个独立分支。

专业课辅导+真题资料+微信公众号：432应统书院

1.2 回归方程与回归名称的由来

回归分析是处理变量 x 与 y 之间的关系的一种统计方法和技术。这里所研究的变量之间的关系就是上述的统计关系，即当给定 x 的值，y 的值不能确定，只能通过一定的概率分布来描述。于是，我们称给定 x 时 y 的条件数学期望

$$f(x)=E(y|x) \tag{1.1}$$

为随机变量 y 对 x 的回归函数，或称为随机变量 y 对 x 的均值回归函数。(1.1) 式从平均意义上刻画了变量 x 与 y 之间的统计规律。

在实际问题中，我们把 x 称为自变量，y 称为因变量。如果要由 x 预测 y，就是要利用 x，y 的观察值，即样本观测值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) \tag{1.2}$$

来建立一个函数，当给定 x 值后，代入此函数中算出一个 y 值，这个值就称为 y 的预测值。如何建立这个函数，这就要从样本观测值（x_i，y_i）出发，观察（x_i，y_i）在平面直角坐标系上的分布情况，图 1.2 就是居民收入与商品销售量的散点图。由这个图可看出样本点基本上分布在一条直线的周围，因而要确定商品销售量 y 与居民收入 x 的关系，可考虑用一个线性函数来描述。图 1.2 中的直线即线性方程

$$E(y|x)=\alpha+\beta x \tag{1.3}$$

方程（1.3）式中的参数 α，β 尚不知道，这就需要由样本数据（1.2）式去进行估计。具体如何估计参数 α，β，我们在第 2 章中将详细介绍。

当我们由样本数据（1.2）式估计出 α，β 的值后，以估计值 $\hat{\alpha}$，$\hat{\beta}$ 分别代替（1.3）式中的 α，β，得方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x \tag{1.4}$$

方程（1.4）式就称为回归方程。这里因为因变量 y 与自变量 x 呈线性关系，故称（1.4）式为 y 对 x 的线性回归方程。又因（1.4）式的建立依赖于观察或试验积累的数据（1.2）式，所以又称（1.4）式为经验回归方程。相对这种叫法，我们把（1.3）式称为理论回归方程。理论回归方程是设想把所研究问题的总体中每一个体的（x，y）值都测量了，利用其全部结果而建立的回归方程，这在实际中是办不到的。理论回归方程中的 α 是方程（1.3）式所画出的直线在 y 轴上的截距，β 为直线的斜率，它们分别称为回归常数和回归系数。而方程（1.4）式中的参数 $\hat{\alpha}$，$\hat{\beta}$ 称为经验回归常数和经验回归系数。

回归分析的基本思想和方法以及“回归”（regression）名称的由来归功于英国统计学家 F. 高尔顿（F. Galton，1822—1911）。高尔顿和他的学生、现代统计学的

专业课辅导+真题资料+微信公众号：432应统书院

奠基者之一 K. 皮尔逊（K. Pearson，1856—1936）在研究父母身高与其子女身高的遗传问题时，观察了1 078对夫妇，以每对夫妇的平均身高作为 x，而取他们的一个成年儿子的身高作为 y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为

$$\hat{y}=33.73+0.516x \tag{1.5}$$

这种趋势及回归方程总的表明父母平均身高 x 每增加一个单位，其成年儿子的身高 y 也平均增加 0.516 个单位。这个结果表明，虽然高个子父辈确有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。反之，矮个子父辈确有生矮个子儿子的趋势，但父辈身高减少一个单位，儿子身高仅减少半个单位左右。通俗地说，一群特高个子父辈（例如排球运动员）的儿子们在同龄人中平均仅为高个子，一群高个子父辈的儿子们在同龄人中平均仅为略高个子；一群特矮个子父辈的儿子们在同龄人中平均仅为矮个子，一群矮个子父辈的儿子们在同龄人中平均仅为略矮个子，即子代的平均高度向中心回归了。正是因为子代的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。这个例子生动地说明了生物学中“种”的概念的稳定性。正是为了描述这种有趣的现象，高尔顿引进了“回归”这个名词来描述父辈身高 x 与子辈身高 y 的关系。尽管“回归”这个名称的由来具有其特定的含义，人们在研究大量的问题中，其变量 x 与 y 之间的关系并不总是具有这种“回归”的含义，但借用这个名词把研究变量 x 与 y 间统计关系的量化方法称为“回归”分析也算是对高尔顿这个伟大的统计学家的纪念。

1.3　回归分析的主要内容及其一般模型

一、回归分析研究的主要内容

回归分析研究的主要对象是客观事物变量间的统计关系，它是建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的统计方法。回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态及进行模型预测的一种有效的工具。

回归分析方法在生产实践中的广泛应用是它发展和完善的根本动力。如果从 19 世纪初（1809 年）高斯（Gauss）提出最小二乘法算起，回归分析的历史已有 200 年。从经典的回归分析方法到近代的回归分析方法，它们所研究的内容已非常丰富。如果按研究的方法来划分，回归分析研究的范围大致如下：

专业课辅导+真题资料+微信公众号：432应统书院

- 回归分析
  - 线性回归
    - 一元线性回归
    - 多元线性回归
    - 多个因变量与多个自变量的回归
  - 回归诊断
    - 讨论如何从数据推断回归模型基本假设的合理性
    - 当基本假设不成立时如何对数据进行修正
    - 判定回归方程拟合的效果
    - 选择回归函数的形式
  - 回归变量的选择
    - 自变量选择的准则
    - 逐步回归分析方法
  - 参数估计方法的改进
    - 岭回归
    - 主成分回归
    - 偏最小二乘法
  - 非线性回归
    - 一元非线性回归
    - 分段回归
    - 多元非线性回归
  - 含有定性变量的回归
    - 自变量含定性变量的情况
    - 因变量是定性变量的情况

二、回归模型的一般形式

如果变量 x_1，x_2，…，x_p 与随机变量 y 之间存在着相关关系，通常就意味着每当 x_1，x_2，…，x_p 取定值后，y 便有相应的概率分布与之对应。随机变量 y 与相关变量 x_1，x_2，…，x_p 之间的概率模型为

$$y=f(x_1,x_2,\cdots,x_p)+\varepsilon \tag{1.6}$$

其中，随机变量 y 称为被解释变量（因变量）；x_1，x_2，…，x_p 称为解释变量（自变量）。在计量经济学中，也称因变量为内生变量，自变量为外生变量。$f(x_1$，x_2，…，$x_p)$ 为一般变量 x_1，x_2，…，x_p 的确定性关系；ε 为随机误差。正是因为随机误差项 ε 的引入，才将变量之间的关系描述为一个随机方程，使得我们可以借助随机数学方法研究 y 与 x_1，x_2，…，x_p 的关系。由于客观经济现象是错综复杂的，一种经济现象很难用有限个因素来准确说明，随机误差项可以概括表示由于人们的认识以及其他客观原因的局限而没有考虑的种种偶然因素。随机误差项主要包括下列因素的影响：

（1）由于人们认识的局限或时间、费用、数据质量等的制约未引入回归模型但又对回归被解释变量 y 有影响的因素。

（2）样本数据的采集过程中变量观测值的观测误差。

（3）理论模型设定的误差。

（4）其他随机因素。

模型（1.6）式清楚地表达了变量 x_1，x_2，…，x_p 与随机变量 y 的相关关系，它由两部分组成：一部分是确定性函数关系，由回归函数 $f(x_1$，x_2，…，$x_p)$ 给出；另一部分是随机误差项 ε。由此可见模型（1.6）式准确地表达了相关关系既有联系又不确定的特点。

专业课辅导+真题资料+微信公众号：432应统书院

当概率模型（1.6）式中回归函数为线性函数时，即有

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{1.7}$$

其中，β_0，β_1，β_2，…，β_p 为未知参数，常称它们为回归系数。线性回归模型的“线性”是针对未知参数 $\beta_i(i=0,1,2,\cdots,p)$ 而言的。回归解释变量的线性是非本质的，因为解释变量是非线性时，常可以通过变量的替换把它转化成线性的。

如果（x_{i1}，x_{i2}，…，x_{ip}；y_i）$(i=1,2,\cdots,n)$ 是（1.7）式中变量（x_1，x_2，…，x_p；y）的一组观测值，则线性回归模型可表示为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i,\quad i=1,2,\cdots,n \tag{1.8}$$

为了估计模型参数的需要，古典线性回归模型通常应满足以下几个基本假设：

（1）解释变量 x_1，x_2，…，x_p 是非随机变量，观测值 x_{i1}，x_{i2}，…，x_{ip} 是常数。

（2）等方差及不相关的假定条件为

$$\begin{cases}E(\varepsilon_i)=0, & i=1,2,\cdots,n\\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j\\ 0, i\neq j\end{cases} & i,j=1,2,\cdots,n\end{cases}$$

这个条件称为高斯-马尔柯夫（Gauss-Markov）条件，简称 G-M 条件。在此条件下，便可以得到关于回归系数的最小二乘估计及误差项方差 σ^2 估计的一些重要性质，如回归系数的最小二乘估计是回归系数的最小方差线性无偏估计等。

（3）正态分布的假定条件为

$$\begin{cases}\varepsilon_i\sim N(0,\sigma^2), & i=1,2,\cdots,n\\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立}\end{cases}$$

在此条件下便可得到关于回归系数的最小二乘估计及 σ^2 估计的进一步的结果，如它们分别是回归系数及 σ^2 的最小方差无偏估计等，并且可以进行回归的显著性检验及区间估计。

（4）通常为了便于数学上的处理，还要求 $n>p$，即样本量的个数要多于解释变量的个数。

在整个回归分析中，线性回归的统计模型最为重要。一方面是因为线性回归的应用最广泛；另一方面是只有在回归模型为线性的假定下，才能得到比较深入和一般的结果；再就是有许多非线性的回归模型可以通过适当的转化变为线性回归问题进行处理。因此，线性回归模型的理论和应用是本书研究的重点。

对线性回归模型通常要研究的问题有：

（1）如何根据样本（x_{i1}，x_{i2}，…，x_{ip}；y_i）$(i=1,2,\cdots,n)$ 求出 β_0，β_1，β_2，…，β_p 及方差 σ^2 的估计。

（2）对回归方程及回归系数的种种假设进行检验。

（3）如何根据回归方程进行预测和控制，以及如何进行实际问题的结构分析。

专业课辅导+真题资料+微信公众号：432应统书院

1.4 建立实际问题回归模型的过程

在实际问题回归分析模型的建立和分析中有几个重要的阶段，为了给读者一个整体印象，我们以经济模型的建立为例，先用逻辑框图表示回归模型的建模过程（见图 1.3）。

下面按逻辑框图顺序叙述每个阶段要做的工作以及应注意的问题。

图 1.3 回归建模步骤流程图

一、根据研究的目的设置指标变量

回归分析模型主要是揭示事物间相关变量的数量联系。首先要根据所研究问题的目的设置因变量 y，然后再选取与 y 有统计关系的一些变量作为自变量。

通常情况下，我们希望因变量与自变量之间具有因果关系。尤其是在研究某种经济活动或经济现象时，必须根据具体的经济现象的研究目的，利用经济学理论，从定性角度来确定某种经济问题中各因素之间的因果关系。当把某一经济变量作为“果”之后，接着更重要的是正确选择作为“因”的变量。在经济问题回归模型中，前者称为“内生变量”或“被解释变量”，后者称为“外生变量”或“解释变量”。变量的正确选择关键在于能否正确把握所研究的经济活动的经济学内涵。这就要求研究者对所研究的经济问题及其背景有足够的了解。例如，要研究中国通货膨胀问题，必须懂得一些金融理论。通常把全国零售物价总指数作为衡量通货膨胀的重要指标，那么，全国零售物价总指数作为被解释变量，影响全国零售物价总指数的有关因素就作为解释变量。

专业课辅导+真题资料+微信公众号：432应统书院

对一个具体的经济问题，当研究目的确定之后，被解释变量就容易确定下来，被解释变量一般直接表达研究的目的。而对被解释变量有影响的解释变量的确定就不太容易：一是由于我们的认识有限，可能并不知道对被解释变量有重要影响的因素。二是为了保证模型参数估计的有效性，设置的解释变量之间应该是不相关的，而我们很难确定哪些变量是相关的，哪些是不相关的，因为在经济问题中很难找到影响同一结果的相互独立的因素。这就看我们如何在多个变量中确定几个重要的且不相关的变量。三是从经济关系角度考虑，非常重要的变量应该引进，但是在实际中并没有这样的统计数据。这一点，在我国建立经济模型时经常会遇到。这时，可以考虑用相近的变量代替，或者由其他几个指标复合成一个新的指标。

在选择变量时要注意与一些专门领域的专家合作。研究金融模型，就要与金融专家和具体业务人员合作；研究粮食生产问题，就要与农业部门的专家合作；研究医学问题，就要与医学专家密切合作。这样做可以帮助我们更好地确定模型变量。

另外，不要认为一个回归模型所涉及的解释变量越多越好。一个经济模型，如果把一些主要变量漏掉肯定会影响模型的应用效果，但如果细枝末节一起进入模型也未必就好。当引入的变量太多时，可能选择了一些与问题无关的变量，还可能由于一些变量的相关性很强，它们所反映的信息有较大的重叠，从而出现共线性问题。当变量太多时，计算工作量太大，计算误差也大，估计出的模型参数精度自然不高。

总之，回归变量的确定是一个非常重要的问题，是建立回归模型最基本的工作。这项工作一般一次并不能完全确定，通常要经过反复试算，最终找出最适合的一些变量。这在今天计算机和相关的统计软件的帮助下，已变得不太困难。

二、收集、整理统计数据

回归模型的建立是基于回归变量的样本统计数据。当确定好回归模型的变量之后，就要对这些变量收集、整理统计数据。数据的收集是建立经济问题回归模型的重要一环，是一项基础性工作，样本数据的质量如何，对回归模型的水平有至关重要的影响。

常用的样本数据分为时间序列数据和横截面数据。

顾名思义，时间序列数据就是按时间顺序排列的统计数据。如新中国建立以来历年的工农业总产值、国民收入、发电量、钢产量、粮食产量等都是每年有一个对应的数据，那么到 2010 年每种指标就有 60 个按时间顺序排列的数据，它们都是时间序列数据。研究宏观经济问题，这方面的时间序列数据来自国家统计局或专业部委的统计年鉴。如果研究微观经济现象，如研究某企业的产值与能耗，数据就要在这个企业的计划统计科获取。

对于收集到的时间序列资料要特别注意数据的可比性和数据的统计口径问题。如历年的国民收入数据，是否按可比价格计算。中国在改革开放前，几十年物价不变，而从 20 世纪 80 年代初开始，物价几乎是直线上升。那么你所获得的数据是否

专业课辅导+真题资料+微信公众号：432应统书院

具有可比性？这就需认真考虑。如在宏观经济研究中，国内生产总值（GDP）与国民生产总值（GNP）二者在内容上是一致的，但在计算口径上不同。国民生产总值按国民原则计算，反映一国常住居民当期在国内外所从事的生产活动；国内生产总值则以国土为计算原则，反映一国国土范围内所发生的生产活动量。对于没有可比性和统计口径不一致的统计数据就要作认真调整，这个调整过程就是数据整理过程。

时间序列数据容易产生模型中随机误差项的序列相关，这是因为许多经济变量的前后期之间总是有关联的。如在建立需求模型时，人们的消费习惯、商品短缺程度等具有一定的延续性，它们对相当一段时间的需求量有影响，这样就产生随机误差项的序列相关。对于具有随机误差项序列相关的情况，就要通过对数据的某种计算整理来消除序列相关性。最常用的处理方法是差分法，我们将在后面的章节中详细介绍。

横截面数据即在同一时间截面上的统计数据。如同一年在不同的地块上测得的施肥量与小麦产量试验的统计数据就是截面数据。再如某一年的全国人口普查数据、工业普查数据、同一年份全国 35 个大中城市的物价指数等都是截面数据。当用横截面数据作样本时，容易产生异方差性。这是因为一个回归模型往往涉及众多解释变量，如果其中某一因素或一些因素随着解释变量观测值的变化而对被解释变量产生不同影响，就产生异方差性。如在研究城镇居民收入与购买消费品的关系时，用 x_i 表示第 i 户的收入量；y_i 表示第 i 户的购买量，购买回归模型为

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\quad i=1,2,\cdots,n \tag{1.9}$$

在此模型中，随机项 ε_i 就具有不同的方差。因为在购买行为中，低收入的家庭购买的差异性比较小，大都购买生活必需品；高收入的家庭购买行为差异很大，高档消费品很多，他们的选择余地很大，这样购买物品所花费用的差异就较大。因而，随机获取的样本数据用来建立回归模型，它的随机项 ε_i 就具有异方差性。

对于具有异方差性的建模问题，数据整理就要注意消除异方差性，这常与模型参数估计方法结合起来考虑。我们将在后面的章节中详细介绍。

不论是时间序列数据还是横截面数据的收集，样本量的多少一般要与设置的解释变量数目相配套。通常为了使模型的参数估计更有效，要求样本量 n 大于解释变量个数 p。样本量的个数小于解释变量数目时，普通的最小二乘估计方法失效。n 与 p 到底应该有怎样一个比例？英国的统计学家 M. 肯德尔（M. Kendall）在他的《多元分析》中指出，样本量 n 应是解释变量个数 p 的 10 倍。如果 p 较大，按肯德尔的说法 n 就很大，这在许多经济问题中是办不到的，尤其新中国才建国 60 多年，统计数据不全是普遍的现象。但由肯德尔的观点我们看到，样本量应比解释变量个数大一些才好，这告诉我们在收集数据时应尽可能多收集一些样本数据。

统计数据的整理中不仅要把一些变量数据进行折算、差分，甚至要把数据对数化、标准化等，有时还须注意剔除个别特别大或特别小的“野值”。在统计数据质

专业课辅导+真题资料+微信公众号：432应统书院

量不高时，经常会碰到这种情况。当然，有时还须利用插值的方法把空缺的数据补齐。

三、确定理论回归模型的数学形式

当收集到所设置的变量的数据之后，就要确定适当的数学形式来描述这些变量之间的关系。绘制变量 y_i 与 $x_i(i=1, 2, \cdots, n)$ 的样本散点图是选择数学模型形式的重要一环。一般我们把（x_i，y_i）所对应的点在平面直角坐标系上画出来，看看散点图的分布状况。如果 n 个样本点大致分布在一条直线的周围，可考虑用线性回归模型去拟合这条直线，也即选择线性回归模型。如果 n 个样本点的分布大致在一条指数曲线的周围，就可选择指数形式的理论回归模型去描述它。

经济回归模型的建立，通常要依据经济理论和一些数理经济学结果。数理经济学中已对投资函数、生产函数、需求函数、消费函数给出了严格的定义，把它们分别用公式表示出来。借用这些理论，我们给它们的公式中增加随机误差项，就可把问题转化为用随机数学工具处理的回归模型。如数理经济学中最有名的生产函数 C-D 生产函数是 20 世纪 30 年代初美国经济学家查尔斯·W·柯布（Charles W. Cobb）和保罗·H·道格拉斯（Paul H. Douglas）根据历史统计数据建立的，资本 K 和劳动 L 与产出被确切地表达为

$$y=AK^{\alpha}L^{\beta} \tag{1.10}$$

其中，α，β 分别为 K 和 L 对产出 y 的弹性。C-D 生产函数指出了厂商行为的一种模式，在函数中变量之间的关系是准确实现的。但是由计量经济学的观点，变量之间的关系并不符合数理经济学所拟定的准确关系模式，而是有随机偏差的。因而给 C-D 生产函数增加一个随机项 U，将变量之间的关系描述为一个随机模型，然后用随机数学方法加以研究，以得出非确定的概率性结论，这更能反映出经济问题的特点。随机模型为

$$y=AK^{\alpha}L^{\beta}U \tag{1.11}$$

或

$$\ln y=\ln A+\alpha\ln K+\beta\ln L+\ln U \tag{1.12}$$

模型（1.11）式是一个非线性的回归模型；模型（1.12）式是一个对数线性回归模型。我们在研究工业生产和农业生产问题时就可考虑用上述的理论模型。

有时候，我们无法根据所获信息确定模型的形式，这时可以采用不同的形式进行计算机模拟，对于不同的模拟结果，选择较好的一个作为理论模型。

尽管模型中待估的未知参数要到参数估计、检验之后才能确定，但在很多情况下可以根据所研究的经济问题对未知参数所带的符号以及大小范围事先给予确定。如 C-D 生产函数（1.11）式中的待估参数 A，α，β 都应为正数。

专业课辅导+真题资料+微信公众号：432应统书院

四、模型参数的估计

回归理论模型确定之后，利用收集、整理的样本数据对模型的未知参数给出估计是回归分析的重要内容。未知参数的估计方法最常用的是普通最小二乘法，它是经典的估计方法。对于不满足模型基本假设的回归问题，人们给出了种种新方法，如岭回归、主成分回归、偏最小二乘估计等。但它们都是以普通最小二乘法为基础，这些具体方法是我们后边一些章节研究的重点。这里要说明的是，参数估计当变量及样本较多时，计算量很大，只有依靠计算机才能得到可靠的准确结果。现在这方面的计算机软件很多，如 MINITAB，SPSS，SAS 等都是参数估计的基本软件。

五、模型的检验与修改

当模型的未知参数估计出来后，就初步建立了一个回归模型。建立回归模型的目的是应用它来研究经济问题，但如果马上就用这个模型去作预测、控制和分析，显然是不够慎重的。因为这个模型是否真正揭示了被解释变量与解释变量之间的关系，必须通过对模型的检验才能决定。一般需要进行统计检验和模型经济意义的检验。

统计检验通常是对回归方程的显著性检验，以及回归系数的显著性检验，还有拟合优度的检验，随机误差项的序列相关检验，异方差性检验，解释变量的多重共线性检验等。这些内容都将在后边的章节中详细讨论。

在经济问题回归模型中，往往还碰到回归模型通过了一系列统计检验，可就是得不到合理的经济解释的情形。例如，国民收入与工农业总产值之间应该是正相关，回归模型中工农业总产值变量前的系数应该为正，但有时候由于样本量的限制或数据质量的问题，可能估计出的系数是负的。如此这般，这个回归模型就没有意义，也就谈不上进一步应用了。可见回归方程经济意义的检验同样是非常重要的。

如果一个回归模型没有通过某种统计检验，或者通过了统计检验而没有合理的经济意义，就需要对回归模型进行修改。模型的修改有时要从设置变量是否合理开始，是不是把某些重要的变量忘记了，变量间是否具有很强的依赖性，样本量是不是太少，理论模型是否合适。譬如某个问题本应用曲线方程去拟合，而我们误用直线方程去拟合，当然通不过检验。这就要重新构造理论模型。

模型的建立往往要反复修改几次，特别是建立一个实际经济问题的回归模型，要反复修正才能得到一个理想模型。

六、回归模型的运用

当一个经济问题的回归模型通过了各种统计检验，且模型具有合理的经济意义

专业课辅导+真题资料+微信公众号：432应统书院

时，就可以运用这个模型来进一步研究经济问题了。

经济变量的因素分析是回归模型的一个重要应用。应用回归模型对经济变量之间的关系作出了度量，从模型的回归系数可发现经济变量的结构关系，给出政策评价的一些量化依据。

既然回归模型揭示经济变量间的因果关系，那么可以考虑给定被解释变量值来控制解释变量值。比如把某年的通货膨胀指标定为全国零售物价指数增长 5%以下，那么，根据通货膨胀的回归模型可以确定货币的发行量、银行的存款利率等。这就是对经济变量的一种控制。

进行经济预测是回归模型的另一个重要应用。比如我国 2015 年的国民收入是多少？通过建立国民经济的宏观经济模型就可以对未来作出预测。用回归模型进行经济预测在我国已有不少成功例子。

在回归模型的运用中，我们还强调定性分析和定量分析的有机结合。这是因为数理统计方法只是从事物的数量表面去研究问题，不涉及事物质的规定性。单纯的表面上的数量关系是否反映事物的本质？这本质究竟如何？必须依靠专门学科的研究才能下定论。所以，在经济问题的研究中，我们不能仅凭样本数据估计的结果就不加分析地说长道短，必须把参数估计的结果和具体经济问题以及现实情况紧密结合，这样才能保证回归模型在经济问题研究中的正确运用。

1.5　回归分析应用与发展述评

从高斯提出最小二乘法算起，回归分析已有 200 年的历史。回归分析的应用非常广泛，我们大概很难找到不用它的领域，这也正是 200 年来其经久不衰、生命力强大的根本原因。

这里仅介绍回归分析在经济领域的广泛应用。我们知道计量经济学是现代经济学中影响最大的一门独立学科，诺贝尔经济学奖获得者萨缪尔森曾经说过：第二次世界大战后的经济学是计量经济学的时代。然而，计量经济学中的基本计量方法就是回归分析，计量经济学的一个重要理论支柱是回归分析理论。

自从 1969 年设立诺贝尔经济学奖以来，已有近 60 位学者获奖，其中绝大部分获奖者是统计学家、计量经济学家、数学家。从大多数获奖者的著作看，他们对统计学及回归分析方法的应用都有娴熟的技巧，这足以说明统计学方法在现代经济研究中的重要作用。

矩阵理论和计算机技术的发展为回归分析模型在经济研究中的应用提供了极大的方便。国民经济是一个错综复杂的系统，一个宏观经济问题常需要涉及几十个甚至几千个变量和方程，如果没有先进的计算机和求解线性方程组的矩阵计算理论，要研究复杂的经济问题是不可想象的。比如一个 20 阶的线性方程组要用克莱姆法则去求解，就需要 10^{22} 次乘法运算，这可是一个天文数字。然而用矩阵变换的方法只需 6 000 次乘法运算。也正是由于计算方法的改进和现代计算机的发展，使得过

专业课辅导+真题资料+微信公众号：432应统书院

去不可想象的事情变成了现实。计量经济学研究中涉及的变量和方程也越来越多，例如英国剑桥大学的多部门动态模型涉及多达 2 759 个方程、7 484 个变量；由诺贝尔经济学奖获得者克莱因发起的国际连接系统，使用了 7 447 个方程和 3 368 个外生变量。

模型技术在经济问题研究中的应用在我国也盛行起来。自 20 世纪 80 年代初期以来，每年都有许多国家级和省部级鉴定的计量经济应用成果诞生。特别是在一些省级以上的重点经济课题和经济学学位论文中，如果没有模型技术的应用，给人的印象总是分量不足。这些足以说明模型技术的应用在我国备受重视。这里要强调说明的是，回归分析方法是模型技术中最基本的内容。

回归分析的理论和方法研究 200 年来也得到不断发展，统计学中的许多重要方法都与回归分析有着密切的联系，如时间序列分析、判别分析、主成分分析、因子分析、典型相关分析等。这些都极大地丰富了统计学方法的宝库。

回归分析方法自身的完善和发展至今是统计学家研究的热点课题。例如自变量的选择、稳健回归、回归诊断、投影寻踪、分位回归、非参数回归模型等近年仍有大量研究文献出现。

在回归模型中，当自变量代表时间、因变量不独立并且构成平稳序列时，这种回归模型的研究就是统计学中的另一个重要分支——时间序列分析。它提供了一系列动态数据的处理方法，帮助人们科学地研究分析所获得的动态数据，从而建立描述动态数据的统计模型，以达到预测、控制的目的。

在前面的回归模型（1.7）式中，因变量 y 和自变量 x 都是一维时，称它为一元回归模型；当 x 是多维，y 是一维时，则它为多元回归模型；若 x 是多维，y 也是多维，则称它为多重回归模型。特别是当因变量观察矩阵 $\boldsymbol{Y}$ 的诸行向量假定是独立的，而列向量假定是相关的，就称为半相依回归方程系统。

对于满足基本假设的回归模型，它的理论已经成熟，但对于违背基本假设的回归模型的参数估计问题近些年仍有较多研究。

在实际问题的研究应用中，人们发现经典的最小二乘估计的结果并不总是令人满意，统计学家从多方面进行努力试图克服经典方法的不足。例如，为了克服设计矩阵的病态性，提出了以岭估计为代表的多种有偏估计。斯泰因（Stein）于 1955 年证明了当维数 p 大于 2 时，正态均值向量最小二乘估计的不可容性，即能够找到另一个估计在某种意义上一致优于最小二乘估计，从此之后人们提出了许多新的估计，其中主要有岭估计、压缩估计、主成分估计、Stein 估计，以及特征根估计。这些估计的共同点是有偏的，即它们的均值并不等于待估参数，于是人们把这些估计称为有偏估计。当设计矩阵 $\boldsymbol{X}$ 呈病态时，这些估计都改进了最小二乘估计。

为了解决自变量个数较多的大型回归模型的自变量的选择问题，人们提出了许多关于回归自变量选择的准则和算法；为了克服最小二乘估计对异常值的敏感性，人们提出了各种稳健回归；为了研究模型假设条件的合理性及样本数据对统计推断影响的大小，产生了回归诊断；为了研究回归模型（1.7）中未知参数非线性的问题，人们提出了许多非线性回归方法，这其中有利用数学规划理论提出的非线性回

专业课辅导+真题资料+微信公众号：432应统书院

归参数估计方法、样条回归方法、微分几何方法等；为了分析和处理高维数据，特别是高维非正态数据，产生了投影寻踪回归、切片回归等。

近年来，新的研究方法不断出现，如非参数统计、自助法、刀切法、经验贝叶斯估计等方法都对回归分析起着渗透和促进作用。

由此看来，回归模型技术随着它自身的不断完善和发展以及应用领域的不断扩大，必将在统计学中占有更重要的位置，也必将为人类社会的发展起着它独到的作用。

□ 思考与练习

1.1　变量间统计关系和函数关系的区别是什么？

1.2　回归分析与相关分析的区别与联系是什么？

1.3　回归模型中随机误差项 ε 的意义是什么？

1.4　线性回归模型的基本假设是什么？

1.5　回归变量设置的理论根据是什么？在设置回归变量时应注意哪些问题？

1.6　收集、整理数据包括哪些内容？

1.7　构造回归理论模型的基本根据是什么？

1.8　为什么要对回归模型进行检验？

1.9　回归模型有哪几个方面的应用？

1.10　为什么强调运用回归分析研究经济问题要定性分析和定量分析相结合？

专业课辅导+真题资料+微信公众号：432应统书院

第 2 章 Chapter 2 一元线性回归

一元线性回归是描述两个变量之间统计关系的最简单的回归模型。一元线性回归虽然简单，但通过一元线性回归模型的建立过程，我们可以了解回归分析方法的基本统计思想以及它在实际问题研究中的应用原理。本章将详细讨论一元线性回归的建模思想、最小二乘估计及其性质、回归方程的有关检验、预测和控制的理论及其应用。

2.1 一元线性回归模型

一、一元线性回归模型的实际背景

在实际问题的研究中，经常需要研究某一现象与影响它的某一最主要因素的关系。如影响粮食产量的因素非常多，但在众多因素中，施肥量是一个最重要的因素，我们往往需要研究施肥量这一因素与粮食产量之间的关系；在消费问题的研究中，影响消费的因素很多，但我们可以只研究国民收入与消费额之间的关系，因为国民收入是影响消费的最主要因素；保险公司在研究火灾损失的规律时，把火灾发生地与最近的消防站的距离作为一个最主要因素，研究火灾损失与火灾发生地和最近的消防站的距离之间的关系。

上述几个例子都是研究两个变量之间的关系，它们的一个共同点是：两个变量之间有着密切的关系，但它们之间密切的程度并不能由一个变量唯一确定另一个变量，即它们间的关系是一种非确定性的关系。那么它们之间到底有什么样的关系呢？这就是下面要进一步研究的问题。

通常我们对所研究的问题首先要收集与它有关的 n 组样本数据（x_i，y_i）

专业课辅导+真题资料+微信公众号：432应统书院

（$i=1, 2, \cdots, n$）。为了直观地发现样本数据的分布规律，我们把（x_i，y_i）看成平面直角坐标系中的点，画出这 n 个样本点的散点图。

例 2.1

假定一保险公司希望确定居民住宅区火灾造成的损失数额与该住户到最近的消防站的距离之间的相关关系，以便准确地定出保险金额。表 2.1 列出了 15 起火灾事故的损失及火灾发生地与最近的消防站的距离。图 2.1 给出了 15 个样本点的分布状况。

表 2.1　火灾损失表

距消防站距离 x（km）	3.4	1.8	4.6	2.3	3.1	5.5	0.7	3.0
火灾损失 y（千元）①	26.2	17.8	31.3	23.1	27.5	36.0	14.1	22.3
距消防站距离 x（km）	2.6	4.3	2.1	1.1	6.1	4.8	3.8	
火灾损失 y（千元）	19.6	31.3	24.0	17.3	43.2	36.4	26.1	

图 2.1

例 2.2

在研究我国城镇人均支出和人均收入之间关系的问题中，把城镇家庭平均每人全年消费性支出记作 y（元）；把城镇家庭平均每人可支配收入记作 x（元）。我们收集到 1990—2009 年 20 年的样本数据（x_i，y_i）（$i=1, 2, \cdots, n$）。数据见表 2.2；样本分布情况见图 2.2。

表 2.2　城镇人均收支表

年份	人均支出 y（元）	人均收入 x（元）	年份	人均支出 y（元）	人均收入 x（元）
1990	1 278.89	1 510.16	1992	1 671.7	2 026.6
1991	1 453.8	1 700.6	1993	2 110.8	2 577.4

① 本书中使用了一些不规范的单位如千、百万等。因原统计数据如此，书中所作回归分析亦使用了这些数据，无法更改，故保持原貌。

专业课辅导+真题资料+微信公众号：432应统书院

续前表

年份	人均支出 y（元）	人均收入 x（元）	年份	人均支出 y（元）	人均收入 x（元）
1994	2 851.3	3 496.2	2002	6 029.92	7 702.8
1995	3 537.57	4 282.95	2003	6 510.94	8 472.2
1996	3 919.5	4 838.9	2004	7 182.1	9 421.6
1997	4 185.6	5 160.3	2005	7 942.88	10 493
1998	4 331.6	5 425.1	2006	8 696.55	11 759.5
1999	4 615.9	5 854	2007	9 997.47	13 785.8
2000	4 998	6 279.98	2008	11 242.85	15 780.76
2001	5 309.01	6 859.6	2009	12 264.55	17 174.65

图 2.2

从图 2.1 和图 2.2 看到，上面两个例子的样本数据点（x_i，y_i）大致分别落在一条直线附近。这说明变量 x 与变量 y 之间具有明显的线性关系。从图上还可以看到，这些样本点又不都在一条直线上，这表明 x 与 y 的关系并没有确切到给定 x 就可以唯一确定 y 的程度。事实上，对 y 产生影响的因素还有许多，如上年收入、消费习惯、银行利率、物价指数等，这些对 y 的取值都有随机影响。每个样本点与直线的偏差就可看做其他随机因素的影响。

二、一元线性回归模型的数学形式

像上面两个例子都是只考虑两个变量间的关系，描述上述 x 与 y 间线性关系的数学结构式可看作上章中回归模型（1.7）式的特例，即当（1.7）式中 $p=1$ 时的情况，亦即

$$y=\beta_0+\beta_1 x+\varepsilon \tag{2.1}$$

专业课辅导+真题资料+微信公众号：432应统书院

（2.1）式将实际问题中变量 y 与 x 之间的关系用两个部分描述：一部分是由于 x 的变化引起的 y 的线性变化，即 $\beta_0+\beta_1 x$；另一部分是由其他一切随机因素引起的，记为 ε。（2.1）式确切地表达了经济变量 x 与 y 之间密切相关，但并没有到由 x 唯一确定 y 的程度。

（2.1）式称为变量 y 对 x 的一元线性理论回归模型。一般我们称 y 为被解释变量（因变量），x 为解释变量（自变量）。式中，β_0 和 β_1 是未知参数，称 β_0 为回归常数，β_1 为回归系数；ε 表示其他随机因素的影响。在（2.1）式中一般假定 ε 是不可观测的随机误差，它是一个随机变量，通常假定 ε 满足

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases} \tag{2.2}$$

这里 $E(\varepsilon)$ 表示 ε 的数学期望，$\mathrm{var}(\varepsilon)$ 表示 ε 的方差。对（2.1）式两端求条件期望，得

$$E(y|x)=\beta_0+\beta_1 x \tag{2.3}$$

称（2.3）式为回归方程。以下把条件期望 $E(y|x)$ 简记为 $E(y)$。

一般情况下，对我们所研究的某个实际问题，如果获得的 n 组样本观测值 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) 符合模型（2.1）式，则

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i, \quad i=1,2,\cdots,n \tag{2.4}$$

由（2.2）式有

$$\begin{cases} E(\varepsilon_i)=0 \\ \mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases} \quad i=1,2,\cdots,n \tag{2.5}$$

通常我们还假定 n 组数据是独立观测的，因而 y_1，y_2，…，y_n 与 ε_1，ε_2，…，ε_n 都是相互独立的随机变量。而 $x_i(i=1, 2, \cdots, n)$ 是确定性变量，其值是可以精确测量和控制的。我们称（2.4）式为一元线性样本回归模型。

（2.1）式的理论回归模型与（2.4）式的样本回归模型是等价的，因而我们常不加区分地将两者统称为一元线性回归模型。

对（2.4）式两边分别求数学期望和方差，得

$$E(y_i)=\beta_0+\beta_1 x_i, \quad \mathrm{var}(y_i)=\sigma^2, \quad i=1,2,\cdots,n \tag{2.6}$$

（2.6）式表明随机变量 y_1，y_2，…，y_n 的期望不等，方差相等，因而 y_1，y_2，…，y_n 是独立的随机变量，但并不是同分布的。而 ε_1，ε_2，…，ε_n 是独立同分布的随机变量。

$E(y_i)=\beta_0+\beta_1 x_i$ 从平均意义上表达了变量 y 与 x 的统计规律性。关于这一点，在应用上非常重要，因为我们经常关心的是这个平均值。例如，在消费 y 与收入 x 的研究中，我们所关心的正是当国民收入达到某个水平时，人均消费能达到多少；在小麦亩产 y 与施肥量 x 的关系中，我们所关心的也正是当施肥量 x 确定后，小麦

专业课辅导+真题资料+微信公众号：432应统书院

的平均产量是多少。

回归分析的主要任务就是通过 n 组样本观测值 $(x_i, y_i)(i=1, 2, \cdots, n)$，对 β_0，β_1 进行估计。一般用 $\hat{\beta}_0$，$\hat{\beta}_1$ 分别表示 β_0，β_1 的估计值，则称

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x \tag{2.7}$$

为 y 关于 x 的一元线性经验回归方程。

通常 $\hat{\beta}_0$ 表示经验回归直线在纵轴上的截距。如果模型范围里包括 $x=0$，则 $\hat{\beta}_0$ 是 $x=0$ 时 y 概率分布的均值；如果不包括 $x=0$，$\hat{\beta}_0$ 只是作为回归方程中的分开项，没有别的具体意义。$\hat{\beta}_1$ 表示经验回归直线的斜率，$\hat{\beta}_1$ 在实际应用中表示自变量 x 每增加一个单位时因变量 y 的平均增加数量。

在实际问题的研究中，为了方便地对参数作区间估计和假设检验，我们还假定模型（2.1）式中误差项 ε 服从正态分布，即

$$\varepsilon \sim N(0,\sigma^2) \tag{2.8}$$

由于 ε_1，ε_2，$\cdots$，ε_n 是 ε 的独立同分布的样本，因而有

$$\varepsilon_i \sim N(0,\sigma^2), \quad i=1,2,\cdots,n \tag{2.9}$$

在 ε_i 服从正态分布的假定下，进一步有随机变量 y_i 也服从正态分布

$$y_i \sim N(\beta_0+\beta_1 x_i,\sigma^2), \quad i=1,2,\cdots,n \tag{2.10}$$

为了在今后的讨论中充分利用矩阵这个处理线性关系的有力工具，我们这里将一元线性回归的一般形式（2.1）式用矩阵表示。令

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{x}=\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2.11}$$

于是模型（2.1）式表示为

$$\begin{cases} \boldsymbol{y}=\boldsymbol{x\beta}+\boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon})=0 \\ \mathrm{var}(\boldsymbol{\varepsilon})=\sigma^2 \boldsymbol{I}_n \end{cases} \tag{2.12}$$

其中，$\boldsymbol{I}_n$ 为 n 阶单位矩阵。

专业课辅导+真题资料+微信公众号：432应统书院

2.2　参数 β_0，β_1 的估计

一、普通最小二乘估计

为了由样本数据得到回归参数 β_0 和 β_1 的理想估计值，我们将使用普通最小二乘估计（ordinary least square estimation，OLSE）。对每一个样本观测值（x_i，y_i），最小二乘法考虑观测值 y_i 与其回归值 $E(y_i)=\beta_0+\beta_1 x_i$ 的离差越小越好，综合考虑 n 个离差值，定义离差平方和为

$$\begin{aligned} Q(\beta_0,\beta_1) &= \sum_{i=1}^{n}(y_i-E(y_i))^2 \\ &= \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \end{aligned} \tag{2.13}$$

所谓最小二乘法，就是寻找参数 β_0，β_1 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，使（2.13）式定义的离差平方和达到极小，即寻找 $\hat{\beta}_0$，$\hat{\beta}_1$，满足

$$\begin{aligned} Q(\hat{\beta}_0,\hat{\beta}_1) &= \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 \\ &= \min_{\beta_0,\beta_1}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \end{aligned} \tag{2.14}$$

依照（2.14）式求出的 $\hat{\beta}_0$，$\hat{\beta}_1$ 就称为回归参数 β_0，β_1 的最小二乘估计。称

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i \tag{2.15}$$

为 $y_i(i=1,2,\cdots,n)$ 的回归拟合值，简称回归值或拟合值。称

$$e_i=y_i-\hat{y}_i \tag{2.16}$$

为 $y_i(i=1,2,\cdots,n)$ 的残差。

从几何关系上看，用一元线性回归方程拟合 n 个样本观测点（x_i，y_i）$(i=1,2,\cdots,n)$，就是要求回归直线 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ 位于这 n 个样本点中间，或者说这 n 个样本点最靠近这条回归直线。由图 2.3 可以直观地理解这种思想。

残差平方和

$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 \tag{2.17}$$

从整体上刻画了 n 个样本观测点（x_i，y_i）$(i=1,2,\cdots,n)$ 到回归直线 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$ 距离的长短。

专业课辅导+真题资料+微信公众号：432应统书院

图 2.3

从（2.14）式中求出 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是一个求极值问题。由于 Q 是关于 $\hat{\beta}_0$，$\hat{\beta}_1$ 的非负二次函数，因而它的最小值总是存在的。根据微积分中求极值的原理，$\hat{\beta}_0$，$\hat{\beta}_1$ 应满足下列方程组

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)x_i=0 \end{cases} \tag{2.18}$$

经整理后，得正规方程组

$$\begin{cases} n\hat{\beta}_0+\left(\sum\limits_{i=1}^{n}x_i\right)\hat{\beta}_1=\sum\limits_{i=1}^{n}y_i \\ \left(\sum\limits_{i=1}^{n}x_i\right)\hat{\beta}_0+\left(\sum\limits_{i=1}^{n}x_i^2\right)\hat{\beta}_1=\sum\limits_{i=1}^{n}x_iy_i \end{cases} \tag{2.19}$$

求解以上正规方程组得 β_0，β_1 的最小二乘估计为

$$\begin{cases} \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \\ \hat{\beta}_1=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} \end{cases} \tag{2.20}$$

其中

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

记

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n(\bar{x})^2 \tag{2.21}$$

$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y} \tag{2.22}$$

专业课辅导+真题资料+微信公众号：432应统书院

则（2.20）式可简写为

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}\\ \hat{\beta}_1=L_{xy}/L_{xx}\end{cases}\tag{2.23}$$

易知，$\hat{\beta}_1$ 可以等价地表示为

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\tag{2.24}$$

或

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n(\bar{x})^2}\tag{2.25}$$

由 $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ 可知

$$\bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}\tag{2.26}$$

可见回归直线 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$ 是通过点（$\bar{x}$，$\bar{y}$）的，这对回归直线的作图很有帮助。从物理学的角度看，（$\bar{x}$，$\bar{y}$）是 n 个样本值（x_i，y_i）的重心，也就是说，回归直线通过样本的重心。

利用上述公式就可以具体计算回归方程的参数。下面以例 2.1 的数据为例，建立火灾损失与距消防站距离之间的回归方程。根据表 2.1 的数据计算得

$$\bar{x}=\frac{49.2}{15}=3.28,\quad \bar{y}=\frac{396.2}{15}=26.413$$

$$\begin{aligned}L_{xx}&=\sum_{i=1}^{n}x_i^2-n(\bar{x})^2\\&=196.16-15\times(3.28)^2=34.784\end{aligned}$$

$$\begin{aligned}L_{xy}&=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}\\&=1\,470.65-1\,299.536=171.114\end{aligned}$$

代入（2.23）式得

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=26.413-4.919\times3.28=10.279\\ \hat{\beta}_1=L_{xy}/L_{xx}=171.114/34.784=4.919\end{cases}$$

于是回归方程为

$$\hat{y}=10.279+4.919x$$

由图 2.1 看出，回归直线与 15 个样本数据点都很接近，这从直观上说明回归

专业课辅导+真题资料+微信公众号：432应统书院

直线对数据的拟合效果是好的。

由（2.18）式可以得到由（2.16）式定义的残差的一个有用的性质

$$\begin{cases}\sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0\end{cases} \tag{2.27}$$

即残差的平均值为 0，残差以自变量 x 的加权平均值为 0。

我们要确定的回归直线就是使它与所有样本数据点都比较靠近，为了刻画这种靠近程度，人们曾设想用绝对残差和，即

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.28}$$

来度量观测值与回归直线的接近程度。显然，绝对残差和越小，回归直线就与所有数据点越近。然而，绝对残差和 $\sum |e_i|$ 在数学处理上比较麻烦，所以在经典的回归分析中，都用残差平方和（2.17）式来描述因变量观测值 $y_i(i=1, 2, \cdots, n)$ 与回归直线的偏离程度。

二、最大似然估计

除了上述的最小二乘估计外，最大似然估计（maximum likelihood estimation，MLE）也可以作为回归参数的估计方法。最大似然估计是利用总体的分布密度或概率分布的表达式及其样本所提供的信息求未知参数估计量的一种方法。

最大似然估计的直观想法可用下面的例子说明：设有一事件 A，已知其发生的概率 p 只可能是 0.01 或 0.1。若在一次试验中事件 A 就发生了，自然应当认为事件 A 发生的概率 p 是 0.1 而不是 0.01。把这种考虑问题的方法一般化，就得到最大似然准则。

当总体 X 为连续型分布时，设其分布密度族为 $\{f(x, \theta), \theta \in \Theta\}$，假设总体 X 的一个独立同分布的样本为 $x_1, x_2, \cdots, x_n$。其似然函数为

$$L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) \tag{2.29}$$

最大似然估计应在一切 θ 中选取使随机样本（$X_1, X_2, \cdots, X_n$）落在点（$x_1, x_2, \cdots, x_n$）附近的概率最大的 $\hat{\theta}$ 为未知参数 θ 真值的估计值，即选取 $\hat{\theta}$ 满足

$$L(\hat{\theta}; x_1, x_2, \cdots, x_n) = \max_{\theta} L(\theta; x_1, x_2, \cdots, x_n) \tag{2.30}$$

对连续型随机变量，似然函数就是样本的联合分布密度函数；对离散型随机变量，似然函数就是样本的联合概率函数。似然函数的概念并不局限于独立同分布的样本，只要样本的联合密度的形式是已知的，就可以应用最大似然估计。

专业课辅导+真题资料+微信公众号：432应统书院

对于一元线性回归模型参数的最大似然估计，如果已经得到样本观测值 $(x_i, y_i)(i=1, 2, \cdots, n)$，其中，$x_i$ 为非随机变量；$y_1, y_2, \cdots, y_n$ 为随机样本，那么在假设 $\varepsilon_i \sim N(0, \sigma^2)$ 时，由（2.10）式知 y_i 服从如下正态分布

$$y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2) \tag{2.31}$$

y_i 的分布密度为

$$f_i(y_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}[y_i-(\beta_0+\beta_1 x_i)]^2\right\}, \quad i=1,2,\cdots,n \tag{2.32}$$

于是 $y_1, y_2, \cdots, y_n$ 的似然函数为

$$\begin{aligned} L(\beta_0,\beta_1,\sigma^2) &= \prod_{i=1}^{n} f_i(y_i) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2\right\} \end{aligned} \tag{2.33}$$

由于 L 的极大化与 $\ln(L)$ 的极大化是等价的，所以取对数似然函数为

$$\ln(L)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2 \tag{2.34}$$

求（2.34）式的极大值，等价于对 $\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2$ 求极小值，到此又与最小二乘原理完全相同。因而 $\hat{\beta}_0$，$\hat{\beta}_1$ 的最大似然估计就是（2.20）式的最小二乘估计。另外，由最大似然估计还可以得到 σ^2 的估计值为

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)]^2 \end{aligned} \tag{2.35}$$

这个估计量是 σ^2 的有偏估计。在实际应用中，常用无偏估计量

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)]^2 \end{aligned} \tag{2.36}$$

作为 σ^2 的估计量。

在此需要注意的是，以上最大似然估计是在 $\varepsilon_i \sim N(0, \sigma^2)$ 的正态分布假设下求得的，而最小二乘估计则对分布假设没有要求。另外，$y_1, y_2, \cdots, y_n$ 是独立的正态分布样本，但并不是同分布的。期望值 $E(y_i)=\beta_0+\beta_1 x_i$ 不相等，但这并不妨碍最大似然方法的应用。

专业课辅导+真题资料+微信公众号：432应统书院

2.3 最小二乘估计的性质

一、线性

所谓线性就是估计量 $\hat{\beta}_0$，$\hat{\beta}_1$ 为随机变量 y_i 的线性函数。由（2.24）式得

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}y_i \tag{2.37}$$

其中，$\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$ 是 y_i 的常数，所以 $\hat{\beta}_1$ 是 y_i 的线性组合。同理可以证明 $\hat{\beta}_0$ 是 y_i 的线性组合，证明过程请读者自己完成。

因为 y_i 为随机变量，所以作为 y_i 的线性组合，$\hat{\beta}_0$，$\hat{\beta}_1$ 亦为随机变量，因此各有其概率分布、均值、方差、标准差及两者的协方差。

二、无偏性

下面我们讨论 $\hat{\beta}_0$，$\hat{\beta}_1$ 的无偏性。由于 x_i 是非随机变量，$y_i=\beta_0+\beta_1x_i+\varepsilon_i$，$E(\varepsilon_i)=0$，因而有

$$E(y_i)=\beta_0+\beta_1x_i \tag{2.38}$$

再由（2.37）式可得

$$\begin{aligned}E(\hat{\beta}_1)&=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}E(y_i)\\&=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}(\beta_0+\beta_1x_i)=\beta_1\end{aligned} \tag{2.39}$$

证得 $\hat{\beta}_1$ 是 β_1 的无偏估计，其中用到 $\sum(x_i-\bar{x})=0$，$\sum(x_i-\bar{x})x_i=\sum(x_i-\bar{x})^2$。同理可证 $\hat{\beta}_0$ 是 β_0 的无偏估计，证明过程请读者自己完成。

无偏估计的意义是，如果屡次变更数据，反复求 β_0，β_1 的估计值，这两个估计量没有高估或低估的系统趋向，它们的平均值将趋于 β_0，β_1。

进一步有

专业课辅导+真题资料+微信公众号：432应统书院

$$
\begin{aligned}
E(\hat{y}) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \beta_0 + \beta_1 x_i \\
&= E(y)
\end{aligned}
\tag{2.40}
$$

这表明回归值 $\hat{y}$ 是 $E(y)$ 的无偏估计，也说明 $\hat{y}$ 与真实值 y 的平均值是相同的。

三、$\hat{\beta}_0$，$\hat{\beta}_1$ 的方差

一个估计量是无偏的，只揭示了估计量优良性的一个方面。我们通常还关心估计量本身的波动状况，这就需进一步研究它的方差。

由 y_1，y_2，…，y_n 相互独立，$\mathrm{var}(y_i)=\sigma^2$ 及（2.37）式，得

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}_1) &= \sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right]^2 \mathrm{var}(y_i) \\
&= \frac{\sigma^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}
\end{aligned}
\tag{2.41}
$$

我们知道，方差的大小表示随机变量取值波动的大小，因而 $\mathrm{var}(\hat{\beta}_1)$ 反映了估计量 $\hat{\beta}_1$ 的波动大小。假设我们反复抽取容量为 n 的样本建立回归方程，每次计算的 $\hat{\beta}_1$ 的值是不相同的，$\mathrm{var}(\hat{\beta}_1)$ 正是反映了这些 $\hat{\beta}_1$ 的差异程度。

由 $\mathrm{var}(\hat{\beta}_1)$ 的表达式我们能得到对实际应用有指导意义的思想。从（2.41）式中看到，回归系数 $\hat{\beta}_1$ 不仅与随机误差的方差 σ^2 有关，而且与自变量 x 的取值离散程度有关。如果 x 的取值比较分散，即 x 的波动较大，则 $\hat{\beta}_1$ 的波动就小，β_1 的估计值 $\hat{\beta}_1$ 就比较稳定。反之，如果原始数据 x 是在一个较小的范围内取值，则 β_1 的估计值稳定性就差，当然也就很难说精确了。这一点显然对我们收集原始数据有重要的指导意义。类似地，有

$$
\mathrm{var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]\sigma^2 \tag{2.42}
$$

由（2.42）式可知，回归常数 $\hat{\beta}_0$ 的方差不仅与随机误差的方差 σ^2 和自变量 x 的取值离散程度有关，而且同样本数据的个数 n 有关。显然 n 越大，$\mathrm{var}(\hat{\beta}_0)$ 越小。

总之，由（2.41）式和（2.42）式可以看到，要想使 β_0，β_1 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$ 更稳定，在收集数据时，就应该考虑 x 的取值尽可能分散一些，不要挤在一块，样本量也应尽可能大一些，样本量 n 太小时，估计量的稳定性肯定不会太好。

由前边 $\hat{\beta}_0$，$\hat{\beta}_1$ 线性的讨论我们知道 $\hat{\beta}_0$，$\hat{\beta}_1$ 都是 n 个独立正态随机变量 y_1，y_2，…，y_n 的线性组合，因而 $\hat{\beta}_0$，$\hat{\beta}_1$ 也服从正态分布。由上边 $\hat{\beta}_0$，$\hat{\beta}_1$ 的均值和方差的结果，有

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n}+\frac{(\bar{x})^2}{L_{xx}}\right)\sigma^2\right) \tag{2.43}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right) \tag{2.44}$$

另外，还可得到 $\hat{\beta}_0$，$\hat{\beta}_1$ 的协方差

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{L_{xx}}\sigma^2 \tag{2.45}$$

（2.45）式说明，在 $\bar{x}=0$ 时，$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不相关，在正态假定下独立；在 $\bar{x}\neq 0$ 时不独立。它揭示了回归系数之间的关系状况。

在前边我们曾给出回归模型随机误差项 ε_i 等方差及不相关的假定条件，这个条件称为高斯-马尔柯夫条件，即

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j \\ 0, \ i\neq j\end{cases} \quad i,j=1,2,\cdots,n \end{cases} \tag{2.46}$$

在此条件下可以证明，$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 分别是 β_0 与 β_1 的最佳线性无偏估计（best linear unbiased estimator，BLUE），也称为最小方差线性无偏估计。BLUE 即指在 β_0 和 β_1 的一切线性无偏估计中，它们的方差最小。此结论本书不予证明，在第 3 章的多元线性回归中也有这个重要结论，其证明请参见参考文献［2］。

进一步可知，对固定的 x_0 来讲

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{2.47}$$

也是 y_1，y_2，…，y_n 的线性组合，且

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right) \tag{2.48}$$

由此可见，$\hat{y}_0$ 是 $E(y_0)$ 的无偏估计，且 $\hat{y}_0$ 的方差随给定的 x_0 值与 $\bar{x}$ 的距离 $|x_0-\bar{x}|$ 的增大而增大。即当给定的 x_0 与 x 的样本平均值 $\bar{x}$ 相差较大时，$\hat{y}_0$ 的估计值波动就增大。这说明在实际应用回归方程进行控制和预测时，给定的 x_0 值不能偏离样本均值太大，否则，用回归方程无论是作因素分析还是作预测，效果都不会理想。

2.4 回归方程的显著性检验

当我们得到一个实际问题的经验回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$ 后，还不能马上就用它去作分析和预测，因为 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$ 是否真正描述了变量 y 与 x 之间的统计规律性，还需运用统计方法对回归方程进行检验。在对回归方程进行检验时，通常需

专业课辅导+真题资料+微信公众号：432应统书院

要正态性假设 $\varepsilon_i \sim N(0, \sigma^2)$，以下的检验内容若无特别声明，都是在此正态性假设下进行的。下面我们介绍几种检验方法。

一、t 检验

t 检验是统计推断中常用的一种检验方法，在回归分析中，t 检验用于检验回归系数的显著性。检验的原假设是

$$H_0: \beta_1 = 0 \tag{2.49}$$

对立假设是

$$H_1: \beta_1 \neq 0 \tag{2.50}$$

回归系数的显著性检验就是要检验自变量 x 对因变量 y 的影响程度是否显著。如果原假设 H_0 成立，则因变量 y 与自变量 x 之间并没有真正的线性关系，也就是说，自变量 x 的变化对因变量 y 并没有影响。由（2.44）式知，$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$，因而当原假设 $H_0: \beta_1 = 0$ 成立时有

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{L_{xx}}\right) \tag{2.51}$$

此时 $\hat{\beta}_1$ 在零附近波动，构造 t 统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \tag{2.52}$$

其中

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.53}$$

是 σ^2 的无偏估计，称 $\hat{\sigma}$ 为回归标准差。由（2.41）式和（2.52）式可以看出，t 统计量就是回归系数的最小二乘估计值除以其标准差的样本估计值。

当原假设 $H_0: \beta_1 = 0$ 成立时，（2.52）式构造的 t 统计量服从自由度为 $n-2$ 的 t 分布。给定显著性水平 α，双侧检验的临界值为 $t_{\alpha/2}$。当 $|t| \geqslant t_{\alpha/2}$ 时拒绝原假设 $H_0: \beta_1 = 0$，认为 β_1 显著不为零，因变量 y 对自变量 x 的一元线性回归成立；当 $|t| < t_{\alpha/2}$ 时接受原假设 $H_0: \beta_1 = 0$，认为 β_1 为零，因变量 y 对自变量 x 的一元线性回归不成立。

以上的公式计算可以手工完成，但在计算机高速发展的今天，许多手工工作都已经被计算机所取代，并且很多有关多元回归的复杂计算不可能手工完成，因此本书的计算工作都是用统计软件实现的。我们会对有关的统计软件结合例题作简要的介绍。

专业课辅导+真题资料+微信公众号：432应统书院

二、用统计软件计算

SPSS 是软件英文名称的首字母缩写，其最初为 Statistical Package for the Social Sciences 的缩写，即“社会科学统计软件包”。2009 年 3 月 19 日，SPSS 公司将 SPSS 四大系列产品（Statistics Family，Modeling Family，Data Collection Family，Deployment Family）整合到一个综合分析平台，把四大类产品统一加上 PASW（为 Predictive Analysis Software 的首字母）前缀，喻义 SPSS 产品的发展方向为预测分析领域。在本书中，还是采用简单的名称，即 SPSS。

目前国际上通用的统计软件主要包括 SPSS 和 SAS 两种，SPSS 的优点是操作界面友好，输出结果美观。它将几乎所有的功能都以统一、规范的界面展现出来，使用 Windows 窗口展示出各种管理和分析数据的功能，以对话框方式展示出各种功能选择项。用户只要掌握一定的 Windows 操作技能，粗通统计分析原理，就可以使用该软件为特定的科研工作服务，或者进行企业级的数据分析。同时只要安装了 R 插件，在 SPSS 中就可以调用 R 的所有统计分析程序，这大大地扩展了 SPSS 统计分析软件的功能。SAS 的优点是功能非常齐全，应用普遍，易于处理大型数据，但缺点是使用相对困难，对于基本统计课程则不那么方便。另外，非常普及的 Excel 软件也有简单的统计计算功能，适合非统计专业人士做一些简单的统计分析。本书的计算主要使用 SPSS 软件。

1. 用 Excel 软件计算

如果工具下拉菜单中没有数据分析项，首先要加入 Excel 的数据分析功能，方法是依次点击工具→加载宏→分析工具库→确定，这时工具下拉菜单中就增加了数据分析项。按要求录入数据，点击数据分析项，在对话框中选择回归，根据对话框的提示作必要的操作，例 2.1 火灾损失数据的输出结果见输出结果 2.1。

输出结果 2.1

回 归 统 计	
Multiple R	0.960978
R Square	0.923478
Adjusted R Square	0.917592
标准误差	2.316346
观测值	15

方差分析

	df	SS	MS	F	Significance F
回归分析	1	841.7664	841.7664	156.8862	1.25E-08
残差	13	69.75098	5.36546		
总计	14	911.5173			

专业课辅导+真题资料+微信公众号：432应统书院

	Coefficients	标准误差	t Stat	P-value	Lower 95 %	Upper 95 %
Intercept	10.27793	1.420278	7.236562	6.59E-06	7.209605	13.34625
X Variable 1	4.919331	0.392748	12.52542	1.25E-08	4.070851	5.76781

输出结果 2.1 中，Intercept 是截距，即回归常数项 β_0，X Variable 1 指第一个自变量，本例为一元线性回归，只有一个自变量 x。Coefficients 列中的两个数值即 $\hat{\beta}_0=10.277\,93$，$\hat{\beta}_1=4.919\,331$，这与例 2.1 的手工计算结果是一致的。另外，$\hat{\beta}_0$ 的标准差为 $\sqrt{\mathrm{var}(\hat{\beta}_0)}=1.420\,278$，$\hat{\beta}_1$ 的标准差为 $\sqrt{\mathrm{var}(\hat{\beta}_1)}=0.392\,748$。(2.52) 式的 t 值为 $t=12.525\,42$，它等于 $4.919\,331/0.392\,748$。取显著性水平 $\alpha=0.05$，自由度为 $n-2=15-2=13$，查 t 分布表得临界值 $t_{\alpha/2}(13)=2.160$，由 $|t|=12.525\,42>2.160$ 可知，应拒绝原假设 $H_0:\beta_1=0$，认为火灾损失 y 对距消防站距离 x 的一元线性回归的效果显著。

另外，从输出结果 2.1 中可以看到回归标准误差 $\hat{\sigma}=2.316\,346$。对回归常数项 β_0 的显著性检验的 t 值为 $t=7.236\,562$，由 $7.236\,562>2.160$ 可知常数项 β_0 显著不为零。不过，我们主要关心的是回归系数 β_1 的显著性，这决定 y 对 x 的回归是否成立，而对回归常数项 β_0 的显著性并不关心。所谓检验显著是指原假设显著不成立，这里原假设是 $H_0:\beta_1=0$，对立假设是 $H_1:\beta_1\neq0$，所以 β_1 显著是指 β_1 显著不等于零。

输出结果 2.1 中的 P-value 是显著性概率值（significance probability value），通常称为 P 值，对回归系数 β_1 的显著性检验的 P 值 $=1.25\text{E-}08=1.25\times10^{-8}\approx0$，检验统计量 t 值与 P 值的关系是

$$P(|t|>|t\text{ 值}|)=P\text{ 值} \tag{2.54}$$

其中，t 为检验统计量，是随机变量，本例中 t 服从自由度为 $n-2$ 的 t 分布。t 值是 t 统计量的样本值，回归系数 β_1 的 $|t\text{ 值}|=12.525\,42$，因而(2.54)式即

$$P(|t|>12.525\,42)=1.25\times10^{-8}$$

其中，$t\sim t(13)$。可以看出，P 值越小，$|t\text{ 值}|$ 越大；P 值越大，$|t\text{ 值}|$ 越小。当 P 值$\leqslant\alpha$ 时，$|t\text{ 值}|\geqslant t_{\alpha/2}$，此时应拒绝原假设 H_0；当 P 值$>\alpha$ 时，$|t\text{ 值}|<t_{\alpha/2}$，此时应接受原假设 H_0；因而可以用 P 值代替 t 值作判定。另外，当 $|t\text{ 值}|=t_{\alpha/2}$ 时，必有 P 值$=\alpha$。

用 P 值代替 t 值作判定有几方面的优越性：

第一，用 P 值作检验不需要查表，只需直接用 P 值与显著性水平 α 相比，当 P 值$\leqslant\alpha$ 时即拒绝原假设 H_0，当 P 值$>\alpha$ 时即接受原假设 H_0，而用 t 值作检验需要查 t 分布表求临界值。

第二，用 P 值作检验具有可比性，而用 t 值作检验与自由度有关，可比性差。

第三，用 P 值作检验可以准确地知道检验的显著性，实际上 P 值就是犯弃真错误的真实概率，也就是检验的真实显著性。

专业课辅导+真题资料+微信公众号：432应统书院

用 P 值作检验的缺点是难以手工计算，但计算机软件可以方便地算出 P 值。

2. 用 SPSS 软件计算

本教材主要采用 SPSS 18.0 版本。进入 SPSS 软件后的界面有两张工作表，第一张是数据工作表 Data View，第二张是变量工作表 Variable View。首先在变量工作表中定义变量的名称、类型、数据宽度、小数位数、变量标签等，其中变量名称要求以英文字母或中文汉字打头不超过 64 个字符，一个中文汉字相当于两个字符。例如 x_1，火灾损失（千元）等名称，必要时还可以用变量标签 Label 对变量做详细说明。可以用 Edit 菜单中的 Options 选项的 General 选项卡选择是显示变量名称还是变量标签。如果没有使用变量标签，则输出结果中就只使用变量名称。早期版本的 SPSS 要求变量名称不超过 8 个字符，常需要使用变量标签对变量做详细说明。SPSS 18.0 版本将变量名称增加到 64 个字符，变量标签的作用就不大了。例 2.1 的因变量的名称为 y，自变量的名称为 x。数据类型等与通常的各种软件都是相同的，这里不再解释。

完成变量定义后再回到数据工作表录入数据，如果已经建立了 Excel 数据文件，则可以把数据直接粘贴到 SPSS 中的数据工作表中。SPSS 软件也提供了直接导入其他数据库文件的功能，方法是点击 File→Open Database→New Query 进入数据导入界面，选择所需要的数据库类型，例如 Excel Files，然后按照软件的提示操作即可。如果没有事先在变量工作表中定义变量而直接在数据工作表中录入数据，这时变量名称默认为 var00001，var00002，…，然后再进入变量工作表中修改变量名称等。建立好数据文件后把它存盘保护好。

对建立好的火灾数据文件，依次点击 Analyze→Regression→Linear 进入线性回归窗口，这是一个对话框形式的窗口。左侧是变量列表，在数据工作表中建立的所有变量名称都列在了这个表中，例 2.1 中只有 y 和 x 两个变量。选中 y，点击右侧 Dependent（因变量）框条旁的箭头按钮，变量 y 即进入此框条（从框条中剔除变量的方法是选中框条中的变量，框条左侧的箭头按钮即转向左侧，点此按钮即可）。用同样的方法把自变量 x 选入 Independent 框条中，再点右侧 OK 按钮，即可得到以下输出结果 2.2。其中 Coefficients 系数表格中 X 所对应的 Sig 值 .000 就是对 β_1 显著性检验的 P 值。对 $\alpha=0.05$，由 Sig=.000<0.05 应拒绝原假设 $H_0:\beta_1=0$，认为 β_1 显著非零。表中数据只保留了 3 位小数，可以双击 SPSS 的输出表格，对选定的数据用右键的 Cell Properties 功能改变数据的显示形式，增加小数位数。Sig 的精确显示数值是 $1.25E\text{-}08=1.25\times10^{-8}$。

输出结果 2.2

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X[a]		Enter

a. All requested variables entered.

b. Dependent Variable: Y.

专业课辅导+真题资料+微信公众号：432应统书院

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.961[a]	.923	.918	2.3163

a. Predictors: (Constant), X.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	841.766	1	841.766	156.886	.000[a]
	Residual	69.751	13	5.365		
	Total	911.517	14			

a. Predictors: (Constant), X.

b. Dependent Variable: Y.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	X	4.919	.393	.961	12.525	.000

a. Dependent Variable: Y.

可以看出，两种软件的计算结果是一致的，只是输出项目和项目名称略有不同。以上只是软件默认的输出项目，还可以根据需要增加输出项目。

三、*F* 检验

对线性回归方程显著性的另外一种检验是 F 检验，F 检验是根据平方和分解式，直接从回归效果检验回归方程的显著性。平方和分解式是

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2.55}$$

其中，$\sum_{i=1}^{n}(y_i-\bar{y})^2$ 称为总离差平方和，简记为 SST 或 $S_{总}$ 或 L_{yy}，SST 表示 Sum of Squares for Total。

$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 称为回归平方和，简记为 SSR 或 $S_{回}$，R 表示 Regression。

$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 称为残差平方和，简记为 SSE 或 $S_{残}$，E 表示 Error。

因而平方和分解式可以简写为

$$SST=SSR+SSE$$

请读者根据（2.27）式自己证明平方和分解式。

总平方和反映因变量 y 的波动程度或称不确定性，在建立了 y 对 x 的线性回归方程后，总平方和 SST 就分解成回归平方和 SSR 与残差平方和 SSE 这两个组成

专业课辅导+真题资料+微信公众号：432应统书院

部分，其中 SSR 是由回归方程确定的，也就是由自变量 x 的波动引起的，SSE 是不能由自变量解释的波动，是由 x 之外的未加控制的因素引起的。这样，总平方和 SST 中，能够由自变量解释的部分为 SSR，不能由自变量解释的部分为 SSE。这样，回归平方和 SSR 越大，回归的效果就越好，可以据此构造 F 检验统计量如下

$$F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)} \tag{2.56}$$

在正态假设下，当原假设 $H_0:\beta_1=0$ 成立时，F 服从自由度为（1，$n-2$）的 F 分布。当 F 值大于临界值 $F_\alpha(1,n-2)$ 时，拒绝 H_0，说明回归方程显著，x 与 y 有显著的线性关系。也可以根据 P 值作检验，具体检验过程可以放在方差分析表中进行，如表 2.3 所示。

表 2.3　　一元线性回归方差分析表

方差来源	自由度	平方和	均方	F 值	P 值
回归	1	SSR	SSR/1	$\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}$	$P(F>F\text{值})=P\text{值}$
残差	$n-2$	SSE	SSE/($n-2$)		
总和	$n-1$	SST			

对例 2.1 的数据，Excel 软件计算的结果见输出结果 2.1 的方差分析表，由表中看到 $F=156.886\,2$，P 值$=1.25\times10^{-8}$。SPSS 软件的输出结果 2.2 中 ANOVA 即方差分析表，ANOVA 表示 Analysis of Variance。两个软件的结果是一致的。

四、相关系数的显著性检验

由于一元线性回归方程讨论的是变量 x 与变量 y 之间的线性关系，所以可以用变量 x 与 y 之间的相关系数来检验回归方程的显著性。设（x_i，y_i）($i=1,2,\cdots,n$）是（x，y）的 n 组样本观测值，我们称

$$\begin{aligned} r&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}\\ &=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \end{aligned} \tag{2.57}$$

为 x 与 y 的简单相关系数，简称相关系数。其中，L_{xy}，L_{xx}，L_{yy} 与前面的定义相同。相关系数 r 表示 x 和 y 的线性关系的密切程度。相关系数的取值范围为 $|r|\leqslant1$。相关系数的直观意义如图 2.4 所示。

专业课辅导+真题资料+微信公众号：432应统书院

图 2.4

图 2.4 中的（a）、（b）和（c）、（d）是四种极端情况，即当 x 与 y 有精确的线性关系时，$r=1$ 或 $r=-1$。$r=1$ 表示 x 与 y 之间完全正相关，所有的对应点都在一条直线上；$r=-1$ 表示 x 与 y 之间完全负相关，对应点也都在一条直线上。这实际上就是一种确定的线性函数关系，它并不是统计学中研究的主要内容。图中（c）这种极端情况，说明所有的样本点分布杂乱无章，变量 x 与 y 之间没有相关关系，即 $r=0$。在实际中 $r=0$ 的情况很少，往往我们拿来毫不相干的两个变量序列，计算相关系数绝对值都会大于零。图中（d）这种情况，表明 x 与 y 有确定的非线性函数关系，或称曲线函数关系。此时 $|r|<1$，并不等于 1，这是因为简单相关系数只是反映两个变量间的线性关系，并不能反映变量间的非线性关系。因而，即使 $r=0$ 也不能说明 x 与 y 无任何关系。

当变量 x 与 y 之间有线性统计关系时，$0<|r|<1$，如图 2.4 中（e）、（f）所示。统计学中主要研究这种非确定性的统计关系。（e）图表示 x 与 y 是正的线性相关，（f）图表示 x 与 y 是负的线性相关。我们在实际问题中经常碰到的是这两种情况。

由（2.57）式和回归系数 $\hat{\beta}_1$ 的表达式可得

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\hat{\beta}_1\sqrt{\frac{L_{xx}}{L_{yy}}} \tag{2.58}$$

专业课辅导+真题资料+微信公众号：432应统书院

由上式可以得到一个很有用的结论，即一元线性回归的回归系数 $\hat{\beta}_1$ 的符号与相关系数 r 的符号相同。

这里需要指出的是，相关系数有个明显的缺点，就是它接近 1 的程度与数据组数 n 有关，这样容易给人一种假象。因为，当 n 较小时，相关系数的绝对值容易接近 1；当 n 较大时，相关系数的绝对值容易偏小。特别是当 $n=2$ 时，相关系数的绝对值总为 1。因此在样本量 n 较小时，我们仅凭相关系数较大就说变量 x 与 y 之间有密切的线性关系，就显得过于草率。在第 3 章的多元线性回归中还将进一步讨论这个问题。

本书附录中有相关系数的检验表，表中是相关系数绝对值的临界值。当我们计算变量 x 与 y 的相关系数绝对值大于表中之值时，才可以认为 x 与 y 有线性关系。通常当 $|r|$ 大于表中 $\alpha=5\%$ 相应的值，但小于表中 $\alpha=1\%$ 相应的值时，称 x 与 y 有显著的线性关系；如果 $|r|$ 大于表中 $\alpha=1\%$ 相应的值，称 x 与 y 有高度显著的线性关系；如果 $|r|$ 小于表中 $\alpha=5\%$ 相应的值，就认为 x 与 y 没有明显的线性关系。

相关系数的计算也可以用软件完成，在 Excel 软件的数据分析对话框中，选择相关系数即可算出例 2.1 数据的相关系数为 $r=0.960\,978$。

例 2.1 中 $n=15$，表中 $\alpha=5\%(n-2=13)$ 相应的值为 0.514，$\alpha=1\%$ 相应的值为 0.641，而 $r=0.961>0.641$。因此说明距消防站距离同火灾损失之间有高度显著的线性依赖关系。

用 SPSS 软件计算相关系数有两种方法。第一种方法是点击 Analyze→Correlate→Bivariate 进入相关系数对话框，点击 Pearson 计算出 x 与 y 的简单相关系数，其中 Bivariate 是二项的含义，表示计算两个变量的相关系数，Pearson 相关系数就是（2.57）式定义的简单相关系数。另外，对话框中还有选项 Two-tail 与 One-tail，代表对相关系数作双侧检验与单侧检验，检验的统计量为

$$t=\frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \tag{2.59}$$

当 $|t|>t_{\alpha/2}(n-2)$ 时，认为 y 与 x 的简单回归系数显著不为零，软件中没有给出 t 值，而是直接给出了 P 值（Sig），对例 2.1 火灾损失的数据，计算出 y 与 x 的相关系数输出结果见输出结果 2.3。同样得到 y 与 x 的相关系数 $r=0.961$，由 P 值近似为零，可知 y 与 x 的简单相关系数显著不为零。

用 SPSS 软件对简单相关系数的另外一种检验方法是直接在线性回归对话框内完成，点击线性回归对话框下面的 Statistics（统计量）选项，进入统计量选项对话框，可以看到默认选项为 Estimates 和 Model fit 两项，再点击 Descriptive，点击右边的 Continue，回到线性回归对话框，计算的输出结果就增加了 y 与 x 的简单回归系数 r 及单侧检验的 P 值。对于对称分布的统计量，单侧检验的 P 值的 2 倍就是双侧检验的 P 值。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 2.3

Correlations

		Y	X
Y	Pearson Correlation	1.000	.961
	Sig. (2-tailed)	.	.000
	N	15	15
X	Pearson Correlation	.961	1.000
	Sig. (2-tailed)	.000	.
	N	15	15

(2.57) 式的相关系数 r 是用样本计算的，也称为样本相关系数。假设我们观测了变量对 (x, y) 的所有取值，此时计算出的相关系数称为总体相关系数，记做 ρ，它反映两变量之间的真实相关程度。样本相关系数 r 是总体相关系数 ρ 的估计值，这个估计值是有误差的。

一般来说，可以将两变量间相关程度的强弱分为以下几个等级：当 $|\rho|\geqslant 0.8$ 时，视为高度相关；当 $0.5\leqslant|\rho|<0.8$ 时，视为中度相关；当 $0.3\leqslant|\rho|<0.5$ 时，视为低度相关；当 $0<|\rho|<0.3$ 时，表明两个变量之间的相关程度极弱，在实际应用中可视为不相关；当 $\rho=0$ 时，两个变量不相关。

在实际应用中，我们往往只能得到样本相关系数 r，而无法得到总体相关系数 ρ。用样本相关系数 r 判定两变量间相关程度的强弱时一定要注意样本量的大小，只有当样本量较大时用样本相关系数 r 判定两变量间相关程度的强弱才可信服。

需要正确区分相关系数显著性检验与相关程度强弱的关系，相关系数的 t 检验显著只是表示总体相关系数 ρ 显著不为零，并不能表示相关程度高。如果有 A，B 两位同学，A 同学计算出 $r=0.8$，但是显著性检验没有通过；B 同学计算出 $r=0.1$，而声称此相关系数高度显著，你能肯定这两位同学都出错了吗？这个问题的回答同样与样本量有关。观察检验统计量 (2.59) 式，可以看到 t 值不仅与样本相关系数 r 有关，同时与样本量 n 有关，对同样的相关系数 r，样本量 n 大时 $|t|$ 就大，样本量 n 小时 $|t|$ 就小。实际上，对任意固定的非零的 r 值，只要样本量 n 充分大就能使 $|t|$ 足够大，从而得到相关系数高度显著的结论。明白这个道理后你就会相信 A，B 两位同学说的都可能是正确的。

在样本量充分大时，可以把样本相关系数 r 作为总体相关系数 ρ，而不必关心显著性检验的结果。你所需要做的事情是结合数据的实际背景判定这样一个 r 值是表示高度相关、中度相关、低度相关，还是视为不相关？前面提到，当 $|\rho|<0.3$ 时可视为不相关，果真是这样吗？如果你被告之，食用含有苏丹红的食品与患癌症之间的相关系数只有 0.2，你是否就可以放心地食用这些食品？如果你得知食用某保健品与健康长寿的相关系数只有 0.2，你是否打算拒绝这种保健品？

五、三种检验的关系

前面介绍了回归系数显著性的 t 检验、回归方程显著性的 F 检验、相关系数的

专业课辅导+真题资料+微信公众号：432应统书院

显著性检验。那么这三种检验之间是否存在一定的关系？回答是肯定的。对一元线性回归这三种检验的结果是完全一致的。可以证明，回归系数显著性的 t 检验与相关系数的显著性检验是完全等价的，（2.52）式与（2.59）式是相等的，而（2.56）式的 F 统计量则是这两个 t 统计量的平方。因而对一元线性回归实际只需要作其中的一种检验即可。然而对多元线性回归这三种检验所考虑的问题不同，所以并不等价，是三种不同的检验。

六、决定系数

由回归平方和与残差平方和的意义我们知道，如果在总离差平方和中回归平方和所占的比重越大，则线性回归效果就越好，这说明回归直线与样本观测值的拟合优度越好；如果残差平方和所占的比重大，则回归直线与样本观测值拟合得就不理想。这里把回归平方和与总离差平方和之比定义为决定系数（coefficient of determination），也称为判定系数，确定系数，记为 r^2，即

$$r^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{2.60}$$

由关系式

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2.61}$$

可以证明（2.60）式的 r^2 正好是（2.57）式中相关系数 r 的平方。即

$$r^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=\frac{L_{xy}^2}{L_{xx}L_{yy}}=(r)^2 \tag{2.62}$$

决定系数 r^2 是一个回归直线与样本观测值拟合优度的相对指标，反映了因变量的变异中能用自变量解释的比例。其数值在 0～1 之间，可以用百分数表示。如果决定系数 r^2 接近 1，说明因变量不确定性的绝大部分能由回归方程解释，回归方程拟合优度就好；反之，如 r^2 不大，说明回归方程的效果不好，应进行修改，可以考虑增加新的自变量或者使用曲线回归。需要注意以下几个方面：

第一，当样本量较小时，与前面在讲述相关系数时所强调的一样，此时即使得到一个大的决定系数，这个决定系数也很可能是虚假现象。为此可以结合样本量和自变量个数对决定系数做调整，计算调整的决定系数。具体计算方法在 5.2 节中讲述。

第二，即使样本量并不小，决定系数很大，例如 0.9，也不能肯定自变量与因变量之间的关系就是线性的，这是因为有可能曲线回归的效果更好。尤其是当自变量的取值范围很窄时，线性回归的效果通常较好，这样的线性回归方程是不能用于外推预测的。可以用模型失拟检验（lack of fit test）来判定因变量与自变

专业课辅导+真题资料+微信公众号：432应统书院

量之间的真实函数关系，到底是线性关系还是曲线关系，如果是曲线关系到底是哪一种曲线关系。这种检验需要对自变量有重复观测数据，而经济数据建模通常不能得到重复观测，这时可以用下面一节介绍的残差分析方法来判定回归方程的正确性。

第三，当你算出一个很小的决定系数 r^2，例如 $r^2=0.1$ 时，与相关系数的显著性检验相似，这时如果样本量 n 不大，就会得到线性回归不显著的检验结论，而在样本量 n 很大时，会得出线性回归显著的结论。不论检验结果是否显著，这时都应该尝试改进回归的效果，例如增加自变量，改用曲线回归等。

对例 2.1 火灾损失的数据，SPSS 输出结果 2.2 中的 R Square 即决定系数 r^2，其值为 $r^2=0.923\,478=(0.960\,978)^2$。Excel 输出结果 2.1 也得到同样的结果。$r^2=0.923\,478$ 表明在 y 值与 $\bar{y}$ 的偏离的平方和中有 92.35%可以通过距消防站距离 x 来解释，这也说明了 y 与 x 之间高度的线性相关关系。

2.5　残差分析

一个线性回归方程通过了 t 检验或 F 检验，只是表明变量 x 与 y 之间的线性关系是显著的，或者说线性回归方程是有效的，但不能保证数据拟合得很好，也不能排除由于意外原因而导致的数据不完全可靠，比如有异常值出现、周期性因素干扰等。只有当与模型中的残差项有关的假定满足时，我们才能放心地运用回归模型。因此，在利用回归方程作分析和预测之前，应该用残差图帮助我们诊断回归效果与样本数据的质量，检查模型是否满足基本假定，以便对模型作进一步的修改。

一、残差的概念与残差图

残差 $e_i=y_i-\hat{y}_i$ 的定义已由（2.16）式给出，n 对数据产生 n 个残差值。残差是实际观测值 y 与通过回归方程给出的回归值之差，残差 e_i 可以看做误差项 ε_i 的估计值。残差 $e_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_i$，误差项 $\varepsilon_i=y_i-\beta_0-\beta_1x_i$，比较两个表达式可以正确区分残差 e_i 与误差项 ε_i 的异同。

以自变量 x 作横轴（或以因变量回归值 $\hat{y}$ 作横轴），以残差作纵轴，将相应的残差点画在直角坐标系上，就可得到残差图，残差图可以帮助我们对数据质量作一些分析。图 2.5 给出常见的一些残差图，这些残差图各不相同，它们分别说明样本数据的不同表现情况。

一般认为，如果一个回归模型满足所给出的基本假定，所有残差应在 $e=0$ 附近随机变化，并在变化幅度不大的一个区域内，见图 2.5 中（a）的情况。反之，这种情况的残差图表明回归模型满足基本假设。

专业课辅导+真题资料+微信公众号：432应统书院

图 2.5 中（b）的情况表明 y 的观测值的方差并不相同，而是随着 x 的增加而增加。这种方差不同的情况的处理将专门在第 4 章中详细讨论。

图 2.5 中（c）的情况表明 y 和 x 之间的关系并非线性关系，而是曲线关系。这就需考虑用另外的曲线方程去拟合样本观测值 y。另外一种可能性是 y 存在自相关。

图 2.5 中（d）的情形称为蛛网现象，表明 y 存在自相关。

图 2.5 残差图

下面对例 2.1 的火灾损失数据作残差分析，首先计算残差。残差 e_i 可以用软件在做回归时直接计算出来，在 SPSS 软件的线性回归对话框中，点击下面的 Save 框条进入 Save 对话框，即可保留所需的中间变量，在 Save 对话框中，点击 Residuals 下的 Unstandardized 选项，再点击右面的 Continue 回到线性回归对话框，继续做回归。回归完成后，在原始数据表格中即可看到新增加了一列变量 res_1，此即残差 e_i。用 Excel 软件同样可以算出残差 e_i。表 2.4 列出了火灾损失数据的残差。

表 2.4 火灾损失数据的残差

序号	x	y	$\hat{y}$	e	ZRE	SRE
1	0.70	14.10	13.721 46	0.378 54	0.163 42	0.189 72
2	1.10	17.30	15.689 19	1.610 81	0.695 41	0.779 10
3	1.80	17.80	19.132 72	−1.332 72	−0.575 36	−0.616 72
4	2.10	24.00	20.608 52	3.391 48	1.464 15	1.549 12
5	2.30	23.10	21.592 39	1.507 61	0.650 86	0.683 89
6	2.60	19.60	23.068 19	−3.468 19	−1.497 27	−1.560 97

专业课辅导+真题资料+微信公众号：432应统书院

续前表

序号	x	y	$\hat{y}$	e	ZRE	SRE
7	3.00	22.30	25.035 92	−2.735 92	−1.181 14	−1.224 07
8	3.10	27.50	25.527 85	1.972 15	0.851 40	0.881 73
9	3.40	26.20	27.003 65	−0.803 65	−0.346 95	−0.359 21
10	3.80	26.10	28.971 39	−2.871 39	−1.239 62	−1.288 50
11	4.30	31.30	31.431 05	−0.131 05	−0.056 58	−0.059 52
12	4.60	31.30	32.906 85	−1.606 85	−0.693 70	−0.738 13
13	4.80	36.40	33.890 72	2.509 28	1.083 29	1.163 48
14	5.50	36.00	37.334 25	−1.334 25	−0.576 01	−0.647 39
15	6.10	43.20	40.285 85	2.914 15	1.258 08	1.498 66

计算出残差后，以自变量 x 为横轴，以残差 e_i 为纵轴画散点图即得残差图。图 2.6 是用 SPSS 软件画出的火灾损失数据的残差图。从残差图上看出，残差是围绕 $e=0$ 随机波动的，从而可以判定模型的基本假定是满足的。

图 2.6　火灾损失数据残差图

二、有关残差的性质

性质 1　$E(e_i)=0$

证明：$E(e_i)=E(y_i)-E(\hat{y}_i)$

$$=(\beta_0+\beta_1 x_i)-(\beta_0+\beta_1 x_i)=0$$

性质 2

$$\mathrm{var}(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}\right]\sigma^2$$

$$=(1-h_{ii})\sigma^2 \tag{2.63}$$

其中，$h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}$，称为杠杆值。$0<h_{ii}<1$，当 x_i 靠近 $\bar{x}$ 时，h_{ii} 的值接近 0，相应的残差方差就大。当 x_i 远离 $\bar{x}$ 时，h_{ii} 的值接近 1，相应的残差方差就小。

专业课辅导+真题资料+微信公众号：432应统书院

也就是说，靠近 $\bar{x}$ 附近的点相应的残差方差较大，远离 $\bar{x}$ 附近的点相应的残差方差较小，这条性质可能令读者感到意外。实际上，远离 $\bar{x}$ 的点数目必然较少，回归线容易“照顾”到这样的少数点，使得回归线接近这些点，因而远离 $\bar{x}$ 附近的 x_i 相应的残差方差较小。

性质 3 残差满足约束条件：$\sum_{i=1}^{n} e_i = 0$，$\sum_{i=1}^{n} x_i e_i = 0$，此关系式已在（2.27）式中给出。这表明残差 e_1，e_2，…，e_n 是相关的，不是独立的。

三、改进的残差

在残差分析中，一般认为超过 $\pm 2\hat{\sigma}$ 或 $\pm 3\hat{\sigma}$ 的残差为异常值，考虑到普通残差 e_1，e_2，…，e_n 的方差不等，用 e_i 作判断和比较会带来一定的麻烦，因此人们引入标准化残差和学生化残差的概念，以改进普通残差的性质，分别定义如下：

标准化残差

$$\mathrm{ZRE}_i = \frac{e_i}{\hat{\sigma}} \tag{2.64}$$

学生化残差

$$\mathrm{SRE}_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{2.65}$$

标准化残差使残差具有可比性，$|\mathrm{ZRE}_i|>3$ 的相应观测值即判定为异常值，这简化了判定工作，但是没有解决方差不等的问题。学生化残差则进一步解决了方差不等的问题，因而在寻找异常值时，用学生化残差优于用普通残差，认为 $|\mathrm{SRE}_i|>3$的相应观测值为异常值。学生化残差的构造公式类似于 t 检验公式，而 t 分布则是 Student（学生）分布的简称，因而把（2.65）式称为学生化残差。在第 4 章我们还将介绍删除残差与学生化残差删除。火灾损失数据的标准化残差与学生化残差已经列在表 2.4 中。

2.6 回归系数的区间估计

当我们用最小二乘法得到 β_0，β_1 的点估计后，在实际应用中往往还希望给出回归系数的估计精度，即给出其置信水平为 $1-\alpha$ 的置信区间。换句话说，就是分别给出以 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 为中心的一个区间，这个区间以 $1-\alpha$ 的概率包含参数 β_0，β_1。置信区间的长度越短，说明估计值 $\hat{\beta}_0$，$\hat{\beta}_1$ 与 β_0，β_1 接近的程度越好，估计值就越精确；置信区间的长度越长，说明估计值 $\hat{\beta}_0$，$\hat{\beta}_1$ 与 β_0，β_1 接近的程度越差，估计值就越不精确。

专业课辅导+真题资料+微信公众号：432应统书院

在实际应用中，我们主要关心回归系数 $\hat{\beta}_1$ 的精度，因而这里只推导 $\hat{\beta}_1$ 的置信区间。根据（2.44）式 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$ 可得

$$t=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\hat{\sigma}^2/L_{xx}}}=\frac{(\hat{\beta}_1-\beta_1)\sqrt{L_{xx}}}{\hat{\sigma}} \tag{2.66}$$

服从自由度为 $n-2$ 的 t 分布。因而

$$P\left(\left|\frac{(\hat{\beta}_1-\beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}\right|<t_{\alpha/2}(n-2)\right)=1-\alpha \tag{2.67}$$

上式等价于

$$P\left(\hat{\beta}_1-t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}<\beta_1<\hat{\beta}_1+t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}\right)=1-\alpha \tag{2.68}$$

即得 β_1 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\hat{\beta}_1-t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}, \hat{\beta}_1+t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}\right) \tag{2.69}$$

在 SPSS 软件中，回归系数的区间估计不是默认的输出结果。在线性回归对话框中，点击下面的统计量 Statistics 框条进入统计量对话框，再点击 Confidence interval，这样在输出的回归系数表中就增加了回归系数的区间估计。用 SPSS 软件计算出的 β_0 和 β_1 的置信度为 95%的置信区间分别为（7.210，13.346）和（4.071，5.768）。

2.7　预测和控制

建立回归模型的目的是为了应用，而预测和控制是回归模型最重要的应用。下面我们专门讨论回归模型在预测和控制方面的应用。

一、单值预测

单值预测就是用单个值作为因变量新值的预测值。比如我们研究某地区小麦亩产量 y 与施肥量 x 的关系时，在 n 块面积为一亩的地块上各施肥 x_i(kg)，最后测得相应的产量 y_i，建立回归方程 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$。当某农户在一亩地块上施肥 $x=x_0$ 时，该地块预期的小麦产量为

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$$

此即因变量新值 $y_0=\beta_0+\beta_1 x_0+\varepsilon_0$ 的单值预测。这里预测目标 y_0 是一个随机变量，因而这个预测不能用普通的无偏性来衡量。根据（2.40）式 $E(\hat{y}_0)=E(y_0)=\beta_0+$

专业课辅导+真题资料+微信公众号：432应统书院

$\beta_1 x_0$，说明预测值 $\hat{y}_0$ 与目标值 y_0 有相同的均值。

二、区间预测

以上的单值预测 $\hat{y}_0$ 只是这个地块小麦产量的大概值。仅知道这一点意义并不大，对于预测问题，除了知道预测值外，还希望知道预测的精度，这就需要作区间预测，也就是给出小麦产量的一个预测值范围。给一个预测值范围比只给出单个值 $\hat{y}_0$ 更可信，这个问题也就是对于给定的显著性水平 α，找一个区间（T_1，T_2），使对应于某特定的 x_0 的实际值 y_0 以 $1-\alpha$ 的概率被区间（T_1，T_2）所包含，用公式表示就是

$$P(T_1<y_0<T_2)=1-\alpha \tag{2.70}$$

对因变量的区间预测又分为两种情况：一种是因变量新值的区间预测；另一种是因变量新值的平均值的区间预测。

1. 因变量新值的区间预测

为了给出新值 y_0 的置信区间，需要首先求出其估计值 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$ 的分布。由于 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 都是 y_1，y_2，…，y_n 的线性组合，因而 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$ 也是 y_1，y_2，…，y_n 的线性组合，在正态假定下 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$ 服从正态分布，其期望值为 $E(\hat{y}_0)=\beta_0+\beta_1 x_0$，以下计算其方差，首先

$$\begin{aligned}\hat{y}_0 &= \hat{\beta}_0+\hat{\beta}_1 x_0\\ &= \bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1 x_0\\ &= \sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{L_{xx}}\right]y_i\end{aligned} \tag{2.71}$$

因而有

$$\begin{aligned}\mathrm{var}(\hat{y}_0) &= \sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{L_{xx}}\right]^2\mathrm{var}(y_i)\\ &= \left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right]\sigma^2\end{aligned} \tag{2.72}$$

从而得

$$\hat{y}_0\sim N\left\{\beta_0+\beta_1 x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right]\sigma^2\right\} \tag{2.73}$$

记

$$h_{00}=\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}} \tag{2.74}$$

为新值 x_0 的杠杆值，则上式简写为

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{y}_0 \sim N(\beta_0+\beta_1 x_0, h_{00}\sigma^2) \tag{2.75}$$

$\hat{y}_0$ 是先前独立观测到的随机变量 y_1，y_2，…，y_n 的线性组合，现在小麦产量的新值 y_0 与先前的观测值是独立的，所以 y_0 与 $\hat{y}_0$ 是独立的。因而

$$\begin{aligned}\mathrm{var}(y_0-\hat{y}_0)&=\mathrm{var}(y_0)+\mathrm{var}(\hat{y}_0)\\&=\sigma^2+h_{00}\sigma^2\end{aligned} \tag{2.76}$$

再由（2.40）式知 $E(y_0-\hat{y}_0)=0$，于是有

$$y_0-\hat{y}_0 \sim N(0,(1+h_{00})\sigma^2) \tag{2.77}$$

进而可知统计量

$$t=\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\,\hat{\sigma}} \sim t(n-2) \tag{2.78}$$

可得

$$P\left(\left|\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\,\hat{\sigma}}\right| \leqslant t_{\alpha/2}(n-2)\right)=1-\alpha \tag{2.79}$$

由此可以求得 y_0 的置信概率为 $1-\alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1+h_{00}}\,\hat{\sigma} \tag{2.80}$$

当样本量 n 较大，$|x_0-\bar{x}|$ 较小时，h_{00} 接近零，y_0 的置信度为 95%的置信区间近似为

$$\hat{y}_0 \pm 2\hat{\sigma} \tag{2.81}$$

由（2.80）式可看到，对给定的显著性水平 α，样本量 n 越大，$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$ 越大，x_0 越靠近 $\bar{x}$，则置信区间长度越短，此时的预测精度就高。所以，为了提高预测精度，样本量 n 应越大越好，采集数据 x_1，x_2，…，x_n 不能太集中。在进行预测时，所给定的 x_0 不能偏离 $\bar{x}$ 太大，否则预测结果肯定不好；如果给定值 $x_0=\bar{x}$，置信区间长度最短，这时的预测结果最好。因此，如果在自变量观测值之外的范围作预测，精度就较差。这种情况进一步说明当 x 的取值发生较大变化时，即 $|x_0-\bar{x}|$ 很大时，预测就不准。所以在作预测时一定要看 x_0 与 $\bar{x}$ 相差多大，相差太大，效果肯定不好。尤其是在经济问题的研究中作长期预测时，x 的取值 x_0 肯定距当时建模时采集样本的 $\bar{x}$ 相差很大。比如，我们用人均国民收入 1 000 元左右的数据建立的消费基金模型，只适合近期人均收入 1 000 元左右的消费基金预测，而若干年后人均国民收入增长幅度变化较大时，以及人的消费观念发生较大变化时，用原模型去作预测肯定不准。

2. 因变量新值的平均值的区间预测

（2.80）式给出的是因变量单个新值的置信区间，我们关心的另外一种情况是

专业课辅导+真题资料+微信公众号：432应统书院

因变量新值的平均值的区间估计。对于前面提出的小麦产量问题，如果该地区的一大片麦地每亩施肥量同为 x_0，那么这一大片地小麦的平均亩产如何估计呢？这个问题就是要估计平均值 $E(y_0)$，根据（2.40）式，$E(y_0)$ 的点估计仍为 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$，但是其区间估计却与因变量单个新值 y_0 的置信区间（2.80）式有所不同，由于 $E(y_0)=\beta_0+\beta_1x_0$ 是常数，由（2.73）式知

$$\hat{y}_0-E(y_0)\sim N\left(0,\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right) \tag{2.82}$$

进而可得置信水平为 $1-\alpha$ 的置信区间为

$$\hat{y}_0\pm t_{\alpha/2}(n-2)\sqrt{h_{00}}\,\hat{\sigma} \tag{2.83}$$

用 SPSS 软件可以直接计算出因变量单个新值 y_0 与平均值 $E(y_0)$ 的置信区间，方法是在计算回归之前，把自变量新值 x_0 输入样本数据中，而因变量的相应值空缺，然后在 Save 对话框中点击 Mean 计算因变量平均值 $E(y_0)$ 的置信区间，或点击 Individual 计算因变量单个新值 y_0 的置信区间，也可以二者同时点击，还可以选择置信水平。另外，点击 Predicted values 下面的 Unstandardized 对话框可以计算出点估计值 $\hat{y}_0$，计算结果列在原始数据表中。对例 2.1 的火灾损失数据，假设保险公司希望预测一个距离最近的消防队 $x_0=3.5$ 公里的居民住宅如果失火的损失额，用 SPSS 软件计算出点估计值 $\hat{y}_0$ 以及置信水平为 95%的置信区间为：

点估计值 $\hat{y}_0$：27.50（千元）

单个新值：(22.32，32.67)

平均值 $E(y_0)$：(26.19，28.80)

用（2.81）式的近似公式计算单个新值置信水平为 95%的近似置信区间为

$$\begin{aligned}(\hat{y}_0-2\hat{\sigma},\hat{y}_0+2\hat{\sigma})&=(27.50-2\times2.316,27.50+2\times2.316)\\&=(22.87,32.13)\end{aligned}$$

这个近似的置信区间与精确的置信区间（22.32，32.67）很接近。如果用手工计算，多数场合可以用近似区间。

有的软件和教材中把因变量平均值的区间预测称为置信区间（confidence interval），把因变量单个值的区间预测称为预测区间（prediction interval）。

三、控制问题

控制问题相当于预测的反问题。预测和控制有着密切的关系。在许多经济问题中，我们要求 y 在一定的范围内取值。比如在研究近年的经济增长率时，我们希望经济增长能保持在 8%～12%；在控制通货膨胀问题中，我们希望全国零售物价指数增长在 5%以内等。这些问题用数学表达式描述，即要求

$$T_1<y<T_2$$

专业课辅导+真题资料+微信公众号：432应统书院

问题是如何控制 x 呢？对于前面谈到的经济问题，即如何控制影响经济增长和通货膨胀的最主要因素呢？在统计学中进一步要讨论如何控制自变量 x 的值才能以 $1-\alpha$ 的概率保证把目标值 y 控制在 $T_1<y<T_2$ 中，即

$$P(T_1<y<T_2)=1-\alpha \tag{2.84}$$

其中，α 是事先给定的小的正数，$0<\alpha<1$。

我们通常用近似的预测区间来确定 x。如果 $\alpha=0.05$，根据（2.81）式，可由下述不等式组

$$\begin{cases}\hat{y}(x)-2\hat{\sigma}>T_1\\ \hat{y}(x)+2\hat{\sigma}<T_2\end{cases} \tag{2.85}$$

求出 x 的取值区间，将 $\hat{y}(x)=\hat{\beta}_0+\hat{\beta}_1x$ 代入求得

当 $\hat{\beta}_1>0$ 时，得

$$\frac{T_1+2\hat{\sigma}-\hat{\beta}_0}{\hat{\beta}_1}<x<\frac{T_2-2\hat{\sigma}-\hat{\beta}_0}{\hat{\beta}_1} \tag{2.86}$$

当 $\hat{\beta}_1<0$ 时，得

$$\frac{T_2-2\hat{\sigma}-\hat{\beta}_0}{\hat{\beta}_1}<x<\frac{T_1+2\hat{\sigma}-\hat{\beta}_0}{\hat{\beta}_1} \tag{2.87}$$

控制问题的应用要求因变量 y 与自变量 x 之间有因果关系，经常用在工业生产的质量控制中，这方面的例子参见参考文献［7］。在经济问题中，经济变量之间有强的相关性，形成一个综合的整体，仅控制回归方程中的一个或几个自变量，而忽视回归方程之外的其他变量，往往达不到预期的效果。

2.8　本章小结与评注

本章通过两个例子系统介绍了一元线性回归模型概念引入的实际背景，以及回归模型未知参数的估计、最小二乘估计的性质、回归方程的显著性检验、回归系数的区间估计、残差分析的基本概念和方法、回归模型的主要应用、预测和控制等问题。

一元线性回归模型虽然比较简单，但它的统计思想非常重要。后面将要介绍的多元线性回归的很多内容是一元线性回归结果的直接推广，所以有必要对一元线性回归建模及应用方面多作一些讨论，以便使我们对回归分析方法的思想实质有更深的体会。

专业课辅导+真题资料+微信公众号：432应统书院

一、一元线性回归模型从建模到应用的全过程

第一步，提出因变量与自变量。这里以例 2.2 的数据为例，本例因变量 y 为城镇家庭平均每人全年消费性支出（元），自变量 x 为城镇家庭平均每人可支配收入（元），采用年份数据。

第二步，收集数据。从中经网统计数据库中可查得表 2.2。

第三步，根据表 2.2 的数据画散点图（见图 2.2）。

第四步，设定理论模型。由图 2.2 我们看到，随着人均可支配收入的增加，居民人均消费增加，而且 20 个样本点大致分布在一条直线的周围。因此，用直线回归模型去描述它们是合适的。故可以采用（2.4）式一元线性回归理论模型。

第五步，用软件计算，输出计算结果。

本例使用 SPSS 软件，选中全部输出统计量，得到输出结果 2.4。

输出结果 2.4

Descriptive Statistics

	Mean	Std. Deviation	N
y	5506.5465	3221.17514	20
x	7230.1050	4608.65532	20

Correlations

		y	x
Pearson Correlation	y	1.000	.999
	x	.999	1.000
Sig. (1-tailed)	y	.	.000
	x	.000	.
N	y	20	20
	x	20	20

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.999 [a]	.998	.998	153.82165	.300

a. Predictors: (Constant), x.
b. Dependent Variable: y.

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.967E8	1	1.967E8	8313.963	.000 [a]
	Residual	425899.816	18	23661.101		
	Total	1.971E8	19			

a. Predictors: (Constant), x.
b. Dependent Variable: y.

专业课辅导+真题资料+微信公众号：432应统书院

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	458.596	65.177		7.036	.000	321.665	595.527
	x	.698	.008	.999	91.181	.000	.682	.714

a. Dependent Variable: y.

第六步，回归诊断，分析输出结果。

（1）从 Descriptive Statistics 描述统计表中看到，$\bar{x}=7\ 230.105$，$\bar{y}=5\ 506.546\ 5$，有效样本量 $n=20$。x 的标准差 $S_x=4\ 608.655$，y 的标准差 $S_y=3\ 221.175$。

（2）从 Correlations 相关表中看到，相关系数 $r=0.999$，单侧检验显著性概率 (Sig)=9.514E－26≈0.000，双侧检验 Sig=2×9.514E－26=1.903E－25≈0.000，说明 y 与 x 有显著的线性相关，这与散点图的直观分析是一致的。需要注意的是，SPSS 通常只显示 3 位小数，可以双击输出表格选定所需数据用鼠标右键更改数据显示位数。

（3）从 Model Summary 表中看到，决定系数 $r^2=0.998$，从相对水平上看，回归方程能够减少因变量 y 的 99.8%的方差波动。回归标准差 $\hat{\sigma}=153.822$，从绝对水平上看，y 的标准差从回归前的 3 221.175 减少到回归后的 153.822。

（4）从 ANOVA 方差分析表中看到，$F=8\ 313.963$，Sig=1.903E－25，说明 y 与 x 的线性回归高度显著，这与相关系数的检验结果是一致的。

（5）从 Coefficients 系数表中得到回归方程为 $\hat{y}=458.596+0.698x$，回归系数 β_1 检验的 t 值为 91.181，Sig=1.903E－25，与 F 检验和相关系数 r 的检验结果一致。另外，常数项 β_0 的置信度 95%的区间估计为（321.665，595.527），回归系数 β_1 的置信度 95%的区间估计为（0.682，0.714）。

（6）残差分析。仿照 2.5 节中对例 2.1 数据的残差分析，计算出残差 e_i，标准差 ZRE_i，学生化残差 SRE_i，再以自变量 x 为横轴，学生化残差 SRE_i 为纵轴作残差图。由残差图看到所有的点都在±3 内，没有异常值，但是残差有自相关趋势，

专业课辅导+真题资料+微信公众号：432应统书院

这一点将在 4.4 节自相关中继续讨论。由以上分析可认为本例的样本数据基本正常，理论模型的基本假定是合适的。

第七步，模型的应用。当所建模型通过所有检验之后，就可结合实际经济问题进行应用。最常见的应用之一就是因素分析。我们由回归方程可知，当城镇人均可支配收入增长 1 元时，平均约有 0.698 元用于消费，人均可支配收入的增长与人均消费支出的增长成正相关关系，这大致符合现阶段的实际情况。这个结果可为现阶段制定宏观调控政策提供量化依据，另外还可仿照 2.7 节做所需的预测。

回归分析方法的应用要特别注意定性分析与定量分析相结合。当现阶段的实际情况与建模时所用数据资料的背景发生较大变化时，不能仍机械地死套公式，这时就应对模型进行修改。修改包括重新收集数据，尽可能使用近期数据；还包括考虑是否要增加新的自变量，因为影响某种经济现象的因素可能发生了变化，可能还有一些重要的因素需要考虑等。这些问题都是本书后面几章要重点讨论的内容。

二、有关回归假设检验问题

对于一元线性回归方程显著性的检验，我们介绍的一个主要方法是 F 检验，即 $H_0:\beta_1=0$，$H_1:\beta_1\neq0$。那么接受 H_0 或拒绝 H_0 意味着什么？前面在作 F 检验时，假定 y 对 x 的回归形式为线性关系，而不是曲线关系。这时如果拒绝 H_0，就说明 x 与 y 之间有显著的线性关系，回归方程刻画了 x 与 y 的这种线性关系。然而，对于一个实际问题，变量 x 与 y 之间到底是一个什么样的关系，我们并不十分清楚。另外，样本数据是否存在异常值，是否存在周期性，往往从数据的表面并不能明显看出。运用普通最小二乘法估计模型的参数在模型满足一些基本假定时才有效，如果模型的基本假定明显出错，可能导致模型结论严重歪曲。

一般情况下，当 $H_0:\beta_1=0$ 被接受时，表明 y 的取值倾向不随 x 的值按线性关系变化。这种状况的原因可能是变量 y 与 x 之间的相关关系不显著，也可能虽然变量 y 与 x 之间的相关关系显著，但这种相关关系不是线性的而是非线性的。

当 $H_0:\beta_1=0$ 被拒绝时，如果没有其他信息，只能认为因变量 y 对自变量 x 的线性回归是有效的，但是并没有说明回归的有效程度，不能断言 y 与 x 之间就一定是线性相关关系，而不是曲线关系或其他的关系。这些问题还需要借助决定系数、散点图、残差图等工具作进一步分析。

为了说明上述问题，1973 年安斯库姆（Anscombe）构造了四组数据（参见参考文献 [2]），见表 2.5。用这四组数据得到的经验回归方程是相同的，都是 $\hat{y}=3.00+0.500x$，决定系数都是 $r^2=0.667$，相关系数 $r=0.816$。这四组数据所建的回归方程是相同的，决定系数 r^2、F 统计量也都相同，且均通过显著性检验，说明这四组数据 y 与 x 之间都有显著的线性相关关系。然而，变量 y 与 x 之间是否有相同的线性相关关系呢？由上述四组数据的散点图（见图 2.7）可以看到，变量 y 与 x 之间的关系大不相同。

专业课辅导+真题资料+微信公众号：432应统书院

表 2.5　四组数据

第一组		第二组		第三组		第四组	
x	y	x	y	x	y	x	y
4	4.26	4	3.10	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.44	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.10	14	8.84	19	12.5

由图 2.7 的（a）图可知，由直线作为 y 与 x 间关系的拟合是合适的，回归方程刻画出了变量 y 与 x 间的线性相关关系。

由图 2.7 的（b）图可知，变量 y 与 x 之间的相关关系应当是曲线关系，尽管回归方程也通过了显著性检验，但由直线方程去揭示它们的相关关系很不合适。如果用 y 对 x 作曲线回归，必然可以大幅提高决定系数 r^2，如果进一步作残差分析会发现残差点的分布不具有随机性。

图 2.7

专业课辅导+真题资料+微信公众号：432应统书院

由图 2.7 的（c）图可知，变量 y 与 x 之间存在着线性关系，但用直线 $\hat{y}=3.00+0.500x$ 去拟合这种关系不太理想。因为第三组数据中第 10 对数据（13，12.73）远离回归直线，可以认为是异常值。如果将它剔除，用其余 10 对数据重新计算得经验回归方程为 $\hat{y}=4.00+0.346x$，拟合效果非常好，决定系数接近 1，回归标准误差接近零。

由图 2.7 的（d）图可知，回归直线的斜率完全取决于（19，12.50）这一个点，这样得到的经验回归方程是很不可信的。实际上，自变量 x 只取了 8 和 19 这两个不同的值，因而不能断言 y 与 x 之间是何种关系。对这种情况，可以说数据收集得不理想，应该对自变量 x 在 8～19 这个区间上再收集一些不同的数据。

这个例子告诉我们，当拒绝假设 $H_0:\beta_1=0$ 时，我们说 y 与 x 之间存在线性相关关系，但是并不能完全肯定线性关系就是 y 与 x 之间关系的最好描述，很可能 y 与 x 之间更准确的关系应该是曲线关系，或者存在异常值等原因造成 y 与 x 之间虚假的线性关系。在实际应用中，不应局限于一种方法去分析判断。要得到确实可信的结果，应该 F 检验、决定系数、散点图、残差分析等一起使用，得到一致的结果时才可下定论。

三、回归系数的解释问题

对于回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$$

一般情况下，我们把回归系数 $\hat{\beta}_1$ 解释为：当自变量 x 增加或减少一个单位时，平均地说，y 增加或减少 $\hat{\beta}_1$ 个单位。不过这种说法并不总是正确的，在实际问题的分析时，应根据具体情况而定，在下一章中再详细讨论。

四、回归方程的预测问题

对于回归方程的应用，很重要的一个方面就是用回归方程预测未来。如果在预测时，自变量的取值在建模时样本数据 x 的取值范围之内，这种预测称为内插预测，内插预测的效果通常较好，预测误差小。如果自变量 x 的取值超出了建模时样本数据 x 的取值范围，这种预测称为外推预测，外推预测的效果可能不好。因为我们所建立的回归方程是直线方程，而理论上回归方程一般并不是严格的直线方程。如果用经验回归方程去预测，可能导致较大的误差。

在实际问题的研究中，如果从定性的角度认为回归方程为线性这一点有充分的理论根据，那么外推预测的效果不会太差。预测的结果肯定是有误差的，在实际应用时，要尽可能使误差较小。自变量 x 的取值距 $\bar{x}$ 明显过大时，预测效果一般不好。就像我们用 20 世纪 80 年代的人均国民收入与人均消费额数据建立模型作长期预测，预测 2015 年的人均消费额肯定误差很大。因为 2015 年

专业课辅导+真题资料+微信公众号：432应统书院

的经济情况与 20 世纪 80 年代肯定有很大差别。所以，用回归方程作长期预测一定要慎重。

□ 思考与练习

2.1　一元线性回归模型有哪些基本假定？

2.2　考虑过原点的线性回归模型

$$y_i=\beta_1 x_i+\varepsilon_i,\quad i=1,2,\cdots,n$$

误差 ε_1，ε_2，…，ε_n 仍满足基本假定。求 β_1 的最小二乘估计。

2.3　证明（2.27）式，$\sum_{i=1}^{n} e_i=0$，$\sum_{i=1}^{n} x_i e_i=0$。

2.4　回归方程 $E(y)=\beta_0+\beta_1 x$ 的参数 β_0，β_1 的最小二乘估计与最大似然估计在什么条件下等价？给出证明。

2.5　证明 $\hat{\beta}_0$ 是 β_0 的无偏估计。

2.6　证明（2.42）式 $\mathrm{var}(\hat{\beta}_0)=\left[\frac{1}{n}+\frac{(\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]\sigma^2$ 成立。

2.7　证明平方和分解式 SST=SSR+SSE。

2.8　验证三种检验的关系，即验证：

(1) $t=\frac{\hat{\beta}_1\sqrt{L_{xx}}}{\hat{\sigma}}=\frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$

(2) $F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}=\frac{\hat{\beta}_1^2\cdot L_{xx}}{\hat{\sigma}^2}=t^2$

2.9　验证（2.63）式：

$$\mathrm{var}(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}\right]\sigma^2$$

2.10　用第 9 题证明 $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 是 σ^2 的无偏估计。

2.11　验证决定系数 r^2 与 F 值之间的关系式

$$r^2=\frac{F}{F+n-2}$$

以上表达式说明 r^2 与 F 值是等价的，那么我们为什么要分别引入这两个统计量，而不是只使用其中的一个？

2.12　如果把自变量观测值都乘以 2，回归参数的最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 会发生什么变化？如果把自变量观测值都加上 2，回归参数的最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 会发生什么变化？

专业课辅导+真题资料+微信公众号：432应统书院

2.13 如果回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$ 相应的相关系数 r 很大，则用它预测时，预测误差一定较小。这一结论成立吗？请说明理由。

2.14 为了调查某广告对销售收入的影响，某商店记录了 5 个月的销售收入 y（万元）和广告费用 x（万元），数据见表 2.6：

表 2.6

月份	1	2	3	4	5
x	1	2	3	4	5
y	10	10	20	20	40

(1) 画散点图。

(2) x 与 y 之间是否大致呈线性关系？

(3) 用最小二乘估计求出回归方程。

(4) 求回归标准误差 $\hat{\sigma}$。

(5) 给出 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的置信度为 95%的区间估计。

(6) 计算 x 与 y 的决定系数。

(7) 对回归方程作方差分析。

(8) 作回归系数 β_1 的显著性检验。

(9) 作相关系数的显著性检验。

(10) 对回归方程作残差图并作相应的分析。

(11) 求当广告费用为 4.2 万元时，销售收入将达到多少，并给出置信度为 95%的置信区间。

2.15 一家保险公司十分关心其总公司营业部加班的程度，决定认真调查一下现状。经过 10 周时间，收集了每周加班时间的数据和签发的新保单数目，x 为每周签发的新保单数目，y 为每周加班时间（小时），数据见表 2.7：

表 2.7

周序号	1	2	3	4	5	6	7	8	9	10
x	825	215	1 070	550	480	920	1 350	325	670	1 215
y	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

(1) 画散点图。

(2) x 与 y 之间是否大致呈线性关系？

(3) 用最小二乘估计求出回归方程。

(4) 求回归标准误差 $\hat{\sigma}$。

(5) 给出 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的置信度为 95%的区间估计。

(6) 计算 x 与 y 的决定系数。

(7) 对回归方程作方差分析。

(8) 作回归系数 β_1 的显著性检验。

(9) 作相关系数的显著性检验。

专业课辅导+真题资料+微信公众号：432应统书院

(10) 对回归方程作残差图并作相应的分析。

(11) 该公司预计下一周签发新保单 $x_0=1\,000$ 张，需要的加班时间是多少？

(12) 给出 y_0 的置信度为 95%的精确预测区间和近似预测区间。

(13) 给出 $E(y_0)$ 的置信度为 95%的区间估计。

2.16　表 2.8 是 1985 年美国 50 个州和哥伦比亚特区公立学校中教师的人均年工资 y（美元）和对学生的人均经费投入 x（美元）。

(1) 绘制 y 对 x 的散点图，可以用直线回归描述两者之间的关系吗？

(2) 建立 y 对 x 的线性回归。

(3) 用线性回归的 Plots 功能绘制标准残差的直方图和正态概率图，检验误差项的正态性假设。

表 2.8

序号	y	x	序号	y	x	序号	y	x
1	19 583	3 346	18	20 816	3 059	35	19 538	2 642
2	20 263	3 114	19	18 095	2 967	36	20 460	3 124
3	20 325	3 554	20	20 939	3 285	37	21 419	2 752
4	26 800	4 542	21	22 644	3 914	38	25 160	3 429
5	29 470	4 669	22	24 624	4 517	39	22 482	3 947
6	26 610	4 888	23	27 186	4 349	40	20 969	2 509
7	30 678	5 710	24	33 990	5 020	41	27 224	5 440
8	27 170	5 536	25	23 382	3 594	42	25 892	4 042
9	25 853	4 168	26	20 627	2 821	43	22 644	3 402
10	24 500	3 547	27	22 795	3 366	44	24 640	2 829
11	24 274	3 159	28	21 570	2 920	45	22 341	2 297
12	27 170	3 621	29	22 080	2 980	46	25 610	2 932
13	30 168	3 782	30	22 250	3 731	47	26 015	3 705
14	26 525	4 247	31	20 940	2 853	48	25 788	4 123
15	27 360	3 982	32	21 800	2 533	49	2 9132	3 608
16	21 690	3 568	33	22 934	2 729	50	41 480	8 349
17	21 974	3 155	34	1 8443	2 305	51	25 845	3 766

专业课辅导+真题资料+微信公众号：432应统书院

第 3 章 Chapter 3 多元线性回归

在第 2 章我们介绍了被解释变量 y 只与一个解释变量 x 有关的线性回归问题，但在许多实际问题中，一元线性回归只不过是回归分析中的一种特例，它通常是我们对影响某种现象的许多因素进行简化考虑的结果。如某公司管理人员要预测来年该公司的销售额 y，研究认为影响销售额的因素不只是广告宣传费 x_1，还有个人可支配收入 x_2、价格 x_3、研发费用 x_4、各种投资 x_5、销售费用 x_6 等。这样因变量 y 就与多个自变量 x_1，x_2，x_3，x_4，x_5，x_6 有关。因此，我们就需要进一步讨论多元线性回归问题。

本章将重点介绍多元线性回归模型及其基本假设、回归模型未知参数的估计及其性质、回归方程及回归系数的显著性检验等。从这一章起将使用矩阵工具进行讨论。多元回归的计算量要比一元回归大得多，手工计算已不现实，需要使用计算机软件完成计算。

3.1 多元线性回归模型

一、多元线性回归模型的一般形式

设随机变量 y 与一般变量 x_1，x_2，…，x_p 的线性回归模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{3.1}$$

其中，β_0，β_1，…，β_p 是 $p+1$ 个未知参数，β_0 称为回归常数，β_1，…，β_p 称为回归系数。y 称为被解释变量（因变量），x_1，x_2，…，x_p 是 p 个可以精确测量并控制的一般变量，称为解释变量（自变量）。$p=1$ 时，（3.1）式即上一章的一元线性回归模型（2.1）式；$p\geqslant 2$ 时，我们就称（3.1）式为多元线性回归模型。ε 是随机

专业课辅导+真题资料+微信公众号：432应统书院

误差，与一元线性回归一样，对随机误差项我们常假定

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases} \tag{3.2}$$

称

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p \tag{3.3}$$

为理论回归方程。

对一个实际问题，如果我们获得 n 组观测数据（x_{i1}，x_{i2}，…，x_{ip}；y_i）($i=1$，2，…，n)，则线性回归模型（3.1）式可表示为

$$\begin{cases} y_1=\beta_0+\beta_1 x_{11}+\beta_2 x_{12}+\cdots+\beta_p x_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1 x_{21}+\beta_2 x_{22}+\cdots+\beta_p x_{2p}+\varepsilon_2 \\ \cdots\cdots \\ y_n=\beta_0+\beta_1 x_{n1}+\beta_2 x_{n2}+\cdots+\beta_p x_{np}+\varepsilon_n \end{cases} \tag{3.4}$$

写成矩阵形式为

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \tag{3.5}$$

其中

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{3.6}$$

$$\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\boldsymbol{X}$ 是一个 $n\times(p+1)$ 阶矩阵，称为回归设计矩阵或资料矩阵。在实验设计中，$\boldsymbol{X}$ 的元素是预先设定并可以控制的，人的主观因素可作用其中，因而称 $\boldsymbol{X}$ 为设计矩阵。

二、多元线性回归模型的基本假定

为了方便地进行模型的参数估计，对回归方程（3.4）式有如下一些基本假定：

（1）解释变量 x_1，x_2，…，x_p 是确定性变量，不是随机变量，且要求 $\mathrm{rank}(\boldsymbol{X})=p+1<n$。这里的 $\mathrm{rank}(\boldsymbol{X})=p+1<n$，表明设计矩阵 $\boldsymbol{X}$ 中的自变量列之间不相关，样本量的个数应大于解释变量的个数，$\boldsymbol{X}$ 是一满秩矩阵。

（2）随机误差项具有零均值和等方差，即

专业课辅导+真题资料+微信公众号：432应统书院

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j \\ 0, i\neq j\end{cases} \quad i,j=1,2,\cdots,n \end{cases} \tag{3.7}$$

这个假定常称为高斯-马尔柯夫条件。$E(\varepsilon_i)=0$，即假设观测值没有系统误差，随机误差项 ε_i 的平均值为零。随机误差项 ε_i 的协方差为零，表明随机误差项在不同的样本点之间是不相关的（在正态假定下即为独立的），不存在序列相关，并且有相同的精度。

（3）正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), \quad i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立} \end{cases} \tag{3.8}$$

对于多元线性回归的矩阵模型（3.5）式，这个条件便可表示为

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n) \tag{3.9}$$

由上述假定和多元正态分布的性质可知，随机向量 $\boldsymbol{y}$ 服从 n 维正态分布，回归模型（3.5）式的期望向量

$$E(\boldsymbol{y})=\boldsymbol{X\beta} \tag{3.10}$$

$$\operatorname{var}(\boldsymbol{y})=\sigma^2\boldsymbol{I}_n \tag{3.11}$$

因此

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n) \tag{3.12}$$

三、多元线性回归方程的解释

为了给多元线性回归方程及其回归系数一个解释，下面以 $p=2$ 的一个微观经济问题为例，给出回归方程的几何解释和回归系数的经济意义。在建立空调机销售量的预测模型时，用 y 表示空调机的销售量，x_1 表示空调机的价格，x_2 表示消费者可用于支配的收入，则可建立二元线性回归模型

$$\begin{cases} y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon \\ E(y)=\beta_0+\beta_1x_1+\beta_2x_2 \end{cases} \tag{3.13}$$

在（3.13）式中，假如 x_2 保持不变，为一常数，则有

$$\frac{\partial E(y)}{\partial x_1}=\beta_1 \tag{3.14}$$

即 β_1 可解释为在消费者收入 x_2 保持不变时，空调机价格 x_1 每增加一个单位，空调机销售量 y 的平均增加程度。一般来说，随着空调机价格的提高，销售量是减少的，因此 β_1 将是负的。

在（3.13）式中，假如 x_1 保持不变，为一常数，则有

专业课辅导+真题资料+微信公众号：432应统书院

$$\frac{\partial E(y)}{\partial x_2}=\beta_2 \tag{3.15}$$

即 β_2 可解释为在空调机价格 x_1 保持不变时，消费者收入 x_2 每增加一个单位，空调机销售量 y 的平均增加程度。一般来说，随着消费者收入的增加，空调机的需求量是增加的，因此 β_2 应该是正的。

对一般情况下含有 p 个自变量的多元线性回归而言，每个回归系数 β_i 表示在回归方程中其他自变量保持不变的情况下，自变量 x_i 每增加一个单位时因变量 y 的平均增加程度。因此也把多元线性回归的回归系数称为偏回归系数（partial regression coefficient），本书则仍简称为回归系数。

再用一个例子说明回归系数的含义。考虑国内生产总值（GDP）和三次产业增加值的关系，本章思考与练习中表 3.10 给出了历史数据。这个问题中 GDP$=x_1+x_2+x_3$是确定性的函数关系，可以看作误差项为 0 的特殊的回归关系。3 个回归系数都是 1，对 $\beta_2=1$ 的解释为，第二产业增加值 x_2每增加 1 亿元，GDP 也增加 1 亿元。现在做 GDP 对 x_2的一元线性回归，得回归方程 $\hat{y}=5\,289.9+1.855\,4x_2$，对这个方程回归系数的解释是，第二产业增加值每增加 1 亿元，GDP 增加 1.855 4 亿元。两个回归方程对同样的经济现象给出了不同的解释，问题出在什么地方？前面强调过，多元回归系数表示在回归方程中其他自变量保持不变的情况下，相应自变量每增加一个单位时因变量的平均增加程度。因此在用多元回归方程 GDP$=x_1+x_2+x_3$解释 $\beta_2=1$ 时，一定要强调是在 x_1和 x_3保持不变的情况下，x_2每增加 1 亿元，GDP 也增加 1 亿元。在用一元回归方程 $\hat{y}=5\,289.9+1.855\,4x_2$ 解释回归系数时，要强调的是在方程之外的有关变量也相应变化时，x_2每增加 1 亿元，GDP 增加 1.855 4 亿元。GDP 增加的 1.855 4 亿元中 x_2的直接贡献只有 1 亿元，回归方程外的 x_1和 x_3的贡献是 0.855 4 亿元。

还有一个问题，为什么回归方程外的 x_1和 x_3的贡献是 0.855 4 亿元，而不是 2 亿元？仔细观察表 3.10 的数据你会发现，x_2的增加幅度远大于 x_1和 x_3的增加幅度，假如 x_2增加 1 亿元，x_1和 x_3相应的增加幅度都达不到 1 亿元。

需要说明的是，2005 年我国进行了第一次全国经济普查，国家统计局根据经济普查资料对 2004 年 GDP 进行了重新核算，并对 1993 年以来的 GDP 历史数据进行了修订。根据经济普查资料计算出我国 2004 年 GDP 现价总量为 159 878 亿元，与 2004 年全国 GDP 年快报核算数据相比增加了 23 002 亿元，提高了 16.8%。其中第一产业数据变化不大，重新核算后增加了 188 亿元；第二产业增加值增加了 1 517 亿元，提高了 2.1%；第三产业增加值增加了 21 297 亿元，提高了 48.7%。在 GDP 总量多出的 2.3 万亿元中，有 93%是第三产业带来的。

回归方程（3.13）式的图形，不像一元线性回归那样是一条直线，而是一个回归平面。而对一般情况下的回归方程（3.3）式，当 $p>2$ 时，回归方程是一个超平面，无法用几何图形表示。

专业课辅导+真题资料+微信公众号：432应统书院

3.2 回归参数的估计

一、回归参数的普通最小二乘估计

多元线性回归方程未知参数β_0，β_1，…，β_p 的估计与一元线性回归方程的参数估计原理一样，仍然可以采用最小二乘估计。对于（3.5）式表示的回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，所谓最小二乘法，就是寻找参数$\beta_0$，$\beta_1$，$\beta_2$，…，$\beta_p$ 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，使离差平方和 $Q(\beta_0,\beta_1,\beta_2,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$ 达到极小，即寻找 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 满足

$$
\begin{aligned}
Q(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p)&=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})^2\\
&=\min_{\beta_0,\beta_1,\beta_2,\cdots,\beta_p}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2
\end{aligned}
\tag{3.16}
$$

依照（3.16）式求出的 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 就称为回归参数β_0，β_1，β_2，…，β_p 的最小二乘估计。

从（3.16）式中求出 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 是一个求极值问题。由于 Q 是关于β_0，β_1，β_2，…，β_p 的非负二次函数，因而它的最小值总是存在的。根据微积分中求极值的原理，$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 应满足下列方程组

$$
\begin{cases}
\left.\dfrac{\partial Q}{\partial\beta_0}\right|_{\beta_0=\hat{\beta}_0}=-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})=0\\
\left.\dfrac{\partial Q}{\partial\beta_1}\right|_{\beta_1=\hat{\beta}_1}=-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{i1}=0\\
\left.\dfrac{\partial Q}{\partial\beta_2}\right|_{\beta_2=\hat{\beta}_2}=-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{i2}=0\\
\cdots\cdots\\
\left.\dfrac{\partial Q}{\partial\beta_p}\right|_{\beta_p=\hat{\beta}_p}=-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{ip}=0
\end{cases}
\tag{3.17}
$$

以上方程组经整理后，得出用矩阵形式表示的正规方程组

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

移项得

专业课辅导+真题资料+微信公众号：432应统书院

$$X'X\hat{\boldsymbol{\beta}}=X'y$$

当 $(X'X)^{-1}$ 存在时，即得回归参数的最小二乘估计为

$$\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'y \tag{3.18}$$

称

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p \tag{3.19}$$

为经验回归方程。

二、回归值与残差

在求出回归参数的最小二乘估计后，可以用经验回归方程（3.19）式计算因变量的回归值与残差。称

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip} \tag{3.20}$$

为观测值 $y_i(i=1,2,\cdots,n)$ 的回归拟合值，简称回归值或拟合值。相应地，称向量 $\hat{y}=X\hat{\boldsymbol{\beta}}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n)'$ 为因变量向量 $y=(y_1,y_2,\cdots,y_n)'$ 的回归值。由 $\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'y$ 可得

$$\hat{y}=X\hat{\boldsymbol{\beta}}=X(X'X)^{-1}X'y \tag{3.21}$$

由（3.21）式看到，矩阵 $X(X'X)^{-1}X'$ 的作用是把因变量向量 y 变为拟合值向量 $\hat{y}$，从形式上看是给 y 戴上了一顶帽子“^”，因而形象地称矩阵 $X(X'X)^{-1}X'$ 为帽子矩阵，记为 H，于是 $\hat{y}=Hy$。显然帽子矩阵 $H=X(X'X)^{-1}X'$ 是 n 阶对称矩阵，同时还是幂等矩阵，即 $H=H^2$。帽子矩阵 H 也是一个投影阵，从代数学的观点看，$\hat{y}$ 是 y 在自变量 X 生成的空间上的投影，这个投影过程就是把 y 左乘矩阵 H，因此称 H 为投影阵。帽子矩阵 $H=X(X'X)^{-1}X'$ 的主对角线元素记为 h_{ii}，可以证明，帽子矩阵 H 的迹为

$$\mathrm{tr}(H)=\sum_{i=1}^{n}h_{ii}=p+1 \tag{3.22}$$

（3.22）式的证明只需根据迹的性质 $\mathrm{tr}(AB)=\mathrm{tr}(BA)$，因而

$$\begin{aligned}\mathrm{tr}(H)&=\mathrm{tr}(X(X'X)^{-1}X')\\&=\mathrm{tr}((X'X)^{-1}X'X)\\&=\mathrm{tr}(I_{p+1})=p+1\end{aligned}$$

称

$$e_i=y_i-\hat{y}_i \tag{3.23}$$

为 $y_i(i=1,2,\cdots,n)$ 的残差。称 $e=(e_1,e_2,\cdots,e_n)'=y-\hat{y}$ 为回归残差向量。将 $\hat{y}=Hy$ 代入得，$e=y-Hy=(I-H)y$。记 $\mathrm{cov}(e,e)=(\mathrm{cov}(e_i,e_j))_{n\times n}$ 为残差向

专业课辅导+真题资料+微信公众号：432应统书院

量 $\boldsymbol{e}$ 的协方差阵，或称为方差阵，记为 $D(\boldsymbol{e})$。因而

$$\begin{aligned}D(\boldsymbol{e})&=\operatorname{cov}(\boldsymbol{e},\boldsymbol{e})\\&=\operatorname{cov}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y},(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y})\\&=(\boldsymbol{I}-\boldsymbol{H})\operatorname{cov}(\boldsymbol{y},\boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})'\\&=\sigma^2(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{I}_n(\boldsymbol{I}-\boldsymbol{H})'\\&=\sigma^2(\boldsymbol{I}-\boldsymbol{H})\end{aligned}$$

于是有

$$D(e_i)=(1-h_{ii})\sigma^2,\quad i=1,2,\cdots,n \tag{3.24}$$

根据（3.17）式可知，残差满足关系式

$$\begin{cases}\sum e_i=0\\\sum e_ix_{i1}=0\\\cdots\cdots\\\sum e_ix_{ip}=0\end{cases} \tag{3.25}$$

即残差的平均值为 0，残差对每个自变量的加权平均为 0。（3.25）式可以用矩阵表示为 $\boldsymbol{X}'\boldsymbol{e}=\boldsymbol{0}$ 。

误差项方差 σ^2 的无偏估计为

$$\begin{aligned}\hat{\sigma}^2&=\frac{1}{n-p-1}\mathrm{SSE}=\frac{1}{n-p-1}(\boldsymbol{e}'\boldsymbol{e})\\&=\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2\end{aligned} \tag{3.26}$$

（3.26）式的证明只需注意 $E(\sum_{i=1}^{n}e_i^2)=\sum_{i=1}^{n}D(e_i)$，然后再用（3.24）式和（3.22）式即可。

前边在由正规方程组求 $\hat{\boldsymbol{\beta}}$ 时，要求 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 必须存在，即 $\boldsymbol{X}'\boldsymbol{X}$ 是一非奇异矩阵

$$|\boldsymbol{X}'\boldsymbol{X}|\neq 0$$

由线性代数可知，$\boldsymbol{X}'\boldsymbol{X}$ 为 $p+1$ 阶满秩矩阵

$$\operatorname{rank}(\boldsymbol{X}'\boldsymbol{X})=p+1$$

必须有

$$\operatorname{rank}(\boldsymbol{X})\geqslant p+1$$

而 $\boldsymbol{X}$ 为 $n\times(p+1)$ 阶矩阵，于是应有

专业课辅导+真题资料+微信公众号：432应统书院

$$n \geqslant p+1$$

这是一个重要的结论，我们在多元线性回归模型的基本假定中用过它，这里就更清楚这个假定的重要意义。结论说明，要想用普通最小二乘法估计多元线性回归模型的未知参数，样本量必须不少于模型中参数的个数。在后面关于回归方程的假设检验中也少不了这一假设，否则检验无任何意义。

三、回归参数的最大似然估计

多元线性回归参数的最大似然估计与一元线性回归参数的最大似然估计的思想一致。对于（3.5）式所表示的模型

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon}\sim N(0,\sigma^2\boldsymbol{I}_n)$$

即 $\boldsymbol{\varepsilon}$ 服从多变量正态分布，那么 $\boldsymbol{y}$ 的概率分布为

$$\boldsymbol{y}\sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$$

这时，似然函数为

$$L=(2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right) \tag{3.27}$$

其中的未知参数是 $\boldsymbol{\beta}$ 和 σ^2，最大似然估计就是选取使似然函数 L 达到最大的 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\sigma}^2$。要使 L 达到最大，对（3.27）式两边同时取自然对数，得

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{3.28}$$

在（3.28）式中，仅在最后一项中含有 $\boldsymbol{\beta}$，显然使（3.28）式达到最大，等价于使

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

达到最小，这又完全与普通最小二乘估计一样。故在正态假定下，回归参数 $\boldsymbol{\beta}$ 的最大似然估计与普通最小二乘估计完全相同，即

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

误差项方差 σ^2 的最大似然估计为

$$\hat{\sigma}^2=\frac{1}{n}\mathrm{SSE}=\frac{1}{n}(\boldsymbol{e}'\boldsymbol{e}) \tag{3.29}$$

这是 σ^2 的有偏估计，但它满足一致性。在大样本的情况下，这是 σ^2 的渐近无偏估计。

专业课辅导+真题资料+微信公众号：432应统书院

四、实例分析

例 3.1

现实生活中，影响一个地区居民消费的因素有很多，例如，一个地区的人均生产总值、收入水平、消费价格指数、生活必需品的花费等。本例选取 9 个解释变量研究城镇居民家庭平均每人全年的消费性支出 y，解释变量为：x_1 居民的食品花费，x_2 居民的服装花费，x_3 居民的居住花费，x_4 居民的医疗花费，x_5 居民的教育花费，x_6 地区的职工平均工资，x_7 地区的人均 GDP，x_8 地区的消费价格指数，x_9 地区的失业率。本例选取 2009 年《中国统计年鉴》我国 30 个省、市、自治区（西藏地区失业率数据缺失，因此从样本中剔除西藏）2008 年的数据，以居民的消费性支出（元）为因变量，以如上 9 个变量为自变量作多元线性回归。数据见表 3.1，其中，自变量 $x_1 \sim x_7$ 单位为元，x_9 单位为%。

表 3.1

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
北　京	5 562	1 102	1 286	1 563	809.3	55 844	6 060	105.1	1.82	16 460
天　津	5 005	812.5	1 528	1 221	715.2	39 990	3 111	105.4	3.6	13 423
河　北	3 155	754.8	1 097	808.9	386.6	24 276	8 478	106.2	3.96	9 087
山　西	2 975	820.8	1 251	769.8	570.8	25 489	3 571	107.2	3.29	8 807
内蒙古	3 553	1 178	1 028	869.7	568.4	25 949	3 041	105.7	4.1	10 829
辽　宁	4 378	722	1 271	913.1	670.1	27 179	6 672	104.6	3.9	11 232
吉　林	3 307	871.4	1 285	914.5	576.2	23 294	3 122	105.1	3.98	9 729
黑龙江	3 128	830.5	941.3	864.9	551.7	21 764	4 751	105.6	4.23	8 623
上　海	7 109	1 135	1 646	755.3	1 166	52 122	8 073	105.8	4.2	19 398
江　苏	4 545	856.6	1 042	794.6	751	31 297	15 004	105.4	3.25	11 978
浙　江	5 523	1 184	1 334	933.1	1 179	33 622	11 649	105	3.49	15 158
安　徽	3 905	679.3	988.1	633.9	633.5	25 703	4 759	106.2	3.92	9 524
福　建	5 079	862.2	1 300	540.6	686.4	25 555	5 763	104.6	3.86	12 501
江　西	3 633	739.2	851.2	484	388.5	20 597	3 457	106.1	3.42	8 717
山　东	3 699	996.9	1 247	799.8	580.1	26 234	15 022	105.3	3.7	11 007
河　南	3 080	842	963.6	790.9	464.4	24 438	8 554	107	3.4	8 837
湖　北	3 996	816.2	914.3	675.3	571	22 384	5 633	106.3	4.2	9 478
湖　南	3 970	787.9	960.8	791	543.5	24 146	5 642	106	4.2	9 946
广　东	5 867	727	1 748	836.4	720.6	33 282	18 865	105.6	2.56	15 528
广　西	4 083	557.2	891.3	529.4	483.6	24 798	3 434	107.8	3.75	9 627
海　南	4 227	379.9	1 106	536.4	459.7	21 767	798.9	106.9	3.72	9 408
重　庆	4 418	947.7	1 097	878.3	536.4	26 640	2 693	105.6	3.96	11 147
四　川	4 255	770.8	819.3	564.9	422.1	24 725	6 380	105.1	4.57	9 679
贵　州	3 598	590.2	836.5	471.4	436.2	23 979	1 678	107.6	3.98	8 349
云　南	4 272	728.6	739.2	606.9	294.3	23 305	3 082	105.7	4.21	9 077
陕　西	3 586	753.9	1 008	862.7	697.3	25 478	3 176	106.4	3.91	9 772
甘　肃	3 184	715.4	846.3	654.8	428.4	23 632	1 688	108.2	3.23	8 309
青　海	3 316	680.9	802.7	610	389	30 101	466.1	110.1	3.8	8 193
宁　夏	3 353	868.5	1 069	816.9	465.1	30 050	537.2	108.5	4.35	9 558
新　疆	3 236	877.7	781.9	643.5	417.1	24 686	2 209	108.1	3.7	8 669

专业课辅导+真题资料+微信公众号：432应统书院

用 SPSS 18.0 计算出的回归系数见输出结果 3.1。

输出结果 3.1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2869.723	7962.399		-.360	.722
	x1	1.467	.150	.520	9.772	.000
	x2	2.577	.421	.167	6.115	.000
	x3	2.003	.321	.187	6.233	.000
	x4	.005	.503	.000	.011	.992
	x5	.894	.473	.066	1.888	.074
	x6	.066	.019	.198	3.509	.002
	x7	.028	.015	.045	1.828	.082
	x8	12.776	70.233	.006	.182	.857
	x9	-161.486	121.816	-.032	-1.326	.200

a. Dependent Variable: y.

因而 y 对 9 个自变量的线性回归方程为

$$\hat{y}=-2\,869.723+1.467x_1+2.577x_2+2.003x_3+0.005x_4+0.894x_5+0.066x_6+0.028x_7+12.776x_8-161.486x_9$$

从回归方程中可以看到，x_1，…，x_8 对居民的消费性支出起正影响，x_9 对居民的消费性支出起负影响，这与定性分析的结果一致。

从回归方程中可以看出，对城镇居民消费性支出有显著影响的，即通过回归系数显著性检验的是居民在食品、服装、居住上的花费，以及平均工资，而且回归系数的符号都为正。很显然，居民在食品、服装、居住上的花费越多，其消费性支出越多。

根据凯恩斯的消费理论，随着收入的增加，消费也会增加。在回归方程中的体现就是，平均工资的系数为正，工资越多意味着收入越多，从而消费就会增加。地区的人均 GDP 也会引起居民收入的增加，从而导致居民消费的增加。消费价格指数越高，一定程度上虽然会影响居民的消费意愿，但是由于居民对食品、服装等必需品的消费具有刚性，因此消费性支出也会增加。但当一个地区失业率高时，居民的收入会减少，同时失业的压力会影响居民对未来收入的预期和消费信心，因此消费性支出会随之减少。

但是，这一回归方程并不理想，所选自变量数目过多，部分回归系数的显著性检验不能通过，这里只是作为多元线性回归的参数估计的一个例子，后面我们将要

专业课辅导+真题资料+微信公众号：432应统书院

进一步完善这一问题模型的建立。

3.3 参数估计量的性质

性质 1 $\hat{\boldsymbol{\beta}}$是随机向量 $\boldsymbol{y}$ 的一个线性变换。

在多元线性回归中，无论应用普通最小二乘估计还是最大似然估计，得到回归系数向量 $\boldsymbol{\beta}$ 的估计量为

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.30}$$

根据回归模型假设知，$\boldsymbol{X}$ 是固定的设计矩阵，因此，$\hat{\boldsymbol{\beta}}$是 $\boldsymbol{y}$ 的一个线性变换。

性质 2 $\hat{\boldsymbol{\beta}}$是 $\boldsymbol{\beta}$ 的无偏估计。

证明：

$$\begin{aligned}E(\hat{\boldsymbol{\beta}})&=E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y})\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\\&=\boldsymbol{\beta}\end{aligned}$$

这一性质与一元线性回归 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 无偏的性质相同。

性质 3 $D(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (3.31)

证明：

$$\begin{aligned}D(\hat{\boldsymbol{\beta}})&=\mathrm{cov}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})\\&=\mathrm{cov}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{cov}(\boldsymbol{y},\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')'\\&=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\\&=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\\&=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned}$$

当 $p=1$ 时即一元线性回归的情况，此时

$$\boldsymbol{X}'\boldsymbol{X}=\begin{bmatrix}n & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2\end{bmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}=\frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}\begin{bmatrix}\sum_{i=1}^{n}x_i^2 & -\sum_{i=1}^{n}x_i\\ -\sum_{i=1}^{n}x_i & n\end{bmatrix}$$

专业课辅导+真题资料+微信公众号：432应统书院

$$
= \frac{1}{nL_{xx}} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{1}{nL_{xx}} \sum_{i=1}^{n} x_i^2 & -\frac{\bar{x}}{L_{xx}} \\ -\frac{\bar{x}}{L_{xx}} & \frac{1}{L_{xx}} \end{bmatrix} \tag{3.32}
$$

再由

$$
D(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathrm{var}(\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{var}(\hat{\beta}_1) \end{pmatrix} \tag{3.33}
$$

即可得（2.41）、（2.42）、（2.45）式。

$\hat{\boldsymbol{\beta}}$ 的方差阵 $D(\hat{\boldsymbol{\beta}})$ 也记为 $\mathrm{cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})$，因而也称做 $\hat{\boldsymbol{\beta}}$ 的协方差阵，它是回归系数 $\hat{\beta}_1$ 方差的推广，反映了估计量 $\hat{\boldsymbol{\beta}}$ 的波动大小。由于 $D(\hat{\boldsymbol{\beta}})$ 是 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 乘上 σ^2，而 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 一般为非对角阵，所以 $\hat{\boldsymbol{\beta}}$ 的各分量 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 之间有一定的联系，根据 $D(\hat{\boldsymbol{\beta}})$ 可以分析 $\hat{\boldsymbol{\beta}}$ 各分量的波动以及各分量之间的相关程度。

由此性质还可看出，回归系数向量 $\hat{\boldsymbol{\beta}}$ 的稳定状况不仅与随机误差项的方差 σ^2 有关，还与设计矩阵 $\boldsymbol{X}$ 有关，这种情况与一元线性回归中一样，即要想使估计量的方差小，采集样本数据时就不能太集中。这对设计矩阵的构造有一定的指导意义。

为了分析 $\hat{\boldsymbol{\beta}}$ 各分量之间的相关程度，更方便的工具是采用 $\hat{\boldsymbol{\beta}}$ 的相关阵。以一元线性回归为例，$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ 的相关阵为

$$
R(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1 & \frac{\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\mathrm{var}(\hat{\beta}_0)}\sqrt{\mathrm{var}(\hat{\beta}_1)}} \\ \frac{\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\mathrm{var}(\hat{\beta}_0)}\sqrt{\mathrm{var}(\hat{\beta}_1)}} & 1 \end{bmatrix}
$$

$$
= \begin{bmatrix} 1 & -\frac{\bar{x}}{\sqrt{\frac{1}{n}\sum x_i^2}} \\ -\frac{\bar{x}}{\sqrt{\frac{1}{n}\sum x_i^2}} & 1 \end{bmatrix} \tag{3.34}
$$

利用 SPSS 软件可以方便地计算出 $\hat{\boldsymbol{\beta}}$ 的协方差阵与相关阵，$\hat{\boldsymbol{\beta}}$ 的协方差阵与相关阵不属于默认输出值，在 Linear Regression 对话框中点击 Statistics→Covariance matrix 即可。根据例 3.1 的数据计算出的 $\hat{\boldsymbol{\beta}}$ 的相关阵与协方差阵如表 3.2、表 3.3 所示。

专业课辅导+真题资料+微信公众号：432应统书院

表 3.2 相关系数阵

	$\hat{\beta}_9$	$\hat{\beta}_8$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_7$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
$\hat{\beta}_9$	1.000	0.202	−0.262	0.019	0.456	−0.034	0.329	−0.074	0.119
$\hat{\beta}_8$	0.202	1.000	0.251	0.677	0.244	−0.031	0.666	−0.092	−0.681
$\hat{\beta}_2$	−0.262	0.251	1.000	0.327	−0.211	0.202	−0.062	−0.427	−0.339
$\hat{\beta}_1$	0.019	0.677	0.327	1.000	−0.027	−0.203	0.676	−0.334	−0.819
$\hat{\beta}_7$	0.456	0.244	−0.211	−0.027	1.000	−0.259	0.214	−0.086	0.106
$\hat{\beta}_3$	−0.034	−0.031	0.202	−0.203	−0.259	1.000	−0.285	−0.321	0.057
$\hat{\beta}_4$	0.329	0.666	−0.062	0.676	0.214	−0.285	1.000	−0.059	−0.703
$\hat{\beta}_5$	−0.074	−0.092	−0.427	−0.334	−0.086	−0.321	−0.059	1.000	0.031
$\hat{\beta}_6$	0.119	−0.681	−0.339	−0.819	0.106	0.057	−0.703	0.031	1.000

表 3.3 协方差阵

	$\hat{\beta}_9$	$\hat{\beta}_8$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_7$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
$\hat{\beta}_9$	14 839.023	1 730.889	−13.459	0.355	0.838	−1.315	20.139	−4.260	0.273
$\hat{\beta}_8$	1 730.889	4 932.699	7.417	7.142	0.259	−0.704	23.517	−3.064	−0.904
$\hat{\beta}_2$	−13.459	7.417	0.178	0.021	−0.001	0.027	−0.013	−0.085	−0.003
$\hat{\beta}_1$	0.355	7.142	0.021	0.023	0.000	−0.010	0.051	−0.024	−0.002
$\hat{\beta}_7$	0.838	0.259	−0.001	0.000	0.000	−0.001	0.002	−0.001	0.000
$\hat{\beta}_3$	−1.315	−0.704	0.027	−0.010	−0.001	0.103	−0.046	−0.049	0.000
$\hat{\beta}_4$	20.139	23.517	−0.013	0.051	0.002	−0.046	0.253	−0.014	−0.007
$\hat{\beta}_5$	−4.260	−3.064	−0.085	−0.024	−0.001	−0.049	−0.014	0.224	0.000
$\hat{\beta}_6$	0.273	−0.904	−0.003	−0.002	0.000	0.000	−0.007	0.000	0.000

性质 4　高斯-马尔柯夫（G-M）定理

在实际应用中，我们关心的一个主要问题是预测。预测函数

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{10}+\hat{\beta}_2x_{20}+\cdots+\hat{\beta}_px_{p0} \tag{3.35}$$

是 $\hat{\boldsymbol{\beta}}$ 的线性函数，因而我们希望 $\hat{\boldsymbol{\beta}}$ 的线性函数的波动越小越好。设 $\boldsymbol{c}$ 为任一 $p+1$ 维常数向量，我们希望回归系数向量 $\boldsymbol{\beta}$ 的估计值 $\hat{\boldsymbol{\beta}}$ 具有如下性质：

（1）$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计。

（2）$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 的方差要小。

下面的一个重要性质告诉我们普通最小二乘估计 $\hat{\boldsymbol{\beta}}$ 正好满足上述条件。

高斯-马尔柯夫定理：在假定 $E(\boldsymbol{y})=\boldsymbol{X\beta}$，$D(\boldsymbol{y})=\sigma^2\boldsymbol{I}_n$ 时，$\boldsymbol{\beta}$ 的任一线性函数 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计（BLUE）为 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$，其中，$\boldsymbol{c}$ 是任一 $p+1$ 维常数向量，$\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计。

证明参见参考文献［5］。

此定理说明了用普通最小二乘估计得到的 $\hat{\boldsymbol{\beta}}$ 是理想的估计量。关于这条性质，请读者注意以下四点：

第一，取常数向量 $\boldsymbol{c}$ 的第 j（$j=0$，1，…，p）个分量为 1，其余分量为 0，这

专业课辅导+真题资料+微信公众号：432应统书院

时 G-M 定理表明最小二乘估计 $\hat{\beta}_j$ 是 β_j 的最小方差线性无偏估计。

第二，可能存在 y_1，y_2，…，y_n 的非线性函数，作为 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 的方差更小。

第三，可能存在 $\boldsymbol{c}'\boldsymbol{\beta}$ 的有偏估计，在某种意义（例如均方误差最小）上比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 更好。

第四，在正态假定下，$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差无偏估计。也就是说，既不可能存在 y_1，y_2，…，y_n 的非线性函数，也不可能存在 y_1，y_2，…，y_n 的其他线性函数，作为 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计，比最小二乘估计 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 的方差更小。

性质 5　$\mathrm{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e})=\boldsymbol{0}$。

证明参见参考文献［5］。

此性质说明与 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 不相关，在正态假定下，$\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 不相关等价于 $\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{e}$ 独立，从而 $\hat{\boldsymbol{\beta}}$ 与 $\mathrm{SSE}=\boldsymbol{e}'\boldsymbol{e}$ 独立。

性质 6　当 $\boldsymbol{y}\sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ 时，则

（1）$\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$。

（2）$\mathrm{SSE}/\sigma^2\sim\chi^2(n-p-1)$。

由前 3 个性质易证明性质 1，性质 2 的证明参见参考文献［7］。

性质 5 和性质 6 在构造 t 统计量和 F 统计量时有用。这两条性质对一元线性回归当然也是成立的，只是为了教材的系统性，在第 2 章一元线性回归中没有提出。

3.4　回归方程的显著性检验

在实际问题的研究中，事先并不能断定随机变量 y 与变量 x_1，x_2，…，x_p 之间确有线性关系，在进行回归参数的估计前，我们用多元线性回归方程去拟合随机变量 y 与变量 x_1，x_2，…，x_p 之间的关系，只是根据一些定性分析所作的一种假设。因此，当求出线性回归方程后，还需对回归方程进行显著性检验。多元线性回归方程的显著性检验与一元线性回归方程的显著性检验既有相同之处，也有不同之处。

下面介绍两种统计检验方法，一是回归方程显著性的 F 检验，另一个是回归系数显著性的 t 检验。同时介绍衡量回归拟合程度的拟合优度检验。

一、F 检验

对多元线性回归方程的显著性检验就是要看自变量 x_1，x_2，…，x_p 从整体上对随机变量 y 是否有明显的影响。为此提出原假设

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0$$

专业课辅导+真题资料+微信公众号：432应统书院

如果 H_0 被接受，则表明随机变量 y 与 x_1，x_2，…，x_p 之间的关系由线性回归模型表示不合适。类似一元线性回归检验，为了建立对 H_0 进行检验的 F 统计量，仍然利用总离差平方和的分解式，即

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

简写为

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}$$

此分解式的证明只需利用（3.25）式即可。构造 F 检验统计量如下

$$F=\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)} \tag{3.36}$$

在正态假设下，当原假设 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$ 成立时，F 服从自由度为（p，$n-p-1$）的 F 分布。于是，可以利用 F 统计量对回归方程的总体显著性进行检验。对于给定的数据，$i=1$，2，…，n，计算出 SSR 和 SSE，进而得到 F 的值，其计算过程列在表 3.4 的方差分析表中，再由给定的显著性水平 α 查 F 分布表，得临界值 $F_\alpha(p,n-p-1)$。

表 3.4　　方差分析表

方差来源	自由度	平方和	均方	F 值	P 值
回归	p	SSR	SSR/p	$\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)}$	$P(F>F$ 值$)=P$ 值
残差	$n-p-1$	SSE	SSE/$(n-p-1)$		
总和	$n-1$	SST			

当 $F>F_\alpha(p,n-p-1)$ 时，拒绝原假设 H_0，认为在显著性水平 α 下，y 与 x_1，x_2，…，x_p 有显著的线性关系，即回归方程是显著的。更通俗一些说，就是接受“自变量全体对因变量 y 产生线性影响”这一结论犯错误的概率不超过 α；反之，当 $F\leqslant F_\alpha(p,n-p-1)$ 时，则认为回归方程不显著。

与一元线性回归一样，也可以根据 P 值作检验。当 P 值$<\alpha$ 时，拒绝原假设 H_0；当 P 值$\geqslant\alpha$ 时，接受原假设 H_0。

对例 3.1 的数据，用 SPSS 软件计算出的方差分析表见输出结果 3.2。

输出结果 3.2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.162E8	9	2.402E7	368.330	.000[a]
	Residual	1304187.292	20	65209.365		
	Total	2.175E8	29			

a. Predictors: (Constant), x9, x8, x2,x1, x7, x3, x4, x5, x6.

b. Dependent Variable: y.

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 3.2 中，Sig 即显著性 P 值，由 P 值$=3.257\mathrm{E}-20\approx0.000$，可知此回归方程高度显著，即做出 9 个自变量整体对因变量 y 产生显著线性影响的判断所犯错误的概率仅为 $3.257\mathrm{E}-20\approx0.000$。

二、t 检验

在多元线性回归中，回归方程显著并不意味着每个自变量对 y 的影响都显著，我们总想从回归方程中剔除那些次要的、可有可无的变量，重新建立更为简单的回归方程，所以需要对每个自变量进行显著性检验。

显然，如果某个自变量 x_j 对 y 的作用不显著，那么在回归模型中，它的系数 β_j 就取值为零。因此，检验变量 x_j 是否显著，等价于检验假设

$$H_{0j}:\beta_j=0,\quad j=1,2,\cdots,p$$

如果接受原假设 H_{0j}，则 x_j 不显著；如果拒绝原假设 H_{0j}，则 x_j 是显著的。

由 3.3 中性质 6 知

$$\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{3.37}$$

记

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}=(c_{ij}),\quad i,j=0,1,2,\cdots,p \tag{3.38}$$

于是有

$$E(\hat{\beta}_j)=\beta_j,\mathrm{var}(\hat{\beta}_j)=c_{jj}\sigma^2$$
$$\hat{\beta}_j\sim N(\beta_j,c_{jj}\sigma^2),\quad j=0,1,2,\cdots,p \tag{3.39}$$

据此可以构造 t 统计量

$$t_j=\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \tag{3.40}$$

其中

$$\hat{\sigma}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3.41}$$

是回归标准差。

当原假设 $H_{0j}:\beta_j=0$ 成立时，(3.40) 式构造的 t_j 统计量服从自由度为$n-p-1$的 t 分布。给定显著性水平 α，查出双侧检验的临界值 $t_{\alpha/2}$。当 $|t_j|\geqslant t_{\alpha/2}$时，拒绝原假设 $H_{0j}:\beta_j=0$，认为 β_j 显著不为零，自变量 x_j 对因变量 y 的线性效果显著；当 $|t_j|<t_{\alpha/2}$时，接受原假设 $H_{0j}:\beta_j=0$，认为 β_j 为零，自变量 x_j 对因变量 y 的线性效果不显著。

对于例 3.1 的城镇居民消费性支出的例子，由 F 检验知道回归方程的整体是显

专业课辅导+真题资料+微信公众号：432应统书院

著的，即 9 个自变量作为一个整体对因变量 y 有十分显著的影响。那么，每一个自变量 $x_j(j=1, 2, \cdots, 9)$ 是否对 y 有显著影响呢？

利用 SPSS 软件计算的关于 β_j 的 t 统计量 $t_j(j=1, 2, \cdots, 9)$ 及其相应的 P 值见输出结果 3.1。我们可以发现在显著性水平 $\alpha=0.05$ 时只有 x_1，x_2，x_3，x_6 通过了显著性检验。这个例子说明，尽管回归方程的显著性检验高度显著，但也会出现有某些自变量 x_j（甚至每个 x_j）对 y 无显著影响的情况。

由于某些自变量不显著，因而在多元回归中并不是包含在回归方程中的自变量越多越好，这个问题将在第 5 章中做详细讨论。在此仅简单介绍一种剔除多余变量的方法——后退法。当有多个自变量对因变量 y 无显著影响时，由于自变量之间的交互作用，不能一次剔除掉所有不显著的变量。原则上每次只剔除一个变量，先剔除其中 $|t|$ 值最小的（或 P 值最大的）一个变量，然后再对求得的新的回归方程进行检验，有不显著的变量再剔除，直到保留的变量都对 y 有显著影响为止。也可以根据对问题的定性分析选择 t 值较小的变量先剔除。本例中 P 值最大的 $p_4=0.992$。从定性分析看，居民在医疗上的花费对居民的消费性支出的影响应该很小。首先剔除 x_4，用 y 与其余 8 个自变量作回归，计算结果见输出结果 3.3。

输出结果 3.3

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2811.502	5681.148		-.495	.626
	x1	1.466	.108	.520	13.581	.000
	x2	2.577	.410	.167	6.279	.000
	x3	2.004	.301	.187	6.667	.000
	x5	.894	.461	.066	1.939	.066
	x6	.066	.013	.198	5.067	.000
	x7	.028	.014	.045	1.915	.069
	x8	12.275	51.120	.006	.240	.813
	x9	-161.916	112.266	-.032	-1.442	.164

a. Dependent Variable: y.

剔除 x_4 后，仍然有不显著的自变量，此时最大的 P 值为 $p_8=0.813$，因此进一步剔除 x_8，用 y 与其余 6 个自变量作回归。依次剔除 P 值最大的自变量，直到最后所有的自变量在显著性水平 $\alpha=0.05$ 时都显著。最终方程中保留 x_1，x_2，x_3，x_6，x_7，其回归系数表见输出结果 3.4。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 3.4

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2393.625	302.082		-7.924	.000
	x1	1.490	.088	.528	16.934	.000
	x2	2.717	.337	.176	8.061	.000
	x3	2.208	.289	.207	7.634	.000
	x6	.078	.011	.231	7.394	.000
	x7	.037	.014	.061	2.716	.012

a. Dependent Variable: y.

在一元线性回归中，回归系数显著性的 t 检验与回归方程显著性的 F 检验是等价的，而在多元线性回归中，这两种检验是不等价的。F 检验显著，说明 y 对自变量 x_1，x_2，…，x_p 整体的线性回归效果是显著的，但不等于 y 对每个自变量 x_i 的效果都显著。反之，某个或某几个 x_i 的系数不显著，回归方程显著性的 F 检验仍有可能是显著的。

可以从另外一个角度考虑自变量 x_j 的显著性。y 对自变量 x_1，x_2，…，x_p 线性回归的残差平方和为 SSE，回归平方和为 SSR，在剔除掉 x_j 后，用 y 对其余的 $p-1$ 个自变量作回归，记所得的残差平方和为 $\mathrm{SSE}_{(j)}$，回归平方和为 $\mathrm{SSR}_{(j)}$，则自变量 x_j 对回归的贡献为 $\Delta\mathrm{SSR}_{(j)}=\mathrm{SSR}-\mathrm{SSR}_{(j)}$，称为 x_j 的偏回归平方和。由此构造偏 F 统计量

$$F_j=\frac{\Delta\mathrm{SSR}_{(j)}/1}{\mathrm{SSE}/(n-p-1)} \tag{3.42}$$

当原假设 $H_{0j}:\beta_j=0$ 成立时，（3.42）式的偏 F 统计量 F_j 服从自由度为（1，$n-p-1$）的 F 分布，此 F 检验与（3.40）式的 t 检验是一致的，可以证明 $F_j=t_j^2$，当从回归方程中剔除变元时，回归平方和减少，残差平方和增加。根据平方和分解式可知，$\Delta\mathrm{SSR}_{(j)}=\Delta\mathrm{SSE}_{(j)}=\mathrm{SSE}_{(j)}-\mathrm{SSE}$。反之，往回归方程中引入变元时，回归平方和增加，残差平方和减少，两者的增减量同样相等。

三、回归系数的置信区间

当我们有了参数向量 $\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}$ 时，$\hat{\boldsymbol{\beta}}$ 与 $\boldsymbol{\beta}$ 的接近程度如何？这就需构造 β_j 的一个区间，以 $\hat{\beta}_j$ 为中心的区间，该区间以一定的概率包含 β_j。

由（3.39）式可知

$$t_j=\frac{\hat{\beta}_j-\beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}}\sim t(n-p-1) \tag{3.43}$$

专业课辅导+真题资料+微信公众号：432应统书院

仿照（2.69）式一元线性回归系数区间估计的推导过程，可得 β_j 的置信度为 $1-\alpha$ 的置信区间为

$$(\hat{\beta}_j - t_{\alpha/2}\sqrt{c_{jj}}\,\hat{\sigma},\ \hat{\beta}_j + t_{\alpha/2}\sqrt{c_{jj}}\,\hat{\sigma}) \tag{3.44}$$

用 SPSS 软件可计算出例 3.1 数据的回归系数区间估计。

四、拟合优度

拟合优度用于检验回归方程对样本观测值的拟合程度。在一元线性回归中，定义了样本决定系数 $r^2=\mathrm{SSR}/\mathrm{SST}$，在多元线性回归中，同样可以定义样本决定系数为

$$R^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=1-\frac{\mathrm{SSE}}{\mathrm{SST}} \tag{3.45}$$

样本决定系数 R^2 的取值在［0，1］区间内，R^2 越接近 1，表明回归拟合的效果越好；R^2 越接近 0，表明回归拟合的效果越差。与 F 检验相比，R^2 可以更清楚直观地反映回归拟合的效果，但是并不能作为严格的显著性检验。

称

$$R=\sqrt{R^2}=\sqrt{\frac{\mathrm{SSR}}{\mathrm{SST}}} \tag{3.46}$$

为 y 关于 x_1，x_2，…，x_p 的样本复相关系数。在两个变量的简单相关系数中，相关系数有正负之分，而复相关系数表示的是因变量 y 与全体自变量之间的线性关系，它的符号不能由某一个自变量的回归系数的符号来确定，因而都取正号。与一元线性回归方程中曾定义的相关系数 r 一样，在多元线性回归的实际应用中，人们用复相关系数 R 来表示回归方程对原有数据拟合程度的好坏，它衡量作为一个整体的 x_1，x_2，…，x_p 与 y 的线性关系。

在实际应用中，样本决定系数 R^2 到底多大时，才算通过了拟合优度检验？这要根据具体情况来定。在此需要指出的是，拟合优度并不是检验模型优劣的唯一标准，有时为了使得模型从结构上有较合理的经济解释，在 n 较大时，R^2 等于 0.7 左右我们也给回归模型以肯定的态度。在后面的回归变量选择中，还将会看到 R^2 与回归方程中自变量的数目以及样本量 n 有关，当样本量 n 与自变量的个数接近时，R^2 易接近 1，其中隐含着一些虚假成分。因此，由 R^2 决定模型优劣时还需慎重。

3.5 中心化和标准化

在多元线性回归分析中，由于涉及多个自变量，自变量的单位往往不同，给利

专业课辅导+真题资料+微信公众号：432应统书院

用回归方程进行结构分析带来一定困难；由于多元回归涉及的数据量很大，可能因为舍入误差而使计算结果不理想。尽管计算机能使我们保留更多位的小数，但舍入误差肯定还会出现。因此，对原始数据进行一些处理，尽量避免大的误差是有实际意义的。

产生舍入误差有两个主要原因：一是回归分析计算中数据量级有很大差异，比如数据 892 976 与 0.582 这样的大小相差悬殊的数据出现在同一个计算中；二是设计矩阵 $\boldsymbol{X}$ 的列向量近似线性相关时，$\boldsymbol{X}'\boldsymbol{X}$ 为病态矩阵，其逆矩阵 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 就会产生较大的误差。

一、中心化

多元线性回归模型的一般形式（3.1）式为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon$$

其经验回归方程（3.19）式为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$$

此经验回归方程经过样本中心（$\bar{x}_1$，$\bar{x}_2$，…，$\bar{x}_p$；$\bar{y}$），将坐标原点移至样本中心，即作坐标变换

$$\boxed{\begin{aligned} &x'_{ij}=x_{ij}-\bar{x}_j, \quad i=1,2,\cdots,n; \quad j=1,2,\cdots,p \\ &y'_i=y_i-\bar{y}, \quad i=1,2,\cdots,n \end{aligned}} \tag{3.47}$$

上述经验方程式即转变为

$$\hat{y}'=\hat{\beta}_1x'_1+\hat{\beta}_2x'_2+\cdots+\hat{\beta}_px'_p \tag{3.48}$$

（3.48）式即中心化经验回归方程。中心化经验回归方程的常数项为 0，而回归系数的最小二乘估计值 $\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 保持不变，这一点是容易理解的。这是因为坐标系的平移变换只改变直线的截距，不改变直线的斜率。

中心化后的经验回归方程（3.48）式只包含 p 个参数估计值 $\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，比（3.19）式的一般经验回归方程减少了一个未知参数。在变量较多时，减少一个未知参数，计算工作量会减少许多，对手工计算尤其重要。因而在用手工计算求解线性回归方程时，通常先对数据中心化，求出中心化经验回归方程（3.48）式，再由

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2-\cdots-\hat{\beta}_p\bar{x}_p$$

求出常数项估计值 $\hat{\beta}_0$。

二、标准化回归系数

在上述中心化的基础上，我们可进一步给出变量的标准化和标准化回归系数。

专业课辅导+真题资料+微信公众号：432应统书院

在用多元线性回归方程描述某种经济现象时，由于自变量 x_1，x_2，…，x_p 所用的单位大都不同，数据的大小差异也往往很大，这就不利于在同一标准上进行比较。为了消除量纲不同和数量级的差异所带来的影响，就需要将样本数据作标准化处理，然后用最小二乘法估计未知参数，求得标准化回归系数。

样本数据的标准化公式为

$$\boxed{\begin{aligned} x_{ij}^* &= \frac{x_{ij}-\bar{x}_j}{\sqrt{L_{jj}}}, \quad i=1,2,\cdots,n; \quad j=1,2,\cdots,p \\ y_i^* &= \frac{y_i-\bar{y}}{\sqrt{L_{yy}}}, \quad i=1,2,\cdots,n \end{aligned}} \tag{3.49}$$

其中

$$L_{jj}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 \tag{3.50}$$

是自变量 $x_j(j=1,2,\cdots,p)$ 的离差平方和。用最小二乘法求出标准化的样本数据（x_{i1}^*，x_{i2}^*，…，x_{ip}^*；y_i^*）的经验回归方程，记为

$$\hat{y}^*=\hat{\beta}_1^* x_1^*+\hat{\beta}_2^* x_2^*+\cdots+\hat{\beta}_p^* x_p^* \tag{3.51}$$

其中，$\hat{\beta}_1^*$，$\hat{\beta}_2^*$，…，$\hat{\beta}_p^*$ 为 y 对自变量 x_1，x_2，…，x_p 的标准化回归系数。标准化包括了中心化，因而标准化的回归常数项为 0。容易验证，标准化回归系数与普通最小二乘回归系数之间存在关系式

$$\hat{\beta}_j^*=\frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j, \quad j=1,2,\cdots,p \tag{3.52}$$

普通最小二乘估计 $\hat{\beta}_j$ 表示在其他变量不变的情况下，自变量 x_j 的每单位的绝对变化引起的因变量均值的绝对变化量。标准化回归系数 $\hat{\beta}_j^*$ 表示自变量 x_j 的 1%相对变化（相对于 $\sqrt{L_{jj}}$）引起的因变量均值的相对变化百分数（相对于 $\sqrt{L_{yy}}$）。

当自变量所使用的单位不同时，用普通最小二乘估计建立的回归方程，其回归系数不具有可比性，得不到合理的解释。例如有一回归方程为

$$\hat{y}=200+2\,000x_1+2x_2$$

如果不管 x_1，x_2 的单位是什么，人们会很自然地认为 x_1 对因变量 y 的影响最重要，因为 x_1 的系数 2 000 比 x_2 的系数 2 大得多。可是，如果 x_1 的单位是吨，x_2 的单位是公斤，那么 x_1 与 x_2 的重要性实际上是相同的。这是因为 x_1 增加 1 吨时 y 增加 2 000 个单位，x_2 增加 1 公斤时 y 增加 2 个单位，那么 x_2 增加 1 吨时 y 同样增加 2 000 个单位，x_1 增加 1 吨对 y 的影响程度与 x_2 增加 1 吨对 y 的影响程度是相同的。

标准化回归系数是比较自变量对 y 影响程度的相对重要性的一种较为理想的方

专业课辅导+真题资料+微信公众号：432应统书院

法，有了标准化回归系数后，变量的相对重要性就容易进行比较了。但是，我们仍提醒人们对回归系数的解释须采取谨慎的态度，这是因为当自变量相关时会影响标准化回归系数的大小．有关内容在第 6 章中详细讨论。

3.6　相关阵与偏相关系数

一、样本相关阵

复相关系数 R 反映了 y 与一组自变量的相关性，是整体和共性指标；简单相关系数反映的是两个变量间的相关性，是局部和个性指标。我们在分析问题时，应该本着整体与局部相结合、共性与个性相结合的原则。

由样本观测值 x_{i1}，x_{i2}，…，x_{ip}（$i=1$，2，…，n），分别计算 x_i 与 x_j 之间的简单相关系数 r_{ij}，得自变量样本相关阵

$$\boldsymbol{r}=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{3.53}$$

注意相关阵是对称矩阵。记

$$\boldsymbol{X}^*=(x_{ij}^*)_{n\times p}$$

表示中心标准化的设计阵，则相关阵可表示为

$$\boldsymbol{r}=(\boldsymbol{X}^*)'\boldsymbol{X}^* \tag{3.54}$$

进一步求出 y 与每个自变量 x_i 的相关系数 r_{yi}，得增广的样本相关阵为

$$\tilde{\boldsymbol{r}}=\begin{bmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{py} & r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{3.55}$$

用 SPSS 软件计算出的例 3.1 城镇居民消费性支出数据的增广样本相关矩阵如表 3.5 所示，其中表的格式已略作修改。

可以看出，y 与 x_1 的相关系数最大 $r_{y1}=0.932$。$r_{y9}=-0.389$，也可以作为对自变量 x_9 的系数为负数的一个解释。某些自变量间的相关性也很高，例如 $r_{15}=0.766$，说明自变量之间可能存在多重共线性，回归模型还需要进行优化。SPSS 软件同时可以计算出相关系数显著性单侧和双侧检验的 P 值，限于篇幅，本书不在此列出了（操作流程为点击 Analyze→Correlate→Bivariate）。

专业课辅导+真题资料+微信公众号：432应统书院

表 3.5 Correlations（样本相关阵）

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9
y	1	.932	.581	.811	.532	.864	.860	.515	-.503	-.389
x1	.932	1	.353	.690	.302	.766	.746	.445	-.492	-.271
x2	.581	.353	1	.376	.552	.602	.526	.310	-.404	-.123
x3	.811	.690	.376	1	.514	.724	.615	.519	-.480	-.360
x4	.532	.302	.552	.514	1	.464	.669	.203	-.392	-.501
x5	.864	.766	.602	.724	.464	1	.689	.470	-.471	-.253
x6	.860	.746	.526	.615	.669	.689	1	.247	-.201	-.481
x7	.515	.445	.310	.519	.203	.470	.247	1	-.497	-.363
x8	-.503	-.492	-.404	-.480	-.392	-.471	-.201	-.497	1	.059
x9	-.389	-.271	-.123	-.360	-.501	-.253	-.481	-.363	.059	1

二、偏决定系数

前面介绍了复相关系数与简单相关系数，以下介绍变量间的另一种相关性——偏相关。在多元线性回归分析中，当其他变量被固定后，给定的任两个变量之间的相关系数叫偏相关系数。偏相关系数可以度量 $p+1$ 个变量 y，x_1，x_2，…，x_p 之中任意两个变量的线性相关程度，而这种相关程度是在固定其余$p-1$个变量的影响下的线性相关。例如，我们在研究粮食产量与农业投入资金、粮食产量与劳动力投入之间的关系时，农业投入资金的多少会影响粮食产量，劳动力投入的多少也会影响粮食产量。由于资金投入数量的变化，劳动力投入的多少也经常在变化，用简单相关系数往往不能说明现象间的关系程度如何。这就需要在固定其他变量影响的情况下来计算两个变量之间的关系程度，计算出的这种相关系数就称为偏相关系数。我们在研究粮食产量与劳动力投入的关系时可以假定投入资金数量不变；在研究粮食产量与投入资金的关系时可以假定劳动力投入不变。复决定系数 R^2 测量回归中一组自变量 x_1，x_2，…，x_p 使因变量 y 的变差的相对减少量。相应地，偏决定系数测量在回归方程中已包含若干个自变量时，再引入某一个新的自变量时，y 的剩余变差的相对减少量，它衡量某自变量对 y 的变差减少的边际贡献。在讲偏相关系数之前，首先引入偏决定系数。

1. 两个自变量的偏决定系数

二元线性回归模型为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i,\quad i=1,2,\cdots,n$$

记 SSE(x_2) 是模型中只含有自变量 x_2 时 y 的残差平方和，SSE(x_1，x_2) 是模型中同时含有自变量 x_1 和 x_2 时 y 的残差平方和。因此，模型中已含有 x_2 时，再加入 x_1 使 y 的剩余变差的相对减小量为

专业课辅导+真题资料+微信公众号：432应统书院

$$r^2_{y1\cdot 2}=\frac{\mathrm{SSE}(x_2)-\mathrm{SSE}(x_1,x_2)}{\mathrm{SSE}(x_2)} \tag{3.56}$$

此即模型中已含有 x_2 时，y 与 x_1 的偏决定系数。

同样地，模型中已含有 x_1 时，y 与 x_2 的偏决定系数为

$$r^2_{y2\cdot 1}=\frac{\mathrm{SSE}(x_1)-\mathrm{SSE}(x_1,x_2)}{\mathrm{SSE}(x_1)} \tag{3.57}$$

2. 一般情况

当模型中已含有 x_2，…，x_p 时，y 与 x_1 的偏决定系数为

$$r^2_{y1\cdot 2,\cdots,p}=\frac{\mathrm{SSE}(x_2,\cdots,x_p)-\mathrm{SSE}(x_1,x_2,\cdots,x_p)}{\mathrm{SSE}(x_2,\cdots,x_p)} \tag{3.58}$$

其余情况依此类推。由思考与练习题 3.9 知，偏决定系数与回归系数显著性检验的偏 F 值是等价的。

三、偏相关系数

偏决定系数的平方根称为偏相关系数，其符号与相应的回归系数的符号相同。偏相关系数与回归系数显著性检验的 t 值是等价的。

例 3.2

为了研究北京市各经济开发区经济发展与招商投资的关系，我们以因变量 y 为各开发区的销售收入（百万元），选取两个自变量：x_1 为截至 1998 年底各开发区累计招商数目，x_2 为招商企业注册资本（百万元）。表 3.6 列出了至 1998 年底招商企业注册资本 x_2 在 5 亿元～50 亿元的 15 个开发区的数据。以 y 对 x_1 和 x_2 建立二元线性回归，用 SPSS 软件计算出回归系数及偏相关系数，见输出结果 3.5。

表 3.6　北京开发区数据

x_1	x_2	y	x_1	x_2	y
25	3 547.79	553.96	7	671.13	122.24
20	896.34	208.55	532	2 863.32	1 400.00
6	750.32	3.10	75	1 160.00	464.00
1 001	2 087.05	2 815.40	40	862.75	7.50
525	1 639.31	1 052.12	187	672.99	224.18
825	3 357.70	3 427.00	122	901.76	538.94
120	808.47	442.82	74	3 546.18	2 442.79
28	520.27	70.12			

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 3.5

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Correlations		
	B	Std.Error	Beta	t	Sig.	Zero-order	Partial	Part
(Constant)	-327.039	218.001		-1.500	.159			
×1	2.036	.438	.594	4.649	.001	.807	.802	.534
×2	.468	.123	.485	3.799	.003	.746	.739	.436

a.Dependent Variable:y.

从输出结果 3.5 中看到，两个偏相关系数分别为 $r_{y1;2}=0.802$，$r_{y2;1}=0.739$，进一步计算偏决定系数 $r^2_{y1;2}=(0.802)^2=0.643$，$r^2_{y2;1}=(0.739)^2=0.546$。表中相关系数栏的 Zero-order 为 y 与 x_i 的简单相关系数，分别为 $r_{y1}=0.807$，$r_{y2}=0.746$，两个决定系数分别为 $r^2_{y1}=(0.807)^2=0.651$，$r^2_{y2}=(0.746)^2=0.557$。

表中的 Part 为部分相关系数。$\Delta SSR(x_2)=SSR(x_1, x_2)-SSR(x_1)$ 为先引入 x_1，再引入 x_2 时 SSR 的增量，那么 y 关于 x_2 的部分相关系数$=\sqrt{\dfrac{\Delta SSR(x_2)}{SST}}$。

以上数据表明，用 y 与 x_1 作一元线性回归时，x_1 能消除 y 的变差 SST 的比例为 $r^2_{y1}=0.651=65.1\%$，再引入 x_2 时，x_2 能消除剩余变差 $SSE(x_1)$ 的比例为 $r^2_{y2;1}=0.546=54.6\%$，因而自变量 x_1 和 x_2 消除 y 变差的总比例为 $1-(1-r^2_{y1})(1-r^2_{y2;1})=1-(1-0.651)\times(1-0.546)=0.842=84.2\%$。这个值 84.2%恰好是 y 对 x_1 和 x_2 二元线性回归的决定系数 R^2，这一点请读者自己验证。

相应地，用 y 与 x_2 作一元线性回归时，x_2 能消除 y 的变差 SST 的比例为$r^2_{y2}=0.557=55.7\%$，再引入 x_1 时，x_1 能消除剩余变差 $SSE(x_2)$ 的比例为$r^2_{y1;2}=0.643=64.36\%$，因而自变量 x_1 和 x_2 消除 y 变差的总比例为 $1-(1-r^2_{y2})(1-r^2_{y1;2})=1-(1-0.557)(1-0.643)=0.842=84.2\%$。这个值同样是 y 对 x_1 和 x_2 二元线性回归的决定系数 R^2。

偏相关系数反映的是变量间的相关性，因而并不需要有处于特殊地位的变量 y，我们可以对任意 p 个变量 x_1，x_2，…，x_p 定义它们之间的偏相关系数。记

$$r_{ij}=\frac{L_{ij}}{\sqrt{L_{ii}\cdot L_{jj}}} \tag{3.59}$$

表示两个变量 x_i，x_j 之间的简单相关系数，$\boldsymbol{r}=(r_{ij})_{p\times p}$ 为 x_1，x_2，…，x_p 的相关阵，则在固定 x_3，…，x_p 保持不变时，x_1 与 x_2 之间的偏相关系数为

$$r_{12;3,\cdots,p}=\frac{-\Delta_{12}}{\sqrt{\Delta_{11}\cdot\Delta_{22}}} \tag{3.60}$$

其余变量间偏相关系数的定义依此类推，这个定义与用（3.58）式的平方根的定义是等价的。

其中符号 Δ_{ij} 表示相关阵 $(r_{ij})_{p\times p}$ 第 i 行第 j 列元素的代数余子式，注意相关阵

专业课辅导+真题资料+微信公众号：432应统书院

$(r_{ij})_{p\times p}$是对称矩阵。容易验证以下关系

$$r_{12;3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \tag{3.61}$$

再用一个例子说明偏相关系数和简单相关系数的关系。分别以 x_1表示某种商品的销售量，x_2表示消费者人均可支配收入，x_3表示商品价格。从经验上看，销售量 x_1与消费者人均可支配收入 x_2之间应该有正相关，简单相关系数 r_{12}应该是正的。但是如果你计算出的 r_{12}是个负数也不要感到惊讶，这是因为还有其他没有被固定的变量在产生影响，例如商品价格 x_3在这期间大幅提高了。反映固定 x_3后 x_1与 x_2相关程度的偏相关系数 $r_{12;3}$会是个正数。如果你计算出的偏相关系数 $r_{12;3}$仍然是个负数的话，想一想会是什么原因？肯定是还有需要考虑而没有考虑的重要变量，也就是没有被固定的变量，会是什么变量？如果这种商品已经进入淘汰期，正在被其他商品所取代，那么你计算出负的 $r_{12;3}$也就不足为奇了。

在多元回归中，应注意简单相关系数只是两变量局部的相关性质，而并非整体的性质。所以在多元线性回归分析中我们并不看重简单相关系数，而认为偏相关系数才是真正反映因变量 y 与自变量 x_i 以及自变量 x_i 与 x_j 的相关性的数值。根据偏相关系数，可以判断哪些自变量对因变量的影响较大，而选择必须考虑的自变量，对于那些对因变量影响较小的自变量，则可以舍去不顾。在剔除某个自变量时，可以结合偏相关系数考虑。

3.7　本章小结与评注

一、多元线性回归模型的建立过程

本章我们结合两个经济问题实例介绍了多元线性回归模型的建立过程，在此，我们再结合一个实例，对多元线性回归模型的建立过程与应用做一个完整的介绍。

例 3.3

中国民航客运量的回归模型。为了研究我国民航客运量的变化趋势及其成因，我们以民航客运量作为因变量 y，以国民收入、消费额、铁路客运量、民航航线里程、来华旅游入境人数为影响民航客运量的主要因素。y 表示民航客运量（万人），x_1 表示国民收入（亿元），x_2 表示消费额（亿元），x_3 表示铁路客运量（万人），x_4 表示民航航线里程（万公里），x_5 表示来华旅游入境人数（万人）。根据《1994 年统计摘要》获得 1978—1993 年统计数据，见表 3.7。

专业课辅导+真题资料+微信公众号：432应统书院

表 3.7

年份	y	x_1	x_2	x_3	x_4	x_5
1978	231	3 010	1 888	81 491	14.89	180.92
1979	298	3 350	2 195	86 389	16.00	420.39
1980	343	3 688	2 531	92 204	19.53	570.25
1981	401	3 941	2 799	95 300	21.82	776.71
1982	445	4 258	3 054	99 922	23.27	792.43
1983	391	4 736	3 358	106 044	22.91	947.70
1984	554	5 652	3 905	110 353	26.02	1 285.22
1985	744	7 020	4 879	112 110	27.72	1 783.30
1986	997	7 859	5 552	108 579	32.43	2 281.95
1987	1 310	9 313	6 386	112 429	38.91	2 690.23
1988	1 442	11 738	8 038	122 645	37.38	3 169.48
1989	1 283	13 176	9 005	113 807	47.19	2 450.14
1990	1 660	14 384	9 663	95 712	50.68	2 746.20
1991	2 178	16 557	10 969	95 081	55.91	3 335.65
1992	2 886	20 223	12 985	99 693	83.66	3 311.50
1993	3 383	24 882	15 949	105 458	96.08	4 152.70

第一步，提出因变量与自变量，收集数据，如例 3.3 所示。

第二步，作相关分析，设定理论模型。用 SPSS 软件计算增广相关阵，见输出结果 3.6。

输出结果 3.6

Correlations

		Y	X1	X2	X3	X4	X5
Y	Pearson Correlation	1.000	.989	.985	.227	.987	.924
	Sig. (2-tailed)	.	.000	.000	.398	.000	.000
	N	16	16	16	16	16	16
X1	Pearson Correlation	.989	1.000	.999	.258	.984	.930
	Sig. (2-tailed)	.000	.	.000	.335	.000	.000
	N	16	16	16	16	16	16
X2	Pearson Correlation	.985	.999	1.000	.289	.978	.942
	Sig. (2-tailed)	.000	.000	.	.278	.000	.000
	N	16	16	16	16	16	16
X3	Pearson Correlation	.227	.258	.289	1.000	.213	.504
	Sig. (2-tailed)	.398	.335	.278	.	.428	.046
	N	16	16	16	16	16	16
X4	Pearson Correlation	.987	.984	.978	.213	1.000	.882
	Sig. (2-tailed)	.000	.000	.000	.428	.	.000
	N	16	16	16	16	16	16
X5	Pearson Correlation	.924	.930	.942	.504	.882	1.00
	Sig. (2-tailed)	.000	.000	.000	.046	.000	.
	N	16	16	16	16	16	16

从相关阵看出，y 与 x_1，x_2，x_4，x_5 的相关系数都在 0.9 以上，说明所选自变量与 y 高度线性相关，用 y 与自变量作多元线性回归是适合的。y 与 x_3 的相关系数 $r_{y3}=0.227$ 偏小，P 值$=0.398$，x_3 是铁路客运量，这说明铁路客运量对民航客

专业课辅导+真题资料+微信公众号：432应统书院

运量无显著影响。一般认为铁路客运量与民航客运量之间应呈负相关，铁路和民航共同拥有旅客，乘了火车就乘不了飞机。但就中国当时的实际情况分析，我国居民的收入还很低，一般人外出旅游、出差都乘火车。近年来乘飞机的人虽逐渐增多，但我国民航客运量最大的一部分是来华旅游入境人数。国内尽管有些旅客乘坐飞机，但对火车客运量不会有大的影响，一是铁路运力十分紧张；二是近年来外出民工增多，而民工主要乘火车，所以不会因民航客运量增加而导致火车客运量下降。因此铁路客运量与民航客运量之间的关系不密切是正常的。那么在回归方程中是否还应该包含 x_3 呢？仅凭简单相关系数的大小是不能决定变量的取舍的，在初步建模时还是应该包含 x_3 在内。请读者注意，现在已进入高铁年代，用现在的数据建模，结合高铁和民航运行的现状，可能会得到另外的解释。

第三步，用软件计算，输出计算结果。本例采用 SPSS 软件对原始数据作回归分析，见输出结果 3.7。

输出结果 3.7

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.999[a]	0.998	0.997	49.4924

a. Predictors: (Constant), x5, x3, x4, x2, x1.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13818876.769	5	2763775.354	1128.303	.000
	Residual	24494.981	10	2449.498		
	Total	13843371.750	15			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	450.909	178.078		2.532	.030
X1	.354	.085	2.447	4.152	.002
X2	-.561	.125	-2.485	-4.478	.001
X3	-7.254E-03	.002	-.083	-3.510	.006
X4	21.578	4.030	.531	5.354	.000
X5	.435	.052	.564	8.440	.000

第四步，回归诊断。

（1）回归方程为

$$\hat{y}=450.9+0.354x_1-0.561x_2-0.007\,3x_3+21.578x_4+0.435x_5 \quad (3.62)$$

（2）复相关系数 $R=0.999$，决定系数 $R^2=0.998$，由决定系数看回归方程高度显著。

（3）方差分析表，$F=1\,128.303$，P 值$=0.000$，表明回归方程高度显著，说

专业课辅导+真题资料+微信公众号：432应统书院

明 x_1，x_2，x_3，x_4，x_5 整体上对 y 有高度显著的线性影响。

（4）回归系数的显著性检验。自变量 x_1，x_2，x_3，x_4，x_5 对 y 均有显著影响，其中 x_3 铁路客运量的 P 值=0.006 最大，但仍在 1%的显著性水平上对 y 高度显著，这充分说明在多元线性回归中不能仅凭简单相关系数的大小而决定变量的取舍。

第五步，回归应用。

因变量新值的点估计为

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{10}+\hat{\beta}_2x_{20}+\cdots+\hat{\beta}_px_{p0} \tag{3.63}$$

其精确置信区间的表达式较为复杂，也不可能用手工计算，可以仿照一元线性回归的情况用 SPSS 软件计算。其置信度为 95%的近似置信区间为

$$(\hat{y}-2\hat{\sigma},\hat{y}+2\hat{\sigma}) \tag{3.64}$$

另外，x_2 的回归系数 −0.561 是负的，x_2 是消费额，负的回归系数显然是不合理的，其原因可能是自变量之间的共线性，因而回归方程（3.62）式还要在第 6 章多重共线性部分作进一步改进，或用其他消除共线性的方法重新建立回归方程，在此暂不作具体的应用。

二、评注

对于多元线性回归模型未知参数向量 $\boldsymbol{\beta}$ 的估计，最主要的方法是普通最小二乘估计。在运用普通最小二乘法估计未知参数时，应首先看具体问题的样本数据是否满足模型的基本假定，只有满足基本假定的模型才能应用普通最小二乘法。前面的几个例子都是假设满足基本假定要求的，在后面几章中我们还将会看到不满足基本假定的情况时，如何估计未知参数。

当回归模型的未知参数估计出来后，我们实际上是由 n 组样本观测数据得到一个经验回归方程，这个经验回归方程是否真正反映了变量 y 和变量 x_1，x_2，…，x_p 之间的线性关系，这就需要进一步对回归方程进行检验。一种检验方法是拟合优度检验，即用样本决定系数的大小来衡量模型的拟合优度。样本决定系数 R^2 越大，说明回归方程拟合原始数据 y 的观测值的效果越好。但由于 R^2 的大小与样本量 n 以及自变量个数 p 有关，当 n 与 p 的数目接近时，R^2 容易接近 1，这说明 R^2 中隐含着一些虚假成分。因此，仅由 R^2 的值很大去推断模型优劣一定要慎重。

对于回归方程的显著性检验，我们用 F 统计量去判断假设 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$ 是否成立。当给定显著性水平 α 时，$F>F_\alpha(p,n-p-1)$，则拒绝假设 H_0，否则接受 H_0。接受假设 H_0 和拒绝假设 H_0 对于回归方程来说意味着什么，仍需慎重对待。

一般来说，当接受假设 H_0 时，认为在给定的显著性水平 α 下，自变量 x_1，x_2，…，x_p 对因变量 y 无显著影响，于是通过 x_1，x_2，…，x_p 去推断 y 也就无多大意义。在这种情况下，一方面可能这个问题本来应该用非线性模型去描述，而我

专业课辅导+真题资料+微信公众号：432应统书院

们误用了线性模型，使得自变量对因变量无显著影响；另一方面可能是在考虑自变量时，由于我们认识上的局限性把一些影响因变量 y 的自变量漏掉了，这就从两个方面提醒我们重新考虑建模问题。

当拒绝了假设 H_0 时，我们也不能过于相信这个检验，认为这个回归模型已经很完美了。其实当拒绝 H_0 时，我们只能认为这个回归模型在一定程度上说明了自变量 x_1，x_2，…，x_p 与因变量 y 的线性关系。因为这时仍不能排除我们漏掉了一些重要的自变量。参考文献［2］的作者认为，此检验只宜用于辅助性的、事后验证性质的目的。研究者在事前根据专业知识及经验，认为已把较重要的自变量选入了，且在一定误差限度内认为模型为线性是合理的。经过样本数据计算后，可以用来验证原先的考虑是否周全。这时，若拒绝 H_0，可认为至少并不与其原来的设想矛盾。如果接受 H_0，可以肯定模型不能反映因变量 y 与自变量 x_1，x_2，…，x_p 的线性关系，这个模型就不能用于实际预测和分析。

当样本量 n 较小，变量个数 p 较大时，F 检验或 t 检验的自由度太小，这时尽管样本决定系数 R^2 很大，但参数估计的效果很不稳定，我们曾发现一个实际应用例子暴露出这方面的问题。某参考文献在研究建筑业降低成本率 y 与流动资金 x_1、固定资金 x_2、优良品率 x_3、竣工面积 x_4、劳动生产率 x_5、施工产值 x_6 的关系时，利用表 3.8 数据建立回归方程，得回归方程

$$\hat{y}=-38.487\,36-0.000\,273\,9x_1-0.002\,833x_2+0.205\,18x_3$$
$$-0.755\,07x_4+0.005\,227x_5+0.002\,657\,4x_6$$

$$\mathrm{SST}=154.764\,6,\quad \mathrm{SSR}=143.45,\quad \mathrm{SSE}=11.314\,6$$

$$F=4.226,\quad R^2=0.926\,79$$

由于 $R^2=0.926\,79$，所以该文献中作者认为上述回归方程非常显著。其实进一步作 F 检验，给定 $\alpha=0.05$，查 F 分布表：$F_{0.05}(p,\ n-p-1)=F_{0.05}(6,\ 2)=19.3$。$F=4.226<F_{0.05}(6,\ 2)=19.3$，回归方程没有通过 F 检验。可是该参考文献当时给错了自由度，查 $F_{0.05}(6,\ 9)=3.37$。结果 $F>F_{0.05}(6,\ 9)$，通过了检验，从而进一步肯定了上述回归方程。

之所以 R^2 在 0.9 以上，已接近 1，方程还通不过 F 检验，这就是样本量个数 n 太小，而自变量又较多造成 R^2 的虚假现象。如果样本量再稍作改变，未知参数就会发生较大变化，即表现出很不稳定的状况。

表 3.8

序号	降低成本率 y(%)	流动资金 x_1(万元)	固定资金 x_2(万元)	优良品率 x_3(%)	竣工面积 x_4(万平方米)	劳动生产率 x_5(元/人)	施工产值 x_6(万元)
1	5.78	1 297.98	1 543.48	62.68	13.828	6 761	3 666.29
2	6.34	2 164.21	1 527.03	64.99	15.228	7 133	4 320.21
3	5.49	1 429.28	1 714.09	66.96	17.211	6 946	4 786.66
4	−6.99	581.38	681.03	40.03	4.304	4 968	1 262.76
5	7.18	981.78	1 134.31	74.72	12.298	6 810	3 062.90

专业课辅导+真题资料+微信公众号：432应统书院

续前表

序号	降低成本率 y(%)	流动资金 x_1(万元)	固定资金 x_2(万元)	优良品率 x_3(%)	竣工面积 x_4(万平方米)	劳动生产率 x_5(元/人)	施工产值 x_6(万元)
6	6.70	601.21	611.98	60.24	7.481	6 416	1 718.70
7	5.00	588.27	802.21	62.93	10.683	6 911	2 369.13
8	6.56	2 975.63	2 403.22	67.59	25.938	7 124	7 797.64
9	5.01	1 096.10	1 908.98	64.49	9.800	6 540	3 494.30

一个回归方程通过了显著性检验，并不能说明这个回归方程中所有自变量都对因变量 y 有显著影响，因此还要对回归系数进行检验。前面的几个例子中，我们看到回归方程尽管通过了检验，但有些回归系数并没有通过检验。对没有通过检验的回归系数，在一定程度上说明它们对应的自变量在方程中可有可无，一般为了使模型简化，需剔除不显著的自变量，重新建立回归方程。但在实际应用时，为了模型的结构合理，我们有时也保留个别对 y 影响不大的自变量，尤其是在建立宏观经济模型时常常如此。

当一个实际经济问题的回归模型通过各种检验之后，模型的形式就随之确定下来，接着就可以运用回归方程去作经济预测和经济分析。我们在一元线性回归方程的应用中强调注意的问题在多元线性回归方程的应用中仍然有效。由于多元线性回归模型所描述的实际问题的复杂性，在作预测和结构分析时应更慎重。

如果自变量 $x_j(j=1, 2, \cdots, p)$ 的取值 $x_{0j}(j=1, 2, \cdots, p)$ 可以人为控制，其取值范围在当初建模时的范围之内，其他条件也没发生太大变化，这时利用回归方程，根据 $x_{0j}(j=1, 2, \cdots, p)$ 的值去推断 y_0 的预测值 $\hat{y}_0$ 是可行的，预测值与真实值的误差也不会太大。这时把 x_j 的回归系数 $\hat{\beta}_j$ 解释为当 x_j 增减 1 单位时，因变量 y 平均增减 $\hat{\beta}_j$ 个单位也是合理的。

在实际应用中，尤其是在经济问题的研究中，我们研究的某种经济现象涉及多个因素，这些因素之间也大都有一定的联系。当回归方程中某一个自变量变动时，往往也会导致其他变量变动。这时，各回归系数的值都是在全体自变量值的联合变动的格局内起作用，如果我们仍认为某一回归系数 $\hat{\beta}_j$ 表示当 x_j 增减 1 单位时，因变量 y 平均增减 $\hat{\beta}_j$ 个单位就不合理了。

回归自变量之间的相关性在经济问题研究中经常存在，只要涉及多个自变量，就很难找出它们当中某些自变量是不相关的。要想找到对某一经济现象既有显著影响，自变量之间又完全不相关的一组自变量几乎是不可能的。问题是我们在建立经济问题的回归模型时，应尽可能地避免自变量间的高度相关。自变量间的高度相关称为多重共线性，它使得最小二乘估计的参数稳健性很差。后面的章节中将专门研究这类问题。

真实的回归函数，特别在较大的范围内，很少是线性的。线性是一种近似，它包含了一种从实际角度看往往不一定合理的假定：各变量的作用与其他变量取什么值无关，且各变量的作用可以叠加。因为，若 $y=\beta_0+\beta_1x_1+\cdots+\beta_px_p$，则不论把

专业课辅导+真题资料+微信公众号：432应统书院

x_2，…，x_p 的值固定在何处，当 x_1 增减 1 单位时，y 总是增减 β_1 个单位。事实并非如此。例如，以 y 记为某公司的销售利润，x_1 为销售量，x_2 为商品价格，x_3 为广告费，x_4 为销售费用，则 x_1 对 y 起的作用与 x_2，x_3，x_4 的值有关。这种现象称为各因素之间的“交互作用”。在这种情况下，单个回归系数意义的解释，也应是基于其他变量的平均而言的。

□ 思考与练习

3.1　写出多元线性回归模型的矩阵表示形式，并给出多元线性回归模型的基本假设。

3.2　讨论样本量 n 与自变量个数 p 的关系，它们对模型的参数估计有何影响？

3.3　证明 $\hat{\sigma}^2=\dfrac{1}{n-p-1}\mathrm{SSE}$ 是误差项方差 σ^2 的无偏估计。

3.4　一个回归方程的复相关系数 $R=0.99$，样本决定系数 $R^2=0.9801$，我们能断定这个回归方程就很理想吗？

3.5　如何正确理解回归方程显著性检验拒绝 H_0 或接受 H_0？

3.6　数据中心化和标准化在回归分析中的意义是什么？

3.7　验证（3.52）式

$$\hat{\beta}_j^*=\frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j,\quad j=1,2,\cdots,p$$

3.8　利用（3.60）式证明（3.61）式成立，即

$$r_{12\cdot3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

3.9　证明 y 与自变量 x_j 的偏决定系数与（3.42）式的偏 F 检验值 F_j 是等价的。

3.10　验证决定系数 R^2 与 F 值之间的关系式

$$R^2=\frac{F}{F+(n-p-1)/p}$$

3.11　研究货运总量 y（万吨）与工业总产值 x_1（亿元）、农业总产值 x_2（亿元）、居民非商品支出 x_3（亿元）的关系。数据见表 3.9。

（1）计算出 y，x_1，x_2，x_3 的相关系数矩阵。

（2）求 y 关于 x_1，x_2，x_3 的三元线性回归方程。

（3）对所求得的方程作拟合优度检验。

（4）对回归方程作显著性检验。

（5）对每一个回归系数作显著性检验。

专业课辅导+真题资料+微信公众号：432应统书院

（6）如果有的回归系数没通过显著性检验，将其剔除，重新建立回归方程，再作回归方程的显著性检验和回归系数的显著性检验。

（7）求出每一个回归系数的置信水平为95%的置信区间。

（8）求标准化回归方程。

（9）求当 $x_{01}=75$，$x_{02}=42$，$x_{03}=3.1$ 时的 $\hat{y}_0$，给定置信水平为95%，用SPSS软件计算精确置信区间，用手工计算近似预测区间。

（10）结合回归方程对问题作一些基本分析。

表 3.9

编号	货运总量 y(万吨)	工业总产值 x_1(亿元)	农业总产值 x_2(亿元)	居民非商品支出 x_3(亿元)
1	160	70	35	1.0
2	260	75	40	2.4
3	210	65	40	2.0
4	265	74	42	3.0
5	240	72	38	1.2
6	220	68	45	1.5
7	275	78	42	4.0
8	160	66	36	2.0
9	275	70	44	3.2
10	250	65	42	3.0

3.12 用表3.10的数据，建立GDP对 x_1 和 x_2 的回归。对得到的二元回归方程 $\hat{y}=2\,914.6+0.607x_1+1.709x_2$，你能够合理地解释两个回归系数吗？如果现在不能给出合理的解释，不妨在学过第6章多重共线性后再来解释这个问题，在学过第7章岭回归后再来改进这个问题。

表 3.10 **GDP和三次产业数据** 单位：亿元

年份	GDP	第一产业增加值 x_1	第二产业增加值 x_2	第三产业增加值 x_3
1990	18 547.9	5 017.0	7 717.4	5 813.5
1991	21 617.8	5 288.6	9 102.2	7 227.0
1992	26 638.1	5 800.0	11 699.5	9 138.6
1993	34 634.4	6 882.1	16 428.5	11 323.8
1994	46 759.4	9 457.2	22 372.2	14 930.0
1995	58 478.1	11 993.0	28 537.9	17 947.2
1996	67 884.6	13 844.2	33 612.9	20 427.5
1997	74 462.6	14 211.2	37 222.7	23 028.7
1998	78 345.2	14 552.4	38 619.3	25 173.5
1999	82 067.5	14 472.0	40 557.8	27 037.7

专业课辅导+真题资料+微信公众号：432应统书院

续前表

年份	GDP	第一产业增加值 x_1	第二产业增加值 x_2	第三产业增加值 x_3
2000	89 468.1	14 628.2	44 935.3	29 904.6
2001	97 314.8	15 411.8	48 750.0	33 153.0
2002	105 172.3	16 117.3	52 980.2	36 074.8
2003	117 390.2	16 928.1	61 274.1	39 188.0
2004	136 875.9	20 768.1	72 387.2	43 720.6

资料来源：2005 年《中国统计年鉴》，未经 2005 年经济普查修订。

专业课辅导+真题资料+微信公众号：432应统书院

第 4 章 Chapter 4 违背基本假设的情况

在回归模型的基本假设中，假定随机误差项 ε_1，ε_2，…，ε_n 具有相同的方差，独立或不相关，即对于所有样本点，有

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,n \\ \mathrm{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, i=j \\ 0, \ i\neq j\end{cases} \quad i,j=1,2,\cdots,n \end{cases}$$

但在建立实际问题的回归模型时，经常存在与此假设相违背的情况，一种是计量经济建模中常说的异方差性，即

$$\mathrm{var}(\varepsilon_i)\neq\mathrm{var}(\varepsilon_j), \quad 当\ i\neq j\ 时$$

另一种是自相关性，即

$$\mathrm{cov}(\varepsilon_i,\varepsilon_j)\neq 0, \quad 当\ i\neq j\ 时$$

本章将结合实例介绍异方差性及自相关性产生的背景和原因，以及给回归建模带来的影响，异方差性及自相关性问题的诊断及处理方法。

4.1 异方差性产生的背景和原因

一、异方差性产生的原因

由于实际问题是错综复杂的，因而在建立实际问题的回归分析模型时，经常会出现某一因素或一些因素随着解释变量观测值的变化而对被解释变量产生不同的影

专业课辅导+真题资料+微信公众号：432应统书院

响，导致随机误差项产生不同方差。通过下面的几个例子，我们可以了解产生异方差性的背景和原因。

例 4.1

在研究城镇居民收入与购买量的关系时，我们知道居民收入与消费水平有着密切的关系。用 x_i 表示第 i 户的收入量，y_i 表示第 i 户的消费额，一个简单的消费模型为

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\quad i=1,2,\cdots,n$$

在此问题中，由于各户的收入、消费观念和习惯不同，通常存在明显的异方差性。一般情况下，低收入的家庭购买差异性比较小，大都购买生活必需品，但高收入的家庭购买行为差异就很大。高档消费品很多，房子、汽车的选择余地也很大，这样购买金额的差异就很大，导致消费模型的随机项 ε_i 具有不同的方差。

例 4.2

利用某行业的不同企业的截面样本数据估计 C-D 生产函数

$$y=AK^{\alpha}L^{\beta}e^{\varepsilon}$$

这里的 ε 表示不同企业的设备、工艺、地理条件、工人素质、管理水平上的差异以及其他因素，对于不同企业，这些因素对产出的影响程度不同，引起 ε_i 偏离 0 均值的程度不同，进而出现了异方差。

利用平均数作为样本数据，也容易出现异方差。鉴于正态分布的普遍性，许多经济变量之间的关系服从正态分布。例如，不同收入水平组的人数随收入增加呈正态分布。以不同收入组的人均数据作为样本时，由于每组中人数不同，观测误差也不同。一般来说，人数多的收入组的人均数据较人数少的收入组的人均数据具有较高的准确性。这些不同的观测误差也会引起异方差，且 $\mathrm{var}(\varepsilon_i)$ 随收入的增加呈先降后升的趋势。参见参考文献［9］。

总之，引起异方差的原因很多，但当样本数据为截面数据时容易出现异方差。

二、异方差性带来的问题

当一个回归问题存在异方差时，如果仍用普通最小二乘法估计未知参数，将引起不良后果，特别是最小二乘估计量不再具有最小方差的优良性，即最小二乘估计的有效性被破坏了。

当存在异方差时，参数向量 $\hat{\boldsymbol{\beta}}$ 的方差大于在同方差条件下的方差，如果用普通最小二乘法估计参数，将出现低估 $\hat{\boldsymbol{\beta}}$ 的真实方差的情况，进一步将导致高估回归系

专业课辅导+真题资料+微信公众号：432应统书院

数的 t 检验值，可能造成本来不显著的某些回归系数变成显著。这将给回归方程的应用效果带来一定影响。

当存在异方差时，普通最小二乘估计存在以下问题：

（1）参数估计值虽是无偏的，但不是最小方差线性无偏估计。

（2）参数的显著性检验失效。

（3）回归方程的应用效果极不理想。

4.2 一元加权最小二乘估计

一、异方差性的检验

关于异方差性的检验，统计学家进行了大量的研究，提出的诊断方法已有 10 多种，但没有一个公认的最优方法。本书介绍残差图分析法与等级相关系数法两种常用方法。

1. 残差图分析法

残差图分析法是一种直观、方便的分析方法。它以残差 e_i 为纵坐标，以其他适宜的变量为横坐标画散点图。常用的横坐标有三种选择：（1）以拟合值 $\hat{y}$ 为横坐标；（2）以 $x_i(i=1, 2, \cdots, p)$ 为横坐标；（3）以观测时间或序号为横坐标。

如果回归模型适合样本数据，那么残差 e_i 应反映 ε_i 所假定的性质，因此可以根据它来判断回归模型是否具有某些性质。一般情况下，当回归模型满足所有假定时，残差图上的 n 个点的散布应是随机的，无任何规律，如图2.5（a)所示。如果回归模型存在异方差，残差图上的点的散布呈现出一定的趋势。图2.5（b)的残差 e 值随 y 值的增大而增大，具有明显的规律，因而可认为模型的随机误差项 ε_i 的方差是非齐性的，存在异方差。另外，残差 e 值也可能随 y 值的增大而减小，这种情况同样属于存在异方差。

2. 等级相关系数法

等级相关系数法又称斯皮尔曼（Spearman）检验，是一种应用较广泛的方法。进行等级相关系数检验通常有三个步骤：

第一步，作 y 关于 x 的普通最小二乘回归，求出 ε_i 的估计值，即 e_i 的值。

第二步，取 e_i 的绝对值，即 $|e_i|$，把 x_i 和 $|e_i|$ 按递增或递减的次序排列后分成等级，按下式计算出等级相关系数

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \tag{4.1}$$

专业课辅导+真题资料+微信公众号：432应统书院

其中，n 为样本量；d_i 为对应于 x_i 和 $|e_i|$ 的等级的差数。

第三步，做等级相关系数的显著性检验。在 $n>8$ 的情况下，用下式对样本等级相关系数 r_s 进行 t 检验。检验统计量为

$$t=\frac{\sqrt{n-2}r_s}{\sqrt{1-r_s^2}} \tag{4.2}$$

如果 $t\leqslant t_{\alpha/2}(n-2)$，可以认为异方差性问题不存在；如果 $t>t_{\alpha/2}(n-2)$，说明 x_i 与 $|e_i|$ 之间存在系统关系，异方差性问题存在。

例 4.3

参见参考文献 [8]，设某地区的居民收入与储蓄额的历史统计数据如表 4.1 所示。

(1) 用普通最小二乘法建立储蓄额 y 与居民收入 x 的回归方程，并画出残差散点图。

(2) 诊断该问题是否存在异方差。

解：(1) 首先用 SPSS 软件建立 y 对 x 的普通最小二乘回归，决定系数 $R^2=0.912$，回归标准差 $\sigma^2=247.62$。方差分析表与回归系数表见输出结果 4.1。

输出结果 4.1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.844E7	1	1.844E7	300.732	.000[a]
	Residual	1778202.735	29	61317.336		
	Total	2.022E7	30			

a. Predictors: (Constant), x.

b. Dependent Variable: y.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-648.124	118.163		-5.485	.000
	x	.085	.005	.955	17.342	.000

a. Dependent Variable: y.

回归方程为 $\hat{y}=-648.124+0.085x$，残差 e_i 列在表 4.1 中，残差图如图 4.1 所示。

专业课辅导+真题资料+微信公众号：432应统书院

表 4.1

序号	储蓄额 y(万元)	居民收入 x(万元)	x_i 等级	残差 e_i	$\lvert e_i \rvert$	残差 $\lvert e_i \rvert$ 等级	d_i	d_i^2
1	264	8 777	1	169	169	16	−15	225
2	105	9 210	2	−26.6	26.6	3	−1	1
3	90	9 954	3	−104.6	104.6	7	−4	16
4	131	10 508	4	−110.5	110.5	8	−4	16
5	122	10 979	5	−159.4	159.4	15	−10	100
6	107	11 912	6	−253.4	253.4	23	−17	289
7	406	12 747	7	−25.1	25.1	2	5	25
8	503	13 499	8	8.2	8.2	1	7	49
9	431	14 269	9	−129	129	9	0	0
10	588	15 522	10	−78	78	4	6	36
11	898	16 730	11	129.7	129.7	10	1	1
12	950	17 663	12	102.7	102.7	6	6	36
13	779	18 575	13	−145.5	145.5	14	−1	1
14	819	19 635	14	−195.3	195.3	19	−5	25
15	1 222	21 163	15	78.4	78.4	5	10	100
16	1 702	22 880	16	413	413	28	−12	144
17	1 578	24 127	17	183.4	183.4	18	−1	1
18	1 654	25 604	18	134.4	134.4	11	7	49
19	1 400	26 500	19	−195.5	195.5	20	−1	1
20	1 829	27 670	21	134.4	134.4	12	9	81
21	2 200	28 300	23	452.1	452.1	29	−6	36
22	2 017	27 430	20	342.8	342.8	27	−7	49
23	2 105	29 560	24	250.4	250.4	22	2	4
24	1 600	28 150	22	−135.2	135.2	13	9	81
25	2 250	32 100	25	180.4	180.4	17	8	64
26	2 420	32 500	26	316.5	316.5	25	1	1
27	2 570	35 250	28	233.7	233.7	21	7	49
28	1 720	33 500	27	−468.2	468.2	30	−3	9
29	1 900	36 000	29	−499.8	499.8	31	−2	4
30	2 100	36 200	30	−316.7	316.7	26	4	16
31	2 300	38 200	31	−286.1	286.1	24	7	49

图 4.1 残差图

专业课辅导+真题资料+微信公众号：432应统书院

从残差图看出，误差项具有明显的异方差性，误差随着 x 的增加呈现出增加的态势。

（2）计算等级相关系数。由表 4.1 得，$\sum d_i^2 = 1\,558$，代入（4.1）式得

$$r_s=1-\frac{6}{31\times(31^2-1)}\times 1\,558=0.685\,9$$

将 $r_s=0.685\,9$ 代入（4.2）式，得

$$t=\frac{\sqrt{31-2}\times 0.685\,9}{\sqrt{1-0.685\,9^2}}=5.076$$

给定显著性水平 $\alpha=0.05$，自由度 $n-2=31-2=29$，查得临界值 $t_{0.025}(29)=2.045$，由 $t=5.076>2.045$，认为残差绝对值 $|e_i|$ 与自变量 x_i 显著相关，误差项存在异方差。

等级相关系数的检验可以用 SPSS 软件实现，首先用 Transform→Compute 命令算出残差绝对值 $|e_i|=\text{abs}(e_i)$，然后在 Analyze 下拉菜单中点击 Correlate→Bivariate→Spearman，计算 $|e_i|$ 与 x_i 的等级相关系数，如输出结果 4.2 所示。

输出结果 4.2

Correlations

			x	abse
Spearman's rho	x	Correlation Coefficient	1.000	.686**
		Sig. (2-tailed)	.	.000
		N	31	31
	abse	Correlation Coefficient	.686**	1.000
		Sig. (2-tailed)	.000	.
		N	31	31

**. Correlation is significant at the 0.01 level (2-tailed).

同样得等级相关系数 $r_s=0.686$，P 值=2.055E－5，认为残差绝对值 $|e_i|$ 与自变量 x_i 显著相关，存在异方差。

计算残差绝对值 $|e_i|$ 与自变量 x_i 的相关性时采用 Spearman 等级相关系数，而不采用 Pearson 简单相关系数，这是由于等级相关系数可以反映非线性相关的情况，而简单相关系数不能如实反映非线性相关的情况。例如 x 与 y 的取值如表 4.2 所示。

表 4.2

序号	1	2	3	4	5	6	7	8	9	10
x	1	2	3	4	5	6	7	8	9	10
y	1	4	9	16	25	36	49	64	81	100

专业课辅导+真题资料+微信公众号：432应统书院

可以看出，y_i 与 x_i 之间的关系为 $y_i=x_i^2(i=1, 2, \cdots, 10)$，具有完全的曲线相关。容易计算出 y 与 x 的简单相关系数 $r=0.974\,6$，而 y 与 x 的等级相关系数 $r_s=1$。与简单相关系数相比，等级相关系数可以更准确地反映非线性相关的情况。等级相关系数可以如实反映单调递增或单调递减趋势的变量间的相关性，而简单相关系数只适宜衡量直线趋势的变量间的相关性。

二、一元加权最小二乘估计

当我们所研究的问题具有异方差性时，就违反了线性回归模型的基本假定。此时，就不能用普通最小二乘法进行参数估计，必须寻求适当的补救方法，对原来的模型进行变换，使变换后的模型满足同方差性假设，然后进行模型参数的估计，即可得到理想的回归模型。消除异方差性的方法通常有加权最小二乘法、BOX-COX 变换法、方差稳定性变换法（参见参考文献［2］）。下面结合例 4.3 介绍加权最小二乘法。加权最小二乘法（weighted least square，WLS）是一种最常用的消除异方差性的方法。

对一元线性回归方程来说，普通最小二乘法的离差平方和为

$$\begin{aligned} Q(\beta_0,\beta_1) &= \sum_{i=1}^{n}(y_i-E(y_i))^2 \\ &= \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \end{aligned} \tag{4.3}$$

其中，每个观测值的权数相同。在等方差的条件下，平方和中的每一项的地位是相同的。然而，在异方差的条件下，平方和中的每一项的地位是不同的，误差项方差 σ_i^2 大的项，在（4.3）式平方和中的作用就偏大，因而普通最小二乘估计的回归线就被拉向方差大的项，而方差小的项的拟合程度就差。加权最小二乘法是在平方和中加入一个适当的权数 w_i，以调整各项在平方和中的作用。一元线性回归的加权最小二乘的离差平方和为

$$\begin{aligned} Q_w(\beta_0,\beta_1) &= \sum_{i=1}^{n}w_i(y_i-E(y_i))^2 \\ &= \sum_{i=1}^{n}w_i(y_i-\beta_0-\beta_1 x_i)^2 \end{aligned} \tag{4.4}$$

其中，w_i 为给定的第 i 个观测值的权数。加权最小二乘估计就是寻找参数 β_0，β_1 的估计值 $\hat{\beta}_{0w}$，$\hat{\beta}_{1w}$，使（4.4）式的离差平方和 Q_w 达到极小。如果所有的权数相等，即 w_i 都等于某个常数，问题就成为普通最小二乘法。可以证明加权最小二乘估计为

$$\begin{cases} \hat{\beta}_{0w}=\bar{y}_w-\hat{\beta}_{1w}\bar{x}_w \\ \hat{\beta}_{1w}=\dfrac{\sum_{i=1}^{n}w_i(x_i-\bar{x}_w)(y_i-\bar{y}_w)}{\sum_{i=1}^{n}w_i(x_i-\bar{x}_w)^2} \end{cases} \tag{4.5}$$

专业课辅导+真题资料+微信公众号：432应统书院

其中，$\bar{x}_w = \frac{1}{\sum w_i} \sum w_i x_i$ 为自变量的加权平均；

$\bar{y}_w = \frac{1}{\sum w_i} \sum w_i y_i$ 为因变量的加权平均。

加权最小二乘估计的计算可以用 SPSS 软件完成。

在使用加权最小二乘法时，为了消除异方差性的影响，使（4.4）式中的各项地位相同，观测值的权数应该是观测值误差项方差的倒数，即

$$w_i = \frac{1}{\sigma_i^2}$$

其中，σ_i^2 为第 i 个观测值误差项的方差。所以误差项方差较大的观测值接受较小的权数；误差项方差较小的观测值接受较大的权数。

在实际问题的研究中，误差项的方差 σ_i^2 通常是未知的，但是，当误差项方差随自变量水平以系统的形式变化时，我们可以利用这种关系。如，已知误差项方差 σ_i^2 与 x_i^2 成比例，那么 $\sigma_i^2 = kx_i^2$，其中 k 为比例系数。

权数 w_i 为

$$w_i = \frac{1}{kx_i^2}$$

因为比例系数 k 在参数估计中可以消去，所以可以直接使用权数

$$w_i = \frac{1}{x_i^2}$$

在社会、经济研究中，经常会遇到这种特殊的权数，即误差项方差与 x 的幂函数 x^m 成比例，其中，m 为待定的未知参数。

此时权函数为

$$w_i = \frac{1}{x_i^m} \tag{4.6}$$

三、寻找最优权函数

利用 SPSS 软件可以确定（4.6）式幂指数 m 的最优取值。对例 4.3 的数据，依次点击 Analyze→Regression→Weight Estimation 进入估计权函数对话框，默认的幂指数 m 的取值为 −2.0，−1.5，−1.0，−0.5，0，0.5，1.0，1.5，2.0，这一默认值可以更改。先将因变量 y 与自变量 x 选入各自的变量框，再把 x 选入 Weight 变量框，幂指数（Power）取默认值，得输出结果 4.3（格式略有变动）。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 4.3

Log-Likelihood Values[b]

Power	-2.000	-224.259
	-1.500	-221.515
	-1.000	-218.832
	-.500	-216.252
	.000	-213.856
	.500	-211.773
	1.000	-210.186
	1.500	-209.316[a]
	2.000	-209.380

a. The corresponding power is selected for further analysis because it maximizes the log-likelihood function.

b. Dependent variable: y, source variable: x.

Best Model Statistics

Model Description

Dependent Variable		y
Independent Variables	1	x
Weight	Source	x
	Power Value	1.500

Model: MOD_1.

Model Summary

Multiple R	.967
R Square	.936
Adjusted R Square	.934
Std. Error of the Estimate	.125
Log-likelihood Function Value	-209.316

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	6.655	1	6.655	423.741	.000
Residual	.455	29	.016		
Total	7.110	30			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta	Std. Error		
(Constant)	-719.123	78.316			-9.182	.000
x	.088	.004	.967	.047	20.585	.000

专业课辅导+真题资料+微信公众号：432应统书院

根据以上输出结果，在 $m=1.5$ 时对数似然函数达到极大，因而幂指数 m 的最优取值为 $m=1.5$。加权最小二乘法的 $R^2=0.936$，F 值$=423.741$；而普通最小二乘法的 $R^2=0.912$，F 值$=300.732$。这说明加权最小二乘估计的效果好于普通最小二乘估计的效果。

比较图 4.1 的普通残差图和图 4.2 的加权最小二乘残差图，我们可能没有看出两张图之间的差异，这是否表明加权最小二乘回归没有达到效果？现在进一步计算出 $n=31$ 组数据的普通残差 e_i 和加权最小二乘残差 e_{iw}，比较两者数值的差异，由此来说明加权最小二乘法的作用。两种残差的具体数值列在了表 4.3 中（鉴于和表 4.1 的连贯性，残差保留一位小数）。

图 4.2　加权最小二乘残差图

表 4.3　两种残差的数值

	序号	y_i	x_i	w_i	e_i	e_{iw}
小方差组	1	264	8 777	1.216 1E－06	169	211.3
	2	105	9 210	1.131 4E－06	－26.6	14.3
	3	90	9 954	1.006 9E－06	－104.6	－66.1
	4	131	10 508	9.283 7E－07	－110.5	－73.9
	5	122	10 979	8.692 7E－07	－159.4	－124.3
	6	107	11 912	7.691 7E－07	－253.4	－221.3
	7	406	12 747	6.948 5E－07	－25.1	4.3
	8	503	13 499	6.376 0E－07	8.2	35.1
	9	431	14 269	5.866 9E－07	－129	－104.6
	10	588	15 522	5.171 0E－07	－78	－57.7

专业课辅导+真题资料+微信公众号：432应统书院

续前表

	序号	y_i	x_i	w_i	e_i	e_{iw}
中等方差组	11	898	16 730	4.621 2E－07	129.7	146
	12	950	17 663	4.259 9E－07	102.7	116
	13	779	18 575	3.950 1E－07	－145.5	－135.2
	14	819	19 635	3.634 6E－07	－195.3	－188.4
	15	1 222	21 163	3.248 1E－07	78.4	80.2
	16	1 702	22 880	2.889 5E－07	413	409.3
	17	1 578	24 127	2.668 4E－07	183.4	175.6
	18	1 654	25 604	2.440 8E－07	134.4	121.7
	19	1 400	26 500	2.318 1E－07	－195.5	－211.1
	20	1 829	27 670	2.172 6E－07	134.4	115.1
	21	2 200	28 300	2.100 5E－07	452.1	430.7
大方差组	22	2 017	27 430	2.201 2E－07	342.8	324.2
	23	2 105	29 560	1.967 6E－07	250.4	224.9
	24	1 600	28 150	2.117 3E－07	－135.2	－156.1
	25	2 250	32 100	1.738 8E－07	180.4	146.5
	26	2 420	32 500	1.706 8E－07	316.5	281.3
	27	2 570	35 250	1.511 0E－07	233.7	189.5
	28	1 720	33 500	1.630 9E－07	－468.2	－506.6
	29	1 900	36 000	1.464 0E－07	－499.8	－546.4
	30	2 100	36 200	1.451 9E－07	－316.7	－364
	31	2 300	38 200	1.339 4E－07	－286.1	－339.9

这个例子共有 31 对数据，把数据分为 3 组，第 1～10 对数据为第 1 组，是小方差组；第 11～21 对数据为第 2 组，是中等方差组；第 22～31 对数据为第 3 组，是大方差组。

从表 4.3 中看到，第 1 组 10 个普通残差 e_i 中有 8 个是负值，说明普通残差图中小残差有整体的负偏。而 10 个加权残差 e_{iw} 中只有 6 个是负值，说明加权残差针对小残差整体负偏的情况已经有了明显的改进。10 个普通残差中绝对值最大的是 $e_6=-253.4$，加权回归后改善为 $e_{6w}=-221.3$。图形是对数值的直观展示，从两张残差图上也可以看出这个现象。

第 3 组 10 个普通残差 e_i 和加权残差 e_{iw} 的正负性相同，正负值各有 5 个，说明普通最小二乘和加权最小二乘对大残差项拟合得都好。仔细观察这组的两种残差还是能发现区别的，10 个普通残差中绝对值最大的是 $e_{29}=-499.8$，加权回归后成为 $e_{29}=-546.4$。不是像小残差组那样得到改善，而是误差变得更大。其道理也很简

专业课辅导+真题资料+微信公众号：432应统书院

单，加权最小二乘估计照顾小残差项是以牺牲大残差项为代价的，有得必有失，也是有一定局限性的。

从上面的分析看到，当回归模型存在异方差时，加权最小二乘估计只是对普通最小二乘估计的改进，这种改进有可能是细微的，不能理解为加权最小二乘估计一定会得到与普通最小二乘估计截然不同的回归方程，或者一定有大幅度的改进。实际上，可以构造出这样的数据，其回归模型具有明显的异方差性，但是普通最小二乘与加权最小二乘所得的回归方程却完全一样。

另外，加权最小二乘以牺牲大方差项的拟合效果为代价改善了小方差项的拟合效果，这也并不总是研究者所需要的。在社会经济现象中，通常变量取值大时方差也大，在以经济总量为研究目标时，更关心的是变量取值大的项，而普通最小二乘恰好能满足这个要求。所以在这样一些特定场合下，即使数据存在异方差，也可以选择使用普通最小二乘估计。

SPSS 软件的加权最小二乘具有普通最小二乘的很多功能，包括共线性诊断、异常值判定、自相关分析、区间预测等，读者在学过这些内容之后很容易使用这些功能。

4.3　多元加权最小二乘估计

一、多元加权最小二乘法

对于一般的多元线性回归模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i,\quad i=1,2,\cdots,n$$

当误差项 ε_i 存在异方差时，加权离差平方和为

$$Q_w=\sum_{i=1}^{n}w_i(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2 \tag{4.7}$$

其中，w_i 为给定的第 i 个观测值的权数。加权最小二乘估计就是寻找参数 β_0，β_1，β_2，…，β_p 的估计值 $\hat{\beta}_{0w}$，$\hat{\beta}_{1w}$，$\hat{\beta}_{2w}$，…，$\hat{\beta}_{pw}$，使（4.7）式的 Q_w 达到极小。记

$$\boldsymbol{W}=\begin{bmatrix} w_1 & & & \vdots \\ & w_2 & & \\ & & \ddots & \\ \vdots & & & w_n \end{bmatrix}$$

可以证明，加权最小二乘估计的矩阵可表达为

$$\hat{\boldsymbol{\beta}}_w=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \tag{4.8}$$

专业课辅导+真题资料+微信公众号：432应统书院

二、权函数的确定方法

多元线性回归有多个自变量，通常取权函数 W 为某个自变量 $x_j(j=1, 2, \cdots, p)$ 的幂函数，即 $W=x_j^m$，在 x_1，x_2，…，x_p 这 p 个自变量中，应该取哪一个自变量呢？只需计算每个自变量 x_j 与普通残差的等级相关系数，选取等级相关系数最大的自变量构造权函数。

例 4.4

续例 3.2，研究北京市各经济开发区经济发展与招商投资的关系，因变量 y 为各开发区的销售收入（百万元），选取两个自变量：x_1 为截至 1998 年底各开发区累计招商数目，x_2 为招商企业注册资本（百万元）。

计算出普通残差的绝对值 ABSE＝$|e_i|$ 与 x_1，x_2 的等级相关系数，见输出结果 4.4。

输出结果 4.4

Correlations

			X1	X2	ABSE
Spearman's rho	X1	Correlation Coefficient	1.000	.432	.443
		Sig. (2-tailed)	.	.108	.098
		N	15	15	15
	X2	Correlation Coefficient	.432	1.000	.721**
		Sig. (2-tailed)	.108	.	.002
		N	15	15	15
	ABSE	Correlation Coefficient	.443	.721**	1.000
		Sig. (2-tailed)	.098	.002	.
		N	15	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

从输出结果 4.4 中看出，残差绝对值与自变量 x_1 的相关系数为 $r_{e1}=0.443$，与自变量 x_2 的相关系数为 $r_{e2}=0.721$，因而选 x_2 构造权函数。

仿照例 4.3，用 Weight Estimate 估计幂指数 m，得 m 的最优值为 $m=2$，由于是在默认范围［－2，2］的边界，因而应该扩大范围重新计算。

取 m 从 1 到 5，步长仍为 0.5，得 m 的最优值为 $m=2.5$，部分输出结果如输出结果 4.5 所示。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 4.5

Model Summary

Multiple R	.922
R Square	.849
Adjusted R Square	.824
Std. Error of the Estimate	.032
Log-likelihood Function Value	-102.683

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	.071	2	.035	33.843	.000
Residual	.013	12	.001		
Total	.084	14			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta	Std. Error		
(Constant)	-266.962	106.742			-2.501	.028
X1	1.696	.404	.587	.140	4.195	.001
X2	.470	.149	.441	.140	3.150	.008

根据以上输出结果，加权最小二乘的 $R^2=0.849$，F 值$=33.843$；而普通最小二乘的 $R^2=0.842$，F 值$=31.96$。这说明对本例的数据加权最小二乘估计的拟合效果略好于普通最小二乘的效果，选用加权最小二乘估计是正确的。

加权最小二乘的回归方程是

$$\hat{y}=-266.962+1.696x_1+0.47x_2$$

普通最小二乘的回归方程为

$$\hat{y}=-327.039+2.036x_1+0.468x_2$$

4.4　自相关性问题及其处理

无论是在介绍一元还是多元线性回归模型时，我们总假定其随机误差项是不相关的，即

$$\mathrm{cov}(\varepsilon_i,\varepsilon_j)=0,\quad i\neq j \tag{4.9}$$

专业课辅导+真题资料+微信公众号：432应统书院

（4.9）式表示不同时点的误差项之间不相关。如果一个回归模型不满足（4.9）式，即 $\mathrm{cov}(\varepsilon_i, \varepsilon_j)\neq 0$，则称随机误差项之间存在自相关现象。这里的自相关现象不是指两个或两个以上的变量之间的相关关系，而是指一个变量前后期数值之间的相关关系。

本节主要讨论自相关现象产生的背景和原因，自相关现象对回归分析带来的影响，诊断自相关是否存在的方法，以及如何克服自相关现象产生的影响。

一、自相关性产生的背景和原因

在实际问题的研究中，经常遇到时间序列出现正的序列相关的情形。产生序列自相关的背景及其原因通常有以下几个方面：

（1）遗漏关键变量时会产生序列的自相关性。在回归分析的建模过程中，如果忽略了一个或几个重要的变量，而这些遗漏的关键变量在时间顺序上的影响是正相关的，回归模型中的误差项就会具有明显的正相关，这是因为误差包含了遗漏变量的影响。例如，我们利用新中国成立以来的有关统计数据建立我国居民消费模型时，居民可支配收入是一个重要的变量，它对居民的消费产生重要的影响，如果把这个重要的变量漏掉了，就可能使得误差项正自相关，因为居民可支配收入对居民消费的影响很可能是时间上正相关的。

（2）经济变量的滞后性会给序列带来自相关性。许多经济变量都会产生滞后影响，例如物价指数、基建投资、国民收入、消费、货币发行量等都有一定的滞后性。如前期消费额对后期消费额一般会有明显的影响。有时，经济变量的这种滞后表现出一种不规则的循环波动，当经济情况处于衰退的谷底时，经济扩张期随之开始，这时，大多数经济时间序列上升得快一些。在经济扩张期，经济时间序列内部有一种内在的冲力，受此影响，时间序列一直上升到循环的顶点，在顶点时刻，经济收缩随之开始。因此，在这样的时间序列数据中，顺序观测值之间的相关现象是很自然的。经济现象中的自相关一般是正的。

（3）采用错误的回归函数形式也可能引起自相关性。例如，假定某实际问题的正确回归函数应由指数形式

$$y=\beta_0\exp(\beta_1 x+\varepsilon)$$

来表示，但研究者误用线性回归模型

$$y=\beta_0+\beta_1 x+\varepsilon'$$

表示，这时，误差项 ε' 也表现为自相关性。

（4）蛛网现象（cobweb phenomenon）可能带来序列的自相关性。蛛网现象是微观经济学中研究商品市场运行规律用的一个名词，它表示某种商品的供给量因受前一期价格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。规律性的作用使得所用回归模型的误差项不再是随机的，而产生了某种自相关

专业课辅导+真题资料+微信公众号：432应统书院

性。例如，许多农产品的供给呈现出蛛网现象，即供给量受前一期的价格的影响。这样，今年某种产品的生产和供给计划取决于上一年的价格。因此，农产品的供给函数可表示为

$$S_t=\beta_0+\beta_1 P_{t-1}+\varepsilon_t,\quad t=1,2,\cdots,n$$

其中，S_t 为 t 期农产品供给量；P_{t-1} 为 $t-1$ 期农产品的价格。

假定 t 期的农产品价格 P_t 低于 $t-1$ 期的农产品价格 P_{t-1}，那么，$t+1$ 期的农产品供给量将低于 t 期的供给量。在这种情况下，干扰项 ε_t 不能预测，成为随机的，因为农民在第 t 年多生产了，很可能导致他们在第 $t+1$ 年少生产。比如我们都有过上年某种农产品的价格低，本年这种农产品就供应紧张、价格上涨的经验。

(5) 因对数据加工整理而导致误差项之间产生自相关性。在回归分析建模中，经常要对原始数据进行一些处理，如在具有季节性时序资料的建模中，我们常常要消除季节性，对数据作修匀处理。但如果采用了不恰当的差分变换，也会带来序列的自相关性。

自相关问题在时序资料的建模中会经常碰到，在截面样本数据中有时也会存在。大多数经济时间序列由于受经济波动规律的作用，一般随着时间的推移有一种向下或向上变动的趋势，所以，随机误差项 ε_t 一般表现为正自相关情形。负自相关的情形有时也会出现，但并不多见。

二、自相关性带来的问题

当一个线性回归模型的随机误差项存在序列相关时，就违背了线性回归方程的基本假设，如果仍然直接用普通最小二乘法估计未知参数，将会产生严重后果，一般情况下，序列相关性会带来下列问题：

(1) 参数的估计值不再具有最小方差线性无偏性。

(2) 均方误差（MSE）可能严重低估误差项的方差。

(3) 容易导致对 t 值评价过高，常用的 F 检验和 t 检验失效。如果忽视这一点，可能导致得出回归参数统计检验为显著，但实际上并不显著的严重错误结论。

(4) 当存在序列相关时，$\hat{\boldsymbol{\beta}}$ 仍然是 $\boldsymbol{\beta}$ 的无偏估计量，但在任一特定的样本中，$\hat{\boldsymbol{\beta}}$ 可能严重歪曲 $\boldsymbol{\beta}$ 的真实情况，即最小二乘估计量对抽样波动变得非常敏感。

(5) 如果不加处理地运用普通最小二乘法估计模型参数，用此模型进行预测和结构分析将会带来较大的方差甚至错误的解释。

三、自相关性的诊断

由于随机扰动项存在序列相关给普通最小二乘法的应用带来了非常严重的后果，因此，如何诊断随机扰动项是否存在序列相关就成为一个极其重要的问题。下

专业课辅导+真题资料+微信公众号：432应统书院

面介绍两种主要的诊断方法。

1. 图示检验法

图示检验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参数，求出残差项 e_t，e_t 作为随机项 ε_t 的真实值的估计值，再描绘 e_t 的散点图，根据 e_t 的相关性来判断随机项 ε_t 的序列相关性。残差 e_t 的散点图通常有两种绘制方式。

（1）绘制 e_t，e_{t-1} 的散点图。用（e_t，e_{t-1}）($t=2, 3, \cdots, n$）作为散布点绘图。如果大部分点落在第Ⅰ，Ⅲ象限，表明随机扰动项 ε_t 存在正的序列相关，如图 4.3（a)所示；如果大部分点落在第Ⅱ，Ⅳ象限，表明随机扰动项 ε_t 存在负相关，如图 4.3（b）所示。

图 4.3

（2）按照时间顺序绘制回归残差项 e_t 的图形。如果 e_t($t=1, 2, \cdots, n$）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，就可断言 e_t 存在相关，表明 ε_t 存在序列相关。如果 e_t 随着 t 的变化逐次变化并不断地改变符号，如图 4.4（a）所示，那么随机扰动项 ε_t 存在负的序列相关，这种现象称为蛛网现象。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，则表明随机扰动项 ε_t 存在正的序列相关，如图 4.4（b）所示。

2. 自相关系数法

误差序列 ε_1，ε_2，…，ε_n 的自相关系数定义为

$$\rho=\frac{\sum_{t=2}^{n}\varepsilon_t\varepsilon_{t-1}}{\sqrt{\sum_{t=2}^{n}\varepsilon_t^2}\sqrt{\sum_{t=2}^{n}\varepsilon_{t-1}^2}} \tag{4.10}$$

专业课辅导+真题资料+微信公众号：432应统书院

图 4.4

自相关系数 ρ 的取值范围是 [−1，1]，当 ρ 接近 1 时，表明误差序列存在正相关，当 ρ 接近 −1 时，表明误差序列存在负相关。在实际应用中，误差序列 ε_1，ε_2，…，ε_n 的真实值是未知的，需要用其估计值 e_t 代替，得自相关系数的估计值为

$$\hat{\rho}=\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\sum_{t=2}^{n} e_t^2}\sqrt{\sum_{t=2}^{n} e_{t-1}^2}} \tag{4.11}$$

$\hat{\rho}$ 作为自相关系数 ρ 的估计值与样本量有关，需要做统计显著性检验才能确定自相关性的存在，通常采用下面介绍的 DW 检验代替对 $\hat{\rho}$ 的检验。

3. DW 检验

DW 检验是杜宾（J. Durbin）和沃特森（G. S. Watson）于 1951 年提出的适用于小样本的一种检验方法。DW 检验只能用于检验随机扰动项具有一阶自回归形式的序列相关问题。这种检验方法是建立计量经济学模型中最常用的方法，一般的计算机软件都可以计算出 DW 值。

随机扰动项的一阶自回归形式为

$$\varepsilon_t=\rho\varepsilon_{t-1}+u_t \tag{4.12}$$

为了检验序列的相关性，构造的假设是

$$H_0:\rho=0$$

为了检验上述假设，构造 DW 统计量首先要求算出回归估计式的残差 e_t，定义 DW 统计量为

$$\mathrm{DW}=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=2}^{n} e_t^2} \tag{4.13}$$

其中，$e_t=y_t-\hat{y}_t(t=1,2,\cdots,n)$。

专业课辅导+真题资料+微信公众号：432应统书院

下面我们推导出 DW 值的取值范围。由（4.13）式有

$$\mathrm{DW}=\frac{\sum_{t=2}^{n}e_t^2+\sum_{t=2}^{n}e_{t-1}^2-2\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2} \tag{4.14}$$

如果认为 $\sum_{t=2}^{n}e_t^2$ 与 $\sum_{t=2}^{n}e_{t-1}^2$ 近似相等，则由（4.14）式得

$$\mathrm{DW}\approx 2\left[1-\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2}\right] \tag{4.15}$$

同样，在认为 $\sum_{t=2}^{n}e_t^2$ 与 $\sum_{t=2}^{n}e_{t-1}^2$ 近似相等时，则由（4.11）式得

$$\hat{\rho}\approx\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2} \tag{4.16}$$

因此，（4.15）式可以写为

$$\mathrm{DW}\approx 2(1-\hat{\rho}) \tag{4.17}$$

因而 DW 值与 $\hat{\rho}$ 的对应关系如表 4.4 所示。

表 4.4

$\hat{\rho}$	DW	误差项的自相关性
−1	4	完全负自相关
(−1，0)	(2，4)	负自相关
0	2	无自相关
(0，1)	(0，2)	正自相关
1	0	完全正自相关

由上述讨论可知 DW 的取值范围为：$0\leqslant \mathrm{DW}\leqslant 4$。

根据样本量 n 和解释变量的数目 k（这里包括常数项）查 DW 分布表，得临界值 d_L 和 d_U，然后依下列准则考察计算得到的 DW 值，以决定模型的自相关状态，见表 4.5。

表 4.5

$0\leqslant \mathrm{DW}\leqslant d_L$	误差项 ε_1，ε_2，…，ε_n 间存在正自相关
$d_L<\mathrm{DW}\leqslant d_U$	不能判定是否有自相关
$d_U<\mathrm{DW}<4-d_U$	误差项 ε_1，ε_2，…，ε_n 间无自相关
$4-d_U\leqslant \mathrm{DW}<4-d_L$	不能判定是否有自相关
$4-d_L\leqslant \mathrm{DW}\leqslant 4$	误差项 ε_1，ε_2，…，ε_n 间存在负自相关

专业课辅导+真题资料+微信公众号：432应统书院

上述判别准则结合图 4.5 容易记忆。由图 4.5 可看到 DW＝2 的左右有一个较大的无自相关区，所以，通常当 DW 的值在 2 左右时，则无须查表，即可放心地认为模型不存在序列的自相关性。

图 4.5

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，只有增大样本量或选取其他方法。

（2）DW 统计量的上、下界表要求 $n>15$，这是因为样本如果再小，利用残差就很难对自相关的存在性作出比较正确的诊断。

（3）DW 检验不适合随机项具有高阶序列相关的检验。

四、自相关问题的处理方法

当一个回归模型存在序列相关性时，首先要查明序列相关产生的原因。如果是回归模型选用不当，则应改用适当的回归模型；如果是缺少重要的自变量，则应增加自变量；如果以上两种方法都不能消除序列相关性，则需采用迭代法、差分法等方法处理。

1. 迭代法

以一元线性回归模型为例，设一元线性回归模型的误差项存在一阶自相关

$$y_t=\beta_0+\beta_1 x_t+\varepsilon_t \tag{4.18}$$

$$\varepsilon_t=\rho\varepsilon_{t-1}+u_t \tag{4.19}$$

$$\begin{cases} E(u_t)=0, \quad t=1,2,\cdots,n \\ \operatorname{cov}(u_t,u_s)=\begin{cases}\sigma^2, t=s \\ 0, \ t\neq s\end{cases} \quad t,s=1,2,\cdots,n \end{cases} \tag{4.20}$$

（4.19）式表明误差项 ε_t 存在一阶自相关，（4.20）式表明 u_t 满足关于随机扰

专业课辅导+真题资料+微信公众号：432应统书院

动项的基本假设。

根据回归模型（4.18）式，有

$$y_{t-1}=\beta_0+\beta_1 x_{t-1}+\varepsilon_{t-1} \tag{4.21}$$

将（4.21）式两端乘以 ρ，用（4.18）式减去乘以 ρ 的（4.21）式，则有

$$(y_t-\rho y_{t-1})=(\beta_0-\rho\beta_0)+\beta_1(x_t-\rho x_{t-1})+(\varepsilon_t-\rho\varepsilon_{t-1}) \tag{4.22}$$

在（4.22）式中，令

$$\begin{aligned} y_t'&=y_t-\rho y_{t-1} \\ x_t'&=x_t-\rho x_{t-1} \\ \beta_0'&=\beta_0(1-\rho),\beta_1'=\beta_1 \end{aligned} \tag{4.23}$$

于是（4.22）式变成

$$y_t'=\beta_0'+\beta_1' x_t'+u_t \tag{4.24}$$

模型（4.24）式有独立随机误差项，它满足线性回归模型的基本假设，用普通最小二乘法估计的参数估计量具有通常的优良性。

由于（4.23）式中的自相关系数 ρ 是未知的，需要用（4.17）式对 ρ 作估计。根据（4.17）式，$\hat{\rho}\approx 1-\frac{1}{2}\mathrm{DW}$，计算出 ρ 的估计值 $\hat{\rho}$ 后，代入（4.23）式，计算变换因变量 y_t' 与变换自变量 x_t'，然后用（4.24）式作普通最小二乘回归。如果误差项确实是（4.19）式的一阶自相关模型，那么通过以上变换，模型（4.24）式已经消除自相关，迭代法到此结束。

在实际问题中，有时误差项并不是简单的一阶自相关，而是更复杂的自相关形式，（4.24）式的误差项 u_t 可能仍然存在自相关，这就需要进一步对（4.24）式的误差项 u_t 做 DW 检验，以判断 u_t 是否存在自相关。如果检验表明误差项 u_t 不存在自相关，迭代法到此结束。如果检验表明误差项 u_t 存在自相关，那么对回归模型（4.24）式重复用迭代法，这个过程可能要重复几次，直至最终消除误差项的自相关。这种通过迭代消除自相关的过程正是迭代法名称的由来。

2. 差分法

差分法就是用增量数据代替原来的样本数据，将原来的回归模型变为差分形式的模型。一阶差分法通常适用于原模型存在较高程度的一阶自相关的情况。

在迭代法（4.22）式中，当 $\rho=1$ 时，得

$$(y_t-y_{t-1})=\beta_1(x_t-x_{t-1})+(\varepsilon_t-\varepsilon_{t-1}) \tag{4.25}$$

以 $\Delta y_t=y_t-y_{t-1}$，$\Delta x_t=x_t-x_{t-1}$ 代之，得

$$\Delta y_t=\beta_1\Delta x_t+u_t \tag{4.26}$$

专业课辅导+真题资料+微信公众号：432应统书院

(4.26) 式不存在序列的自相关，它是以差分数据 Δy_t 和 Δx_t 为样本的回归方程。

对 (4.26) 式这样不带有常数项的回归方程仍用最小二乘法，但它与前面的带有常数项的情形稍有不同，它是回归直线过原点的回归方程。根据第 2 章习题 2.2 得

$$\hat{\beta}_1=\frac{\sum_{t=2}^{n}\Delta y_t\Delta x_t}{\sum_{t=2}^{n}\Delta x_t^2}$$

一阶差分法的应用条件是自相关系数 $\rho=1$，在实际应用中，ρ 接近 1 时就采用差分法而不用迭代法。这有两个原因：(1) 迭代法需要用样本估计自相关系数 ρ，对 ρ 的估计误差会影响迭代法的使用效率；(2) 差分法比迭代法简单，人们在建立时序数据的回归模型时，更习惯于用差分法。

五、自相关实例分析

例 4.5

续例 2.2，表 2.2 的数据是时间序列数据，因变量 y 为城镇家庭平均每人全年消费性支出，自变量 x 为城镇家庭平均每人可支配收入。输出结果 2.4 中计算出 DW=0.3，查 DW 表，$n=20$，$k=2$，显著性水平 $\alpha=0.05$，得 $d_L=1.20$，$d_U=1.41$。由 DW=0.3<1.2，可知残差存在正的自相关。由输出结果 2.4 的图中可以看到残差有明显的趋势变动，表明误差项存在自相关。自相关系数 $\hat{\rho}\approx1-\frac{1}{2}\mathrm{DW}=1-\frac{1}{2}\times0.3=0.85$，说明误差项存在高度自相关。

(1) 用迭代法消除自相关。依照 (4.23) 式计算变换因变量 y_t' 与变换自变量 x_t'，见表 4.6. 然后用 y_t' 对 x_t' 作普通最小二乘回归，计算结果见输出结果 4.6，残差 e_t' 列在表 4.6 中。从输出结果 4.6 中看到，新回归残差 e_t' 的 DW=2.327，查 DW 表，$n=19$，$k=2$，显著性水平 $\alpha=0.05$，得 $d_L=1.18$，$d_U=1.40$。由于 $d_U<2.327<4-d_U$，因而 DW 值落入无自相关区域。误差项 u_t 的标准差 $\hat{\sigma}_u=71.419\,10$，小于 ε_t 的标准差 $\hat{\sigma}=153.822$。y_t' 对 x_t' 的回归方程为

$$y_t'=138.021+0.663x_t'$$

把 $y_t'=y_t-0.85y_{t-1}$，$x_t'=x_t-0.85x_{t-1}$ 代入，还原为原始变量的方程

$$\begin{aligned}\hat{y}_t&=138.021+0.85y_{t-1}+0.663(x_t-0.85x_{t-1})\\&=138.021+0.85y_{t-1}+0.663x_t-0.564x_{t-1}\end{aligned}$$

专业课辅导+真题资料+微信公众号：432应统书院

表 4.6

年份	序号	x_t	y_t	e_t	x'_t	y'_t	e'_t
1990	1	1 510.16	1 278.89	−234.076 94			
1991	2	1 700.6	1 453.8	−192.129 29	416.964	366.74	−47.548 88
1992	3	2 026.6	1 671.7	−201.837 58	581.09	435.97	−87.063 70
1993	4	2 577.4	2 110.8	−147.297 86	854.79	689.86	−14.518 89
1994	5	3 496.2	2 851.3	−48.290 20	1 305.41	1 057.12	54.174 19
1995	6	4 282.95	3 537.57	88.682 79	1 311.18	1 113.97	107.201 16
1996	7	4 838.9	3 919.5	82.456 86	1 198.393	912.57	−19.469 65
1997	8	5 160.3	4 185.6	124.160 22	1 047.235	854.03	22.142 99
1998	9	5 425.1	4 331.6	85.280 84	1 038.845	773.84	−52.488 06
1999	10	5 854	4 615.9	70.129 31	1 242.665	934.04	−27.332 91
2000	11	6 279.98	4 998	154.816 48	1 304.08	1 074.49	72.425 40
2001	12	6 859.6	5 309.01	61.144 52	1 521.617	1 060.71	−85.487 91
2002	13	7 702.8	6 029.92	193.344 96	1 872.14	1 517.26	138.816 36
2003	14	8 472.2	6 510.94	137.181 46	1 924.82	1 385.51	−27.837 79
2004	15	9 421.6	7 182.1	145.484 66	2 220.23	1 647.80	38.722 65
2005	16	10 493	7 942.88	158.229 29	2 484.64	1 838.10	53.832 73
2006	17	11 759.5	8 696.55	27.648 04	2 840.45	1 945.10	−74.916 01
2007	18	13 785.8	9 997.47	−86.164 14	3 790.225	2 605.40	−43.907 65
2008	19	15 780.76	11 242.85	−233.635 20	4 062.83	2 745.00	−84.927 32
2009	20	17 174.65	12 264.55	−185.128 23	3 761.004	2 708.13	78.183 28

输出结果 4.6

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.995[a]	.991	.990	71.41910	2.327

a. Predictors: (Constant), xite.

b. Dependent Variable: yite.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9471587.537	1	9471587.537	1856.924	.000[a]
	Residual	86711.694	17	5100.688		
	Total	9558299.231	18			

a. Predictors: (Constant), xite.

b. Dependent Variable: yite.

专业课辅导+真题资料+微信公众号：432应统书院

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	138.021	32.566		4.238	.001
	xd	.663	.015	.995	43.092	.000

a. Dependent Variable: yite.

（2）用一阶差分法消除自相关。首先计算差分 $\Delta y_t=y_t-y_{t-1}$，$\Delta x_t=x_t-x_{t-1}$，差分结果列在表 4.7 中，然后用 Δy_t 对 Δx_t 作过原点的最小二乘回归，计算结果见输出结果 4.7，残差 e'_t列在表 4.7 中。从输出结果 4.7 中看到，新回归残差 e'_t的 DW=2.322，查 DW 表，$n=19$，$k=2$，显著性水平 $\alpha=0.05$，得 $d_L=1.18$，$d_U=1.4$，由于 $d_U<2.322<4-d_U$，可知 DW 值落入无自相关区域。误差项 u_t 的标准差 $\hat{\sigma}_u=77.441\,39$，小于 ε_t 的标准差 $\hat{\sigma}=153.822$。Δy_t 对 Δx_t 的回归方程为

$$\Delta y_t=0.628\Delta x_t+60.796$$

将 $\Delta y_t=y_t-y_{t-1}$，$\Delta x_t=x_t-x_{t-1}$代入，还原为原始变量的方程

$$y_t=60.796+y_{t-1}+0.628(x_t-x_{t-1}) \tag{4.27}$$

表 4.7

年份	序号	x_t	y_t	e_t	Δx_t	Δy_t	e'_t
1990	1	1 510.16	1 278.89	−234.076 94			
1991	2	1 700.6	1 453.8	−192.129 29	190.44	174.91	−5.399 92
1992	3	2 026.6	1 671.7	−201.837 58	326	217.9	−47.483 10
1993	4	2 577.4	2 110.8	−147.297 86	550.8	439.1	32.639 50
1994	5	3 496.2	2 851.3	−48.290 20	918.8	740.5	103.094 30
1995	6	4 282.95	3 537.57	88.682 79	786.75	686.27	131.734 72
1996	7	4 838.9	3 919.5	82.456 86	555.95	381.93	−27.762 48
1997	8	5 160.3	4 185.6	124.160 22	321.4	266.1	3.603 71
1998	9	5 425.1	4 331.6	85.280 84	264.8	146	−80.975 91
1999	10	5 854	4 615.9	70.129 31	428.9	284.3	−45.659 90
2000	11	6 279.98	4 998	154.816 48	425.98	382.1	53.972 60
2001	12	6 859.6	5 309.01	61.144 52	579.62	311.01	−113.537 02
2002	13	7 702.8	6 029.92	193.344 96	843.2	720.91	130.948 48
2003	14	8 472.2	6 510.94	137.181 46	769.4	481.02	−62.626 97
2004	15	9 421.6	7 182.1	145.484 66	949.4	671.16	14.550 70
2005	16	10 493	7 942.88	158.229 29	1 071.4	760.78	27.607 35
2006	17	11 759.5	8 696.55	27.648 04	1 266.5	753.67	−101.941 26
2007	18	13 785.8	9 997.47	−86.164 14	2 026.3	1 300.92	−31.517 79
2008	19	15 780.76	11 242.85	−233.635 20	1 994.96	1 245.38	−67.389 80
2009	20	17 174.65	12 264.55	−185.128 23	1 393.89	1 021.7	86.142 79

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 4.7

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.976[a]	.952	.949	77.44139	2.322

a. Predictors: (Constant), xdif.
b. Dependent Variable: ydif.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2033546.791	1	2033546.791	339.084	.000[a]
	Residual	101951.862	17	5997.168		
	Total	2135498.653	18			

a. Predictors: (Constant), xdif.
b. Dependent Variable: ydif.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60.796	33.243		1.829	.085
	xp	.628	.034	.976	18.414	.000

a. Dependent Variable: ydif.

（3）预测。SPSS 软件提供的三种方法可以直接保存回归预测值 $\hat{y}_t$ 和残差 e'_t。而使用迭代法和差分法则需要手工计算回归预测值 $\hat{y}_t$。计算 $\hat{y}_t$ 有两种方法，下面以迭代法为例说明回归预测值 $\hat{y}_t$ 和残差 e'_t 的计算方法。

在自相关回归中，回归预测值 $\hat{y}_t$ 不是使用估计值 $\hat{\beta}_0+\hat{\beta}_1x_t$ 计算，而是用（4.27）式计算，其一般性的公式为

$$\hat{y}_t=\hat{\beta}'_0+\hat{\rho}\,y_{t-1}+\hat{\beta}'_1(x_t-\hat{\rho}\,x_{t-1}) \tag{4.28}$$

计算出 $\hat{y}_t$ 后，再用 $y_t-\hat{y}_t$ 计算 e'_t，这里 e'_t 是随机误差项 u_t 的估计值。

另外一种计算 $\hat{y}_t$ 的方法是对 $\hat{\beta}_0+\hat{\beta}_1x_t$ 做修正。在误差项没有自相关时，我们实际上就是直接用估计值 $\hat{\beta}_0+\hat{\beta}_1x_t$ 作为回归预测值 $\hat{y}_t$。现在误差项存在自相关 $\varepsilon_t=\rho\varepsilon_{t-1}+u_t$，需要从残差 e_t 中提取出有用的信息对估计值 $\hat{\beta}_0+\hat{\beta}_1x_t$ 做修正，其中 $e_t=y_t-(\hat{\beta}_0+\hat{\beta}_1x_t)$ 是误差项 ε_t 的估计值。注意其中的系数估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是按照关系式 $\hat{\beta}_0=\hat{\beta}'_0/(1-\hat{\rho})$ 和 $\hat{\beta}_1=\hat{\beta}'_1$ 根据迭代法的参数估计值推算的，并不是普

专业课辅导+真题资料+微信公众号：432应统书院

通最小二乘的估计值，残差 e_t 也不是普通最小二乘的残差。计算过程如下：

$$\begin{aligned}&t=1\text{ 时，取 }\hat{y}_1=\hat{\beta}_0+\hat{\beta}_1x_1,e_1=y_1-(\hat{\beta}_0+\hat{\beta}_1x_1)\\&t\geqslant 2\text{ 时，取 }\hat{y}_t=\hat{\beta}_0+\hat{\beta}_1x_t+\hat{\rho}e_{t-1},e_t=y_t-(\hat{\beta}_0+\hat{\beta}_1x_t)\end{aligned}\tag{4.29}$$

例如，预计 2010 年城镇居民人均收入是 $x_{21}=19\,000$（元），则用迭代法计算的人均消费额的预测值是：

$$\begin{aligned}y_{21}&=138.021+0.85\times 12\,264.55+0.663(19\,000-0.85\times 17\,174.65)\\&=13\,481.11(\text{元})\end{aligned}$$

用第二种方法：

$$e_{20}=12\,264.55-(920.14+0.663\times 17\,174.65)=-42.382\,95(\text{元})$$

$$y_{21}=920.14+0.663\times 19\,000+0.85\times(-42.382\,95)=13\,481.11(\text{元})$$

两种方法得到的结果完全一样。

4.5　BOX-COX 变换

BOX-COX 变换是由 Box 与 Cox 在 1964 年提出的一种应用非常广泛的变换，它是对因变量 y 所作的如下变换：

$$y^{(\lambda)}=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln y, & \lambda=0\end{cases}$$

其中，λ 是待定参数。此变换要求 y 的各分量都大于 0，否则可用下面推广的 BOX-COX 变换：

$$y^{(\lambda)}=\begin{cases}\dfrac{(y+a)^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln(y+a), & \lambda=0\end{cases}$$

即先对 y 做平移，使得 $y+a$ 的各个分量都大于 0 后再做 BOX-COX 变换。

对于不同的 λ，所作的变换也不同，所以这是一个变换族。它包含了一些常用变换，如对数变换（$\lambda=0$），平方根变换（$\lambda=1/2$）和倒数变换（$\lambda=-1$）。

通过此变换，我们寻找合适的 λ，使得变换后

$$y^{(\lambda)}=\begin{pmatrix}y_1^{(\lambda)}\\ y_2^{(\lambda)}\\ \vdots\\ y_n^{(\lambda)}\end{pmatrix}\sim N_n(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$$

从而符合线性回归模型的各项假设：误差各分量等方差、不相关等。事实上，

专业课辅导+真题资料+微信公众号：432应统书院

BOX-COX 变换不仅可以处理异方差问题，还能处理自相关、误差非正态、回归函数非线性等情况。

经过计算可得 λ 的最大似然估计（参见参考文献［2］）：

$$L_{\max}(\lambda)=(2\pi e\hat{\sigma}_{\lambda}^{2})^{-\frac{n}{2}}|J|$$

其中，$\hat{\sigma}_{\lambda}^{2}=\frac{1}{n}\mathrm{SSE}(\lambda,y^{(\lambda)})$，$|J|=\prod_{i=1}^{n}\left|\frac{\mathrm{d}y_i^{(\lambda)}}{\mathrm{d}y_i}\right|=\prod_{i=1}^{n}y_i^{\lambda-1}$

令 $y^{(\lambda)}=\frac{y^{(\lambda)}}{|J|}$，对 $L_{\max}(\lambda)$ 取对数并略去与 λ 无关的常数项，可得：

$$\ln L_{\max}(\lambda)=-\frac{n}{2}\ln\mathrm{SSE}(\lambda,z^{(\lambda)})$$

为找出 λ，使得 $L_{\max}(\lambda)$ 达到最大，只需使 $\mathrm{SSE}(\lambda,z^{(\lambda)})$ 达到最小即可。它的解析解比较难找，通常是给出一系列 λ 的值，计算对应的 $\mathrm{SSE}(\lambda,z^{(\lambda)})$，取使得 $\mathrm{SSE}(\lambda,z^{(\lambda)})$ 达到最小的 λ 即可。用 SAS 软件可以直接给出各 λ 值对应的 $\ln L(\lambda)$。

1. 消除异方差

下面用 SAS 软件中的 BOX-COX 变换继续讨论例 3.2。

程序：

```
data eg3_2;
input x1 x2 y;
cards;
25    3547.79    553.96
20    896.34   208.55
6     750.32   3.10
1001  2087.05 2815.40
525   1639.31 1052.12
825   3357.70 3427.00
120   808.47   442.82
28    520.27   70.12
7     671.13   122.24
532   2863.32 1400.00
75    1160.00    464.00
40    862.75   7.50
187   672.99   224.18
122   901.76   538.94
```

专业课辅导+真题资料+微信公众号：432应统书院


```
74    3546.18     2442.79
;
proc transreg ss2;model boxcox(y)=identity(x1 x2);
/*调用 transreg 过程，对经 box-cox 变换后的新因变量关于 x1，x2 做线性回归*/
run;
```

输出结果 4.8

The TRANSREG Procedure

Transformation Information for BoxCox(y)

Lambda	R-Square	Log Like	
-3.00	0.07	-256.428	
-2.75	0.07	-240.492	
-2.50	0.07	-224.684	
-2.25	0.07	-209.034	
-2.00	0.08	-193.583	
-1.75	0.08	-178.389	
-1.50	0.09	-163.536	
-1.25	0.10	-149.141	
-1.00	0.11	-135.365	
-0.75	0.14	-122.426	
-0.50	0.21	-110.592	
-0.25	0.34	-100.178	
0.00	0.56	-91.555	
0.25	0.76	-85.437	
0.50 +	0.85	-83.468	<
0.75	0.86	-86.508	
1.00	0.84	-92.473	
1.25	0.81	-99.439	
1.50	0.78	-106.728	
1.75	0.75	-114.163	
2.00	0.72	-121.704	
2.25	0.70	-129.340	
2.50	0.68	-137.067	
2.75	0.66	-144.880	
3.00	0.65	-152.777	

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda

The TRANSREG Procedure Hypothesis Tests for BoxCox(y)

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	2	15212.43	7606.213	34.23	>= <.0001
Error	12	2666.49	222.207		
Corrected Total	14	17878.91			

The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.

Root MSE	14.90662	R-Square	0.8509
Dependent Mean	47.80761	Adj R-Sq	0.8260
Coeff Var	31.18043	Lambda	0.5000

Univariate Regression Table Based on the Usual Degrees of Freedom

Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Liberal p
Intercept	1	6.30117059	189.10	189.10	0.85	>= 0.3744
Identity(x1)	1	0.06130763	4435.08	4435.08	19.96	>= 0.0008
Identity(x2)	1	0.01658142	4094.64	4094.64	18.43	>= 0.0010

专业课辅导+真题资料+微信公众号：432应统书院

根据以上输出结果，本例中的 λ 取 0.5，$y^{(0.5)}$ 对 x_1，x_2 的回归方程为

$$\hat{y}^{(0.5)}=6.3012+0.0613x_1+0.0166x_2$$

将 $\hat{y}^{(0.5)}=\dfrac{\sqrt{\hat{y}}-1}{0.5}$ 代入，还原为原始变量的方程

$$\hat{y}=(4.1506+0.03065x_1+0.0083x_2)^2$$

并求得残差绝对值与 x_1 和 x_2 的等级相关系数 t 检验的 P 值分别为 0.206，0.108，都不显著，故可认为异方差被消除。

另外，经过 BOX-COX 变换后的 $R^2=0.8509$，F 值=34.23；而普通最小二乘的 $R^2=0.842$，F 值=31.96；加权最小二乘的 $R^2=0.849$，F 值=33.84。这说明用 BOX-COX 变换和加权最小二乘估计都能消除异方差，但对于本例的数据用 BOX-COX 变换比用加权最小二乘估计的拟合效果要好。对于不同的问题，结果可能不同，在实际应用时可以两种方法都试一下。

2. 消除自相关

下面我们对例 2.2 用 BOX-COX 变换消除残差序列自相关。

SAS 程序：

```
data eg2_2;
input year y x;
cards;
1990   1278.89   1510.16
……
2008   11242.85   15780.76
2009   12264.55   17174.65
;
proc transreg ss2;model boxcox(y/lambda=1 to 2 by 0.05)=identity(x);
run;
```

输出结果 4.9

```
The TRANSREG Procedure

Transformation Information
for BoxCox(y)

Lambda      R-Square    Log Like

  1.00          1.00    -100.716
  1.05          1.00     -94.006
  1.10          1.00     -85.068
  1.15          1.00     -77.602 <
  1.20          1.00     -83.995
  1.25          1.00     -93.447
  1.30          1.00    -100.671
  1.35          1.00    -106.228
  1.40          1.00    -110.706
  1.45          0.99    -114.456
  1.50          0.99    -117.688
```

专业课辅导+真题资料+微信公众号：432应统书院

1.55	0.99	-120.535
1.60	0.99	-123.086
1.65	0.98	-125.404
1.70	0.98	-127.534
1.75	0.98	-129.509
1.80	0.97	-131.356
1.85	0.97	-133.093
1.90	0.97	-134.738
1.95	0.96	-136.303
2.00	0.96	-137.798

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda

The TRANSREG Procedure Hypothesis Tests for BoxCox(y)

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	1	2.6729E9	2.6729E9	90975.5	>= <.0001
Error	18	528857	29381.0		
Corrected Total	19	2.6735E9			

The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.

Root MSE	171.40872	R-Square	0.9998
Dependent Mean	17916	Adj R-Sq	0.9998
Coeff Var	0.95673	Lambda	1.1500

Univariate Regression Table Based on the Usual Degrees of Freedom

Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Liberal p
Intercept	1	-691.45229	2663025	2663025	90.64	>= <.0001
Identity(x)	1	2.57362	2.673E9	2.673E9	90975.5	>= <.0001

根据以上输出结果，本例中的 λ 取 1.15，$y^{(1.15)}$ 对 x_1，x_2 的回归方程为

$$\hat{y}^{(1.15)}=-691.4523+2.5736x$$

将 $\hat{y}^{(1.15)}=\dfrac{\hat{y}^{1.15}-1}{1.15}$ 代入，还原为原始变量的方程

$$\hat{y}=(-794.17+2.96x)^{\frac{1}{1.15}}$$

变换后，可计算得新回归残差的 DW=2.167，查 DW 表，$n=20$，$k=2$，显著性水平 $\alpha=0.05$，得 $d_L=1.20$，$d_U=1.41$，而 $2<2.167<4-d_U$，即新的残差序列不存在自相关，这表明 BOX-COX 变换成功地消除了序列自相关。

由 $\hat{y}=(-794.17+2.96x)^{\frac{1}{1.15}}$，将 $x_{21}=19\,000$ 代入得

$$\hat{y}_{21}=13\,338.4$$

这与其他方法计算的结果近似。

4.6　异常值与强影响点

在回归分析的应用中，数据时常包含一些异常的或极端的观测值，这些观测值

专业课辅导+真题资料+微信公众号：432应统书院

与其他数据远远分开，可能引起较大的残差，极大地影响回归拟合的效果。在一元回归的情况下，用散点图或残差图就可以方便地识别出异常值，而在多元回归的情况下，用简单画图法识别异常值就很困难，需要更有效的方法。

异常值分为两种情况：一种是关于因变量 y 异常；另一种是关于自变量 x 异常。以下分别讨论这两种情况。

一、关于因变量 y 的异常值

在残差分析中，认为超过 $\pm 3\hat{\sigma}$ 的残差为异常值。由于普通残差 e_1，e_2，…，e_n 的方差 $D(e_i)=(1-h_{ii})\sigma^2$ 不等，用 e_i 作判断会带来一定的麻烦。类似于一元线性回归，在多元线性回归中，同样可以引入标准化残差 ZRE_i 和学生化残差 SRE_i 的概念，以改进普通残差的性质。定义形式与（2.64）式和（2.65）式完全相同，分别为：

标准化残差

$$\mathrm{ZRE}_i=\frac{e_i}{\hat{\sigma}} \tag{4.30}$$

学生化残差

$$\mathrm{SRE}_i=\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{4.31}$$

其中，h_{ii} 为帽子矩阵 $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的主对角线元素。标准化残差使残差具有可比性，$|\mathrm{ZRE}_i|>3$ 的相应观测值即判定为异常值，这简化了判定工作，但是没有解决方差不等的问题。学生化残差则进一步解决了方差不等的问题，比标准化残差又有所改进。但是当观测数据中存在关于 y 的异常观测值时，普通残差、标准化残差、学生化残差这三种残差都不再适用。这是由于异常值把回归线拉向自身，使异常值本身的残差减少，而其余观测值的残差增大，这时回归标准差 $\hat{\sigma}$ 也会增大，因而用“3σ”准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。

删除残差的构造思想是：在计算第 i 个观测值的残差时，用删除掉这第 i 个观测值的其余 $n-1$ 个观测值拟合回归方程，计算出第 i 个观测值的删除拟合值 $\hat{y}_{(i)}$，这个删除拟合值与第 i 个值无关，不受第 i 个值是否为异常值的影响，由此定义第 i 个观测值的删除残差为

$$e_{(i)}=y_i-\hat{y}_{(i)} \tag{4.32}$$

删除残差 $e_{(i)}$ 较普通残差更能如实反映第 i 个观测值的异常性。可以证明

$$e_{(i)}=\frac{e_i}{1-h_{ii}}$$

进一步，我们可以给出第 i 个观测值的删除学生化残差，记为 $\mathrm{SRE}_{(i)}$。删除学生化

专业课辅导+真题资料+微信公众号：432应统书院

残差 $SRE_{(i)}$ 的公式推导比较复杂，本书在此不加证明直接给出其表达式

$$SRE_{(i)}=SRE_i\left(\frac{n-p-2}{n-p-1-SRE_i^2}\right)^{\frac{1}{2}} \tag{4.33}$$

（4.33）式的证明请参见参考文献［2］。在实际应用中，我们可以直接用 SPSS 软件计算出删除学生化残差 $SRE_{(i)}$ 的数值，$|SRE_{(i)}|>3$ 的观测值即判定为异常值。

二、关于自变量 x 的异常值对回归的影响

由（3.24）式有 $D(e_i)=(1-h_{ii})\sigma^2$，其中，$h_{ii}$ 为帽子矩阵中主对角线的第 i 个元素，它是调节 e_i 方差大小的杠杆，因而称 h_{ii} 为第 i 个观测值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值 h_{ii} 也是表示自变量的第 i 次观测值与自变量平均值之间距离的远近。根据（3.24）式，较大的杠杆值的残差偏小，这是因为杠杆值大的观测点远离样本中心，能够把回归方程拉向自身，因而把杠杆值大的样本点称为强影响点。

强影响点并不一定是 y 值的异常值点，因而强影响点并不总会对回归方程造成不良影响。但是强影响点对回归效果通常有较强的影响，我们对强影响点应该有足够的重视，这是由于以下两个原因：（1）在实际问题中，因变量与自变量的线性关系只是在一定的范围内成立，强影响点远离样本中心，因变量与自变量之间可能不再是线性函数关系，因而在选择回归函数的形式时，要侧重于强影响点；（2）即使线性回归形式成立，但是强影响点远离样本中心，能够把回归方程拉向自身，使回归方程产生偏移。

由于强影响点并不总是 y 的异常值点，因而不能单纯根据杠杆值 h_{ii} 的大小判断强影响点是否异常。为此，我们引入库克距离，用来判断强影响点是否为 y 的异常值点。库克距离的计算公式为

$$D_i=\frac{e_i^2}{(p+1)\hat{\sigma}^2}\cdot\frac{h_{ii}}{(1-h_{ii})^2} \tag{4.34}$$

由（4.34）式可以看出，库克距离反映了杠杆值 h_{ii} 与残差 e_i 大小的一个综合效应。

根据（3.22）式，$\mathrm{tr}(\boldsymbol{H})=\sum_{i=1}^{n}h_{ii}=p+1$，则杠杆值 h_{ii} 的平均值为

$$\bar{h}=\frac{1}{n}\sum_{i=1}^{n}h_{ii}=\frac{p+1}{n} \tag{4.35}$$

这样，一个杠杆值 h_{ii} 如果大于 2 倍或 3 倍的 $\bar{h}$ 就认为是大的。

对于库克距离大小标准的判定比较复杂，较精确的方法请参见参考文献［2］。一个粗略的标准是：当 $D_i<0.5$ 时，认为不是异常值点；当 $D_i>1$ 时，认为是异常值点。

专业课辅导+真题资料+微信公众号：432应统书院

在用 SPSS 软件计算杠杆值时，计算的是中心化杠杆值 ch_{ii}，也就是自变量中心化后生成的帽子矩阵的主对角线元素，由参考文献［2］可知

$$ch_{ii}=h_{ii}-1/n$$

因此，$\sum_{i=1}^{n} ch_{ii}=p$，中心化杠杆值 ch_{ii} 的平均值是

$$\overline{ch}=\frac{1}{n}\sum_{i=1}^{n} ch_{ii}=\frac{p}{n}$$

三、异常值实例分析

以下我们以例 3.2 的北京各经济开发区的数据为例，作异常值的诊断分析。分别计算普通残差 e_i，学生化残差 SRE_i，删除残差 $e_{(i)}$，删除学生化残差 $SRE_{(i)}$，杠杆值 ch_{ii}，库克距离 D_i，见表 4.8。

表 4.8

序号	x_1	x_2	y	e_i	SRE_i	$e_{(i)}$	$SRE_{(i)}$	ch_{ii}	D_i
1	25	3 547.79	553.96	−832	−2.340	−1 490	−3.038	0.375	1.445
2	20	896.34	208.55	75	0.167	84	0.160	0.043	0.001
3	6	750.32	3.10	−34	−0.075	−38	−0.072	0.054	0.000
4	1 001	2 087.05	2 815.40	127	0.376	253	0.363	0.432	0.047
5	525	1 639.31	1 052.12	−458	−1.034	−529	−1.037	0.068	0.055
6	825	3 357.70	3 427.00	502	1.305	768	1.348	0.280	0.302
7	120	808.47	442.82	147	0.326	164	0.313	0.036	0.004
8	28	520.27	70.12	96	0.218	112	0.209	0.070	0.003
9	7	671.13	122.24	121	0.271	138	0.261	0.060	0.004
10	532	2 863.32	1 400.00	−697	−1.606	−837	−1.735	0.100	0.172
11	75	1 160.00	464.00	95	0.209	104	0.201	0.021	0.001
12	40	862.75	7.50	−151	−0.336	−169	−0.323	0.040	0.005
13	187	672.99	224.18	−145	−0.324	−164	−0.312	0.052	0.005
14	122	901.76	538.94	195	0.431	216	0.416	0.029	0.007
15	74	3 546.18	2 442.79	958	2.613	1 613	3.810	0.339	1.555

从表 4.8 中看到，绝对值最大的学生化残差为 $SRE_{15}=2.613$，小于 3，因而根据学生化残差诊断认为数据不存在异常值。绝对值最大的删除学生化残差为 $SRE_{(15)}=3.810$，因而根据删除学生化残差诊断认为第 15 个数据为异常值。其中心化杠杆值 $ch_{ii}=0.339$ 位居第三，库克距离 $D_i=1.555$ 位居第一。由于

$$\overline{ch}=\frac{p}{n}=\frac{2}{15}=0.133\,33$$

第 15 个数据 $ch_{ii}=0.339>2\overline{ch}$，因而从杠杆值看第 15 个数据是自变量的异常

专业课辅导+真题资料+微信公众号：432应统书院

值，同时库克距离 $D_{15}=1.555>1$，这样第 15 个数据为异常值是由自变量异常与因变量异常两个原因共同引起的。

诊断出异常值后，进一步要判断引起异常值的原因。引起异常值的原因通常有几条，具体见表 4.9。

表 4.9

异常值原因	异常值消除方法
1. 数据登记误差，存在抄写或录入的错误	重新核实数据
2. 数据测量误差	重新测量数据
3. 数据随机误差	删除或重新观测异常值数据
4. 缺少重要自变量	增加必要的自变量
5. 缺少观测数据	增加观测数据，适当扩大自变量取值范围
6. 存在异方差	采用加权线性回归
7. 模型选用错误，线性模型不适用	改用非线性回归模型

对引起异常值的不同原因，需要采取不同的处理方法。对本例的数据，通过核实认为不存在登记误差和测量误差。删除第 15 组数据，用其余 14 组数据拟合回归方程，发现第 6 组数据的删除学生化残差增加为 $SRE_{(6)}=4.418$，仍然存在异常值现象，因而认为异常值不是由于数据的随机误差引起的。实际上，在 4.3 节中已经诊断出本例数据存在异方差，应该采用加权最小二乘回归。权数为 $W_i=x_2^{-2.5}$，用 SPSS 软件计算出加权最小二乘回归的有关变量值如表 4.10 所示。

表 4.10

序号	x_1	x_2	y	e_i	SRE_i	$e_{(i)}$	$SRE_{(i)}^*$	ch_{ii}	D_i
1	25	3 547.79	553.96	−890	−1.149	−1 165	−1.165 8	0.234 1	0.136 0
2	20	896.34	208.55	20	0.135	23	0.129 3	0.060 4	0.000 9
3	6	750.32	3.10	−93	−0.795	−110	−0.782 4	0.050 1	0.038 5
4	1 001	2 087.05	2 815.40	403	1.175	716	1.196 3	0.429 4	0.358 1
5	525	1 639.31	1 052.12	−343	−1.135	−429	−1.149 8	0.186 4	0.108 1
6	825	3 357.70	3 427.00	715	0.937	841	0.932 0	0.147 1	0.051 5
7	120	808.47	442.82	126	0.949	139	0.944 8	0.009 3	0.031 8
8	28	520.27	70.12	45	0.717	74	0.701 5	0.133 9	0.111 5
9	7	671.13	122.24	62	0.617	76	0.600 8	0.046 3	0.028 7
10	532	2 863.32	1 400.00	−582	−0.926	−677	−0.919 9	0.136 6	0.046 6
11	75	1 160.00	464.00	58	0.281	65	0.270 2	0.074 8	0.003 3
12	40	862.75	7.50	−199	−1.391	−223	−1.454 4	0.032 4	0.076 5
13	187	672.99	224.18	−143	−1.611	−224	−1.742 4	0.227 2	0.495 1
14	122	901.76	538.94	175	1.137	189	1.152 8	0.011 2	0.036 0
15	74	3 546.18	2 442.79	916	1.173	1 179	1.193 9	0.220 9	0.131 7

从表 4.10 中看到，采用加权最小二乘回归后，删除学生化残差 $SRE_{(i)}$ 的绝对值最大者为 $|SRE_{(13)}|=1.742\,4$，库克距离小于 0.5，说明数据没有异常值。这个例子也说明了用加权最小二乘法处理异方差性问题的有效性。

专业课辅导+真题资料+微信公众号：432应统书院

4.7 本章小结与评注

一、异方差问题

本章介绍了模型随机误差项是否存在异方差性的诊断以及克服异方差性的方法。关于异方差性诊断的方法很多，至于哪种检验方法最好，目前还没有一致的看法。残差图分析法直观但较粗糙，等级相关系数法要比残差图检验方法更为可取。当残差散点图呈现无任何规律的分布时，我们可认为无异方差；如果残差点分布有明显的规律，可认为存在异方差。对于既无明显分布规律，分布似乎又不随机的情况，我们就要慎重了，这时，需要借助等级相关系数检验或其他方法来判断异方差性。

当根据某种检验方法认为存在异方差时，可以用自变量的幂函数作为权函数，作加权最小二乘回归，以解决异方差性带来的问题。多元线性回归有多个自变量，应该取哪一个自变量构造权函数呢？只需计算每个自变量 x_j 与残差绝对值的等级相关系数，选取等级相关系数最大的自变量构造权函数。

在实际应用中，实际工作者可能更多地从实际背景方面去分析和判断是否可能存在异方差以及权函数的形式。有人认为，对异方差的检验及权函数的选择，依赖于人们对可能的异方差形式的先验认识。

检验方法尽管不同，但它们有一个共同的思路。各种检验都是设法检验 ε_i 的方差与解释变量 x_j 的相关性，一般是通过 ε_i 的估计值 e_i 来进行这些检验。如果 ε_i 与某一 x_j 之间存在相关性，则模型存在异方差。

需要注意的是，加权最小二乘估计并不能消除异方差，只是能够消除或减弱异方差的不良影响。当存在异方差时，普通最小二乘估计不再具有最小方差线性无偏估计等良好的性质，而加权最小二乘估计可以改进估计的性质。加权最小二乘估计给误差项方差小的项加一个大的权数，给误差项方差大的项加一个小的权数，因此加强了小方差项的地位，使离差平方和中各项的作用相同。如果把误差项加权，那么加权的误差项 $\sqrt{w_i}\varepsilon_i$ 是等方差的。从残差图来看，普通最小二乘估计只能照顾到残差大的项，而小残差项往往有整体的正偏或负偏。加权最小二乘估计的残差图，对大残差和小残差拟合得都好，大残差和小残差都没有整体的正偏或负偏。

出现异方差时，消除异方差影响的方法也较多，用得最多的是加权最小二乘法。如果你使用的软件没有加权最小二乘功能，可以通过对数据做变换，把第 i 组观测数据同乘以 $\sqrt{w_i}$ 后，对变换后的数据做普通最小二乘，可以得到与加权最小二乘等价的回归方程。只是使用这种方法时，变换后数据的回归方程中可能不含有回归常数项，给回归的拟合优度检验带来麻烦。具体方法参见参考文献［16］。当模型存在异方差时，人们往往还考虑对因变量作变换，使得对变换过后的数据，误差方差能够近似相等，即方差比较稳定，所以通常称这种变换为方差稳定变换（参见

专业课辅导+真题资料+微信公众号：432应统书院

参考文献［2］）。常见的变量变换有如下几种：

（1）如果 σ_i^2 与 $E(y_i)$ 存在一定的比例关系，使用 $y'=\sqrt{y}$。

（2）如果 σ_i 与 $E(y_i)$ 存在一定的比例关系，使用 $y'=\ln(y)$。

（3）如果 $\sqrt{\sigma_i}$ 与 $E(y_i)$ 存在一定的比例关系，使用 $y'=\frac{1}{y}$。

方差稳定变换在改变误差项方差的同时，也会改变误差项的分布和回归函数的形式。因而当误差项服从正态分布，因变量与自变量之间遵从线性回归函数关系，只是误差项存在异方差时，应该采用加权最小二乘估计，以消除异方差的影响。当误差项不仅存在异方差，而且误差项不服从正态分布，因变量与自变量之间也不遵从线性回归函数关系时，应该采用方差稳定变换。

二、自相关问题

在 4.4 节中我们讨论了模型随机误差项存在序列相关时带来的严重后果，并给出了两种诊断方法，介绍了几种克服序列相关的方法。就诊断方法而言，残差图方法直观，但不够严谨；DW 检验是最常用的一种方法，许多统计软件中都有 DW 值，用起来很方便，但 DW 检验也有一些局限性。尤其是 DW 检验有两个不能确定结果的区域，对于这种状况，一般需要增大样本量。但在实际问题的研究中，样本量的获取往往受到一定限制。为了克服 DW 检验的这一局限，J. Durbin 和 G. S. Watson 在参考文献［26］中给出了一个近似的检验，在使用下界 d_L 和上界 d_U 的 DW 检验得不到确定结果时可以使用。

另外在 DW 表中，变量个数较多，样本量 n 较小时会出现 $d_U>2$ 的情形，这正是这种方法的一个不太合理的地方。在多元线性回归中，一定要注意 n 与 p 的匹配问题。

回归检验法也很受人们的推崇。回归检验法需要首先应用普通最小二乘法估计模型并求出 ε 的估计值 e，然后以 e_t 为被解释变量，以各种可能的相关量，诸如 e_{t-1}，e_{t-2} 等作为解释变量分别进行线性拟合

$$e_t=\beta e_{t-1}+u_t$$

$$e_t=\beta_1 e_{t-1}+\beta_2 e_{t-2}+u_t$$

……

对各种拟合形式进行统计检验，选择显著的最优拟合形式作为序列相关的具体形式。这种方法的优点是确定了序列相关性存在时，也就确定了相关的形式，而且它适用于任何形式的序列相关检验。参考文献［9］中详细介绍了这种方法的应用。

用迭代法处理序列相关并不总是有效。主要原因是当误差项正自相关时，（4.16）式往往低估自相关参数 ρ。如果这种偏差严重，就会显著地降低迭代法的效率。

对于误差项一阶自相关回归模型（4.18）式，用迭代法得到的（4.28）式回归方程适于作短期预测。如果要作长期预测，可直接使用回归方程 $\hat{y}_t=\hat{\beta}_0+\hat{\beta}_1 x_t$，

专业课辅导+真题资料+微信公众号：432应统书院

这里 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 不是普通最小二乘估计值，而是根据（4.23）式，用公式 $\hat{\beta}_0=\hat{\beta}_0'/(1-\hat{\rho})$ 和 $\hat{\beta}_1=\hat{\beta}_1'$ 转换得到的。

一阶差分法是自相关参数 $\rho=1$ 时的迭代法，一阶差分模型的一个重要特征是它没有截距项，得到的差分回归线通过原点。

一阶差分法是对原始数据的一种修正，有时一阶差分法可能会过度修正，使得差分数据中出现负自相关的误差项。因此，从一定意义上说，差分法要慎用。只有当 $\rho=1$ 或者接近 1 时，差分法的效果才会好。

三、异常值问题

对异常值的分析是得到优良回归方程的一个必要组成部分，这项工作可以借助计算机软件实现，但是并不能由计算机软件自动完成，它需要统计分析人员进行有效的判断。

本书介绍了用删除学生化残差、杠杆值、库克距离等识别异常值的方法，在识别出异常值后，必须决定对这些异常观测值应该采取什么措施。对异常观测值，不能总是简单地剔除了事，有时异常观测值是正确的，它说明了回归模型为什么失败。失败的原因可能是遗漏了一个重要的自变量，或者是选择了不正确的回归函数形式。

如果一个异常值数据是准确的，但是找不到对它的合理解释，与剔除这个观测值相比，一个更稳健的方法是抑制它的影响。最小绝对离差和法是一种稳健估计方法，它具有对异常值和不合适模型不敏感的性质。最小绝对离差和法是寻找参数 β_0，β_1，β_2，…，β_p 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，使绝对离差和达到极小，即寻找 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，满足

$$
\begin{aligned}
Q(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p) &= \sum_{i=1}^{n}\left|y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip}\right| \\
&= \min_{\beta_0,\beta_1,\beta_2,\cdots,\beta_p}\sum_{i=1}^{n}\left|y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip}\right|
\end{aligned}
\tag{4.36}
$$

依照（4.36）式求出的 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$ 就称为回归参数 β_0，β_1，β_2，…，β_p 的最小绝对离差和估计，在 SPSS 软件中可以使用非线性回归功能计算。

□ 思考与练习

4.1 试举例说明产生异方差的原因。

4.2 异方差性带来的后果有哪些？

4.3 简述用加权最小二乘法消除一元线性回归中异方差性的思想与方法。

专业课辅导+真题资料+微信公众号：432应统书院

4.4　简述用加权最小二乘法消除多元线性回归中异方差性的思想与方法。

4.5　验证（4.5）式一元加权最小二乘回归系数估计公式。

4.6　验证（4.8）式多元加权最小二乘回归系数估计公式。

4.7　有同学认为当数据存在异方差时，加权最小二乘回归方程与普通最小二乘回归方程之间必然有很大的差异，异方差越严重，两者之间的差异就越大。你是否同意这位同学的观点？说明原因。

4.8　对例 4.3 的数据，用公式 $e'_{iw}=\sqrt{w_i}\cdot e_{iw}$ 计算出加权变换残差 e'_{iw}，绘制加权变换残差图，根据绘制出的图形说明加权最小二乘估计的效果。

4.9　参见参考文献[2]，表 4.11 是用电高峰每小时用电量 y 与每月总用电量 x 的数据。

（1）用普通最小二乘法建立 y 与 x 的回归方程，并画出残差散点图。

（2）诊断该问题是否存在异方差。

（3）如果存在异方差，用幂指数型的权函数建立加权最小二乘回归方程。

（4）用方差稳定变换 $y'=\sqrt{y}$ 消除异方差。

表 4.11

用户序号	x	y	用户序号	x	y
1	679	0.79	28	1 748	4.88
2	292	0.44	29	1 381	3.48
3	1 012	0.56	30	1 428	7.58
4	493	0.79	31	1 255	2.63
5	582	2.70	32	1 777	4.99
6	1 156	3.64	33	370	0.59
7	997	4.73	34	2 316	8.19
8	2 189	9.50	35	1 130	4.79
9	1 097	5.34	36	463	0.51
10	2 078	6.85	37	770	1.74
11	1 818	5.84	38	724	4.10
12	1 700	5.21	39	808	3.94
13	747	3.25	40	790	0.96
14	2 030	4.43	41	783	3.29
15	1 643	3.16	42	406	0.44
16	414	0.50	43	1 242	3.24
17	354	0.17	44	658	2.14
18	1 276	1.88	45	1 746	5.71
19	745	0.77	46	468	0.64
20	435	1.39	47	1 114	1.90
21	540	0.56	48	413	0.51
22	874	1.56	49	1 787	8.33
23	1 543	5.28	50	3 560	14.94
24	1 029	0.64	51	1 495	5.11
25	710	4.00	52	2 221	3.85
26	1 434	0.31	53	1 526	3.93
27	837	4.20			

专业课辅导+真题资料+微信公众号：432应统书院

4.10 试举一可能产生随机误差项序列相关的经济例子。

4.11 序列相关性带来的严重后果是什么？

4.12 总结 DW 检验的优缺点。

4.13 表 4.12 中是某软件公司月销售额数据，其中，x 为总公司的月销售额（万元）；y 为某分公司的月销售额（万元）。

(1) 用普通最小二乘法建立 y 与 x 的回归方程。

(2) 用残差图及 DW 检验诊断序列的自相关性。

(3) 用迭代法处理序列相关，并建立回归方程。

(4) 用一阶差分法处理数据，并建立回归方程。

(5) 比较以上各方法所建回归方程的优良性。

表 4.12

序号	x	y	序号	x	y
1	127.3	20.96	11	148.3	24.54
2	130.0	21.40	12	146.4	24.28
3	132.7	21.96	13	150.2	25.00
4	129.4	21.52	14	153.1	25.64
5	135.0	22.39	15	157.3	26.46
6	137.1	22.76	16	160.7	26.98
7	141.1	23.48	17	164.2	27.52
8	142.8	23.66	18	165.6	27.78
9	145.5	24.10	19	168.7	28.24
10	145.3	24.01	20	172.0	28.78

4.14 某乐队经理研究其乐队 CD 盘的销售额（y），两个有关的影响变量是每周演出场次 x_1 和乐队网站的周点击率 x_2，数据见表 4.13。

(1) 用普通最小二乘法建立 y 与 x_1 和 x_2 的回归方程，用残差图及 DW 检验诊断序列的自相关性。

(2) 用迭代法处理序列相关，并建立回归方程。

(3) 用一阶差分法处理数据，并建立回归方程。

(4) 比较以上各方法所建回归方程的优良性。

表 4.13

周次	销售额 y	每周演出场次 x_1	周点击率 x_2	周次	销售额 y	每周演出场次 x_1	周点击率 x_2
1	893.93	5	292	9	866.43	5	266
2	1 091.27	5	252	10	603	5	253
3	1 229.97	5	267	11	343.52	5	315
4	1 045.85	5	379	12	472.1	6	271
5	997.24	5	318	13	171.79	4	166
6	1 495.14	6	393	14	135.79	4	204
7	1 200.56	5	331	15	925.95	5	335
8	747.24	4	204	16	1 574.01	5	352

专业课辅导+真题资料+微信公众号：432应统书院

续前表

周次	销售额 y	每周演出场次 x_1	周点击率 x_2	周次	销售额 y	每周演出场次 x_1	周点击率 x_2
17	1 405.33	5	274	35	982.94	6	391
18	971.27	4	333	36	722.28	4	279
19	1 165.2	5	302	37	1 337.44	5	322
20	597.85	4	324	38	1 150.51	4	231
21	490.34	4	327	39	1 514.84	6	368
22	709.59	5	206	40	1 442.08	5	357
23	987.3	5	310	41	767.64	5	260
24	954.6	6	306	42	1 020.03	5	298
25	1 216.89	6	350	43	1 067.49	5	350
26	1 491.52	5	275	44	1 484.12	6	320
27	668.3	4	173	45	957.68	4	227
28	915.03	5	360	46	1 344.91	5	261
29	565.92	4	340	47	1 361.78	5	303
30	1 267.98	5	380	48	1 424.69	6	263
31	930.24	6	285	49	1 158.21	4	215
32	379.38	4	232	50	827.56	4	294
33	500.74	5	294	51	803.16	4	288
34	83.65	5	220	52	1 447.46	6	257

4.15　说明引起异常值的原因和消除异常值的方法。

4.16　对第 3 章思考与练习第 11 题做异常值检验。

专业课辅导+真题资料+微信公众号：432应统书院

第 5 章 Chapter 5 自变量选择与逐步回归

回归自变量的选择无疑是建立回归模型的一个极为重要的问题。在建立一个实际问题的回归模型时，首先碰到的问题便是如何确定回归自变量，一般情况下，我们大都是根据所研究问题的目的，结合经济理论罗列出对因变量可能有影响的一些因素作为自变量。如果遗漏了某些重要的变量，回归方程的效果肯定不好；如果担心遗漏了重要的变量，而考虑过多的自变量，在这些变量中，某些自变量对问题的研究可能并不重要，有些自变量数据的质量可能很差，有些变量可能和其他变量有很大程度的重叠，结果，不仅计算量增大许多，而且得到的回归方程稳定性也很差，直接影响到回归方程的应用。

从 20 世纪 60 年代开始，关于回归自变量的选择成为统计学中研究的热点问题。统计学家提出了许多回归选元的准则，并提出了许多行之有效的选元方法。本章我们从回归选元对回归参数估计和预测的影响开始，介绍自变量选择常用的几个准则；扼要介绍所有子集回归选元的几个方法；详细讨论逐步回归方法及其应用。

5.1 自变量选择对估计和预测的影响

一、全模型和选模型

设我们研究的某一实际问题涉及对因变量有影响的因素共有 m 个，由因变量 y 和 m 个自变量 x_1，x_2，…，x_m 构成的回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{5.1}$$

专业课辅导+真题资料+微信公众号：432应统书院

因为模型（5.1）式是因变量 y 与所有自变量 x_1，x_2，…，x_m 的回归模型，故称（5.1）式为全回归模型。

如果从所有可供选择的 m 个变量中挑选出 p 个，记为 x_1，x_2，…，x_p，由所选的 p 个自变量组成的回归模型为

$$y=\beta_{0p}+\beta_{1p}x_1+\beta_{2p}x_2+\cdots+\beta_{pp}x_p+\varepsilon_p \tag{5.2}$$

相对全模型而言，我们称模型（5.2）式为选模型。选模型（5.2）式的 p 个自变量 x_1，x_2，…，x_p 并不一定是全体 m 个自变量 x_1，x_2，…，x_m 中的前 p 个，x_1，x_2，…，x_p 是在 m 个自变量 x_1，x_2，…，x_m 中按某种规则挑选出的 p 个，不过为了方便，我们不妨认为 x_1，x_2，…，x_p 就是 x_1，x_2，…，x_m 中的前 p 个。

自变量的选择问题可以看成对一个实际问题是用（5.1）式全模型还是用（5.2）式选模型去描述。如果应该用（5.1）式全模型去描述实际问题，而误选了（5.2）式选模型，这就说明在建模时丢掉了一些有用的变量；如果应该选用（5.2）式选模型，而误选了（5.1）式全模型，这就说明我们把一些不必要的自变量引进了模型。

模型选择不当会给参数估计和预测带来什么影响？下面将分别给予讨论。

为了方便，把模型（5.1）式的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

$$\hat{\boldsymbol{\beta}}_m=(\boldsymbol{X}'_m\boldsymbol{X}_m)^{-1}\boldsymbol{X}'_m\boldsymbol{y} \tag{5.3}$$

$$\hat{\sigma}_m^2=\frac{1}{n-m-1}\mathrm{SSE}_m \tag{5.4}$$

把模型（5.2）式的参数向量 $\boldsymbol{\beta}$ 和 σ^2 的估计记为

$$\hat{\boldsymbol{\beta}}_p=(\boldsymbol{X}'_p\boldsymbol{X}_p)^{-1}\boldsymbol{X}'_p\boldsymbol{y} \tag{5.5}$$

$$\hat{\sigma}_p^2=\frac{1}{n-p-1}\mathrm{SSE}_p \tag{5.6}$$

二、自变量选择对预测的影响

假设全模型（5.1）式与选模型（5.2）式不同，即要求 $p<m$，$\beta_{p+1}x_{p+1}+\cdots+\beta_m x_m$ 不恒为 0。在此条件下，当全模型（5.1）式正确而误用了选模型（5.2）式时，本书不加证明地引用以下性质。

性质 1　在 x_j 与 x_{p+1}，…，x_m 的相关系数不全为 0 时，选模型回归系数的最小二乘估计是全模型相应参数的有偏估计，即 $E(\hat{\beta}_{jp})=\beta_{jp}\neq\beta_j(j=1,2,\cdots,p)$。

性质 2　选模型的预测是有偏的。给定新自变量值，$\boldsymbol{x}_{0m}=(x_{01},x_{02},\cdots,x_{0m})'$，因变量新值为 $y_0=\beta_0+\beta_1x_{01}+\beta_2x_{02}+\cdots+\beta_m x_{0m}+\varepsilon_0$，用选模型的预测值 $\hat{y}_{0p}=\hat{\beta}_{0p}+\hat{\beta}_{1p}x_{01}+\hat{\beta}_{2p}x_{02}+\cdots+\hat{\beta}_{pp}x_{0p}$，作为 y_0 的预测值是有偏的，即 $E(\hat{y}_{0p}-y_0)\neq 0$。

性质 3　选模型的参数估计有较小的方差。选模型的最小二乘参数估计为 $\hat{\boldsymbol{\beta}}_p=$

专业课辅导+真题资料+微信公众号：432应统书院

$(\hat{\beta}_{0p}, \hat{\beta}_{1p}, \cdots, \hat{\beta}_{pp})'$，全模型的最小二乘参数估计为 $\hat{\boldsymbol{\beta}}_m=(\hat{\beta}_{0m}, \hat{\beta}_{1m}, \cdots, \hat{\beta}_{mm})'$，这一性质说明 $D(\hat{\beta}_{jp})\leqslant D(\hat{\beta}_{jm})(j=0, 1, \cdots, p)$。

性质 4 选模型的预测残差有较小的方差。选模型的预测残差为 $e_{0p}=\hat{y}_{0p}-y_0$，全模型的预测残差为 $e_{0m}=\hat{y}_{0m}-y_0$，其中 $y_0=\beta_0+\beta_1x_{01}+\beta_2x_{02}+\cdots+\beta_mx_{0m}+\varepsilon$，则有 $D(e_{0p})\leqslant D(e_{0m})$。

性质 5 记 $\boldsymbol{\beta}_{m-p}=(\beta_{p+1}, \cdots, \beta_m)'$，用全模型对 $\boldsymbol{\beta}_{m-p}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}}_{m-p}=(\hat{\beta}_{p+1}, \cdots, \hat{\beta}_m)'$，则在 $D(\hat{\boldsymbol{\beta}}_{m-p})\geqslant \boldsymbol{\beta}_{m-p}\boldsymbol{\beta}'_{m-p}$ 的条件下，$E(e_{0p})^2=D(e_{0p})+(E(e_{0p}))^2\leqslant D(e_{0m})$，即选模型预测的均方误差比全模型预测的方差更小。

以上性质的证明参见参考文献［2］。

性质 1 和性质 2 表明，当全模型（5.1）式正确时，而我们舍去了 $m-p$ 个自变量，用剩下的 p 个自变量去建立选模型（5.2）式，参数估计值是全模型相应参数的有偏估计，用其作预测，预测值也是有偏的。这是误用选模型产生的弊端。

性质 3 和性质 4 表明，用选模型去作预测，残差的方差比用全模型去作预测的方差小，尽管用选模型所作的预测是有偏的，但得到的预测残差的方差下降了。这说明尽管全模型正确，误用选模型是有弊也有利的。

性质 5 说明即使全模型正确，但如果其中有一些自变量对因变量影响很小或回归系数方差过大，我们丢掉这些变量之后，用选模型去预测可以提高预测的精度。由此可见，如果模型中包含了一些不必要的自变量，模型的预测精度就会下降。

上述结论告诉我们，一个好的回归模型，并不是考虑的自变量越多越好。在建立回归模型时，选择自变量的基本指导思想是少而精。哪怕我们丢掉了一些对因变量 y 有些影响的自变量，由选模型估计的保留变量的回归系数的方差，要比由全模型所估计的相应变量的回归系数的方差小。对于所预测的因变量的方差来说也是如此。丢掉了一些对因变量 y 有影响的自变量后，所付出的代价是估计量产生了有偏性。然而，尽管估计量是有偏的，但预测偏差的方差会下降。因此，自变量的选择有重要的实际意义。在建立实际问题的回归模型时，应尽可能剔除那些可有可无的自变量。

5.2 所有子集回归

一、所有子集的数目

设在一个实际问题的回归建模中，有 m 个可供选择的变量 $x_1, x_2, \cdots, x_m$，由于每个自变量都有入选和不入选两种情况，这样 y 关于这些自变量的所有可能的回归方程就有 2^m-1 个，这里 -1 是要求回归模型中至少包含一个自变量，即减去模型中只包含常数项的这种情况。如果把回归模型中只包含常数项的这种情况也算

专业课辅导+真题资料+微信公众号：432应统书院

在内，那么所有可能的回归方程就有 2^m 个。

从另一个角度看，选模型包含的自变量数目 p 有从 0 到 m 共 $m+1$ 种不同情况，而对选模型中恰包含 p 个自变量的情况，从全部 m 个自变量中选出 p 个的方法共有组合数 C_m^p 个（或记为 $\begin{pmatrix} m \\ p \end{pmatrix}$），因而所有选模型的数目为

$$C_m^0 + C_m^1 + \cdots + C_m^m = 2^m$$

二、关于自变量选择的几个准则

对于有 m 个自变量的回归建模问题，一切可能的回归子集有 2^m 个，在这些回归子集中如何选择一个最优的回归子集，衡量最优子集的标准是什么，这是我们这一节要讨论的问题。

在第 3 章，我们从数据与模型拟合优劣的角度出发，认为残差平方和 SSE 最小的回归方程就是最好的，还用复相关系数 R 来衡量回归拟合的好坏。然而下面的讨论将会看到上述两种方法都有明显的不足。

我们把选模型（5.2）式的残差平方和记为 SSE_p，当再增加一个新的自变量 x_{p+1}时，相应的残差平方和记为 SSE_{p+1}。根据最小二乘估计的原理，增加自变量时残差平方和将减少，减少自变量时残差平方和将增加。因此有

$$\mathrm{SSE}_{p+1} \leqslant \mathrm{SSE}_p$$

又记它们的复决定系数分别为：$R_{p+1}^2 = 1 - \mathrm{SSE}_{p+1}/\mathrm{SST}$，$R_p^2 = 1 - \mathrm{SSE}_p/\mathrm{SST}$。由于 SST 是因变量的离差平方和，与自变量无关，因而

$$R_{p+1}^2 \geqslant R_p^2$$

即当自变量子集在扩大时，残差平方和随之减少，而复决定系数 R^2 随之增大。因此，如果按残差平方和越小越好的原则来选择自变量子集，或者为提高复决定系数，不论什么变量只要多取就行，则毫无疑问选的变量越多越好。这样由于变量的多重共线性，给变量的回归系数估计值带来不稳定性，加上变量的测量误差积累和参数数目增加，将使估计值的误差增大。如此构造的回归模型稳定性差，使得为增大复相关系数 R 而付出了模型参数估计稳定性差的代价。因此残差平方和、复相关系数或样本决定系数都不能作为选择变量的准则。

下面从不同的角度给出几个常用的准则。

准则 1　自由度调整复决定系数达到最大。

前面我们已看到，当给模型增加自变量时，复决定系数也随之逐步增大，然而复决定系数增大的代价是残差自由度的减少，因为残差自由度等于样本个数与自变量个数之差。自由度小意味着估计和预测的可靠性低。这表明一个回归方程涉及的自变量很多时，回归模型的拟合从表面上看是良好的，而区间预测和区间估计的幅度则变大，以至失去实际意义。这里回归模型的拟合良好掺杂了一些虚假成分。为

专业课辅导+真题资料+微信公众号：432应统书院

了克服样本决定系数的这一缺点，我们设法对 R^2 给予适当的修正，使得只有加入有意义的变量时，经过修正的样本决定系数才会增加，这就是所谓的自由度调整复决定系数。

设 R_a^2 为调整的复决定系数，n 为样本量，p 为自变量的个数，则

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) \tag{5.7}$$

显然有 $R_a^2 \leqslant R^2$，R_a^2 随着自变量的增加并不一定增大。由（5.7）式可以看到，尽管 $1-R^2$ 随着变量的增加而减少，但由于其前面的系数 $(n-1)/(n-p-1)$ 起折扣作用，才使 R_a^2 随着自变量的增加并不一定增大。当所增加的自变量对回归的贡献很小时，R_a^2 反而可能减少。

在一个实际问题的回归建模中，自由度调整复决定系数 R_a^2 越大，所对应的回归方程越好。从拟合优度的角度追求最优，则所有回归子集中 R_a^2 最大者对应的回归方程就是最优方程。

从另外一个角度考虑回归的拟合效果，回归误差项方差 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\mathrm{SSE}$$

此无偏估计式中也加入了惩罚因子 $n-p-1$，$\hat{\sigma}^2$ 实际上就是用自由度 $n-p-1$ 作平均的平均残差平方和。当自变量个数从 0 开始增加时，SSE 逐渐减小，作为除数的惩罚因子 $n-p-1$ 也随之减小。一般来说，当自变量个数从 0 开始增加时，$\hat{\sigma}^2$ 先是开始下降，而后稳定下来，当自变量个数增加到一定数量后，$\hat{\sigma}^2$ 又开始增加。这是因为刚开始时，随着自变量个数的增加，SSE 能够快速减小，虽然作为除数的惩罚因子 $n-p-1$ 也随之减小，但由于 SSE 减小的速度更快，因而 $\hat{\sigma}^2$ 是趋于减小的。当自变量数目增加到一定程度，重要的自变量基本都选上了，这时再增加自变量，SSE 减少的幅度不大，以至于抵消不了除数 $n-p-1$ 的减小，最终又导致了 $\hat{\sigma}^2$ 的增加。

由以上分析可知，用平均残差平方和 $\hat{\sigma}^2$ 作为自变量选元准则是合理的，那么它和调整的复决定系数 R_a^2 准则有什么关系呢？实际上，这两个准则是等价的，容易证明以下关系式成立

$$R_a^2 = 1 - \frac{n-1}{\mathrm{SST}}\hat{\sigma}^2 \tag{5.8}$$

由于 SST 是与回归无关的固定值，因而 R_a^2 与 $\hat{\sigma}^2$ 是等价的。

准则 2 赤池信息量 AIC 达到最小。

AIC 准则是日本统计学家赤池（Akaike）于 1974 年根据最大似然估计原理提出的一种模型选择准则，人们称它为赤池信息量准则（Akaike information criterion，AIC）。AIC 准则既可用来作回归方程自变量的选择，又可用于时间序列分析中自回归模型的定阶。该方法的广泛应用使得赤池乃至日本统计学家在世界的该领

专业课辅导+真题资料+微信公众号：432应统书院

域中声名鹊起。

对一般情况，设模型的似然函数为 $L(\boldsymbol{\theta}, \boldsymbol{x})$，$\boldsymbol{\theta}$ 的维数为 p，$\boldsymbol{x}$ 为随机样本（在回归分析中随机样本为 $\boldsymbol{y}=(y_1, y_2, \cdots, y_n)'$），则 AIC 定义为

$$\mathrm{AIC} = -2\ln L(\hat{\boldsymbol{\theta}}_L, \boldsymbol{x}) + 2p \tag{5.9}$$

其中，$\hat{\boldsymbol{\theta}}_L$ 为 $\boldsymbol{\theta}$ 的最大似然估计；p 为未知参数的个数。式中右边第一项是似然函数的对数乘以 -2，第二项惩罚因子是未知参数个数的 2 倍。我们知道，似然函数越大的估计量越好，而 AIC 是似然函数的对数乘以 -2 再加上惩罚因子 $2p$，因而使 AIC 达到最小的模型是最优模型。

下面我们讨论把 AIC 用于回归模型的选择。假定回归模型的随机误差项 ε 服从正态分布，即

$$\varepsilon \sim N(0, \sigma^2)$$

在这个正态假定下，回归参数的最大似然估计已在 3.2 节中给出，根据（3.28）式

$$\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2}\mathrm{SSE}$$

将 $\hat{\sigma}_L^2 = \frac{1}{n}\mathrm{SSE}$ 代入得

$$\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{\mathrm{SSE}}{n}\right) - \frac{n}{2}$$

将上式代入（5.9）式中，这里似然函数中的未知参数个数为 $p+2$，略去与 p 无关的常数，得回归模型的 AIC 公式为

$$\mathrm{AIC} = n\ln(\mathrm{SSE}) + 2p \tag{5.10}$$

在回归分析的建模过程中，对每一个回归子集计算 AIC，其中 AIC 最小者所对应的模型是最优回归模型。

准则 3　C_p 统计量达到最小。

1964 年马洛斯（Mallows）从预测的角度提出一个可以用来选择自变量的统计量，这就是我们常说的 C_p 统计量。根据性质 5，即使全模型正确，但仍有可能选模型有更小的预测误差。C_p 正是根据这一原理提出来的。

考虑在 n 个样本点上，用选模型（5.2）式作回归预测，预测值与期望值的相对偏差平方和为

$$\begin{aligned} J_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{ip} - E(y_i))^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2 \end{aligned}$$

可以证明，J_p 的期望值是

专业课辅导+真题资料+微信公众号：432应统书院

$$E(J_p)=\frac{E(\mathrm{SSE}_p)}{\sigma^2}-n+2(p+1)$$

对以上证明有兴趣的读者请参见参考文献［5］。略去无关的常数 2，据此构造出 C_p 统计量为

$$\begin{aligned}C_p&=\frac{\mathrm{SSE}_p}{\hat{\sigma}^2}-n+2p\\&=(n-m-1)\frac{\mathrm{SSE}_p}{\mathrm{SSE}_m}-n+2p\end{aligned}\tag{5.11}$$

其中，$\hat{\sigma}^2=\frac{1}{n-m-1}\mathrm{SSE}_m$ 为全模型中 σ^2 的无偏估计。这样我们得到一个选择变量的 C_p 准则：选择使 C_p 最小的自变量子集，这个自变量子集对应的回归方程就是最优回归方程。

上面从不同角度介绍了三个准则，自变量选择的准则还有一些，我们就不一一列举了。下面用一个例子对所有回归子集计算上述三个准则，综合比较一下最优回归子集的选择。

例 5.1

y 表示某种消费品的销售额，x_1 表示居民可支配收入，x_2 表示该类消费品的价格指数，x_3 表示其他消费品平均价格指数。表 5.1 给出了某地区 18 年某种消费品销售情况资料，试建立该地区该消费品销售额预测方程。

表 5.1

序号	x_1（元）	x_2（%）	x_3（%）	y（百万元）
1	81.2	85.0	87.0	7.8
2	82.9	92.0	94.0	8.4
3	83.2	91.5	95.0	8.7
4	85.9	92.9	95.5	9.0
5	88.0	93.0	96.0	9.6
6	99.9	96.0	97.0	10.3
7	102.0	95.0	97.5	10.6
8	105.3	95.6	97.0	10.9
9	117.7	98.9	98.0	11.3
10	126.4	101.5	101.2	12.3
11	131.2	102.0	102.5	13.5
12	148.0	105.0	104.0	14.2
13	153.0	106.0	105.9	14.9
14	161.0	109.0	109.5	15.9
15	170.0	112.0	111.0	18.5
16	174.0	112.5	112.0	19.5
17	185.0	113.0	112.3	19.9
18	189.0	114.0	113.0	20.5

专业课辅导+真题资料+微信公众号：432应统书院

在例 5.1 中，$n=18$，$m=3$，所有的自变量子集有 $2^m-1=7$ 个，即有 7 个回归子集。用 SPSS 软件计算这 7 个回归子集及其上述统计量。对统计量列表，见表 5.2。

表 5.2

自变量子集	R^2	R_a^2	AIC	C_p
x_1	0.972 8	0.971 1	40.06	4.134
x_2	0.956 6	0.953 9	48.48	16.151
x_3	0.950 8	0.947 7	50.74	20.452
x_1，x_2	0.974 7	0.971 4	40.76	4.734
x_1，x_3	0.978 4	0.975 5	37.93	2.005
x_2，x_3	0.957 6	0.951 9	50.09	17.461
x_1，x_2，x_3	0.981 1	0.977 1	37.52	2.000

由表 5.2 的几项指标均可看到 x_1，x_2，x_3 是最优子集，x_1，x_3 是次优子集。则回归方程分别为

$$\hat{y}=-10.148\,9+0.100\,8x_1-0.310\,4x_2+0.411\,0x_3$$

$$\hat{y}=-14.049+0.076\,41x_1+0.117\,8x_3$$

因为这个实际问题所涉及的自变量较少，只有 3 个，所以根据以上几个准则判断全模型是最优的。这种情况在自变量只有少数几个时是常见的，但当涉及的自变量数目较多时，很少见到全模型是最优的。我们讲的最优是相对而言的，在实际问题的选模中应综合考虑，或根据实际问题的研究目的从不同角度来考虑。如有时希望模型各项衡量准则较优，得到的模型又能给出合理的经济解释；有时只从拟合角度考虑，有时只从预测角度考虑，并不计较回归方程能否有合理解释；有时要求模型的各个衡量准则较优，而模型最好简单一些，涉及变量少一些；有时还要看回归模型参数估计的标准误差大小等。因此，上述准则只给了我们选择模型的一些参考，最终的选择既应以上述几个准则作基本参考根据，又要考虑实际问题的性质和需要。

三、用 SAS 软件寻找最优子集

SAS 软件是以子过程（procedure）的形式生成的组合式软件包，在使用时需要编写简单的程序命令，不像 SPSS 软件那样可以用菜单的方式运行。SAS 软件包含的统计方法比 SPSS 更为丰富，想全面学习掌握 SAS 软件是很困难的，如果只是想学会某几种特殊方法的应用则并非难事。在应用中，我们可以以 SPSS 为主，对于 SPSS 解决不了的问题，再用 SAS 软件解决。SPSS 没有用所有子集回归选择最优子集的功能，SAS 软件中提供了用 R_a^2 准则和 C_p 准则选择最优子集的功能，没有提供用 AIC 准则选元的功能。以下结合例 3.1 的数据，介绍用 SAS 软件寻找最优子集的方法。

专业课辅导+真题资料+微信公众号：432应统书院

进入 SAS 软件后，屏幕显示如下：

File Edit View Locals Globals Options Window Help 快捷键
LOG
PROGRAM

第一行是主菜单，第二行是快捷键，下面的 LOG 和 PROGRAM 是两个窗口。

例 5.2

对例 3.1 的数据，用调整的复决定系数 R_a^2 准则选择最优子集回归模型。

将光标定位在程序编辑窗口（PROGRAM EDITOR），输入以下内容：

```
data data1;
input x1-x9 y;
cards;
5561.54   1101.49   1286.32   1563.1 809.25 55844  6060.28
   105.06 1.82   16460.26
5005.09   812.52 1528.28   1220.92   715.24 39990  3110.97   105.4
   3.6 13422.47
3155.4 754.78 1097.41   808.88 386.6  24276  8477.63   106.2  3.96
   9086.73
2974.76   820.81 1250.87   769.79 570.79 25489  3571.37   107.19
   3.29   8806.55
3553.48   1178.43   1028.19   869.71 568.35 25949  3041.07
   105.74 4.1 10828.62
…
;
proc reg;
model y=x1-x9/selection=adjrsq;
run;
```

第一行的 data data1；是给生成的数据集命名为 data1，这个名称可以自己随意起，也可以不命名数据集，直接是 data，行的结尾加分号。

第二行的 input x1-x9 y；是给变量赋值，其中 x1-x9 是一种简单的表示方法，等同于 x1 x2 x3 x4 x5 x6 x7 x8 x9。

第三行的 cards；是数据卡片，在其下输入数据，数据间用空格分开，依次输入变量的 n 个观测值。本例 $n=30$，每个观测值包含 10 个数据（9 个自变量和 1 个

专业课辅导+真题资料+微信公众号：432应统书院

因变量)。数据可以直接从 SPSS 或 Excel 或记事本数据文件中粘贴过来。本例数据量较大。只列出了前 5 个观测值，在数据的后面加分号。

数据下面的 proc reg；表示调用回归分析子过程。proc 是 procedure 的简写，reg 是 regression 的简写。

倒数第二行的 model y=x1-x9 指明回归方程中的变量，/后面的 selection=adjrsq 指明调用调整的复决定系数准则作所有子集回归。

最后一行 run；表示程序结束。

下一步就可以运行程序了，点击快捷键中的 submit 发送键，该键是个跑动的人形，也可以选中程序后直接用功能键 F8 或 F3 来运行程序。

部分输出结果见输出结果 5.1。

输出结果 5.1 中，第一个子集回归模型就是用 R_a^2 准则选出的最优子集回归模型，下面按照 R_a^2 从大到小的顺序列出了部分近似最优子集回归模型作为备选。最优子集为 x_1，x_2，x_3，x_5，x_6，x_7，x_9，$R_a^2=0.9921$。

输出结果 5.1

Number in Model	Adjusted R-Square	R-Square	Variables in Model
7	0.9921	0.9940	x1 x2 x3 x5 x6 x7 x9
8	0.9917	0.9940	x1 x2 x3 x5 x6 x7 x8 x9
8	0.9917	0.9940	x1 x2 x3 x4 x5 x6 x7 x9
6	0.9917	0.9934	x1 x2 x3 x5 x6 x7
7	0.9913	0.9934	x1 x2 x3 x4 x5 x6 x7
7	0.9913	0.9934	x1 x2 x3 x5 x6 x7 x8
9	0.9913	0.9940	x1 x2 x3 x4 x5 x6 x7 x8 x9
6	0.9911	0.9930	x1 x2 x3 x5 x6 x9
6	0.9910	0.9929	x1 x2 x3 x6 x7 x9
8	0.9910	0.9935	x1 x2 x3 x4 x5 x6 x7 x8
5	0.9908	0.9924	x1 x2 x3 x6 x7
7	0.9907	0.9930	x1 x2 x3 x4 x5 x6 x9
7	0.9907	0.9930	x1 x2 x3 x5 x6 x8 x9

专业课辅导+真题资料+微信公众号：432应统书院

7	0.9907	0.9929	x1 x2 x3 x6 x7 x8 x9
7	0.9906	0.9929	x1 x2 x3 x4 x6 x7 x9
6	0.9904	0.9924	x1 x2 x3 x6 x7 x8
6	0.9904	0.9924	x1 x2 x3 x4 x6 x7
8	0.9903	0.9930	x1 x2 x3 x4 x5 x6 x8 x9
8	0.9902	0.9929	x1 x2 x3 x4 x6 x7 x8 x9

输出结果列在 OUTPUT 窗口中，SAS 软件共有三个基本窗口，分别为：

（1）程序编辑窗口（PROGRAM EDITOR），用来编辑程序。用功能键 F5 可以直接切换到程序编辑窗口。

（2）日志窗口（LOG），显示已执行的语句和系统信息，错误信息用红色字体显示。用功能键 F6 可以直接切换到日志窗口。

（3）输出窗口（OUTPUT），显示程序运行结果。用功能键 F7 可以直接切换到输出窗口。

用主菜单的 Window 命令可以实现在三个窗口间的转换。进入日志窗口查看，没有错误信息，R_a^2 选元法到此结束。

例 5.3

对例 3.1 的数据，用 C_p 准则寻找最优子集回归模型。

将窗口切换到程序编辑窗口，只需将倒数第二行的 selection=adjrsq；改为 selection=cp，然后运行。若上次运行完后，关闭了程序编辑窗口，用主菜单的 Window 命令进入程序编辑窗口，这时发现窗口为空白，前面编写的用 R_a^2 准则选出最优子集回归模型的程序不见了。此时可以用主菜单 Locals 下的 Recall Text 命令重新调入。当 SAS 为汉语版本时，点击“运行”→“重新调用上一次提交”，即可重现上次的程序。部分输出结果见输出结果 5.2。

输出结果 5.2

Number in Model	C(p)	R-Square	Variables in Model
6	6.0220	0.9934	x1 x2 x3 x5 x6 x7

专业课辅导+真题资料+微信公众号：432应统书院

7	6.0555	0.9940	x1 x2 x3 x5 x6 x7 x9
5	7.3842	0.9924	x1 x2 x3 x6 x7
6	7.4948	0.9930	x1 x2 x3 x5 x6 x9
6	7.7182	0.9929	x1 x2 x3 x6 x7 x9
7	7.9691	0.9934	x1 x2 x3 x4 x5 x6 x7
7	7.9809	0.9934	x1 x2 x3 x5 x6 x7 x8
8	8.0001	0.9940	x1 x2 x3 x5 x6 x7 x8 x9
8	8.0336	0.9940	x1 x2 x3 x4 x5 x6 x7 x9
6	9.2719	0.9924	x1 x2 x3 x6 x7 x8
6	9.3458	0.9924	x1 x2 x3 x4 x6 x7
7	9.4177	0.9930	x1 x2 x3 x4 x5 x6 x9
7	9.4941	0.9930	x1 x2 x3 x5 x6 x8 x9
7	9.5808	0.9929	x1 x2 x3 x6 x7 x8 x9
5	9.6561	0.9917	x1 x2 x3 x6 x9
7	9.6947	0.9929	x1 x2 x3 x4 x6 x7 x9

……

用 C_p 准则寻找的最优子集回归模型为 x_1，x_2，x_3，x_5，x_6，x_7，$C_p=6.0220$，比用 R_a^2 准则选出的最优子集回归模型少一个变量 x_9。

5.3　逐步回归

在第 3 章的多元线性回归分析中，我们看到并不是所有自变量都对因变量 y 有显著的影响，这就存在着如何挑选出对因变量有显著影响的自变量的问题。自变量的所有可能子集构成 2^m-1 个回归方程，当可供选择的自变量不太多时，用前面的方法可以求出一切可能的回归方程，然后用几个选元准则去挑出最优的方程，但是当自变量的个数较多时，要求出所有可能的回归方程是非常困难的。为此，人们提出了一些较为简便、实用、快速的选择最优方程的方法。人们所给出的方法各有优

专业课辅导+真题资料+微信公众号：432应统书院

缺点，至今还没有绝对最优的方法，目前常用的方法有前进法、后退法、逐步回归法，而逐步回归法最受推崇。

在后面的讨论中，无论是从回归方程中剔除某个自变量，还是给回归方程增加某个自变量，都要利用（3.42）式的偏 F 检验，这个偏 F 检验与（3.40）式的 t 检验是等价的，F 检验的定义式的统计意义更为明了，并且容易推广到对多个自变量的显著性检验，因而采用 F 检验。

一、前进法

前进法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。具体做法是首先将全部 m 个自变量分别对因变量 y 建立 m 个一元线性回归方程，并分别计算这 m 个一元回归方程的 m 个回归系数的 F 检验值，记为 $\{F_1^1, F_2^1, \cdots, F_m^1\}$，选其最大者记为

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著性水平 α，若 $F_j^1 \geqslant F_\alpha(1, n-2)$，则首先将 x_j 引入回归方程，为了方便，设 x_j 就是 x_1。

接下来因变量 y 分别与 (x_1, x_2)，(x_1, x_3)，…，(x_1, x_m) 建立 $m-1$ 个二元线性回归方程，对这 $m-1$ 个回归方程中 x_2，x_3，…，x_m 的回归系数进行 F 检验，计算 F 值，记为 $\{F_2^2, F_3^2, \cdots, F_m^2\}$，选其最大者记为

$$F_j^2 = \max\{F_2^2, F_3^2, \cdots, F_m^2\}$$

若 $F_j^2 \geqslant F_\alpha(1, n-3)$，则接着将 x_j 引入回归方程。

依上述方法接着做下去，直至所有未被引入方程的自变量的 F 值均小于 $F_\alpha(1, n-p-1)$ 时为止。这时，得到的回归方程就是最终确定的方程。

每步检验中的临界值 $F_\alpha(1, n-p-1)$ 与自变量数目 p 有关，在用软件计算时，我们实际使用的是显著性 P 值（或记为 Sig）作检验。

例 5.4

对例 3.1 城镇居民消费性支出 y 关于 9 个自变量作回归的数据，用前进法作变量选择，取显著性水平 $\alpha_{entry}=0.05$。

首先进入线性回归对话框，将 y 与 x_1，x_2，x_3，x_4，x_5，x_6，x_7，x_8，x_9 分别选入各自的变量框，然后在 Model 下拉框中选前进法 Forward，点击 Options 选项看到默认的显著性水平 $\alpha_{entry}=0.05$。部分运行结果见输出结果 5.3。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 5.3

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-45.984	810.304		-.057	.955
	x1	2.630	.192	.932	13.661	.000
2	(Constant)	-2511.784	697.167		-3.603	.001
	x1	2.344	.140	.831	16.736	.000
	x2	4.438	.768	.287	5.782	.000
3	(Constant)	-3120.888	476.858		-6.545	.000
	x1	1.883	.122	.668	15.433	.000
	x2	3.836	.523	.248	7.339	.000
	x3	2.747	.467	.257	5.880	.000
4	(Constant)	-2697.889	314.311		-8.584	.000
	x1	1.557	.095	.552	16.434	.000
	x2	2.942	.366	.190	8.033	.000
	x3	2.484	.304	.232	8.183	.000
	x6	.069	.011	.205	6.152	.000
5	(Constant)	-2393.625	302.082		-7.924	.000
	x1	1.490	.088	.528	16.934	.000
	x2	2.717	.337	.176	8.061	.000
	x3	2.208	.289	.207	7.634	.000
	x6	.078	.011	.231	7.394	.000
	x7	.037	.014	.061	2.716	.012

a. Dependent Variable: y.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.932[a]	.870	.865	1006.62631
2	.970[b]	.942	.937	685.17084
3	.987[c]	.975	.972	457.45931
4	.995[d]	.990	.988	294.22252
5	.996[e]	.992	.991	262.61898

专业课辅导+真题资料+微信公众号：432应统书院

ANOVA[f]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.891E8	1	1.891E8	186.617	.000[a]
	Residual	2.837E7	28	1013296.531		
	Total	2.175E8	29			
2	Regression	2.048E8	2	1.024E8	218.119	.000[b]
	Residual	1.268E7	27	469459.074		
	Total	2.175E8	29			
3	Regression	2.120E8	3	7.068E7	337.731	.000[c]
	Residual	5440994.529	26	209269.020		
	Total	2.175E8	29			
4	Regression	2.153E8	4	5.383E7	621.793	.000[d]
	Residual	2164172.269	25	86566.891		
	Total	2.175E8	29			
5	Regression	2.158E8	5	4.316E7	625.837	.000[e]
	Residual	1655249.521	24	68968.730		
	Total	2.175E8	29			

从输出结果 5.3 中看到，前进法依次引入了 x_1，x_2，x_3，x_6，x_7，最优回归模型为

$$\hat{y}=-2\,393.625+1.490x_1+2.717x_2+2.208x_3+0.078x_6+0.037x_7$$

复决定系数 $R^2=0.992$，调整的复决定系数 $R_a^2=0.991$，全模型的复决定系数 $R^2=0.994$，调整的复决定系数 $R_a^2=0.991$。

二、后退法

后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。在第 3 章的回归系数的显著性检验中，用的就是这种思想，将回归系数检验的 F 值最小者对应的自变量剔除。

设对 m 个回归系数进行 F 检验，记求得的 F 值为 $\{F_1^m, F_2^m, \cdots, F_m^m\}$，选其最小者记为

$$F_j^m=\min\{F_1^m, F_2^m, \cdots, F_m^m\}$$

给定显著性水平 α，若 $F_j^m \leqslant F_\alpha(1, n-m-1)$，则首先将 x_j 从回归方程中剔除，为了方便，设 x_j 就是 x_m。

接着对剩下的 $m-1$ 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_j^{m-1}，如果又有 $F_j^{m-1} \leqslant F_\alpha(1, n-(m-1)-1)$，则剔除 x_j，重

专业课辅导+真题资料+微信公众号：432应统书院

新建立 y 关于 $m-2$ 个自变量的回归方程，依此类推，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 $F_\alpha(1, n-p-1)$，没有可剔除的自变量为止。这时，得到的回归方程就是最终确定的方程。

续例 5.4

对例 3.1 城镇居民消费性支出 y 关于 9 个自变量作回归的数据，用后退法作变量选择，取显著性水平 $\alpha_{removal}=0.10$。

首先进入线性回归对话框，将 y 与 x_1，x_2，x_3，x_4，x_5，x_6，x_7，x_8，x_9 分别选入各自的变量框，然后在 Model 下拉框中选前进法 Backward，点击 Options 选项看到默认的显著性水平 $\alpha_{removal}=0.10$。

部分运行结果见输出结果 5.4。

输出结果 5.4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2869.723	7962.399		-.360	.722
	x1	1.467	.150	.520	9.772	.000
	x2	2.577	.421	.167	6.115	.000
	x3	2.003	321	.187	6.233	.000
	x6	.066	.019	.198	3.509	.002
	x7	.028	.015	.045	1.828	.082
	x4	.005	.503	.000	.011	.992
	x5	.894	.473	.066	1.888	.074
	x8	12.776	70.233	.006	.182	.857
	x9	-161.486	121.816	-.032	-1.326	.200
	…	…	…	…	…	…
4	(Constant)	-2088.577	330.316		-6.323	.000
	x1	1.411	.094	.500	15.062	.000
	x2	2.393	.365	.155	6.563	.000
	x3	2.019	.293	.189	6.889	.000
	x6	.077	.010	.230	7.739	.000
	x7	.036	.013	.059	2.769	.011
	x5	.863	.461	.064	1.874	.074

专业课辅导+真题资料+微信公众号：432应统书院

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.994	.991	255.36124
2	.997[b]	.994	.992	249.20777
3	.997[c]	.994	.992	243.81209
4	.997[d]	.993	.992	249.87022

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.162E8	9	2.402E7	368.330	.000[a]
	Residual	1304187.292	20	65209.365		
	Total	2.175E8	29			
2	Regression	2.162E8	8	2.702E7	435.087	.000[b]
	Residual	1304194.782	21	62104.513		
	Total	2.175E8	29			
3	Regression	2.162E8	7	3.088E7	519.485	.000[c]
	Residual	1307775.395	22	59444.336		
	Total	2.175E8	29			
4	Regression	2.160E8	6	3.601E7	576.692	.000[d]
	Residual	1436007.929	23	62435.127		
	Total	2.175E8	29			

其中，模型 1 是全模型，从模型 2 到模型 4 依次剔除变量 x_4，x_8，x_9，最优回归子集模型 4 的回归方程为

$$\hat{y}=-2\,088.577+1.411x_1+2.393x_2+2.019x_3+0.077x_6+0.036x_7+0.863x_5$$

复决定系数 $R^2=0.993$，调整的复决定系数 $R_a^2=0.992$，全模型的复决定系数 $R^2=0.994$，调整的复决定系数 $R_a^2=0.991$。

前进法和后退法显然都有明显的不足。前进法可能存在这样的问题，即不能反映引进新的自变量后的变化情况。因为某个自变量开始可能是显著的，但当引入其他自变量后它就变得不显著了，但是也没有机会将其剔除，即一旦引入，就是“终身制”的。这种只考虑引入而没有考虑剔除的做法显然是不全面的。我们在许多例子中会发现可能最先引入的某个自变量——当其他自变量相继引入后——会变得对因变量 y 很不显著。

后退法的明显不足是，一开始把全部自变量引入回归方程，这样计算量很大。

专业课辅导+真题资料+微信公众号：432应统书院

如果有些自变量不太重要，一开始就不引入，就可减少一些计算量；再就是一旦某个自变量被剔除，它就再也没有机会重新进入回归方程。

如果说我们的问题涉及的自变量 x_1，x_2，…，x_m 是完全独立的（或不相关），那么在取 $\alpha_{进}=\alpha_{出}$ 时，前进法与后退法所建的回归方程是相同的。然而在实际中很难碰到自变量间真正无关的情况，尤其是经济问题中，我们所研究的绝大部分问题，自变量间都有一定的相关性。这就使得随着回归方程中变量的增加和减少，某些自变量对回归方程的影响也会发生变化。这是因为自变量间的不同组合，由于它们相关的原因，对因变量 y 的影响可能大不一样。如果几个自变量的联合效应对 y 有重要作用，但是单个自变量对 y 的作用都不显著，那么前进法就不能引入这几个自变量，而后退法却可以保留这几个自变量，这是后退法的一个优点。

根据前进法和后退法的思想及方法以及它们的不足，人们比较自然地想构造一种方法，即吸收前进法和后退法的优点，克服它们的不足，把两者结合起来，这就产生了逐步回归法。

三、逐步回归法

逐步回归的基本思想是有进有出。具体做法是将变量一个一个引入，每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。引入一个变量或从回归方程中剔除一个变量，为逐步回归的一步，每一步都要进行 F 检验，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。这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中剔除为止。这样就避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是最优回归子集。

在逐步回归法中需要注意的一个问题是引入自变量和剔除自变量的显著性水平 α 值是不同的，要求引入自变量的显著性水平 $\alpha_{进}$ 小于剔除自变量的显著性水平 $\alpha_{出}$，否则可能产生“死循环”。也就是当 $\alpha_{进}\geqslant\alpha_{出}$ 时，如果某个自变量的显著性 P 值在 $\alpha_{进}$ 与 $\alpha_{出}$ 之间，那么这个自变量将被引入、剔除，再引入、再剔除，循环往复，以至无穷。

逐步回归的计算过程可以利用 SPSS 软件在计算机上自动完成，我们要求关心应用的读者一定要通过前面的叙述掌握逐步回归方法的思想，这样才能用对用好逐步回归法。

例 5.5

本例为回归分析中经典的 Hald 水泥问题。某种水泥在凝固时放出的热量 y（卡/克，cal/g）与水泥中的四种化学成分的含量（%）有关，这四种化学成分分别是 x_1 铝酸三钙（$3CaO\cdot Al_2O_3$），x_2 硅酸三钙（$3CaO\cdot SiO_2$），x_3 铁铝酸四钙（$4CaO\cdot Al_2O_3\cdot Fe_2O_3$），$x_4$ 硅酸二钙（$2CaO\cdot SiO_2$）。现观测到 13 组数据，如表 5.3 所示。本例用逐

专业课辅导+真题资料+微信公众号：432应统书院

步回归法作变量选择，希望从中选出主要的变量，建立 y 关于四种成分的线性回归方程。

表 5.3

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

取显著性水平 $\alpha_{entry}=0.1$，$\alpha_{removal}=0.15$。首先进入线性回归对话框，将 y 与 x_1，x_2，x_3，x_4 分别选入各自的变量框，然后在 Model 下拉框中选前进法 Stepwise，点击 Options 选项看到默认的显著性水平 $\alpha_{entry}=0.1$，$\alpha_{removal}=0.15$。部分运行结果见输出结果 5.5。

输出结果 5.5

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.821[a]	.675	.645	8.96390
2	.986[b]	.972	.967	2.73427
3	.991[c]	.982	.976	2.30874
4	.989[d]	.979	.974	2.40634

a. Predictors: (Constant), x4.

b. Predictors: (Constant), x4, x1.

c. Predictors: (Constant), x4, x1, x2.

d. Predictors: (Constant), x1, x2.

专业课辅导+真题资料+微信公众号：432应统书院

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1831.896	1	1831.896	22.799	.001[a]
	Residual	883.867	11	80.352		
	Total	2715.763	12			
2	Regression	2641.001	2	1320.500	176.627	.000[b]
	Residual	74.762	10	7.476		
	Total	2715.763	12			
3	Regression	2667.790	3	889.263	166.832	.000[c]
	Residual	47.973	9	5.330		
	Total	2715.763	12			
4	Regression	2657.859	2	1328.929	229.504	.000[d]
	Residual	57. 904	10	5.790		
	Total	2715.763	12			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	117.568	5.262		22.342	.000
	x4	-.738	.155	-.821	-4.775	.001
2	(Constant)	103.097	2.124		48.540	.000
	x4	-.614	.049	-.683	-12.621	.000
	x1	1.440	.138	.563	10.403	.000
3	(Constant)	71.648	14.142		5.066	.001
	x4	-.237	.173	-.263	-1.365	.205
	x1	1.452	.117	.568	12.410	.000
	x2	.416	.186	.430	2.242	.052
4	(Constant)	52.577	2.286		22.998	.000
	x1	1.468	.121	.574	12.105	.000
	x2	.662	.046	.685	14.442	.000

a. Dependent Variable: y

从输出结果 5.5 看到，逐步回归的最优子集为模型 4，回归方程为

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{y}=52.577+1.468x_1+0.662x_2$$

由回归方程可以看出，对水泥凝固时放热量有显著影响的是水泥中铝酸三钙和硅酸三钙的含量，回归方程中两个自变量的系数都为正，即，水泥中铝酸三钙和硅酸三钙的含量越高，每克水泥凝固时放出的热量越多。具体地说，在 x_2 含量保持不变时，x_1 含量每增加一个百分点，每克水泥凝固时放出的热量平均增多 1.468cal；在 x_1 含量保持不变时，x_2 含量每增加一个百分点，每克水泥凝固时放出的热量平均增多 0.662cal。

同时，在输出结果 5.5 中可以看到逐步回归的选元过程。本例逐步回归的选元过程为第一步引入 x_4，第二步引入 x_1，第三步引入 x_2，第四步剔除 x_4。从本例逐步回归的选元过程可以看出逐步回归方法的有进有出的思想，在第一步引入的自变量 x_4，在第四步又被剔除了。这种有进有出的结果说明变量之间有相关性，如果自变量之间是完全不相关的，那么引入的自变量就不会再次被剔除，而剔除的自变量也就不会再次被引入，这时逐步回归方法与前进法是相同的。在实际问题中，自变量之间的通常存在相关性，当相关性程度严重时称为多重共线性。自变量之间的多重共线性会对回归产生极大的影响，第 6 章将详细讨论多重共线性问题。

5.4 本章小结与评注

一、逐步回归实例分析

为了系统掌握逐步回归的思想及其应用，再举一个实际经济问题用逐步回归方法建模的例子。

例 5.6

为了研究香港股市的变化规律，此例以恒生指数为例，建立回归方程，分析影响股票价格趋势变动的因素。这里研究的股票价格指数，并非某一种股票的价格，它是综合反映股票市场上所有上市股票价格整体水平变化的指标。这里我们选了 7 个影响股票价格指数的经济变量：x_2 为 99 金价（港元/两）（两为香港现行黄金计量单位，因本题为实际课题，故不作变动）；x_3 为港汇指数；x_4 为人均生产总值（现价港元）；x_5 为建筑业总开支（现价百万港元）；x_6 为房地产买卖金额（百万港元）；x_7 为优惠利率（最低%）。其中，x_2，x_3，x_7 分别从贵金属、汇率和利率方面反映金融环境的影响；x_4，x_5，x_6 则从不同方面反映了整体经济状况。由于市场环境状况对股价也有十分重要的影响，我们选择成交额 x_1（百万港元）来反映市场状况。y 为恒生指数。我们收集了以上变量 1974—1988 年 15 年的数据资料，见表 5.4。逐步回归的部分输出结果见输出结果 5.6。

专业课辅导+真题资料+微信公众号：432应统书院

表 5.4

年份	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1974	172.9	11 246	681	105.9	10 183	4 110	11 242	9
1975	352.94	10 335	791	107.4	10 414	3 996	12 693	6.5
1976	447.67	13 156	607	114.4	13 134	4 689	16 681	6
1977	404.02	6 127	714	110.8	15 033	6 876	22 131	4.75
1978	409.51	27 419	911	99.4	17 389	8 636	31 353	4.75
1979	619.71	25 633	1 231	91.1	21 715	12 339	43 528	9.5
1980	1 121.17	95 684	2 760	90.8	27 075	16 623	70 752	10
1981	1 506.84	105 987	2 651	86.3	31 827	19 937	125 989	16
1982	1 105.79	46 230	2 105	125.3	35 393	24 787	99 468	10.5
1983	933.03	37 165	3 030	107.4	38 832	25 112	82 478	10.5
1984	1 008.54	48 787	2 810	106.6	46 079	24 414	54 936	8.5
1985	1 567.56	75 808	2 649	115.7	47 871	22 970	87 135	6
1986	1 960.06	123 128	3 031	110.1	54 372	24 403	129 884	6.5
1987	2 884.88	371 406	3 644	105.8	65 602	30 531	153 044	5
1988	2 556.72	198 569	3 690	101.6	74 917	37 861	215 033	5.25

输出结果 5.6

相关系数阵

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.0000	0.9171	0.8841	−0.0425	0.9382	0.8786	0.9372	−0.0955
X1	0.9171	1.0000	0.7375	−0.1293	0.7842	0.6973	0.7817	−0.1732
X2	0.8841	0.7375	1.0000	−0.1083	0.9195	0.9477	0.8747	0.1517
X3	−0.0425	−0.1293	−0.1083	1.0000	0.0725	0.0469	−0.0952	−0.4164
X4	0.9382	0.7842	0.9195	0.0725	1.0000	0.9601	0.9137	−0.1409
X5	0.8786	0.6973	0.9477	0.0469	0.9601	1.0000	0.9167	0.0666
X6	0.9372	0.7817	0.8747	−0.0952	0.9137	0.9167	1.0000	0.0617
X7	−0.0955	−0.1732	0.1517	−0.4164	−0.1409	0.0666	0.0617	1.0000

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.938[a]	.880	.871	295.5760
2	.983[b]	.966	.960	164.5798
3	.991[c]	.981	.976	126.4937

a. Predictors: (Constant), x4.

b. Predictors: (Constant), x4, x1.

c. Predictors: (Constant), x4, x1, x6.

专业课辅导+真题资料+微信公众号：432应统书院

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8347442.1	1	8347442.1	95.547	.000
	Residual	1135747.1	13	87365.165		
	Total	9483189.3	14			
2	Regression	9158151.1	2	4579040.362	169.054	.000
	Residual	325038.20	12	27086.517		
	Total	9483189.3	14			
3	Regression	9307181.9	3	3102394.0	193.892	.000
	Residual	176007.34	11	16000.667		
	Total	9483189.3	14			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-147.176	151.913		-.969	.350
	X4	3.78E-02	.004	.938	9.775	.000
2	(Constant)	38.563	91.146		.423	.680
	X4	2.29E-02	.003	.569	6.604	.000
	X1	4.01E-03	.001	.471	5.471	.000
3	(Constant)	75.807	71.109		1.066	.309
	X4	1.29E-02	.004	.319	3.038	.011
	X1	3.55E-03	.001	.417	6.086	.000
	X6	4.42E-03	.001	.319	3.052	.011

下面根据回归方程和相关系数矩阵，从定性与定量的结合上分析股票价格指数的成因。

在相关矩阵中，我们看到

$$r_{y,3}=-0.0425,\ r_{y,7}=-0.0955$$

这说明港汇指数和优惠利率对恒生指数的影响不大。中国香港作为国际金融中心之一，它的证券市场是向国际开放的。事实上，1987 年以前，香港证券市场上的股份所有权有 50%以上掌握在外国经营机构手中，因此，从理论上讲，作为反映港币汇率水平的主要指标港汇指数应该与股票价格高度相关，但事实并非如此。原因何在？观察 1974—1988 年的港汇指数值，可以看出除 1981 年、1982 年出现大起大落外，港汇指数一直处于比较平稳的状态，说明港币比较坚挺。（至于 1981 年、1982 年应把它们视为特殊年份，1981 年提出香港回归问题，1982 年英首相访华，正是这一连串的政治事件造成了港币汇率的大幅波动。）由于汇率波动不大，自然对股价不会产生很大的影响。其次，优惠利率 x_7 指的是贷款利率，而从股份所有权估计看，香港股市投资活动中绝大部分是由国外经营机构和私人进行的，因而优惠利率对股价影响不大。这样看来，回归方程中没有引进 x_3 和 x_7 也是合乎情理的。

从自变量之间的关系来看，有两组自变量之间都高度相关。一组是 x_4，x_5，x_6，其中

专业课辅导+真题资料+微信公众号：432应统书院

$$r_{4,5}=0.960\,1,\quad r_{4,6}=0.913\,7,\quad r_{5,6}=0.916\,7$$

先看建筑业总开支 x_5 与房地产买卖金额 x_6，建筑业的发达自然会引起房地产买卖的兴旺，同样房地产炒得热也会刺激建筑业的发展，二者存在正相关。人均生产总值 x_4 综合反映香港地区的经济发展水平，自然也包括了建筑业的发展。所以 x_4，x_5 是包含与被包含的关系，可以认为 x_5 所反映的内容包含于 x_4 反映的内容之中，x_5 即成为多余变量。由于 x_4 与 x_5，x_5 与 x_6 高度相关，必然引起 x_4 与 x_6 高度相关，从实际意义来看，可以得到同样的结论。由于不动产是香港投资商致富的主要源泉，房地产买卖是香港经济十分重要的组成部分，因而它与经济总水平必然高度相关，这就造成了 x_4，x_6 的相关系数较高。另一组高度相关的变量是 x_2 与 x_4，x_5，x_6，其中

$$r_{2,4}=0.919\,5,\quad r_{2,5}=0.947\,7,\quad r_{2,6}=0.874\,7$$

99 金价 x_2 变动频繁，黄金市场对环境因素的影响十分敏感，这都造成了金价与外部经济因素密切相关。而 x_4，x_5，x_6 都是反映经济状况的指标，因而都与 x_2 密切相关。

综上所述，可以看到逐步回归方程只保留 x_1，x_4，x_6 是合适的，即

$$y=75.807+0.003\,55x_1+0.012\,9x_4+0.004\,42x_6$$

如果进一步作回归诊断，可以发现该回归模型满足正态性假设，无异方差、无序列相关等。因此，运用该回归方程可以对恒生指数的变动成因作一些分析。

从上述回归方程来看，影响恒生指数的主要因素为成交额、人均生产总值和房地产买卖金额。成交额作为反映市场因素的主要指标对股票价格有着重要的影响。香港股市上，成交额每增长 100 万港元，恒生指数平均上涨 0.003 55 个百分点。人均生产总值是反映经济状况的主要指标，它代表了经济环境对股票价格的影响，香港人均生产总值每上升 100 港元，恒生指数平均上涨 1.29 个百分点。另外，房地产买卖金额每增加 100 万港元，恒生指数平均上涨 0.004 42 个百分点，这是香港股市区别于其他股票市场的一大特色。香港的证券市场反映了香港的财政与贸易活动，但证券市场的大部分资金却投入了房地产部门，因为不动产是香港投资商致富的主要源泉。因此，房地产事业相应地对股票市场产生了重大影响，它的影响程度甚至大于其他所有因素。这是香港股市的一大特色。

例 5.6 是我们运用逐步回归方法分析实际经济问题的一个较为典型的例子，从这个例子读者可看到用逐步回归选择自变量是一种较为实用的方法。

二、评注

从本章 5.1 节讨论的自变量选择对参数估计和预测的影响来看，自变量的选择是回归分析建模中的一个非常重要的基本问题。在对一个实际经济问题建立回归模型时，我们首先根据经济理论和采集样本数据的条件限制，来定性地确定一些对所

专业课辅导+真题资料+微信公众号：432应统书院

研究的经济现象有重要影响的因素，这些因素就是所谓的自变量。由于人们认识事物的水平的局限，从事物的表面很难分清哪些自变量对因变量有重要影响，哪些自变量间存在着密切的相关性。人们通常认为研究某个经济现象的回归问题，考虑得越细越周到肯定越好，自然就会罗列出很多自变量。通过自变量选择对参数估计和预测的影响分析，我们得到的重要结论是，回归方程并非自变量越多越好，当一些对因变量影响不大的自变量进入回归方程后，反而会使参数估计的稳定性变差，预测误差的方差增大。因此，回归模型中应该保留对因变量影响最显著的变量，即变量的个数和质量要求是少而精。

由于变量之间的相关性，自变量间不同的组合对因变量 y 的影响是不一样的，到底哪些自变量子集对应的回归方程是最优的方程，这就要根据我们介绍的几个衡量准则在所有自变量子集中去挑选。当所研究的问题有 m 个自变量时，就有 2^m-1 个自变量子集，每一个自变量子集对应一个回归方程，这个回归方程称为回归子集。挑选最优的回归方程就是选择最优自变量子集。这里的最优实际上是指一个相对好的回归方程，没有绝对的最优。我们所选的最优回归方程也是根据研究问题的性质和目的，用不同的准则来衡量的结果。同一个回归子集在不同的准则衡量下结果可能是不一样的。

选择哪一个回归子集，用哪一个衡量准则要根据我们研究问题的目的来决定。回归模型常用的三个方面是：结构分析、预测、控制。如果我们想通过回归模型去研究经济变量之间的相互联系，即作结构分析，则在选元时可考虑适当放宽选元标准，让回归方程保留较多的自变量，但这时需注意回归系数的正负号，看它们是否符合经济意义。如果我们希望回归方程简单明了，易于理解，则应采用较严的选元标准。比如在逐步回归选元中，给显著性水平 $\alpha_{进}$ 赋一个较小的值，就可使得回归方程中保留较少最重要、最能说明问题的自变量。如果我们建立回归方程的目的是用于控制，就应采用能使回归参数的估计标准误差值尽可能小的准则。如果建立回归方程的目的是用于预测，就应考虑使得预测值的均方误差尽量小的准则，如 C_p 准则。

一般来说，一个好的回归方程往往在几个准则衡量下都较优，如例 5.1 中 y 关于 x_1，x_3 的回归方程，y 关于 x_1，x_2，x_3 的回归方程，它们分别用 R_a^2，AIC，C_p 准则衡量，从表 5.2 中看到这些指标相对较好，说明 y 关于 x_1，x_2，x_3 的方程是最优的，y 关于 x_1，x_3 的回归方程是次优的。

当所研究的问题涉及的自变量较多时，即使针对某一给定的用途，根据某种准则也往往会发现自变量子集有几组几乎同样好，这时就要附加其他信息。整个选择过程应该注重实效，并要进行大量的主观判断。有学者认为统计学是研究、分析数据的艺术。实际上是说我们不应过于依赖什么准则，不应单纯地机械搬用，要注意运用的技巧，综合各方面信息，选择最优回归模型。

还需说明的是，由所选择的自变量子集并不能完全决定要使用的模型，还必须作其他的判定，如自变量是否是线性的，是否要用变换的形式或者是否要用二次项，以及模型是否应该包含交互作用项等。比如有三个基本变量 x_1，x_2，x_3，还可

专业课辅导+真题资料+微信公众号：432应统书院

考虑 $x_4=x_1x_3, x_5=x_2^2, x_6=\ln x_2$ 等，这些问题将在第 9 章非线性回归中进一步讨论。在本章中所介绍的选元方法假定研究人员已考虑好了回归关系的函数形式，自变量或者因变量是否要首先进行变换，以及是否要包括交互作用项。这些工作都可看做数据的预处理，如上面的 $x_4=x_1x_3, x_5=x_2^2, x_6=\ln x_2$ 等。在上述前提下，使用选元方法，以达到寻求最优回归方程的目的。

对 p 个自变量的线性回归问题，所有可能的回归方程有 2^m-1 个，从 2^m-1 个回归中如何选择出某种准则意义上的最优回归方程，计算方法是十分重要的。20 世纪 60 年代，一些统计学家提出的一些算法基本上只能处理含 10～12 个自变量的回归问题。而弗尼尔（Furnial）和威尔逊（Wilson）提出的算法较完美地解决了节省计算量、存储量以及减少计算误差的问题，它可以计算含 30 多个自变量的所有可能的子集回归，而所需的计算时间与逐步回归大体相当（参见参考文献［2］）。弗尼尔和威尔逊的方法尽管设计得很巧妙，但自变量多于 30 的大型回归问题，其计算量仍然很大。逐步回归目前被认为是研究多个自变量建模较为理想的方法，其应用已非常普遍。

在第 3 章中我们曾强调在建立一个实际问题的回归方程时，样本量的个数 n 一定要大于自变量的个数 p，即 $n>p$，如果不满足这一条件，则无法运用普通最小二乘估计，F 检验、t 检验也无法进行。然而，逐步回归是每次只引入或剔除一个自变量，所以对 $n<p$ 的情形也可进行回归子集的选择。参考文献[1]的第 12 章有一个研究我国国债的例子，涉及 18 个自变量，而我国国债的发行时间不长，只收集了 10 年的样本数据，即 $n=10$，$p=18$，$n<p$。在这种情形下无法使用普通最小二乘估计，但用逐步回归选择了一个含 4 个自变量的回归模型，用各种准则衡量都不错，它的平均残差平方和比主成分回归还要好，预测效果也较为理想，这是逐步回归又一个好的例证。

□ 思考与练习

5.1　自变量选择对回归参数的估计有何影响？

5.2　自变量选择对回归预测有何影响？

5.3　如果所建模型主要用于预测，应该用哪个准则来衡量回归方程的优劣？

5.4　试述前进法的思想、方法。

5.5　试述后退法的思想、方法。

5.6　前进法、后退法各有哪些优缺点？

5.7　试述逐步回归法的思想、方法。

5.8　在运用逐步回归法时，$\alpha_{进}$ 与 $\alpha_{出}$ 的赋值原则是什么？如果希望回归方程中多保留一些自变量，$\alpha_{进}$ 应如何赋值？

5.9　在研究国家财政收入时，我们把财政收入按收入形式分为：各项税收收入、企业收入、债务收入、国家能源交通重点建设基金收入、基本建设贷款归还收

专业课辅导+真题资料+微信公众号：432应统书院

入、国家预算调节基金收入、其他收入等。为了建立国家财政收入回归模型，我们以财政收入 y（亿元）为因变量，自变量如下：x_1 为农业增加值（亿元）；x_2 为工业增加值（亿元）；x_3 为建筑业增加值（亿元）；x_4 为人口数（万人）；x_5 为社会消费总额（亿元）；x_6 为受灾面积（万公顷）。据《中国统计年鉴》获得 1978—1998 年共 21 个年份的统计数据，见表 5.5。由定性分析知，所选自变量都与变量 y 有较强的相关性，分别用后退法和逐步回归法作自变量选元。

表 5.5

年份	农业增加值 x_1	工业增加值 x_2	建筑业增加值 x_3	人口数 x_4	社会消费总额 x_5	受灾面积 x_6	财政收入 y
1978	1 018.4	1 607.0	138.2	96 259	2 239.1	50 760	1 132.3
1979	1 258.9	1 769.7	143.8	97 542	2 619.4	39 370	1 146.4
1980	1 359.4	1 996.5	195.5	98 705	2 976.1	44 530	1 159.9
1981	1 545.6	2 048.4	207.1	100 072	3 309.1	39 790	1 175.8
1982	1 761.6	2 162.3	220.7	101 654	3 637.9	33 130	1 212.3
1983	1 960.8	2 375.6	270.6	103 008	4 020.5	34 710	1 367.0
1984	2 295.5	2 789.0	316.7	104 357	4 694.5	31 890	1 642.9
1985	2 541.6	3 448.7	417.9	105 851	5 773.0	44 370	2 004.8
1986	2 763.9	3 967.0	525.7	107 507	6 542.0	47 140	2 122.0
1987	3 204.3	4 585.8	665.8	109 300	7 451.2	42 090	2 199.4
1988	3 831.0	5 777.2	810.0	111 026	9 360.1	50 870	2 357.2
1989	4 228.0	6 484.0	794.0	112 704	10 556.5	46 990	2 664.9
1990	5 017.0	6 858.0	859.4	114 333	11 365.2	38 470	2 937.1
1991	5 288.6	8 087.1	1 015.1	115 823	13 145.9	55 470	3 149.5
1992	5 800.0	10 284.5	1 415.0	117 171	15 952.1	51 330	3 483.4
1993	6 882.1	14 143.8	2 284.7	118 517	20 182.1	48 830	4 349.0
1994	9 457.2	19 359.6	3 012.6	119 850	26 796.0	55 040	5 218.1
1995	11 993.0	24 718.3	3 819.6	121 121	33 635.0	45 821	6 242.2
1996	13 844.2	29 082.6	4 530.5	122 389	40 003.9	46 989	7 408.0
1997	14 211.2	32 412.1	4 810.6	123 626	43 579.4	53 429	8 651.1
1998	14 599.6	33 429.8	5 262.0	124 810	46 405.9	50 145	9 876.0

5.10　表 5.6 的数据是 1968—1983 年间美国与电话线制造有关的数据，各变量的含义如下：

x_1——年份；

x_2——国民生产总值（10 亿美元）；

x_3——新房动工数（单位：1 000）；

x_4——失业率（%）；

x_5——滞后 6 个月的最惠利率；

x_6——用户用线增量（%）；

y——年电话线销量（百万尺双线）。

（1）建立 y 对 $x_2 \sim x_6$ 的线性回归方程。

（2）用后退法选择自变量。

专业课辅导+真题资料+微信公众号：432应统书院

(3) 用逐步回归法选择自变量。

(4) 根据以上计算结果分析后退法与逐步回归法的差异。

表 5.6

x_1	x_2	x_3	x_4	x_5	x_6	y
1968	1 051.8	1 503.6	3.6	5.8	5.9	5 873
1969	1 078.8	1 486.7	3.5	6.7	4.5	7 852
1970	1 075.3	1 434.8	5.0	8.4	4.2	8 189
1971	1 107.5	2 035.6	6.0	6.2	4.2	7 494
1972	1 171.1	2 360.8	5.6	5.4	4.9	8 534
1973	1 235.0	2 043.9	4.9	5.9	5.0	8 688
1974	1 217.8	1 331.9	5.6	9.4	4.1	7 270
1975	1 202.3	1 160.0	8.5	9.4	3.4	5 020
1976	1 271.0	1 535.0	7.7	7.2	4.2	6 035
1977	1 332.7	1 961.8	7.0	6.6	4.5	7 425
1978	1 399.2	2 009.3	6.0	7.6	3.9	9 400
1979	1 431.6	1 721.9	6.0	10.6	4.4	9 350
1980	1 480.7	1 290.8	7.2	14.9	3.9	6 540
1981	1 510.3	1 100.0	7.6	16.6	3.1	7 675
1982	1 492.2	1 039.0	9.2	17.5	0.6	7 419
1983	1 535.4	1 200.0	8.8	16.0	1.5	7 923

专业课辅导+真题资料+微信公众号：432应统书院

第 6 章 Chapter 6 多重共线性的情形及其处理

多元线性回归模型有一个基本假设，就是要求设计矩阵 $\boldsymbol{X}$ 的秩 $\mathrm{rank}(\boldsymbol{X})=p+1$，即要求 $\boldsymbol{X}$ 中的列向量之间线性无关。如果存在不全为零的 $p+1$ 个数 c_0，c_1，c_2，…，c_p，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}=0,\quad i=1,2,\cdots,n \tag{6.1}$$

则自变量 x_1，x_2，…，x_p 之间存在完全多重共线性。在实际问题中，完全的多重共线性并不多见，常见的是（6.1）式近似成立的情况，即存在不全为零的 $p+1$ 个数 c_0，c_1，c_2，…，c_p，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}\approx 0,\quad i=1,2,\cdots,n \tag{6.2}$$

当自变量 x_1，x_2，…，x_p 存在（6.2）式的关系时，称自变量 x_1，x_2，…，x_p 之间存在多重共线性（multi-collinearity），也称为复共线性。在实际经济问题的多元回归分析中，多重共线性的情形很多，如何诊断变量间的多重共线性，多重共线性会给多元线性回归分析带来什么影响，如何克服多重共线性的影响，这些问题就是我们在本章要讨论的主要内容。

6.1 多重共线性产生的背景和原因

解释变量之间完全不相关的情形是非常少见的，尤其是研究某个经济问题时，涉及的自变量较多，我们很难找到一组自变量，它们之间互不相关，而且它们又都对因变量有显著影响。客观地说，当某一经济现象涉及多个影响因素时，这些影响因素之间大都有一定的相关性。当它们之间的相关性较弱时，我们一般就认

专业课辅导+真题资料+微信公众号：432应统书院

为符合多元线性回归模型设计矩阵的要求；当这一组变量间有较强的相关性时，就认为是一种违背多元线性回归模型基本假设的情形。

当所研究的经济问题涉及时间序列资料时，由于经济变量随时间往往存在共同的变化趋势，使得它们之间容易出现共线性。例如，我国近年来的经济增长态势很好，经济增长对各种经济现象都产生影响，使得多种经济指标相互密切关联。比如要研究我国居民消费状况，影响居民消费的因素很多，一般有职工平均工资、农民平均收入、银行利率、全国零售物价指数、国债利率、货币发行量、储蓄额、前期消费额等，这些因素显然既对居民消费产生重要影响，它们之间又有着很强的相关性。

对于许多利用截面数据建立回归方程的问题，常常也存在自变量高度相关的情形。例如，以企业的截面数据为样本估计生产函数，由于投入要素资本 K、劳动力投入 L、科技投入 S、能源供应 E 等都与企业的生产规模有关，所以它们之间存在较强的相关性。

再如，有人在建立某地区粮食产量回归模型时，以粮食产量为因变量 y，以化肥用量 x_1，水浇地面积 x_2，农业资金投入 x_3 等作为自变量。从表面上我们看到 x_1，x_2，x_3 都是影响粮食产量 y 的重要因素，可是建立的 y 关于 x_1，x_2，x_3 的回归方程效果很差，原因是什么？后来发现尽管所选自变量 x_1，x_2，x_3 都是影响因变量 y 的重要因素，但是农业投入资金 x_3 与化肥用量 x_1、水浇地面积 x_2 有很强的相关性，农业资金投入主要用于购买化肥和开发水利，也就是说，资金投入的效应已被化肥用量和水浇地面积体现出来。进一步计算 x_3 分别与 x_1，x_2 的简单相关系数得 $r_{13}=0.98$，$r_{23}=0.99$，呈现高度相关。剔除 x_3 后重新建立回归模型，结果无论从预测还是结构分析来看都十分理想。

在研究社会、经济问题时，鉴于问题本身的复杂性，涉及的因素往往很多。在建立回归模型时，由于研究者认识水平的局限性，很难在众多因素中找到一组互不相关又对因变量 y 有显著影响的变量，不可避免地会出现所选自变量相关的情形。当自变量之间有较强的相关性时，会给回归模型的参数估计带来什么样的后果，这就是下面我们要讨论的问题。

6.2　多重共线性对回归模型的影响

设回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

存在完全的多重共线性，即设计矩阵 $\boldsymbol{X}$ 的列向量存在不全为零的一组数 c_0，c_1，c_2，…，c_p，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}=0,\quad i=1,2,\cdots,n$$

专业课辅导+真题资料+微信公众号：432应统书院

设计矩阵 $\boldsymbol{X}$ 的秩 $\mathrm{rank}(\boldsymbol{X})<p+1$，此时 $|\boldsymbol{X}'\boldsymbol{X}|=0$，正规方程组 $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{y}$ 的解不唯一，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$不存在，回归参数的最小二乘估计表达式 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$不成立。

在实际问题的研究中，经常见到的是近似共线性的情形，即存在不全为零的一组数 c_0，c_1，c_2，…，c_p，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}\approx 0,\quad i=1,2,\cdots,n$$

此时设计矩阵 $\boldsymbol{X}$ 的秩 $\mathrm{rank}(\boldsymbol{X})=p+1$ 虽然成立，但是此时$|\boldsymbol{X}'\boldsymbol{X}|\approx 0$，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$的对角线元素很大，$\hat{\boldsymbol{\beta}}$ 的方差阵 $D(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$的对角线元素很大，而$D(\hat{\boldsymbol{\beta}})$的对角线元素即$\mathrm{var}(\hat{\beta}_0)$，$\mathrm{var}(\hat{\beta}_1)$，…，$\mathrm{var}(\hat{\beta}_p)$，因而 β_0，β_1，…，β_p 的估计精度很低。这样，虽然用普通最小二乘估计能得到 $\boldsymbol{\beta}$ 的无偏估计，但估计量 $\hat{\boldsymbol{\beta}}$ 的方差很大，不能正确判断解释变量对被解释变量的影响程度，甚至导致估计量的经济意义无法解释。这样的情况在进行实际问题的回归分析时会经常碰到。

从下面的二元回归的简单例子的讨论中，能够看到当自变量间的相关性从小到大增加时，估计量的方差增大得很快。

作 y 对两个自变量 x_1，x_2 的线性回归，假定 y 与 x_1，x_2 都已经中心化，此时回归常数项为零，回归方程为

$$\hat{y}=\hat{\beta}_1x_1+\hat{\beta}_2x_2$$

记 $L_{11}=\sum_{i=1}^{n}x_{i1}^2$，$L_{12}=\sum_{i=1}^{n}x_{i1}x_{i2}$，$L_{22}=\sum_{i=1}^{n}x_{i2}^2$，则 x_1 与 x_2 之间的相关系数为

$$r_{12}=\frac{L_{12}}{\sqrt{L_{11}L_{22}}}$$

$\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\hat{\beta}_2)'$ 的协方差阵为

$$\mathrm{cov}(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\boldsymbol{X}'\boldsymbol{X}=\begin{bmatrix}L_{11} & L_{12}\\ L_{12} & L_{22}\end{bmatrix}$$

$$\begin{aligned}(\boldsymbol{X}'\boldsymbol{X})^{-1}&=\frac{1}{|\boldsymbol{X}'\boldsymbol{X}|}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}\\&=\frac{1}{L_{11}L_{22}-L_{12}^2}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}\\&=\frac{1}{L_{11}L_{22}(1-r_{12}^2)}\begin{bmatrix}L_{22} & -L_{12}\\ -L_{12} & L_{11}\end{bmatrix}\end{aligned}$$

由此可得

$$\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{(1-r_{12}^2)L_{11}}\tag{6.3}$$

专业课辅导+真题资料+微信公众号：432应统书院

$$\mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{(1-r_{12}^2)L_{22}} \tag{6.4}$$

可知，随着自变量 x_1 与 x_2 的相关性增强，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差将逐渐增大。当 x_1 与 x_2 完全相关时，$r=1$，方差将变为无穷大。

当给定不同的 r_{12} 值时，我们由表 6.1 可看出方差增大的速度。为了方便，我们假设 $\sigma^2/L_{11}=1$，相关系数从 0.5 变为 0.9 时，回归系数的方差增加了 295%，相关系数从 0.5 变为 0.95 时，回归系数的方差增加了 671%。回归自变量 x_1 与 x_2 的相关程度越高，多重共线性越严重，回归系数的估计值方差就越大，回归系数的置信区间就变得很宽，估计的精确性大幅度降低，使估计值稳定性变得很差，进一步致使在回归方程整体高度显著时，一些回归系数通不过显著性检验，回归系数的正负号也可能出现倒置，使得回归方程无法得到合理的经济解释，直接影响到最小二乘法的应用效果，降低回归方程的应用价值。

表 6.1

r_{12}	0.2	0.50	0.70	0.80	0.90	0.95	0.99	1.00
$\mathrm{var}(\hat{\beta}_1)$	1.04	1.33	1.96	2.78	5.26	10.26	50.25	∞

在第 3 章例 3.3 中，我们建立的中国民航客运量回归方程为

$$\hat{y}=450.9+0.354x_1-0.561x_2-0.007\,3x_3+21.578x_4+0.435x_5$$

其中，y 为民航客运量（万人）；x_1 为国民收入（亿元）；x_2 为消费额（亿元）；x_3 为铁路客运量（万人）；x_4 为民航航线里程（万公里）；x_5 为来华旅游入境人数（万人）。五个自变量都通过了 t 检验，但是 x_2 的回归系数是负值，x_2 是消费额，从经济学的定性分析看，消费额与民航客运量应该是正相关，负的回归系数无法解释。问题出在哪里？这正是自变量之间的多重共线性造成的。

由上述实际例子，我们看到当自变量存在多重共线性时，利用普通最小二乘估计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，会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，甚至导致回归系数的正负号得不到合理的经济解释。但从中国民航客运量一例的回归方程来看，尽管有的回归系数得不到合理的经济解释，但它们对历史数据拟合得很好，其复决定系数 $R^2=0.998$。

以上的分析表明，如果利用模型去作经济结构分析，要尽可能避免多重共线性；如果利用模型去作经济预测，只要保证自变量的相关类型在未来时期中保持不变，即未来时期自变量间仍具有当初建模时数据的联系特征，即使回归模型中包含有严重多重共线性的变量，也可以得到较好的预测结果；如果不能保证自变量的相关类型在未来时期中保持不变，那么多重共线性就会对回归预测产生严重的影响。

专业课辅导+真题资料+微信公众号：432应统书院

6.3 多重共线性的诊断

从前面的例子我们已能大致体会到诊断变量间多重共线性的思想。一般情况下，当回归方程的解释变量之间存在很强的线性关系，回归方程的检验高度显著时，有些与因变量 y 的简单相关系数绝对值很高的自变量，其回归系数不能通过显著性检验，甚至出现有的回归系数所带符号与实际经济意义不符，这时我们就认为变量间存在多重共线性。近年来，关于多重共线性的诊断及多重共线性严重程度的度量是统计学家讨论的热点，他们已经提出了许多可行的判断方法，下面我们只介绍两种主要方法。

一、方差扩大因子法

对自变量作中心标准化，则 $\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}=(r_{ij})$ 为自变量的相关阵。记

$$\boldsymbol{C}=(c_{ij})=(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*})^{-1} \tag{6.5}$$

称其主对角线元素 $\mathrm{VIF}_j=c_{jj}$ 为自变量 x_j 的方差扩大因子（variance inflation factor，VIF）。根据（3.31）式可知

$$\mathrm{var}(\hat{\beta}_j)=c_{jj}\sigma^2/L_{jj},\quad j=1,2,\cdots,\mathrm{p} \tag{6.6}$$

其中，L_{jj} 为 x_j 的离差平方和，由（6.6）式可知，用 c_{jj} 作为衡量自变量 x_j 的方差扩大程度的因子是恰如其分的。记 R_j^2 为自变量 x_j 对其余 $p-1$ 个自变量的复决定系数，可以证明

$$c_{jj}=\frac{1}{1-R_j^2} \tag{6.7}$$

（6.7）式也可以作为方差扩大因子 VIF_j 的定义，由此式可知 $\mathrm{VIF}_j\geqslant 1$。（6.7）式的证明参见参考文献［7］。

R_j^2 度量了自变量 x_j 与其余 $p-1$ 个自变量的线性相关程度，这种相关程度越强，说明自变量之间的多重共线性越严重，R_j^2 就越接近 1，VIF_j 就越大。反之，x_j 与其余 $p-1$ 个自变量的线性相关程度越弱，自变量间的多重共线性就越弱，R_j^2 就越接近零，VIF_j 就越接近 1。由此可见，VIF_j 的大小反映了自变量之间是否存在多重共线性，因此可由它来度量多重共线性的严重程度。经验表明，当 $\mathrm{VIF}_j\geqslant 10$ 时，就说明自变量 x_j 与其余自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值。

也可以用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性。当

$$\overline{\mathrm{VIF}}=\frac{1}{p}\sum_{j=1}^{p}\mathrm{VIF}_j \tag{6.8}$$

专业课辅导+真题资料+微信公众号：432应统书院

远远大于 1 时，就表示存在严重的多重共线性问题。

对于只含两个解释变量 x_1 和 x_2 的回归方程，判断它们是否存在多重共线性，实际上就是计算 x_1 和 x_2 的样本决定系数 R_{12}^2，如果 R_{12}^2很大，则认为 x_1 与 x_2 可能存在严重的多重共线性。为什么我们只说可能存在严重的多重共线性，而没有下定论呢？这是因为 R^2 和样本量 n 有关，当样本量较小时，R^2 容易接近 1，就像我们曾说的，$n=2$ 时，两点总能连成一条直线，$R^2=1$。所以我们认为当样本量还不算小，而 R^2 接近 1 时，可以肯定存在严重的多重共线性。

当某自变量 x_j 对其余 $p-1$ 个自变量的复决定系数 R_j^2 超过一定界限时，SPSS 软件将拒绝这个自变量 x_j 进入回归模型。称 $\mathrm{Tol}_j=1-R_j^2$ 为自变量 x_j 的容忍度(tolerance)，SPSS 软件的默认容忍度为0.000 1。也就是说，当 $R_j^2>0.999\ 9$ 时，自变量 x_j 将被自动拒绝在回归方程之外，除非我们修改容忍度的默认值。

以下用 SPSS 软件诊断例 3.3 中国民航客运量一例中的多重共线性问题。在线性回归对话框的 Statistics 选项框中点击 Collinearity Diagnostic 共线性诊断选项，然后作回归，得到输出结果 6.1。

输出结果 6.1

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	459.909	178.078		2.532	0.030		
X1	0.354	0.085	2.447	4.152	0.002	0.001	1963
X2	−0.561	0.125	−2.485	−4.478	0.001	0.001	1741
X3	−7.E−03	0.002	−0.083	−3.510	0.006	0.315	3.171
X4	21.578	4.030	0.531	5.354	0.000	0.018	55.5
X5	0.435	0.052	0.564	8.440	0.000	0.040	25.2

a. Dependent Variable: Y.

从输出结果 6.1 看到，x_1，x_2 的方差扩大因子很大，分别为 $\mathrm{VIF}_1=1\ 963$，$\mathrm{VIF}_2=1\ 741$，远远超过 10，说明民航客运量回归方程也存在着严重的多重共线性。x_1 是国民收入，x_2 是消费额，两者的简单相关系数 $r_{12}=0.998\ 9$，高度相关。另外，这两个自变量与其余自变量之间也可能存在严重的共线性。

一般情况下，当一个回归方程存在严重的多重共线性时，有若干个自变量所对应的方差扩大因子大于 10，这个回归方程多重共线性的存在就是由方差扩大因子超过 10 的这几个变量引起的，说明这几个自变量间有一定的多重共线性的关系存在。知道了这一点，对于我们消除回归方程的多重共线性非常有用。

二、特征根判定法

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 $|\boldsymbol{X}'\boldsymbol{X}|\approx 0$ 时，矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零。反之可以证明，当矩

专业课辅导+真题资料+微信公众号：432应统书院

阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零时，$\boldsymbol{X}$ 的列向量间必存在多重共线性，证明如下：

记 $\boldsymbol{X}=(\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_p)$，其中 $\boldsymbol{X}_i(i=0, 1, \cdots, p)$ 为 $\boldsymbol{X}$ 的列向量，$\boldsymbol{X}_0=(1, 1, \cdots, 1)'$ 是元素全为 1 的 n 维列向量。λ 是矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的一个近似为零的特征根，$\lambda\approx 0$，$\boldsymbol{c}=(c_0, c_1, \cdots, c_p)'$ 是对应于特征根 λ 的单位特征向量，则

$$\boldsymbol{X}'\boldsymbol{X}\,\boldsymbol{c}=\lambda\,\boldsymbol{c}\approx 0$$

上式两边左乘 $\boldsymbol{c}'$，得

$$\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\,\boldsymbol{c}\approx 0$$

从而有

$$\boldsymbol{X}\boldsymbol{c}\approx 0$$

即

$$c_0\boldsymbol{X}_0+c_1\boldsymbol{X}_1+\cdots+c_p\boldsymbol{X}_p\approx 0$$

写成分量形式即

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}\approx 0,\quad i=1,2,\cdots,n \tag{6.9}$$

这正是（6.2）式定义的多重共线性关系。

如果矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有多个特征根近似为零，在上面的证明中，取每个特征根的特征向量为标准化正交向量，即可证明：$\boldsymbol{X}'\boldsymbol{X}$ 有多少个特征根接近零，设计矩阵 $\boldsymbol{X}$ 就有多少个多重共线性关系，并且这些多重共线性关系的系数向量就等于接近零的那些特征根对应的特征向量。

2. 条件数

特征根分析表明，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有一个特征根近似为零时，设计矩阵 $\boldsymbol{X}$ 的列向量间必存在多重共线性，并且 $\boldsymbol{X}'\boldsymbol{X}$ 有多少个特征根接近零，$\boldsymbol{X}$ 就有多少个多重共线性关系。那么特征根近似为零的标准如何确定呢？可以用下面介绍的条件数确定。记 $\boldsymbol{X}'\boldsymbol{X}$ 的最大特征根为 λ_m，我们称

$$k_i=\sqrt{\frac{\lambda_m}{\lambda_i}},\quad i=0,1,2,\cdots,p \tag{6.10}$$

为特征根 λ_i 的条件数（condition index）。在其他一些书中，条件数定义为 $k_i=\lambda_m/\lambda_i$，没有开平方根，SPSS 软件是采用的（6.10）式开平方根的，这一点请读者注意。

条件数度量了矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的特征根的散布程度，可以用它来判断多重共线性是否存在以及多重共线性的严重程度。通常认为 $0<k<10$ 时，设计矩阵 $\boldsymbol{X}$ 没有多重共线性；$10\leqslant k<100$ 时，存在较强的多重共线性；$k\geqslant 100$ 时，存在严重的多重共线性。

专业课辅导+真题资料+微信公众号：432应统书院

对例 3.3 中国民航客运量的例子，用 *SPSS* 软件计算出特征根与条件数如输出结果 6.2 所示。

输出结果 6.2

Collinearity Diagnostics

Dimension	Eigenvalue	Condition Index	Variance Proportions					
			(Constant)	X1	X2	X3	X4	X5
1	5.578	1.000	.00	.00	.00	.00	.00	0.00
2	.378	3.842	.00	.00	.00	.00	.00	0.00
3	3.745E-02	12.205	.01	.00	.00	.00	.03	0.19
4	4.203E-03	36.431	.17	.00	.01	.09	.50	0.04
5	1.939E-03	53.643	.72	.00	.01	.66	.15	0.71
6	8.080E-05	262.762	.10	.99	.99	.25	.31	0.06

a. Dependent Variable: Y.

从条件数看到，最大的条件数 $k_6=262.762$，说明自变量间存在严重的多重共线性，这与方差扩大因子法的结果一致。输出结果 6.2 中的特征根是按从大到小的顺序排列的，不是按自变量的顺序排列的，这与方差扩大因子法不同。那么如何判定究竟哪几个自变量间存在共线性呢？这可以由输出结果 6.2 右边的方差比例（Variance Proportions）粗略判定。如果某几个自变量的方差比例值在某一行同时较大（接近 1），则这几个自变量间就存在多重共线性。表中第 6 行 x_1，x_2 的系数都为 0.99，说明 x_1，x_2 之间存在较强的多重共线性；表中第 5 行 x_0（常数项），x_3，x_5 的系数分别为 0.72，0.66，0.71，说明 x_0（常数项），x_3，x_5 之间存在多重共线性。

细心的读者可能已经发现，本例自变量的数目只有 5 个，而特征根与条件数却有 6 个。原因很简单，这是由于设计矩阵 $\boldsymbol{X}$ 的第一列有一列 1，代表常数项，$\boldsymbol{X}$ 共有 $p+1$ 列，$\boldsymbol{X}'\boldsymbol{X}$ 是 $p+1$ 阶方阵。当一个自变量的取值范围很小，接近常数时，这个自变量就与常数项存在多重共线性。在（6.2）式多重共线性的定义式中，如果 $c_2=c_3=\cdots=c_p=0$，而 $c_0\neq0$，$c_1\neq0$，这时自变量 x_1 就与常数项存在多重共线性。

最后再说说输出结果 6.2 的方差比例。方差比例是用于判断哪几个自变量之间存在共线性的。实际上共线性关系可以根据（6.9）式直接从特征向量中看出来，只是 SPSS 软件在线性回归模块中没有输出特征向量阵。把特征向量按照特征值由大到小排成行向量，每个数值平方后再除以特征值，再把每列数据除以列数据之和，使得每列数据之和为 1，这样就得到了输出结果 6.2 的方差比例。

三、直观判定法

上述方法是为了诊断共线性是否存在的专门方法，此外，还有一些在建模过程

专业课辅导+真题资料+微信公众号：432应统书院

中顺便直观判断的非正规方法。

（1）当增加或剔除一个自变量，或者改变一个观测值时，回归系数的估计值发生较大变化，就认为回归方程存在严重的多重共线性。

（2）从定性分析角度来看，当一些重要的自变量在回归方程中没有通过显著性检验时，可初步判断存在严重的多重共线性。

（3）当有些自变量的回归系数所带正负号与定性分析结果违背时，认为存在多重共线性。

（4）自变量的相关矩阵中，当自变量间的相关系数较大时，认为可能存在多重共线性。

（5）当一些重要的自变量的回归系数的标准误差较大时，认为可能存在多重共线性。

6.4 消除多重共线性的方法

当通过某种检验发现解释变量中存在严重的多重共线性时，就要设法消除这种共线性。消除多重共线性的方法很多，常用的有下面几种。

一、剔除一些不重要的解释变量

通常在经济问题的建模中，由于我们认识水平的局限，容易考虑过多的自变量。当涉及自变量较多时，大多数回归方程都受到多重共线性的影响。这时，最常用的办法是首先用第 5 章的方法作自变量的选元，剔除一些自变量。当回归方程中的全部自变量都通过显著性检验后，回归方程中仍然存在严重的多重共线性，有几个变量的方差扩大因子大于 10，我们可把方差扩大因子最大者所对应的自变量首先剔除，再重新建立回归方程，如果仍然存在严重的多重共线性，再继续剔除方差扩大因子最大者所对应的自变量，直到回归方程中不再存在严重的多重共线性为止。

有时，根据所研究的问题的需要，当回归方程中仍然存在严重的多重共线性时，也可以首先剔除方差扩大因子最大者所对应的自变量，依次剔除，直到消除了多重共线性为止，然后再作自变量的选元。或者根据所研究问题的经济意义，决定保留或剔除某自变量。

总之，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自变量的经济含义结合起来考虑，以引进或剔除变量。

在民航客运量一例中，五个自变量都通过了回归系数的显著性检验，但仍然存在着严重的多重共线性。x_1 的方差扩大因子 $\mathrm{VIF}_1=1\,963$ 为最大，因此剔除 x_1，建立 y 对四个自变量 x_2，x_3，x_4，x_5 的回归方程。有关计算结果如输出结果 6.3 所示。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 6.3

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	695.039	264.525		2.627	.024		
X2	-5.257E-02	.042	-.233	-1.262	.233	.013	77.546
X3	-1.170E-02	.003	-.134	-4.207	.001	.431	2.319
X4	32.037	4.951	.788	6.471	.000	.030	33.812
X5	.399	.080	.517	4.988	.000	.041	24.469

从输出结果 6.3 中看到，x_2 的方差扩大因子 $\mathrm{VIF}_2=77.5$ 为最大，远大于 10，并且 x_2 的回归系数 $\hat{\beta}_2=-0.053$ 仍然为负值，说明此回归模型仍然存在强多重共线性，应该继续剔除变元。

剔除 x_2，用 y 与剩下的三个自变量 x_3，x_4，x_5 建立回归方程，有关计算结果如输出结果 6.4 所示。

输出结果 6.4

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	591.876	257.730		2.296	.040		
X3	-1.037E-02	.003	-.119	-3.934	.002	.504	1.984
X4	26.436	2.249	.650	11.754	.000	.150	6.650
X5	.317	.048	.411	6.568	.000	.117	8.514

从输出结果 6.4 中看到，3 个方差扩大因子都小于 10，回归系数也都有合理的经济解释，说明此回归模型不存在强多重共线性，可以作为最终回归模型。回归方程为：

$$\hat{y}=591.876-0.010\,37x_3+26.436x_4+0.317x_5 \tag{6.11}$$

标准化回归方程为

$$\hat{y}^*=-0.119x_3^*+0.650x_4^*+0.411x_5^*$$

由标准化回归系数看到，对民航客运量影响最大的因素是民航航线里程 x_4，其次是来华旅游入境人数 x_5。民航航线里程每增加 1%，民航客运量会平均增加 0.650%。来华旅游入境人数每增加 1%，民航客运量会平均增加 0.411%。而铁路客运量 x_3 对民航客运量的影响相对较小，铁路客运量每增加 1%，民航客运量会平均减少 0.119%。

此回归方程的样本决定系数 $R^2=0.994$，调整的样本决定系数 $R_a^2=0.993$。而 y 对 5 个自变量的全模型的样本决定系数 $R^2=0.998$，调整的样本决定系数 $R_a^2=0.997$。与全模型相比，(6.11) 式的拟合优度仍然很高，并且回归系数有合理的经

专业课辅导+真题资料+微信公众号：432应统书院

济解释。

二、增大样本量

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，也容易产生多重共线性。譬如，我们的问题涉及两个自变量 x_1 和 x_2，假设 x_1 和 x_2 都已经中心化。由（6.3）式和（6.4）式

$$\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{(1-r_{12}^2)L_{11}}$$

$$\mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{(1-r_{12}^2)L_{22}}$$

其中，r_{12} 为 x_1 和 x_2 的相关系数，$L_{11}=\sum_{i=1}^{n}x_{i1}^2$，$L_{22}=\sum_{i=1}^{n}x_{i2}^2$。可以看到，在 r_{12} 固定不变时，当样本量 n 增大时，L_{11} 和 L_{22} 都会增大，两个回归系数估计值的方差均可减小，从而减弱多重共线性对回归方程的影响。因此，增大样本量也是消除多重共线性的一个途径。

在实践中，当我们所选的变量个数接近样本量 n 时，自变量间就容易产生共线性。所以在运用回归分析研究经济问题时，要尽可能使样本量 n 远大于自变量个数 p。

增大样本量的方法在有些经济问题中是不现实的，因为在经济问题中，许多自变量是不受控制的，或由于种种原因不可能再得到一些新的样本数据。在有些情况下，虽然可以增加一些样本数据，但当自变量个数较多时，我们往往难以确定增加什么样的数据才能克服多重共线性。

有时，增加了样本数据，但可能新数据距离原来样本数据的平均值较大，会产生一些新的问题，使模型拟合变差，没有收到增加样本数据期望的效果。

三、回归系数的有偏估计

消除多重共线性对回归模型的影响是近几十年来统计学家关注的热点课题之一，除以上方法被人们应用外，统计学家还致力于改进古典的最小二乘法，提出以采用有偏估计为代价来提高估计量稳定性的方法，如岭回归法、主成分法、偏最小二乘法等，这些方法已有不少应用效果很好的经济例子，而且在计算机如此发达的今天，具体计算也不难实现。我们将在本书的第 7 章中详细介绍岭回归法，在第 8 章中介绍主成分回归和偏最小二乘。

6.5 本章小结与评注

因为大多数经济变量在时间上有共同的变化趋势，所以在建立经济问题的回归

专业课辅导+真题资料+微信公众号：432应统书院

模型时经常会遇到多重共线性的诊断和处理。

本章从共线性产生的经济背景谈起，介绍了多重共线性对回归系数估计值和回归方程预测值的影响，给出了几种诊断共线性的方法，并就如何消除共线性对回归方程的影响介绍了几种方法。

关于多重共线性对回归参数的影响，我们认为这不仅取决于自变量中多重共线性的强弱程度，还取决于存在多重共线性的自变量在整个回归方程中的重要性。如果对因变量有着重要影响的自变量中出现严重的多重共线性，那么给模型参数估计带来的危害要比次要因素中存在严重的多重共线性大得多。我们应尽量避免主要自变量中存在多重共线性。如果各自变量的取值可人为控制，可使设计矩阵 $\boldsymbol{X}$ 达到回归模型基本假设的要求。如果无法克服自变量间的多重共线性，回归方程中尽量少引进一些解释变量是一种有效的方法，但这样得到的回归方程可能不利于结构分析。

我们在前面看到，有时一个回归模型存在严重的多重共线性，回归系数可能通不过显著性检验，回归系数的正负号不符合经济意义，但用这个方程去作预测，拟合效果还相当好，甚至比不存在共线性时还好。如果建模的目的就是预测，只要保证自变量的相关类型在预测期不变，即当初建模时自变量间共同的相关趋势在预测时仍基本保持，用具有较强的多重共线性的方程去作预测效果仍会不错。但这里我们要强调，如果自变量的相关类型在预测期发生了变化，若用具有很强共线性的模型去作预测，效果肯定不好。

在建立经济问题的回归模型时，如果发现解释变量之间的简单相关系数很大，可以断定自变量间存在着严重的多重共线性，但是，当一个回归方程存在严重的多重共线性时，并不能完全肯定解释变量之间的简单相关系数就一定很大。例如对含有三个自变量的回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon \tag{6.12}$$

假定三个自变量之间有完全确定的关系

$$x_1=x_2+x_3$$

因为 x_1 可由 x_2 和 x_3 线性表示，所以变量 x_1 与 x_2 和 x_3 的复决定系数 $R^2_{1\cdot 23}=1$，回归方程存在完全的多重共线性。再假定 x_2 与 x_3 的简单相关系数 $r_{23}=-0.5$，x_2 与 x_3 的离差平方和 $L_{22}=L_{33}=1$，此时

$$L_{23}=r_{23}\sqrt{L_{22}L_{33}}=-0.5$$

$$\begin{aligned}L_{11}&=\sum(x_1-\bar{x}_1)^2\\&=\sum(x_2+x_3-(\bar{x}_2+\bar{x}_3))^2\\&=\sum((x_2-\bar{x}_2)+(x_3-\bar{x}_3))^2\\&=\sum(x_2-\bar{x}_2)^2+\sum(x_3-\bar{x}_3)^2+2\sum(x_2-\bar{x}_2)(x_3-\bar{x}_3)\end{aligned}$$

专业课辅导+真题资料+微信公众号：432应统书院

$$= 1+1+2\times(-0.5) = 1$$

$$\begin{aligned} L_{12} &= \sum(x_1-\bar{x}_1)(x_2-\bar{x}_2) \\ &= \sum(x_2+x_3-(\bar{x}_2+\bar{x}_3))(x_2-\bar{x}_2) \\ &= \sum((x_2-\bar{x}_2)+(x_3-\bar{x}_3))(x_2-\bar{x}_2) \\ &= L_{22}+L_{23} \\ &= 1-0.5 = 0.5 \end{aligned}$$

因而 $r_{12} = L_{12}/\sqrt{L_{11}L_{22}} = 0.5$

同理 $r_{13} = 0.5$

在这里我们看到三个自变量的简单相关系数的绝对值都是 0.5，都不高，但是三者之间却存在完全的多重共线性。

由此看到，当回归方程中的自变量数目超过 2 个时，并不能由自变量间的简单相关系数不高，就断定它们不存在多重共线性。如果回归方程中只有 2 个自变量，则由它们的简单相关系数可判断是否存在多重共线性。

关于多重共线性的诊断我们在 6.3 节中介绍了一些正规方法和非正规方法。一般说来非正规方法比较直观，往往在建模的过程中就会发现。介绍的几个正规方法都要通过一定的运算，但通过它们可以发现多重共线性的严重程度。要想知道存在多重共线性的程度，就需用条件数和方差扩大因子来度量，现在已有不少统计软件都可直接计算出来。

关于消除共线性的方法，除了 6.4 节中介绍的外，还有逐步回归法、岭回归法、主成分法、特征根法、偏最小二乘法等。至今如何消除多重共线性仍是研究的热点，有许多这方面的问题需要研究，而且还没有哪一个方法占绝对优势，从运用的效果还很难说明哪个方法最优。各人可以根据自己的知识水平和计算机软件的运用水平来选择合适的方法。

□ 思考与练习

6.1 试举一个产生多重共线性的经济实例。

6.2 多重共线性对回归参数的估计有何影响？

6.3 具有严重多重共线性的回归方程能否用来作经济预测？

6.4 多重共线性的产生与样本量的个数 n、自变量的个数 p 有无关系？

6.5 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设计矩阵 $\boldsymbol{X}$ 才可能避免多重共线性的出现？

6.6 对第 5 章思考与练习第 9 题财政收入的数据，分析数据的多重共线性，并根据多重共线性剔除变量，将所得结果与用逐步回归法所得的选元结果相比较。

专业课辅导+真题资料+微信公众号：432应统书院

第 7 章 Chapter 7 岭回归

在第 6 章中我们已经看到，当设计矩阵 $\boldsymbol{X}$ 呈病态时，$\boldsymbol{X}$ 的列向量之间有较强的线性相关性，即解释变量间出现严重的多重共线性。这种情况下，用普通最小二乘法估计模型参数，往往参数估计方差太大，使普通最小二乘法的效果变得很不理想。为了解决这一问题，统计学家从模型和数据的角度考虑，采用回归诊断和自变量选择来克服多重共线性的影响。近 40 年来，人们还对普通最小二乘估计提出了一些改进方法。目前，岭回归就是最有影响的一种新的估计方法。本章将系统介绍岭回归估计的定义及性质，并结合实际例子给出岭回归的应用。

7.1 岭回归估计的定义

一、普通最小二乘估计带来的问题

多元线性回归模型的矩阵形式为 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，参数 $\boldsymbol{\beta}$ 的普通最小二乘估计为 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$。在第 6 章多重共线性部分讲到，当自变量 x_j 与其余自变量间存在多重共线性时，$\mathrm{var}(\hat{\beta}_j)=c_{jj}\sigma^2/L_{jj}$ 很大，$\hat{\beta}_j$ 就很不稳定，在具体取值上与真值有较大的偏差，甚至有时会出现与实际经济意义不符的正负号，在第 3 章的例 3.3 民航客运的例子中我们已经看到这种现象。下面进一步用参考文献［5］的一个例子来说明这一点。

专业课辅导+真题资料+微信公众号：432应统书院

例 7.1

我们作回归拟合时，总是希望拟合的经验回归方程与真实的理论回归方程能够很接近。基于这个想法，这里举一个模拟的例子。假设 x_1，x_2 与 y 的关系服从线性回归模型

$$y = 10 + 2x_1 + 3x_2 + \varepsilon \tag{7.1}$$

给定 x_1，x_2 的 10 个值，见表 7.1 的第（1）、（2）两行。

表 7.1

序号		1	2	3	4	5	6	7	8	9	10
(1)	x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
(2)	x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
(3)	ε_i	0.8	−0.5	0.4	−0.5	0.2	1.9	1.9	0.6	−1.5	−1.5
(4)	y_i	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

然后用模拟的方法产生 10 个正态随机数，作为误差项 ε_1，ε_2，…，ε_{10}，见表 7.1 的第（3）行。再由回归模型 $y_i = 10 + 2x_{i1} + 3x_{i2} + \varepsilon_i$ 计算出 10 个 y_i 值，列在表 7.1 的第（4）行。

现在假设回归系数与误差项是未知的，用普通最小二乘法求回归系数的估计值得

$$\hat{\beta}_0 = 11.292, \hat{\beta}_1 = 11.307, \hat{\beta}_2 = -6.591$$

而原模型的参数

$$\beta_0 = 10, \beta_1 = 2, \beta_2 = 3$$

看来相差太大。计算 x_1，x_2 的样本相关系数得 $r_{12} = 0.986$，表明 x_1 与 x_2 之间高度相关。这里我们看到解释变量之间高度相关时，普通最小二乘估计明显变坏的又一例证。

二、岭回归的定义

必须要认真学习每一个知识点方可做出判断

针对出现多重共线性时，普通最小二乘法明显变坏的问题，霍尔（A. E. Hoerl）在 1962 年首先提出一种改进最小二乘估计的方法，叫岭估计（ridge estimate），后来 Hoerl 和 Kennard 于 1970 年（见参考文献［18］）给予了详细讨论。

岭回归（ridge regression，RR）提出的想法是很自然的。当自变量间存在多重共线性，$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ 时，我们设想给 $\boldsymbol{X}'\boldsymbol{X}$ 加上一个正常数矩阵 $k\boldsymbol{I}(k>0)$，那么 $\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}$ 接近奇异的程度就会比 $\boldsymbol{X}'\boldsymbol{X}$ 接近奇异的程度小得多。考虑到变量的量纲问题，先对数据作标准化，为了计算方便，标准化后的设计阵仍然用 $\boldsymbol{X}$ 表示，定义为

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{7.2}$$

我们称（7.2）式为 $\boldsymbol{\beta}$ 的岭回归估计，其中，k 称为岭参数。由于假设 $\boldsymbol{X}$ 已经标准化，所以 $\boldsymbol{X}'\boldsymbol{X}$ 就是自变量样本相关阵。（7.2）式中 $\boldsymbol{y}$ 可以标准化也可以不标准化，如果 $\boldsymbol{y}$ 也经过标准化，那么（7.2）式计算的实际是标准化岭回归估计。$\hat{\boldsymbol{\beta}}(k)$ 作为 $\boldsymbol{\beta}$ 的估计应比最小二乘估计 $\hat{\boldsymbol{\beta}}$ 稳定，当 $k=0$ 时的岭回归估计 $\hat{\boldsymbol{\beta}}(0)$ 就是普通最小二乘估计。

因为岭参数 k 不是唯一确定的，所以得到的岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 实际是回归参数 $\boldsymbol{\beta}$ 的一个估计族。

例如对例 7.1 可以算得不同 k 值时的 $\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$，见表 7.2。

表 7.2

k	0	0.1	0.15	0.2	0.3	0.4	0.5	1.0	1.5	2	3
$\hat{\beta}_1(k)$	11.31	3.48	2.99	2.71	2.39	2.20	2.06	1.66	1.43	1.27	1.03
$\hat{\beta}_2(k)$	−6.59	0.63	1.02	1.21	1.39	1.46	1.49	1.41	1.28	1.17	0.98

以 k 为横坐标，$\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$ 为纵坐标画成图 7.1。从图上可看到，当 k 较小时，$\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$ 很不稳定；当 k 逐渐增大时，$\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$ 趋于零。k 取何值时，对应的 $\hat{\beta}_1(k)$，$\hat{\beta}_2(k)$ 才是一个优于普通最小二乘估计的估计呢？这是后面将要讨论的重点问题。

图 7.1

7.2 岭回归估计的性质

无需作太多关注

在本节关于岭回归估计的性质的讨论中，假定（7.2）式中因变量观测向量 $\boldsymbol{y}$ 未经标准化。

性质 1 $\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计。

证明：$E[\hat{\boldsymbol{\beta}}(k)]=E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y})$

专业课辅导+真题资料+微信公众号：432应统书院

$$= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y})$$
$$= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

显然只有当 $k=0$ 时，$E[\hat{\boldsymbol{\beta}}(0)]=\boldsymbol{\beta}$；当 $k\neq 0$ 时，$\hat{\boldsymbol{\beta}}(k)$ 是 $\boldsymbol{\beta}$ 的有偏估计。要特别强调的是 $\hat{\boldsymbol{\beta}}(k)$ 不再是 $\boldsymbol{\beta}$ 的无偏估计，有偏性是岭回归估计的一个重要特性。

性质 2 在认为岭参数 k 是与 $\boldsymbol{y}$ 无关的常数时，$\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，也是 $\boldsymbol{y}$ 的线性函数。

因为 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$

$$=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\,\hat{\boldsymbol{\beta}}$$

所以，岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，根据定义式 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 知 $\hat{\boldsymbol{\beta}}(k)$ 也是 $\boldsymbol{y}$ 的线性函数。

这里需要注意的是，在实际应用中，由于岭参数 k 总是要通过数据来确定，因而 k 也依赖于 $\boldsymbol{y}$，因此从本质上说，$\hat{\boldsymbol{\beta}}(k)$ 并非 $\hat{\boldsymbol{\beta}}$ 的线性变换，也不是 $\boldsymbol{y}$ 的线性函数。

性质 3 对任意 $k>0$，$\|\hat{\boldsymbol{\beta}}\|\neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\|<\|\hat{\boldsymbol{\beta}}\|$$

这里 $\|\cdot\|$ 是向量的模，等于向量各分量的平方和。这个性质表明 $\hat{\boldsymbol{\beta}}(k)$ 可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种向原点的压缩。从 $\hat{\boldsymbol{\beta}}(k)$ 的表达式可以看到，当 $k\to\infty$ 时，$\hat{\boldsymbol{\beta}}(k)\to\boldsymbol{0}$，即 $\hat{\boldsymbol{\beta}}(k)$ 化为零向量。

性质 4 以 MSE 表示估计向量的均方误差，则存在 $k>0$，使得

$$\mathrm{MSE}[\hat{\boldsymbol{\beta}}(k)]<\mathrm{MSE}(\hat{\boldsymbol{\beta}})$$

即

$$\sum_{j=1}^{p}E[\hat{\boldsymbol{\beta}}_j(k)-\beta_j]^2<\sum_{j=1}^{p}D(\hat{\boldsymbol{\beta}}_j)$$

7.3 岭迹分析

当岭参数 k 在 $(0,\infty)$ 内变化时，$\hat{\beta}_j(k)$ 是 k 的函数，在平面坐标系上把函数 $\hat{\beta}_j(k)$ 描画出来，画出的曲线称为岭迹。在实际应用中，可以根据岭迹曲线的变化形状来确定适当的 k 值和进行自变量的选择。下面根据参考文献［2］来介绍岭迹分析。

在岭回归中，岭迹分析可用来了解各自变量的作用及自变量间的相互关系。下面根据图 7.2 所反映的几种有代表性的情况来说明岭迹分析的作用。

（1）在图 7.2（a）中，$\hat{\beta}_j(0)=\hat{\beta}_j>0$，且比较大。从古典回归分析的观点看，应将 x_j 看作对 y 有重要影响的因素。但 $\hat{\beta}_j(k)$ 的图形显示出相当的不稳定，

专业课辅导+真题资料+微信公众号：432应统书院

当 k 从零开始略增加时，$\hat{\beta}_j(k)$ 显著地下降，而且迅速趋于零，因而失去预测能力。从岭回归的观点看，x_j 对 y 不起重要作用，甚至可以剔除这个变量。

（2）与图 7.2（a）相反的情况见图 7.2（b），$\hat{\beta}_j=\hat{\beta}_j(0)>0$，但很接近 0。从古典回归分析的观点看，$x_j$ 对 y 的作用不大。但随着 k 略增加，$\hat{\beta}_j(k)$ 骤然变为负值，从岭回归的观点看，x_j 对 y 有显著影响。

（3）在图 7.2（c）中，$\hat{\beta}_j=\hat{\beta}_j(0)>0$，说明 x_j 比较显著，但当 k 增加时，$\hat{\beta}_j(k)$迅速下降，且稳定为负值。从古典回归分析的观点看，x_j 是对 y 有正影响的显著因素。从岭回归的观点看，x_j 被看做对 y 有负影响的因素。

（4）在图 7.2（d）中，$\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 都很不稳定，但其和却大体上稳定。这种情况往往发生在自变量 x_1 和 x_2 的相关性很大的场合，即在 x_1 和 x_2 之间存在多重共线性。因此，从变量选择的观点看，两者只要保存一个就够了。这可用来解释某些回归系数估计的符号不合理的情形，从实际观点看，β_1 和 β_2 不应有相反的符号。岭回归分析的结果对这一点提供了一种解释。

（5）从全局看，岭迹分析可用来估计在某一具体实例中最小二乘估计是否适用。把所有回归系数的岭迹都描在一张图上，如果这些岭迹线的不稳定性很大，整个系统呈现比较“乱”的局面，往往就使人怀疑最小二乘估计是否很好地反映了真实情况，如图 7.2（e）所示。如果情况如图 7.2（f）那样，则我们对最小二乘估计可以有更大的信心。当情况介于（e）和（f）之间时，我们必须适当地选择 k 值。

图 7.2

专业课辅导+真题资料+微信公众号：432应统书院

7.4 岭参数 k 的选择

我们的目的是要选择使 MSE($\hat{\boldsymbol{\beta}}(k)$) 达到最小的 k，最优 k 值依赖于未知参数 $\boldsymbol{\beta}$ 和 σ^2，因而在实际应用中必须通过样本来确定。究竟如何确定 k 值，在理论上尚未得到满意的答案。问题的关键是最优 k 值对未知参数 $\boldsymbol{\beta}$ 和 σ^2 的依赖关系与函数形式不清楚，但这个问题在应用上又特别重要，因此有不少统计学者进行相应的研究。近几十年来，他们相继提出了许多确定 k 值的原则和方法，这些方法一般都基于直观考虑，有些通过计算机模拟试验，具有一定的应用价值，但目前尚未找到一个公认的最优方法。

下面介绍几种常用的选择方法。

一、岭迹法

岭迹法的直观考虑是，如果最小二乘估计看起来有不合理之处，如估计值以及正负号不符合经济意义，则希望能通过采用适当的岭估计 $\hat{\boldsymbol{\beta}}(k)$来加以一定程度的改善，岭参数 k 值的选择就显得尤为重要。选择 k 值的一般原则是：

（1）各回归系数的岭估计基本稳定。

（2）用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理。

（3）回归系数没有不合乎经济意义的绝对值。

（4）残差平方和增加不太多。

例如在图 7.3 中，当 k 取 k_0 时，各回归系数的估计值基本上都能相对稳定。当然，上述种种要求并不是总能达到的。如在例 7.1 中由图 7.1 看到，取 $k=0.5$，岭迹已算平稳，从而 $\hat{\beta}_1(0.5)=2.06$，$\hat{\beta}_2(0.5)=1.49$。$\hat{\beta}_1(0.5)$ 已相当接近真值 $\beta_1=2$，但 $\hat{\beta}_2(0.5)$ 与 $\beta_2=3$ 还相差很大。

图 7.3

专业课辅导+真题资料+微信公众号：432应统书院

岭迹法与传统的基于残差的方法相比，从概念上来说是完全不同的。因此，它为我们分析问题提供了一种新的思想方法，对于分析各变量之间的作用和关系是有帮助的。

岭迹法确定 k 值缺少严格的令人信服的理论依据，存在着一定的主观性，这似乎是岭迹法的一个明显的缺点。但从另一方面说，岭迹法确定 k 值的这种主观性正好有助于实现定性分析与定量分析有机结合。

二、方差扩大因子法

在 6.3 节中，我们给出方差扩大因子的概念，方差扩大因子 c_{jj} 可以度量多重共线性的严重程度，一般当 $c_{jj}>10$ 时，模型就有严重的多重共线性。计算岭估计 $\hat{\boldsymbol{\beta}}(k)$ 的协方差阵，得

$$
\begin{aligned}
D(\hat{\boldsymbol{\beta}}(k)) &= \operatorname{cov}(\hat{\boldsymbol{\beta}}(k),\hat{\boldsymbol{\beta}}(k)) \\
&= \operatorname{cov}((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\operatorname{cov}(\boldsymbol{y},\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1} \\
&= \sigma^2\boldsymbol{c}(k)
\end{aligned}
$$

其中，矩阵 $\boldsymbol{c}(k)=(\boldsymbol{X}'\boldsymbol{X}+kI)^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}$，其对角元素 $c_{jj}(k)$ 为岭估计的方差扩大因子。不难看出，$c_{jj}(k)$ 随着 k 的增大而减少。应用方差扩大因子选择 k 的经验做法是：选择 k 使所有方差扩大因子 $c_{jj}(k)\leqslant 10$。当 $c_{jj}(k)\leqslant 10$ 时，所对应的 k 值的岭估计 $\hat{\boldsymbol{\beta}}(k)$ 就会相对稳定。

三、由残差平方和确定 k 值

我们知道岭估计 $\hat{\boldsymbol{\beta}}(k)$ 在减小均方误差的同时增大了残差平方和，我们希望岭回归的残差平方和 SSE (k)的增加幅度控制在一定的限度以内，从而可以给定一个大于 1 的 c 值，要求

$$\mathrm{SSE}(k)<c\mathrm{SSE} \tag{7.3}$$

寻找使（7.3）式成立的最大的 k 值。

7.5　用岭回归选择变量　剔除变量

岭回归的一个重要应用是选择变量，选择变量通常的原则是：

（1）在岭回归的计算中，假定设计矩阵 $\boldsymbol{X}$ 已经中心化和标准化，这样可以直接比较标准化岭回归系数的大小。我们可以剔除掉标准化岭回归系数比较稳定且绝对

专业课辅导+真题资料+微信公众号：432应统书院

值很小的自变量。

（2）当 k 值较小时，标准化岭回归系数的绝对值并不很小，但是不稳定，随着 k 的增加迅速趋于零。像这样岭回归系数不稳定、振动趋于零的自变量，我们也可以予以剔除。

（3）剔除标准化岭回归系数很不稳定的自变量。如果有若干个岭回归系数不稳定，究竟剔除几个，剔除哪几个，并无一般原则可循，需根据剔除某个变量后重新进行岭回归分析的效果来确定。

下面通过参考文献［2］引用参考文献［19］的例子来说明如何用岭回归选择变量。

例 7.2

空气污染问题。在参考文献［19］中 Mcdonald 和 Schwing 曾研究死亡率与空气污染、气候以及社会经济状况等因素的关系。考虑了 15 个解释变量：

x_1——年平均降雨量；

x_2——1 月份平均气温；

x_3——3 月份平均气温；

x_4——年龄在 65 岁以上的人口占总人口的百分数；

x_5——每家的人口数；

x_6——中学毕业年龄；

x_7——住房符合标准的家庭比例数；

x_8——每平方公里居民数；

x_9——非白种人占总人口的比例；

x_{10}——白领阶层中受雇百分数；

x_{11}——收入在 300 美元以上的家庭百分数；

x_{12}——碳氢化合物的相对污染势；

x_{13}——氮氧化物的相对污染势；

x_{14}——二氧化硫的相对污染势；

x_{15}——相对湿度；

y——每 10 万人中的死亡人数。

这个问题收集了 60 组样本数据。根据样本数据，计算 $\boldsymbol{X}'\boldsymbol{X}$ 的 15 个特征根为：

4.527 2　2.754 7　2.054 5　1.348 7　1.222 7　0.960 5　0.612 4

0.472 9　0.370 8　0.216 3　0.166 5　0.127 5　0.114 2　0.046 0　0.004 9

后面两个特征根很接近零，由第 6 章中介绍的条件数可知

$$k=\sqrt{\lambda_1/\lambda_{15}}=\sqrt{4.527\,2/0.004\,9}=\sqrt{923.918}=30.396$$

专业课辅导+真题资料+微信公众号：432应统书院

说明设计矩阵 $\boldsymbol{X}$ 具有较严重的多重共线性。

进行岭迹分析，把 15 个回归系数的岭迹绘成图 7.4，从图中看到，当 $k=0.20$ 时，岭迹大体上达到稳定。按照岭迹法，应取 $k=0.2$。若用方差扩大因子法，当 k 为 0.02～0.08 时，方差扩大因子小于 10，故建议在此范围选取 k。由此也看到不同的方法选取的 k 值是不同的。

图 7.4

在用岭回归进行变量选择时，因为从岭迹看到自变量 x_4，x_7，x_{10}，x_{11} 和 x_{15} 有较稳定且绝对值比较小的岭回归系数，根据变量选择的第一条原则，这些自变量可以剔除。又因为自变量 x_{12} 和 x_{13} 的岭回归系数很不稳定，且随着 k 的增加很快趋于零，根据上面的第二条原则这些自变量也应该剔除。还可根据第三条原则剔除变量 x_3 和 x_5。这个问题最后剩的变量是 x_1，x_2，x_6，x_8，x_9，x_{14}，即可用这些自变量建立一个回归方程。

例 7.3

Gorman-Torman 例子（见参考文献 [2]）。本例共有 10 个自变量，$\boldsymbol{X}$ 已经中心化和标准化了，$\boldsymbol{X}'\boldsymbol{X}$ 的特征根为：

3.692　1.542　1.293　1.046　0.972

0.659　0.357　0.220　0.152　0.068

最后一个特征根 $\lambda_{10}=0.068$，较接近零。

$$k=\sqrt{\lambda_1/\lambda_{10}}=\sqrt{3.692/0.068}=\sqrt{54.294}=7.368$$

条件数 $k=7.368<10$。从条件数的角度看，似乎设计矩阵 $\boldsymbol{X}$ 没有多重共线性。但下面的研究表明，作岭回归还是必要的。关于条件数，这里附带说明它的一个缺陷，就是当 $\boldsymbol{X}'\boldsymbol{X}$ 的所有特征根都较小时，虽然条件数不大，但多重共线性却存在。本例就是一个证明。

专业课辅导+真题资料+微信公众号：432应统书院

下面作岭回归分析。对 15 个 k 值算出 $\hat{\boldsymbol{\beta}}(k)$，画出岭迹，如图 7.5（a）所示。由图 7.5（a）可看到，最小二乘估计的稳定性很差。这反映在当 k 与 0 略有偏离时，$\hat{\boldsymbol{\beta}}(k)$ 与 $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}(0)$ 就有较大的差距，特别是 $|\hat{\boldsymbol{\beta}}_5|$ 与 $|\hat{\boldsymbol{\beta}}_6|$ 变化最明显。当 k 从 0 上升到 0.1 时，$\|\hat{\boldsymbol{\beta}}(k)\|^2$ 下降到 $\|\hat{\boldsymbol{\beta}}(0)\|^2$ 的 59%，而在正交设计的情形下只下降 17%。这些现象在直观上就使人怀疑最小二乘估计 $\hat{\boldsymbol{\beta}}$ 是否反映了 $\boldsymbol{\beta}$ 的真实情况。

图 7.5

另外，因素 x_5 的回归系数的最小二乘估计 $\hat{\beta}_5$ 为负回归系数中绝对值最大的，但当 k 增加时，$\hat{\beta}_5(k)$ 迅速上升且变为正的。与此相反，对因素 x_6，$\hat{\beta}_6$ 为正的，且绝对值最大，但当 k 增加时，$\hat{\beta}_6(k)$ 迅速下降。再考虑到 x_5，x_6 的样本相关系数达到 0.84，因此这两个因素可近似地合并为一个因素。

再来看 x_7，它的回归系数估计 $\hat{\beta}_7$ 的绝对值偏高，当 k 增加时，$\hat{\beta}_7(k)$ 很快接近零，这意味着 x_7 实际上对 y 无多大影响。至于 x_1，其回归系数的最小二乘估计

专业课辅导+真题资料+微信公众号：432应统书院

绝对值看来有点偏低，当 k 增加时，$|\hat{\beta}_1(k)|$ 首先迅速上升，成为对因变量有负影响的最重要的自变量。当 k 较大时，$|\hat{\beta}_1(k)|$ 稳定地缓慢趋于零。这意味着，通常的最小二乘估计对 x_1 的重要性估计过低。

从整体上看，当 k 达到 0.2～0.3 的范围时，各个 $\hat{\beta}_j(k)$ 大体上趋于稳定，因此，在这一区间取一个 k 值作岭回归可能得到较好的结果。本例中 $\hat{\beta}_5(k)$和 $\hat{\beta}_7(k)$ 当 k 从零略微增加时，很快趋于零，于是它们很自然应该被剔除。剔除它们之后，重作岭回归分析，岭迹基本稳定，如图 7.5 (b)所示。因此剔除 x_5 和 x_7 是合理的。

以上两个例子是引用的有关参考文献的实例，只引用了计算结果，没有给出计算过程，目的在于使读者对岭回归的运用方法有个全面的了解。下面结合例 3.3 民航客运的数据，介绍如何用 SPSS 软件做岭回归分析。

例 7.4

第 6 章我们采用剔除变量的方法解决民航客运数据的多重共线性问题，现在再用岭回归方法处理多重共线性问题。

SPSS 软件的岭回归功能要用语法命令实现，菜单对话框中没有此功能。运行岭回归程序的步骤如下：

(1) 进入 SPSS 软件，录入变量数据或调入已有的数据文件。

(2) 进入 Syntax 语法窗口。方法是依次点击 File→New→Syntax。

(3) 录入如下的语法命令：

INCLUDE'c: \ Program Files \ SPSS \ Ridge regression. sps'.

RIDGEREG DEP=y/ENTER $x1$ $x2$ $x3$ $x4$ $x5$.

(4) 运行。方法是在 Syntax 窗口的主菜单 Run 命令下选择 All。

很快软件就会运行并输出计算结果，见表 7.3，岭迹图见图 7.6 (a)。

表 7.3

k	RSQ	x_1	x_2	x_3	x_4	x_5
0.000 00	0.998 23	2.447 386	−2.485 10	−0.083 140	0.530 538	0.563 537
0.050 00	0.990 37	0.223 417	0.179 160	−0.083 524	0.370 635	0.250 365
0.100 00	0.988 73	0.239 543	0.214 116	−0.073 407	0.324 407	0.227 824
0.150 00	0.987 29	0.243 335	0.224 695	−0.065 966	0.303 476	0.218 969
0.200 00	0.985 71	0.243 539	0.228 490	−0.059 635	0.290 298	0.213 730
0.250 00	0.983 93	0.242 291	0.229 504	−0.054 040	0.280 606	0.209 925
0.300 00	0.981 95	0.240 355	0.229 147	−0.049 021	0.272 829	0.206 823
0.350 00	0.979 78	0.238 068	0.228 038	−0.044 482	0.266 245	0.204 118
0.400 00	0.977 43	0.235 599	0.226 490	−0.040 356	0.260 473	0.201 660
0.450 00	0.974 93	0.233 041	0.224 676	−0.036 590	0.255 291	0.199 367
0.500 00	0.972 28	0.230 447	0.222 700	−0.033 141	0.250 558	0.197 194

专业课辅导+真题资料+微信公众号：432应统书院

续前表

k	RSQ	x_1	x_2	x_3	x_4	x_5
0.550 00	0.969 49	0.227 850	0.220 626	−0.029 975	0.246 181	0.195 112
0.600 00	0.966 59	0.225 269	0.218 497	−0.027 060	0.242 095	0.193 102
0.650 00	0.963 59	0.222 719	0.216 340	−0.024 372	0.238 253	0.191 151
0.700 00	0.960 48	0.220 206	0.214 174	−0.021 887	0.234 619	0.189 253
0.750 00	0.957 29	0.217 735	0.212 012	−0.019 587	0.231 166	0.187 400
0.800 00	0.954 02	0.215 309	0.209 865	−0.017 453	0.227 873	0.185 588
0.850 00	0.950 67	0.212 930	0.207 737	−0.015 471	0.224 721	0.183 814
0.900 00	0.947 26	0.210 599	0.205 634	−0.013 627	0.221 696	0.182 075
0.950 00	0.943 80	0.208 316	0.203 560	−0.011 910	0.218 788	0.180 369
1.000 00	0.940 28	0.206 080	0.201 515	−0.010 308	0.215 985	0.178 695

图 7.6

专业课辅导+真题资料+微信公众号：432应统书院

表 7.3 中的第 1 列为岭参数 k，软件默认值 k 从 0 到 1，步长为 0.05，共有 21 个 k 值。第 2 列是决定系数 R^2，第 3～7 列是标准化岭回归系数，其中第 1 行 $k=0$ 的数值就是普通最小二乘估计的标准化回归系数。

可以看到，变量 x_2 的岭回归系数 $\hat{\beta}_2(k)$ 从负值迅速变为正值，$|\hat{\beta}_1(k)|$ 和 $|\hat{\beta}_2(k)|$ 都迅速减少，两者之和比较稳定。从岭回归的角度看，x_1 与 x_2 只要保留一个就可以了。x_3，x_4，x_5 的岭回归系数则相对稳定。

通过上面的分析，我们决定剔除 x_1，用 y 与其余 4 个自变量作岭回归。把岭参数步长改为 0.02，范围减小到 0.2。这需要增加一句语法程序，点击主菜单的 Window→Syntax →Editor 返回语法窗口，语法命令如下：

```
INCLUDE'c:\Program Files\SPSS\Ridge regression.sps'.
RIDGEREG DEP=y /ENTER x2 x3 x4 x5
/START=0.0/STOP=0.2/INC=0.02.
```

然后在 Run 命令下选择 Current 运行，输出结果见表 7.4。

表 7.4

k	RSQ	x_2	x_3	x_4	x_5
0.000 00	0.995 18	−0.232 694	−0.134 119	0.787 697	0.516 538
0.020 00	0.992 73	0.191 301	−0.104 683	0.518 190	0.333 153
0.040 00	0.991 61	0.260 930	−0.097 765	0.464 546	0.305 788
0.060 00	0.990 84	0.287 851	−0.093 137	0.438 607	0.295 463
0.080 00	0.990 12	0.301 168	−0.089 162	0.422 128	0.289 970
0.100 00	0.989 38	0.308 468	−0.085 489	0.410 136	0.286 374
0.120 00	0.988 59	0.312 599	−0.082 010	0.400 676	0.283 666
0.140 00	0.987 73	0.314 871	−0.078 687	0.392 810	0.281 425
0.160 00	0.986 80	0.315 969	−0.075 501	0.386 027	0.279 453
0.180 00	0.985 80	0.316 284	−0.072 442	0.380 023	0.277 644
0.200 00	0.984 74	0.316 058	−0.069 503	0.374 602	0.275 942

由表 7.4 看到，剔除 x_1 后岭回归系数变化幅度减小，$\hat{\beta}_2(0)=-0.232\ 694$，虽然仍为负值，但与剔除 x_1 前的 −2.485 1 相比，负的程度已经大大减小。从岭迹图 7.6（b）看，岭参数 k 在 0.04～0.10 之间时，岭参数已经基本稳定，当 $k=0.08$ 时，$R^2=0.990\ 12$ 仍然很大，因而可以选取岭参数 $k=0.08$。然后给定 $k=0.08$，重新作岭回归，语法命令如下：

```
INCLUDE'c:\Program Files\SPSS\Ridge regression.sps'.
RIDGEREG DEP=y /ENTER x2 x3 x4 x5
/k=0.08.
```

计算结果如输出结果 7.1 所示。

输出结果 7.1　　$k=0.08$ 的岭回归

```
Mult R          .9950480
R Square        .9901205
Adj R Squ       .9865280
SE            111.5042453
```

专业课辅导+真题资料+微信公众号：432应统书院

ANOVA table

	df	SS	MS	F value	Sig F
Regress	4	13706607	3426651.6	275.6050375	.0000000
Residual	11	136765.16	12433.197		

	B	SEB	Beta	T	Sig T
X2	.0680451	.0061320	.3011683	11.0967696	.0000003
X3	-0.0077794	.0026574	-.0891621	-2.9274825	.0137522
X4	17.1686519	1.5487144	.4221279	11.0857441	.0000003
X5	.2239279	.0322228	.2899703	6.9493613	.0000243
Constant	424.7925673	262.7290140	.0000000	1.6168468	.1342035

得到 y 对 x_2，x_3，x_4，x_5 的标准化岭回归方程为

$$\hat{y} = 0.301\,2x_2 - 0.089\,16x_3 + 0.422\,1x_4 + 0.290\,0x_5$$

未标准化的岭回归方程为

$$\hat{y} = 424.79 + 0.068\,0x_2 - 0.007\,779x_3 + 17.169x_4 + 0.223\,9x_5$$

与第 6 章剔除变量法相比，岭回归方法多保留了自变量 x_2，如果希望回归方程中多保留一些自变量，那么岭回归方法是很有用的方法。

现在进一步计算出含有全部 5 个自变量的岭回归，与普通最小二乘的结果做一个比较。取岭参数 $k=0.10$，得岭回归方程为

$$\hat{y} = 323.3 + 0.034\,64x_1 + 0.048\,38x_2 - 0.006\,405x_3 + 13.19x_4 + 0.175\,9x_5$$

普通最小二乘回归方程为

$$\hat{y} = 450.9 + 0.354x_1 - 0.561x_2 - 0.007\,3x_3 + 21.578x_4 + 0.435x_5$$

显然岭回归方程比普通最小二乘回归方程的实际意义解释更为容易。

这一节的最后，我们再用一个例子说明共线性的影响和岭回归的作用。

例 7.5

回顾第 3 章思考与练习第 12 题，建立 GDP 对第一产业增加值 x_1 和第二产业增加值 x_2 的二元线性回归，得到二元回归方程 $\hat{y} = 2\,914.6 + 0.607\,x_1 + 1.709x_2$。其中回归系数 $\hat{\beta}_1 = 0.607$ 的解释是：当第二产业增加值 x_2 保持不变时，第一产业增加值 x_1 每增加 1 亿元，GDP 增加 0.607 亿元，这个解释明显不合理。这是因为第一产业增加值 x_1 每增加 1 亿元已经使 GDP 增加 1 亿元，同时还要考虑到没有包含在回归方程中的第三产业增加值 x_3 也要做相应调整，因此合理的 $\hat{\beta}_1$ 的数值应该大于 1。实际上，x_3 的年增长幅度大于 x_1 的年增长幅度，在假设 x_1 增加 1 亿元时，x_3 的增加幅度应大于 1 亿元，这时 x_3 为 GDP 带来的增加幅度要大于 1。因此合理的 $\hat{\beta}_1$ 的数值不仅是大于 1，而是要大于 2。

专业课辅导+真题资料+微信公众号：432应统书院

这个问题产生的原因仍然是存在共线性，在此省略掉计算过程，得岭参数$k=0.03$时岭回归方程为 $\hat{y}=-2\,250.6+2.270\,x_1+1.271\,x_2$，与普通最小二乘回归方程相差很大。岭回归系数 $\hat{\beta}_1=2.270$ 与前面的分析是吻合的，其解释是当第二产业增加值 x_2保持不变时，第一产业增加值 x_1每增加 1 亿元，GDP 增加 2.270 亿元，这个解释是合理的。

7.6 本章小结与评注

本章较系统地介绍了岭回归的思想和方法，并结合实际例子说明了岭回归方法在自变量选择和克服多重共线性方面的应用。

岭回归方法与普通最小二乘法的一个质的区别是，岭回归估计不再是无偏估计。长期以来，人们普遍认为一个好的估计应该满足无偏性，普通最小二乘估计就具有无偏性的重要特点，但当设计矩阵 $\boldsymbol{X}$ 退化时，最小二乘估计变得很不理想。岭回归估计就是针对一些实际问题的最小二乘估计明显变坏而提出的一种新的估计方法。岭回归法实际上是通过对最小二乘法的改进，允许回归系数的有偏估计量存在而解决多重共线性问题的方法之一。

如果一个估计量只有很小的偏差，但它的精度大大高于无偏估计量，人们可能更愿意选择这个估计量，因为它接近真实参数值的可能性较大，图 7.7 说明了这种情况。由图 7.7 可看到，估计量 $\hat{\boldsymbol{\beta}}$ 是无偏的，但不精确，而估计量 $\hat{\boldsymbol{\beta}}'$ 精度高却有小的偏差。$\hat{\boldsymbol{\beta}}'$ 落在真值 $\boldsymbol{\beta}$ 附近的概率远远大于无偏估计量 $\hat{\boldsymbol{\beta}}$。

图 7.7

岭回归估计的回归系数 $\hat{\beta}_j(k)$ 是有偏的，但往往比普通最小二乘估计量更稳定。因此，当回归模型有严重的多重共线性时，普通最小二乘法很不理想，人们就更多地推崇岭回归方法。这里需要注意的是，虽然 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是 $\boldsymbol{y}$ 的线性估计形式，但在实际应用时，总是要通过数据来确定 k，因而 k 也依赖于 $\boldsymbol{y}$，也是随机的，因此从本质上说 $\hat{\boldsymbol{\beta}}(k)$ 实为非线性估计。在实际应用中，只有当对最

专业课辅导+真题资料+微信公众号：432应统书院

小二乘估计的结果不满意时，才考虑使用岭回归。

岭回归计算程序 Ridge regression. sps' 是 SPSS 软件的附加功能，在 Windows 下点击：开始→程序，我们看到在 SPSS for Windows 运行程序下面还有一个 SPSS Production Facility 程序，其中 Production Facility 的含义是设备制造厂，表示能够编制程序。进入 SPSS Production Facility，点击 Add 按钮，我们看到几个程序名，其中的 Ridge regression 就是我们作岭回归所用的，这是 SPSS 软件附带的程序。

SPSS Production Facility 程序通常安装在计算机的 C：\ Program Files \ SPSS 目录下，岭回归计算程序第一行 INCLUDE'c：\ Program Files \ SPSS \ Ridge regression. sps'. 要指明此路径。如果你的计算机中 SPSS 软件是安装在其他硬盘上，例如 D 盘，则要把岭回归计算程序第一行改为 INCLUDE'd：\ SPSS \ Ridge regression. sps'.，其余情况依此类推。

Hoerl 和 Kennard 于 1970 年还提出了岭估计的一种推广形式，称为广义岭估计。普通的岭回归估计是给样本相关阵的主对角线加上相同的常数 k，广义岭回归是给样本相关阵的主对角线加上各不相同的常数 k_j，有兴趣的读者请参见参考文献 [2] 和 [5]。

□ 思考与练习

7.1 岭回归估计是在什么情况下提出的？

7.2 岭回归估计的定义及其统计思想是什么？

7.3 选择岭参数 k 有哪几种主要方法？

7.4 用岭回归方法选择自变量应遵从哪些基本原则？

7.5 对第 5 章思考与练习第 9 题的数据，逐步回归的结果只保留了 3 个自变量 x_1，x_2，x_5，用 y 对这 3 个自变量作岭回归分析。

7.6 对第 3 章思考与练习第 12 题的问题，分别用普通最小二乘和岭回归建立 GDP 对第二产业增加值 x_2 和第三产业增加值 x_3 的二元线性回归，解释所得到的回归系数。

7.7 一家大型商业银行有多家分行，近年来，该银行的贷款额平稳增长，但不良贷款额也有较大比例的提高。为弄清楚不良贷款形成的原因，希望利用银行业务的有关数据做些定量分析，以便找出控制不良贷款的办法。表 7.5 是该银行所属 25 家分行 2002 年的有关业务数据。

(1) 计算 y 与其余 4 个变量的简单相关系数。

(2) 建立不良贷款 y 对 4 个自变量的线性回归方程，所得的回归系数是否合理？

(3) 分析回归模型的共线性。

(4) 采用后退法和逐步回归法选择变量，所得回归方程的回归系数是否合理，是否还存在共线性？

专业课辅导+真题资料+微信公众号：432应统书院

(5) 建立不良贷款 y 对 4 个自变量的岭回归。

(6) 对第 (4) 步剔除变量后的回归方程再做岭回归。

(7) 某研究人员希望做 y 对各项贷款余额、本年累计应收贷款、贷款项目个数这 3 个自变量的回归，你认为这样做是否可行？如果可行应该如何做？

表 7.5　银行不良贷款数据

分行编号	不良贷款 y（亿元）	各项贷款余额 x_1（亿元）	本年累计应收贷款 x_2（亿元）	贷款项目个数 x_3（个）	本年固定资产投资额 x_4（亿元）
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173.0	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	1.0	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3
11	0.3	64.2	2.1	11	42.7
12	4.0	132.2	11.2	23	76.7
13	0.8	58.6	6.0	14	22.8
14	3.5	174.6	12.7	26	117.1
15	10.2	263.5	15.6	34	146.7
16	3.0	79.3	8.9	15	29.9
17	0.2	14.8	0.6	2	42.1
18	0.4	73.5	5.9	11	25.3
19	1.0	24.7	5.0	4	13.4
20	6.8	139.4	7.2	28	64.3
21	11.6	368.2	16.8	32	163.9
22	1.6	95.7	3.8	10	44.5
23	1.2	109.6	10.3	14	67.9
24	7.2	196.2	15.8	16	39.7
25	3.2	102.2	12.0	10	97.1

专业课辅导+真题资料+微信公众号：432应统书院

第 8 章 Chapter 8 主成分回归与偏最小二乘

对不满足模型基本假设的回归建模，这一章主要介绍另外两种改进方法，即主成分回归和偏最小二乘。

8.1 主成分回归

主成分回归（principal components regression，PCR）是对普通最小二乘估计的一种改进，它的参数估计是一种有偏估计。马西（W. F. Massy）于 1965 年根据多元统计分析中的主成分分析提出了主成分回归。为了使读者更容易理解主成分回归，本节首先介绍有关主成分分析的基本思想和性质，然后用实例介绍主成分回归的应用。

一、主成分的基本思想

主成分分析（principal components analysis，PCA）也称主分量分析，首先由霍特林（Hotelling）于 1933 年提出。主成分分析是用一种降维的思想，在损失很少信息的前提下把多个指标利用正交旋转变换转化为几个综合指标的多元统计分析方法。通常把转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关。这样在研究复杂问题时就可以只考虑少数几个主成分且不至于损失太多信息，从而更容易抓住主要矛盾，揭示事物内部变量之间的规律性，同时使问题得到简化，提高分析效率。

设对某一事物的研究涉及 p 个指标，分别用 X_1，X_2，…，X_p 表示，这 p 个指标构成的 p 维随机向量为 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$。设随机向量 $\boldsymbol{X}$ 的均值为 μ，协

专业课辅导+真题资料+微信公众号：432应统书院

方矩阵为 $\boldsymbol{\Sigma}$。

对 $\boldsymbol{X}$ 进行线性变换，可以形成新的综合变量，用 $\boldsymbol{Y}$ 表示，也就是说，新的综合变量可以由原来的变量线性表示，即满足下式：

$$\begin{cases} Y_1=\mu_{11}X_1+\mu_{12}X_2+\cdots+\mu_{1p}X_p \\ Y_2=\mu_{21}X_1+\mu_{22}X_2+\cdots+\mu_{2p}X_p \\ \cdots\cdots \\ Y_p=\mu_{p1}X_1+\mu_{p2}X_2+\cdots+\mu_{pp}X_p \end{cases}$$

由于可以任意地对原始变量进行上述线性变换，得到的综合变量 $\boldsymbol{Y}$ 的统计特性也不尽相同，因此为了取得较好的效果，我们总是希望 $Y_i=\boldsymbol{\mu}_i'\boldsymbol{X}$ 的方差尽可能大且各 Y_i 之间互相独立，由于

$$\mathrm{var}(Y_i)=\mathrm{var}(\boldsymbol{\mu}_i'\boldsymbol{X})=\boldsymbol{\mu}_i'\boldsymbol{\Sigma}\boldsymbol{\mu}_i$$

而对于任意常数 c，有

$$\mathrm{var}(c\boldsymbol{\mu}_i'\boldsymbol{X})=c\boldsymbol{\mu}_i'\boldsymbol{\Sigma}\boldsymbol{\mu}_i c=c^2\boldsymbol{\mu}_i'\boldsymbol{\Sigma}\boldsymbol{\mu}_i$$

因此，对 $\boldsymbol{\mu}_i$ 不加限制时，可使 $\mathrm{var}(Y_i)$ 任意增大，问题将变得没有意义。我们将线性变换约束在下面的原则之下：

(1) $\boldsymbol{\mu}_i'\boldsymbol{\mu}_i=1$，即 $\mu_{i1}^2+\mu_{i2}^2+\cdots+\mu_{ip}^2=1(i=1,2,\cdots,p)$。

(2) Y_i 与 Y_j 不相关 $(i\neq j;\ i,j=1,2,\cdots,p)$。

(3) Y_1 是 $X_1,X_2,\cdots,X_p$ 的所有满足原则 1 的线性组合中方差最大者；Y_2 是与 Y_1 不相关的 $X_1,X_2,\cdots,X_p$ 的所有线性组合中方差最大者；……；Y_p 是与 $Y_1,Y_2,\cdots,Y_{p-1}$ 都不相关的 $X_1,X_2,\cdots,X_p$ 的所有线性组合中方差最大者。

基于以上三条原则决定的综合变量 $Y_1,Y_2,\cdots,Y_p$ 分别称为原始变量的第一、第二……第 p 个主成分。其中，各综合变量在总方差中占的比重依次递减。在实际研究工作中，通常只挑前几个方差最大的主成分，从而达到简化系统结构、抓住问题实质的目的。

二、主成分的基本性质

引论：设矩阵 $\boldsymbol{A}'=\boldsymbol{A}$，将 $\boldsymbol{A}$ 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 依大小顺序排列，不妨设 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$，$\gamma_1,\gamma_2,\cdots,\gamma_p$ 为矩阵 $\boldsymbol{A}$ 各特征值对应的标准正交向量，则对任意向量 $\boldsymbol{x}$，有

$$\max_{x\neq 0}\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}=\lambda_1,\cdots,\min_{x\neq 0}\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}=\lambda_p$$

结论：设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 的协方差矩阵为 $\boldsymbol{\Sigma}$，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 为 $\boldsymbol{\Sigma}$ 的特征值，$\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_p$ 为矩阵 $\boldsymbol{\Sigma}$ 各特征值对应的标准正交向量，则第 i 个主成分为

专业课辅导+真题资料+微信公众号：432应统书院

$$Y_i=\gamma_{1i}X_1+\gamma_{2i}X_2+\cdots+\gamma_{pi}X_p,\ i=1,2,\cdots,p$$

此时，

$$\mathrm{var}(Y_i)=\boldsymbol{\gamma}'_i\boldsymbol{\Sigma}\boldsymbol{\gamma}_i=\lambda_i$$
$$\mathrm{cov}(Y_i,Y_j)=\boldsymbol{\gamma}'_i\boldsymbol{\Sigma}\boldsymbol{\gamma}_j=0,\ i\neq j$$

由以上结论，我们把 X_1，X_2，…，X_p 的协方差阵 $\boldsymbol{\Sigma}$ 的非零特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$ 对应的标准化特征向量 $\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$，…，$\boldsymbol{\gamma}_p$ 分别作为系数向量，$Y_1=\boldsymbol{\gamma}'_1\boldsymbol{X}$，$Y_2=\boldsymbol{\gamma}'_2\boldsymbol{X}$，…，$Y_p=\boldsymbol{\gamma}'_p\boldsymbol{X}$ 分别称为随机向量 $\boldsymbol{X}$ 的第一主成分、第二主成分……第 p 主成分。

性质 1 $\boldsymbol{Y}$ 的协方差阵为对角矩阵 $\boldsymbol{\Lambda}$。其中对角线上的值为 λ_1，λ_2，…，λ_p。

性质 2 记 $\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有 $\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$。

称 $\alpha_k=\dfrac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$（$k=1$，2，…，$p$）为第 k 个主成分 Y_k 的方差贡献率，称 $\dfrac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$ 为主成分 Y_1，Y_2，…，Y_m 的累积贡献率。

性质 3 $\rho(Y_k,X_i)=\mu_{ki}\sqrt{\lambda_k}/\sqrt{\sigma_{ii}}$（$k$，$i=1$，2，…，$p$）。

其中，第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k,X_i)$ 称为因子负荷量。因子负荷量是主成分解释中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。

性质 4 $\sum_{i=1}^{p}\rho^2(Y_k,X_i)\sigma_{ii}=\lambda_k$

性质 5 $\sum_{i=1}^{p}\rho^2(Y_k,X_i)=\dfrac{1}{\sigma_{ii}}\sum_{k=1}^{p}\lambda_k\mu_{ki}^2=1$

X_i 与前 m 个主成分 Y_1，Y_2，…，Y_m 的全相关系数平方和称为 Y_1，Y_2，…，Y_m 对原始变量 X_i 的方差贡献率 ν_i，即 $\nu_i=\dfrac{1}{\sigma_{ii}}\sum_{k=1}^{m}\lambda_k\mu_{ki}^2(i=1,2,\cdots,p)$。这一定义说明前 m 个主成分提取了原始变量 X_i 中 ν_i 的信息，由此可以判断提取的主成分说明原始变量的能力。

三、主成分回归的实例

为了避免变量的量纲不同所产生的影响，先将数据中心标准化，中心标准化后的自变量样本观测数据矩阵 $\boldsymbol{X}^*$ 就是 n 行 p 列的矩阵，$r=(\boldsymbol{X}^*)'\boldsymbol{X}^*$ 就是相关阵。

专业课辅导+真题资料+微信公众号：432应统书院

例 8.1

下面以例 3.3 民航客运量的数据为例介绍主成分回归方法。首先对 5 个自变量计算主成分，点击 Analyze→Dimension Reduction→Factor 进入因子分析（包含主成分分析）对话框。把 $x_1 \sim x_5$ 这 5 个变量选入 Variables 变量框条中，点击 Extraction 把抽取因子（主成分）数目（Fixed number of factors）设为 5，也就是原始变量数目。点击 Scores 保存因子得分。返回主对话框，其他设置都用默认值，点击 OK 运行。需要注意计算过程中数据的精确度，SPSS 默认的输出只显示 3 位小数，可以双击输出表格选定所需数据，用鼠标右键选 Cell Properties 更改数据显示位数。

输出结果 8.1 中有 5 个主成分的特征值（Eigenvalues），最大的是 $\lambda_1 = 3.991\,319\,6$，最小的是 $\lambda_5 = 0.000\,271\,6$。方差百分比（% of Variance）反映主成分所能解释数据变异的比例，也就是包含原数据的信息比例。第一个主成分 Factor1 的方差百分比=79.826%，含有原始 5 个变量近 80%的信息量；前两个主成分累积含有原始 5 个变量近 98.47%的信息量。因此取两个主成分已经足够。

输出结果 8.1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.9913196	79.826	79.826	3.991	79.826	79.826
2	0.9320642	18.641	98.468	.932	18.641	98.468
3	0.0651585	1.303	99.771	.065	1.303	99.771
4	0.0111861	.224	99.995	.011	.224	99.995
5	0.0002716	.005	100.000	.000	.005	100.000

Extraction Method: Principal Component Analysis.

输出结果 8.2 是因子载荷阵，也就是由 $\rho(Y_k, X_i) = \mu_{ki}\sqrt{\lambda_k}/\sqrt{\sigma_{ii}}$ $(k, i=1, 2, \cdots, p)$ 构成的 p 阶方阵。不妨设因子载荷阵为矩阵 $\boldsymbol{A}$，由 $\mu_{ki}(i, k=1, 2, \cdots, p)$ 构成的矩阵为 $\boldsymbol{U}$。由于分析时由相关阵出发，故 $\sqrt{\sigma_{ii}}=1(i=1, 2, \cdots, p)$，所以 $\rho(Y_k, X_i)=\mu_{ki}\sqrt{\lambda_k}(k, i=1, 2, \cdots, p)$。即每个主成分的表达式中，$X_i^*$ 前的系数为第 i 列的因子载荷阵除以 $\sqrt{\lambda_i}$。由于此时得到的是 X_i^* 的系数，我们还得再转换成 X_i 的系数。

SPSS 在做因子分析时可以输出因子得分，表 8.1 是其数值。因此，我们可以通过因子得分的关系来算出主成分得分。用主成分法做因子分析得到的对第 i 个样本的因子得分为 $f_{(i)} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}_{(i)}(i=1, 2, \cdots, n)$，用主成分法得到的对第 i 个样本的主成分得分是 $z_{(i)} = \boldsymbol{U}\boldsymbol{X}_{(i)}$。可以看到，$z_{(ik)} = \sqrt{\lambda_k} f_{(ik)}(i=1, 2, \cdots, n; k=1, 2, \cdots, p)$。由于前两个主成分的方差累积贡献率已经达到 98.468%，故只

专业课辅导+真题资料+微信公众号：432应统书院

需保留前两个主成分。点击 Transform→Compute Variable 进入变量计算。在 Target Variable 中分别输入 prin1 和 prin2，在 Numeric Expression 中分别输入 sqrt(λ_1 的值) * Fac1 _ 1 和 sqrt（λ_2 的值）* Fac2 _ 1，这样就得出主成分，表 8.2 是其数值。

输出结果 8.2

Component Matrix[a]

	Component				
	1	2	3	4	5
x1	0.9848782	-0.1654724	.018	.047	.012
x2	0.9896541	-0.1320637	.000	.055	-.011
x3	0.4128518	0.9083730	.066	.007	.000
x4	0.9631263	-0.2137298	.150	-.064	-.001
x5	0.9715248	0.1281417	-.195	-.043	.000

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

表 8.1

FAC1 _ 1	FAC2 _ 1	FAC3 _ 1	FAC4 _ 1	FAC5 _ 1
−1.289 40	−1.481 00	−0.274 58	0.424 56	1.355 96
−1.154 66	−1.051 09	−0.277 13	0.281 88	1.047 86
−1.002 50	−0.574 32	0.256 10	−0.084 16	0.001 32
−0.884 60	−0.315 39	0.278 78	−0.619 88	−0.885 59
−0.796 64	0.064 41	0.820 47	−0.247 31	−1.143 11
−0.676 95	0.604 04	0.994 81	0.382 79	−0.235 07
−0.472 73	0.949 85	0.921 27	0.056 03	0.642 51
−0.233 79	1.074 89	0.107 51	0.199 28	0.568 35
−0.050 87	0.726 89	−0.921 33	−1.388 13	−1.027 69
0.234 03	0.983 23	−0.856 20	−2.108 90	0.067 58
0.594 55	1.835 45	−1.119 74	0.832 75	1.616 44
0.580 79	0.807 78	0.810 17	2.152 83	−1.387 08
0.587 62	−0.849 24	−1.176 01	0.818 98	−1.512 73
0.907 33	−0.992 31	−2.051 51	0.568 83	−0.193 42
1.483 15	−1.021 30	1.288 66	−1.180 53	−0.036 30
2.174 68	−0.761 90	1.198 73	−0.089 03	1.120 98

表 8.2

prin1	prin2
−2.576 00	−1.429 81
−2.306 81	−1.014 76

专业课辅导+真题资料+微信公众号：432应统书院

续前表

prin1	prin2
−2.002 82	−0.554 47
−1.767 27	−0.304 49
−1.591 56	0.062 18
−1.352 42	0.583 16
−0.944 44	0.917 02
−0.467 08	1.037 74
−0.101 63	0.701 76
0.467 54	0.949 25
1.187 80	1.772 01
1.160 32	0.779 86
1.173 96	−0.819 88
1.812 68	−0.958 01
2.963 08	−0.986 00
4.344 63	−0.735 57

现在用 y 对前两个主成分 prin1 和 prin2 做普通最小二乘回归，得主成分回归方程为

$$\hat{y}=1\,159.125+468.900\text{prin1}-192.531\text{prin2}$$

我们想要得到 prin1 和 prin2 用原始自变量表示的线性组合，然后代回上式的回归方程即可。可以分别用两个主成分 prin1 和 prin2 做因变量，以 5 个原始自变量为自变量做线性回归，所得的回归系数就是所需要的线性组合的系数。这个回归中的残差为 0，这是因为主成分就是自变量的线性组合，是确定的函数关系，所做的回归相当于解一个线性方程组。得到

$$\begin{aligned}\text{prin1}=&-4.923\,39+0.000\,074\,20x_1+0.000\,116\,50x_2+0.000\,018\,77x_3\\&+0.020\,409\,94x_4+0.000\,390\,91x_5\\\text{prin2}=&-8.134\,88-0.000\,025\,80x_1-0.000\,032\,17x_2+0.000\,085\,45x_3\\&-0.009\,372\,58x_4+0.000\,106\,70x_5\end{aligned}$$

还原后的主成分回归方程为

$$\hat{y}=416.76+0.039\,76x_1+0.060\,82x_2-0.007\,65x_3+11.37x_4+0.162\,8x_5$$

每个回归系数的解释也都合理。

一个需要注意的问题是，SPSS 软件保存的因子得分 Factor 与真正意义上的主成分 prin 相差一个常数倍，即 prin$=\sqrt{\lambda}$Factor。在实际应用中如果只是为了得到最终的还原为原始变量表述的主成分回归方程，可以直接用 y 对保留的 Factor 做回归，这样可以省略计算 prin 的步骤，具体算法可参照本章思考与练习第 3 题的参考答案。

专业课辅导+真题资料+微信公众号：432应统书院

8.2 偏最小二乘

一、偏最小二乘的基本原理

在经济问题的研究中遇到的回归问题往往有两个特点：一是自变量 x_1，x_2，…，x_k 的数目比较多，常会碰到有几十个 x_i，而观察的时点并不多。二是回归方程建立后主要的应用是预测。用符号来表示，就是对因变量 y 和自变量 x_1，x_2，…，x_k 观测了 n 组数据：

$$(y_t, x_{t1}, x_{t2}, \cdots, x_{tk}), \quad t=1,2,\cdots,n \tag{8.1}$$

假定它们之间有关系式

$$y=\beta_0+\beta_1 x_{t1}+\beta_2 x_{t2}+\cdots+\beta_k x_{tk}+\varepsilon_t \tag{8.2}$$

其中，ε_t 为误差。我们要用观测值（8.1）式去求（8.2）式中 β_i 的估计值 $\hat{\beta}_i$，从而得到回归方程

$$y=\hat{\beta}_0+\hat{\beta}_1 x_{t1}+\hat{\beta}_2 x_{t2}+\cdots+\hat{\beta}_k x_{tk} \tag{8.3}$$

当 $n>k$ 时，最小二乘法就可以求出 $\hat{\beta}_i$，从而得到（8.3）式。然而现在的问题是 $k>n$，通常的最小二乘法无法进行。

从（8.2）式来看，我们并不需要很多自变量，实际上只要 x_1，x_2，…，x_k 的一个线性函数 $\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$ 就行了。通常的最小二乘法，就是寻求 $\{x_i\}$ 的线性函数中与 y 的相关系数绝对值达到最大的一个。这时须求 $\boldsymbol{X}'\boldsymbol{X}$ 的逆矩阵，$\boldsymbol{X}$ 是由自变量 x_1，x_2，…，x_k 的观测值组成的矩阵，即

$$\boldsymbol{X}=(x_{ti})=\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

当 $k>n$ 时，$\boldsymbol{X}'\boldsymbol{X}$ 是一个奇异矩阵，无法求逆。主成分回归（PCR）就不求 $\boldsymbol{X}'\boldsymbol{X}$ 的逆，而直接求 $\boldsymbol{X}'\boldsymbol{X}$ 的特征根。把它的非零的特征根记为 λ_i，如果有 r 个，r 就是 $\boldsymbol{X}'\boldsymbol{X}$ 的秩，将它们按大小顺序排出，得 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$，相应的特征向量分别记为 $\boldsymbol{\alpha}_1$，$\boldsymbol{\alpha}_2$，…，$\boldsymbol{\alpha}_r$，它们均为 $k\times 1$ 向量，令 $\boldsymbol{\alpha}_i$ 的分量为 α_{ij}，即 $\boldsymbol{\alpha}'_1=(\alpha_{i1}, \cdots, \alpha_{ik})$，又令

$$z_i=\boldsymbol{\alpha}'_i\boldsymbol{X}=\alpha_{i1}x_1+\alpha_{i2}x_2+\cdots+\alpha_{ik}x_k, \quad i=1,2,\cdots,r \tag{8.4}$$

则 z_1，z_2，…，z_r 都是 x_1，x_2，…，x_k 的线性函数，$r<k$，且 $r<n$，因此将 y 对 z_1，z_2，…，z_r 或 z_1，z_2，…，z_r 的一部分做回归就可以了，这就是 PCR 的主要想法。

专业课辅导+真题资料+微信公众号：432应统书院

PCR 虽然解决了 $k>n$ 这一矛盾，但它选 z_i 的方法与因变量 y 无关，只在自变量 x_1，x_2，…，x_k 中去寻找有代表性的 z_1，z_2，…，z_r。偏最小二乘（partial least squares，PLS）在这一点上就与 PCR 不同，它寻找 x_1，x_2，…，x_k 的线性函数时，考虑与 y 的相关性，选择与 y 相关性较强又能方便算得的 x_1，x_2，…，x_k 的线性函数。它的算法是最小二乘，但是它只选 x_1，x_2，…，x_k 中与 y 有相关性的变量，不考虑全部 x_1，x_2，…，x_k 的线性函数，只考虑偏向与 y 有关的一部分，所以称为偏最小二乘。具体的选法与最小二乘法有关，所以先回忆一下最小二乘法的公式很有好处，便于理解 PLS。

（y，x）共观测了 n 组数据（y_1，x_1），…，（y_n，x_n），于是 x，y 的线性回归方程为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$$

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}, \quad \bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i, \quad \bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i \\ \hat{\beta}_1=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\end{cases}$$

当 x_i，y_i 这些数据的均值为 0 时，$\hat{\beta}_0=0$，$\hat{\beta}_1$ 就有简单的形式，即有

$$\begin{cases}y=\hat{\beta}x \\ \hat{\beta}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i}{\sum\limits_{i=1}^{n}x_i^2}=\dfrac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}}\end{cases} \tag{8.5}$$

其中，$\boldsymbol{x}=\begin{pmatrix}x_1\\ \vdots \\ x_n\end{pmatrix}$，$\boldsymbol{y}=\begin{pmatrix}y_1\\ \vdots \\ y_n\end{pmatrix}$为观测值向量。PLS 就是反复利用（8.5）式。

首先将数据

$$\boldsymbol{y}=\begin{pmatrix}y_1\\ \vdots \\ y_n\end{pmatrix}, \quad \boldsymbol{X}=\begin{pmatrix}x_{11} & \cdots & x_{1k}\\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk}\end{pmatrix}$$

中心化，中心化之后得到的 $\tilde{y}_t$，$\tilde{x}_{ti}$ 相应的各自的均值都是 0。因此我们总假定原始数据 $\boldsymbol{y}$ 及 $\boldsymbol{X}$ 均已中心化，这样书写公式、算法时符号比较简单，即 $\boldsymbol{y}$ 和 $\boldsymbol{X}=(x_{ti})$满足

$$\sum_{t=1}^{n}y_t=0, \quad \sum_{t=1}^{n}x_{ti}=0, \quad i=1,2,\cdots,k \tag{8.6}$$

将 y 对每个自变量 x_i 单独作回归，用（8.5）式就得

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{y}(x_i)=\frac{\boldsymbol{x}_i'\boldsymbol{y}}{\boldsymbol{x}_i'\boldsymbol{x}_i}x_i,\quad x_i=\begin{bmatrix}x_{1i}\\ \vdots\\ x_{ni}\end{bmatrix},\quad i=1,2,\cdots,k \tag{8.7}$$

我们用 $\boldsymbol{x}_i$ 表示资料向量，x_i 表示自变量（不是数据）。(8.7) 式告诉我们与 y 有关的 x_i 的线性组合，应该是 (8.7) 式右端的量，将 (8.7) 式右端的量加权后，用 ω_i 记相应的权，就得到

$$\sum_{i=1}^{k}\omega_i\frac{\boldsymbol{x}_i'\boldsymbol{y}}{\boldsymbol{x}_i'\boldsymbol{x}_i}x_i$$

权 ω_i 可以有很多种选择，比较简单的是 $\omega_i=\boldsymbol{x}_i'\boldsymbol{x}_i$，代入上式就得 $\sum_{i=1}^{k}(\boldsymbol{x}_i'\boldsymbol{y})x_i$，可见这个 x_i 的线性组合是应入选的变量。令

$$t_1=\sum_{i=1}^{k}(\boldsymbol{x}_i'\boldsymbol{y})x_i \tag{8.8}$$

它相应的 n 个资料是

$$\boldsymbol{t}_1=\sum_{i=1}^{k}(\boldsymbol{x}_i'\boldsymbol{y})\boldsymbol{x}_i$$

容易看出，(8.8) 式的 t_1 中，它的系数与 y 有关，而不像 PCR 与 y 无关。将 t_1 作为自变量，对 y 求回归，用 (8.5) 式就有

$$\hat{y}(t_1)=\frac{\boldsymbol{t}_1'\boldsymbol{y}}{\boldsymbol{t}_1'\boldsymbol{t}_1}t_1$$

利用上式预测 y，得预测值向量 $\hat{\boldsymbol{y}}(t_1)$，即有

$$\hat{\boldsymbol{y}}(t_1)=\frac{\boldsymbol{t}_1'\boldsymbol{y}}{\boldsymbol{t}_1'\boldsymbol{t}_1}\boldsymbol{t}_1$$

于是得残差 $\boldsymbol{y}^{(1)}=\boldsymbol{y}-\hat{\boldsymbol{y}}(t_1)$。考虑到残差 $\boldsymbol{y}^{(1)}$ 中不再含 t_1 的信息，因此各个自变量 x_i 的作用对 y 而言，含 t_1 的部分已不具新的信息，都应删去。也就是将每个自变量 x_i 对 t_1 求回归，得回归方程（还是用 (8.5) 式）和预测值

$$\hat{\boldsymbol{x}}_i(t_1)=\frac{\boldsymbol{t}_1'\boldsymbol{x}_i}{\boldsymbol{t}_1'\boldsymbol{t}_1}\boldsymbol{t}_1,\quad i=1,2,\cdots,k$$

$\boldsymbol{x}_i$ 相应的残差，$\boldsymbol{x}_i^{(1)}=\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i(t_1)$（$i=1,2,\cdots,k$）。于是将 $\boldsymbol{y}^{(1)}$，$\boldsymbol{x}_i^{(1)}$，…，$\boldsymbol{x}_k^{(1)}$ 作为新的原始资料，重复上述步骤，逐步求得 t_1，t_2，…，t_r，r 是 $\boldsymbol{X}'\boldsymbol{X}$ 的秩。最后利用 y 对 t_2，t_3，…，t_r 用普通最小二乘做回归，经过变量间的转换，最终可得到 y 对 x_1，x_2，…，x_k 的回归方程，这种求得回归方程的方法就称为 PLS 法，即偏最小二乘法。

二、偏最小二乘的算法

从上面构造 t_1 的过程可得如下的算法（$\boldsymbol{X}$，$\boldsymbol{y}$ 资料已中心化，$\mathrm{rk}(\boldsymbol{X})=r$）：

专业课辅导+真题资料+微信公众号：432应统书院

Wold 算法

(1) $y \to y_0$，$X \to X_0$，$0 \to \hat{y}_0$，$0 \to \hat{X}_0$
(2) 对 $a=1$ 到 r 做：
(3) $t_a = X_{a-1} X'_{a-1} y_{a-1}$
(4) $\hat{y}_a = \frac{t_a t'_a}{t'_a t_a} y_{a-1} + \hat{y}_{a-1}$
(5) $y_a = y_{a-1} - \frac{t_a t'_a}{t'_a t_a} y_{a-1}$
(6) $\hat{X}_a = \frac{t_a t'_a}{t'_a t_a} X_{a-1}$
(7) $X_a = X_{a-1} - \hat{X}_a$
(8) $X'_a X_a$ 中主对角元素近似 0，就退出

上述算法完全体现了 PLS 的想法。1988 年 Helland 证明了下列事实，导出了一个更为简单的算法。这个证明利用了回归方程是观测向量 y 在自变量资料向量所张成的子空间中的投影，所以逐次求出 t_1，t_2，…，t_r 的投影矩阵是

$$\boldsymbol{P}_{t_i} = \boldsymbol{t}_i (\boldsymbol{t}'_i \boldsymbol{t}_i)^{-1} \boldsymbol{t}'_i = \frac{\boldsymbol{t}_i \boldsymbol{t}'_i}{\boldsymbol{t}'_i \boldsymbol{t}_i}$$

$$\boldsymbol{P}_{t_i} \boldsymbol{y} = \frac{\boldsymbol{t}'_i \boldsymbol{y}}{\boldsymbol{t}'_i \boldsymbol{t}_i} \boldsymbol{t}_i$$

我们用（t_1），（t_1，t_2），…，（t_1，t_2，…，t_r）分别表示由 t_1 张成的子空间，t_1，t_2 张成的子空间，等等，$\hat{y}_a$ 就是在（t_1，t_2，…，t_a）上的 $\boldsymbol{y}$ 的投影。如引入记号

$$\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}, \quad \boldsymbol{s} = \boldsymbol{X}'\boldsymbol{y}, \quad \boldsymbol{s}_1 = \boldsymbol{s}, \quad \boldsymbol{s}_r = \boldsymbol{S}^{k-1}\boldsymbol{s}, \quad k = 1, 2, \cdots, r$$

Helland 证明了

$$(\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_a) = (\boldsymbol{X}_{s_1}, \boldsymbol{X}_{s_2}, \cdots, \boldsymbol{X}_{s_a})$$

对 $a=1$，2，…，r 都成立。于是 PLS 算法可改为：

Helland 算法

(1) $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}$，$\boldsymbol{s} = \boldsymbol{X}'\boldsymbol{y}$
(2) 对 $a=1$ 到 r 做：
(3) $\boldsymbol{S}_a = \boldsymbol{S}'^{a-1}\boldsymbol{s}$
(4) y 对 $\boldsymbol{X}_{s_1}$，$\boldsymbol{X}_{s_2}$，…，$\boldsymbol{X}_{s_a}$ 做普通最小二乘回归得 $\hat{y}_a$
(5) 选择合适的 $\hat{y}_a$

上述算法中都存在一个问题，就是这个算法何时结束，什么是合适的 a，是否一定要算到某个 $\boldsymbol{X}_a$ 中的一列是 0 为止？

一般来说，可以自己规定一个你认为最切合你所研究的问题的标准。在已有的运用 PLS 的情况中，大部分都使用交叉验证（cross-validation）法。这个方法是这样的：

现在从资料 $\boldsymbol{X}$，$\boldsymbol{y}$ 中删去第 l 组资料，即删去（y_l，x_{l1}，…，x_{lk}），删去后的 $\boldsymbol{X}$，$\boldsymbol{y}$ 用 $\boldsymbol{X}(-l)$，$\boldsymbol{y}(-l)$ 表示；用 $\boldsymbol{X}(-l)$，$\boldsymbol{y}(-l)$ 作为原始资料，用 PLS 方法算出预测方程中 $\hat{y}_a$ 的表达式，然后用 $\hat{y}_a(-l)$ 表示这个预测方程的预测值，将

专业课辅导+真题资料+微信公众号：432应统书院

x_{l1}，…，x_{lk}代入$\hat{y}_a(-l)$，得它的预测值为$\hat{y}_{al}(-l)$，残差$\hat{y}_l-\hat{y}_{al}(-l)$就反映了第$a$步预测方程的好坏在第$l$组资料上的体现，于是

$$\sum_{l=1}^{n}(y_l-\hat{y}_{al}(-l))^2$$

就在整体上反映了第a步预测方程的好坏，把这个值记为损失$L(a)$，自然应该选a使$L(a)$达到最小，即应该选a_*，使

$$L(a_*)=\min_{1\leqslant a\leqslant r}L(a)$$

所以$L(a)$的计算没有必要添加新的程序，实际上重复使用就行了，当n不大时，更为方便。正因为使用了这个交叉验证方法，选出的预测方程往往效果比较好。

三、偏最小二乘的应用

在此介绍使用SPSS（需要安装PLS模块）对发电量模型运用偏最小二乘回归的方法。

例 8.2

对发电量需求和工业产量的关系进行建模，因变量y为发电量产量（亿千瓦小时），自变量x_1为原煤产量（亿吨），x_2为原油产量（万吨），x_3为天然气产量（亿立方米），x_4为生铁产量（万吨），x_5为纱产量（万吨），x_6为硫酸产量（万吨），x_7为烧碱（折100%）产量（万吨），x_8为纯碱产量（万吨），x_9为农用化肥产量（万吨），x_{10}为水泥产量（万吨），x_{11}为平板玻璃产量（万重量箱），x_{12}为钢产量（万吨），x_{13}为成品钢材产量（万吨）。数据见表8.3。

表 8.3

年份	y	x_1	x_2	x_3	x_4	x_5	x_6
1997	11 355.53	13.88	16 074.14	227.03	11 511.41	559.83	2 036.90
1998	11 670.00	13.32	16 100.00	232.79	11 863.67	542.00	2 171.00
1999	12 393.00	13.64	16 000.00	251.98	12 539.24	567.00	2 356.00
2000	13 556.00	13.84	16 300.00	272.00	13 101.48	657.00	2 427.00
2001	14 808.02	14.72	16 395.87	303.29	15 554.25	760.68	2 696.30
2002	16 540.00	15.50	16 700.00	326.61	17 084.60	850.00	3 050.40
2003	19 105.75	18.35	16 959.98	350.15	21 366.68	983.58	3 371.20
2004	22 033.09	21.23	17 587.33	414.60	26 830.99	1 291.34	3 928.90
2005	25 002.60	23.50	18 135.29	493.20	34 375.19	1 450.54	4 544.70
2006	28 657.26	25.29	18 476.57	585.53	41 245.19	1 742.96	5 033.20
2007	32 815.53	26.92	18 631.82	692.40	47 651.63	2 068.17	5 412.60
2008	34 957.61	28.02	19 043.06	802.99	47 824.42	2 170.92	5 098.00
2009	37 146.51	29.73	18 948.96	852.69	55 283.46	2 393.46	5 960.90

专业课辅导+真题资料+微信公众号：432应统书院

续前表

年份	y	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1997	11 355.53	574.40	725.76	2 820.96	51 173.80	16 630.70	10 894.20	9 978.93
1998	11 670.00	539.37	744.00	3 010.00	53 600.00	17 194.03	11 559.00	10 737.80
1999	12 393.00	580.14	766.00	3 251.00	57 300.00	17 419.79	12 426.00	12 109.78
2000	13 556.00	667.88	834.00	3 186.00	59 700.00	18 352.20	12 850.00	13 146.00
2001	14 808.02	787.96	914.37	3 383.01	66 103.99	20 964.12	15 163.40	16 067.61
2002	16 540.00	877.97	1 033.15	3 791.00	72 500.00	23 445.56	18 236.60	19 251.59
2003	19 105.75	945.27	1 133.56	3 881.31	86 208.11	27 702.60	22 233.60	24 108.01
2004	22 033.09	1 041.12	1 334.70	4 804.82	96 681.99	37 026.17	28 291.10	31 975.72
2005	25 002.60	1 239.98	1 421.08	5 177.86	106 884.79	40 210.24	35 324.00	37 771.14
2006	28 657.26	1 511.78	1 560.03	5 345.05	123 676.48	46 574.70	41 914.90	46 893.36
2007	32 815.53	1 759.29	1 765.00	5 824.98	136 117.25	53 918.07	48 928.80	56 560.87
2008	34 957.61	1 926.01	1 854.60	6 028.05	142 355.73	59 890.39	50 305.80	60 460.29
2009	37 146.51	1 832.37	1 944.77	6 385.01	164 397.78	58 574.07	57 218.20	69 405.40

资料来源：中经网。

我们首先用逐步回归法来对这个方程进行回归，部分运行结果见输出结果 8.3。

输出结果 8.3

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4413.777	302.213		14.605	.000
	x5	13.884	.217	.999	63.908	.000
2	(Constant)	1108.697	1067.302		1.039	.323
	x5	8.707	1.642	.626	5.301	.000
	x8	7.859	2.481	.374	3.167	.010

a. Dependent Variable: y.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.036E9	1	1.036E9	4084.265	.000[a]
	Residual	2789222.131	11	253565.648		
	Total	1.038E9	12			
2	Regression	1.037E9	2	5.185E8	3723.741	.000[b]
	Residual	1392451.660	10	139245.166		
	Total	1.038E9	12			

a. Predictors: (Constant), x5.

b. Predictors: (Constant), x5, x8.

c. Dependent Variable: y.

专业课辅导+真题资料+微信公众号：432应统书院

得到的回归方程为 $y=1\,108.697+8.707x_5+7.859x_8$，从 t 检验和 F 检验的结果来看，这个模型非常好。然而，政府工作人员对这个结果很不满意，回归方程中仅剩下 x_5（纱产量）和 x_8（纯碱产量），而这两个产业并不是需求电量最大的，最依赖于发电量的众多重工业却没能进入方程，因此，他们很难信服这个方程。

若是运用全模型，却不能满足 $n\geqslant k$ 的条件。在这种情况下，我们可以运用偏最小二乘法：

点击 Analyze→Regression→Partial Least Squares Regression 进入偏最小二乘的对话框，将 y 选入 Dependent Variables 框内，将 x_1，x_2，…，x_{13} 选入 Independent Variables 框内，Maximum number of latent factors 填 3（即设定上文算法中的 a 最大到 3，用户可自行选择合适的数，一般情况下只需迭代 3～4 次）。SPSS 的 PLS 模块是自行将数据标准化（记为 y^*，x_k^*）后处理的，运行得到的部分结果见输出结果 8.4。

输出结果 8.4

Proportion of Variance Explained

Latent Factors		Statistics				
		X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
dimension0	1	.989	.989	.998	.998	.998
	2	.005	.994	.001	.999	.999
	3	.001	.995	.000	1.000	1.000

Weights

Variables		Latent Factors		
		1	2	3
dimension0	x1	.277	-.327	-.214
	x2	.276	-.352	-.070
	x3	.277	.509	-.126
	x4	.278	-.092	-.470
	x5	.279	.110	-.126
	x6	.275	-.358	.175
	x7	.277	.432	.421
	x8	.278	.105	.583
	x9	.276	-.327	-.042
	x10	.278	.161	.374
	x11	.278	.005	-.120
	x12	.278	-.037	-.269
	x13	.278	.171	-.115
	y	.279	.121	.162

专业课辅导+真题资料+微信公众号：432应统书院

SPSS 中的 PLS 并不能直接给出回归方程的系数，而是给出全部的 t_k 和 y^* 对 t_k 的回归方程。如上表所示，我们得到的结果为：

$$
\begin{aligned}
t_1 = {} & 0.277x_1^* + 0.276x_2^* + 0.277x_3^* + 0.278x_4^* + 0.279x_5^* \\
& + 0.275x_6^* + 0.277x_7^* + 0.278x_8^* + 0.276x_9^* + 0.278x_{10}^* \\
& + 0.278x_{11}^* + 0.278x_{12}^* + 0.278x_{13}^* \\
t_2 = {} & -0.327x_1^* - 0.352x_2^* + 0.509x_3^* - 0.092x_4^* + 0.110x_5^* \\
& - 0.358x_6^* + 0.432x_7^* + 0.105x_8^* - 0.327x_9^* + 0.161x_{10}^* \\
& + 0.005x_{11}^* - 0.037x_{12}^* + 0.171x_{13}^* \\
t_3 = {} & -0.214x_1^* - 0.070x_2^* - 0.126x_3^* - 0.470x_4^* - 0.126x_5^* \\
& + 0.175x_6^* + 0.421x_7^* + 0.583x_8^* - 0.042x_9^* + 0.374x_{10}^* \\
& - 0.120x_{11}^* - 0.269x_{12}^* - 0.115x_{13}^* \\
y^* = {} & 0.279t_1 + 0.121t_2 + 0.162t_3
\end{aligned}
$$

将 t_1，t_2，t_3 代入 y^* 得

$$
\begin{aligned}
y^* = {} & 0.002\,7x_1^* + 0.022\,8x_2^* + 0.118\,3x_3^* - 0.001\,0x_4^* + 0.070\,6x_5^* \\
& + 0.061\,7x_6^* + 0.197\,8x_7^* + 0.185\,0x_8^* + 0.030\,6x_9^* + 0.157\,8x_{10}^* \\
& + 0.058\,5x_{11}^* + 0.029\,5x_{12}^* + 0.079\,8x_{13}^*
\end{aligned}
$$

将回归方程中的变量还原为原始变量：此部分如何得来的

$$
\begin{aligned}
y = {} & -2\,567.100\,7 + 4.025\,9x_1 + 0.180\,6x_2 + 4.997\,6x_3 - 0.005\,8x_4 \\
& + 0.981\,4x_5 + 0.418\,3x_6 + 3.642\,1x_7 + 3.885\,0x_8 + 0.225\,1x_9 \\
& + 0.038\,5x_{10} + 0.032\,8x_{11} + 0.016\,4x_{12} + 0.035\,5x_{13}
\end{aligned}
$$

粗略地看一下所求的回归方程，或许有人感觉到，有些自变量对因变量的影响解释不通，比如 x_4（生铁产量）前面的系数是负的。考虑到经济变量之间的关系，生铁产量与钢产量和成品钢材产量之间有部分重叠的关系，所以生铁产量对因变量的影响可能已经通过钢和成品钢材反映出来。从预测的角度来说，采用这三个自变量比只采用其中一个效果要好。

8.3　本章小结与评注

1. 主成分回归

这一章首先介绍的是主成分回归，由于主成分回归是根据主成分分析的思想提出的，而主成分分析是多元统计分析中的一个主要方法（一般来说，多元分析课程在回归分析课程之后开设），所以在本章中用较大篇幅介绍了主成分分析及其在经济问题研究中的应用，这对于没接触过主成分分析的读者是有必要的，可以使读者体会主成分回归的思想。

专业课辅导+真题资料+微信公众号：432应统书院

在介绍完主成分分析之后，介绍了主成分估计以及该估计的几个基本性质，结合经济分析实例具体介绍了主成分估计的应用。

主成分回归方程使我们看到主成分分析在简化结构、消除变量之间的相关性方面起到了明显的效果，但也给回归方程的解释带来一定的复杂性。它并没有像原解释变量的边缘效应那样简单的解释。因此，我们通常仅将主成分回归作为分析多重共线性问题的一种方法。为了得到最终的估计结果，必须把主成分还原成原始的变量。

主成分估计与前面介绍的岭估计一样是一种有偏估计，大量的实际例子和计算机模拟研究表明，在回归分析中，当设计矩阵 $\boldsymbol{X}$ 呈病态时，或者说存在多重共线性时，有偏估计在均方误差意义下改进了最小二乘估计，但是至今没有一个被公认为最优的有偏估计方法。在实际应用中，一定要根据具体问题选择合适的估计方法，不要简单认为这里介绍的几种有偏估计总会对最小二乘估计有改进作用。我们这里强调的是当设计矩阵 $\boldsymbol{X}$ 呈病态时，有偏估计会对最小二乘估计有所改进，但并不是在任何情况下都比最小二乘估计好。经过自变量多重共线性的检查，对不存在多重共线性的问题，应尽可能运用普通最小二乘法。

这里值得一提的是 1974 年韦伯斯特（J. T. Webster）、冈斯特（R. F. Gunst）和梅森（R. L. Mason）提出了特征根回归，它是主成分估计的一种推广。在主成分回归中，我们只是对自变量计算其特征根和特征向量，而在特征根回归中，把因变量 y 也考虑进来。近年来，该方法得到人们的关注，有兴趣的读者请参见参考文献[2]、[21]。

2. 偏最小二乘法

这里需要说明的是，偏最小二乘法所得的回归系数不再是因变量资料 y 的线性函数，它与普通最小二乘法不同，正是这一点引起了统计学家的兴趣。偏最小二乘法的良好的效果与非线性函数估计量的哪些统计性质有关？这一谜底至今尚未完全揭开，Frank L. E. and Friedman（1993）在这方面作了系统的评述，有兴趣的读者可以参阅。

Frank L. E. and Friedman（1993）比较了各种回归方法的应用所需的假设（见表 8.4）。

表 8.4　　各种回归方法的假设条件

普通最小二乘法、岭回归、变量选择	主成分回归、偏最小二乘法
自变量是独立的	自变量可以是相关的
自变量的值必须是精确的	自变量的值可以有误差
残差必须是随机的	残差可以有一定的结构

从上表可以看出，偏最小二乘法和主成分回归所需的假设条件较少，与实际更为接近，因而相对较优。

专业课辅导+真题资料+微信公众号：432应统书院

以上我们只简单比较了各种回归方法。Frank L. E. and Friedman（1993）详细比较了常用的几种分析方法，如普通最小二乘法、岭回归、主成分回归、偏最小二乘法和变量选择（VSS）、最佳子集回归、逐步筛选回归，从模型和选变量的准则等方面作了仔细分析，阐述了各种方法在什么情况下使用较好，并进行了数值模拟的比较。一般情况下，偏最小二乘法和岭回归是相对较好的。

思考与练习

8.1　试总结主成分回归建模的思想与步骤。

8.2　试总结偏最小二乘建模的思想与步骤。

8.3　对例 5.5 的 Hald 水泥问题用主成分回归方法建立模型，并与其他方法的结果进行比较。

8.4　对例 5.5 的 Hald 水泥问题用偏最小二乘方法建立模型，并与其他方法的结果进行比较。

专业课辅导+真题资料+微信公众号：432应统书院

第 9 章 Chapter 9 非线性回归

在许多实际问题中，变量之间的关系并不都是线性的。通常我们会碰到某些现象的被解释变量与解释变量之间呈现某种曲线关系。对于曲线形式的回归问题，显然不能照搬前面线性回归的建模方法。本章首先讨论可化为线性回归的曲线回归问题，然后讨论一种多项式回归方法，再讨论一般非线性回归模型的参数估计方法和建模过程。

9.1 可化为线性回归的曲线回归

实际问题中，有许多回归模型的被解释变量 y 与解释变量 x 之间的关系都不是线性的，其中一些回归模型通过对自变量或因变量的函数变换可以转化为线性关系，利用线性回归求解未知参数，并作回归诊断。如有下列模型

$$y=\beta_0+\beta_1 e^{bx}+\varepsilon \quad (b\text{ 已知}) \tag{9.1}$$

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_p x^p+\varepsilon \tag{9.2}$$

$$y=a e^{bx} e^{\varepsilon} \tag{9.3}$$

$$y=a e^{bx}+\varepsilon \tag{9.4}$$

对于（9.1）式，只需令 $x'=e^{bx}$ 即可化为 y 关于 x' 的线性形式

$$y=\beta_0+\beta_1 x'+\varepsilon$$

需要指出的是，新引进的自变量只能依赖于原始变量，而不能与未知参数有关。如当（9.1）式中的 b 未知时，则不能通过变量替换化为线性形式。

对于（9.2）式，可以令 $x_1=x$，$x_2=x^2$，…，$x_p=x^p$，于是得到 y 关于 x_1，x_2，…，x_p 的线性表达式

专业课辅导+真题资料+微信公众号：432应统书院

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{9.5}$$

(9.2) 式本来只有一个自变量 x，是一元 p 次多项式回归，线性化后变为 p 元线性回归。

对于 (9.3) 式，等式两边同时取自然对数，得

$$\ln y = \ln a + bx + \varepsilon$$

令 $y' = \ln y$，$\beta_0 = \ln a$，$\beta_1 = b$，于是得到 y' 关于 x 的一元线性回归模型

$$y' = \beta_0 + \beta_1 x + \varepsilon$$

对于 (9.4) 式，不能通过等式两边同时取自然对数的方法将回归模型线性化，只能用非线性最小二乘方法求解。

回归模型 (9.3) 式可以线性化，而回归模型 (9.4) 式不可以线性化，两个回归模型有相同的回归函数 ae^{bx}，只是误差项 ε 的形式不同。(9.3) 式的误差项称为乘性误差项，(9.4) 式的误差项称为加性误差项。因而一个非线性回归模型是否可以线性化，不仅与回归函数的形式有关，而且与误差项的形式有关，误差项的形式还可以有其他多种形式。

(9.3) 式与 (9.4) 式的回归参数的估计值是有差异的。误差项的形式，首先应该由数据的经济意义来确定，然后由回归拟合效果作检验。过去，由于没有非线性回归软件，人们总是希望非线性回归模型可以线性化，因而误差项的形式就假定为可以使模型线性化的形式。现在利用计算机软件可以容易地解决非线性回归问题，因而对误差项形式应该做正确的选择。

在对非线性回归模型线性化时，总是假定误差项的形式就是能够使回归模型线性化的形式，为了方便，常常省去误差项，仅写出回归函数的形式。例如把回归模型 (9.3) 式简写为 $y = ae^{bx}$。

在 SPSS 软件中给出了 10 种常见的可线性化的曲线回归方程。其中，自变量以 t 表示，见表 9.1。

表 9.1

英文名称	中文名称	方程形式
Linear	线性函数	$y = b_0 + b_1 t$
Logarithm	对数函数	$y = b_0 + b_1 \ln t$
Inverse	逆函数	$y = b_0 + b_1 / t$
Quadratic	二次曲线	$y = b_0 + b_1 t + b_2 t^2$
Cubic	三次曲线	$y = b_0 + b_1 t + b_2 t^2 + b_3 t^3$
Power	幂函数	$y = b_0 t^{b_1}$
Compound	复合函数	$y = b_0 b_1^t$
S	S 形函数	$y = \exp(b_0 + b_1 / t)$
Logistic	逻辑函数	$y = \dfrac{1}{\dfrac{1}{u} + b_0 b_1^t}$ u 是预先给定的常数

专业课辅导+真题资料+微信公众号：432应统书院

续前表

英文名称	中文名称	方程形式
Growth	增长曲线	$y=\exp(b_0+b_1t)$
Exponent	指数函数	$y=b_0\exp(b_1t)$

在以上曲线回归函数中，复合函数 $y=b_0b_1^t$，增长曲线 $y=\exp(b_0+b_1t)$，指数函数 $y=b_0\exp(b_1t)$ 这三个曲线方程实际上是等价的，只是表达形式不同，这一点请读者注意。

对以上各种曲线回归，在 SPSS 软件中点击 Regression→Curve Estimation 命令，即可方便地直接拟合各种曲线回归，而不必作任何变量变换。

除了以上 SPSS 软件中收录的几种曲线回归外，还有几种常用的曲线回归：

1. 双曲函数

$$y=\frac{x}{ax+b}$$

或等价地表示为

$$\frac{1}{y}=a+b\frac{1}{x}$$

双曲函数曲线图见图 9.1（a）。

(a) 双曲函数

(b) S形曲线

图 9.1

专业课辅导+真题资料+微信公众号：432应统书院

2. S 形曲线Ⅱ

$$y=\frac{1}{a+be^{-x}} \tag{9.6}$$

此 S 形曲线Ⅱ，当 $a>0$，$b>0$ 时，是 x 的增函数。当 $x\to+\infty$时，$y\to1/a$；当$x\to-\infty$时，$y\to0$。$y=0$ 与 $y=1/a$ 是这条曲线的两条渐进线。S 形曲线有多种，这里介绍的 S 形曲线Ⅱ是一种简单情况，其共同特点是曲线首先是缓慢增长，在达到某点后迅速增长，在超过某点后又变为缓慢增长，并且趋于一个稳定值。S 形曲线在社会经济等很多领域都有应用，例如某种产品的销售量与时间的关系，树木、农作物的生长与时间的关系等。S 形曲线Ⅱ的图形见图 9.1（b），有关 S 形曲线的进一步介绍请参见参考文献［5］。

另外，SPSS 软件中的 S 形曲线 $y=\exp(b_0+b_1/t)$ 在 $b_1<0$ 时是 t 的增函数，当 t 从右侧趋于 0 时，曲线趋于 0；当 $t\to+\infty$时，曲线以 $y=\exp(b_0)$ 为渐近线，属于通常意义上的 S 形曲线。在 $b_1>0$ 时，曲线在 t 的正实轴上是 t 的减函数，不是通常意义上的 S 形曲线。

SPSS 软件中的逻辑函数在 $0<b_1<1$ 时也是 S 形曲线。

例 9.1

对国内生产总值（GDP）的拟合。我们选取 GDP 指标为因变量，单位为亿元，拟合 GDP 关于时间 t 的趋势曲线。以 1991 年为基准年，取值为 $t=1$，2009 年 $t=19$，1991—2009 年的数据见表 9.2。

表 9.2

年份	t	y	$\hat{y}$	e	$y'=\ln y$
1991	1	21 781.5	28 514.2	−6 732.73	9.99
1992	2	26 923.5	32 745.6	−5 822.06	10.20
1993	3	35 333.9	37 604.8	−2 270.90	10.47
1994	4	48 197.9	43 185.1	5 012.79	10.78
1995	5	60 793.7	49 593.5	11 200.19	11.02
1996	6	71 176.6	56 952.9	14 223.72	11.17
1997	7	78 973	65 404.3	13 568.67	11.28
1998	8	84 402.3	75 109.9	9 292.37	11.34
1999	9	89 677.1	86 255.8	3 421.33	11.40
2000	10	99 214.6	99 055.6	159.02	11.51
2001	11	109 655	113 755	−4 099.61	11.61
2002	12	120 333	130 635	−10 302.61	11.70
2003	13	135 823	150 021	−14 197.98	11.82
2004	14	159 878	172 283	−12 404.62	11.98
2005	15	184 937	197 849	−12 911.23	12.13
2006	16	216 314	227 208	−10 893.72	12.28
2007	17	265 810	260 924	4 885.91	12.49
2008	18	314 045	299 644	14 401.47	12.66
2009	19	340 507	344 109	−3 602.31	12.74

专业课辅导+真题资料+微信公众号：432应统书院

方法一：直接用 SPSS 软件进行计算，首先画出 GDP 对变量 t 的散点图，见图 9.2。

图 9.2 GDP 时间趋势图

从散点图中看到，GDP 大致为指数函数形式，从经济学角度看，当 GDP 的年增长速度大致相同时，其趋势线就是指数函数形式。前面说过，复合函数 $y=b_0b_1^t$，增长曲线 $y=\exp(b_0+b_1t)$，指数函数 $y=b_0\exp(b_1t)$ 这三个曲线方程实际上是等价的。在本例中，复合函数 $y=b_0b_1^t$ 的形式与经济意义更吻合。自变量为时间变量时，点击 Analyze→Curve Estimation，默认模型为 Linear（简单线性回归），我们再点击 Compound（复合函数回归），计算结果如输出结果 9.1 所示。

输出结果 9.1

Linear

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1.429E11	1	1.429E11	127.091	.000
Residual	1.912E10	17	1.124E9		
Total	1.620E11	18			

The independent variable is T.

专业课辅导+真题资料+微信公众号：432应统书院

Compound

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	10.912	1	10.912	614.628	.000
Residual	.302	17	.018		
Total	11.214	18			

The independent variable is T.

Model Summary and Parameter Estimates

Dependent Variable:Y

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.882	127.091	1	17	.000	-28668.749	15834.125
Compound	.973	614.628	1	17	.000	24829.662	1.148

The independent variable is T.

线性回归 SSE＝1.912E10，R^2＝0.882。复合函数回归 SSE＝0.302，R^2＝0.973 是按线性化后的回归模型计算的，两者的残差不能直接相比。为了与线性回归的拟合效果直接相比，可以先存储复合函数的残差序列（见表 9.2），然后计算出复合函数回归的 SSE＝1.727E9，可推知复合函数拟合效果明显优于线性回归，从模型拟合图中，也可直观得到这一结论，故在解决此类问题时应采用复合函数回归。

专业课辅导+真题资料+微信公众号：432应统书院

复合函数回归系数为 $b_0=24\,829.662$，等比系数 $b_1=1.148$，回归方程为

$$\hat{y}=24\,829.662\times(1.148)^t$$

其中，$b_1=1.148=114.8\%$表示 GDP 的平均发展速度，平均增长速度为 14.8%。这里 GDP 用的是当年现价，包含物价上涨因素在内。本例只是作为计算非线性回归的示例。在实际工作中，如果需要对 GDP 作趋势拟合或预测，应对此模型作一些改进，例如用不变价格代替现价，对误差项的自相关作相应的处理；考虑到 GDP 的年增长速度会有减缓趋势，可以给回归函数增加适当的阻尼因子，或采用 S 形曲线拟合等改进方法。

方法二：线性化求解法。前面方法一介绍的直接用 SPSS 软件的 Curve Estimation 命令中的复合函数回归的计算过程实际上是按以下的线性化求解法实施的。复合函数 $y=b_0b_1^t$ 两端取自然对数，得

$$\ln y=\ln b_0+t\ln b_1$$

令 $y'=\ln y$，$\beta_0=\ln b_0$，$\beta_1=\ln b_1$，于是得到 y'关于 t 的线性回归方程

$$y'=\beta_0+\beta_1 t$$

计算出 $y'=\ln y$ 的值列在表 9.2 中，用 y'对 t 作一元线性回归，得输出结果 9.2。

输出结果 9.2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.986[a]	.973	.972	.13325	.300

a. Predictors: (Constant), T.
b. Dependent Variable: lny.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.924	1	10.924	615.238	.000[a]
	Residual	.302	17	.018		
	Total	11.226	18			

a. Predictors: (Constant), T.
b. Dependent Variable: lny.

专业课辅导+真题资料+微信公众号：432应统书院

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.119	.064		159.017	.000
	T	.138	.006	.986	24.804	.000

a. Dependent Variable: lny.

其中，$\hat{\beta}_0=10.119$，$\hat{\beta}_1=0.138$，得 $b_0=e^{10.119}=24\ 809.948$；$b_1=e^{0.138}=1.148$，这与前述方法中直接用 SPSS 软件的 Curve Estimation 命令计算的结果大体一致。

9.2　多项式回归

多项式回归模型是一种重要的曲线回归模型，这种模型通常容易转化成一般的多元线性回归来作处理，因而它的应用也十分广泛。

一、几种常见的多项式回归模型

回归模型

$$y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\varepsilon_i$$

称为一元二阶（或一元二次）多项式模型。其中 $i=1, 2, \cdots, n$，在以下的回归模型中不再一一注明。

为了反映回归系数所对应的自变量次数的类型，我们通常将多项式回归模型中的系数表示成下面模型中的情形

$$y_i=\beta_0+\beta_1 x_i+\beta_{11} x_i^2+\varepsilon_i \tag{9.7}$$

模型（9.7）式的回归函数 $y_i=\beta_0+\beta_1 x_i+\beta_{11} x_i^2$ 是一条抛物线方程，通常称为二项式回归函数。回归系数 β_1 为线性效应系数，β_{11} 为二次效应系数。

相应地，回归模型

$$y_i=\beta_0+\beta_1 x_i+\beta_{11} x_i^2+\beta_{111} x_i^3+\varepsilon_i$$

称为一元三次多项式模型。

当自变量的幂次超过 3 时，回归系数的解释变得困难起来，回归函数也变得很不稳定，回归模型的应用会受到影响。因而，幂次超过 3 的多项式回归模型不常使用。

以上两个多项式回归模型都是只含有一个自变量 x，在实际应用中，我们常遇到含有两个或两个以上自变量的情况。称回归模型

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_{11} x_{i1}^2+\beta_{22} x_{i2}^2+\beta_{12} x_{i1} x_{i2}+\varepsilon_i$$

专业课辅导+真题资料+微信公众号：432应统书院

为二元二阶多项式回归模型。它的回归系数中分别含有两个自变量的线性项系数 β_1 和 β_2，二次项系数 β_{11} 和 β_{22}，并含有交叉乘积项系数 β_{12}。交叉乘积项表示 x_1 与 x_2 的交互作用，系数 β_{12} 通常称为交互影响系数。

类似上面的情况，我们还可给出多元高阶多项式回归模型，有兴趣的读者请参见参考文献［3］。

二、一个应用例子

下面利用参考文献［3］的一个例子来说明二元多项式回归的应用。

例 9.2

表 9.3 列出的数据是关于 18 个 35～44 岁经理的前两年年平均收入 x_1（千美元）、风险反感度 x_2 和人寿保险额 y（千美元）。风险反感度是根据发给每个经理的标准调查表估算得到的，它的数值越大，风险反感度就越高。

表 9.3

序号	x_{i1}	x_{i2}	y_i
1	66.290	7	196
2	40.964	5	63
3	72.996	10	252
4	45.010	6	84
5	57.204	4	126
6	26.852	5	14
7	38.122	4	49
8	35.840	6	49
9	75.796	9	266
10	37.408	5	49
11	54.376	2	105
12	46.186	7	98
13	46.130	4	77
14	30.366	3	14
15	39.060	5	56
16	79.380	1	245
17	52.766	8	133
18	55.916	6	133

研究人员想研究给定年龄组内的经理年平均收入、风险反感度和人寿保险额之间的关系。研究者预计，在经理的收入和人寿保险额之间存在二次关系，并有把握地认为风险反感度对人寿保险额只有线性效应，而没有二次效应。但是，研究者对两个自变量是否对人寿保险额有交互效应，心中没底。因此，研究者拟合了一个二阶多项式回归模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_{11}x_{i1}^2+\beta_{22}x_{i2}^2+\beta_{12}x_{i1}x_{i2}+\varepsilon_i$$

并打算先检验是否有交互效应，然后检验风险反感度的二次效应。

专业课辅导+真题资料+微信公众号：432应统书院

回归采用逐个引入自变量的方式，这样可以清楚地看到各项对回归的贡献，使显著性检验更加明确。依次引入自变量 x_1，x_2，x_1^2，x_2^2，x_1x_2，方法如下：在线性回归对话框中，选入 y 与 x_1，然后点 Block 1 of Next，这时自变量框变为空白，再把 x_1，x_2 同时选入自变量框中，然后再点 Block 2 of Next，自变量框又变为空白，再把 x_1，x_2，x_1^2 同时选入自变量框中，如此依次引入自变量。取显著性水平 $\alpha=0.05$，方差分析表见表 9.4。

表 9.4

变量	偏平方和	残差平方和	检验系数	偏 F 值
x_1	104 474	3 567	β_1	—
$x_2 \mid x_1$	2 284	1 283	β_2	—
$x_1^2 \mid x_1, x_2$	1 238	45	β_{11}	1 238/(45/14)=385
$x_2^2 \mid x_1, x_2, x_1^2$	3	42	β_{22}	3/(42/13)=0.93
$x_1x_2 \mid x_1, x_2, x_1^2, x_2^2$	6	36	β_{12}	6/(36/12)=2.00
合计	108 005			

其中，$x_2 \mid x_1$ 表示模型中已含有 x_1 再加入 x_2 的回归模型，其余依此类推。全模型的 SST=108 041，SSE=36，SSE 的自由度 $df=n-p-1=18-5-1=12$。采用（3.42）式的偏 F 检验，对交互影响系数 β_{12} 的显著性检验的偏 F 值=2.00，临界值 $F_{0.05}(1, 12)=4.75$，交互影响系数 β_{12} 不能通过显著性检验，认为 $\beta_{12}=0$，回归模型中不应该包含交互作用项 x_1x_2。这个结果与人们的经验相符合，有了此结果，两个自变量的效应也就容易解释了。此时，研究者暂时决定使用无交互效应的模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_{11}x_{i1}^2+\beta_{22}x_{i2}^2+\epsilon_i$$

但仍想检验风险反感度的二次效应是否存在。这相当于检验二次效应系数 β_{22} 的显著性，这个检验的偏 F 值=0.93，临界值 $F_{0.05}(1, 13)=4.67$，二次效应系数 β_{22} 不能通过显著性检验，认为 $\beta_{22}=0$，回归模型中不应该包含二次效应项 x_2^2。此时，研究者决定使用简化的回归模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_{11}x_{i1}^2+\epsilon_i$$

进一步检验年平均收入的二次效应是否存在，这相当于检验二次效应系数 β_{11} 的显著性，这个检验的偏 F 值=385，临界值 $F_{0.05}(1, 14)=4.60$，二次效应系数 β_{22} 通过了显著性检验，认为 $\beta_{22}\neq0$，回归模型中应该包含二次效应项 x_1^2。得最终的回归方程为

$$\hat{y}=-62.349+\underset{(0.164)}{0.840x_1}+\underset{(0.164)}{5.685x_2}+\underset{(0.785)}{0.037\,1x_1^2}$$

其中，括号中的数值是标准化回归系数。这样，研究者可用这个回归方程来进一步研究经理的年平均收入和风险反感度对人寿保险额的效应。从标准化回归系数看到，年平均收入的二次效应对人寿保险额的影响程度最大。

由这个例子我们可看到利用回归方程分析问题的一些思路，如回归系数的假设

专业课辅导+真题资料+微信公众号：432应统书院

检验、交互效应、二次效应等的实际意义。相信这个例子会对读者扩展回归分析的应用有所启发。

多项式回归常用于分析试验设计的数据。在试验设计中，目标变量 y 与试验因子间的函数关系往往是未知的，因而常用多项式回归近似 y 与试验因子的关系。在用正交设计与均匀设计等方法的试验中，试验次数一般很少，而多项式回归的参数个数较多。例如三元二次多项式回归在考虑全部交互作用时，共有 10 个未知参数。这样样本量 n 常少于未知参数的数目，这与最小二乘法的要求 $n>p+1$ 相违背。不过，我们可以方便地解决这个问题，只需采用逐步回归方法。请看下面的例题。

例 9.3

维生素 C 注射液因长期放置会渐变成微黄色，中国药典规定可以用焦亚硫酸钠等作为抗氧剂。本实验考虑 3 个因素，分别是 EDTA(x_1)、无水碳酸钠（x_2）、焦亚硫酸钠（x_3），每个因素各取 7 个水平，选用 $U_7(7^4)$ 均匀设计表，实验响应变量是吸收度 y，取值越小越好，试验设计与结果见表 9.5。

表 9.5　试验设计与结果

实验号	EDTA x_1(g)	无水碳酸钠 x_2(g)	焦亚硫酸钠 x_3(g)	吸收度 y	$1/y$
1	0.00	30	0.6	1.160	0.862
2	0.02	38	1.2	0.312	3.205
3	0.04	46	0.4	0.306	3.263
4	0.06	26	1.0	1.318	0.759
5	0.08	34	0.2	0.877	1.140
6	0.10	42	0.8	0.147	6.803
7	0.12	50	1.4	0.204	4.902

直接看的好条件是第 6 号实验的条件，EDTA（x_1）取 0.10g，无水碳酸钠（x_2）取 42g，焦亚硫酸钠（x_3）取 0.8g。以下用回归分析方法进一步寻找最优条件。

首先做线性回归，得回归方程

$$\hat{y}=2.63+0.77x_1-0.052\,4x_2-0.087x_3$$

回归方程 F 检验的 P 值=0.104；决定系数 $R^2=83.9\%$；调整的决定系数 $R_a^2=67.8\%$。可见线性回归的效果不够好，以下使用二次多项式回归，回归方程的具体形式是：

$$\hat{y}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_{11}x_1^2+\beta_{22}x_2^2+\beta_{33}x_3^2+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\beta_{23}x_2x_3+\varepsilon$$

对此回归方程按下面的方式做变量替换：

专业课辅导+真题资料+微信公众号：432应统书院

$$x_{11}=x_1^2,\ x_{22}=x_2^2,\ x_{33}=x_3^2,\ x_{12}=x_1x_2,\ x_{13}=x_1x_3,\ x_{23}=x_2x_3$$

转化为 y 对 9 个自变量的线性回归，变换结果见表 9.6。

表 9.6　　回归变量表

x_1	x_2	x_3	x_{11}	x_{22}	x_{33}	x_{12}	x_{13}	x_{23}	y
0.00	30	0.6	0.000 0	900	0.360	0.00	0.000	18.0	1.160
0.02	38	1.2	0.000 4	1 444	1.440	0.76	0.024	45.6	0.312
0.04	46	0.4	0.001 6	2 116	0.160	1.84	0.016	18.4	0.306
0.06	26	1.0	0.003 6	676	1.000	1.56	0.060	26.0	1.318
0.08	34	0.2	0.006 4	1 156	0.040	2.72	0.016	6.8	0.877
0.10	42	0.8	0.010 0	1 764	0.640	4.20	0.080	33.6	0.147
0.12	50	1.4	0.014 4	2 500	1.960	6.00	0.168	70.0	0.204

这个线性回归只有 7 组观测数据却有 10 个未知参数，需要使用逐步回归逐个引入变量。在 SPSS 软件的逐步回归模块默认的进入变量 P 值=0.05，剔除变量 P 值=0.10 的条件下，逐步回归只进行一步就结束了，只选入了自变量 x_2。为了更全面地了解回归的效果，可以把进入变量的条件放宽一些。用 Option 选项把进入变量 P 值改为 0.30，剔除变量 P 值改为 0.50，重新做逐步回归。输出结果 9.3 列出了逐步回归每步的拟合优度模型概括和每步的回归系数。

输出结果 9.3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.912	.831	.798	.219 96
2	.960	.921	.882	.168 11
3	.985	.971	.942	.117 49
4	.994	.987	.962	.095 40
5	1.000	1.000	.999	.012 24
6	1.000	1.000	.	.

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	2.579	.404		6.390	.001
x2	-.052	.010	-.912	-4.966	.004
2 (Constant)	5.957	1.612		3.696	.021
x2	-.238	.087	-4.199	-2.716	.053
x22	.002	.001	3.301	2.135	.100
3 (Constant)	7.311	1.274		5.740	.011
x2	-.303	.068	-5.360	-4.487	.021
x22	.003	.001	4.531	3.751	.033
x3	-.292	.128	-.258	-2.278	.107

专业课辅导+真题资料+微信公众号：432应统书院

4 (Constant)	7.873	1.092		7.207	.019
x2	-.313	.055	-5.523	-5.663	.030
x22	.003	.001	4.358	4.416	.048
x3	-1.115	0.526	-0.985	-2.121	.168
x23	.021	.013	.889	1.597	.251
5 (Constant)	9.165	.183		50.097	.013
x2	-0.379	.009	-6.694	-40.733	.016
x22	.004	.000	5.475	33.717	.019
x3	-1.430	.073	-1.263	-19.518	.033
x23	.032	.002	1.372	16.362	.039
x13	-2.332	.212	-.278	-10.982	.058
6 (Constant)	9.111	.000		.	.
x2	-.376	.000	-6.641	.	.
x22	.004	.000	5.439	.	.
x3	-1.444	.000	-1.276	.	.
x23	.030	.000	1.300	.	.
x13	-2.291	.000	-.273	.	.
x33	.051	.000	.074	.	.

从输出结果中看到，逐步回归共进行了 6 步，依次选入了 x_2，$x_{22}=x_2^2$，x_3，$x_{23}=x_2x_3$，$x_{13}=x_1x_3$，$x_{33}=x_3^2$ 共 6 个变量，没有变量被剔除。

首先看引入变量的顺序，第 1 个回归模型最先选入的是 x_2，说明无水碳酸钠的含量是最重要的影响因素；第 2 个回归模型再选入的是 $x_{22}=x_2^2$，进一步说明无水碳酸钠的含量是最重要的影响因素，并且说明 y 与 x_2 的关系是非线性的，第二个回归方程是：

$$\hat{y}=5.957-0.237\,5x_2+0.002\,45x_2^2$$

容易求出此方程在 $x_2=0.237\,5/(2\times0.002\,45)\approx48$ 时达极小值 $\hat{y}=0.197$，比第 6 号实验值 $y=0.147$ 略高。

再看第 3 个回归方程：

$$\hat{y}=7.311-0.303x_2+0.003\,36x_2^2-0.29x_3$$

为使 y 值最小，x_3 应该最大，取 $x_3=1.4$，x_2 的取值与 x_3 无关，容易求出此方程在 $x_2=45.1\approx45$，$x_3=1.4$ 时达极小值 $\hat{y}=0.074$，低于第 6 号实验值 $y=0.147$。

第 4 个回归方程是：

$$\hat{y}=7.873-0.312\,6x_2+0.003\,23x_2^2-1.115x_3+0.020\,6x_2x_3$$

在回归方程含有 x_3 的两项 $-1.115x_3+0.020\,6x_2x_3$ 中，当 $x_2\leqslant54$ 时，y 是 x_3 的减函数，极值在 x_3 的边界达到。根据对第 2 和第 3 两个回归方程的分析，两个方程中 x_2 的最优解分别是 48 和 45，所以有理由认为 $x_2\leqslant54$，y 是 x_3 的减函数，x_3 越

专业课辅导+真题资料+微信公众号：432应统书院

大 y 越小，因此取 $x_3=1.4$。

把 $x_3=1.4$ 带入方程，解得 x_2 的极小值点是 $x_2=43.9\approx44$，所以第 4 个回归方程的最优组合是 $x_2=44$，$x_3=1.4$，此时最优预测值 $\hat{y}=0.080$，与第 3 个回归方程的最优解基本相同。

第 5 个方程：

$$\hat{y}=9.16-0.379x_2+0.004\,06x_2^2-1.43x_3+0.031\,7x_2x_3-2.33x_1x_3$$

其中包含了变量 x_1，并且是作为与 x_3 的交互作用形式出现的，说明 EDTA 对实验指标本身没有影响，只是通过焦亚硫酸钠对实验产生弱的影响。仿照对第 4 个回归方程求最优解的方法，首先确定 x_1 和 x_3 是 y 的减函数，分别取最大值 $x_1=0.12$ 和 $x_3=1.4$，然后解得 $x_2=41.2\approx41$。最优预测值 $\hat{y}=-0.128<0$，可以视为接近 0。

第 6 个方程引入的自变量数目太多，不予考虑。

比较第 3，4，5 这三个回归方程，回归方程的决定系数分别是：97.11%，98.73%，99.99%，从回归的效果看，第 5 个回归的效果最好，但是有 6 个估计参数，而 y 的数据只有 7 个，所以估计的误差会较大。第 3，4 两个回归方程的实验条件基本相同，预测值也很接近，约为 0.080，明显小于第 6 号实验的吸收度 $y=0.147$，是一组稳定的好条件，见表 9.7。

表 9.7　　吸收度的最优实验条件

回归模型	最优搭配			最优预测值
	x_1(g)	x_2(g)	x_3(g)	
2	0.00	48	0.0	0.197
3	0.00	45	1.4	0.074
4	0.00	44	1.4	0.080
5	0.12	41	1.4	0.000

本例的参考文献［13］对吸收度 y 值先取倒数作为实验指标，其数值越大越好，然后建立回归方程。这样做的一个好处是避免了本例第 5 个回归方程预测值为负值的情况，但是回归方程的效果不好。文献中得到的最优条件是 $x_1=0.12$，$x_2=38$，$x_3=1.4$，和本例第 5 个方程相差不大。

9.3　非线性模型

一、非线性最小二乘

非线性回归模型一般可记为

$$y_i=f(\boldsymbol{x}_i,\boldsymbol{\theta})+\varepsilon_i,\quad i=1,2,\cdots,n \tag{9.8}$$

专业课辅导+真题资料+微信公众号：432应统书院

其中，y_i 为因变量；非随机向量 $\boldsymbol{x}_i=(x_{i1}, x_{i2}, \cdots, x_{ik})'$ 是自变量；$\boldsymbol{\theta}=(\theta_0, \theta_1, \cdots, \theta_p)'$ 为未知参数向量；ε_i 为随机误差项并且满足独立同分布假定，即

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, & i=j \\ 0, & i\neq j\end{cases} \end{cases} \qquad i,j=1,2,\cdots,n$$

如果 $f(\boldsymbol{x}_i, \boldsymbol{\theta})=\theta_0+x_1\theta_1+x_2\theta_2+\cdots+x_p\theta_p$，那么（9.8）式就是前面讨论的线性模型，而且必然有 $k=p$；对于一般情况的非线性模型，参数的数目与自变量的数目并没有一定的对应关系，不要求 $k=p$。

对非线性回归模型（9.8）式，仍使用最小二乘法估计参数 $\boldsymbol{\theta}$，即求使

$$Q(\boldsymbol{\theta})=\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i,\boldsymbol{\theta}))^2 \tag{9.9}$$

达到最小的 $\hat{\boldsymbol{\theta}}$，称 $\boldsymbol{\theta}$ 为非线性最小二乘估计。在假定 f 函数对参数 $\boldsymbol{\theta}$ 连续可微时，可以利用微分法建立正规方程组，求使 $Q(\boldsymbol{\theta})$ 达到最小的 $\hat{\boldsymbol{\theta}}$。将 Q 函数对参数 $\boldsymbol{\theta}_j$ 求偏导，并令其为 0，得 $p+1$ 个方程

$$\left.\frac{\partial Q}{\partial \theta_j}\right|_{\theta_j=\hat{\theta}_j}=-2\sum_{i=1}^{n}(\boldsymbol{y}_i-f(\boldsymbol{x}_i,\hat{\theta}))\left.\frac{\partial f}{\partial \theta_j}\right|_{\theta_j=\hat{\theta}_j}=0 \tag{9.10}$$
$$j=0,1,2,\cdots,p$$

非线性最小二乘估计 $\hat{\boldsymbol{\theta}}$ 就是（9.10）式的解，（9.10）式称为非线性最小二乘估计的正规方程组，它是未知参数的非线性方程组。一般用 Newton 迭代法求解此正规方程组。也可以直接极小化残差平方和 $Q(\boldsymbol{\theta})$，求出未知参数 $\boldsymbol{\theta}$ 的非线性最小二乘估计值 $\hat{\boldsymbol{\theta}}$。

在实际应用中，SPSS 软件可以直接求出未知参数 $\boldsymbol{\theta}$ 的非线性最小二乘估计值 $\hat{\boldsymbol{\theta}}$。

对于非线性最小二乘估计，我们仍然需要作参数的区间估计、显著性检验、回归方程的显著性检验等回归诊断，这需要知道有关统计量的分布。在非线性最小二乘中，一些精确分布是很难得到的，在大样本时，可以得到近似的分布。计算机软件在求出参数 $\boldsymbol{\theta}$ 的非线性最小二乘估计值 $\hat{\boldsymbol{\theta}}$ 的同时，还给出近似的参数的区间估计、显著性检验、回归方程的显著性检验等回归诊断。

在非线性回归中，平方和分解式 SST=SSR+SSE 不再成立。类似于线性回归中的复决定系数，定义非线性回归的相关指数

$$R^2=1-\frac{\mathrm{SSE}}{\mathrm{SST}} \tag{9.11}$$

专业课辅导+真题资料+微信公众号：432应统书院

二、非线性回归模型的应用

例 9.4

一位药物学家使用下面的非线性模型对药物反应拟合回归模型

$$y_i=c_0-\frac{c_0}{1+\left(\frac{x_i}{c_2}\right)^{c_1}}+\varepsilon_i \tag{9.12}$$

其中，自变量 x 为药剂量，用级别表示；因变量 y 为药物反应程度，用百分数表示。3 个参数 c_0，c_1，c_2 都是非负的，根据专业知识，c_0 的上限是 100%，3 个参数的初始值取为 $c_0=100$，$c_1=5$，$c_2=4.8$。测得 9 个反应数据如表 9.8 所示。

表 9.8　反应数据

x	1	2	3	4	5	6	7	8	9
y (%)	0.5	2.3	3.4	24.0	54.7	82.1	94.8	96.2	96.4

请拟合 (9.12) 式的回归方程。

这是一个一元非线性回归，首先用 SPSS 软件画出散点图，见图 9.3。

图 9.3　药物反应程度散点图

从图 9.3 看到，y 与 x 之间确实呈非线性关系。在 SPSS 的 Regression 菜单下点击 Nonlinear，进入非线性回归对话框，将 y 选入因变量框，在 Model Expression 框中输入回归函数 c0−c0/(1+(x/c2)**c1)，然后点击 Parameters 进入参数设置框给未知参数赋初值。

这里有两点需要注意的是：第一，输入回归函数时，一定不要在回归函数前加上“y=”，如果在 Model Expression 框中输入 y=c0−c0/(1+(x/c2)**c1)，那么软件将不进行迭代运算而直接停止；第二，未知参数的符号可以是随意的，本例程序中以 c0，c1，c2 代替加下标的参数 c_0，c_1，c_2。

专业课辅导+真题资料+微信公众号：432应统书院

经过 6 步迭代后收敛，相关指数 $R^2=0.998\ 65$，说明非线性回归拟合效果很好。总平方和 SST=14 917（即输出结果 9.4 中的 Corrected Total），这个 SST 也就是 y_1，y_2，…，y_n 的离差平方和。

输出结果 9.4 中的回归平方和 37 839，是 n 个回归值 $\hat{y}_1$，$\hat{y}_2$，…，$\hat{y}_n$的平方和 $\hat{y}_1^2+\hat{y}_2^2+\cdots+\hat{y}_n^2$，不是我们在线性回归中用的离差平方和。Uncorrected Total 的平方和 37 860 是 $y_1^2+y_2^2+\cdots+y_n^2$，这里 Uncorrected（未修正）是指数据直接平方，没有减去均值。Corrected Total 的平方和是 14 917，它是 y_1，y_2，…，y_n 的离差平方和 $\mathrm{SST}=\sum_{i=1}^{n}(y_i-\bar{y})^2=14\ 917$，这里 Corrected（修正）是指数据减去均值的修正。表 9.9 中列出了有关的平方和及离差平方和的数值，以供读者对比。

输出结果 9.4

Iteration	Residual SS	C0	C1	C2
1	172.7877170	100.000000	5.00000000	4.80000000
1.1	32.60704344	97.7943996	6.57938197	4.74460195
2	32.60704344	97.7943996	6.57938197	4.74460195
2.1	20.20240372	99.5785656	6.73691756	4.80074972
3	20.20240372	99.5785656	6.73691756	4.80074972
3.1	20.18814307	99.5334852	6.76307026	4.79941696
4	20.18814307	99.5334852	6.76307026	4.79941696
4.1	20.18803580	99.5411768	6.76104089	4.79966204
5	20.18803580	99.5411768	6.76104089	4.79966204
5.1	20.18803473	99.5404448	6.76127044	4.79964160
6	20.18803473	99.5404448	6.76127044	4.79964160
6.1	20.18803472	99.5405197	6.76124802	4.79964382

Nonlinear Regression Summary Statistics

Source	DF	Sum of Squares	Mean Square
Regression	3	37839.85197	12613.28399
Residual	6	20.18803	3.36467
Uncorrected Total	9	37860.04000	
(Corrected Total)	8	14917.88889	

R squared = 1 − Residual SS /Corrected SS = .99865

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
C0	99.540519687	1.567325937	95.705411276	103.37562810
C1	6.761248019	.421980049	5.728700036	7.793796003
C2	4.799643816	.050165521	4.676893208	4.922394423

表 9.9

序号	x	y	$\hat{y}$	e	$\hat{y}-\bar{y}$
1	1	0.5	0	0.5	−50.488 89
2	2	2.3	0.27	2.03	−50.218 89
3	3	3.4	3.98	−0.58	−46.508 89
4	4	24	22.48	1.52	−28.008 89
5	5	54.7	56.61	−1.91	6.121 11

专业课辅导+真题资料+微信公众号：432应统书院

续前表

序号	x	y	$\hat{y}$	e	$\hat{y}-\bar{y}$
6	6	82.1	81.52	0.58	31.031 11
7	7	94.8	92.34	2.46	41.851 11
8	8	96.2	96.49	−0.29	46.001 11
9	9	96.4	98.14	−1.74	47.651 11
均值	5	50.488 89	50.203 33	0.285 556	−0.285 56
离差平方和	60	14 917.89	15 156.55	19.431 62	15 156.55
平方和	285	37 860.04	37 839.85	20.188 03	15 157.28

本例回归离差平方和 SSR＝15 156.55，而总离差平方和 SST＝14 917.89＜SSR，可见非线性回归不再满足平方和分解式，即 SST≠SSR＋SSE。另外，非线性回归的残差和不等于零，本例残差均值为 0.285 556≠0。当然，如果回归拟合的效果好，残差的均值会接近零。

通过以上分析可以认为，药物反应程度 y 与药剂量 x 符合下面的非线性回归方程

$$\hat{y}=99.541-\frac{99.541}{1+\left(\frac{x}{4.799\,6}\right)^{6.761\,2}}$$

例 9.5

龚珀兹（Gompertz）模型是计量经济中的一个常用模型，用来拟合社会经济现象的发展趋势，龚珀兹曲线形式为：

$$y_t=k\cdot a^{b^t} \tag{9.13}$$

其中，k 为变量的增长上限，$0<a<1$ 和 $0<b<1$ 是未知参数。当 k 未知时，龚珀兹模型不能线性化，可以用非线性最小二乘法求解。表 9.10 是我国民航国内航线里程数据，以下用龚珀兹模型拟合这个数据。

表 9.10　我国民航国内航线里程数据　单位：万公里

年份	t	y	年份	t	y
1980	1	11.41	1993	14	68.21
1981	2	13.55	1994	15	69.37
1982	3	13.28	1995	16	78.08
1983	4	12.92	1996	17	78.02
1984	5	15.28	1997	18	92.06
1985	6	17.12	1998	19	100.14
1986	7	21.67	1999	20	99.89
1987	8	24.02	2000	21	99.45
1988	9	24.55	2001	22	103.67
1989	10	30.55	2002	23	106.32
1990	11	34.04	2003	24	103.42
1991	12	38.17	2004	25	115.52
1992	13	53.36			

专业课辅导+真题资料+微信公众号：432应统书院

用 SPSS 软件的非线性最小二乘法功能求解，依照 Analyze→Regression→Nonlinear 的顺序进入非线性回归对话框。在回归函数框条中输入 k * a** (b**t)。一个主要的工作是给出参数的初值，初值并不要求很准确，所以很多时候可以凭经验给定。龚珀兹中的参数 k 是变量的发展上限，应该取其初值略大于最大观测值。本例最大观测值是 115.52，不妨取 k 的初值为 120。a 和 b 都是 0～1 之间的数，可以取其初值为 0.5。经过 31 步迭代后收敛，见输出结果 9.5。

输出结果 9.5

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	.01243	.006	.000	.025
b	.8927	.015	.862	.923
k	150.0	15.814	117.162	182.756

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	114 521.478	3	381 73.826
Residual	819.818	22	37.264
Uncorrected Total	115 341.296	25	
Corrected Total	342 22.087	24	

Dependent variable: y.

a. R squared = 1 − (Residual Sum of Squares) /(Corrected Sum of Squares) = 0.976.

用非线性最小二乘法求得的三个参数估计值为 $k=150.0$，$a=0.012\,43$，$b=0.892\,7$，其中 $k=150.0$ 为回归模型估计的国内航线里程增长上限。图 9.4 是用 Excel 绘制的国内航线里程趋势预测图，其中粗实线是观测值，细虚线是预测值。

图 9.4 龚珀兹曲线拟合国内航线里程趋势图

专业课辅导+真题资料+微信公众号：432应统书院

上面的过程中回归迭代给出了一些警告错误，这是由于回归迭代过程中的参数取值超出了允许范围，可以通过对参数取值范围增加一些限制来解决，点击非线性回归对话框中的 Constraints 按钮即可实现。另外，如果参数的初值给得不够准确，回归迭代不收敛，对参数增加一些限制可能就收敛了。例如本例中把 k 的初值改为 130，回归迭代不能收敛，但是增加一个 $k>116$ 的限制，回归迭代就收敛了。

龚珀兹模型和几种常见的非线性回归模型可以用三和值法求解，见参考文献[15] 第 13 章。在正态误差假定下，非线性回归的最小二乘估计与最大似然估计是相同的，而最大似然估计具有良好的大样本性质，例如渐近无偏性、渐近正态性、一致性等。因而非线性最小二乘估计值比三和值更精确，可以把三和值法的参数估计值作为求解非线性最小二乘的初值。

例 9.6

表 9.11 是我国 1950—2005 年历年大陆总人口数，试用威布尔（Weibull）曲线拟合数据并做预测。威布尔曲线如下：

$$y=k-ab^t \tag{9.14}$$

其中，参数 k 是变量发展的上限；参数 $a>0$，$0<b<1$，$c>0$。

表 9.11　　**我国历年大陆总人口数**　　单位：亿人

年份	t	y	年份	t	y
1950	1	5.519 6	1966	17	7.454 2
1951	2	5.63	1967	18	7.636 8
1952	3	5.748 2	1968	19	7.853 4
1953	4	5.879 6	1969	20	8.067 1
1954	5	6.026 6	1970	21	8.299 2
1955	6	6.146 5	1971	22	8.522 9
1956	7	6.282 8	1972	23	8.717 7
1957	8	6.465 3	1973	24	8.921 1
1958	9	6.599 4	1974	25	9.085 9
1959	10	6.720 7	1975	26	9.242
1960	11	6.620 7	1976	27	9.371 7
1961	12	6.585 9	1977	28	9.497 4
1962	13	6.729 5	1978	29	9.625 9
1963	14	6.917 2	1979	30	9.754 2
1964	15	7.049 9	1980	31	9.870 5
1965	16	7.253 8	1981	32	10.007 2

专业课辅导+真题资料+微信公众号：432应统书院

续前表

年份	t	y	年份	t	y
1982	33	10.154 1	1994	45	11.985
1983	34	10.249 5	1995	46	12.112 1
1984	35	10.347 5	1996	47	12.238 9
1985	36	10.453 2	1997	48	12.362 6
1986	37	10.572 1	1998	49	12.476 1
1987	38	10.724	1999	50	12.578 6
1988	39	10.897 8	2000	51	12.674 3
1989	40	11.270 4	2001	52	12.762 7
1990	41	11.433 3	2002	53	12.845 3
1991	42	11.582 3	2003	54	12.922 7
1992	43	11.717 1	2004	55	12.998 8
1993	44	11.851 7	2005	56	13.075 6

根据人口学的专业预测，我国人口上限为 16 亿人，因此取 k 的初值=16，取 b 的初值=0.5，取 c 的初值=1。对以上初值把 $t=1$ 时（即 1950 年）$y_1=5.519\,6$ 代入，得 $a=2(k-y_1)\approx 2(16-5.5)=21$，用以上初值做非线性最小二乘，得下面的输出结果 9.6。从中看到，人口上限为 $k=15.76$ 亿人，这与人口学预测的人口上限 16 亿人基本是一致的。图 9.5 是用 Excel 绘制的人口趋势预测图，其中粗实线是观测值，细虚线是预测值。

输出结果 9.6

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
k	15.760	.650	14.455	17.064
a	10.135	.693	8.746	11.525
b	.997	.000	0.996	.998
c	1.551	.071	1.408	1.694

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	5 266.738	4	1 316.685
Residual	0.884	52	0.017
Uncorrected Total	5 267.622	56	
Corrected Total	319.677	55	

Dependent variable: y.

a. R squared = 1 - (Residual Sum of Squares)/(Corrected Sum of Squares) = 0.997.

专业课辅导+真题资料+微信公众号：432应统书院

图 9.5　威布尔模型预测我国人口趋势图

例 9.7

柯布-道格拉斯生产函数研究。在计量经济学中有一种熟知的C-D（Cobb-Douglas）生产函数

$$y=AK^{\alpha}L^{\beta} \tag{9.15}$$

其中，y 为产出；K(资本)，L(劳动力）为两个投入要素；$A>0$ 为效率系数；α 和 β 为 K 和 L 的产出弹性；A，α，β 均为待估参数。

α 是产出对资本投入的弹性系数，度量在劳动力投入保持不变时资本投入增加 1% 时产出平均增加的百分比。

β 是产出对劳动力投入的弹性系数，度量在资本投入保持不变时劳动力投入增加 1% 时产出平均增加的百分比。

两个弹性系数之和 $\alpha+\beta$ 表示规模报酬（returns to scale）。$\alpha+\beta=1$ 表示规模报酬不变，即 1 倍的投入带来 1 倍的产出；$\alpha+\beta<1$ 表示规模报酬递减，即 1 倍的投入带来少于 1 倍的产出；$\alpha+\beta>1$ 表示规模报酬递增，即 1 倍的投入带来大于 1 倍的产出。

我们假定误差项 ε_t 满足基本假设（3.7）式的高斯-马尔柯夫条件，对模型（9.15）式可以按两种形式设定随机误差项：

（1）乘性误差项，模型形式为 $y=AK^{\alpha}L^{\beta}e^{\varepsilon}$。

（2）加性误差项，模型形式为 $y=AK^{\alpha}L^{\beta}+\varepsilon$。

对乘性误差项，模型可通过两边取对数转化成线性模型

$$\ln y=\ln A+\alpha\ln K+\beta\ln L+\varepsilon \tag{9.16}$$

令 $y'=\ln y$，$\beta_0=\ln A$，$x_1=\ln K$，$x_2=\ln L$，则转化为线性回归方程

$$y'=\beta_0+\alpha x_1+\beta x_2+\varepsilon$$

专业课辅导+真题资料+微信公众号：432应统书院

以下我们分别用乘性误差项模型和加性误差项模型拟合 C-D 生产函数，选取的数据见表 9.12。

表 9.12

年份	t	GDP	K	L	lnGDP	ln$K$	ln$L$
1978	1	3 624.1	1 377.9	40 152	8.195 361	7.228 316	10.600 43
1979	2	4 038.2	1 474.2	41 024	8.303 554	7.295 871	10.621 91
1980	3	4 517.8	1 590.0	42 361	8.415 780	7.371 489	10.653 98
1981	4	4 862.4	1 581.0	43 725	8.489 287	7.365 813	10.685 68
1982	5	5 294.7	1 760.2	45 295	8.574 462	7.473 183	10.720 95
1983	6	5 934.5	2 005.0	46 436	8.688 538	7.603 399	10.745 83
1984	7	7 171.0	2 468.6	48 197	8.877 800	7.811 406	10.783 05
1985	8	8 964.4	3 386.0	49 873	9.101 016	8.127 405	10.817 24
1986	9	10 202.2	3 846.0	51 282	9.230 359	8.254 789	10.845 10
1987	10	11 962.5	4 322.0	52 783	9.389 532	8.371 474	10.873 94
1988	11	14 928.3	5 495.0	54 334	9.611 014	8.611 594	10.902 91
1989	12	16 909.2	6 095.0	55 329	9.735 613	8.715 224	10.921 05
1990	13	18 547.9	6 444.0	64 749	9.828 112	8.770 905	11.078 27
1991	14	21 617.8	7 517.0	65 491	9.981 272	8.924 922	11.089 67
1992	15	26 638.1	9 636.0	66 152	10.190 10	9.173 261	11.099 71
1993	16	34 634.4	14 998.0	66 808	10.452 60	9.615 672	11.109 58
1994	17	46 759.4	19 260.6	67 455	10.752 77	9.865 817	11.119 22
1995	18	58 478.1	23 877.0	68 065	10.976 41	10.080 67	11.128 22
1996	19	67 884.6	26 867.2	68 950	11.125 56	10.198 66	11.141 14
1997	20	74 462.6	28 457.6	69 820	11.218 05	10.256 17	11.153 68
1998	21	78 345.2	29 545.9	70 637	11.268 88	10.293 70	11.165 31
1999	22	82 067.5	30 701.6	71 394	11.315 30	10.332 07	11.175 97
2000	23	89 468.1	32 611.4	72 085	11.401 64	10.392 42	11.185 60
2001	24	97 314.8	37 460.8	73 025	11.485 71	10.531 05	11.198 56
2002	25	105 172.3	42 355.4	73 740	11.563 36	10.653 85	11.208 30

其中，y 是 GDP（亿元）；K 是资金投入，包括固定资产投资和库存占用资金（亿元）；L 是就业总人数（万人）。

（1）假设随机误差项为相乘的，用两边取对数的办法，按照（9.16）式将模型转化为线性形式，对数变换后的数据见表 9.12，用 SPSS 作线性回归得输出结果 9.7。

输出结果 9.7

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32.236	2	16.118	5 917.774	.000
	Residual	.060	22	.003		
	Total	32.296	24			

专业课辅导+真题资料+微信公众号：432应统书院

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-2.086	1.903		-1.096	.285
lnK	.902	.035	.936	25.863	.000
lnL	.361	.201	.065	1.794	.087

a. Dependent Variable: lnGDP.

得两个弹性系数为 $\alpha=0.902$，$\beta=0.361$，资金的贡献率大于劳动力的贡献率。规模报酬 $\alpha+\beta=0.902+0.361=1.263>1$ 表示规模报酬递增。效率系数 $A=e^{-2.086}=0.1242$。其中系数 β 的显著性概率 P 值 $=0.087$，显著性较弱。得乘性误差项的 C-D 生产函数为

$$\hat{y}=0.1242K^{0.902}L^{0.361}$$

（2）对加性误差项模型，不能通过变量变换转化成线性模型，只能用非线性最小二乘法求解未知参数。以上面乘性误差项的参数为初值做非线性最小二乘，经过 81 步迭代得输出结果 9.8。其中参数 β 的置信度为 95% 的置信区间为(-0.555，1.565)，包含 0 在内，因而不能认为 β 非 0，显著性较弱。

得加性误差项的 C-D 生产函数为

$$\hat{y}=0.020K^{0.922}L^{0.505}$$

输出结果 9.8

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
A	.020	.104	-.196	.236
alpha	.922	.064	.789	1.056
beta	.505	.511	-.555	1.565

乘性误差项模型和加性误差项模型所得的结果有一定差异，其中乘性误差项模型认为 y_t 本身是异方差的，而 $\ln y_t$ 是等方差的。加性误差项模型认为 y_t 是等方差的。从统计性质看两者的差异，前者淡化了 y_t 值大的项（近期数据）的作用，强化了 y_t 值小的项（早期数据）的作用，对早期数据拟合的效果较好，而后者则对近期数据拟合的效果较好。

影响模型拟合效果的统计性质主要是异方差、自相关、共线性这三个方面。异方差可以通过选择乘性误差项模型和加性误差项模型解决，必要时还可以使用加权最小二乘。时间序列数据通常都存在自相关，使用自回归方法可以改进模型的拟合效果。在经济数据中，对参数估计影响最大的往往是共线性。

专业课辅导+真题资料+微信公众号：432应统书院

C-D 生产函数是柯布-道格拉斯于 1928 年提出的经济模型，目前对此模型的结构和应用条件都有很多的改进。在模型结构方面，最常用的改进是增加技术进步因素。在应用条件方面，对 GDP 和资本投入使用可比价格，剔除通货膨胀的影响。见本章思考与练习第 5～6 题。另外，使用横截面数据与使用动态数据的结果也会有所不同。如果对三次产业分别建立 C-D 生产函数，也会得到不同的弹性系数，而我国三次产业结构正在不断调整中，第三产业所占比重不断增大，这也会导致建立全国的 C-D 生产函数时弹性系数不稳定。

三、其他形式的非线性回归

前面介绍了用非线性最小二乘方法求解非线性回归方程，非线性最小二乘是使（9.9）式的残差平方和 $Q(\boldsymbol{\theta})=\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i,\boldsymbol{\theta}))^2$ 达到极小的方法。从决策学的观点看，$Q(\boldsymbol{\theta})$ 是关于残差的损失函数，这种平方损失函数的优点是数学性质好，数学上容易处理，在一定条件下也具有统计学的一些优良性质，但其不足之处是缺乏稳健性。当数据存在异常值时，参数的估计效果变得很差。因而在一些场合，我们希望用一些更稳健的残差损失函数代替平方损失函数，例如绝对值损失函数。绝对值残差损失函数为

$$Q(\boldsymbol{\theta})=\sum_{i=1}^{n}|y_i-f(\boldsymbol{x}_i,\boldsymbol{\theta})|$$

SPSS 的非线性回归程序，可以用数值计算法求解多种损失函数的回归估计值。对于非最小二乘估计，即使是线性回归，也要用数值方法计算。在第 4 章处理异常值问题中提到，当数据存在异常值但又不宜剔除时，可以改用最小绝对值法等更稳健的方法求解。以下以例 3.2 北京开发区数据为例，演示用 SPSS 软件的最小绝对值法求解。步骤如下：

（1）进入非线性回归对话框，在因变量框中选入 y，在 Model Expressions 框中输入回归方程表达式 $b0+b1*x1+b2*x2$。

（2）给参数赋初值，以普通最小二乘估计值为初始值，初始值为 $b_0=-327.039$，$b_1=2.036$，$b_2=0.468$，点击 Continue 返回非线性回归对话框。

（3）点击 Options 选项进入 Options 选项框选择数值计算方法，默认的计算方法是 Levenberg-Marquardt 法，将其改为 Sequential quadratic program 法，点击 Continue 返回非线性回归对话框。用自定义损失函数计算时必须作这个改动。

（4）点击 Loss 进入 Loss Function 对话框给出损失函数，默认的损失函数是 Sum of squared residuals，将其改为 User-defined loss function，然后输入 ABS($y-b0-b1*x1-b2*x2$)，点击 Continue 返回非线性回归对话框。

（5）点击 Save 保存残差变量和预测值。

（6）点击 OK 运行。

专业课辅导+真题资料+微信公众号：432应统书院

经过 21 步迭代得到输出结果 9.9。

输出结果 9.9

Iteration History[b]

Iteration Number[a]		Value of Loss Function	Parameter		
			b0	b1	b2
dimension0	0.1	4633.052	-327.039	2.036	.468
	1.1	4438.306	-327.039	2.055	.540
	2.1	4421.891	-327.039	2.034	.551
	3.1	4375.291	-326.792	1.985	.553
	4.1	4373.563	-326.721	1.986	.553
	5.1	4373.447	-326.707	1.986	.553
	6.1	4373.417	-326.609	1.985	.553
	7.1	4373.168	-326.444	1.986	.553
	8.1	4373.126	-326.383	1.986	.553
	9.1	4373.083	-326.202	1.986	.553
	10.1	4355.280	-221.622	2.101	.448
	11.1	4355.278	-221.610	2.101	.448
	12.1	4355.278	-221.610	2.101	.448
	13.1	4354.934	-216.456	2.116	.438
	14.1	4354.203	-214.585	2.121	.434
	15.1	4354.194	-214.397	2.122	.434
	16.1	4353.881	-208.933	2.138	.424
	17.1	4353.431	-200.905	2.161	.409
	18.1	4353.431	-200.901	2.161	.409
	19.1	4353.431	-200.901	2.161	.409
	20.1	4353.431	-200.901	2.161	.409

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 21 iterations because it could not improve on the current point.

从上表得最小绝对值法的经验回归方程为

$$\hat{y}_a = -200.901 + 2.161x_1 + 0.409x_2$$

最小绝对值法残差 e_{ia} 与预测值 $\hat{y}_{ia}$ 见表 9.13，为了便于比较，把普通最小二乘法的残差也列在了表 9.13 中。

专业课辅导+真题资料+微信公众号：432应统书院

表 9.13

序号	x_1	x_2	y_i	$\hat{y}_{ia}$	e_{ia}	$\hat{y}_i$	e_i
1	25	3 547.79	553.96	1 302.68	−748.72	1 385.63	−831.67
2	20	896.34	208.55	208.55	0	133.52	75.03
3	6	750.32	3.1	118.63	−115.53	36.62	−33.52
4	1 001	2 087.05	2 815.4	2 815.38	0.02	2 688.57	126.83
5	525	1 639.31	1 052.12	1 603.62	−551.5	1 509.71	−457.59
6	825	3 357.7	3 427	2 954.13	472.87	2 925.40	501.60
7	120	808.47	442.82	388.79	54.03	295.96	146.86
8	28	520.27	70.12	72.19	−2.07	−26.34	96.46
9	7	671.13	122.24	88.44	33.8	1.57	120.67
10	532	2 863.32	1 400	2 118.85	−718.85	2 097.28	−697.28
11	75	1 160	464	435.15	28.85	369.00	95.00
12	40	862.75	7.5	238.05	−230.55	158.51	−151.01
13	187	672.99	224.18	478.25	−254.07	368.92	−144.74
14	122	901.76	538.94	431.23	107.71	343.73	195.21
15	74	3 546.18	2 442.79	1 407.93	1 034.86	1 484.64	958.15

从表 9.13 中看到，普通最小二乘法的最大残差 $e_{15}=958.15$，最小绝对值法的最大残差 $\hat{e}_{15}=1\,034.86$，最小绝对值法的最大残差比普通最小二乘法的最大残差更大，这是否与最小绝对值法的稳健性相矛盾呢？

其实这正说明了最小绝对值法的稳健性。这是因为最小绝对值法受异常值的影响程度小，回归线向异常值靠拢的程度也小，因而异常值的残差反而大。

对一般的非最小二乘损失函数，参数估计的性质难以确定，因此软件只给出了参数估计值，没有给出方差分析表与回归系数的显著性检验表，这是用非最小二乘损失函数的不足之处。

本例是对线性回归采用最小绝对值损失函数求解，对非线性回归的一般非最小二乘损失函数的求解方法完全相似。

9.4 本章小结与评注

非线性回归的内容非常丰富，特别是线性回归的问题日趋成熟，许多统计学家把精力投入非线性问题的研究。非线性问题要比线性问题复杂得多，今后一个相当长的时期非线性问题是人们关注的热点。

对于可化为线性模型的曲线回归问题，通常的处理方法都是先化为线性模型，然后用普通最小二乘法求出参数的估计值，最后经过适当的变换就得到所求的回归曲线。通过对因变量变换使曲线回归线性化的方法，当然对估计参数的性质会产生影响，比如不具有无偏性等（参见参考文献［5］）。

根据实际观测数据配以合适的曲线模型一般有两个重要的步骤：

专业课辅导+真题资料+微信公众号：432应统书院

一是确定曲线类型。对一个自变量的情况，确定曲线类型一般是把样本观测值画成散点图，根据散点图的形状来大体确定曲线类型。再就是根据专业知识来确定曲线类型，如商品的销售量与广告费用之间的关系，一般用 S 形曲线来描述；在农业生产中，粮食的产量与种植密度往往服从抛物线关系。对于由专业知识可以确定的曲线类型，就用相应的模型去试着拟合，如果拟合的效果可以，问题就解决了。有时对一个问题需要用不同的曲线模型来试验，以求得一个最好的模型。

二是参数估计问题。如果可将曲线模型转化为线性模型，就可用普通最小二乘法估计未知参数；如果不能转化成线性模型，则参数的估计就要采用非线性最小二乘法。非线性最小二乘法比普通最小二乘法要复杂得多，一般都是用迭代方法。现在流行的 SPSS 软件包中就有非线性最小二乘法，所以非线性最小二乘法的参数估计也变得容易起来。

由于任一连续函数都可用分段多项式来逼近，所以在实际问题中，不论变量 y 与其他变量的关系如何，在相当大的范围内我们总可以用多项式来拟合。例如在一元回归中，如果变量 y 与 x 的关系假定为 p 次多项式（9.2）式，就可以转化为多元线性回归模型（9.5）式来处理。利用多项式回归模型可能会把已有的数据拟合得十分理想，但是，如果对较大的 x 作外推预测，这种多项式回归函数就可能会得到很差的结果，预测值可能会朝着意想不到的方向折转，从而与实际情况严重不符。所有类型的多项式回归函数，尤其是高阶多项式回归函数都具有这种外推风险。特别地，对于一元回归，只要用一元 $n-1$ 次多项式就可以把 n 对观测数据完全拟合，多项式曲线通过所有 $n-1$ 个点，残差平方和为零，但是这样的回归拟合却没有任何实际意义。因此，人们必须谨慎地使用高阶多项式回归模型，因为得到的回归函数只是数据的良好拟合，而并不能如实地表明 x 与 y 之间回归关系的基本特征，并且会导致不规则的外推。我们建议在应用多项式回归时，阶数一般不要超过三阶。

在多项式回归中，自变量 x_i 常用围绕均值 $\bar{x}$ 的离差 $x_i-\bar{x}$ 表示，这样做的原因是 x_i 与其高次幂项 x_i^2，x_i^3 等往往高度相关，产生共线性。参数估计时会出现计算上的麻烦，尤其是用手工计算时，数据的舍入误差会对计算结果造成很大的影响。把自变量表示成与其均值的离差，可以降低变量间的多重共线性，有助于避免计算方面的困难。现在的计算软件都采用双精度计算，x_i 与其高次幂项相关所造成的计算误差影响一般不大，因而不必总是把自变量表示成与其均值的离差 $x_i-\bar{x}$ 的形式。多项式回归的内容非常丰富，有兴趣的读者可参见参考文献［3］、［5］、［7］。

在一元线性回归中，我们用相关系数 r 检验回归方程的可靠性。对于一元非线性回归问题，许多书上用类似于相关系数的相关指数来衡量所配曲线效果的好坏。在实际应用中，相关指数 R^2 用于一元非线性强度不高的回归方程的评价还没有碰到什么问题。然而，相关指数 R^2 能否直接用于非线性强度很高的回归方程的评价，还需进一步探讨。我们经常会见到人们毫无顾忌地使用相关指数 R^2，笔者认为，对非线性强度很高的回归方程在使用相关指数 R^2 时应更慎重一些。1990 年就有人

专业课辅导+真题资料+微信公众号：432应统书院

（参考文献［25］）对这一问题提出质疑，认为 R^2 不能用于非线性回归方程的评价，目前这一问题的研究在国内已引起一些学者的关注。一般地说，当非线性回归模型选择正确，回归拟合效果好时，相关指数 R^2 能够如实反映回归拟合效果；而当回归拟合效果差时，相关指数 R^2 则不能够如实反映回归拟合效果，甚至可能取负值。

□ 思考与练习

9.1 在非线性回归线性化时，对因变量作变换应注意什么问题？

9.2 为了研究生产率与废料率之间的关系，记录了如表 9.14 所示的数据，请画出散点图，根据散点图的趋势拟合适当的回归模型。

表 9.14

生产率 x（单位/周）	1 000	2 000	30 000	3 500	4 000	4 500	5 000
废品率 y（%）	5.2	6.5	6.8	8.1	10.2	10.3	13.0

9.3 已知变量 x 与 y 的样本数据如表 9.15 所示，画出散点图，试用 $\alpha e^{\beta/x}$ 来拟合回归模型，假设：

（1）乘性误差项 $y=\alpha e^{\beta/x}\epsilon^{\varepsilon}$。

（2）加性误差项 $y=\alpha e^{\beta/x}+\varepsilon$。

表 9.15

序号	x	y	序号	x	y
1	4.20	0.086	9	2.60	0.220
2	4.06	0.090	10	2.40	0.240
3	3.80	0.100	11	2.20	0.350
4	3.60	0.120	12	2.00	0.440
5	3.40	0.130	13	1.80	0.620
6	3.20	0.150	14	1.60	0.940
7	3.00	0.170	15	1.40	1.620
8	2.80	0.190			

9.4 Logistic 回归函数常用于拟合某种消费品的拥有率，表 9.16 是北京市每百户家庭平均拥有的照相机数，试针对以下两种情况拟合 Logistic 回归函数

$$y=\frac{1}{\frac{1}{u}+b_0b_1^t}$$

（1）已知 $u=100$，用线性化方法拟合。

（2）u 未知，用非线性最小二乘法拟合。根据经济学的意义知道，u 是拥有率的上限，初值可取 100；$b_0>0$，$0<b_1<1$，初值请读者自己选择。

专业课辅导+真题资料+微信公众号：432应统书院

表 9.16

年份	t	y	年份	t	y
1978	1	7.5	1988	11	59.6
1979	2	9.8	1989	12	62.2
1980	3	11.4	1990	13	66.5
1981	4	13.3	1991	14	72.7
1982	5	17.2	1992	15	77.2
1983	6	20.6	1993	16	82.4
1984	7	29.1	1994	17	85.4
1985	8	34.6	1995	18	86.8
1986	9	47.4	1996	19	87.2
1987	10	55.5			

9.5　表 9.17 数据中 GDP 和投资额 K 都是用定基居民消费价格指数（CPI）缩减后的，1978 年的价格指数为 100。

(1) 用线性化的乘性误差项模型拟合 C-D 生产函数。

(2) 用非线性最小二乘拟合加性误差项模型的 C-D 生产函数。

(3) 对线性化回归检验自相关，如果存在自相关则用自回归方法改进。

(4) 对线性化回归检验共线性，如果存在共线性则用岭回归方法改进。

表 9.17

年份	t	CPI	GDP	K	L
1978	1	100.00	3 624.1	1 377.9	40 152
1979	2	101.90	3 962.9	1 446.7	41 024
1980	3	109.54	4 124.2	1 451.5	42 361
1981	4	112.28	4 330.6	1 408.1	43 725
1982	5	114.53	4 623.1	1 536.9	45 295
1983	6	116.82	5 080.2	1 716.4	46 436
1984	7	119.97	5 977.3	2 057.7	48 197
1985	8	131.13	6 836.3	2 582.2	49 873
1986	9	139.65	7 305.4	2 754.0	51 282
1987	10	149.85	7 983.2	2 884.3	52 783
1988	11	178.02	8 385.9	3 086.8	54 334
1989	12	210.06	8 049.7	2 901.5	55 329
1990	13	216.57	8 564.3	2 975.4	64 749
1991	14	223.94	9 653.5	3 356.8	65 491
1992	15	238.27	11 179.9	4 044.2	66 152
1993	16	273.29	12 673.0	5 487.9	66 808
1994	17	339.16	13 786.9	5 679.0	67 455
1995	18	397.15	14 724.3	6 012.0	68 065
1996	19	430.12	15 782.8	6 246.5	68 950
1997	20	442.16	16 840.6	6 436.0	69 820
1998	21	438.62	17 861.6	6 736.1	70 637
1999	22	432.48	18 975.9	7 098.9	71 394
2000	23	434.21	20 604.7	7 510.5	72 085
2001	24	437.25	22 256.0	8 567.3	73 025
2002	25	433.75	24 247.0	9 764.9	73 740

专业课辅导+真题资料+微信公众号：432应统书院

9.6 对上题的数据，拟合含有技术进步的 C-D 生产函数：

$$y = Ae^{\mu t}K^{\alpha}L^{\beta}$$

其中，$e^{\mu t}$ 代表技术进步对产出的影响。

(1) 用线性化的乘性误差项模型拟合。

(2) 用非线性最小二乘对加性误差项模型做拟合。

(3) 对线性化回归检验自相关，如果存在自相关则用自回归方法改进。

(4) 对线性化回归检验共线性，如果存在共线性则用岭回归方法改进。

专业课辅导+真题资料+微信公众号：432应统书院

第 10 章 Chapter 10 含定性变量的回归模型

在实际问题的研究中，经常会碰到一些非数量型的变量，如品质变量：性别；正常年份、干旱年份；战争与和平；改革前、改革后等。在建立一个经济问题的回归方程时，经常需要考虑这些品质变量，如建立粮食产量预测方程就应考虑到正常年份与受灾年份的不同影响。我们把这些品质变量也称为定性变量。定性变量的回归问题已有不少研究（参见参考文献［6］），本章主要介绍自变量含定性变量的回归模型和因变量是定性变量的回归模型两大类。

10.1 自变量含定性变量的回归模型

在回归分析中，我们对一些自变量是定性变量的情形先给予数量化处理，处理方法是引进只取 0 和 1 两个值的虚拟自变量将定性变量数量化。当某一属性出现时，虚拟变量取值 1，否则取值 0。虚拟变量也称为哑变量。

一、简单情况

首先讨论定性变量只取两类可能值的情况，例如研究粮食产量问题，y 为粮食产量，x 为施肥量，另外再考虑气候问题，分为正常年份和干旱年份两种情况，对这个问题的数量化方法是引入一个 0—1 型变量 D，令

$D_i=1$，表示正常年份

$D_i=0$，表示干旱年份

粮食产量的回归模型为

$$y_i=\beta_0+\beta_1 x_i+\beta_2 D_i+\varepsilon_i \tag{10.1}$$

专业课辅导+真题资料+微信公众号：432应统书院

其中，$i=1, 2, \cdots, n$，在以下回归模型中不再一一注明。干旱年份的粮食平均产量为

$$E(y_i \mid D_i=0)=\beta_0+\beta_1 x_i$$

正常年份的粮食平均产量为

$$E(y_i \mid D_i=1)=(\beta_0+\beta_2)+\beta_1 x_i$$

这里有一个前提条件，就是认为干旱年份与正常年份回归直线的斜率 β_1 是相等的；也就是说，不论是干旱年份还是正常年份，施肥量 x 每增加一个单位，粮食产量 y 平均都增加相同的数量 β_1。对（10.1）式的参数估计仍采用普通最小二乘法。

例 10.1

某经济学家想调查文化程度对家庭储蓄的影响，在一个中等收入的样本框中，随机调查了 13 户高学历家庭与 14 户中低学历家庭。因变量 y 为上一年家庭储蓄增加额，自变量 x_1 为上一年家庭总收入，自变量 x_2 为家庭学历。高学历家庭 $x_2=1$，低学历家庭 $x_2=0$，调查数据见表 10.1。

表 10.1

序号	y（元）	x_1（万元）	x_2	e_i	de_i
1	235	2.3	0	−588	455
2	346	3.2	1	−220	−2 372
3	365	2.8	0	−2 371	−1 047
4	468	3.5	1	−1 246	−3 229
5	658	2.6	0	−1 313	−101
6	867	3.2	1	301	−1 851
7	1 085	2.6	0	−886	326
8	1 236	3.4	1	−96	−2 135
9	1 238	2.2	0	797	1 784
10	1 345	2.8	1	2 309	−67
11	2 365	2.3	0	1 542	2 585
12	2 365	3.7	1	−115	−1 985
13	3 256	4.0	1	−371	−2 074
14	3 256	2.9	0	137	1 517
15	3 265	3.8	1	403	−1 412
16	3 265	4.6	1	−2 658	−4 023
17	3 567	4.2	1	−826	−2 416
18	3 658	3.7	1	1 178	−692
19	4 588	3.5	0	−827	891
20	6 436	4.8	1	−252	−1 505
21	9 047	5.0	1	1 593	453
22	7 985	4.2	0	−108	2 002
23	8 950	3.9	0	2 005	3 947
24	9 865	4.8	0	−524	1 924
25	9 866	4.6	0	243	2 578
26	10 235	4.8	0	−154	2 294
27	10 140	4.2	0	2 047	4 157

专业课辅导+真题资料+微信公众号：432应统书院

建立 y 对 x_1，x_2 的线性回归，计算结果见输出结果 10.1，残差值 e_i 列在表 10.1 中。

输出结果 10.1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.938[a]	.879	.869	1288.68

a. Predictors: (Constant), x2, x1.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	290372875.924	2	145186437.962	87.425	.000
	Residual	39856639.705	24	1660693.321		
	Total	330229515.630	26			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-7976.809	1093.445		-7.295	.000
x1	3826.129	304.591	.921	12.562	.000
x2	-3700.330	513.445	-.529	-7.207	.000

两个自变量 x_1 与 x_2 的系数都是显著的，复决定系数 $R^2=0.879$，回归方程为

$$\hat{y}=-7\,976+3\,826x_1-3\,700x_2$$

这个结果表明，中等收入的家庭每增加 1 万元收入，平均拿出 3 826 元作为储蓄。高学历家庭每年的平均储蓄额少于低学历的家庭，平均少 3 700 元。

如果不引入家庭学历定性变量 x_2，仅用 y 对家庭年收入 x_1 作一元线性回归，得决定系数 $R^2=0.618$，说明拟合效果不好。y 对 x_1 的一元线性回归的残差 de_i 也列在了表 10.1 中。

家庭年收入 x_1 是连续型变量，它对回归的贡献也是不可缺少的。如果不考虑家庭年收入这个自变量，13 户高学历家庭的平均年储蓄增加额为 3 009.31 元，14 户低学历家庭的平均年储蓄增加额为 5 059.36 元，这样会认为高学历家庭每年的储蓄额比低学历的家庭平均少 5 059.36－3 009.31＝2 050.05 元，而用回归法算出的数值是 3 700 元，两者并不相等。

用回归法算出的高学历家庭每年的平均储蓄额比低学历的家庭平均少 3 700 元，这是在假设两者的家庭年收入相等的基础上的储蓄差值，或者说是消除了家庭年收入的影响后的差值，因而反映不同学历家庭储蓄额的真实差异。而直接由样本计算的差值 2 050.05 元是包含家庭年收入影响在内的差值，是虚假的差值。所调查的 13 户高学历家庭的平均年收入额为 3.838 5 万元，14 户低学历家庭的平均年收入额为 3.407 1 万元，两者并不相等。

专业课辅导+真题资料+微信公众号：432应统书院

通过本例的分析我们看到，在一些问题的分析中，仅依靠平均数是不够的，很可能得到虚假的数值。只有通过对数据的深入分析，才能得到正确结果。

需要指出的是，虽然虚拟变量取某一数值，但这一数值没有任何数量大小的意义，它仅仅用来说明观察单位的性质或属性。

以上定性自变量只取两个可能值：干旱或正常；高学历或低学历。一般情况就是取是或否两个值，只需用一个 0—1 型自变量表示即可。以下把这种只取两个值的情况推广到取多个值的情况。

二、复杂情况

某些场合下，定性自变量可能取多类值，例如某商厦策划营销方案，需要考虑销售额的季节性影响，季节因素分为春、夏、秋、冬四种情况。为了用定性自变量反映春、夏、秋、冬四季，我们初步设想引入如下四个 0—1 型自变量：

$$\begin{cases} x_1=1, \text{春季} \\ x_1=0, \text{其他} \end{cases} \quad \begin{cases} x_2=1, \text{夏季} \\ x_2=0, \text{其他} \end{cases}$$

$$\begin{cases} x_3=1, \text{秋季} \\ x_3=0, \text{其他} \end{cases} \quad \begin{cases} x_4=1, \text{冬季} \\ x_4=0, \text{其他} \end{cases}$$

可是这样做却产生了一个新的问题，即四个自变量 x_1，x_2，x_3，x_4 之和恒等于 1，即 $x_1+x_2+x_3+x_4=1$，构成完全多重共线性。解决这个问题的方法很简单，我们只需去掉一个 0—1 型变量，只保留三个 0—1 型自变量即可。例如去掉 x_4，只保留 x_1，x_2，x_3。

一般情况下，当一个定性变量有 k 类可能的取值时，需要引入 $k-1$ 个0—1型自变量。当 $k=2$ 时，只需要引入一个 0—1 型自变量即可。

包含多个 0—1 型自变量的计算，仍然是采用普通的线性最小二乘回归方法，在此就不举例了。

10.2 自变量含定性变量的回归模型的应用

一、分段回归

在实际问题中，我们会碰到某些变量在影响因素的不同范围内变化趋势截然不同，例如经济问题在经济政策有较大调整时，调整前与调整后的变化幅度会有很大不同。对于这种问题，有时用多种曲线拟合效果仍不能令人满意。如果作残差分析，会发现残差不是随机的，而具有一定的系统性。对这类问题，人们自然考虑到用分段回归的方法来处理。

专业课辅导+真题资料+微信公众号：432应统书院

例 10.2

表 10.2 给出某工厂生产批量 x 与单位成本 y（美元）的数据。试用分段回归建立回归模型（参见参考文献 [3]）。

表 10.2

序号	y	x（$=x_1$）	x_2
1	2.57	650	150
2	4.4	340	0
3	4.52	400	0
4	1.39	800	300
5	4.75	300	0
6	3.55	570	70
7	2.49	720	220
8	3.77	480	0

这是一个生产批量与生产成本的问题，单位成本 y 对生产批量的回归在某 x_p 点以内服从一种线性关系，而在生产批量超过 x_p 时可能服从另一种线性关系。由图 10.1 可看出数据在生产批量 $x_p=500$ 时发生较大变化，即批量大于 500 时成本明显下降。我们考虑由两段构成的分段线性回归，这可以通过引入一个 0—1 型虚拟自变量实现。假定回归直线的斜率在 $x_p=500$ 处改变，建立回归模型

$$y_i=\beta_0+\beta_1 x_i+\beta_2(x_i-500)D_i+\epsilon_i \tag{10.2}$$

来拟合，其中

$$\begin{cases} D_i=1，当\ x_i>500 \\ D_i=0，当\ x_i\leqslant 500 \end{cases}$$

图 10.1　单位成本与批量的散点图

专业课辅导+真题资料+微信公众号：432应统书院

回归模型（10.2）式实际上是一个二元线性回归模型，为了更清楚起见，引入两个新的自变量 x_1，x_2。有

$$x_{i1}=x_i$$
$$x_{i2}=(x_i-500)D_i$$

其中，x_1 为生产批量；x_2 的数值列在表 10.2 中。这样回归模型（10.2）式转化为标准形式的二元线性回归模型

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i \tag{10.3}$$

（10.3）式可以分解为两个线性回归方程：

当 $x_1\leqslant 500$ 时，得到

$$E(y)=\beta_0+\beta_1 x_1 \tag{10.4}$$

当 $x_1>500$ 时，得到

$$E(y)=(\beta_0-500\beta_2)+(\beta_1+\beta_2)x_1 \tag{10.5}$$

于是 β_1 和 $\beta_1+\beta_2$ 分别是两条回归线（10.4）式和（10.5）式的斜率，β_0 和 $(\beta_0-500\beta_2)$ 是两个 y 截距，如图 10.2 所示。

图 10.2

用普通最小二乘法拟合模型（10.3）式得回归方程

$$\hat{y}=5.895-0.003\,95x_1-0.003\,89x_2 \tag{10.6}$$

利用此模型可说明生产批量小于 500 时，每增加 1 个单位批量，单位成本降低 0.003 95 美元；当生产批量大于 500 时，每增加 1 个单位批量，单位成本降低 0.003 95+0.003 89=0.007 84 美元。

上面是参考文献［3］的分析过程。笔者认为，以上只是根据散点图从直观上判断本例数据应该用折线回归拟合，这一点还需要作统计的显著性检验，只需对（10.2）式的回归系数 β_2 作显著性检验即可。回归方程（10.6）式的有关计算结果如输出结果 10.2 所示。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.985[a]	.969	.957	.2449

a. Predictors: (Constant), x_2, x.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.486	2	4.743	79.059	.000
	Residual	.300	5	5.999E-02		
	Total	9.786	7			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	5.895	.604		9.757	.000
x	-3.954E-03	.001	-.611	-2.65	.045
x2	-3.893E-03	.002	-.388	-1.69	.153

复决定系数 $R^2=0.969$，拟合效果很好。对 β_2 的显著性检验的 t 值$=-1.69$，显著性概率（Sig）$=0.153$，β_2 没有通过显著性检验，不能认为 β_2 非零。这样，根据显著性检验，还不能认为本例数据适合折线回归拟合。

用 y 对 x 作一元线性回归，计算结果如输出结果 10.3 所示。

输出结果 10.3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.976[a]	.952	.944	.2800

a. Predictors: (Constant), x.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.316	1	9.316	118.839	.000
	Residual	0.470	6	7.839E-02		
	Total	9.786	7			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	6.795	.324		20.963	.000
x	-6.318E-03	.001	-.976	-10.90	.000

专业课辅导+真题资料+微信公众号：432应统书院

y 对 x 的一元线性回归的决定系数 $R^2=0.952$，回归方程为

$$\hat{y}=6.795-0.006\,318x \tag{10.7}$$

（10.7）式说明，生产批量每增加一件，单位成本平均下降 0.006 318 美元，这个结论在自变量的样本范围 300～800 内都是适用的。

二、回归系数相等的检验

例 10.3

回到例 10.1 的问题，例 10.1 引入 0—1 型自变量的方法是假定储蓄增加额 y 对家庭总收入的回归斜率 β_1 与家庭学历无关，家庭学历只影响回归常数项 β_0，这个假设是否合理，还需要作统计检验。

检验方法是引入如下含有交互效应的回归模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i1}x_{i2}+\varepsilon_i \tag{10.8}$$

其中，y 为上一年家庭储蓄增加额；x_1 为上一年家庭总收入；x_2 为家庭学历，高学历家庭 $x_2=1$，低学历家庭 $x_2=0$。回归模型（10.8）式可以分解为对高学历和对低学历家庭的两个线性回归模型，分别为

高学历家庭 $x_2=1$：

$$\begin{aligned}y_i&=\beta_0+\beta_1x_{i1}+\beta_2+\beta_3x_{i1}+\varepsilon_i\\&=(\beta_0+\beta_2)+(\beta_1+\beta_3)x_{i1}+\varepsilon_i\end{aligned} \tag{10.9}$$

低学历家庭 $x_2=0$：

$$y_i=\beta_0+\beta_1x_{i1}+\varepsilon_i \tag{10.10}$$

可见，高学历家庭的回归常数为 $\beta_0+\beta_2$，回归系数为 $\beta_1+\beta_3$；低学历家庭的回归常数为 β_0，回归系数为 β_1。要检验两个回归方程的回归系数相等，等价于对（10.8）式的回归模型作参数的假设检验

$$H_0:\beta_3=0$$

当拒绝 H_0 时，认为 $\beta_3\neq0$，这时高学历家庭与低学历家庭的储蓄回归模型实际上被拆分为两个不同的回归模型（10.9）式和（10.10）式。当不拒绝 H_0 时，认为 $\beta_3=0$，这时高学历与低学历家庭的储蓄回归模型是如下形式的联合回归模型

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i \tag{10.11}$$

这正是例 10.1 所建立的回归模型。拟合（10.8）式的回归模型，回归系数检验表如输出结果 10.4 所示。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.4

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-8763.936	1270.878		-6.896	.000
x1	4057.151	359.284	0.977	11.292	.000
x2	-776.939	2514.459	-.111	-.309	.760
x3	-787.564	663.367	-.443	-1.187	.247

从输出结果 10.4 中看到，对 β_3 显著性检验的显著性概率（Sig）=0.247，应该不拒绝原假设 $H_0:\beta_3=0$，例 10.1 采用的回归模型（10.11）式是正确的。

另外，输出结果 10.4 中 x_2 的回归系数 β_2 的显著性概率为 0.760，也没有通过显著性检验，并且比 β_3 的显著性更低，是否应该首先剔除 x_2 而保留 x_3？回答是否定的，因为这样做与经济意义不符。对回归模型（10.9）式与（10.10）式，当 $\beta_2=0$，表明两个回归方程的常数项相等；当 $\beta_3=0$，表明两个回归方程的斜率相等。经济学家首先关心的是两个回归方程的斜率是否相等，其次才关心常数项是否相等。通常认为，回归常数项是在自变量为零时 y 的平均值，但在本例中则没有这个现实意义。这是因为本例是对中等收入家庭的储蓄分析，收入为零的家庭的储蓄额超出了本模型所包含的范围。本例的回归常数项仅是与储蓄的平均值有关的一个数值。

10.3　因变量是定性变量的回归模型

在许多社会经济问题中，所研究的因变量往往只有两个可能结果，这样的因变量也可用虚拟变量来表示，虚拟变量可取 0 或 1。

例如，在一次住房展销会上，与房地产商签订初步购房意向书的顾客中，在随后的 3 个月的时间内，只有一部分顾客确实购买了房屋。确实购买了房屋的顾客记为 1，没有购买房屋的顾客记为 0。

再如，在是否参加赔偿责任保险的研究中，根据户主的年龄、流动资产额和户主的职业，因变量 y 规定有两种可能的结果：户主有赔偿责任保险单，户主没有赔偿责任保险单。这种结果也可以用虚拟变量 1 或 0 来表示。

再如，在一项社会安全问题的调查中，一个人在家是否害怕生人来，因变量 $y=1$ 表示害怕，$y=0$ 表示不怕（参见参考文献［10］）。

上面的例子说明，因变量的结果只取两种可能情况的应用很广泛。

一、定性因变量的回归方程的意义

设因变量 y 是只取 0，1 两个值的定性变量，考虑简单线性回归模型

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i \tag{10.12}$$

专业课辅导+真题资料+微信公众号：432应统书院

在这种 y 只取 0，1 两个值的情况下，因变量均值 $E(y_i)=\beta_0+\beta_1 x_i$ 有着特殊的意义。

由于 y_i 是 0—1 型贝努利随机变量，得如下概率分布

$$P(y_i=1)=\pi_i$$
$$P(y_i=0)=1-\pi_i$$

根据离散型随机变量期望值的定义，可得

$$E(y_i)=1(\pi_i)+0(1-\pi_i)=\pi_i \tag{10.13}$$

进而得到

$$E(y_i)=\pi_i=\beta_0+\beta_1 x_i$$

所以，作为由回归函数给定的因变量均值，$E(y_i)=\beta_0+\beta_1 x_i$ 是自变量水平为 x_i 时 $y_i=1$ 的概率。对因变量均值的这种解释既适用于这里的简单线性回归函数，也适用于复杂的多元回归函数。当因变量是 0，1 时，因变量均值总是代表给定自变量时 $y=1$ 的概率。

二、定性因变量回归的特殊问题

（1）离散非正态误差项。对一个取值为 0 和 1 的因变量，误差项 $\varepsilon_i=y_i-(\beta_0+\beta_1 x_i)$ 只能取两个值：

当 $y_i=1$ 时，$\varepsilon_i=1-\beta_0-\beta_1 x_i=1-\pi_i$

当 $y_i=0$ 时，$\varepsilon_i=-\beta_0-\beta_1 x_i=-\pi_i$

显然，误差项 ε_i 是两点型离散分布，当然正态误差回归模型的假定就不适用了。

（2）零均值异方差性。当因变量是定性变量时，误差项 ε_i 仍然保持零均值，这时出现的另一个问题是误差项 ε_i 的方差不相等。由于 y_i 与 ε_i 只是相差一个常数 $\beta_0+\beta_1 x_i$，因而 y_i 与 ε_i 的方差是相等的。0—1 型随机变量 ε_i 的方差为

$$\begin{aligned} D(\varepsilon_i)&=D(y_i)\\ &=\pi_i(1-\pi_i)\\ &=(\beta_0+\beta_1 x_i)(1-\beta_0-\beta_1 x_i) \end{aligned} \tag{10.14}$$

由（10.14）式可看到，ε_i 的方差依赖于 x_i，误差项方差随着 x 的不同水平而变化，是异方差，不满足线性回归方程的基本假定，最小二乘估计的效果也就不会好。

（3）回归方程的限制。当因变量为 0，1 虚拟变量时，回归方程代表概率分布，所以因变量均值受到如下限制

$$0\leqslant E(y_i)=\pi_i\leqslant 1$$

专业课辅导+真题资料+微信公众号：432应统书院

一般的回归方程本身并不具有这种限制，线性回归方程 $y_i=\beta_0+\beta_1 x_i$ 将会超出这个限制范围。

对于普通的线性回归所具有的上述三个问题，我们需要构造出能够满足以上限制的回归模型。

10.4　Logistic 回归模型

一、分组数据的 Logistic 回归模型

针对 0—1 型因变量产生的问题，我们对回归模型应该做两个方面的改进。

第一，回归函数应该改用限制在［0，1］区间内的连续曲线，而不能再沿用直线回归方程。限制在［0，1］区间内的连续曲线有很多，例如所有连续型随机变量的分布函数都符合要求，常用的是 Logistic 函数与正态分布函数。Logistic 函数的形式为

$$f(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}} \tag{10.15}$$

Logistic 函数的中文名称是逻辑斯谛函数，或简称逻辑函数。

第二，因变量 y_i 本身只取 0，1 两个离散值，不适于直接作为回归模型中的因变量。由于回归函数 $E(y_i)=\pi_i=\beta_0+\beta_1 x_i$ 表示在自变量为 x_i 的条件下 y_i 的平均值，而 y_i 是 0—1 型随机变量，因而 $E(y_i)=\pi_i$ 就是在自变量为 x_i 的条件下 y_i 等于 1 的比例。这提示我们可以用 y_i 等于 1 的比例代替 y_i 本身作为因变量。

下面通过一个例子来说明 Logistic 回归模型的应用。

例 10.4

在一次住房展销会上，与房地产商签订初步购房意向书的共有 $n=313$ 名顾客，在随后的 3 个月的时间内，只有一部分顾客确实购买了房屋。购买了房屋的顾客记为 1，没有购买房屋的顾客记为 0。以顾客的年家庭收入为自变量 x，对表 10.3 所示的数据，建立 Logistic 回归模型。

表 10.3

序号	年家庭收入（万元）x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i=m_i/n_i$	逻辑变换 $p'_i=\ln(\frac{p_i}{1-p_i})$	权重 $w_i=n_i p_i(1-p_i)$
1	1.5	25	8	0.320 000	−0.753 77	5.440
2	2.5	32	13	0.406 250	−0.379 49	7.719
3	3.5	58	26	0.448 276	−0.207 64	14.345

专业课辅导+真题资料+微信公众号：432应统书院

续前表

序号	年家庭收入（万元）x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i=m_i/n_i$	逻辑变换 $p_i'=\ln(\frac{p_i}{1-p_i})$	权重 $w_i=n_ip_i(1-p_i)$
4	4.5	52	22	0.423 077	−0.310 15	12.692
5	5.5	43	20	0.465 116	−0.139 76	10.698
6	6.5	39	22	0.564 103	0.257 829	9.590
7	7.5	28	16	0.571 429	0.287 682	6.857
8	8.5	21	12	0.571 429	0.287 682	5.143
9	9.5	15	10	0.666 667	0.693 147	3.333

Logistic 回归方程为

$$p_i=\frac{\exp(\beta_0+\beta_1x_i)}{1+\exp(\beta_0+\beta_1x_i)},\quad i=1,2,\cdots,c \tag{10.16}$$

其中，c 为分组数据的组数，本例中 $c=9$。将以上回归方程作线性化变换，令

$$p_i'=\ln(\frac{p_i}{1-p_i}) \tag{10.17}$$

（10.17）式的变换称为逻辑（logit）变换，变换后的线性回归模型为

$$p_i'=\beta_0+\beta_1x_i+\varepsilon_i \tag{10.18}$$

（10.18）式是一个普通的一元线性回归模型。（10.16）式没有给出误差项的形式，我们认为其误差项的形式就是作线性化变换所需要的形式。对表 10.3 的数据，算出经验回归方程为

$$\hat{p}'=-0.886+0.156x \tag{10.19}$$

决定系数 $r^2=0.924\,3$，显著性检验 P 值≈0，高度显著。将（10.19）式还原为（10.16）式的 Logistic 回归方程为

$$\hat{p}=\frac{\exp(-0.886+0.156x)}{1+\exp(-0.886+0.156x)} \tag{10.20}$$

利用（10.20）式可以对购房比例作预测，例如对 $x_0=8$ 可得

$$\hat{p}_0=\frac{\exp(-0.886+0.156\times8)}{1+\exp(-0.886+0.156\times8)}$$
$$=\frac{1.436}{1+1.436}=0.590$$

可知年收入 8 万元的家庭预计实际购房比例为 59%。

我们用 Logistic 回归模型成功地拟合了因变量为定性变量的回归模型，但是仍然存在一个不足之处，就是异方差性并没有得到解决，（10.18）式的回归模型不是等方差的，应该对（10.18）式用加权最小二乘估计。当 n_i 较大时，p_i'的近似方差为

专业课辅导+真题资料+微信公众号：432应统书院

$$D(p_i')\approx\frac{1}{n_i\pi_i(1-\pi_i)} \tag{10.21}$$

（10.21）式的证明请参见参考文献［6］。其中，$\pi_i=E(y_i)$，因而选取权数为

$$w_i=n_ip_i(1-p_i) \tag{10.22}$$

对例 10.4 重新用加权最小二乘作估计，计算结果见输出结果 10.5。

输出结果 10.5

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.939[a]	.881	.864	.3862

a. Predictors: (Constant), x.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.754	1	7.754	51.983	.000
	Residual	1.044	7	.149		
	Total	8.798	8			

c. Weighted Least Squares Regression-Weighted by W.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-.849	.114		-7.474	.000
x	.149	.021	.939	7.210	.000

用加权最小二乘法得到的 Logistic 回归方程为

$$\hat{p}=\frac{\exp(-0.849+0.149x)}{1+\exp(-0.849+0.149x)} \tag{10.23}$$

利用（10.23）式可以对购房比例作预测，例如对 $x_0=8$ 可得

$$\hat{p}_0=\frac{\exp(-0.849+0.149\times 8)}{1+\exp(-0.849+0.149\times 8)}$$

$$=\frac{1.409}{1+1.409}=0.585$$

可知年收入 8 万元的家庭预计实际购房比例为 58.5%，这个结果与未用加权法的结果很接近。

以上的例子是只有一个自变量的情况，分组数据的 Logistic 回归模型可以很方便地推广到有多个自变量的情况，在此就不举例说明了。

分组数据的 Logistic 回归只适用于样本量大的分组数据，对样本量小的未分组数据不适用，并且以组数 c 为回归拟合的样本量，拟合的精度低。实际上，我们可以用最大似然估计直接拟合未分组数据的 Logistic 回归模型，以下就介绍这个方法。

专业课辅导+真题资料+微信公众号：432应统书院

二、未分组数据的 Logistic 回归模型

设 y 是 0—1 型变量，x_1，x_2，…，x_p 是与 y 相关的确定性变量，n 组观测数据为（x_{i1}，x_{i2}，…，x_{ip}；y_i）（$i=1$，2，…，n），其中，y_1，y_2，…，y_n 是取值 0 或 1 的随机变量，y_i 与 x_{i1}，x_{i2}，…，x_{ip} 的关系如下：

$$E(y_i)=\pi_i=f(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})$$

其中，函数 $f(x)$ 是值域在［0，1］区间内的单调增函数。对于 Logistic 回归

$$f(x)=\frac{e^x}{1+e^x}$$

于是 y_i 是均值为 $\pi_i=f(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})$ 的 0—1 型分布，概率函数为

$$P(y_i=1)=\pi_i$$
$$P(y_i=0)=1-\pi_i$$

可以把 y_i 的概率函数合写为

$$P(y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i},\quad y_i=0,1;i=1,2,\cdots,n \tag{10.24}$$

于是，y_1，y_2，…，y_n 的似然函数为

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{10.25}$$

对似然函数取自然对数，得

$$\begin{aligned}\ln L&=\sum_{i=1}^{n}[y_i\ln\pi_i+(1-y_i)\ln(1-\pi_i)]\\&=\sum_{i=1}^{n}[y_i\ln\frac{\pi_i}{(1-\pi_i)}+\ln(1-\pi_i)]\end{aligned}$$

对于 Logistic 回归，将

$$\pi_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})}{1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})}$$

代入得

$$\begin{aligned}\ln L=\sum_{i=1}^{n}[&y_i(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\\&-\ln(1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}))]\end{aligned} \tag{10.26}$$

最大似然估计就是选取 β_0，β_1，β_2，…，β_p 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，使（10.26）式达到极大。求解过程需要用数值计算，SPSS 软件提供了求解功能。

专业课辅导+真题资料+微信公众号：432应统书院

例 10.5

在一次关于公共交通的社会调查中，一个调查项目为“是乘坐公交车上下班，还是骑自行车上下班”。因变量 $y=1$ 表示主要乘坐公交车上下班，$y=0$ 表示主要骑自行车上下班。自变量 x_1 是年龄，作为连续型变量；x_2 是月收入；x_3 是性别，$x_3=1$ 表示男性，$x_3=0$ 表示女性。调查对象为工薪层群体，数据见表 10.4，试建立 y 与自变量间的 Logistic 回归。

表 10.4

序号	x_3	x_1	x_2（元）	y	序号	x_3	x_1	x_2（元）	y
1	0	18	850	0	15	1	20	1 000	0
2	0	21	1 200	0	16	1	25	1 200	0
3	0	23	850	1	17	1	27	1 300	1
4	0	23	950	1	18	1	28	1 500	0
5	0	28	1 200	1	19	1	30	950	1
6	0	31	850	0	20	1	32	1 000	0
7	0	36	1 500	1	21	1	33	1 800	0
8	0	42	1 000	1	22	1	33	1 000	0
9	0	46	950	1	23	1	38	1 200	0
10	0	48	1 200	0	24	1	41	1 500	0
11	0	55	1 800	1	25	1	45	1 800	1
12	0	56	2 100	1	26	1	48	1 000	0
13	0	58	1 800	0	27	1	52	1 500	1
14	1	18	850	0	28	1	56	1 800	1

依次点击 SPSS 软件的 Analyze→Regression→Binary Logistic 命令，进入 Logistic 回归对话框，选入变量，点击 OK 运行，输出结果 10.6 是部分运行结果。

输出结果 10.6

Variable	B	S. E.	Wald	df	Sig	R	Exp(B)
SEX	-2. 5016	1. 1578	4. 6689	1	. 0307	-. 2627	. 0820
AGE	. 0822	. 0521	2. 4853	1	. 1149	. 1120	1. 0856
X2	. 0015	. 0019	. 6613	1	. 4161	. 0000	1. 0015
Constant	-3. 6547	2. 0911	3. 0545	1	. 0805		

输出结果 10.6 中，SEX（性别）、AGE（年龄）、X2（月收入）是三个自变量，S. E. 是 Std. Error 的简写，表示回归系数估计量 $\hat{\beta}$ 的标准差。Wald 是回归系数检验的统计量值

$$\text{Wald}=\left(\frac{\text{B}}{\text{S. E.}}\right)^2=\left[\frac{\hat{\beta}_j}{\sqrt{D(\hat{\beta}_j)}}\right]^2 \tag{10.27}$$

Sig 是 Wald 检验的显著性概率，R 是偏相关系数。可以看到，X2（月收入）不显著，决定将其剔除。用 y 对性别与年龄两个自变量作回归，得输出结果 10.7。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.7

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SEX	-2.2239	1.0476	4.5059	1	.0338	-.2546	.1082
AGE	.1023	.0458	4.9856	1	.0256	.2778	1.1077
Constant	-2.6285	1.5537	2.8620	1	.0907		

可以看到，SEX，AGE 两个自变量都是显著的，因而最终的回归方程为

$$\hat{p}=\frac{\exp(-2.6285-2.2239\text{SEX}+0.1023\text{AGE})}{1+\exp(-2.6285-2.2239\text{SEX}+0.1023\text{AGE})}$$

SPSS 软件没有给出 Logistic 回归的标准化回归系数，对于 Logistic 回归，回归系数也没有普通线性回归那样的解释，因而计算标准化回归系数并不重要。如果要考虑每个自变量在回归方程中的重要性，不妨直接比较 Wald 值（或 Sig 值），Wald 值大者（或 Sig 值小者）显著性高，也就更重要。当然这里假定自变量间没有强的多重共线性，否则回归系数的大小及其显著性概率都没有意义。

三、Probit 回归模型

Probit 回归称为单位概率回归，与 Logistic 回归相似，也是拟合 0—1 型因变量回归的方法，其回归函数是

$$\Phi^{-1}(\pi_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip} \tag{10.28}$$

用样本比例 p_i 代替概率 π_i，表示为样本回归模型

$$\Phi^{-1}(p_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+\varepsilon_i \tag{10.29}$$

例 10.6

使用例 10.4 的购房数据，首先计算出 $\Phi^{-1}(p_i)$ 的数值，见表 10.5，SPSS 和 Excel 软件都有计算 $\Phi^{-1}(\cdot)$ 函数的功能。以 $\Phi^{-1}(p_i)$ 为因变量，以年家庭收入 x 为自变量做普通最小二乘线性回归，得回归方程

$$\Phi^{-1}(\hat{p})=-0.552+0.0970x \tag{10.30}$$

或等价地表示为

$$\hat{p}=\Phi(-0.552+0.0970x) \tag{10.31}$$

对 $x_0=8$，$\hat{p}_0=\Phi(-0.552+0.0970\times8)=\Phi(0.224)=0.589$，与用 Logistic 回归计算的预测值很接近。

专业课辅导+真题资料+微信公众号：432应统书院

表 10.5

序号	年家庭收入（万元）x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i=m_i/n_i$	Probit 变换 $p_i'=\Phi^{-1}(p_i)$
1	1.5	25	8	0.320 000	−0.467 70
2	2.5	32	13	0.406 250	−0.237 20
3	3.5	58	26	0.448 276	−0.130 02
4	4.5	52	22	0.423 077	−0.194 03
5	5.5	43	20	0.465 116	−0.087 55
6	6.5	39	22	0.564 103	0.161 38
7	7.5	28	16	0.571 429	0.180 01
8	8.5	21	12	0.571 429	0.180 01
9	9.5	15	10	0.666 667	0.430 73

SPSS 软件提供了 Probit 回归功能，用于对分组数据拟合 Probit 回归。依照 Analyze→Regression→Probit 的顺序进入 Probit 回归对话框，把实际购房人数 m_i 选入 Response Frequency 框条中，把签订意向书人数 n_i 选入 Total Observed 框条中，把分组自变量年家庭收入 x 选入 Covariate 框条中，点击 OK 运行，得输出结果 10.8。

输出结果 10.8

```
********** PROBIT ANALYSIS **********
Parameter estimates converged after 8 iterations.
Optimal solution found.
Parameter Estimates (PROBIT model: (PROBIT (p)) = Intercept + BX):
          Regression   Coeff.  Standard Error    Coeff. /S. E.
               x       0.093 54     0.033 09        2.827 19
                     Intercept  Standard Error  Intercept/S. E.
                     -0.531 77     0.181 51       -2.929 79
Pearson Goodness-of-Fit Chi Square = 1.043 DF = 7 P = 0.994
Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity
factor is used in the calculation of confidence limits.
```

由输出结果 10.8 得回归方程

$$\Phi^{-1}(\hat{p})=-0.531\,77+0.093\,54x \tag{10.32}$$

SPSS 软件采用的是最大似然估计，用的是数值计算方法，与前面用最小二乘法所得到的回归方程（10.30）很接近。输出结果 10.8 中的拟合优度检验（goodness-of-fit）是对模型失拟的检验，原假设是回归模型正确，显著性概率 P 值 $=0.994>0.05$，不能拒绝原假设，说明回归模型正确。

在 SPSS 软件的 Probit 回归对话框中可以看到一个 Logit 选项，用这个选项可以对分组数据做 Logistic 回归。对此例计算出的 Logistic 回归方程是

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{p}'=-0.851\,78+0.149\,82x \tag{10.33}$$

这也是使用数值计算的最大似然估计，与用最小二乘法所得到的 Logistic 回归方程（10.19）式很接近。

10.5 多类别 Logistic 回归

当定性因变量 y 取 k 个类别时，记为 1，2，…，k，这里的数字 1，2，…，k 只是名义代号，并没有大小顺序的含义。因变量 y 取值于每个类别的概率与一组自变量 x_1，x_2，…，x_p 有关，对于样本数据（x_{i1}，x_{i2}，…，x_{ip}；y_i）($i=1$，2，…，n)，多类别 Logistic 回归模型第 i 组样本的因变量 y_i 取第 j 个类别的概率为

$$\pi_{ij}=\frac{\exp(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip})}{\exp(\beta_{01}+\beta_{11}x_{i1}+\cdots+\beta_{p1}x_{ip})+\cdots+\exp(\beta_{0k}+\beta_{1k}x_{i1}+\cdots+\beta_{pk}x_{ip})}$$

$$i=1,2,\cdots,n;\quad j=1,2,\cdots,k \tag{10.34}$$

上式中各回归系数不是唯一确定的，每个回归系数同时加减一个常数后 π_{ij} 的数值保持不变。为此，把分母的第一项 $\exp(\beta_{01}+\beta_{11}x_{i1}+\cdots+\beta_{p1}x_{ip})$ 中的系数都设为 0，得到回归函数的表达式

$$\pi_{ij}=\frac{\exp(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip})}{1+\exp(\beta_{02}+\beta_{12}x_{i1}+\cdots+\beta_{p2}x_{ip})+\cdots+\exp(\beta_{0k}+\beta_{1k}x_{i1}+\cdots+\beta_{pk}x_{ip})}$$

$$i=1,2,\cdots,n;\quad j=1,2,\cdots,k \tag{10.35}$$

这个表达式中每个回归系数都是唯一确定的，第一个类别的回归系数都取0，其他类别的回归系数数值的大小都以第一个类别为参照。

例 10.7

本例数据选自 SPSS 软件自带的数据文件 telco. sav，该文件在 \ SPSS \ tutorial \ sample _ files 文件夹内。一个电信商要分析顾客选择服务类别的影响因素，因变量是顾客类别（Customer category），变量名为 custcat，共取四个类别：

1="Basic service"；2="E-service"；3="Plus service"；4="Total service"

telco. sav 数据文件中包含了众多的变量，变量的详细信息可以在变量工作表中查找，数据的样本量 $n=1\,000$。

可以用 Edit 菜单中的 Options 选项的 General 选项卡选择显示变量标签。依照 Analyze→Regression→Multinomial Logistic 的顺序进入多类别 Logistic 回归对话框。

把因变量 Customer category [custcat] 选入 Dependent 框条中，这里 Customer category 是变量标签，custcat 是变量名称。注意变量数目很多，这个变量在变

专业课辅导+真题资料+微信公众号：432应统书院

量列表的后面。

把定性自变量 Marital status [marital]，Level of education [ed]，Retired [retire] 和 Gender [gender] 选入 factors 框条中。

把数值型自变量 Age in years [age]，Years at current address [address]，Household income in thousands [income]，Years with current employer [employ] 和 Number of people in household [reside] 选入 Covariates 框条中。

在因变量框条的下面有一个 Reference category 按钮，点击进入，选择以 First category 为参照类别，也就是选择（10.35）式的回归方程。点击 Continue 返回到多类别 Logistic 回归对话框，点击 OK 运行。

输出结果 10.9 是变量总结表，给出了分类变量各类别的频数和频率。因变量四个类别的频率分别为 26.6%，21.7%，28.1%和 23.6%。

输出结果 10.9

Case Processing Summary

		N	Marginal Percentage
Customer category	Basic service	266	26.6%
	E-service	217	21.7%
	Plus service	281	28.1%
	Total service	236	23.6%
Marital status	Unmarried	505	50.5%
	Married	495	49.5%
Level of education	Did not complete high school	204	20.4%
	High school degree	287	28.7%
	Some college	209	20.9%
	College degree	234	23.4%
	Post-undergraduate degree	66	6.6%
Retired	No	953	95.3%
	Yes	47	4.7%
Gender	Male	483	48.3%
	Female	517	51.7%
Valid		1000	100.0%
Missing		0	
Total		1000	
Subpopulation		1000	

输出结果 10.10 是用最大似然比方法做的变量显著性检验表，只有 Years at current address [address]，Years with current employer [employ]，Level of education [ed] 这三个自变量是显著的，其中对定性自变量的检验是整体检验。由于很多自变量不显著，下面改用逐步回归选择自变量。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.10

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	2507.081 (a)	0.000	0	.
age	2509.205	2.124	3	.547
address	2519.010	11.929	3	.008
income	2511.549	4.468	3	.215
employ	2518.361	11.280	3	.010
reside	2512.890	5.809	3	.121
marital	2510.650	3.569	3	.312
ed	2665.388	158.307	12	.000
retire	2509.926	2.845	3	.416
gender	2508.484	1.403	3	.705

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

重新进入多类别 Logistic 回归对话框，点击 Model 按钮进入模型界定对话框，选择 Custom/Stepwise（自定义/逐步回归）方法，然后把 9 个自变量逐一选入 Stepwise Terms 框条中。逐一选入是为了只包含变量的主效应，如果按住 Ctrl 键同时选中几个变量，则是考虑这几个变量的交互效应，本例不考虑交互效应。点击 Continue 返回多类别 Logistic 回归对话框。

点击 Statistics 按钮进入统计量对话框，除了默认选项之外再加上 Classification table 选项。点击 Continue 返回多类别 Logistic 回归对话框，点击 OK 运行。

输出结果 10.11 是逐步回归每步过程总结表，逐步回归进行了四步，依次选入了 ed，employ，reside 和 address 这四个自变量，没有变量被剔除。

输出结果 10.11

Step Summary

Model	Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
			-2 Log Likelihood	Chi-Square	df	Sig.
0	Entered	Intercept	2762.531	.		
1	Entered	ed	2588.305	174.226	12	.000
2	Entered	employ	2544.787	43.518	3	.000
3	Entered	reside	2531.282	13.505	3	.004
4	Entered	address	2519.615	11.667	3	.009

Stepwise Method: Forward Entry.

输出结果 10.12 是逐步回归最终模型的拟合优度信息表，给出了对回归模型有效性的检验结果。原假设是回归模型无效，所有回归系数都是 0，显著性概率(Sig)=0.000，拒绝原假设，认为回归模型显著有效。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.12

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	−2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2762.531			
Final	2519.615	242.916	21	0.000

输出结果 10.13 是逐步回归最终模型的变量显著性检验表，表中各变量的显著性概率（Sig）与输出结果 10.11 中有所不同，这是因为输出结果 10.11 是在每步回归模型下所做的检验，而输出结果 10.13 是在最终回归模型下所做的检验。

输出结果 10.13

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	−2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	2519.615(a)	0.000	0	.
ed	2693.134	173.519	12	.000
address	2531.282	11.667	3	.009
employ	2547.891	28.276	3	.000
reside	2534.909	15.294	3	.002

输出结果 10.14 是回归参数表。本例因变量共有 4 个类别，（10.35）式的回归函数中以第 1 个类别作为基准，回归系数取 0。对第 2，3，4 这 3 个类别，每个类别都需要确定一个线性回归函数，因此每个自变量都有 3 个回归系数，自由度合计为 3。其中每一个定性自变量要用其类别数减 1 个示性变量表示，例如 Level of education [ed] 变量有 5 个取值，要用 4 个示性变量表示，因此有 3×4=12 个回归系数，把 [ed] 变量作为一个整体时其自由度为 12。从输出结果 10.13 中看到，自变量 ed 的显著性概率（Sig）=0.000，这是以 ed 变量的 5 个取值作为一个整体做的检验，整体显著并不意味着每个取值都显著。例如从输出结果 10.14 中看到，对因变量的类别 2(E-service)，当 ed=4 时，Sig=0.129，就不显著。

输出结果 10.14

Parameter Estimates

Customer category (a)		B	Std. Error	Wald	df	Sig.	Exp(B)
E-service	Intercept	.049	.444	.012	1	.912	
	[ed=1]	−2.170	.464	21.862	1	.000	.114
	[ed=2]	−1.508	.439	11.783	1	.001	.221
	[ed=3]	−.971	.446	4.738	1	.030	.379
	[ed=4]	−.676	.446	2.301	1	.129	.508
	[ed=5]	0 (b)	.	.	0	.	.
	address	.037	.011	11.236	1	.001	1.038
	employ	.026	.012	5.111	1	.024	1.027
	reside	.148	.069	4.599	1	.032	1.159

专业课辅导+真题资料+微信公众号：432应统书院

Plus service	Intercept	-1.551	.631	6.043	1	.014	
	[ed=1]	.556	.629	.781	1	.377	1.743
	[ed=2]	.718	.624	1.325	1	.250	2.051
	[ed=3]	.673	.636	1.121	1	.290	1.960
	[ed=4]	.465	.646	.518	1	.472	1.592
	[ed=5]	0 (b)	.	.	0	.	.
	address	.022	.010	4.480	1	.034	1.022
	employ	.051	.010	24.404	1	0.000	1.053
	reside	.084	.065	1.670	1	.196	1.088
Total service	Intercept	.181	.431	.176	1	.675	
	[ed=1]	-3.762	.532	50.070	1	.000	.023
	[ed=2]	-1.959	.427	21.042	1	.000	.141
	[ed=3]	-1.453	.435	11.171	1	.001	.234
	[ed=4]	-.584	.425	1.893	1	.169	.557
	[ed=5]	0 (b)	.	.	0	.	.
	address	.022	.012	3.498	1	.061	1.022
	employ	.042	.012	12.437	1	.000	1.043
	reside	.258	.068	14.418	1	.000	1.294

a. The reference category is: Basic service.

b. This parameter is set to zero because it is redundant.

输出结果 10.15 是用逐步回归最终模型所得的预测结果。用（10.35）式对每个样品计算出因变量 y 取第 j 个类别的概率 π_j，因变量的预测值是 π_j 最大的类别。

输出结果 10.15

Classification

Observed	Predicted				
	Basic service	E-service	Plus service	Total service	Percent Correct
Basic service	122	8	75	61	45.9%
E-service	58	10	68	81	4.6%
Plus service	89	8	133	51	47.3%
Total service	47	12	43	134	56.8%
Overall Percentage	31.6%	3.8%	31.9%	32.7%	39.9%

从输出结果 10.15 中看到预测效果如下：

Basic service 类别的 266 个观测值中，有 122 个预测正确，正确率是 45.9%；

E-service 类别的 217 个观测值中，有 10 个预测正确，正确率是 4.6%；

Plus service 类别的 281 个观测值中，有 133 个预测正确，正确率是 47.3%；

Total service 类别的 236 个观测值中，有 134 个预测正确，正确率是 56.8%；

全部 1 000 个观测值中，有 399 个预测正确，总正确率是 39.9%。

如果没有任何信息资料，对每个类别的预测概率应该都是 1/4，预测的总正确率是 25%。如果对现有资料仅做频数分析，第 3 个类别“Plus service”出现 281 次，频率为 28.1%，出现的频率最高，对每位客户都预测为第 3 个类别“Plus

专业课辅导+真题资料+微信公众号：432应统书院

service”是最佳方案，预测的总正确率是 28.1%。现在通过回归分析，预测的总正确率是 39.9%，分别提高了 14.9%和 11.8%。

4 个类别的预测中，第 2 个类别 E-service 的预测效果最差，正确率仅是 4.6%，说明现有的自变量不能解释这类客户群的特性，应该对这一类别的客户群做进一步研究，找出有关的解释变量。

可以用 Save 按钮保存预测概率和预测值，表 10.6 是前 20 个样品的预测数值。从中看到，对客户 1，实际值是第 1 类，预测概率以 $\pi_1=0.38$ 最大，因此其预测值是第 4 类，是不正确的预测。对客户 2，实际值是第 4 类，预测概率以$\pi_4=0.60$ 最大，因此其预测值是第 4 类，是正确的预测。

表 10.6　　前 20 个样品的预测数值

序号	因变量观测值	因变量预测值	预测概率			
			π_1	π_2	π_3	π_4
1	1	4	0.22	0.26	0.14	0.38
2	4	4	0.07	0.28	0.04	0.60
3	3	3	0.19	0.16	0.61	0.05
4	1	1	0.43	0.18	0.26	0.12
5	3	1	0.47	0.15	0.33	0.05
6	3	3	0.25	0.19	0.39	0.16
7	2	1	0.31	0.18	0.27	0.24
8	4	3	0.31	0.18	0.32	0.19
9	3	4	0.06	0.18	0.17	0.59
10	2	3	0.23	0.17	0.54	0.06
11	1	4	0.25	0.26	0.15	0.34
12	3	3	0.28	0.19	0.38	0.16
13	1	3	0.30	0.19	0.32	0.19
14	4	4	0.06	0.29	0.17	0.47
15	1	1	0.50	0.11	0.35	0.03
16	2	4	0.11	0.34	0.05	0.51
17	3	3	0.10	0.24	0.49	0.18
18	3	1	0.36	0.17	0.29	0.18
19	1	1	0.58	0.10	0.29	0.03
20	4	4	0.21	0.23	0.11	0.45

10.6　因变量顺序数据的回归

当定性因变量 y 取 k 个顺序类别时，记为 1，2，…，k，这里的数字 1，2，…，k 仅表示顺序的大小。例如对居住状况分为非常不满意、不满意、一般、满意、非常满意 5 个顺序类别。因变量 y 取值于每个类别的概率仍与一组自变量 x_1，x_2，…，x_p有关，对于样本数据（x_{i1}，x_{i2}，…，x_{ip}；y_i）($i=1$，2，…，n)，顺序类别回归模型有两种主要类型，一种是位置结构（location component）模型；

专业课辅导+真题资料+微信公众号：432应统书院

另一种是规模结构（scale component）模型。

（1）位置结构模型：

$$\text{link}(\gamma_{ij})=\theta_j-(\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}) \tag{10.36}$$

其中，link(·)是联系函数；$\gamma_{ij}=\pi_{i1}+\cdots+\pi_{ij}$ 是第 i 个样品小于等于 j 的累积概率，由于 $\gamma_{ik}=1$，所以（10.36）式只针对 $i=1, 2, \cdots, n$；$j=1, 2, \cdots, k-1$。θ_j 是类别界限值（threshold）。

（2）规模结构模型：

$$\text{link}(\gamma_{ij})=\frac{\theta_j-(\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}{\exp(\tau_1 z_{i1}+\cdots+\tau_m z_{im})} \tag{10.37}$$

其中，$z_1, \cdots, z_m$ 是 $x_1, x_2, \cdots, x_p$ 的一个子集，作为规模结构解释变量。

联系函数的几种主要类型见表 10.7。

表 10.7

联系函数类型	形式	应用场合
Logit	$\log(\gamma/(1-\gamma))$	各类别均匀分布
Complementary log-log	$\log(-\log(1-\gamma))$	高层类别出现几率大
Negative log-log	$-\log(-\log(\gamma))$	低层类别出现几率大
Probit	$\Phi^{-1}(\gamma)$	正态分布
Cauchit (inverse Cauchy)	$\tan(\pi(\gamma-0.5))$	两端的类别出现几率大

还是使用 SPSS 软件自带的一个数据文件 german _ credit. sav 说明此方法。

例 10.8

一个信贷员想评估信贷业务的风险，选取客户的账户状态（Account status）作为因变量，有 5 个有序类别值，分别是 1——无债务历史；2——目前无债务；3——目前有正在偿还的债务；4——曾拖欠债款；5——危机的账户。解释变量由多个财务和个人资料变量构成。

依照 Analyze→Regression→Ordinal 的顺序进入有序数据回归对话框。

把因变量 Account status [chist] 选入 Dependent 框条中。

把定性自变量 # of existing credits [numcred]，Other installment debts [othnstal] 和 Housing [housng] 选入 factors 框条中。

把数值型自变量 Age in years [age] 和 Duration in months [duration] 选入 Covariates 框条中。

点击 Options 按钮选择 Complementary log-log 类型的联系函数，这是因为通过对因变量的频数分析发现类别 3 和 5 出现的频率很大，属于高层类别出现几率大的分布。各选项都取默认值，点击 OK 运行，输出结果 10.16 是参数估计表。

专业课辅导+真题资料+微信公众号：432应统书院

输出结果 10.16

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.
Threshold	[chist = 1.00]	-3.549	.667	28.323	1	.000
	[chist = 2.00]	-2.720	.656	17.167	1	.000
	[chist = 3.00]	-.137	.649	.044	1	.833
	[chist = 4.00]	.199	.649	.094	1	.759
Location	age	.015	.004	15.128	1	.000
	duration	-.002	.003	.379	1	.538
	[numcred = 1.00]	-1.134	.594	3.645	1	.056
	[numcred = 2.00]	.367	.598	.376	1	.540
	[numcred = 3.00]	.981	.711	1.902	1	.168
	[numcred = 4.00]	0 (a)	.	.	0	.
	[othnstal = 1.00]	-.397	.118	11.389	1	.001
	[othnstal = 2.00]	-.469	.193	5.913	1	.015
	[othnstal = 3.00]	0 (a)	.	.	0	.
	[housng = 1.00]	-.082	.165	.249	1	.617
	[housng = 2.00]	.132	.139	.897	1	.344
	[housng = 3.00]	0 (a)	.	.	0	.

Link function: Complementary Log-log.

a. This parameter is set to zero because it is redundant.

默认功能使用的是位置联系函数，SPSS 的有序回归中没有提供逐步回归的功能，需要根据各自变量的显著性情况决定是否剔除。从输出结果 10.16 中看到，数值型自变量 Duration in months [duration] 和定性自变量 Housing [housng] 的两个取值的 Sig 值都较大，是不显著的自变量，可以剔除。在确定最终的回归模型之后可以点击 Output 按钮保存预测变量。有序回归模块中没有提供预测值与观测值的交叉列表，在保存预测变量后可以自己计算这个表格，这些工作留给读者自己完成。

有序数据比类别数据含有更多的信息量，从理论上说有序数据因变量归的回应该比类别数据因变量的回归效果更好。但是从实际应用效果看，有序数据因变量的效果往往不尽如人意，其回归模型也正在研究和发展中。比较两者的回归函数（10.35）式和（10.36）式可以看到，类别因变量的（10.35）式对每一个类别分别建立一个线性回归方程，而顺序数据因变量的（10.36）式对因变量的不同取值建立了一个共同的线性回归方程，只是其界限值 θ_j（相当于常数项）不同。实际上，对顺序数据因变量也可以对因变量的不同取值分别建立回归方程，只是 SPSS 软件没有提供这个功能。在使用 SPSS 软件时，对有序数据因变量回归问题也可以尝试使用多类别数据因变量回归的方法，尤其是在样本量较大的场合。

10.7　本章小结与评注

在这一章我们主要介绍了自变量含定性变量和因变量是定性变量的两大类回归模型。

专业课辅导+真题资料+微信公众号：432应统书院

对于自变量含定性变量的回归模型，我们用两个例子介绍了这类问题的处理方法。也许有的读者会问，像例 10.1 的问题，为什么不对它分别拟合高学历家庭储蓄回归方程和低学历家庭储蓄回归方程，而是拟合带有一个虚拟变量的回归方程呢？这样做的原因有两个：一是因为模型假设对每类家庭具有相同的斜率和误差方差，把两类家庭放在一起可以对公共斜率 β_1 作出最佳估计；二是用带有一个虚拟变量的回归模型进行其他统计推断也会更加精确，这是因为均方误差的自由度更大。

推断统计中的单因素方差分析、无交互作用的双因素方差分析和有交互作用的双因素方差分析模型，都可以转化为 0—1 型自变量的回归分析模型。以单因素方差分析为例，设 $y_{ij}(i=1, 2, \cdots, n_j)$ 是正态总体 $N(\mu_j, \sigma^2)(j=1, 2, \cdots, c)$ 的样本，原假设为

$$H_0: \mu_1=\mu_2=\cdots=\mu_c \tag{10.38}$$

记 $\varepsilon_{ij}=y_{ij}-\mu_j$，则有 $\varepsilon_{ij}\sim N(0, \sigma^2)$，进而有

$$y_{ij}=\mu_j+\varepsilon_{ij}, \quad i=1,2,\cdots,n_j; j=1,2,\cdots,c \tag{10.39}$$

记 $\mu=\frac{1}{c}\sum_{j=1}^{c}u_j$，$a_j=\mu_j-\mu$，则（10.39）式改写为

$$y_{ij}=\mu+a_j+\varepsilon_{ij}, \quad i=1,2,\cdots,n_j; j=1,2,\cdots,c \tag{10.40}$$

引入 0—1 型自变量 x_{ij}，将（10.40）式表示为

$$y_{ij}=\mu+a_1x_{i1}+a_2x_{i2}+\cdots+a_cx_{ic}+\varepsilon_{ij}, i=1,2,\cdots,n_j; j=1,2,\cdots,c \tag{10.41}$$

其中，

$$\begin{cases} x_{i1}=1, \text{当 } j=1 \\ x_{i1}=0, \text{当 } j\neq 1 \end{cases}$$

$$\begin{cases} x_{i2}=1, \text{当 } j=2 \\ x_{i2}=0, \text{当 } j\neq 2 \end{cases}$$

……

$$\begin{cases} x_{ic}=1, \text{当 } j=c \\ x_{ic}=0, \text{当 } j\neq c \end{cases}$$

（10.41）式即我们熟悉的多元线性回归模型。但是其中还存在一个问题，就是 c 个自变量 x_1，x_2，…，x_c 之和恒等于 1，存在完全的多重共线性。为此，剔除 x_c，建立回归模型

$$y_{ij}=\mu+a_1x_{i1}+a_2x_{i2}+\cdots+a_{c-1}x_{ic-1}+\varepsilon_{ij}$$
$$i=1,2,\cdots,n_j; j=1,2,\cdots,c \tag{10.42}$$

（10.42）式回归方程显著性检验的原假设为

$$H_0: a_1=a_2=\cdots=a_{c-1}=0 \tag{10.43}$$

专业课辅导+真题资料+微信公众号：432应统书院

由 $a_j=\mu_j-\mu=\mu_j-\frac{1}{c}\sum_{j=1}^{c}\mu_j$ 可知，(10.38) 式与 (10.43) 式的两个原假设是等价的。作 (10.43) 式的显著性 F 检验，这个检验与单因素方差分析的 F 检验是等价的。

对于无交互作用的双因素方差分析和有交互作用的双因素方差分析模型，也可以用类似的方法转化为 0—1 型自变量的回归分析模型，在此就不多作介绍了。

如果所建立的回归模型，其中的自变量全是定性变量，我们称这样的回归模型为方差分析模型；如果模型中既包含数量变量，又包含定性变量，其中以定性自变量为主，称这样的模型为协方差模型。例 10.1 实际上就是一个协方差模型，对这些模型有兴趣的读者请参见参考文献 [6]。

分组数据的 Logistic 回归首先要对频率作逻辑变换，变换公式为 $p'_i=\ln\left(\frac{p_i}{1-p_i}\right)$，这个变换要求 $p_i=\frac{m_i}{n_i}\neq 0$ 或 1，即要求 $m_i\neq 0$，$m_i\neq n_i$。当 $m_i=0$ 或 $m_i=n_i$ 时，可以用如下的修正公式计算样本频率

$$p_i=\frac{m_i+0.5}{n_i+1} \tag{10.44}$$

分组数据的 Logistic 回归存在异方差性，需要采用加权最小二乘估计。除了 (10.22) 式给出的权函数 $w_i=n_ip_i(1-p_i)$ 之外，也可以通过两阶段最小二乘法确定权函数：

第一阶段是用普通最小二乘拟合回归模型。

第二阶段是从第一阶段的结果估计出组比例 $\hat{p}_i$，用权数 $w_i=n_i\hat{p}_i(1-\hat{p}_i)$ 作加权最小二乘回归（见参考文献 [3]）。

因变量是定性变量的情况有着广泛的应用，这种情况属于广义线性模型（generalized linear model，GLM）的研究范畴。GLM 的内容很广泛，其基本内容是假定因变量分布中的某个参数与一组自变量有关。例如，以 y 表示产品的缺陷数，x_1，x_2，…，x_p 是与 y 相关的变量，假定 y 服从泊松分布 $P(\mu)$，$\mu=E(y)>0$，$\ln\mu$ 的取值范围是整个实轴，可以建立 $\ln\mu$ 对 x_1，x_2，…，x_p 的线性回归模型

$$\ln\mu=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon$$

这就是所谓对数线性模型的一个例子。

Logistic 回归的应用非常广泛。我们将 Logistic 回归建模方法用于标准化试题的评价也得到很有意义的结果，详见参考文献 [12]。

□ 思考与练习

10.1　一个学生使用含有季节定性自变量的回归模型，对春、夏、秋、冬四个季节引入四个 0—1 型自变量，用 SPSS 软件计算的结果中总是自动剔除了其中的一

专业课辅导+真题资料+微信公众号：432应统书院

个自变量，他为此感到困惑不解。出现这种情况的原因是什么？

10.2 对自变量中含定性变量的问题，为什么不对同一属性分别建立回归模型，而采取设虚拟变量的方法建立回归模型？

10.3 研究者想研究采取某项保险革新措施的速度 y 与保险公司的规模 x_1 和保险公司类型的关系（参见参考文献［3］）。因变量的计量是第一个公司采纳这项革新和给定公司采纳这项革新在时间上先后间隔的月数。第一个自变量公司的规模是数量型的，用公司的总资产额（百万美元）来计量；第二个自变量公司的类型是定性变量，由两种类型构成，即股份公司和互助公司。数据资料见表 10.8，试建立 y 对公司规模和公司类型的回归。

表 10.8

i	y	x_1	公司类型
1	17	151	互助
2	26	92	互助
3	21	175	互助
4	30	31	互助
5	22	104	互助
6	0	277	互助
7	12	210	互助
8	19	120	互助
9	4	290	互助
10	16	238	互助
11	28	164	股份
12	15	272	股份
13	11	295	股份
14	38	68	股份
15	31	85	股份
16	21	224	股份
17	20	166	股份
18	13	305	股份
19	30	124	股份
20	14	246	股份

10.4 表 10.9 的数据是我国历年铁路里程数据，根据散点图观察在 1995 年（$t=16$）有折点，用折线回归拟合这些数据。

表 10.9 **我国历年铁路里程数据** 单位：万公里

年份	t	y	年份	t	y
1980	1	5.33	1985	6	5.50
1981	2	5.39	1986	7	5.57
1982	3	5.29	1987	8	5.58
1983	4	5.41	1988	9	5.61
1984	5	5.45	1989	10	5.69

专业课辅导+真题资料+微信公众号：432应统书院

续前表

年份	t	y	年份	t	y
1990	11	5.78	1998	19	6.64
1991	12	5.78	1999	20	6.74
1992	13	5.81	2000	21	6.87
1993	14	5.86	2001	22	7.01
1994	15	5.90	2002	23	7.19
1995	16	5.97	2003	24	7.30
1996	17	6.49	2004	25	7.44
1997	18	6.60			

10.5　某省统计局 1990 年 9 月在全省范围内进行了一次公众安全感问卷调查，参考文献［10］选取了调查表中的一个问题进行分析。本题对其中的数据作了适当的合并。对 1 391 人填写的问卷统计“一人在家是否害怕生人来”。因变量 $y=1$ 表示害怕，$y=0$ 表示不害怕。两个自变量：x_1 是年龄，x_2 是文化程度。各变量的取值含义如表 10.10 所示。

表 10.10

是否害怕　y	年龄 x_1	文化程度 x_2
害怕　1	16～28 岁　22	文盲　0
不害怕　0	29～45 岁　37	小学　1
	46～60 岁　53	中学　2
	61 岁及以上　68	中专及以上　3

现在的问题是：公民一人在家害怕生人来这个事件，与公民的年龄 x_1、文化程度 x_2 有没有关系？调查数据见表 10.11。

表 10.11

序号	x_1	x_2	n_i	$y=1$	$y=0$	p_i
1	22	0	3	0	3	0.125 00
2	22	1	11	3	8	0.291 67
3	22	2	389	146	243	0.375 64
4	22	3	83	26	57	0.315 48
5	37	0	4	3	1	0.700 00
6	37	1	27	18	9	0.660 71
7	37	2	487	196	291	0.402 66
8	37	3	103	27	76	0.264 42
9	53	0	9	4	5	0.450 00
10	53	1	6	3	3	0.500 00
11	53	2	188	73	115	0.388 89
12	53	3	47	18	29	0.385 42
13	68	0	2	0	2	0.166 67
14	68	1	10	3	7	0.318 18
15	68	2	18	7	11	0.394 74
16	68	3	4	0	4	0.100 00

专业课辅导+真题资料+微信公众号：432应统书院

其中，p_i 是根据（10.44）式计算的。

（1）把公民的年龄 x_1、文化程度 x_2 作为数量型变量，建立 y 对 x_1 和 x_2 的 Logistic 回归。

（2）把公民的年龄 x_1、文化程度 x_2 作为定性变量，用 0—1 型变量将其数量化，建立 y 对公民的年龄和文化程度的 Logistic 回归。

（3）你对回归的效果是否满意？如果不满意，你认为主要的问题是什么？

10.6 研制一种新型玻璃，对其做耐冲击试验。用一个小球从不同的高度 h 对玻璃做自由落体撞击，玻璃破碎记 $y=1$，玻璃未破碎记 $y=0$。试对表 10.12 的数据建立玻璃耐冲击性对高度 h 的 Logistic 回归，并解释回归方程的含义。

表 10.12

序号	$h(m)$	y	序号	$h(m)$	y
1	1.50	0	14	1.76	1
2	1.52	0	15	1.78	0
3	1.54	0	16	1.80	1
4	1.56	0	17	1.82	0
5	1.58	1	18	1.84	0
6	1.60	0	19	1.86	1
7	1.62	0	20	1.88	1
8	1.64	0	21	1.90	0
9	1.66	0	22	1.92	1
10	1.68	1	23	1.94	0
11	1.70	0	24	1.96	1
12	1.72	0	25	1.98	1
13	1.74	0	26	2.00	1

10.7 数据用 SPSS 软件自带的数据文件 bankloan.sav，该文件在 \ SPSS \ tutorial \ sample_files 文件夹内。一家银行研究客户拖欠贷款问题，因变量是客户是否曾经拖欠贷款 Previously defaulted [default]，0="No"，1="Yes"。数据文件中共有 850 条记录，其中前 700 条记录是过去客户的资料，作为回归的样本。后 150 条记录是潜在客户的资料，希望用回归预测其拖欠贷款倾向。建立两类别 Logistic 回归，定性自变量是 Level of education [ed]，用 Categorical 按钮指定；数值型自变量是 Age in years [age]，Years with current employer [employ]，Years at current address [address]，Household income in thousands [income]，Debt to income ratio [debtinc]，Credit card debt in thousands [creddebt] 和 Other debt in thousands [othdebt]。

10.8 用 SPSS 软件自带的数据文件 cereal.sav 做多类别 Logistic 回归。这个数据资料来源是某快餐公司抽选了 880 名顾客品尝公司的 3 种早餐套餐，分别是 1——Breakfast Bar，2——Oatmeal，3——Cereal。每位顾客从中确定自己最喜欢的套餐，公司记录下顾客的年龄、性别、婚姻状况、健身运动状况。以 Preferred breakfast [bfast] 为因变量，以定性变量 Age category [agecat]，Gender [gen-

专业课辅导+真题资料+微信公众号：432应统书院

der]，Marital status [marital]，Lifestyle [active] 为自变量做统计分析。

10.9　对例 10.7，根据输出结果 10.14 的参数估计表，手工计算出表 10.6 中前 2 个样品的预测概率。前 2 个样品的变量取值是

样品号	ed	address	employ	reside
1	4	9	5	2
2	5	7	5	6

10.10　某学校对本科毕业学生的去向做了一个调查，分析影响毕业去向的相关因素，结果见表 10.13，其中毕业去向“1”＝工作，“2”＝读研，“3”＝出国留学。性别“1”＝男生，“0”＝女生。用多类别 Logistic 回归分析影响毕业去向的因素。

表 10.13

序号	专业课 x_1	英语 x_2	性别 x_3	月生活费 x_4	毕业去向 y
1	95	65.0	1	600	2
2	63	62.0	0	850	1
3	82	53.0	0	700	2
4	60	88.0	0	850	3
5	72	65.0	1	750	1
6	85	85.0	0	1 000	3
7	95	95.0	0	1 200	2
8	92	92.0	1	950	2
9	63	63.0	0	850	1
10	78	75.0	1	900	1
11	90	78.0	0	500	1
12	82	83.0	1	750	2
13	80	65.0	1	850	3
14	83	75.0	0	600	2
15	60	90.0	0	650	3
16	75	90.0	1	800	2
17	63	83.0	1	700	1
18	85	75.0	0	750	2
19	73	86.0	0	950	2
20	86	66.0	1	1 500	3
21	93	63.0	0	1 300	2
22	73	72.0	0	850	1
23	86	60.0	1	950	2
24	76	63.0	0	1 100	1
25	96	86.0	0	750	2
26	71	75.0	1	1 000	1
27	63	72.0	1	850	2
28	60	88.0	0	650	1
29	67	95.0	1	500	1

专业课辅导+真题资料+微信公众号：432应统书院

续前表

序号	专业课 x_1	英语 x_2	性别 x_3	月生活费 x_4	毕业去向 y
30	86	93.0	0	550	1
31	63	76.0	0	650	1
32	86	86.0	0	750	2
33	76	85.0	1	650	1
34	82	92.0	1	950	3
35	73	60.0	0	800	1
36	82	85.0	1	750	2
37	75	75.0	0	750	1
38	72	63.0	1	650	1
39	81	88.0	0	850	3
40	92	96.0	1	950	2

10.11　对例 10.8 信贷风险数据，剔除掉 Housing [housng] 和 Duration in months [duration] 两个自变量后重新做回归，分析拟合优度、自变量的显著性，保存预测值，用 SPSS 的 Descriptive Statistics 中的 Crosstabs 做预测值与观测值的交叉列表，分析预测效果。

10.12　对上面的习题 10.11，手工计算前 2 个样品的预测概率。前 2 个样品的变量取值是

样品号	age	numcred	othnstal	因变量 chist
1	67	2	3	5
2	22	1	3	3

专业课辅导+真题资料+微信公众号：432应统书院

部分练习题参考答案

第 2 章

2.2 $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

2.7 提示：

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})$$
$$= \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} e_i = 0$$

2.9 提示：

$$\begin{aligned}
\mathrm{var}(e_i) &= \mathrm{var}(y_i - \hat{y}_i) \\
&= \mathrm{var}(y_i) + \mathrm{var}(\hat{y}_i) - 2\mathrm{cov}(y_i, \hat{y}_i) \\
&= \mathrm{var}(y_i) + \mathrm{var}(\hat{\beta}_0 + \hat{\beta}_1 x_i) - 2\mathrm{cov}(y_i, \bar{y} + \hat{\beta}_1(x_i - \bar{x})) \\
&= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}\right] - 2\sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}\right] \\
&= \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right]\sigma^2
\end{aligned}$$

2.14 $\hat{y} = -1 + 7x$，$x_0 = 4.2$ 时 $y_0 = 28.4$。y_0 置信水平 95%的区间估计是 (6.06，50.7)，$E(y_0)$ 置信水平 95%的区间估计是 (17.1，39.7)。本例样本量 $n=5$ 很小，所以区间估计的误差较大。

2.15 $\hat{y} = 0.118 + 0.003\,59x$，$x_0 = 1\,000$ 时 $y_0 = 3.7$。y_0 置信水平 95%的区间估计是 (2.5，4.9)，y_0 置信水平 95%的近似区间估计是 (2.7，4.7)。$E(y_0)$

专业课辅导+真题资料+微信公众号：432应统书院

置信水平 95%的区间估计是（3.3，4.1）。

2.16 （1）散点图（略），可以用直线反映两变量之间的关系。

（2）$\hat{y}=12\,112+3.314x$

（3）从残差的直方图看，略呈右偏；从正态概率图看，散点基本呈直线趋势，可以认为残差服从正态分布。

第 3 章

3.11 x_3的 P 值=0.284，不显著，予以剔除。

$\hat{y}_0=267.8$，y_0 置信水平 95%的区间估计是（204.4，331.2）。y_0 置信水平 95%的近似区间估计是（219.6，316.0），本例样本量 $n=10$ 较小，所以近似区间估计的误差较大。$E(y_0)$ 置信水平 95%的区间估计是（240.0，295.7）。

3.12 x_1的回归系数 0.607 明显不合理。

第 4 章

4.9 （1）普通最小二乘 $\hat{y}=-0.831+0.003\,68x$，$R^2=0.705$，残差图略。

（2）$|e_i|$ 与 x_i的等级相关系数=0.381，P 值=0.021，存在异方差。

（3）加权最小二乘幂指数 m 的最优取值为 $m=1.5$，得

$$\hat{y}_w=-0.683+0.003\,56x$$

计算出加权变换残差 $e'_{iw}=\sqrt{w_i}\cdot e_{iw}$，绘制加权变换残差图（略），$|e'_{iw}|$与 x_i 的等级相关系数=−0.076，P 值=0.591，说明异方差已经消除。但是加权最小二乘的 $R^2=0.659$，小于普通最小二乘的 $R^2=0.705$，说明加权最小二乘的效果不好。

（4）对因变量做变换 $y'=\sqrt{y}$，得回归方程 $\hat{y}'=0.000\,952\,9+0.582\,2x$，保存预测值 $\hat{y}'_i$，将其平方得到因变量 y_i 的预测值，进而计算出残差。等级相关系数=0.052，P 值=0.710，说明异方差已经消除。用公式 $R^2=1-\mathrm{SSE/SST}$ 计算出$R^2=0.710$，优于普通最小二乘的效果。

4.13 （1）普通最小二乘 $\hat{\sigma}=0.106$。

（2）普通最小二乘 $\mathrm{DW}=0.771<d_L=1.120$，存在正的序列相关。

（3）迭代法，$\hat{\rho}=1-0.710/2=0.645$，$\hat{\sigma}_u=0.077$，$\mathrm{DW}=1.48>d_U=1.41$，消除了序列相关，得方程 $\hat{y}'=-2.01+0.170x'$，还原为原始变量方程

$$\hat{y}_t=-2.01+0.645y_{t-1}+0.170(x_t-x_{t-1})$$

（4）差分法，$\hat{\sigma}_u=0.079$，$\mathrm{DW}=1.44>d_U=1.41$，消除了序列相关，得方程 $\Delta\hat{y}=0.166\Delta x$，还原为原始变量方程

$$\hat{y}_t=y_{t-1}+0.166(x_t-x_{t-1})$$

（5）迭代法最好。

4.14 （1）普通最小二乘 $\hat{\sigma}=329.69$，$\mathrm{DW}=0.745<d_L=1.50$，存在正的序

专业课辅导+真题资料+微信公众号：432应统书院

列相关。各种自回归方法主要结果见下表。

自回归方法	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}'_0$	$\hat{\beta}_1=\hat{\beta}'_1$	$\hat{\beta}_2=\hat{\beta}'_2$	DW	$\hat{\sigma}_u$
迭代法	0.627	—	−179.0	211.1	1.437	1.716	257.8
差分法	—	—	0	210.1	1.397	2.040	281.0

第 5 章

5.9　后退法依次剔除 x_4，x_3，x_6，保留 x_1，x_2，x_5作为最终模型。逐步回归法依次将 x_5，x_1，x_2引入回归模型，两者的最终模型是

$$\hat{y}=874.6-0.611x_1-0.353x_2+0.637x_5$$

但是回归系数的解释不合理。

5.10　(1) 略。

(2) 后退法剔除 x_5，保留 x_2，x_3，x_4，x_6作为最终模型。

(3) 逐步回归法依次将 x_3，x_5，x_4引入回归模型，没有剔除变量，保留 x_3，x_5，x_4作为最终模型。

(4) 两个方法得到的最终模型是不同的，后退法首先剔除了 x_5，而逐步回归法在第 2 步引入 x_5，说明两种方法对自变量重要性的认可是不同的，这与自变量之间的相关性有关联。相比之下，后退法首先做全模型的回归，每个自变量都有机会展示自己的作用，所得结果更值得信服。从本例内容看，x_5是滞后 6 个月的最惠利率，对因变量的影响似乎不大。

第 6 章

6.6　方差扩大因子 $\mathrm{VIF}_2=2\,636$，条件数 $\lambda_7=288$，说明存在严重的多重共线性。先剔除方差扩大因子最大的 x_2，重新做回归；再剔除方差扩大因子最大的 x_5，重新做回归；再剔除方差扩大因子最大的 x_1，而这三个自变量正是后退法和逐步回归法所保留的变量，可见按照共线性剔除变量与常规的后退法和逐步回归法按照 t 值显著性剔除变量会有较大的差别。重新做回归，此时共线性已经消除，再剔除不显著的 x_6，仅保留了 x_3和 x_4两个自变量，其中 x_4的 P 值=0.076，表示 x_4只有较弱的显著性。

第 7 章

7.5　选取岭参数 $k=0.01$，得岭回归方程

$$\hat{y}=750.0+0.055\,26x_1+0.081\,4x_2+0.100\,9x_5$$

回归系数都能有合理解释。

7.6　普通最小二乘 $\hat{y}=4\,352.8+1.438x_2+0.679x_3$。其中，回归系数 $\hat{\beta}_3=0.679$ 明显不合理。

岭参数 $k=0.01$ 时，岭回归方程为 $\hat{y}=3\,980.2+1.091x_2+1.227x_3$。

专业课辅导+真题资料+微信公众号：432应统书院

7.7 采用后退法和逐步回归法，得回归方程 $\hat{y}=-0.443+0.050x_1-0.032x_4$，其中，$x_4$的系数是负数不合理，说明仍然存在共线性。

选取岭参数 $k=0.4$，得岭回归方程 $\hat{y}=0.357+0.025\,8x_1+0.045\,3x_4$，回归系数都能有合理解释。

用 y 对 x_1，x_2，x_3做岭回归，选取岭参数 $k=0.4$，得输出结果：

	B	SE (B)	Beta	B/SE (B)
x1	.016 739 073	.003 359 156	.372 627 316	4.983 118 685
x2	.156 806 656	.047 550 034	.275 213 878	3.297 719 120
x3	.067 110 931	.032 703 990	.159 221 005	2.052 071 673
Constant	-.819 486 727	.754 456 246	.000 000 000	-1.086 195 166

岭回归方程 $\hat{y}=-0.819+0.016\,7x_1+0.157x_2+0.067\,1x_3$，回归系数都能有合理解释。表中 B/SE(B) 是近似的 t 值，$t_1=4.983$，$t_2=3.298$，说明 x_1 和 x_2 都是显著的，$t_3=2.052$，说明 x_3 也是比较显著的，所以做 y 对 x_1，x_2，x_3 的岭回归是可行的。

第 8 章

8.3 输出结果

Variance **Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.236	55.893	55.893	2.236	55.893	55.893
2	1.576	39.402	95.294	1.576	39.402	95.294
3	.187	4.665	99.959	.187	4.665	99.959
4	.002	.041	100.000	.002	.041	100.000

on

Extraction Method: Principal Component Analysis.

Factor1	Factor2	Factor3	Factor4
-0.981 28	-1.515 86	-1.226 91	0.956 18
-1.428 43	-0.189 86	-0.671 76	-0.740 35
0.755 65	-0.146 47	-0.024 8	-2.325 33
-0.441 33	-1.255 98	0.414 84	-0.821 83
0.239 94	-0.385 16	-1.713 33	0.476 16
0.646 48	-0.135 37	0.198 39	-0.301 94
0.622 45	1.700 49	-0.400 45	0.205 86
-1.492 84	0.550 95	1.064 22	0.561
-0.235 09	1.140 85	-0.073 07	-1.116 44
1.111 9	-1.456 17	1.970 45	0.492 28
-1.096 94	1.031 62	1.143 99	0.778 97
1.132	0.312 45	-0.045 86	0.922 81
1.1675	0.348 51	-0.635 71	0.912 64

现在用 y 对前两个主成分 Factor1 和 Factor2 做普通最小二乘回归，得主成分回归的回归方程

专业课辅导+真题资料+微信公众号：432应统书院

$$\hat{y}=95.423+12.777\text{Factor1}+0.157\text{Factor2}$$

分别以两个主成分 Factor1 和 Factor2 做因变量，以四个原始变量为自变量做线性回归，所得的回归系数就是所需要的线性组合的系数。

$$\text{Factor1}=-0.43+0.054x_1+0.024x_2-0.041x_3-0.022x_4$$

$$\text{Factor2}=-0.747-0.069x_1+0.021x_2+0.075x_3-0.021x_4$$

还原后的主成分回归方程为

$$\hat{y}=89.812+0.679x_1+0.31x_2-0.512x_3-0.284x_4$$

逐步回归法得到的回归方程为

$$\hat{y}=52.577+1.468x_1+0.662x_2$$

回归系数对因变量的解释较为合理，然而自变量个数本身就比较少，只有四个，这里舍弃掉了一半。两种方法的主要区别在于：普通最小二乘法认为自变量对因变量直接起作用，故要剔除对因变量作用不大的自变量；而主成分回归方程则是寻找影响自变量的主要因子，关注这些因子对因变量的作用。两种方法的选择应当结合实际研究问题，从实际情况出发，从而选取更优的解决方案。

8.4　普通最小二乘得到的回归方程为

$$\hat{y}=62.405+1.551x_1+0.510x_2+0.102x_3-0.144x_4$$

从系数上看可以发现明显的不合理的地方，x_3 与 y 是负相关的，但它的系数却是正的。

偏最小二乘的结果：

当迭代到第三次时，R-square 开始下降，因此我们只迭代两次：

Proportion of Variance Explained

Latent Factors		Statistics				
		X Variance	Cumulative X Variance	Y Variance	Cumulative Y Variance (R-square)	Adjusted R-square
dimension0	1	.559	.559	.968	.968	.965
	2	.062	.621	.014	.982	.978

Weights

Variables		Latent Factors	
		1	2
dimension 0	x1	.497	.637
	x2	.555	-.277
	x3	-.364	.686
	x4	-.559	-.218
	y	.658	.267

得到的结果为：

$$t_1=0.497x_1^*+0.555x_2^*-0.364x_3^*-0.559x_4^*$$

专业课辅导+真题资料+微信公众号：432应统书院

$$t_2=0.637x_1^*-0.277x_2^*+0.686x_3^*-0.218x_4^*$$

$$y^*=0.658t_1+0.267t_2$$

将 t 代入 y^* 中，得

$$y^*=0.498x_1^*+0.292x_2^*-0.056x_3^*-0.426x_4^*$$

还原为原始变量：

$$\hat{y}=85.402+1.272x_1+0.282x_2-0.132x_3-0.383x_4$$

从系数上看，x_1，x_2 对 y 起正影响，x_3，x_4 对 y 起负影响，与相关分析得到的结果一致，因此偏最小二乘回归系数的解释比普通最小二乘更加合理，又比逐步回归保留了更多的自变量。

第 9 章

9.2 选取二次曲线，得 $\hat{y}=5.843-0.008\,7x+4.47\times10^{-7}x^2$，也可以使用指数曲线。

9.3 (1) 乘性误差项 $\alpha=0.021$，$\beta=6.08$

(2) 加性误差项 $\alpha=0.021$，$\beta=6.06$

9.4 (1) $b_0=0.157$，$b_1=0.768$

(2) $u=91.1$，$b_0=0.211$，$b_1=0.727$

9.5 (1) 线性化模型 $\alpha=0.801$，$\beta=0.404$

(2) 非线性化模型 $\alpha=0.868$，$\beta=0.270$

(3) DW=0.715，存在自相关，用 SPSS 的自回归方法得

精确最大似然 $\alpha=0.706$，$\beta=0.688$

(4) VIF=13.0，取岭回归参数=0.20，得岭回归估计 $\alpha=0.479$，$\beta=1.127$。

9.6 (1) 线性化模型 $\mu=0.042$，$\alpha=0.460$，$\beta=-0.028$

(2) 非线性化模型 $\mu=0.046$，$\alpha=0.395$，$\beta=0.002$

(3) DW=1.285，自相关现象不严重。

(4) 最大的 VIF 值是 55.6，取岭回归参数=0.20，得岭回归估计 $\mu=0.027$，$\alpha=0.323$，$\beta=0.724$。

第 10 章

10.3 $\hat{y}=33.874-0.101\,7x_1+8.055x_2$

10.4 设 $x=\begin{cases}0, & t\leqslant16\\ t-16, & t>16\end{cases}$

得 $\hat{y}=5.183+0.055t+0.106x$，回归系数都显著非零，折线回归成立。

10.5 (1) 未加权回归：回归方程 F 检验的显著性概率（Sig）=0.687，回归方程不显著。x_1，x_2两个自变量的 Sig 分别是 0.619，0.487，也不显著。

专业课辅导+真题资料+微信公众号：432应统书院

加权回归：回归方程 F 检验的显著性概率（Sig）=0.037，回归方程显著。回归系数表如下：

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.146	.309		.472	.645
x1	.002	.005	.086	.398	.697
x2	-.331	.116	-.617	-2.858	.013

x_2 显著，其回归系数为 -0.331，表明文化程度越高越不害怕。而 x_1 不显著，应该剔除。

（2）用 3 个 0—1 型变量表示年龄 x_1。未加权回归：F 检验的显著性概率（Sig）=0.106，再用后退法选择变量，最后只保留了 x_{12}（年龄 29～45 岁组），x_2 被剔除，回归效果不好。

加权回归：用后退法选择变量，最后只保留了 x_2。

（3）对回归的效果不满意，主要问题是对年龄 x_1 是否显著的判定，如果能获得年龄的准确值做 Logistic 回归的最大似然估计，可能会改进回归效果。

10.6　Logistic 回归方程 $\hat{p}=\dfrac{\exp(-14.59+7.981h)}{1+\exp(-14.59+7.981h)}$

10.7　用后退法选择自变量，依次剔除不显著的自变量 ed，othdebt，income 和 age，这个过程可以借助软件提供的后退法完成。700 个观测值预测的效果见下表：

Classification Table[a]

Observed		Predicted		
		Previously defaulted		Percentage Correct（%）
		No	Yes	
Previously defaulted	No	478	39	92.5
	Yes	91	92	50.3
Overall Percentage				81.4

a. The cut value is .500.

10.8　剔除不显著的自变量 gender，点击 Statistics 进入统计量对话框，加上 Classification table 选项，重新做回归，预测的效果见下表，总正确率是 57.4%。

Classification

Observed	Predicted			
	Breakfast Bar	Oatmeal	Cereal	Percent Correct
Breakfast Bar	116	30	85	50.2%
Oatmeal	19	239	52	77.1%
Cereal	81	108	150	44.2%
Overall Percentage	24.5%	42.8%	32.6%	57.4%

专业课辅导+真题资料+微信公众号：432应统书院

10.9 对样品 1，ed=4，由输出结果 10.14 得 π_{1j} 的分母是：

$$1+\exp(0.049-0.676+0.037\text{address}+0.026\text{employ}+0.148\text{reside})$$
$$+\exp(-1.551+0.465+0.022\text{address}+0.051\text{employ}+0.084\text{reside})$$
$$+\exp(0.181-0.584+0.022\text{address}+0.042\text{employ}+0.258\text{reside})$$
$$=1+1.141+0.628+1.684=4.453$$

得 $$\pi_{11}=\frac{1}{4.453}=0.22$$

得 $$\pi_{12}=\frac{1.141}{4.453}=0.26$$

得 $$\pi_{13}=\frac{0.628}{4.453}=0.14$$

得 $$\pi_{14}=\frac{1.684}{4.453}=0.38$$

由于 $\pi_{14}=0.38$ 最大，所以第 1 个样品预测值是第 4 类。

同理计算出第 2 个样品的 $\pi_{24}=0.60$ 最大，预测值是第 4 类。

10.10 参照例 10.7，依照 Analyze→Regression→Multinomial Logistic 的顺序进入多类别 Logistic 回归对话框。把因变量 y 选入 Dependent 框条中，把定性自变量 $x3$ 选入 Factors 框条中。把数值型自变量 x_1，x_2 和 x_4 选入 Covariates 框条中。点击因变量框条下面的 Reference category 按钮，选择以 First category 为参照类别，点击 Continue 返回多类别 Logistic 回归对话框，点击 OK 运行。

在输出的回归参数估计表中看到，性别变量 x_3 在 $y=2$ 和 $y=3$ 两个类别的回归系数的 p 值分别是 0.660 和 0.904，都不显著，因此把性别变量 x_3 剔除，重新做回归，同时点击 Statistics 按钮加上 Classification table 选项。得运行结果：

Parameter Estimates

毕业去向[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
读研	Intercept	-19.116	6.383	8.970	1	.003	
	x1	.167	.057	8.650	1	.003	1.182
	x2	.038	.041	.851	1	.356	1.038
	x4	.004	.003	1.768	1	.184	1.004
出国	Intercept	-18.010	7.016	6.589	1	.010	
	x1	-.012	.065	.031	1	.860	.989
	x2	.122	.059	4.343	1	.037	1.130
	x4	.010	.004	6.406	1	.011	1.010

a. The reference category is: 工作.

从参数估计表中看到，与参加工作的同学相比，读研的同学其专业课成绩更好（x_1 的 P 值=0.003），而外语成绩（x_2 的 P 值=0.356）和经济状况（x_4 的 P 值=0.184）没有显著差异；出国留学的同学其专业课成绩（x_1 的 P 值=0.860）和参加工作的同学没有显著差异，外语成绩（x_2 的 P 值=0.037）和经济状况（x_4 的 P 值=0.011）则更好。

对 $y=2$（读研），

专业课辅导+真题资料+微信公众号：432应统书院

$$\pi_2=\frac{\exp(-19.116+0.167x_1+0.038x_2+0.004x_4)}{1+\exp(-19.116+0.167x_1+0.038x_2+0.004x_4)+\exp(-18.010-0.012x_1+0.122x_2+0.010x_4)}$$

对 $y=3$（出国留学），

$$\pi_3=\frac{\exp(-18.010-0.012x_1+0.122x_2+0.010x_4)}{1+\exp(-19.116+0.167x_1+0.038x_2+0.004x_4)+\exp(-18.010-0.012x_1+0.122x_2+0.010x_4)}$$

从 Classification 分类表中看到，预测的总正确率是 65%。

10.11　回归的参数估计表如下：

		Estimate	Std. Error	Wald	df	Sig.
Threshold	[chist = 1.00]	-3.562	.633	31.684	1	.000
	[chist = 2.00]	-2.732	.622	19.320	1	.000
	[chist = 3.00]	-.153	.614	.062	1	.803
	[chist = 4.00]	.182	.613	.088	1	.766
Location	age	.016	.004	18.170	1	.000
	[numcred=1.00]	-1.143	.593	3.710	1	.054
	[numcred=2.00]	.372	.597	.387	1	.534
	[numcred=3.00]	.954	.709	1.812	1	.178
	[numcred=4.00]	0(a)	.	.	0	.
	[othnstal=1.00]	-.401	.117	11.703	1	.001
	[othnstal=2.00]	-.436	.192	5.154	1	.023
	[othnstal=3.00]	0(a)	.	.	0	.

Link function: Complementary Log-log.

预测的效果见下表，因变量的观测值分为五类，而预测值只有两类。总正确率是（479+222)/1 000=70.1%，仅对样本中频率大的类别预测效果好。

		Predicted Response Category		
		Payments current	Critical account	Total
Account status	No debt history	13	27	40
	No current debt	41	8	49
	Payments current	479	51	530
	Payments delayed	31	57	88
	Critical account	71	222	293
Total		635	365	1000

10.12　对样品 1，数值自变量 age=67，定性自变量 numcred=2，othnstal=3。由上题中的参数估计表得

$$\log(-\log(1-\gamma_{1j}))=\theta_j-(0.016\times 67+0.372+0)$$
$$=\theta_j-1.444,\quad j=1,2,3,4$$

专业课辅导+真题资料+微信公众号：432应统书院

得 $\gamma_{1j}=1-\exp(-\exp(\theta_j-1.444))$， $j=1,2,3,4$

$\gamma_{11}=1-\exp(-\exp(-3.562-1.444))=0.006\,7$

$\gamma_{12}=0.015\,3$，$\gamma_{13}=0.184\,0$，$\gamma_{14}=0.247\,4$，$\gamma_{15}=1.0$

$\pi_{15}=1.0-0.247\,4=0.752\,6$，$\pi_{14}=0.247\,4-0.184\,0=0.063\,4$

$\pi_{13}=0.184\,0-0.015\,3=0.168\,7$，$\pi_{12}=0.015\,3-0.006\,7=0.008\,6$，$\pi_{11}=0.006\,7$

由于 $\pi_{15}=0.752\,6$ 最大，所以第 1 个样品预测值是第 5 类。

同理计算出第 2 个样品的 $\pi_{23}=0.715\,6$ 最大，预测值是第 3 类。

专业课辅导+真题资料+微信公众号：432应统书院

附　录

表 1　简单相关系数临界值表

$n-2$	5%	1%	$n-2$	5%	1%	$n-2$	5%	1%
1	0.997	1.000	16	0.468	0.590	35	0.325	0.418
2	0.950	0.990	17	0.456	0.575	40	0.304	0.393
3	0.878	0.959	18	0.444	0.561	45	0.288	0.372
4	0.811	0.947	19	0.433	0.549	50	0.273	0.354
5	0.754	0.874	20	0.423	0.537	60	0.250	0.325
6	0.707	0.834	21	0.413	0.526	70	0.232	0.302
7	0.666	0.798	22	0.404	0.515	80	0.217	0.283
8	0.632	0.765	23	0.396	0.505	90	0.205	0.267
9	0.602	0.735	24	0.388	0.496	100	0.195	0.254
10	0.576	0.708	25	0.381	0.487	125	0.174	0.228
11	0.553	0.684	26	0.374	0.478	150	0.159	0.208
12	0.532	0.661	27	0.367	0.470	200	0.138	0.181
13	0.514	0.641	28	0.361	0.463	300	0.113	0.148
14	0.497	0.623	29	0.355	0.456	400	0.098	0.128
15	0.482	0.606	30	0.349	0.449	1 000	0.062	0.081

专业课辅导+真题资料+微信公众号：432应统书院

表 2　t 分布表

例：自由度 $f=10$，$P(t>1.812)=0.05$，$P(t<-1.812)=0.05$

f \ α	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.000 5
1	0.100	1.376	1.963	3.076	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.397	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.733	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.362	2.576	3.291

专业课辅导+真题资料+微信公众号：432应统书院

表 3　F 分布表

例：自由度 $n_1=5$，$n_2=10$，$P(F>3.33)=0.05$

$P(F>5.64)=0.01$

n_2 中下面的数字是 1%的显著水平，上面的数字为 5%的显著水平。

分母的自由度 n_2 \ 分子的自由度 n_1	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
	4 052	4 999	5 403	5 625	5 764	5 859	5 928	5 981	6 022	6 056	6 082	6 106
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	18.01	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	20.15	10.05	9.96	9.89
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
	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	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55

专业课辅导+真题资料+微信公众号：432应统书院

续前表

n_2 \ n_1	分子的自由度											
	1	2	3	4	5	6	7	8	9	10	11	12
分母的自由度 17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31
	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	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.23	2.25
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.32	2.28	2.24	2.20	2.16
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.34	3.21	3.13	3.05	2.99
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80
34	4.13	3.28	2.88	2.65	2.49	2.38	3.30	2.23	2.17	2.12	2.08	2.50
	7.44	5.29	4.42	3.93	6.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03
	7.39	5.25	4.38	3.80	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69

专业课辅导+真题资料+微信公众号：432应统书院

续前表

n_2 \ n_1	分子的自由度											
	1	2	3	4	5	6	7	8	9	10	11	12
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93
	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90
	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83
	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40	2.33
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82
	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80
	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78
	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.33
1 000	3.85	3.00	1.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76
	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75
	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18

分母的自由度

专业课辅导+真题资料+微信公众号：432应统书院

续前表

n_2 \ n_1	分子的自由度											
	14	16	20	24	30	40	50	75	100	200	500	∞
1	245	246	248	249	250	251	252	253	253	254	254	254
	6 142	6 169	6 208	6 234	6 258	6 286	6 302	6 323	6 334	6 352	6 361	6 366
2	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
3	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12
4	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46
5	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02
6	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88
7	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65
8	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
9	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
10	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.94
11	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60
12	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
13	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	1.24	2.22	2.21
	3.85	3.78	3.67	3.59	3.15	3.42	3.37	3.30	3.27	3.21	3.18	3.16
14	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
15	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
16	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.89	2.86	2.80	2.77	2.75
17	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
18	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
19	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49

分母的自由度

专业课辅导+真题资料+微信公众号：432应统书院

续前表

n_2 \ n_1	分子的自由度											
	14	16	20	24	30	40	50	75	100	200	500	∞
分母的自由度 20	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
21	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	2.97	2.89	2.78	2.79	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
25	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96
34	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91
36	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	2.62	3.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87
38	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
42	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72

专业课辅导+真题资料+微信公众号：432应统书院

续前表

分母的自由度 n_2 \ 分子的自由度 n_1	14	16	20	24	30	40	50	75	100	200	500	∞
48	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
55	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
60	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
65	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	2.37	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
70	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
125	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37
150	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33
200	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
400	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
1 000	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
∞	1.67	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

专业课辅导+真题资料+微信公众号：432应统书院

表 4 DW 检验上下界表

n 是观测值的数目；k 是解释变量的数目，包括常数项。 5%的上下界

n	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.26	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

专业课辅导+真题资料+微信公众号：432应统书院

续前表　　1%的上下界

n	$k=2$		$k=3$		$k=4$		$k=5$		$k=6$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

专业课辅导+真题资料+微信公众号：432应统书院

参考文献

[1] 何晓群. 回归分析与经济数据建模. 北京：中国人民大学出版社，1997

[2] 陈希孺，王松桂. 近代回归分析. 合肥：安徽教育出版社，1987

[3] [美] 约翰·内特. 张勇，王国明，赵秀珍译. 应用线性回归模型. 北京：中国统计出版社，1990

[4] [美] 达摩达尔·N·古扎拉蒂. 费建平，孙春霞等译. 计量经济学基础（第四版）. 北京：中国人民大学出版社，2005

[5] 方开泰. 实用回归分析. 北京：科学出版社，1988

[6] 张尧庭等. 定性资料的统计分析. 桂林：广西师范大学出版社，1991

[7] 周纪芗. 回归分析. 上海：华东师范大学出版社，1993

[8] 张寿，于清文. 计量经济学. 上海：上海交通大学出版社，1984

[9] 李子奈. 计量经济学——方法和应用. 北京：清华大学出版社，1992

[10] 王国梁等. 问卷调查资料的一种统计分析方法——Logistic 回归模型. 统计研究，1991 (2)

[11] 卢文岱. SPSS for Windows 统计分析（第三版）. 北京：电子工业出版社，2006

[12] 何晓群等. 多元统计分析在考试评价中的应用. 国家教育部课题报告，2000

[13] 陆游等. 维生素 C 注射液抗变色配方的优选．均匀设计应用论文选（第一辑），1995

[14] 徐秀兰等. 均匀设计试验法在内燃机试验中的应用. 农业工程学报，1998 (12)

[15] 贾俊平. 统计学. 北京：中国人民大学出版社，2007

[16] 何晓群，刘文卿. 关于加权最小二乘法的探讨. 统计研究，2006 (4)

[17] 何晓群，刘文卿. 应用回归分析（第二版）. 北京：中国人民大学出版

专业课辅导+真题资料+微信公众号：432应统书院

社，2007

[18] A. E. Hoerl and R. W. Kennard, Ridge Regression: Biased Estimation for Non-orthogonal Problems, Technometrics, 12, 1970, 55-88

[19] Mcdonald, G. C. and Schwing, R. C., Instabilities of Regression Estimates Relating Air Pollution to Mortality, Technometrics, 15, 1973, 463-481

[20] He Xiaoqun, The Applications of Principal Component Estimation to Grain Production Analysis Model, 50th Session of the International Statistical institute, Beijing, 1995

[21] Seber, G. A. F., Linear Regression Analysis, John Wiley, 1977

[22] He Xiaoqun, Multiple Variable Statistical Analysis of the Causes of National Income Growth in China, IS MAA, Hong Kong, 1992

[23] N. R. Draper, H. Smith, Applied Regression Analysis, New York, 1981

[24] Frank, L. E. and Friedman, J. H., A Statistical View of Some Chemometrics Regression Tools, Technometrics, 35, 1993, 109-148

[25] Ratkowsky, D. A., Handbook of Nonlinear Regression Models, Mavcel Dekker, New York, 1990

[26] J. Durbin and G. S. Watson, Testing for Serial Correlation in Least Squares Regression, 1951

专业课辅导+真题资料+微信公众号：432应统书院

[General Information]
书名=应用回归分析
作者=何晓群等编著
丛书名=普通高等教育“十一五”国家级规划教材 21世纪统计学系列教材
页数=288
SS号=12916470
出版日期=2011.09
出版社=中国人民大学出版社
尺寸=26cm
原书定价=32.00
参考文献格式=何晓群等编著.应用回归分析.北京市：中国人民大学出版社,2011.09.
内容提要=回归分析是统计学中的一个重要分支，本书介绍了回归分析的基本理论和方法，结合中国社会、经济、自然和学等领域的研究实例，把回归分析方法与实际应用相联系。本书上一版出版于2007年，截至目前实现销售9000多册。