北京市高等教育精品教材立项项目

光华管理学院
Guanghua School of Management

王汉生 ● 著

北京大学光华管理学院教材

Business Statistics

商务统计系列

应用商务统计分析

APPLIED
BUSINESS
STATISTICAL
ANALYSIS

北京大学出版社
PEKING UNIVERSITY PRESS

应用商务统计分析
金融时间序列分析
数据挖掘与应用

北京大学光华管理学院教材 · 商务统计系列

上架建议：工商管理/统计学

ISBN 978-7-301-12893-0

定价：35.00元

Business Statistics
商务统计系列

应用商务统计分析

APPLIED
BUSINESS
STATISTICAL
ANALYSIS

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

应用商务统计分析/王汉生著. —北京:北京大学出版社,2008.1
(北京大学光华管理学院教材·商务统计系列)
ISBN 978-7-301-12893-0

Ⅰ.应… Ⅱ.王… Ⅲ.商业统计-统计分析-高等学校-教材 Ⅳ.F712.3

中国版本图书馆 CIP 数据核字(2007)第 187087 号

书　　名:应用商务统计分析
著作责任者:王汉生　著
责任编辑:张静波
标准书号:ISBN 978-7-301-12893-0/F·1753
出版发行:北京大学出版社
地　　址:北京市海淀区成府路 205 号　100871
网　　址:http://www.pup.cn
电　　话:邮购部 62752015　发行部 62750672　编辑部 62752926
　　出版部 62754962
电子邮箱:em@pup.pku.edu.cn
印　刷　者:北京飞达印刷有限责任公司
经　销　者:新华书店
　　730 毫米×980 毫米　16 开本　15 印张　273 千字
　　2008 年 1 月第 1 版　2008 年 1 月第 1 次印刷
印　　数:0001—5000 册
定　　价:35.00 元

丛书总序

教材建设是大学人才培养和知识传授的重要组成部分。对管理教育而言，教材建设尤为重要，一流的商学院不仅要有一流的师资力量、一流的生源、一流的教学管理水平，而且必须使用一流的教科书。一流的管理类教科书必须满足以下标准：第一，能把所在领域的基础知识以全面、系统的方式和与读者友好的语言呈献给读者；第二，必须有时代感，能把学科前沿的研究成果囊括进去；第三，必须做到理论和实务（包括案例分析）相结合，有很强的实用性；第四，能够启发学生思考现实的管理问题，培养他们分析问题和解决问题的能力；第五，可以作为研究人员和管理人士的工具书。

中国的管理教育是伴随改革开放而产生的。真正意义上的管理教育在中国不过十多年的历史，但巨大的市场需求使得管理教育成为中国高等教育各学科中发展最快的领域，管理类教科书市场异常繁荣。但总体而言，目前国内市场上管理类教科书的水平仍不能令人满意。国内教科书作者大多数在所涉及领域并没有真正的原创性研究和学术贡献，所撰写的教科书普遍停留在对国外教科书的内容进行中国式排列组合的水平上；国外引进的原版教科书虽然具有学术上的先进性，但由于其写作背景是外国的管理实践和制度安排，案例也都是取自于西方发达国家，对中国读者而言，总有一种隔靴搔痒的感觉。如何写出一流的中国版的管理类教材，是中国管理教育发展面临的重要任务。

北京大学光华管理学院一直重视教材建设工作。1999 年夏，我们曾与经济科学出版社签约，以每本 20 万元的稿酬，向全国征集 MBA 教科书作者。这个计划公布之后，我们收到了十几本教科书的写作方案。遗憾的是，经专家委员会评审，没有一本可以达到我们所期望的水平。究其原因，主要

是当时中国管理学院的教授、学者大多数并没有真正从事有关中国商业实践、管理实践的理论性和实证性研究。我们得出的结论是：没有一流的学者，没有一流的学术研究成果，就不可能写出一流的教科书。国外有大量优秀的教科书，这些教科书都是成千上万的优秀学者在对每一个具体的管理问题进行出类拔萃的研究的基础上写成的，是学术研究的结晶。国内学者如果没有研究的积累，要写出包含中国管理实践的好的教科书是不可能的。所以，我们果断地中断了这个计划。

自1999年以来，师资队伍的建设成为光华管理学院工作的重中之重，除了通过出国培训、合作研究等方式提升原有教师的水平外，我们还从国内外引进了六十多位优秀的新教师，使得光华管理学院成为真正与国际接轨的研究型商学院。我们的绝大多数教师不仅受过良好的科学研究训练，具有很好的理论素养，而且潜心于中国管理实践的研究和教学。不少教师已在国际一流的学术刊物上发表论文，受到国际同行的关注。

今天，我们有充分的信心向社会呈献一套由光华管理学院教师撰写的优秀的管理类系列教科书。本系列教材包括“应用经济学”、“金融学”、“会计学”、“市场营销学”、“战略管理”、“组织管理”、“管理科学与工程”和“信息系统管理”等子系列，涵盖管理学院教学的各个层面。这些教科书都是作者在光华管理学院多次授课的讲义的基础上反复修改写成的，经受过课堂实践的考验。

当然，我们深知，优秀教材的建设是一项长期的任务，不可能一蹴而就，也不是一个学院所能独立完成的，需要所有管理学院的通力合作。我们欢迎兄弟院校的师生和广大读者对教材提出批评和建议，以便我们不断改进。

让众多的没有机会进入光华管理学院的读者分享我们的课堂内容，是我们的荣幸，更是我们的责任。我们将以开放的姿态和与时俱进的精神为管理学教育的发展而努力，将更多更高水平的教材奉献给广大读者！

张维迎

北京大学光华管理学院

2005年7月11日

前　言

在过去的20年中，中国的商学院教育（本科生、普通硕士、MBA、EMBA以及管理学博士）经历了巨大的变化与发展。伴随其中的是统计学在各个项目中广泛深入的应用。一方面，统计学在商学院教学科研工作中的重要作用得到了一致的认可。大量的经济学、金融学以及营销学模型需要通过统计学的方法予以实现；大量的心理学、社会学以及行为学的实验需要运用统计学的分析方法进行研究。但另一方面，商学院的统计学教育却面临着前所未有的挑战。传统的以数学推导为主的教学方式给人们留下了统计学非常有用，但是又非常晦涩难懂的印象。因此，我们必须对传统的统计学教学方法进行合理的改进，否则无法满足商学院的教学及科研需要。笔者认为，商务统计学的教学绩效，从小处看影响学生的学习效果，从大处看则关系着统计学在商学院教育中的地位以及未来发展。

统计学在商学院的教学中会遇到哪些挑战呢？第一，学生背景复杂。传统的统计系面对的是大量具有非常良好数理背景的理科学生。因此，大量复杂的数学推导不成问题。而如今的商学院面对的学生背景极其复杂。虽然其中不乏有数理功底很强的学生，但是更多的还是其他非数理背景的学生，特别是MBA学员。而要让并不具备数理背景的同学也欣赏统计学并不是一件容易的事情。第二，教学目的不同。传统的统计系往往希望能够培养出优秀的统计学专业人才。这里的专业人才指的是不仅可以分析典型数据，而且还具备一定的方法创新能力的人才。但是，商学院对统计学的教学要求是不同的，商学院要培养的是具备良好的经济管理知识和出色的数据分析能力的经济管理人才。对于这类人才，能否创造出新的统计学方法并不是核心要求。而其核心要求是能够通过标准的统计学模型以及软件解决实际的经济管理问题。

上述商学院对统计教学的特殊要求决定了商务统计教学要有自己的特点。具体地说，就是要淡化数学公式的推导，着重讲授统计学思想，并强化其在实际

案例中的应用。而要实现该目标，一个重要的前提就是要有一本合适的商务统计学教材。但是，目前市场上却缺乏一种符合以上所有教学特点和需要的教材。目前市场上的教材大致可以分为两类：第一类是传统的数理统计学教材。此类教材以数学推导为主，主要满足具有较强数理统计背景的学生的需要。第二类是影印或者翻译的国外 MBA 统计学教材。此类教材可以作为相关专业的统计学入门教材，但是它们对标准的回归模型涉足甚浅，因而不能满足更高层次的学习及研究需要。笔者以自己在北京大学光华管理学院多年的商务统计学的教学经验为基础，并结合商学院的统计学教学特点及需要编写了这本教材。本教材有以下几个特色：第一，全部以实际案例背景驱动；第二，没有任何复杂的数学推导；第三，每个案例还配备了完整的分析报告；第四，附有详细的 *R* 程序以及注释。出于对商业版权的考虑，本教材所采用的全部数据都是随机模拟生成的，但是案例背景是真实可靠的。还特别值得一提的是，本教材所使用的 *R* 软件是公开、免费的统计分析软件，可以在网上轻松获得（http://CRAN.R-project.org）。本书融合了以上这些特色，目的是希望能够使读者学习得更加轻松、有趣。

本教材从开始编写到最终定稿历时近两年，其间得到了我的多位教学助理的大力支持，其中要特别感谢罗荣华和赵羿两位同学。是罗荣华同学的帮助，本书才有了完整的分析报告，以及带有注释的 *R* 程序；是赵羿同学的仔细校对及润色，才使得本书的文字大有改进。笔者还要感谢北京大学光华管理学院的涂平老师，他为本书提出了许多建设性的建议。笔者还要感谢北京大学出版社的朱启兵老师还有张静波老师，没有他们的帮助，本书不可能得以如此迅速地出版。由于笔者的能力有限，书中难免有疏漏之处，请多指正。最后，我想将此书献给我的太太还有刚刚来到世间的儿子，希望他们永远幸福健康！

王汉生
北京大学光华管理学院
2008 年 1 月 1 日
hansheng@gsm.pku.edu.cn
http://hansheng.gsm.pku.edu.cn

CONTENTS
目 录

CONTENTS
目 录

CONTENTS
目 录

CONTENTS
目 录

第一章
线性回归

- 案例介绍
- 模型定义
- 描述性分析
- 参数估计
- 假设检验
- 模型诊断
- 变量选择
- 模型预测
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的是通过一个盈利预测的实际案例，详细介绍线性回归这种最重要的统计回归模型。它主要处理的是因变量和解释性变量都是连续型数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用普通线性模型；(2) 线性模型的基本统计学理论；(3) 相关理论在统计学软件 R 中的应用；(4) 相应的统计分析报告的撰写。本章初次涉及的重要统计学概念如下：因变量、解释性变量、普通线性模型、最小二乘估计、F 检验、t 检验、模型诊断、异常值、Cook 距离、模型选择(AIC、BIC)，还有外样本的预测检验等。

第一节　案例介绍

线性回归模型是实际工作中用得最多、最广的统计模型。它不仅为我们提供了一套系统而有效的分析预测方法，而且为我们提供了一套完整的方法论。对线性模型的理解与掌握将极大地有助于以后章节的学习。具体地说，本章将以一个实际应用为例，详细讲解并演示线性模型的各个方面。

我们考虑的具体问题是：如何利用上市公司当年的公开财务指标对其来年的盈利状况予以预测。合理回答该问题对于指导投资者了解企业的盈利模式、风险大小，以及正确投资帮助甚大。类似的问题早已在北美和欧洲的金融市场上被广泛研究，而本章将对中国股市的类似数据予以简略分析。具体地说，我们的目标是如何有效利用上市公司的历史财务数据，对其来年的净资产收益率(return on equity, ROE)予以大概的估计。我们考虑的财务指标有：公司当年的净资产收益率(ROEt)、资产周转率(ATO)、债务资本比率(LEV)、市倍率(PB)、应收账款/主营业务收入(ARR)、主营业务利润/主营业务收入(PM)、主营业务收入增长率(GROWTH)、存货/资产总计(INV)以及对数变换后的资产总计(ASSET)，以后简称为资产总计。

对于这些财务指标的经济意义的详细解释可以在任何一本会计教科书中找到。我们只对以上财务指标的经济意义简述如下。首先，当年表现好的公司，由于惯性效应，其下年度的表现也趋向于较好。所以，很自然地，我们应该考虑公司当年的净资产收益率(ROEt)。其次，债务资本比率(LEV)反映了公

司的基本债务状况，市倍率(PB)与主营业务收入增长率(GROWTH)分别反映了公司预期的未来成长率以及公司已经实现了的当年增长率，应收账款/主营业务收入(ARR)以及主营业务利润率(PM)分别反映了公司的收入质量以及利润状况。最后，存货周转率(INV)用来度量公司的存货状况，而资产总计(ASSET)被用来控制公司规模的影响。以上所考虑的变量都是过去欧美国家同类研究中发现的、非常典型的能够影响公司盈利能力的指标。因此，检验这些指标在我国股票市场上的有效性就变得非常有意义。为方便演示，我们随机抽取了深市和沪市2002、2003年度的各500个样本。其中，分析主要是基于2002年的样本，而2003年的数据主要用来检验模型的预测精度。值得注意的是，考虑到数据的商业性与保密性，以及教学演示的方便，本案例的数据是随机模拟生成的。

第二节　模型定义

根据前一节中的讨论可以看到，我们有一个清晰的目的，那就是利用所给的财务数据(accounting variables)预测下一年度的净资产收益率(ROE)。还可以看到，我们所考虑的这些指标所扮演的角色是不一样的。其中，下年度的净资产收益率(ROE)是可预测的指标。之所以它可预测，是因为它的大小在一定程度上是由其他几个财务指标所决定的。因此，我们称其为因变量(dependent variable)。也就是说，下一年度的净资产收益率会因为其他几个财务指标的改变而改变，或者说下一年度的净资产收益率在一定程度上是可以被其他的财务指标所解释的。一旦我们明白了这一点，就不难理解为什么人们会称其他几个财务指标(如ROEt、LEV、ATO等)为自变量(independent variable)或者解释性变量(predictor, predictive variable)，亦称协变量(covariate)。

为了方便讨论，我们用数学符号来描述我们的数据。具体地说，用(y_i, x_i) $(i=1,\cdots,n)$来代表$n=500$个样本。其中，y_i是来自于第i个公司的下年度净资产收益率(ROE)。另外，用$x_i=(x_{i1},x_{i2},\cdots,x_{ip})'$来表示相关的$p$个解释性变量。在我们的数据中，共有9个解释性变量($p=9$)，它们分别对应了以下9个财务指标：公司当年的净资产收益率(x_{i1} = ROEt)、资产周转率(x_{i2} = ATO)、债务资本比率(x_{i3} = LEV)、市倍率(x_{i4} = PB)、应收账款/主营业务收入(x_{i5} = ARR)、主营业务利润/主营业务收入(x_{i6} = PM)、主营业务收入增长率(x_{i7} = GROWTH)、存货/资产总计(x_{i8} = INV)和资产总计(x_{i9} = ASSET)。

而回归分析的根本目的就是要探寻因变量(下年度的净资产收益率)同自变量(其他财务指标)之间的数量关系。为了达到此目的，我们不可避免地需要

假设 y_i 和 x_i 之间的数量关系满足某种函数形式，而最简单也是最常用的函数形式就是线性函数。这对应了下面这个含有 p 个自变量的一般线性模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

其中，β_0 为常数项，$\beta_j (j = 1, 2, \cdots, p)$ 为第 j 个解释性变量 x_{ij} 的回归系数，它意味着，若 x_{ij} 变化一单位，我们可以预期因变量 y_i 会变化多少个单位。例如，在我们的案例中，如果 x_{i1} 代表了公司当年的净资产收益率，那么 β_1 意味着该公司当年净资产收益率若增加一单位，则来年净资产收益率变化的期望值（平均值）为 β_1 个单位。

当然，这样的预测是不可能完全准确的。这是因为除了我们所考虑的财务指标以外，还有太多其他的因素（如经济周期、公司高管变化）也可能对公司下一年度的盈利产生影响。而且这样的影响是无法用我们所考虑的财务指标来反映的，也是不能预测的。这种影响就构成了我们的残差项 ε_i。残差项对因变量 y_i 影响力的大小，直接反映了自变量 x_i 对因变量的预测能力。比方说，如果 ε_i 对公司下一年度的影响力很大，那么我们就会发现，很难利用现有的财务数据准确预测公司下年度的盈利。否则，我们就会获得一个有效的预测模型。那么，统计上应该怎样度量残差项影响的大小呢？我们将在后面的章节中详细讨论。最后值得一提的是，对于此普通线性模型，技术上我们一般作如下假定。

独立性：这包含两个方面。一方面，不同的观测之间是互相独立的；另一方面，残差项同解释性变量之间是互相独立的。在我们的案例中，这首先意味着不同公司之间是互相独立的，A 公司盈利的好坏不会影响到 B 公司。由此可见，此假设在一定程度上是合理的，符合常识的。但是，我们为什么要说是“一定程度上”呢？这是因为此假设在一定程度上也是不合理的。例如，如果我们的数据跨越多个年份，那么来自同一年份的数据，受当年整个宏观经济的影响，将会表现出一定程度上的相似性。进一步讲，如果我们的数据中有的公司之间是有关联的（如母公司与子公司、主要竞争对手），那么一个公司的盈利状况必然影响另外一个公司。再进一步讲，如果我们的数据中有多个观测来自于同一个公司，那么这些数据之间是高度相关的。那么，这是否意味着线性模型将无法应用？也不是的。在现实生活中，几乎没有任何一个数据能够完美地满足所有的理论假设，因此我们需要一个主观的、经验的判断——这个数据是否极大程度地满足了我们的理论要求，它与理论要求的偏差是否并不会显著地影响我们的结论？如果答案是肯定的，那么我们还是可以继续利用线性模型获得信息。

常方差：即残差项 ε_i 的方差不依赖于自变量 x_i 的取值，为一个常数。对于我们的案例，这个假设意味着，公司盈利状况的波动程度不依赖于我们所考虑

的那些财务指标，如公司规模、成长率等。但是，现实恰恰相反。一般来说，大规模的公司历史悠久，经营稳定，因此盈利也趋向于稳定，即 ε_i 的方差较小。而高速成长的公司经历了快速的发展变化，因此极有可能某年盈利很高而另一年盈利很低。也就是说，这类公司的盈利状况是非常不稳定的，即 ε_i 的方差较大。既然这个假设在实际数据中常常被严重破坏，那么我们是否还可以利用线性模型呢？答案是肯定的。这是因为，人们的理论研究发现，即使常方差的假设遭到破坏，只要样本足够大，我们将来所介绍的各种统计推断方法仍然是有效的。这主要是由于中心极限定理的作用。但是，如果我们的样本并不是很大，那么对常方差假设的破坏往往带来的是低效率的估计量(inefficient estimator)。

正态性：即假设残差项 ε_i 是服从正态分布的。不幸的是，在现实生活中这个假设被破坏得更加普遍而严重。例如，金融市场的数据(如本案例)往往是肥尾的(heavy tailed)。也就是说，它们所产生的极值的概率要大于正态分布的预测。但是，幸运的是，这个假设与常方差假设相比更不重要。同样是由于中心极限定理的存在，使得只要样本量足够大，我们就不用担心此假设。但是，对于小样本，我们必须非常小心。

第三节　描述性分析

对于任何一个数据，在进行正式的统计推断以前，都有必要作一些简单的描述性分析。描述性分析可以帮助我们获得对数据的整体概念，也可以帮助我们很早地发现异常观测，以及重要的趋势。这对后面的正式分析有着非常重要的借鉴意义。假定我们的数据文件名为“roe. txt”，被存放在“D:\Practical Business Data Analysis\case\CH1\roe. txt”，那么在 *R* 的编程环境中输入以下语句，就可以读入数据：

```
> rm(list=ls())
> a=read.table("D:/Practical Business Data Analysis/case/CH1/roe.txt",header=T)
> round(a[1:10,],4)
   year    ROEt     ATO      PM     LEV GROWTH      PB     ARR     INV  ASSET     ROE
1  2002   0.296   0.389   0.215   4.384  0.197   8.048   0.637   0.248 20.873   0.181
2  2002   0.665   0.335   0.407   2.273 -0.084   2.154  -1.217   0.030 21.062   0.899
3  2002  -0.045   0.963   0.084  -5.339  4.155 -12.879  -0.598   0.096 21.474   1.504
4  2002  -0.783   0.437  -0.096   1.016  2.231   3.836   0.056   0.051 19.746  -0.777
5  2002   1.053   0.858  -0.154  -2.066  0.519   1.972   0.293   0.032 22.570   0.728
6  2002   0.590   0.528   0.197  -0.946 -1.441   2.405  -0.986   0.189 20.890   1.141
7  2002  -0.204   0.419   0.324   3.513  1.999   8.827   0.389  -0.060 21.965   0.568
8  2002  -0.845   0.691   0.131   4.628  1.733   7.522  -0.471   0.145 20.377  -0.171
9  2002  -0.065  -0.230   0.071   1.525 -2.436 -14.887   0.958   0.215 20.480   0.415
10 2002   1.001  -0.176   0.123  -0.891  2.257 -12.065  -0.336   0.065 20.455   0.699
```

然后，我们希望对数据进行简单的描述性分析。具体地说，我们需要知道各个自变量以及因变量的均值、最小值、中位数、最大值、标准差，从而获得对数据的一个整体印象。在 *R* 中，可以具体计算如下：

```
> a1=a[a$year==2002,-1]
> Mean=sapply(a1,mean)
> Min=sapply(a1,min)
> Median=sapply(a1,median)
> Max=sapply(a1,max)
> SD=sapply(a1,sd)
> cbind(Mean,Min,Median,Max,SD)
            Mean      Min   Median     Max        SD
ROEt    0.067778  -1.390  0.0800   1.421 0.5189265
ATO     0.429818  -0.928  0.4375   1.927 0.4604041
PM      0.210952  -0.424  0.2180   0.698 0.1810975
LEV     0.708852  -7.941  0.5595   9.362 3.1823500
GROWTH  0.331254  -5.962  0.3790   6.092 2.1204400
PB      2.126802 -20.816  2.2710  32.591 9.5126759
ARR     0.200780  -2.601  0.2230   3.187 0.9466115
INV     0.100044  -0.264  0.1020   0.431 0.1224118
ASSET  21.066002  18.629 21.0565  23.414 0.8547751
ROE     0.410498  -1.161  0.4200   5.285 0.5447802
```

从中我们可以对各个变量予以简单描述。例如，从对 ROE 的描述性统计量我们可以知道，在我们的 500 个训练样本中，公司下年度的净资产收益率处于 -1.161 到 5.285 之间。其平均水平约为 0.4105（平均值）、0.4200（中位数），其标准差约为 0.5448。然后，我们对各个变量之间的相关性分析如下：

```
> round(cor(a),3)
         year   ROEt    ATO     PM    LEV GROWTH     PB    ARR    INV  ASSET    ROE
year    1.000 -0.009  0.002  0.005  0.068 -0.051  0.080 -0.038  0.033  0.032 -0.038
ROEt   -0.009  1.000  0.044  0.098 -0.247  0.042 -0.225 -0.073 -0.029  0.136  0.572
ATO     0.002  0.044  1.000 -0.272 -0.036  0.004 -0.054 -0.117  0.105  0.069  0.033
PM      0.005  0.098 -0.272  1.000 -0.146  0.026 -0.056 -0.402 -0.097  0.023  0.091
LEV     0.068 -0.247 -0.036 -0.146  1.000 -0.028  0.746  0.060  0.036 -0.019 -0.297
GROWTH -0.051  0.042  0.004  0.026 -0.028  1.000 -0.019 -0.034  0.043 -0.030  0.099
PB      0.080 -0.225 -0.054 -0.056  0.746 -0.019  1.000  0.002 -0.026 -0.188 -0.237
ARR    -0.038 -0.073 -0.117 -0.402  0.060 -0.034  0.002  1.000 -0.051 -0.086 -0.091
INV     0.033 -0.029  0.105 -0.097  0.036  0.043 -0.026 -0.051  1.000  0.022 -0.021
ASSET   0.032  0.136  0.069  0.023 -0.019 -0.030 -0.188 -0.086  0.022  1.000  0.085
ROE    -0.038  0.572  0.033  0.091 -0.297  0.099 -0.237 -0.091 -0.021  0.085  1.000
```

从中我们可以看到，相关性最高的是下一年净资产收益率（ROE）和当年净资产收益率（ROEt），其相关系数高达 0.572，这暗示着当年净资产收益率对下一年净资产收益率具有极好的预测能力。这同我们的常识也比较吻合。为了更好地说明公司下一年的净资产收益率（ROE）对当年的净资产收益率（ROEt）的线性依赖关系，我们可以考虑通过下面的命令将它们画在同一张散点图上，结果如图 1-1 所示。

```
> plot(a1$ROEt,a1$ROE)
```

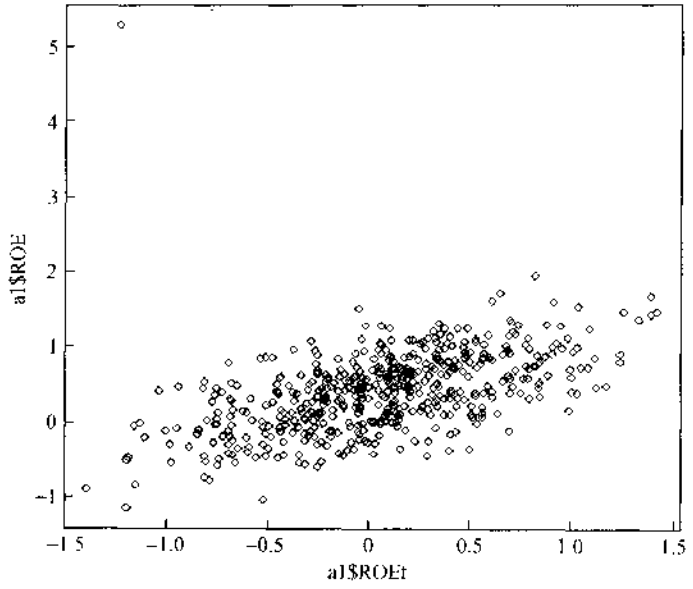


图 1-1　ROE 与 ROEt 的线性关系图

从图 1-1 可以看到，公司下一年的净资产收益率（ROE）和当年的净资产收益率（ROEt）近似地表现出一种线性关系。同时我们注意到，图中还有一个明显的异常值。该观测的取值明显地偏离了主体的数据趋势。在下面的分析中，我们会对该观测再作更加详细的讨论。

第四节　参数估计

在描述分析的基础上，我们希望能够严格地建立一个线性模型，以便于实际应用。但是，为了能够具体应用线性模型，我们至少应该知道回归系数 $\beta=(\beta_1,\cdots,\beta_p)'$ 的取值。例如，在我们的案例中，对于一个给定的公司，我们知道了它当年的各种财务数据，那么要对它下一年度的盈利状况予以预测，我们就必须知道或者近似地知道回归系数 β 的取值。

但是，由于随机噪音 ε_i 的存在，我们永远都无法知道 β 的确切取值。因此，取而代之，我们希望能够对 β 的取值予以合理“猜测”，这就是估计！除了估计 β 以外，为了假设检验的需要，我们还希望估计 ε_i 的方差 σ^2。对于回归系数，我们通常通过以下方法来获得估计。具体地说，我们要找到这样一个统计量 $\hat{\beta}$，它

能够使得目标函数：

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

最小化。换句话说，也就是：

$$S(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 = \min_{\beta} \{S(\beta)\}$$

此估计方法被称为最小二乘法（ordinary least squares，OLS），由此产生的参数估计 $\hat{\beta}$ 称为 β 的最小二乘估计。对于参数 β，我们可以验证，最小二乘估计也是极大似然估计。在 R 中，我们可以通过以下命令轻松获得最小二乘估计：

```
> lm1=lm(ROE~ROEt+ATO+PM+LEV+GROWTH+PB+ARR+INV+ASSET,data=a1)
> summary(lm1)

Call:
lm(formula = ROE ~ ROEt + ATO + PM + LEV + GROWTH + PB + ARR +
    INV + ASSET, data = a1)

Residuals:
      Min        1Q    Median        3Q       Max
-1.170755 -0.280202 -0.007754  0.243568  5.391891

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.454352   0.528317   0.860 0.390210
ROEt         0.487484   0.041052  11.875  < 2e-16 ***
ATO         -0.014700   0.047628  -0.309 0.757720
PM           0.078801   0.133221   0.592 0.554454
LEV         -0.039655   0.010520  -3.770 0.000184 ***
GROWTH       0.019938   0.009641   2.068 0.039159 *
PB           0.003317   0.003476   0.954 0.340558
ARR         -0.026016   0.024327  -1.069 0.285399
INV         -0.019784   0.168340  -0.118 0.906494
ASSET       -0.003111   0.024910  -0.125 0.900649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4557 on 490 degrees of freedom
Multiple R-Squared: 0.3129,     Adjusted R-squared: 0.3003
F-statistic: 24.79 on 9 and 490 DF,  p-value: < 2.2e-16
```

从上面的输出结果可以看到，对应于“Estimate”的一列代表了最小二乘估计值。其中，1 单位的当年净资产收益率的增长，对应于 0.4875 个单位的下年度净资产收益率的增长。其他参数也可以作类似的理解。从中可以看到，因变量（下年度净资产收益率）同公司当年的净资产收益率（ROEt）、主营业务利润/主营业务收入（PM）、主营业务收入增长率（GROWTH）以及市倍率（PB）这四项财务指标正相关（因为它们的回归系数是正的），而同其他财务指标负相关。

但是，我们不能就此认定我们的结论是可靠的，因为我们还不知道这些估计量的精确程度，也就是说，我们需要知道这些估计量的标准差。为方便起见，人们习惯于用以下矩阵形式来表述该线性模型：

$$Y = X\beta + \varepsilon$$

其中，$\beta = (\beta_0, \beta_1, \cdots, \beta_p)'$是回归系数向量，$Y = (y_1, \cdots, y_n)'$是因变量向量，$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$是随机扰动向量，而

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

是设计矩阵(design matrix)。对于我们的案例，我们可以获得部分设计矩阵如下：

```
> round(a1[1:10,],3)
    ROEt    ATO     PM    LEV GROWTH     PB    ARR    INV ASSET    ROE
1  0.296  0.389  0.215  4.384  0.197  8.048  0.637  0.248 20.873  0.181
2  0.665  0.335  0.407  2.273 -0.084  2.154 -1.217  0.030 21.062  0.899
3 -0.045  0.963  0.084 -5.339  4.155 -12.879 -0.598  0.096 21.474  1.504
4 -0.783  0.437 -0.096  1.016  2.231  3.836  0.056  0.051 19.746 -0.777
5  1.053  0.858 -0.154 -2.066  0.519  1.972  0.293  0.032 22.570  0.728
6  0.590  0.528  0.197 -0.946 -1.441  2.405 -0.986  0.189 20.890  1.141
7 -0.204  0.419  0.324  3.513  1.999  8.827  0.389 -0.060 21.965  0.568
8 -0.845  0.891  0.131  4.628  1.733  7.522 -0.471  0.145 20.377 -0.171
9 -0.065 -0.230  0.071  1.525 -2.436 -14.887  0.958  0.215 20.480  0.415
10 1.001 -0.176  0.123 -0.891  2.257 -12.065 -0.336  0.065 20.455  0.699
```

从中可以看到，最后一列就是我们的因变量向量 Y = ROE，而前九列就是设计矩阵 X 的前十行。为了简便起见，我们没有将截距项包含在内，希望能够给大家一个关于 Y 和 X 的直观理解。当$(X'X)^{-1}$存在时，回归参数 β 的最小二乘估计为 $\hat{\beta} = (X'X)^{-1}X'Y$，而因变量的拟合值为 $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$，残差为实际观测值与拟合值之差，即 $e = Y - \hat{Y} = (I - H)Y$，其中 I 为单位矩阵，$H = X(X'X)^{-1}X'$称为帽子矩阵(hat matrix)。在完成对回归模型的矩阵形式的定义之后，我们可以将参数的抽样分布也用矩阵形式给出：

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

即 $\hat{\beta}$ 服从于均值为 β，方差为$(X'X)^{-1}\sigma^2$ 的正态分布。因此，对于一个具体的 β_j，我们知道它的标准差为 $\sqrt{v_j(X)}\sigma$，其中 $v_j(X)$ 是$(X'X)^{-1}$的第 j 个对角元素，可以通过可观测的设计矩阵计算获得。但是，σ^2 常常是未知的，必须通过估计获得。常用的有如下无偏估计：

$$\hat{\sigma}^2 = \frac{1}{n-p-1}SSE = \frac{1}{n-p-1}(e'e) = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

其中，n 为观测的个数，p 为模型中自变量的个数，而

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

对于 $\hat{\sigma}^2$，我们知道有：

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

并且有 $\mathrm{Cov}(\hat{\beta}, e) = 0$，即 $\hat{\beta}$ 和 e 不相关。又因为在正态假定下不相关等价于独立，所以 $\hat{\beta}$ 和 e 独立，从而 $\hat{\beta}$ 和 $e'e$ 独立，再进一步导出 $\hat{\beta}$ 与 $\hat{\sigma}^2$ 独立。

从上表可以看到，有一列输出对应于“Std. Error”，这就是各个最小二乘估计的标准差。比方说，对于当年净资产收益率，其回归系数的最小二乘估计为 0.4875，代表了一种正的相关关系，而此估计的标准误差为 0.0411。请注意，此标准误差同相应的最小二乘估计比非常小，因此，直观地，我们可以比较确信当年的净资产收益率确实同下年度净资产收益率高度正相关。但是，在最终下结论之前，还应该接受严格的假设检验。

在获得参数估计之后，人们往往希望对模型的拟合优度（goodness-of-fit）作一个数量化的判断，从而表明我们所考虑的自变量能够在多大程度上解释因变量。对于我们的案例来说，也就是研究我们所考虑的 9 个解释性变量到底能够在多大程度上决定下一年的净资产收益率。因此，人们定义判决系数（R-square）如下：

$$R^2 = 1 - \frac{SSE}{SST}$$

其中，$SST = \sum (y_i - y)^2$ 为总平方和；$SSE = \sum (y_i - \hat{y}_i)^2$ 为残差平方和；R^2 表示误差能被模型解释的部分占总误差的百分比。一般认为 R^2 越大，模型拟合效果越好。但是，当变量增多的时候，R^2 只会不断增加而不会减少，所以如果仅以 R^2 来选择变量的话，肯定会倾向于选择更多的变量。为了解决这个问题，我们对增加变量进行相应的惩罚，得到调整后的判决系数（adjusted R-square）如下：

$$R_a^2 = 1 - \frac{n-1}{n-p-1}\left(\frac{SSE}{SST}\right)$$

其中，p 为模型中自变量的个数。从以上 R 的输出中我们可以看到，未调整的 R^2 为 31.29%，调整后的 R^2 为 30.03%。对于金融市场的数据，这个 R^2 值已经非常高了。

第五节　假设检验

仅仅得到模型拟合结果是不够的。这是因为我们不知道这些估计量是否具有统计上的显著性。以我们的案例为例，请注意变量 x_9（资产总计）的回归系数非常小，仅为 -0.0031。这就产生了两个可能：第一，有可能资产总计确实对公司下年度盈利有负的影响，只是比较小而已。第二，有可能资产总计对公司下年度盈利根本没有影响，而我们所看到的最小二乘估计完全是由随机噪音 ε_i 引起的。因此，我们需要对该模型以及各个变量的显著性予以严格检验，其顺序如下：

我们首先检验模型的显著性（global hypotheses testing），即检验在所有的自变量中是否有至少一个自变量对因变量有重要的解释性作用。在我们的案例中，这个问题就是：在我们所考虑的 9 个财务指标中，是否有至少一个对公司下年度盈利状况有预测能力。如果答案是肯定的，我们再研究到底是哪一个或者哪几个有预测能力。否则，我们就没有继续分析的必要。具体地说，此目的对应了著名的 F 检验，它的原假设以及对立假设给定如下：

$$H_0:\beta_i = 0 \quad \forall i \quad \text{vs} \quad H_1:\text{存在至少一个}\ \beta_i \neq 0$$

请注意，F 检验的原假设要求每一个回归系数都为 0，也就是说，我们考虑的所有自变量都是不重要的。而对立假设说至少有一个回归系数不为 0，也就是说，至少有一个解释性变量是重要的（但是，我们不知道到底是哪一个）。具体地说，F 检验的检验统计量如下：

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)}$$

如果原假设正确，我们知道该统计量服从一个自由度为 $(p, n-p-1)$ 的 F 分布。因此，对于一个给定的显著性水平 α，我们在 $F > F_{1-\alpha}(p, n-p-1)$ 时拒绝原假设，而接受对立假设。从以上输出中可以看到，我们的案例数据所产生的 F 统计量的值为 24.79，对应于自由度为 (9,490) 的 F 分布。其 p 值远小于 0.01，因此在 0.01 的显著性水平上高度显著。这说明在所考虑的所有财务指标中，有至少一个对预测未来盈利能力是重要的。

但是，F 检验不能告诉我们到底是哪几个财务指标重要，因此我们需要逐一检验到底是哪几个财务指标重要。其检验过程就是 t 检验。对一个给定的自变量 x_j，它的原假设以及对立假设给定如下：

$$H_0:\beta_j = 0 \quad \text{vs} \quad H_1:\beta_j \neq 0$$

检验统计量为：

$$t = \frac{\hat{\beta}_j}{\sqrt{v_j(X)}\hat{\sigma}}$$

其中，$v_j(X)$是$(X'X)^{-1}$的第j个对角元素。如果原假设正确，我们知道t服从自由度为$(n-p-1)$的t分布。因此，对于一个给定的显著性水平α，我们在$|t| > t_{1-\alpha/2}(n-p-1)$时拒绝原假设，而接受对立假设。

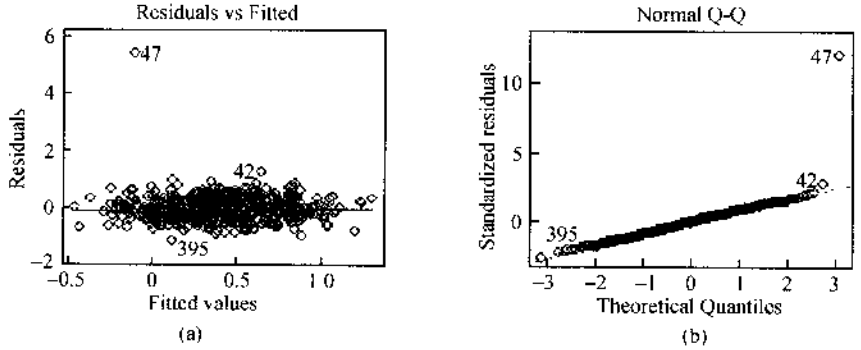
从以上输出可以看到，对应于"Pr(> |t|)"的那一列就是相应t检验的p值。比方说，对于主营业务收入增长率(GROWTH)的t检验的p值为0.0392 < 0.05。这说明，在0.05的显著性水平下，我们可以断定主营业务收入增长率同下一年盈利状况有着显著的正相关关系。类似地，我们还可以看到，当年净资产收益率(ROEt)和债务资本比率(LEV)在0.05的显著性水平下同来年的ROE分别正相关和负相关，而对于其他的变量暂时没有定论。

第六节　模型诊断

接下来，我们会对此模型作必要的模型诊断。我们要检验对于所选取的最优模型，我们在第二节中所提到的各种模型假设是否近似成立。另外，我们还关心数据中有没有异常值或影响点(influential point)。在*R*中，这可以轻松实现如下：

```
> par(mfrow=c(2,2))
> plot(lm3,which=c(1:4))
```

这个命令一共将产生四张图，如图1-2所示：



(a)　(b)

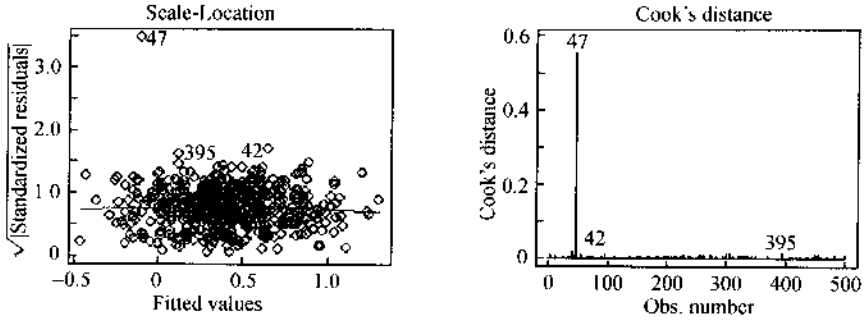


图 1-2 模型诊断图

左边两张是残差图，往往大同小异，因此只详细讨论图 1-2(a)。其中，横轴是对各个观测的拟合值 $\hat{y}_i$，而纵轴是分离出来的残差 $\hat{\varepsilon}_i = y_i - \hat{y}_i$。从图中首先可以看到第 47 个观测值的残差非常大，这说明该观测值很难用我们的线性模型拟合。这也从一个侧面暗示我们该观测值非常异常。其次，还可以看到残差的分布杂乱无序，没有明显趋势。这说明常方差的假设基本成立，并且没有遗漏重要的可预测信息。

下面，我们再检验残差 $\hat{\varepsilon}_i$ 的正态性。这可以通过常见的 QQ 图来简单检验。其基本原理如下：在 $\varepsilon_i \sim N(\mu, \sigma^2)$ 成立的条件下，设 $P(\varepsilon_i < q_\alpha) = \alpha$，则有：

$$P\left(\frac{\varepsilon_i - \mu}{\sigma} < \frac{q_\alpha - \mu}{\sigma}\right) = \alpha$$

由此可以得到 $q_\alpha = z_\alpha \sigma + \mu$（其中 z_α 是标准正态分布 α 分位点），即 q_α 和 z_α 为线性关系，所以在正态性的条件下，ε_i 与 z_α 也为线性关系。我们对样本残差 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 排序得到 $\varepsilon_{(1)} \leqslant \varepsilon_{(2)} \leqslant \cdots \leqslant \varepsilon_{(n)}$，并以 $\varepsilon_{(1)}, \varepsilon_{(2)}, \cdots, \varepsilon_{(n)}$ 为 X 轴，以 $z_{1/n}, z_{2/n}, \cdots, z_{n/n}$ 为 Y 轴作散点图。如果得到的散点图近似直线，则可以认为满足正态性假定；如果得到的散点图较大幅度偏离直线，则可以认为不满足正态性假定。从图 1-2(b)可以看到，除去第 47 个异常值以外，QQ 图基本上成一条直线，这说明正态性假设基本成立。

接下来，我们要检验数据中是否有异常值或影响点。这可以通过计算并比较 Cook 距离来实现。具体地说，Cook 距离定义如下：

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^{\mathrm{T}}(X^{\mathrm{T}}X)(\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2}$$

其中，$\hat{\beta}_{(i)}$ 是删除掉第 i 个观测之后的最小二乘估计。简单地说，Cook 距离 D_i 反映了包含与不包含第 i 个观测所引起的最终估计的差异。Cook 距离从某一

个侧面度量了第 i 个观测对整个估计的影响力。因此，当发现有的观测具有很大的 Cook 距离时，我们需要格外小心。值得注意的是，图 1-2(d)中的 Cook's distance 显示，第 47 个观测数据是一个非常有影响力的点。我们尝试将其删除，重新分析如下(参见图 1-3)：

```
> a1=a1[-47,]
> lm1=lm(ROE~ROE1+ATO+PM+LEV+GROWTH+PB+ARR+INV+ASSET,data=a1)
> plot(lm1,which=c(1:4))
```

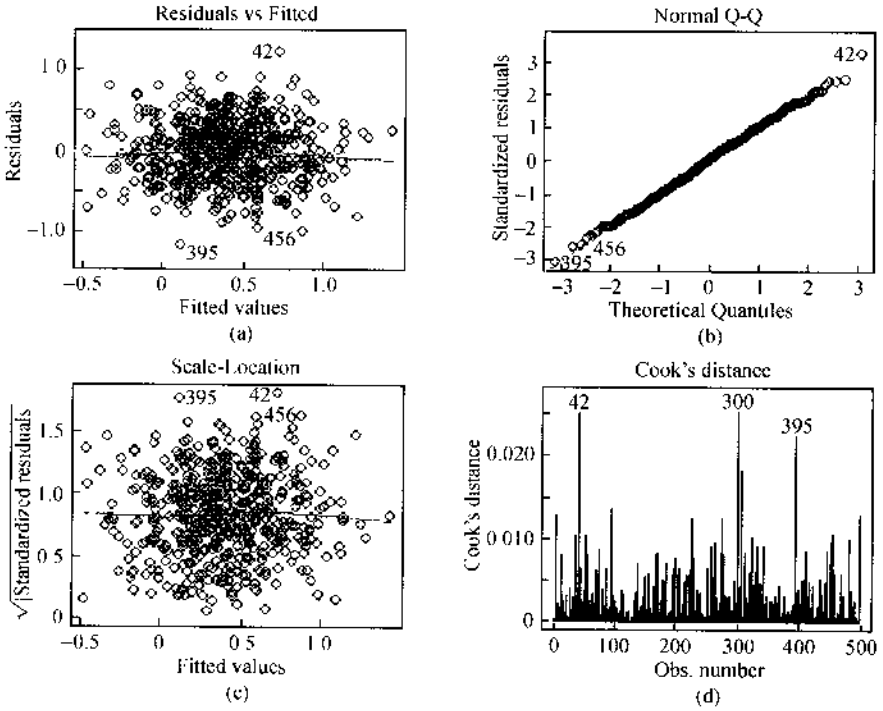


图 1-3 修正后的模型诊断图

结果如图 1-3 所示，从中可以看到各种诊断图形的趋势都变得更好了。具体地说，残差图中残差的大小相互之间更加可比了，而且更加杂乱无章，没有明显的趋势可言；QQ 图更加趋向于一条直线；Cook 距离图中也没有非常异常的点了。

最后，我们要对自变量的多重共线性程度予以分析。我们之所以关心这个问题，是因为如果有一个或者多个自变量可以被其他自变量很好地线性表出，那么我们的估计量的可靠性会很差。统计上对多重共线性程度是通过方差膨胀因子(variance inflation factor, VIF)来度量的，它的具体定义如下：

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

其中，R_j^2 是将 x_j 当做因变量，其他解释性变量作为自变量而拟合的线性模型所产生的判决系数 R^2。VIF_j 反映了在多大程度上第 j 个自变量所包含的信息已经被其他自变量所覆盖。因此，如果一个自变量对应于很大的 VIF 值，那么盲目地在模型中加入此变量，会严重地降低模型的效率。这是因为模型会很难区分该自变量同其他自变量的作用，因此相应的回归系数估计会很不精确。具体地，我们可以通过以下的 R 程序进行分析：

```
> library(car)
> round(vif(lm2),2)
  ROEt    ATO     PM    LEV GROWTH     PB    ARR    INV  ASSET
  1.09   1.15   1.40   2.70   1.00   2.63   1.28   1.02   1.09
```

可以看到所有参数估计的 VIF 值都远远小于 10，并且非常接近 1。因此，这里我们不用担心多重共线性的问题。

综上所述，我们可以认为此模型基本满足线性模型的所有重要假设。这为我们的进一步分析奠定了坚实的基础。

第七节　变量选择

如果我们的分析目的仅仅在于确立一些能够显著影响下年度盈利能力的财务指标的话，我们的分析基本上可以到此为止。但是，如果我们的终极目的是预测的话，就不可以了。这是因为，我们还不知道哪些变量应该被放在模型中作预测，一个简单而直观的答案是，就把以上三个显著性变量放入模型，而删除其他变量就可以了。但是，事实上这并不是最好的办法。这是因为，我们的假设检验只能告诉我们，这三个显著的财务指标非常重要，但是它们无法排除其他变量也有预测能力的可能。因此，我们要跳出假设检验的理论框架，寻求其他工具来搜寻最具有预测能力的模型。在这里，我们将介绍两种最为常用的选择变量的方法，即 AIC 和 BIC。

AIC(Akaike Information Criterion)

这是日本统计学家赤池(Akaike)根据极大似然估计原理提出的一种常用的选择标准：

$$\mathrm{AIC} = n\left\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\right\} + 2 \times (p + 1)$$

其中，RSS 是拟合残差平方和，p 是选入模型的变量个数，n 为样本量。当选入

模型的变量增加时，式中的拟合残差平方和 RSS 是减小的。由于自然对数是单调增函数，所以整个第一项是减小的，但是第二项随着选入模型的变量增加而增大。当由变量增加带来的方差减少的作用大于变量增加带来的惩罚时，AIC 的值逐渐减小。而当变量个数达到一定数目，由变量增加带来的惩罚大于变量增加带来的方差减小时，AIC 的值将会逐渐增加。因此，使用 AIC 选择变量的原则是：使 AIC 达到最小的模型是“最优”模型。

BIC(Bayesian Information Criterion)

$$\mathrm{BIC} = n\left\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\right\} + \log(n) \times (p+1)$$

使用 BIC 选择变量的原则是：使 BIC 达到最小的模型是“最优”模型。与 AIC 准则函数相比，BIC 第二项变为 $\log(n) \times (p+1)$。当 $n \geqslant 8$ 时，$\log(n) > 2$，因此 BIC 的惩罚项比 AIC 的惩罚项的力度要大，往往 BIC 选出来的变量个数少于 AIC 选出的变量个数。相比较而言，AIC 更为保守，选出的模型变量个数往往多于真实模型变量个数，即过拟合；而 BIC 确定的最优模型变量个数往往与真实模型更为接近。

由于这两种方法赋予模型复杂程度的权重不同，因此计算结果有可能不同。但是，在多数情况下都是大同小异的。当两者的结果不一样时，一般而言，AIC 选择的模型更加保守(即包含更多的变量)，而 BIC 恰恰相反。在 R 环境中，具体操作如下：

```
> lm.aic=step(lm1,trace=F)
> summary(lm.aic)

Call:
lm(formula = ROE ~ ROEt + LEV + GROWTH + ARR, data = a1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12744 -0.27107  0.02144  0.24905  1.22067

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.384554   0.018384  20.918  < 2e-16 ***
ROEt         0.546030   0.034258  15.939  < 2e-16 ***
LEV         -0.028790   0.005556   5.182 3.21e-07 ***
GROWTH       0.015699   0.008071   1.945   0.0523 .
ARR         -0.034533   0.018101  -1.908   0.0570 .
---
```

以上所产生的是根据 AIC 标准挑选出的最优模型。从中可以看到，AIC 认为第 1 个变量(ROEt)、第 4 个变量(LEV)、第 5 个变量(GROWTH)以及第 7 个变量(ARR)对于预测来年 ROE 非常重要。所有选出的变量都在 0.10 的水平下显著。如果我们按照 BIC 来选择模型，那么可以在 R 中实现如下：

```
> lm.bic=step(lm2,k=log(length(a1[,1])),trace=F)
> summary(lm.bic)

Call:
lm(formula = ROE ~ ROEt + LEV, data = a1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06999 -0.28107  0.02201  0.26315  1.22279

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.383049   0.017940  21.352  < 2e-16 ***
ROEt         0.549988   0.034408  15.984  < 2e-16 ***
LEV         -0.029561   0.005576  -5.301 1.73e-07 ***
```

从中可以清楚地看到，BIC 对于变量的选择更加苛刻。它和 AIC 一样，认为第 1 个变量（ROEt）和第 4 个变量（LEV）对于预测来年 ROE 非常重要，但是它不认为第 5 个变量（GROWTH）和第 7 个变量（ARR）也很重要。那么，对于一个具体的问题，到底是 AIC 所选择的模型预测精度好，还是 BIC 选择的模型预测精度好？很遗憾，对此问题没有一个一致性的答案，只能够具体数据具体分析。

如果我们从保守的角度，接受 AIC 所选择的模型，那么可以看到：在 0.10 的显著性水平下，当年净资产收益率（ROEt）和主营业务收入增长率（GROWTH）较高的公司，趋向于拥有更好的来年盈利能力。另一方面，如果公司债务资本比率（LEV）过高或者主营业务收入中应收账款所占比例（ARR）过大，那么下年度净资产收益率会受到一定的负面影响。

第八节　模型预测

在充分建立可靠模型的基础上，我们就可以对新的观测进行预测了。比方说，对于我们的案例，如果我们获得一个新的公司当年的财务数据，我们希望能够对它的来年盈利状况予以预测。为了方便起见，我们设新观测点 $x_0=(x_{01},x_{02},\cdots,x_{0p})'$，因变量 y 的预测值：

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\hat{\beta}_2x_{02}+\cdots+\hat{\beta}_px_{0p}=x_0'\hat{\beta}$$

在实际应用中，仅仅给出这样一个点预测（point prediction）往往是不够的，因为我们不知道该预测的精度。因此，人们往往还需要一个关于 $\hat{y}_0$ 的预测区间（prediction interval）：

$$x_0'\hat{\beta}\pm t_{n-p-1}\hat{\sigma}\sqrt{1+x_0'(X'X)^{-1}x_0}$$

其中，t_{n-p-1} 是自由度为 $n-p-1$ 的 t 分布的分位值。而另一方面，如果我们关

心的不是预测区间，而是关于 $\hat{y}_0$ 的期望（即 $E(\hat{y}_0) = x_0'\beta$）的置信区间（confidence interval），那么应该是：

$$x_0'\hat{\beta} \pm t_{n-p-1}\hat{\sigma}\sqrt{x_0'(\overline{X'X})^{-1}x_0}$$

请注意，预测值的置信区间的范围要大于期望值置信区间的范围。这是因为预测值的方差等于期望值的方差加新产生的误差项的方差，因此，其置信区间的范围较大。对于我们的案例，我们用 2003 年的数据作为检验数据，因此首先对该数据准备如下：

```
> a2=a[a$year==2003,-1]
> round(a2[1:5,],3)
      ROEt    ATO    PM    LEV GROWTH     PB    ARR    INV  ASSET    ROE
501 -1.151 -0.331 0.299  4.085  0.188 11.919  0.004  0.078 21.492 -0.403
502  0.338 -0.611 0.300  1.402  5.369 18.418 -0.669  0.167 20.456  0.211
503  0.722  0.794 0.016 -2.929  0.749 -20.886 -0.733  0.327 21.532  1.085
504  0.633  0.980 0.345 -2.832  1.107  -6.290 -0.985  0.346 22.069  0.814
505 -0.039 -0.236 0.366 -0.961  3.340 -11.308 -0.171 -0.174 20.860 -0.011
>
```

基于三个不同的模型（全模型、AIC 选择的最优模型以及 BIC 选择的最优模型），我们可以通过以下命令予以预测：

```
> y1=predict(lm2,a2)
> y2=predict(lm.aic,a2)
> y3=predict(lm.bic,a2)
> y0=a2[,10]
```

为了便于比较，我们还记录了每个公司 2003 年当年的盈利状况，并考虑如果不用线性模型而直接用 2003 年的盈利预测 2004 年，结果会怎样。所有的预测结果如下所示：

```
> r0=y0-a2$ROEt
> r1=y0-y1
> r2=y0-y2
> r3=y0-y3
> resid=abs(as.data.frame(cbind(r0,r1,r2,r3)))
> sapply(resid,mean)
       r0        r1        r2        r3
0.4157320 0.2944518 0.2937247 0.2947862
```

从中不难发现，所有基于模型的预测都要优于仅仅依靠当年净资产收益率的预测。具体地说，如果我们仅仅考虑利用公司当年的盈利能力简单预测来年，那么其平均绝对预测误差为 0.4157。而如果考虑了全模型，那么该预测误差下降为 0.2945。在此基础上经过 AIC 变量选择后的模型预测精度为 0.2937，经过 BIC 变量选择后的模型预测精度为 0.2948。

综上所述，线性模型的预测结果远远优于仅用公司上一年度的净资产收益率进行预测的预测结果。而基于线性模型的三个预测结果相差无几。但是，同全模型相比，AIC 或者 BIC 所使用的模型相对简单，同时也能够为我们深入了

解哪些财务指标重要提供理论依据。

第九节　简单分析报告

上市公司净资产收益率预测分析报告

内容提要　本报告利用上市公司当年的公开财务指标对其来年的盈利状况予以分析和预测。从我们的分析结果发现，公司当年的净资产收益率(ROEt)、债务资本比率(LEV)、主营业务收入增长率(GROWTH)以及应收账款/主营业务收入(ARR)四个财务指标，尤其是前两个财务指标，对于预测公司下一年的净资产收益率(ROE)非常重要。这四个财务指标主要取决于公司的资本结构和主营业务状况。基于本报告的分析结果，投资者和管理者可以利用上市公司当年的公开财务指标了解公司的投资风险、发展状况等信息，从而进行合理的投资规划。

一、研究目的

在金融市场上，如何利用上市公司当年的公开财务指标对其来年的盈利状况予以预测，是一个非常重要的问题。因为对该问题的合理回答，可以对投资者了解企业的盈利模式、风险大小以及进行正确的投资帮助甚大，而管理者可以根据预测结果对企业的发展规划、资源配置等方面进行合理的规划和部署。本报告将对中国股市的数据予以分析，找出对上市公司来年净资产收益率进行预测的方法，并根据分析结果提出有意义的结论和建议。

二、数据来源和相关说明

为实现合理预测上市公司来年的盈利状况这个目标，我们需要有效利用上市公司的历史财务数据，对其来年的净资产收益率(ROE)，即本分析报告中的因变量，予以大概的估计。我们选取了下列财务指标作为本分析报告中的自变量：

- 公司当年的净资产收益率(ROEt)：该财务指标直接反映了公司当年的盈利状况。可以预期，当年表现好的公司，其下年度的表现也趋向于较好。
- 资产周转率(ATO)：该财务指标综合评价了企业全部资产的利用效率。
- 主营业务利润/主营业务收入(PM)：该财务指标反映了公司收入的质量。

- 债务资本比率(LEV):该财务指标反映了公司的基本债务状况。
- 主营业务收入增长率(GROWTH):该增长率指标反映了公司的成长状况。
- 市倍率(PB):该财务指标反映了预期的公司未来成长率。
- 应收账款/主营业务收入(ARR):该财务指标反映了公司当年尚未实现的主营业务收入,从一定程度上说明了公司的盈利质量。
- 存货/资产总计(INV):该比率指标反映了公司的存货状况。
- 对数变换后的资产总计(ASSET):简称资产总计,反映了公司的规模。

我们随机选取深圳股票市场和上海股票市场2002、2003年度的各500个样本来进行分析。其中,模型的建立主要是基于2002年的训练样本,而2003年数据主要用来检验模型的预测精度。

三、描述性分析

为了获得对数据的整体了解,我们对数据进行简单的描述性分析,得到表1-1。

表1-1 样本描述

变量名	均值	最小值	中位数	最大值	标准差
ROEt	0.068	-1.390	0.080	1.421	0.519
ATO	0.430	-0.928	0.438	1.927	0.460
PM	0.211	-0.424	0.218	0.698	0.181
LEV	0.709	-7.941	0.560	9.362	3.182
GROWTH	0.331	-5.962	0.379	6.092	2.120
PB	2.127	-20.816	2.271	32.591	9.513
ARR	0.201	-2.601	0.223	3.187	0.949
INV	0.100	-0.264	0.102	0.431	0.122
ASSET	21.066	18.629	21.057	23.414	0.855
ROE	0.410	-1.161	0.420	5.285	0.545

从表1-1中的描述性统计可以看出:

公司当年的净资产收益率(ROEt)介于-1.390与1.421之间,其平均水平约为0.068(平均值)和0.080(中位数),标准差为0.519。而公司下年度的净资产收益率(ROE)介于-1.161与5.285之间,其平均水平约为0.410(平均值)和0.420(中位数),标准差为0.545。从中位数可以看出,超过半数的公司有正的净资产收益率(ROE);同当年相比,公司下一年的净资产收益率(ROE)

有一定的增长，而且不同公司之间的差距在扩大。

资产周转率（ATO）的均值（0.430）和标准差（0.460）从一个侧面反映出大多数公司资产的平均利用水平。主营业务利润/主营业务收入（PM）的均值（0.211）和标准差（0.181）反映了大多数公司的平均利润水平。值得注意的是，债务资本比率（LEV）的取值范围较大（从 -7.941 到 9.362），而且标准差也较大（3.182），这表明不同的公司之间债务水平差别较大。主营业务收入增长率（GROWTH）也有较大的取值范围（从 -5.962 到 6.092）和标准差（2.120），这表明各个公司所处的发展阶段呈现出多样化。从市倍率（PB）的均值（2.127）和标准差（9.513）可以看出，市场认为大多数公司有较好的发展空间，但不同公司之间的差异比较显著。应收账款/主营业务收入（ARR）的均值（0.201）和标准差（0.949）反映出，在相当多的公司中应收账款在主营业务收入中所占的比例较大。从存货/资产总计（INV）的均值（0.100）和标准差（0.122）以及其取值范围（从 -0.264 到 0.431）可以看出，大多数公司对存货都有较强的控制，从而避免出现高存货率的情况。对数变换后的资产总计（ASSET）的标准差（0.855）反映出不同的公司在资产规模上的差距还是比较大的。

四、数据建模

1. 全模型分析

在本节中，我们用线性回归的分析方法建立模型，以此来寻找自变量（公司当年的 9 项财务指标）和因变量（公司下一年的净资产收益率，ROE）之间的关系。通过考察观测值的 Cook 距离，我们删除了第 47 个强影响点。利用剩余的数据，我们运用包括全部 9 个自变量的全模型对因变量进行估计，得到表 1-2 所示的估计结果。从表 1-2 中我们可以看到，模型 F 检验的 P 值非常小，表明该模型是显著的，即自变量和因变量之间确实存在一定的关系。另外，未调整的判决系数（R-square）为 31.29%，调整后的判决系数（R-square）为 30.03%，这都表明该模型对自变量和因变量之间的关系有一定的解释能力。通过考察各个自变量对应的 t 检验的 P 值，在 0.05 的置信水平下，我们可以断定下一年的净资产收益率（ROE）与当年净资产收益率（ROEt）和主营业务收入增长率（GROWTH）之间有着显著的正相关关系，与债务资本比率（LEV）显著地负相关，而对于其他变量暂时没有定论。

表 1-2 全模型

变量名	系数估计值	标准差	P 值
截距	0.454	0.528	0.390
ROEt	0.487	0.041	0.000
ATO	-0.015	0.048	0.758
PM	0.079	0.133	0.554
LEV	-0.040	0.011	0.000
GROWTH	0.020	0.010	0.039
PB	0.003	0.003	0.341
ARR	-0.026	0.024	0.285
INV	-0.020	0.168	0.906
ASSET	-0.003	0.025	0.901
残差项标准差 0.4557		模型 F 检验 P 值 <0.0001	
判决系数(R-square) 0.3129		调整的判决系数(R-square) 0.3003	

2. 模型选择与预测

从以上全模型的分析结果容易发现，有三个财务指标非常重要，但是我们不能排除其他变量也有预测能力的可能。因此，我们用两种最为常用的选择变量的方法，即 AIC 和 BIC，来选择最具有预测能力的模型。

如果使用 AIC 来选择模型，我们可以得到如下的模型估计结果，如表 1-3 所示。

表 1-3 AIC

变量名	系数估计值	标准差	P 值
截距	0.384554	0.018384	0.00
ROEt	0.546030	0.034258	0.00
LEV	-0.028790	0.005556	0.00
GROWTH	0.015699	0.008071	0.05
ARR	-0.034530	0.018101	0.06
残差项标准差 0.3818		模型 F 检验 P 值 <0.0001	
判决系数(R-square) 0.4205		调整的判决系数(R-square) 0.4158	

从表 1-3 中可以看到，AIC 认为第 1 个变量(ROEt)、第 4 个变量(LEV)、第 5 个变量(GROWTH)以及第 7 个变量(ARR)对于预测下一年的净资产收益率(ROE)非常重要。而且，选出的所有变量都在 0.10 的水平下是显著的，其判决系数(R-square)相对于全模型有所提高。

如果用 BIC 来选择模型，我们可以得到如下的模型估计结果，如表 1-4 所示。从表 1-4 中可以清楚地看到，BIC 认为第 1 个变量(ROEt)和第 4 个变量(LEV)对

于预测来年的净资产收益率非常重要，但是它不认为第 5 个变量（GROWTH）和第 7 个变量（ARR）也很重要。而且，BIC 选出的所有变量都在 0.01 的水平下是显著的。其判决系数相对于全模型也有所提高，但略微低于 AIC 所选出的模型。

表 1-4　BIC

变量名	系数估计值	标准差	P 值
截距	0.383049	0.017940	0.00
ROE1	0.549988	0.034408	0.00
LEV	-0.029560	0.005576	0.00
残差项标准差 0.384		模型 F 检验 P 值 <0.0001	
判决系数（R-square）　0.4116		调整的判决系数（R-square）　0.4092	

为了从三个不同的模型（全模型、AIC 选择的最优模型以及 BIC 选择的最优模型）中选出最具有预测能力的模型，我们用 2003 年的数据来对模型的预测能力进行检验。另外，我们也考虑最简单的预测方法，即仅用 2003 年的盈利预测 2004 年，称之为“直接预测”。通过计算得到模型的预测结果比较，如表 1-5 所示。

表 1-5　预测结果比较

模型	直接预测	全模型	AIC	BIC
平均预测误差	0.4157	0.2945	0.2937	0.2948

从表 1-5 中我们不难发现，所有基于模型的预测都要优于仅仅依靠当年净资产收益率的预测。具体地说，如果我们仅仅利用公司当年的盈利能力来简单预测来年，那么其平均绝对预测误差为 0.4157；而如果考虑了全模型，那么该预测误差下降为 0.2945。在此基础上，经过 AIC 变量选择后的模型预测精度为 0.2937，而经过 BIC 变量选择后的模型预测精度为 0.2948。

综上所述，线性模型的预测结果远远优于仅用公司上一年度的净资产收益率进行预测的预测结果。而基于线性模型的三个预测结果相差无几。但是，同全模型相比，AIC 或 BIC 所使用的模型相对简单，为我们深入了解哪些财务指标对于预测公司下一年的盈利能力更为重要提供了理论依据。进一步讲，从好的预测能力、简单和相对保守的角度来看，AIC 所选择的模型能够提供更多的理论思考。

五、结论及建议

从上述的分析结果可知，我国上市公司的财务信息为预测下一年的盈利能力提供了重要信息，而且表现出较好的预测能力。具体来说，从保守和谨慎的

角度来看，公司当年的净资产收益率（ROEt）、债务资本比率（LEV）、主营业务收入增长率（GROWTH）以及应收账款/主营业务收入（ARR）这四个财务指标，尤其是前两个，对于预测公司下一年的净资产收益率（ROE）非常重要。

进一步可以发现，上述四个财务指标主要取决于公司的资本结构和主营业务状况。我们推测原因如下：较低的负债比率使得公司在下一年的还债压力较轻，因而有更充足的现金流和更大的经营自由，从而容易获得较高的净资产收益率；主营业务状况在很大程度上取决于公司产品所处的生命阶段，处于成熟期的产品对于公司的盈利能力有很强的正向影响，而衰退期的产品容易带来负向的影响。因此，我们的结论验证了保持合理的负债水平、专注于主营业务以及积极开发新产品等策略对于公司发展的积极意义。

投资者和管理者可以利用上述分析结果了解投资风险和公司的发展状况等信息，从而进行合理的投资和管理规划。例如，投资者在进行投资分析时，需要特别关注公司的以上相关财务指标，了解公司的资本结构和主营业务状况，利用相关信息来预测公司下一年的盈利能力，并基于预测结果得出最有利的投资决策；而管理者可以通过控制上述的相关财务指标来有意地透露公司的发展规划等重要信息，进而影响投资者，使其投资决策与公司的发展规划相一致。当然，每一个公司都有其特殊情况，需要针对不同公司的具体情况进行更详细的分析。

[讨论总结]

本章以盈利预测为例，系统演示并讲解了普通线性模型。通过对本章的学习，读者应该能够了解：什么时候可以使用普通线性模型，以及如何使用。在 R 语言学习方面，读者应该了解：(1) 如何读入数据；(2) 如何对数据进行简单操作；(3) 如何作图等。由于篇幅限制，我们不可能对 R 语言作系统详细的介绍。由于 R 语言的语法同商业软件 $S+$ 极其相似，因此，希望深入了解 R 的读者可以查阅 Venables and Ripley(1994)。在统计理论方面，读者应该理解并掌握以下重要概念：因变量、解释性变量、普通线性模型、最小二乘估计、F 检验、t 检验、模型诊断、异常值、Cook 距离、模型选择(AIC、BIC)，还有外样本的预测检验等。渴望深入了解相关统计学理论的读者可以参阅 Rao(1973)以及 Draper and Smith(1981)。

附录　程序及注释

```
rm(list=ls())                                                    # 清理当前工作空间
a=read.table("D:/Practical Business Data Analysis/case/CH1/roe.txt",header=T)
                                                                 # 读入以空格为分隔符,并带有标题行的文本文件
round(a[1:10,],4)                                                # 用4位小数点的格式显示a中前10行的数据
a1=a[a$year==2002,-1]                                            # 从a中选出year为2002的数据,并删除第1列,然后赋值给a1
Mean=sapply(a1,mean)                                             # 计算a1中各列的均值
Min=sapply(a1,min)                                               # 计算a1中各列的最小值
Median=sapply(a1,median)                                         # 计算a1中各列的中位数
Max=sapply(a1,max)                                               # 计算a1中各列的最大值
SD=sapply(a1,sd)                                                 # 计算a1中各列的标准差
cbind(Mean,Min,Median,Max,SD)                                    # 将均值、最小值、中位数、最大值、标准差集中在一起展示
round(cor(a),3)                                                  # 计算相关系数,用3位小数点的格式展示
plot(a1$ROEt,a1$ROE)                                             # 画出ROEt和ROE之间的散点图
lm1=lm(ROE~ROEt+ATO+PM+LEV+GROWTH+PB+ARR-INV+ASSET,data=a1)
                                                                 # 用a1中数据拟合线性回归模型
summary(lm1)                                                     # 给出模型lm1中系数估计值、P值等细节
round(a1[1:10,],3)                                               # 用3位小数点的格式显示a1中前10行的数据
par(mfrow=c(2,2))                                                # 设置画图为2x2的格式
plot(lm1,which=c(1:4))                                           # 画出lm1中对应于模型检验的4张图,包括残差图、QQ图和Cook
                                                                   距离图
```

```
a1=a1[-47,]                                                    # 删除 a1 中第 47 行的观测
lm2=lm(ROE ~ ROEt + ATO + PM + LEV + GROWTH + PB + ARR + INV + ASSET,data=a1)
                                                               # 对新数据 a1 再次拟合回归模型为 lm2
plot(lm2,which=c(1:4))                                         # 画出 lm2 中对应下模型检验的 4 张图,包括残差图、QQ 图和 Cook 距离图
library(car)                                                   # 载入程序包 car
round(vif(lm2),2)                                              # 计算模型 lm2 的方差膨胀因子,用 2 位小数点的格式展示
lm.aic=step(lm2,trace=F)                                       # 根据 AIC 准则选出最优模型,并赋值给 lm.aic
summary(lm.aic)                                                # 给出模型 lm.aic 中系数估计值、P 值等细节
lm.bic=step(lm2,k=log(length(a1[,1])),trace=F)                 # 根据 BIC 准则选出最优模型,并赋值给 lm.bic
summary(lm.bic)                                                # 给出模型 lm.bic 中系数估计值、P 值等细节
a2=a[a$year==2003,-1]                                          # 从数据 a 中选出 year 为 2003 的观测,并删除第一列,赋值给 a2
round(a2[1:5,],3)                                              # 用 3 位小数点的格式展示 a2 的前 5 行数据
y1=predict(lm2,a2)                                             # 用全模型 lm2 对 a2 进行预测
y2=predict(lm.aic,a2)                                          # 用模型 lm.aic 对 a2 进行预测
y3=predict(lm.bic,a2)                                          # 用模型 lm.bic 对 a2 进行预测
y0=a2[,10]                                                     # 选出 a2 中的第 10 列,即当年的 ROE
r0=y0 - a2$ROEt                                                # 用当年 ROE 对下年进行预测的残差
r1=y0 - y1                                                     # 用全模型 lm2 预测的残差
r2=y0 - y2                                                     # 用模型 lm.aic 预测的残差
r3=y0 - y3                                                     # 用模型 lm.bic 预测的残差
resid=abs(as.data.frame(cbind(r0,r1,r2,r3)))                   # 计算残差的绝对值
sapply(resid,mean)                                             # 计算不同模型的平均绝对偏差,即对残差的绝对值取平均
```

第二章
方差分析

- 案例介绍
- 描述性分析
- 单因素方差分析
- 多重比较
- 双因素简单可加模型
- 双因素交互作用模型
- 多因素方差分析
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过一个商品房价格分析的实际案例，详细介绍方差分析这种重要的统计回归模型。它主要处理的是因变量为连续型数据而解释性变量为离散型数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用方差分析；(2) 方差分析的基本统计学理论；(3) 相关理论在统计学软件 *R* 中的应用；(4) 相应的统计分析报告的撰写。本章初次涉及的重要统计学概念如下：离散型解释性变量、方差分解、多重比较、单因素以及多因素方差分析、交互作用。

第一节　案例介绍

方差分析(analysis of variance，ANOVA)是另外一种在实践中被广泛运用的统计方法。它同前一章所讲的回归分析关系密切，但是又不完全相同。具体地说，如果有一个人们感兴趣的指标(因变量)，其变化可能受到众多离散型因素(如性别、种族、职业等)而不是连续型因素(如年龄、收入、价格等)的影响，那么我们应该考虑采用方差分析的方法。此时，线性回归的分析方法不再适用。原因很简单，对于离散型解释性变量，简单地使用回归模型来刻画它们同因变量之间的"线性"关系是不恰当的。因此，我们可以粗略地认为，方差分析就是线性回归的一种延续。

在本章中，我们仍然用一个实际案例来具体讲解并演示方差分析的各个方面。这是一个关于房地产价格的案例。北京市房地产市场是我国房地产市场中最为发达，也是最具有代表性的几个房地产市场之一。根据《2002 年固定资产投资统计年报》，北京市 2002 年商品房销售总额为 813.8 亿元，仅次于上海的 815.0 亿元，位居全国第二。从本案例所收集的数据中可以看出，北京市商品房的销售价格差异巨大，从最低的 4 000 元/平方米一直到最高的 20 000 元/平方米。对于这样一个事实，人们会很自然地产生一个疑问：是什么样的因素在影响着北京市商品房的销售价格，价格上的巨大差异是怎样产生的呢？

为了找出适合北京市情况的商品房定价因素，我们从"搜房网"(www.soufun.com)上随机选取了北京地区 2003—2004 年度开盘的新楼盘共 506 个，其中

相当多的数据不完整或者有明显错误。在进行数据清理后，我们最终得到200个合格的楼盘样本。我们希望能够基于这样一个公开的实际数据，建立恰当的经济计量模型，从数量上刻画房地产销售价格同各个影响因素之间的关系。进一步讲，我们还希望借此模型确定影响商品房销售价格的重要因素，并量化这些影响因素的相对重要性。我们的研究结果可以为商品房购房者提供科学可靠的价格参考依据，也可以为房地产开发商的新楼盘定价提供更多参考，还可以为相关机构的价值评估提供理论依据。具体地说，我们的数据中包含了表2-1中的信息。

表2-1 变量说明

变量类型	变量含义	变量名	水平数
连续型	销售均价(元/平方米)	price	无
	容积率(%)	rong	无
	绿化率(%)	lv	无
	小区总建筑面积(平方米)	area	无
	小区停车位住户比例(位/户)	ratio	无
离散型	所在区县	dis	共七个区县(七个主城区)
	所在环线	ring	共五环(<2、2-3、3-4、4-5、>5)
	物业类别	wuye	共两种(普通、公寓)
	装修状况	fitment	共三种(毛坯、粗装修、精装修)
	建筑类别	contype	共四种(板楼、塔楼、板塔结合、高层)

从表2-1可以看到，在我们的数据所涵盖的所有变量中，有一个是我们特别感兴趣的，试图通过其他变量来解释的，那就是销售均价。这也就是我们方差分析中的因变量(dependent variable)，而其他的所有变量都是自变量(independent variable)。

值得注意的是，我们的解释性变量大致分为两类。一类是连续型的解释性变量(如容积率、绿化率等)。如果我们单纯地想通过这些连续型变量来解释销售均价，那么我们只要使用前一章所学的回归模型就可以了。但是，我们还应该注意到另外一类解释性变量，即离散型变量(如所在区县、所在环线等)。如果我们单纯地想通过这些离散型变量来解释销售均价，那么我们就要使用本章将学到的方差分析。但是，如果我们希望将所有的变量(不论连续型的还是离散型的)都结合起来解释销售均价，那应该怎么做呢？这将是下一章协方差分析(analysis of covariance，ANCOVA)将要讲述的内容。因此，本章将集中精力讨论如何利用这些离散型变量来对销售均价予以合理解释。

我们称每一个我们所考虑的离散型变量为一个因素(factor),如所在区县是影响销售价格的一个重要因素。但是,这个因素(所在区县)有七个不同的取值(朝阳、崇文、东城、西城、宣武、海淀、石景山)。这说明,"所在区县"这个因素总共有七个不同的水平。值得注意的是,这七个水平之间没有必然的数量关系,我们无法对它们进行任何形式的数学运算(如"海淀 - 朝阳 + 石景山 × 宣武")。因此,该因素的本质是将我们所研究的总体(所有北京市的商品房)根据其水平的取值(即所在区县)分成了总共七组。而一个典型的问题就是,商品房在不同区县之间的平均销售价格有没有显著区别?如果有,那么我们就有理由相信,"所在区县"是影响房屋销售价格的一个重要因素。否则,我们就不可以轻易地下此结论。

需要注意的是,我们这里的简单分析仅仅涉及了一个因素(所在区县),因此被称为单因素方差分析(one-way ANOVA)。如果我们同时还考虑了另一个因素(如所在环线),则被称为双因素方差分析(two-way ANOVA)。当然,我们完全可以考虑更多的离散因素(如物业类别等),这些被统称为多因素方差分析(multi-way ANOVA)。在本章中,我们将集中精力详细讨论单因素以及双因素方差模型,其分析方法可以很容易地被推广到多因素方差模型。

第二节 描述性分析

和前一章所讲的线性回归一样,在进行正式的方差分析之前,我们非常有必要进行一些简单的描述性分析。这些简单的描述性分析可以帮助我们获得对数据的整体认识,发现异常数据,并进而指导我们的下一步分析。

假定数据保存在名为"real.csv"的文件中,存放路径为"D:\Practical Business Data Analysis\case\CH2\real.csv",那么我们可以在 *R* 的编程环境中输入以下语句读入数据,并画出散点图(如图 2-1 所示)。

```
> rm(list=ls())
> a=read.csv("D:/Practical Business Data Analysis/case/CH2/real.csv",header=T)
> attach(a)
> pairs(a[,c(1:6)])
```

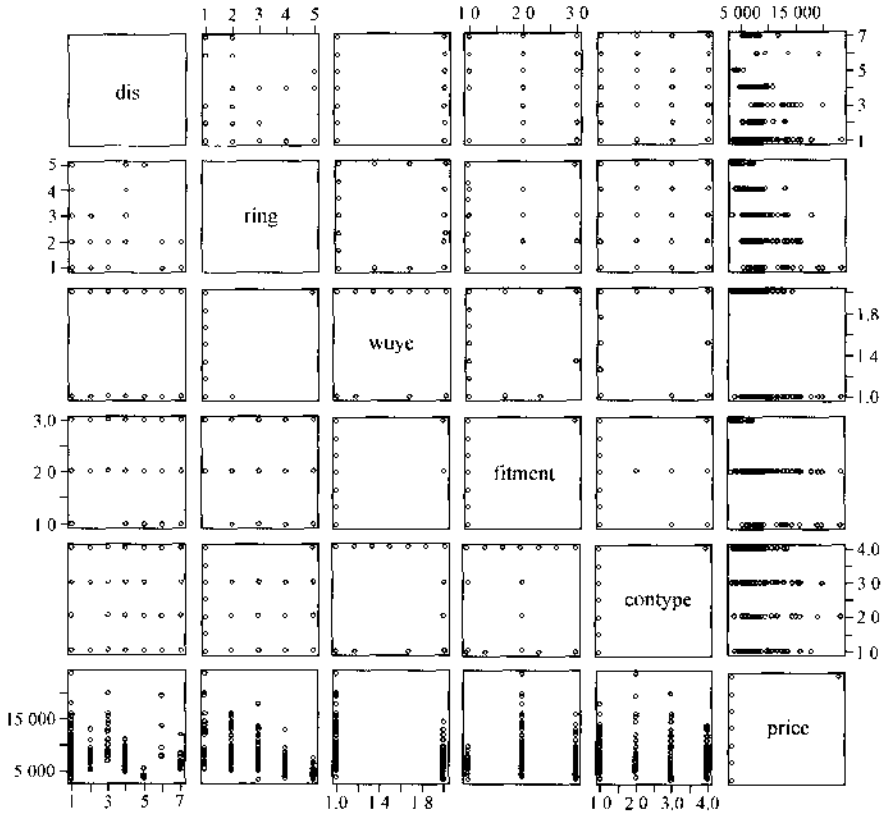


图 2-1　不同变量的散点图

大家可以看到，同连续型解释性变量不同，如果我们的解释性变量是离散型的（如本例所示），那么散点图似乎并不是一种非常有效的展示方法，从中很难看出非常有意义的趋势。因此，我们需要考虑其他的作图方法来更好地观察因变量与这一类型的自变量的关系。以楼盘所处环线为例，我们可以通过以下的简单命令，画出自变量楼盘所处环线（ring）不同水平下的楼盘销售均价（price）分布的盒状图（或称箱图），如图 2-2 所示。

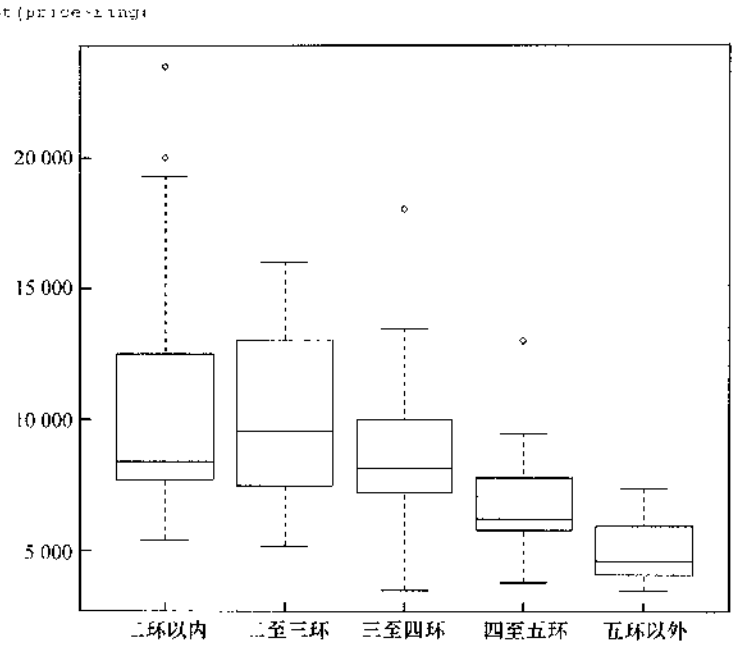


图 2-2　楼盘销售均价的盒状图

在对图 2-2 作解释之前，我们需要先讲一下如何理解盒状图（boxplot）。以图 2-2 为例，我们看到共五个盒状图，每一个对应于一个特定的地理位置（如二环以内）。以二环以内为例（即第一个盒状图），该盒状图最重要的信息来自于盒子中间的横线。该横线在纵轴上的取值代表了二环以内商品房均价的中位数，请注意：不是算术平均值。人们之所以在盒状图中反映中位数而不是均值，主要是因为中位数对异常值非常不敏感。从某种意义上说，中位数的取值反映了该地区房屋价格的某种平均水平（但绝不是算术平均水平）。然后，我们可以看盒子的厚度。每一个盒子的下边缘对应的纵轴上的取值是该水平下因变量的第一个四分位数，即 25% 的分位数；其上边缘就对应了因变量的第三个四分位数，即 75% 的分位数。那么，盒子的厚度（即第三个四分位数与第一个四分位数之差）从某种意义上说就反映了该地区商品房价格的波动情况。最后值得注意的是，每个盒子都有上下两根横线，该横线所画出的范围是在正态假设下数据取值的一个合理范围。如果一个数据落在了该范围以外（如二环以内所示），那么有两种可能的原因：第一，该数据也许是个异常值；第二，因变量并不服从正态假设。但是，无论如何，凡落在该范围以外的数据都值得注意。

在理解了盒状图后，我们现在可以看一下盒状图能为我们提供哪些有用的信息。从图 2-2 可以明显看到两个趋势：第一，随着地理位置从内环向外环延伸，商

品房均价的中位数(盒子中间的横线)呈现出明显的下降趋势。这说明,距离城市中心越远,商品房平均价格越低。第二,随着地理位置从内环向外环延伸,商品房均价的波动程度(盒子的厚度)也呈现明显的下降趋势。所以线性模型的同方差假设是显然不成立的,因此,我们考虑对因变量(商品房单位面积均价)实施对数变换。

这里值得一提的是,随着本书内容的进一步深入,大家会发现对数变换在统计分析中占据了很重要的地位。人们常常通过对数变换来获得更加可靠且准确的统计模型。具体地说,对数变换有以下优点:

第一,对数变换是实际研究中最常见,也最容易被接受的变换;

第二,过去的大量经验以及研究显示,对数变换在很大程度上具有稳定方差(variance stabilization)的作用,因此有望解决异方差的问题;

第三,对数变换能够将非负的变量(如商品房均价)转换成可正可负的实数变量,因此更有可能服从正态分布。

因此,我们可以考虑用对数变换后的商品房均价作为我们的因变量。在 *R* 环境中,可以具体实现如下:

```
> log.price=log(price)
> boxplot(log.price~ring)
```

对数变换后的盒状图如图 2-3 所示。从图中可以看出,异方差的问题得到了很大的修正,而单调下降的趋势得到了保留。因此,接下来的所有分析都是以对数变换后的单价作为因变量。

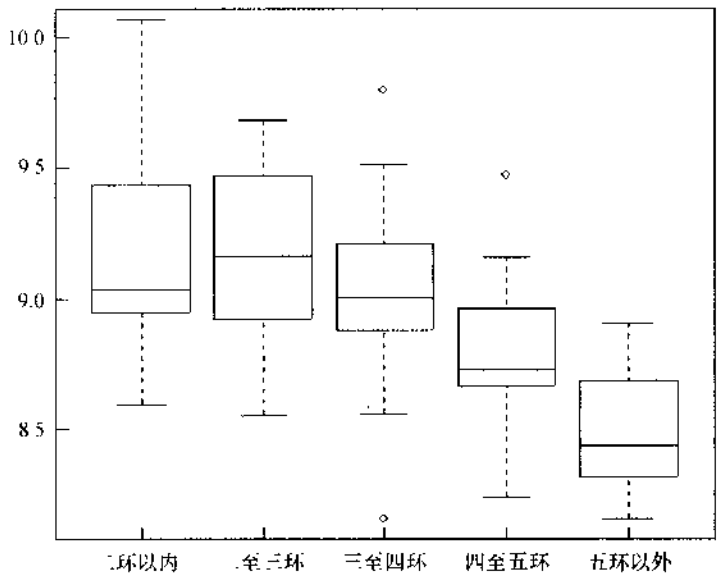


图 2-3 对数变换后的商品房均价盒状图

在对因变量销售均价使用了对数变换后，其他的离散型自变量（如所在区县、物业类别等）也可以通过类似的命令和图形来进行分析，如图 2-4 所示。

```
> par(mfrow=c(2,2))
> boxplot(log.price~dis)
> boxplot(log.price~muye)
> boxplot(log.price~fitment)
> boxplot(log.price~buntype)
```

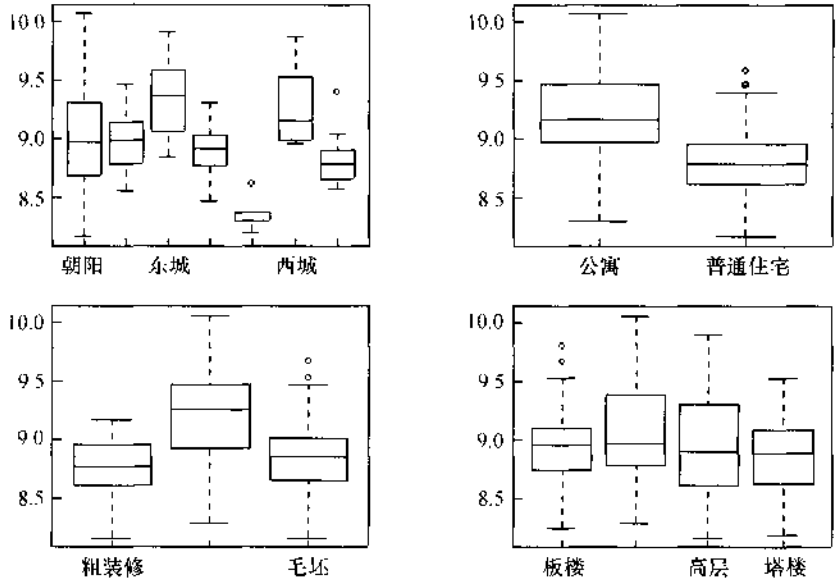


图 2-4　其他变量盒状图

从图 2-4 我们不难发现以下有意义的初步结论：

（1）城区的不同不仅极大地影响着房屋的均价，而且还影响着房屋价格的波动程度。例如，朝阳区（第一个盒子）的房屋价格波动程度远远高于其他地区，此外，东城区（第三个盒子）的房屋平均价格远远高于其他城区。还值得注意的是，有一个城区（第五个盒子）的样本量偏小，以至于中位数同盒子的上边缘重合了。

（2）公寓的平均价格明显高于普通住宅，但波动程度相当。

（3）精装修房屋的平均价格明显高于粗装修以及毛坯房，而且价格波动较大。此外，值得注意的是，粗装修房屋的均价并没有明显高于毛坯房。

（4）不同的房屋类型（如板楼、塔楼、板塔结合）对房屋的平均价格及其波动程度影响都很小。

然后，我们对因变量进行一些简单描述。我们特别想知道样本量在各个因素及其各个水平之间的分布情况，因为一个良好的具有代表性的样本应该满足

一定的样本分布均匀性。在 *R* 中具体实现如下：

```
> summary(a[,c(1:5)])
     dis             ring           wuye        fitment      contype
 朝阳  :86   二环以内:33   公寓     : 83   粗装修: 20   板楼     :66
 崇文  :16   二至三环:48   普通住宅:117   精装修: 67   板塔结合:26
 东城  :16   三至四环:42                  毛坯  :113   高层     :41
 海淀  :53   四至五环:53                               塔楼     :67
 石景山: 9   五环以外:24
 西城  : 5
 宣武  :15
```

从以上的数据中我们可以看到，我们的样本在地理位置（即几环到几环）、房屋类型（公寓还是普通住宅）、装修情况（粗装修、精装修还是毛坯）以及建筑类型（板楼还是塔楼）等方面都分布得较为理想，没有出现某因素的个别水平样本量偏小的情况。但是，值得注意的是，样本在不同城区间的分布是非常不均匀的。绝大多数样本来自于朝阳区和海淀区。这有可能是因为其他城区的商品房开发相对较少，从而客观地反映到了样本分布上。如果实际情况真是如此，那么我们的样本仍然具有很好的代表性。但是，如果这是因为我们数据采集的人为偏差造成的，那么我们就要小心处理了。无论如何，个别城区（如西城）太小的样本量，确实会在一定程度上影响分析的精度。一个可能的解决方法是将样本量太小的城区整合成为一个较大的城区（称为其他城区）。为方便起见，我们仍采用原来的划分。

第三节　单因素方差分析

方差分析的基本目的是要比较不同水平下的自变量的均值是否相等。在我们的案例中，就是要比较不同城区、不同地理位置、不同建筑类别等的商品房价格的平均水平有没有显著差异，如果有，差异多大？为了简化我们的研究，这里先介绍单因素方差分析（one-way ANOVA）。对于某一个因素（如环线位置），为了研究在它的不同水平下因变量的均值是不是相等，我们建立以下模型：

$$y_{ij} = u_i + \varepsilon_{ij}$$

其中，y_{ij}为第 i 个水平下第 j 个观测的因变量值（如第 i 个环线的第 j 个楼盘的销售均价），u_i 是第 i 个水平下因变量的均值（即该环线所有楼盘的总体销售均价），ε_{ij}表示第 i 个水平下第 j 个观测的因变量实际值和均值之间的残差，服从 $N(0,\sigma_i^2)$ 的正态分布。因此，在我们的案例中，该因素（即“所在环线”）共涉及五个水平（即五个不同的环线位置）。此外，根据上一节的描述性分析，我们可

以知道每个水平下的具体样本量分别为 $n_1=33, n_2=48, n_3=42, n_4=53, n_5=24$。在现实生活中，某个因变量不可能仅仅受某一个或某几个自变量的影响，总会有其他的因素或者随机的原因（如政策、自然情况等）无法被完全地考虑到，而这些无法被完全考虑到的因素，统统归结到了残差项 ε_{ij} 中。

根据上面的模型我们可以认为，观测值也就是房屋销售价格之间肯定存在着差异，而差异来源于两个方面：一方面是由因素的不同水平造成的，如环线的不同带来价格的不同，我们称为系统性差异；另一个方面是由于抽选样本的随机性而产生的差异，如即使同一环线不同楼盘的销售价格仍然不同，我们称为随机性差异。两个方面的差异可以用两个方差来计算，一个是水平内部方差，一个是水平之间方差。这两种方差可以通过方差分解的方法得到：

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij}-\bar{y}_i)^2+(\bar{y}_i-\bar{y})^2+2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})] \\ &= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2}_{SSA} \end{aligned}$$

其中，$\bar{y}_i = n_i^{-1}\sum_{j=1}^{n} y_{ij}$，表示第 i 种水平下的样本均值；$\bar{y} = \left(\sum_{i=1}^{k} n_i\right)^{-1}\sum\sum y_{ij}$，表示所有水平下的样本总均值；$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2$，称为总平方和（total sum of square）；$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$，称为残差平方和（sum of square due to error），反映水平内部方差；$SSA = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$，称为因素平方和（sum of squares due to the factor A），反映水平之间方差。

同线性回归模型类似，在使用方差分析的方法分析问题的时候，我们一般有常方差的假定，即 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2$。这表示因变量在自变量的各个水平下方差都相等，且等于总体的方差。也就是说，残差项 ε_i 的方差不依赖于自变量的水平，即为一个常数。对于我们的案例，以自变量“所在环线”为例，该假设意味着不同环线上房屋均价的波动程度是相似的。但是，实际上这种假设往往不能成立。越靠近繁华地区，如三环以内的房屋均价普遍比较高，而五环以外的房屋均价相对较低，这样三环以内的房屋均价波动也就会相对较大，而五环以外的房屋均价波动相对较小。但是我们可以看到，进行对数变换后，这个问题在很大程度上得到了修正。此外，即使在现实中方差相等的假设往往不能

成立，但方差分析也仍然能够给我们提供很多有用的信息。

如果我们的常方差假设是成立的，并且关于残差项的正态性假设也是合理的，那么我们可以通过一个 F 检验来检验如下假设：

$$H_0:\mu_1 = \mu_2 = \cdots = \mu_k \quad vs \quad H_1:\text{not } H_0$$

其中，原假设表示该因素（环线位置）的不同取值对因变量（商品房销售价格）的平均水平没有影响，因此，该因素是不重要的。相反，如果对立假设是正确的，那么我们就知道，该因素至少会有两个水平的因变量平均值是不一样的，即至少有两个环线位置的销售均价是很不一样的。这说明，该因素（环线位置）在统计上是很重要的。

如果我们的原假设是正确的，那么所分离出来的残差平方和（SSE）和因素平方和（SSA）分别具有以下的统计性质：① SSE 和 SSA 相互独立；② $SSE/\sigma^2 \sim \chi^2_{n-k}$；(3) $SSA/\sigma^2 \sim \chi^2_{k-1}$。由以上三条性质就可以得到如下 F 统计量：

$$F = \frac{SSA/(n-k)}{SSE/(k-1)} \sim F(n-k, k-1)$$

我们的方差分析也是利用这一结果。分子是因变量的变化中能被自变量的变化所解释的部分，分母是除自变量的变化外其他随机因素对因变量的变化解释的部分。直观地说，自变量所能解释的变化占因变量总变化的比例越大，它对因变量变化的影响越大，而随机因素带来的变化占因变量总变化的比例越大，则自变量对因变量的影响越小。对于我们的案例，在 R 环境下，我们可以轻松实现 F 检验如下：

```
> lm1=lm(log.price~as.factor(ring))
> library(car)
> Anova(lm1,type="III")
Anova Table (Type III tests)

Response: log.price
                Sum Sq  Df  F value    Pr(>F)
(Intercept)    2763.52   1 31791.58 < 2.2e-16 ***
as.factor(ring)   9.97   4    28.66 < 2.2e-16 ***
Residuals        16.95 195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

上面程序的运行结果中，被圈起的一列即为对应的 SSA 和 SSE，其后还有进行上述假设检验所对应的 F 值。从结果中可以看出，所在环线对楼盘销售均价是有显著影响的，显著性水平接近于零。这说明，环线位置确实是影响商品房销售价格的一个重要因素。换句话说，我们确信至少有两个不同的环线位置，它们的商品房平均销售价格不一样。但是，我们仍然不知道：① 到底哪两个或者哪几个环线位置的商品房均价不一样？② 如果两个环线位置不同，那么它们的商品房均价到底有多大差别？因此，我们需要进行更加具体的估计与分析。

在 *R* 中继续输入命令如下：

```
> summary(lm1)

Call:
lm(formula = log.price ~ as.factor(ring))

Residuals:
     Min       1Q   Median       3Q      Max
-0.86428 -0.17369 -0.04323  0.19992  0.91436

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               9.151119   0.051324 178.302  < 2e-16 ***
as.factor(ring)二至三环   0.005221   0.066671   0.078    0.938
as.factor(ring)三至四环  -0.126323   0.068584  -1.842    0.067 .
as.factor(ring)四至五环  -0.375584   0.065378  -5.745 3.49e-08 ***
as.factor(ring)五环以外  -0.645476   0.079095  -8.161 4.04e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2948 on 195 degrees of freedom
Multiple R-squared: 0.3702,     Adjusted R-squared: 0.3573
F-statistic: 28.66 on 4 and 195 DF,  p-value: < 2.2e-16
```

在我们进行统计推断以前，我们首先需要理解上面的统计报表。首先注意到，环线位置这个因素有五个水平，分别是二环以内、二至三环、三至四环、四至五环以及五环以外。但是，在上面的报表中，“二环以内”却不见了！其实不然，我们没有看到“二环以内”是因为“二环以内”就是我们的截距项（即 9.15）。由此我们知道，二环以内的商品房平均售价大约为 $\exp(9.15)=9414$ 元/平方米。该估计量的标准误差为 0.051。由此可得，二环以内商品房平均售价的 95% 的置信区间大约为：

$$(\exp\{9.15-1.96\times0.051\},\exp\{9.15+1.96\times0.051\})=(8519,10405)$$

那么，二至三环的商品房平均价格应该是多少呢？我们注意到对应于“二至三环”的估计量为 0.005。这能说明二至三环的商品房平均价格是 $\exp(0.005)=1$ 元/平方米吗？当然不可能。这个 0.005 是估计的“二环以内”同“二至三环”的价格差异。该价格差异的标准差为 0.067，相应的统计检验所产生的 *P* 值较大，等于 0.938。这说明，“二环以内”同“二至三环”没有明显的价格差异。那么，三至四环的商品房价格如何呢？我们注意到相应的估计量为 −0.126，其相应的统计检验高度显著。这说明，二至四环的商品房价格明显低于“二环以内”这个标准（请注意：不是“二至三环”）。如果我们一定要估计，那么三至四环的商品房价格大约为 $\exp\{9.151-0.126\}=8300$ 元/平方米。类似地，我们可以知道四至五环的商品房平均价格为 $\exp\{9.151-0.375\}=6477$ 元/平方米，而五环以外的商品房平均价格为 $\exp\{9.151-0.645\}=4944$ 元/平方米。

第四节　多重比较

从以上的分析可以看到，为了更加清晰地了解环线对房价的影响，R 自动进行了以下四对比较：二环以内和二至三环、二环以内和三至四环、二环以内和四至五环以及二环以内和五环以外。每一对比较都有相应的 P 值。我们知道，如果我们只面对一个假设检验问题（例如，我只关心二环以内和三至四环之间有没有显著差异），那么上面所产生的 P 值是可靠的，它能够保证我们犯第一类错误的概率在控制范围（如 5%）以内。但是，如果我们同时考虑很多个假设检验问题（如本案例共四个假设检验），而每一个都有 5% 的概率犯第一类错误，那么我们至少犯一个第一类错误的概率有多大呢？可以肯定地说，不是 5%，而要大于 5%。这说明，如果我们同时面对很多假设检验问题，那么很可能其中相当多的显著性结果是不可靠的。因此，我们需要一种可以控制这种总体错误（experimentwise error）概率的统计方法。

关于如何控制总体错误概率，过去的统计文献中有很多方法，并且已经在很多软件中得到实现。而我们这里将着重介绍一种最简单的，但也是最常用的方法，那就是 Bonferroni 方法。具体地说，假设我们总共有 τ 个感兴趣的假设检验问题，而且我们希望把总体错误概率控制在 α 的水平以下。那么，根据 Bonferroni 方法，我们在对每一个假设做检验的时候，我们不应该采用 α 的显著水平，而应该采用 α/τ 的显著水平。如果我们这样做，那么对任意的一个假设检验问题，我们犯第一类错误的概率为 α/τ。那么，犯总体错误的概率为：

$$P(\text{犯至少一个第一类错误}) \leqslant \tau \times \frac{\alpha}{\tau} = \alpha$$

因此，我们犯总体错误的概率得到了有效的控制。以本案例为例，我们总共有四个假设检验问题。如果我们希望的总体错误概率为 10%，那么我们考察每个假设检验的显著性水平就应该是 0.10/4 = 0.025。由此可以看到，原来我们认为在 10% 的显著性水平可以断定二环以内和三至四环之间的房地产平均价格有明显的差异（因为 P 值 = 0.067）。但是，根据 Bonferroni 方法，我们知道这个证据是不充分的。而另一方面，我们所观测到的二环以内和四至五环以及二环以内和五环以外的房地产价格的显著差异是非常可靠的。

第五节　双因素简单可加模型

在更多的情况下，影响因变量变化的不仅仅是一个自变量，而可能同时有两个或者更多个自变量在共同影响。例如，对于我们的案例，能够影响销售价格的因素除了环线位置以外，还有物业类别（即普通住宅还是公寓）等。这个时候，就涉及多因素的方差分析。我们先来看一个双因素方差分析的模型。假设有因素 A 和因素 B，因素 A 有 g 个水平，因素 B 有 b 个水平，因此共有 gb 个水平组合，对每个水平组合可得到 n 个独立的观测值。y_{lkr} 表示在因素 A 第 l 个水平、因素 B 第 k 个水平下，因变量的第 r 个观测值。此时模型为：

$$y_{lkr} = u + \alpha_l + \beta_k + e_{lkr}$$

其中，μ 代表了某种平均值或者基准平均值，而 $\alpha_l(l=1,2,\cdots,g)$ 代表了因素 A 在水平 l 下的效应（effect）。简单地说，这说明因素 A 在水平 l 下的平均商品房价格为 $\mu+\alpha_l$。类似地，$\beta_k(k=1,2,\cdots,b)$ 代表了因素 B 在水平 k 下的效应。这说明因素 B 在水平 k 下的平均商品房价格为 $\mu+\beta_k$。如果我们同时固定因素 A 的水平为 l，而因素 B 的水平为 k，那么商品房的平均价格应该是多少呢？答案是 $\mu+\alpha_l+\beta_k$。如果我们同时固定因素 A 的水平为 l，而因素 B 的水平从 k_1 变为 k_2，那么商品房的平均价格应该变化多少呢？答案是 $(\mu+\alpha_l+\beta_{k_1})-(\mu+\alpha_l+\beta_{k_2})=\beta_{k_1}-\beta_{k_2}$。

请注意，该差异同因素 A 的水平 l 无关！这说明，在该模型下，因素 B 的水平变化所带来的价格差异同当时因素 A 的水平无关。类似地，我们也可以说，因素 A 的水平变化所带来的价格差异同当时因素 B 的水平无关。这使得我们在预测房屋价格的时候，可以简单地叠加各个因素的效应。因此，我们称此模型为简单可加模型。在 R 中，可以具体分析如下：

```
> lm2.1=lm(log.price ~as.factor(ring)+as.factor(wuye))
> Anova(lm2.1,type="III")
Anova Table (Type III tests)

Response: log.price
                Sum Sq  Df   F value    Pr(>F)
(Intercept)    2263.85   1 34785.143 < 2.2e-16 ***
as.factor(ring)   6.68   4    25.662 < 2.2e-16 ***
as.factor(wuye)   4.32   1    66.454 4.352e-14 ***
Residuals        12.63 194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中可以看到，不管是环线位置还是物业类型，其 P 值都是高度显著的。因此，我们可以得出结论，这两个因素都可以显著地影响商品房销售价格。但是，同单因素方差分析一样，我们还希望知道，到底是哪个因素的哪些水平有差异，并且差异有多大。

```
> summary(lm2.1)

Call:
lm(formula = log.price ~ as.factor(ring) + as.factor(wuye))

Residuals:
     Min       1Q   Median       3Q      Max
-0.73087 -0.18701  0.00767  0.15874  0.72571

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.33977    0.05008 186.506  < 2e-16 ***
as.factor(ring)二至三环 -0.06670    0.05836  -1.143   0.2545
as.factor(ring)三至四环 -0.13710    0.05936  -2.310   0.0220 *
as.factor(ring)四至五环 -0.35868    0.05661  -6.336 1.61e-09 ***
as.factor(ring)五环以外 -0.57473    0.06899  -8.331 1.44e-14 ***
as.factor(wuye)普通住宅 -0.31128    0.03818  -8.152 4.35e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2551 on 194 degrees of freedom
Multiple R-squared: 0.5309,     Adjusted R-squared: 0.5188
F-statistic: 43.91 on 5 and 194 DF,  p-value: < 2.2e-16
```

应该如何解读这张统计报表呢？首先注意到，第一个因素“环线位置”的第一个水平“二环以内”消失了。这说明，截距项 9.34 是在估计“二环以内”。此外，还可以注意到第二个因素的第一个水平“公寓”消失了。这说明，截距项 9.34 是在估计“公寓”。因此，exp(9.34) = 11 384 元/平方米是在估计二环以内公寓的平均售价，这也是整个这张报表的基准值。同单因素方差分析类似，我们没有发现“二环以内”和“二至三环”有显著差异。然后随着环线的增加，房屋价格稳定且显著地下降。此外，我们还注意到“普通住宅”的估计为 −0.31，而且高度显著。这说明，同豪华公寓相比，普通住宅的平均对数价格要低 0.31 个单位。这还说明，给定相同的地理位置，普通住宅的平均销售价格是豪华公寓的 exp(−0.31) = 73.3%。对于一个给定的地区以及房屋类型（如四至五环和普通住宅），其平均销售价格可以估计如下：exp[9.34 − 0.36 − 0.31] = 5 825 元/平方米。

最后，值得注意的是，对于一个实际问题，简单可加模型是不是合理？能不能回答我们关心的问题？那就得视具体问题而论了。例如，某开发商正在考虑将某项目从普通住宅升级为豪华公寓。他有两个不同的项目可以选择，而这两

个项目处在不同的环线位置。简单可加模型会说，选哪一个项目都一样，因为物业类型变化所带来的对数价格变化同环线位置无关，这显然是值得斟酌的。因此，我们下一节将主要介绍带交互作用的双因素模型。

第六节　双因素交互作用模型

和上一节一样，我们假设有因素 A 和因素 B，因素 A 有 g 个水平，因素 B 有 b 个水平，因此共有 gb 个水平组合。对每个水平组合可得到 n 个独立的观测值，y_{lkr} 为因素 A 第 l 个水平、因素 B 第 k 个水平下，因变量的第 r 个观测值。此时，带有交互作用(interaction)的模型为：

$$y_{lkr} = u + \alpha_l + \beta_k + \gamma_{lk} + e_{lkr}$$

其中，μ 仍然代表了某种平均值或者基准平均值，而 $\alpha_l(l=1,2,\cdots,g)$ 代表了因素 A 在水平 l 下的效应，$\beta_k(k=1,2,\cdots,b)$ 代表了因素 B 在水平 k 下的效应。值得注意的是交互作用项 γ_{lk}，它代表了因素 A 和 B 的交互影响作用。

如果我们同时固定因素 A 的水平为 l，而因素 B 的水平为 k，那么商品房的平均价格应该是多少呢？答案是 $\mu+\alpha_l+\beta_k+\gamma_{lk}$。如果我们同时固定因素 A 的水平为 l，而因素 B 的水平从 k_1 变为 k_2，那么商品房的平均价格应该变化多少呢？答案是：

$$(\mu + \alpha_l + \beta_{k_1} + \gamma_{lk_1}) - (\mu + \alpha_l + \beta_{k_2} + \gamma_{lk_2})$$

$$= (\beta_{k_1} - \beta_{k_2}) + (\gamma_{lk_1} - \gamma_{lk_2})$$

请注意，由于交互作用的存在，该差异同因素 A 的水平 l 开始相关了！这说明，在该模型下，因素 B 的水平变化所带来的价格差异同当时因素 A 的水平有关。类似地，我们也可以说，因素 A 的水平变化所带来的价格差异同当时因素 B 的水平有关。这使得我们在预测房屋价格的时候，不可以再简单地叠加各个因素的效应了。因此，我们称此模型为交互作用模型。在 R 中，可以具体分析如下：

```
> lm2.2=lm(log.price~as.factor(ring)*as.factor(wuye))
> Anova(lm2.2,type="III")
Anova Table (Type III tests)

Response: log.price
                                  Sum Sq  Df    F value    Pr(>F)
(Intercept)                      1156.57   1 18645.7008 < 2.2e-16 ***
as.factor(ring)                     4.32   4    17.4312 3.358e-12 ***
as.factor(wuye)                     1.70   1    27.3271 4.503e-07 ***
as.factor(ring):as.factor(wuye)     0.84   4     3.3865   0.01055 *
Residuals                          11.79 190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

和单变量方差分析相似，“Sum Sq”一列分别为所在环线、物业类别、两因素交互作用以及残差的离差平方和。后面的 F 检验显示，这三个因素都是显著的。再输入以下语句便可以获得各个变量的各种水平的具体影响，如下所示：

```
> summary(lm2.2)

Call:
lm(formula = log.price ~ as.factor(ring) * as.factor(wuye))

Residuals:
      Min        1Q    Median        3Q       Max
-0.672785 -0.168384  0.002209  0.152996  0.633248

Coefficients:
                                                    Estimate Std. Error t value
(Intercept)                                          9.43223    0.06908 136.549
as.factor(ring)二至三环                             -0.15679    0.08270  -1.896
as.factor(ring)三至四环                             -0.15211    0.09065  -1.678
as.factor(ring)四至五环                             -0.56648    0.09065  -6.249
as.factor(ring)五环以外                             -0.81495    0.14240  -5.723
as.factor(wuye)普通住宅                             -0.46384    0.08873  -5.228
as.factor(ring)二至三环:as.factor(wuye)普通住宅      0.14622    0.11570   1.264
as.factor(ring)三至四环:as.factor(wuye)普通住宅      0.01702    0.11791   0.144
as.factor(ring)四至五环:as.factor(wuye)普通住宅      0.32723    0.11442   2.860
as.factor(ring)五环以外:as.factor(wuye)普通住宅      0.32987    0.16273   2.027
                                                    Pr(>|t|)
(Intercept)                                          < 2e-16 ***
as.factor(ring)二至三环                              0.05950 .
as.factor(ring)三至四环                              0.09499 .
as.factor(ring)四至五环                             2.65e-09 ***
as.factor(ring)五环以外                             4.03e-08 ***
as.factor(wuye)普通住宅                             4.50e-07 ***
as.factor(ring)二至三环:as.factor(wuye)普通住宅      0.20786
as.factor(ring)三至四环:as.factor(wuye)普通住宅      0.88539
as.factor(ring)四至五环:as.factor(wuye)普通住宅      0.00471 **
as.factor(ring)五环以外:as.factor(wuye)普通住宅      0.04405 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2491 on 190 degrees of freedom
```

与简单可加模型类似，在结果中截距项的估计值对应的是二环以内的公寓的平均售价。那么，二至三环的普通住宅的平均售价应该是多少呢？答案是 exp {9.43223 - 0.15679 - 0.46384 + 0.14622} = 7 768.404 元/平方米。

从上面的简单计算可以看到，交互作用的实质就是表明物业类型对平均售价的影响结果还会受到另外一个因素环线位置的影响，反之亦然。换句话说，也就是公寓同普通住宅的价格差异，会因为环线位置的不同而不同。例如，如果我们将一处房产从普通住宅变为豪华公寓，那么对数价格会改变多少呢？答案是不确定，这取决于该房产处于什么地理位置。如果是二环以内，那么变化为 -0.46；如果是三至四环，那么变化为 -0.46 + 0.02 = -0.44；而如果是五环以外，该变化为 -0.46 + 0.33 = -0.13。这说明什么呢？这说明在城市的中心区域修建豪华公寓所带来的价格上升要远远高于在城区边缘地带修建所带来的。

第七节　多因素方差分析

在实际应用中，我们碰到的情况可能会更加复杂。我们所涉及的因素个数可能远远大于两个。此时，我们就应该采用多因素方差分析的方法。简单地说，多因素方差分析是双因素方差分析的一个简单延伸。在多因素方差分析中，我们同样可以考虑简单可加模型，也可以考虑各种交互作用模型。但是，值得注意的是，交互作用的准确估计需要较大的样本（同简单可加模型相比较），还有，并不是每一种交互作用都有很好的实际意义，所以一般来说，我们不建议在多因素方差分析中盲目地添加交互作用项。

接下来，我们再用北京市房地产数据详细演示多因素方差分析方法。在上一节，我们仅仅考虑了城区以及环线位置两个因素。下面，我们将综合考虑数据所提供给我们的所有因素。这些因素包括：城区、环线位置、物业类型、装修状况以及建筑类型，共五个因素。如果我们考虑所有的交互作用，那么我们就会有十个交互作用项。我们假设研究者就对两个地理位置因素（即城区以及环线位置）特别感兴趣。因此，我们在模型分析中也考虑了它们的交互作用。但是，所有其他因素的效应都以可加的形式放入了模型中。具体如下：

```
> lm4=lm(log.price~as.factor(die)*as.factor(ring)
+ +as.factor(huye)+as.factor fitment)+as.factor(contype))
> summary(lm4)
```

其结果如下所示：

(Intercept)	9.83289	0.21897	44.905	< 2e-16	***
as.factor(dis)崇文	-0.82384	0.22186	-3.713	0.000275	***
as.factor(dis)东城	-0.51725	0.22569	-2.292	0.023103	*
as.factor(dis)海淀	0.30462	0.11131	2.737	0.006848	**
as.factor(dis)石景山	-0.04942	0.10370	-0.477	0.634271	
as.factor(dis)西城	-0.56187	0.23420	-2.399	0.017488	*
as.factor(dis)宣武	-0.85992	0.22207	-3.872	0.000152	***
as.factor(ring)二至三环	-0.62293	0.21626	-2.880	0.004467	**
as.factor(ring)三至四环	-0.80277	0.21586	-3.719	0.000269	***
as.factor(ring)四至五环	-1.01860	0.21482	-4.742	4.38e-06	***
as.factor(ring)五环以外	-1.35810	0.22660	-5.993	1.14e-08	***
as.factor(wuye)普通住宅	-0.16354	0.03586	-4.560	9.57e-06	***
as.factor(fitment)精装修	0.26085	0.06171	4.227	3.81e-05	***
as.factor(fitment)毛坯	0.06802	0.05479	1.241	0.216086	
as.factor(contype)板塔结合	-0.02826	0.05418	-0.522	0.602651	
as.factor(contype)高层	-0.02597	0.04383	-0.593	0.554252	
as.factor(contype)塔楼	-0.03914	0.03841	-1.019	0.309594	
as.factor(dis)崇文:as.factor(ring)二至三环	0.47937	0.24829	1.931	0.055140	.
as.factor(dis)东城:as.factor(ring)二至三环	0.55569	0.24595	2.259	0.025096	*
as.factor(dis)海淀:as.factor(ring)二至三环	-0.47849	0.13791	-3.470	0.000656	***
as.factor(dis)石景山:as.factor(ring)二至三环	NA	NA	NA	NA	
as.factor(dis)西城:as.factor(ring)二至三环	0.50264	0.32048	1.568	0.118594	
as.factor(dis)宣武:as.factor(ring)二至三环	0.35862	0.24109	1.487	0.138686	
as.factor(dis)崇文:as.factor(ring)三至四环	0.50383	0.30571	1.648	0.101127	
as.factor(dis)东城:as.factor(ring)三至四环	NA	NA	NA	NA	
as.factor(dis)海淀:as.factor(ring)三至四环	-0.30933	0.12892	-2.399	0.017474	*
as.factor(dis)石景山:as.factor(ring)三至四环	NA	NA	NA	NA	
as.factor(dis)西城:as.factor(ring)三至四环	NA	NA	NA	NA	
as.factor(dis)宣武:as.factor(ring)三至四环	NA	NA	NA	NA	
as.factor(dis)崇文:as.factor(ring)四至五环	NA	NA	NA	NA	
as.factor(dis)东城:as.factor(ring)四至五环	NA	NA	NA	NA	
as.factor(dis)海淀:as.factor(ring)四至五环	-0.32457	0.12954	-2.506	0.013137	*
as.factor(dis)石景山:as.factor(ring)四至五环	NA	NA	NA	NA	
as.factor(dis)西城:as.factor(ring)四至五环	NA	NA	NA	NA	
as.factor(dis)宣武:as.factor(ring)四至五环	NA	NA	NA	NA	
as.factor(dis)崇文:as.factor(ring)五环以外	NA	NA	NA	NA	
as.factor(dis)东城:as.factor(ring)五环以外	NA	NA	NA	NA	
as.factor(dis)海淀:as.factor(ring)五环以外	NA	NA	NA	NA	
as.factor(dis)石景山:as.factor(ring)五环以外	NA	NA	NA	NA	
as.factor(dis)西城:as.factor(ring)五环以外	NA	NA	NA	NA	
as.factor(dis)宣武:as.factor(ring)五环以外	NA	NA	NA	NA	

从中可见，除了石景山区外，城区的大部分水平都是显著的，表示大部分地区和朝阳区之间存在显著差异；环线位置的所有水平都是显著的，表示其他环线位置和二环以内都存在显著差异，而且随着地理位置从内环向外环延伸，商品房均价的平均水平呈现出显著的下降趋势；物业类型的两个水平之间差异显著，普通住宅的价格要显著低于豪华公寓；在装修类型的不同水平之间，精装修的房屋价格要显著高于粗装修的房屋，但毛坯房与粗装修的房屋之间的价格差异并不显著；相对于板楼，不同建筑类型的房屋，其价格并无显著差异。

如何去解读这样一张复杂的统计报表呢？同前面的单因素方差分析以及双因素方差分析一样，我们首先要搞清楚截距项代表什么。这里我们共涉及四个因素：城区、环线位置、物业类型以及装修状况。那么，我们可以在上面的报表中依次寻找到底各个因素的哪个水平消失了。答案如下：

- 城区：朝阳区

- 环线：二环以内
- 住宅类型：豪华公寓
- 装修类型：粗装修
- 建筑类型：板楼

由此可见，截距项表示的是朝阳区、二环以内、豪华公寓、粗装修、板楼的对数价格。以普通货币作为单位，这样的商品房的平均销售价格为 exp(9.83) = 18 583 元/平方米。

另外，值得注意的是，报表中有大量的缺失值（即 NA）存在，如石景山与二至三环的交互作用。这是为什么呢？在回答这个问题之前，我们首先要明白如果这个参数估计不是缺失值，那么它在估计什么？事实上，它在估计两种房屋的理论平均价格差异。哪两种房屋？一种是处在二环以内的石景山房屋（这当然是不可能的），而另外一种是处在二至三环间的石景山房屋（这也是不可能的）。我们假设这两种房屋的其他因素完全一致（如都是精装修豪华公寓），那么它们的平均价格差异就是该交互作用试图估计的对象。大家可以看到，要做这样一种比较我们的样本必须满足一些条件。具体地说，必须有一部分样本来自于石景山而且地处二环以内，还必须有一部分样本来自于石景山而且地处二到三环之间。缺一不可！而在我们的数据中显然不存在这样的样本，因此该参数无法估计。但是，值得注意的是，我们的样本中确实有的房屋来自于石景山而且地处五环以外，那么为什么它的交互作用的估计也是缺失值呢？这是因为它没有比较的对象，即来自于石景山而且地处二环以内的样本。

这些缺失的估计量会对我们的分析以及实践产生什么影响呢？第一，它不影响我们关于城区和环线位置这两个因素的基本结论，即这两个因素以及它们的交互作用都很重要。这是因为，但凡我们看到一个交互作用重要，那么这两个因素及其交互作用就自然而然地变得重要了。第二，假如真的出现了一套来自于石景山而且地处二环以内的商品房，那么本数据所产生的模型无法对其价格予以预测，这是我们的数据带来的局限。

以上的讨论还产生了一个问题，那就是：为什么在简单可加模型中（如第五节所演示的），所有的参数都是可以估计的呢？请注意，简单可加模型的基本假设就是各个因素对房屋对数价格所产生的影响不依赖于其他因素的水平。因此，虽然我们的样本中根本就没有来自于石景山而地处三环以内的房屋，但是其他区（如崇文区）的类似房屋所表现出来的差异（如崇文区二环以内的房屋对比崇文区二至三环的房屋）就可以为石景山所借鉴。

最后，我们演示一下一个具体的预测问题。假设研究者对一套海淀区地处三至四环之间的精装修、建筑类型为板楼的普通住宅感兴趣（请注意：这是一个

可以估计的房产），那么它的平均销售价格应该是：

$$\exp\{9.83 + 0.30 - 0.80 - 0.16 + 0.26 - 0.30\} = 9\,228 \text{ 元 / 平方米}$$

同第一章类似，我们对模型所分离出来的残差诊断如下（参见图 2-5）：

```
> par(mfrow=c(2,2))
> plot(lm4,which=c(1:4))
> par(mfrow=c(1,1))
```

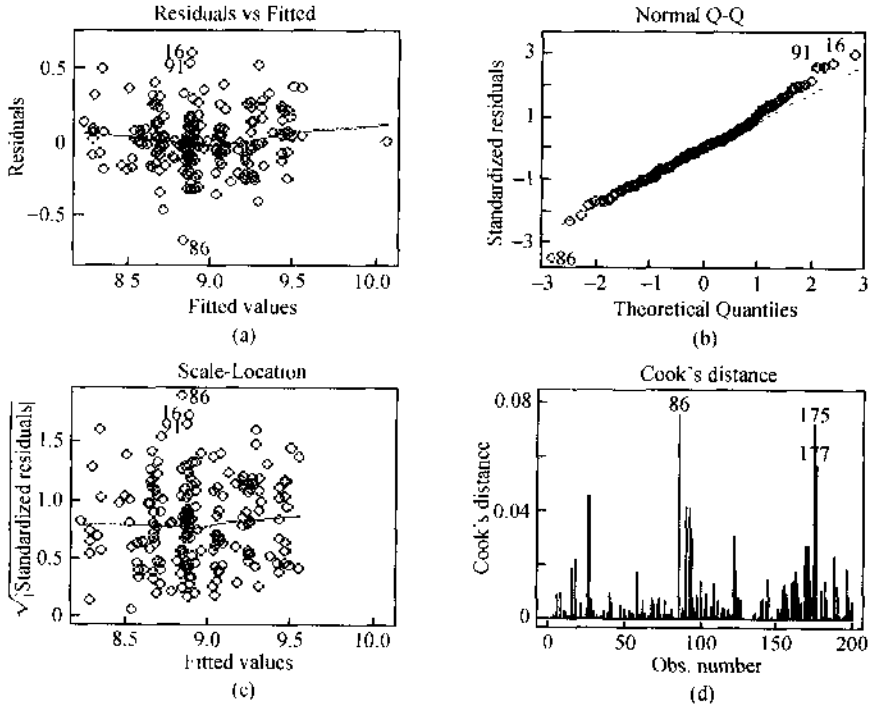


图 2-5　模型诊断图

其结果包含了两张残差图、一张 QQ 图和一张 Cook 距离图，如图 2-5 所示。从中没有发现任何严重背离模型假设的证据。因此，该模型可以接受。假设某开发商正在考虑投资以下项目：

```
> a0=read.csv("D:/Practical Business Data Analysis/case/CH2/new.csv")
> a0=a0[,c(1:5)]
> a0
   dis     ring     wuye fitment contype
1 朝阳 三至四环 普通住宅     毛坯     板楼
2 海淀 三至四环 普通住宅   精装修 板塔结合
3 海淀 四至五环 普通住宅   粗装修     高层
4 崇文 二环以内     公寓     毛坯     塔楼
```

我们可以利用多因素方差分析模型，对其预期的销售价格预测如下：

```
> y.pred=exp(predict.lm(lm4,a0))
Warning message:
用秩缺乏拟合来进行预测的结果很可能不可靠 in: predict.lm(lm4, a0)
> a0$y.pred=y.pred
> a0
   dis      ring      wuye fitment  contype   y.pred
1 朝阳 三至四环 普通住宅     毛坯      板楼 7590.113
2 海淀 三至四环 普通住宅 精装修 板塔结合 8905.805
3 海淀 四至五环 普通住宅 粗装修      高层 5457.969
4 崇文 二环以内      公寓     毛坯      塔楼 8416.420
```

基于此预期的销售价格，再结合各个项目的成本，我们就可以科学地选择项目，并予以开发。值得注意的是，由于我们的模型中有很多交互作用无法估计，这使得模型的设计具有一定的缺陷（即秩缺乏）。因此，*R* 会警告用户该模型的预测结果可能不准确。但是，只要我们确信该房产所涉及的所有参数都是可以估计的，那么该预测结果就是准确的。

第八节　简单分析报告

北京市商品房价格影响因素分析报告

内容提要　本报告利用北京市商品房的销售价格数据，确定影响商品房销售价格的重要因素，并量化这些因素对销售价格的影响。从我们的分析结果发现，影响北京市商品房平均销售价格的主要因素有所在区县、所在环线、物业类别以及装修状况。本报告的分析结果可以为商品房购房者提供科学、可靠的价格参考依据，也可以为相关机构的价值评估提供理论依据，还可以为房地产开发商选择项目以及制定开发策略提供有价值的参考。

一、研究目的

北京市房地产市场是我国房地产市场中最为发达，也是最具有代表性的几个房地产市场之一，而且其商品房的销售价格差异巨大，从最低的 4 000 元/平方米一直到最高的 20 000 元/平方米。因此，找出是什么样的因素在影响着北京市商品房的销售价格，以及为何会产生价格上的巨大差异，是一件非常有意义的事情。本报告利用北京市商品房的销售价格数据，确定影响商品房销售价格的重要因素，并量化了这些因素的影响。通过了解影响商品房销售价格的因素，购房者可以更加理性地选择房屋，价值评估机构可以通过理论模型来评估房屋价值，房地产项目开发商可以预测项目的销售价格，再结合各个项目的成本，从而科学地选择项目进行开发并制定相关的开发策略。

二、数据来源和相关说明

我们从“搜房网”(www.soufun.com)上随机选取了北京地区2003—2004年度开盘的新楼盘共506个。在进行数据清理后,我们最终得到200个合格的楼盘样本。我们希望能够基于这样一个公开的实际数据,建立恰当的经济计量模型,从数量上刻画房地产销售价格同各个影响因素之间的关系。具体地说,我们的数据中包含了表2-2中的信息。

表2-2 变量说明

变量类型	变量	水平数
连续型	销售均价(元/平方米)	无
	容积率(%)	无
	绿化率(%)	无
	小区总建筑面积(平方米)	无
	小区停车位住户比例(位/户)	无
离散型	所在区县	共七个区县(七个主城区)
	所在环线	共五环(<2、2-3、3-4、4-5、>5)
	物业类别	共两种(普通、公寓)
	装修状况	共三种(毛坯、粗装修、精装修)
	建筑类别	共四种(板楼、塔楼、板塔结合、高层)

从表2-2可以看到,在数据所涵盖的所有变量中,有一个是我们特别感兴趣的,试图通过其他变量来解释的,那就是销售均价。这也就是我们的因变量,而其他的所有变量都是自变量。

通过对数据的简单分析,我们发现对销售均价进行对数变换对本报告的数据分析是有帮助的,具有稳定方差的作用。在以后的分析中,我们都采用对数变换后的销售均价作为因变量。

三、描述性分析

为了获得对数据的整体概念,我们利用盒状图对数据进行简单的描述性分析,如图2-6所示。

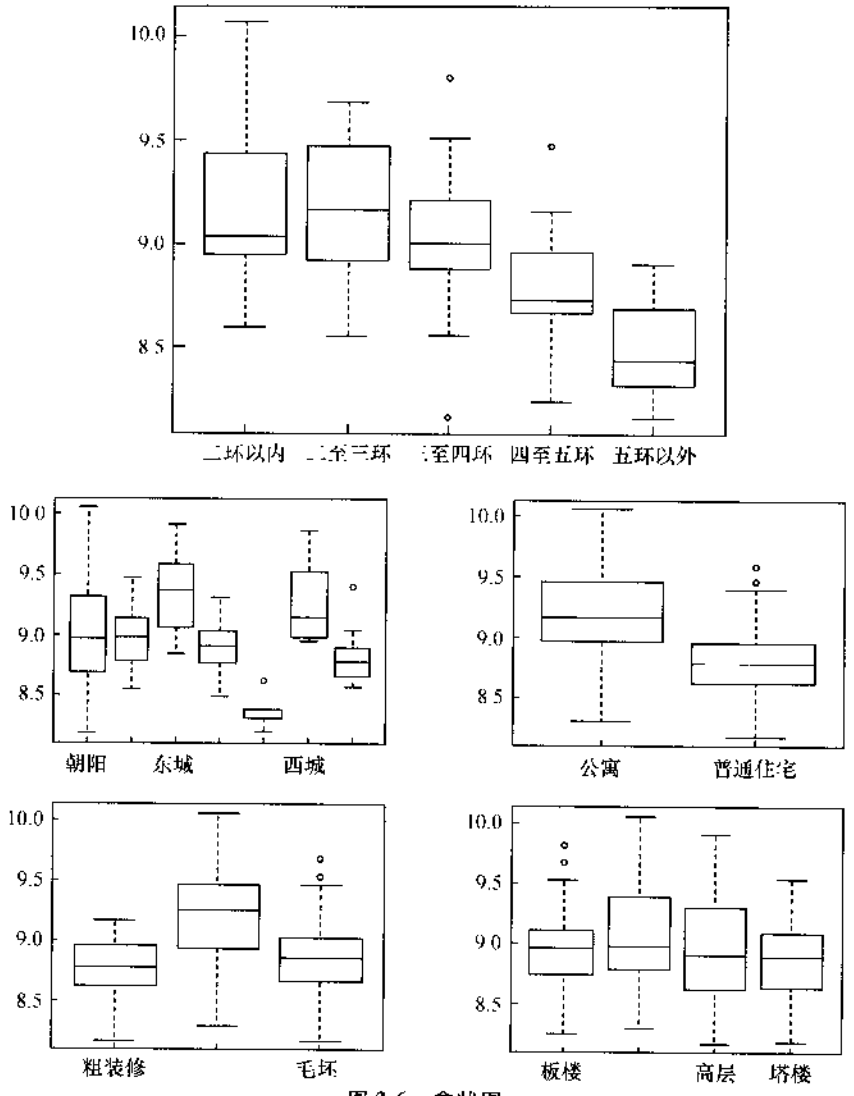


图 2-6 盒状图

从图 2-6 可以看出：

• 地理位置对房屋的均价有显著的影响。随着地理位置从内环向外环延伸，商品房均价的中位数呈现出明显的下降趋势。这说明，距离城市中心越远，

商品房平均价格越低。

• 城区的不同不仅极大地影响着房屋的均价，而且还影响着房屋价格的波动程度。

• 公寓的平均价格明显高于普通住宅，但波动程度相当。

• 精装修房屋的平均价格明显高于粗装修以及毛坯房，而且价格波动较大。此外，粗装修房屋的均价并没有明显高于毛坯房。

• 不同的房屋类型（如板楼、塔楼、板塔结合）对房屋的平均价格及其波动程度影响都很小。

四、数据建模

在本节中，我们考虑的是离散型的自变量对销售价格的影响，因此采用方差分析的方法建立模型，来寻找因变量和自变量之间的关系。具体地说，就是比较不同城区、不同地理位置、不同建筑类别等的商品房价格从平均水平上来说有没有显著差异，并量化这种差异。需要特别指出的是，一般来说，地理位置对房地产价格的影响非常重要，因此我们在模型分析中考虑两个地理位置因素（即所在区县以及所在环线）的交互作用，而对其他的自变量我们采用可加模型的形式。

利用全部数据估计各个因素的不同水平对销售价格的具体影响，得到估计结果如表 2-3 所示。从表 2-3 中可以发现，“所在区县”的不同水平的价格之间近似有如下从大到小的关系：海淀、朝阳、石景山、东城、西城、崇文、宣武。需要注意的是，这种近似关系是通过和朝阳区的比较得出的，而且石景山区和朝阳区之间的差异并不显著。根据“所在环线”的估计结果可以看出，地理位置对房屋的均价有显著的影响，随着地理位置从内环向外环延伸，商品房均价的平均水平呈现出显著的下降趋势。还可以发现，对于不同“物业类别”的房屋，公寓的平均价格显著高于普通住宅。“装修状况”的不同水平之间也有区别，精装修房屋的价格显著高于粗装修的房屋，而毛坯房与粗装修房屋之间的区别并不显著。另外，所在城区和所在环线的交互作用也是非常显著的。因此，通过上述分析，我们可以判断出显著影响北京市商品房平均销售价格的因素有所在区县、所在环线、物业类别以及装修状况。

表 2-3 模型估计结果

变　量	系数估计值	标准差	P 值
截距	9.8329	0.2190	<2e-16
所在区县:崇文	-0.8238	0.2219	0.0003
所在区县:东城	-0.5173	0.2257	0.0231
所在区县:海淀	0.3046	0.1113	0.0068
所在区县:石景山	-0.0494	0.1037	0.6343
所在区县:西城	-0.5619	0.2342	0.0175
所在区县:宣武	-0.8599	0.2221	0.0002
所在环线:二至三环	-0.6229	0.2163	0.0045
所在环线:三至四环	-0.8028	0.2159	0.0003
所在环线:四至五环	-1.0186	0.2148	0.0000
所在环线:五环以外	-1.3581	0.2266	0.0000
物业类别:普通住宅	-0.1635	0.0359	0.0000
装修状况:精装修	0.2609	0.0617	0.0000
装修状况:毛坯	0.0680	0.0548	0.2161
建筑类别:板塔结合	-0.0283	0.0542	0.6027
建筑类别:高层	-0.0260	0.0438	0.5543
建筑类别:塔楼	-0.0391	0.0384	0.3096
所在区县:崇文 * 所在环线:二至三环	0.4794	0.2483	0.0551
所在区县:东城 * 所在环线:二至三环	0.5557	0.2460	0.0251
所在区县:海淀 * 所在环线:二至三环	-0.4785	0.1379	0.0007
所在区县:石景山 * 所在环线:二至三环	NA	NA	NA
所在区县:西城 * 所在环线:二至三环	0.5026	0.3205	0.1186
所在区县:宣武 * 所在环线:二至三环	0.3586	0.2411	0.1387
所在区县:崇文 * 所在环线:三至四环	0.5038	0.3057	0.1011
所在区县:东城 * 所在环线:三至四环	NA	NA	NA
所在区县:海淀 * 所在环线:三至四环	-0.3093	0.1289	0.0175
所在区县:石景山 * 所在环线:三至四环	NA	NA	NA
所在区县:西城 * 所在环线:三至四环	NA	NA	NA
所在区县:宣武 * 所在环线:三至四环	NA	NA	NA
所在区县:崇文 * 所在环线:四至五环	NA	NA	NA
所在区县:东城 * 所在环线:四至五环	NA	NA	NA
所在区县:海淀 * 所在环线:四至五环	-0.3246	0.1295	0.0131
……	…	…	…

五、结论及建议

从上述分析结果可知，影响北京市商品房平均销售价格的主要因素有所在区县、所在环线、物业类别以及装修状况。而且，所在区县与所在环线之间存在着相互影响。商品房购房者可以利用本报告得到的模型，对价格进行合理的预测，比较不同的楼盘，从而为购房决策提供帮助。本报告的分析结果还可以为相关机构的价值评估提供理论依据。对于房地产开发商，本报告提供了更多有用的信息。开发商可以利用该模型来合理地选择项目，并结合成本核算的结果来对项目的开发制定最优的策略。例如，由装修类型不同水平的差别可以发现，精装修房屋的销售价格显著高于毛坯房，但粗装修房屋则与毛坯房之间无显著差异，因此开发商就可以考虑不对房屋进行粗装修，而在成本核算之后根据利润最大化的原则选择精装修或者毛坯房。类似地，对于其他因素的分析也可以为开发商提供有价值的参考。

[讨论总结]

本章以商品房价格为例，系统演示并讲解了方差分析。通过对本章的学习，读者应该能够了解：什么时候可以使用方差分析，以及如何使用。在 R 语言学习方面，读者主要应该了解对离散型解释性变量的处理，即哑变量的生成。在统计理论方面，读者应该理解并掌握以下重要概念：离散型解释性变量、方差分解、多重比较、单因素以及多因素方差分析、交互作用。由于篇幅限制，我们不可能对 R 语言作系统详细的介绍。由于 R 语言的语法同商业软件 $S+$ 极其相似，因此，希望对 R 作深入了解的读者可以查阅 Venables and Ripley(1994)。对相关统计学理论渴望深入了解的读者可以参阅 Rao(1973)、Draper and Smith(1981)，还有 Milliken and Johnson(2002a)。

附录　程序及注释

```
rm(list=ls())                                                    # 清空当前工作空间
a=read.csv("D:/Practical Business Data Analysis/case/CH2/real.csv",header=T)
                                                                 # 读入 csv 格式的数据,并赋值给 a
attach(a)                                                        # 将 a 中各变量加入工作空间,便于直接调用
pairs(a[,c(1:6)])                                                # 对 a 的前 6 列作散点图
boxplot(price ~ ring)                                            # 画出 price 与 ring 之间的盒状图
log.price=log(price)                                             # 对 price 作对数变化,并赋值给 log.price
boxplot(log.price ~ ring)                                        # 画出 log.price 与 ring 之间的盒状图
par(mfrow=c(2,2))                                                # 设置画图模式 2x2 的格式
boxplot(log.price ~ dis)                                         # 画出 log.price 与 dis 之间的盒状图
boxplot(log.price ~ wuye)                                        # 画出 log.price 与 wuye 之间的盒状图
boxplot(log.price ~ fitment)                                     # 画出 log.price 与 fitment 之间的盒状图
boxplot(log.price ~ contype)                                     # 画出 log.price 与 contype 之间的盒状图
summary(a[,c(1:5)])                                              # 给出 a 中前 5 列的描述性分类统计
lm1=lm(log.price ~ as.factor(ring))                              # 以离散型变量 ring 为解释性变量作单因子方差分析
library(car)                                                     # 载入程序包 car
Anova(lm1,type="III")                                            # 对模型 lm1 作三型方差分析
summary(lm1)                                                     # 显示模型 lm1 的各方面细节,包括参数估计值、P 值等
lm2.1=lm(log.price ~ as.factor(ring) + as.factor(wuye))          # 不带交互作用的双因子方差分析
Anova(lm2.1,type="III")                                          # 对模型 lm2.1 作三型方差分析
```

```
summary(lm2.1)                                    # 显示模型 lm2.1 的各方面细节,包括参数估计值、P 值等
lm2.2=lm(log.price ~ as.factor(ring)* as.factor(wuye))   # 带交互作用的双因子方差分析
Anova(lm2.2,type="III")                           # 对模型 lm2.2 作三型方差分析
summary(lm2.2)                                    # 显示模型 lm2.2 的各方面细节,包括参数估计值、P 值等
lm4=lm(log.price ~ as.factor(dis)* as.factor(ring) + as.factor(wuye) + as.factor(fitment) + as.factor(contype))
                                                  # 全模型方差分析
summary(lm4)                                      # 显示模型 lm4 的各方面细节,包括参数估计值、P 值等
par(mfrow=c(2,2))                                 # 设置画图模式为 2x2
plot(lm4,which=c(1:4))                            # 画出 lm4 中模型检验的 4 张图,包括残差图、QQ 图和 Cook 距离图
par(mfrow=c(1,1))                                 # 设置画图模式为 1x1
a0=read.csv("D:/Practical Business Data Analysis/case/CH2/new.csv")
                                                  # 读入新数据,赋值给 a0
a0=a0[,c(1:5)]                                    # 取 a0 的前 5 列
a0                                                # 展示 a0 的数据
y.pred=exp(predict(lm4,a0))                       # 用模型 lm4 对 a0 作预测
a0$y.pred=y.pred                                  # 将预测结果赋值给 a0 中的变量 y.pred
a0                                                # 展示 a0 的数据,包括预测值
```

第三章
协方差分析

- 案例介绍
- 描述性分析
- 单因素可加模型
- 单因素交互作用模型
- 多因素协方差分析
- 模型选择与预测
- 更科学的绩效评估
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过光华教学评估数据分析的实际案例，详细介绍协方差分析这种重要的统计回归模型。它主要处理的是因变量为连续型数据而解释性变量为连续型与离散型混合数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用协方差分析；(2) 协方差分析的基本统计学理论；(3) 相关理论在统计学软件 *R* 中的应用；(4) 相应的统计分析报告的撰写。本章可以看做是普通线性回归和方差分析的结合，因此没有涉及太新的统计学概念。

第一节　案例介绍

在第一章中，我们学习了如何通过回归模型处理连续型协变量。紧接着，我们在第二章中学习了如何通过方差分析处理离散型协变量。那么，如果连续型和离散型协变量同时出现，又应该怎么处理呢？本章将通过一个教学绩效评估的案例，具体演示一种同时处理各种类型协变量的统计方法，即协方差分析(covariance analysis)。从理论上说，第一章的回归模型、第二章的方差分析模型以及本章的协方差分析模型都是普通线性模型。因此，本章在理论上没有任何新颖之处。所以，我们将着重演示整个分析推理的过程。

按照惯例，我们首先对本章所使用的案例简要介绍如下：本章所使用的数据来自于北京大学光华管理学院的教学评估记录。本数据共有 340 条有效记录，其中每一条记录都对应于 2002 年至 2004 年这三年间，在北京大学光华管理学院所教授的某一门课程。因变量是该课程的最终评估得分，其分数是将参与教学评估的全部学员的评估数据汇总后按照一定方法计算而得。同任何一所大专院校一样，该教学评估数据虽然不是唯一的，但却是一种重要的衡量教员教学成果的手段。如果该评估是非常完美的，那么我们就可以通过简单地比较两门课的教学评估成绩来比较两个教员的教学绩效。

但是，在实践中人们很容易发现事情远远没有想象中那么简单。那问题是什么呢？举一个简单的例子：某教员同时给两个不同的班级上课，假设该教员

对每一门课的付出都是基本相同的，但是最终两门课的教学评估成绩却可能差别极大（例如，一门优秀，而另一门低于平均水平）。这是为什么呢？有过教学经历的研究者都会知道，有太多的非教员因素（如班级的大小）会影响教学评估成绩。一般来说，同一个教员在同样的努力程度下，如果班级规模较小，那么学生感受到的教学效果就会较好，从而该教员的教学评估成绩就会较好。此外，同一门课程，同样的内容，如果授课对象不同，评估成绩也会不同。例如，MBA学员和本科生给出的教学评估成绩较低，而普通研究生给出的教学评估成绩较高。因此，一个合理的且更加公平的评估体系应该将各种影响因素尽量考虑在内。而要达到这一步的前提是我们必须清楚：① 哪些非教员因素在影响教学评估成绩？② 这些影响因素对教学评估成绩的影响有多大？为了回答这些问题，在我们的数据中详细收集了以下信息：教员职称（助理教授、副教授、正教授）、教员性别（男、女）、学生类别（MBA、本科生、研究生）、年份（2002、2003、2004）、学期（秋季、春季）以及班级规模。值得注意的是，在我们所考虑的解释性变量中，班级规模（即班级中学生的数目）是一个具有数值意义的变量，可以简单地看做连续型变量，而其他的所有解释性变量都是离散型变量。

第二节　描述性分析

按照惯例，我们先进行描述性分析。因为本案例既涉及连续型协变量，又涉及离散型协变量，因此我们需要对它们分别予以描述。首先，将数据读入如下：

```
> rm(list=ls())
> a=read.csv("D:/Practical Business Data Analysis/case/CH3/teaching.csv",header=T)
> attach(a)
> a[c(1:5),]
  title gender student year semester size score
1 副教授      女      MBA 2002      秋季  114 3.175
2 副教授      女      MBA 2002      秋季   88 3.523
3 副教授      女      MBA 2003      秋季   83 4.458
4 副教授      女      MBA 2002      秋季   66 3.470
5 副教授      女      MBA 2003      秋季   46 4.630
```

从输出结果的第一行可以看到，2002年秋季某位女性副教授讲授了一门114人的MBA课程，其最终的教学评估成绩为3.175分。从第二行可知，另外

一位女性副教授讲授了一门 88 人的 MBA 课程，其最终的教学评估成绩为 3.523 分。现在的问题是，产生这 0.348 分差异的原因是什么？是因为第二位教员教得好，还是因为第一位教员的班级规模过大？类似这样的问题就是本章所研究的重点。我们首先尝试探索连续型协变量班级规模（size）和因变量评估成绩（score）之间的数量关系。

大家可以看到，在图 3-1 这样一张杂乱无章的散点图中，我们很难发现非常有意义的统计规律。其主要原因是噪音太大，我们的肉眼很难对其总体趋势予以准确判断。因此，我们尝试将班级规模从 0 到 140 人等分成七组（例如，第一组：0—20 人的课程；第二组：21—40 人的课程……第七组：121—140 人的课程）。然后，再按照不同的组作盒状图，如图 3-2 所示。

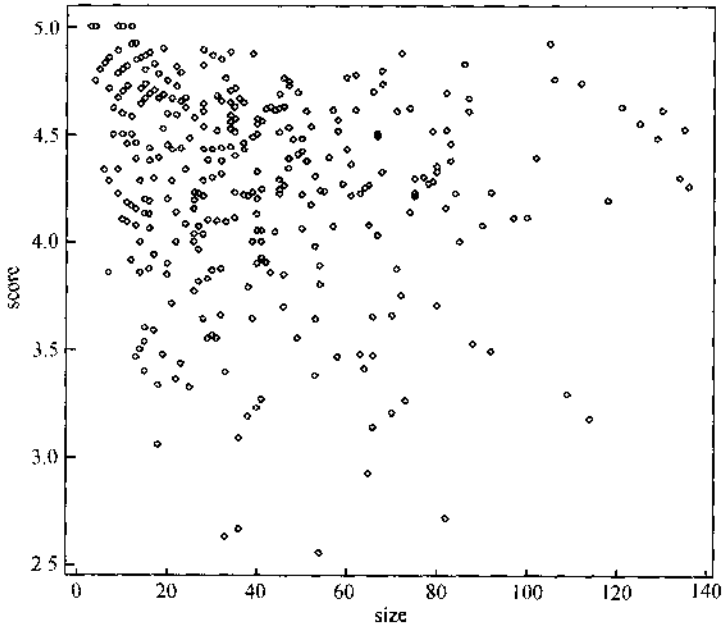


图 3-1　最终评估得分与学生人数的散点图

```
> boxplot(score~ceiling(size/20))
```

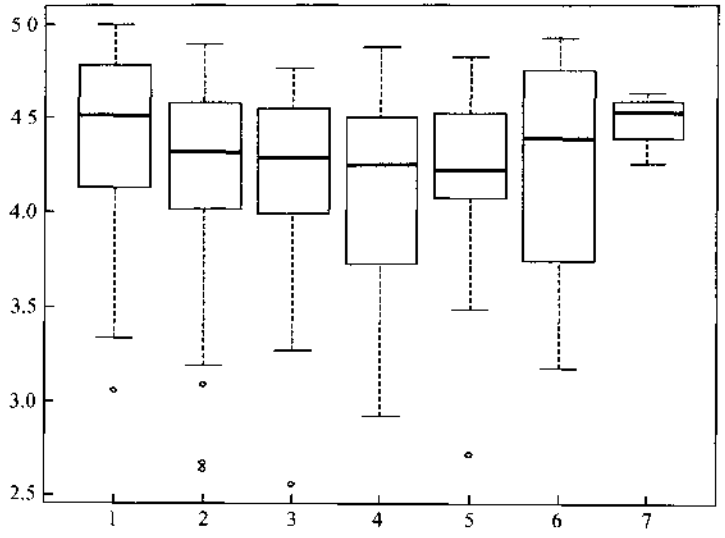


图 3-2　不同分组的盒状图

从图 3-2 我们可以发现以下重要的规律：① 第一组（0—20 人）和第二组（21—40 人）之间的平均教学评估成绩（以中位数计）差异明显；② 第二至五组（41—100 人）之间的平均教学评估成绩没有特别明显的差异；③ 第六、七组（101—140 人）的平均教学评估成绩表现异常，怀疑是由于样本量过小而造成的。下面我们再考察样本在各组之间的具体分布：

```
> table(ceiling(size/20))

  1   2   3   4   5   6   7
 98 112  59  40  17   7   7
```

从中可以看到，第六、七组的样本量太小，因此盒状图中所表现出来的较高的平均教学评估成绩缺乏可信度。但是，第一、二组的样本量非常大，因此可以比较确信第一组和第二组之间确实存在着明显的差异。当然，该差异在统计上是否显著还需要进行后面的严格分析，但是这提示我们，班级规模大约以 20 人为界，前后规律迥异。因此，我们定义哑变量（group）如下：如果班级规模小于等于 20，那么 group = 1，否则 group = 0。在 *R* 中具体实现如下：

```
> group=1*(size<=20)
```

这样，我们就得到总共六个离散型变量：班级规模哑变量、教员职称、教员性别、学生类别、年份以及学期。通过盒状图（如图 3-3 所示）对它们描述如下：

```
> par(mfrow=c(3,2))
> boxplot(score~title,main="职称")
> boxplot(score~gender,main="性别")
> boxplot(score~student,main="学生类别")
> boxplot(score~year,main="年份")
> boxplot(score~semester,main="学期")
> boxplot(score~group,main="班级规模")
> par(mfrow=c(1,1))
```

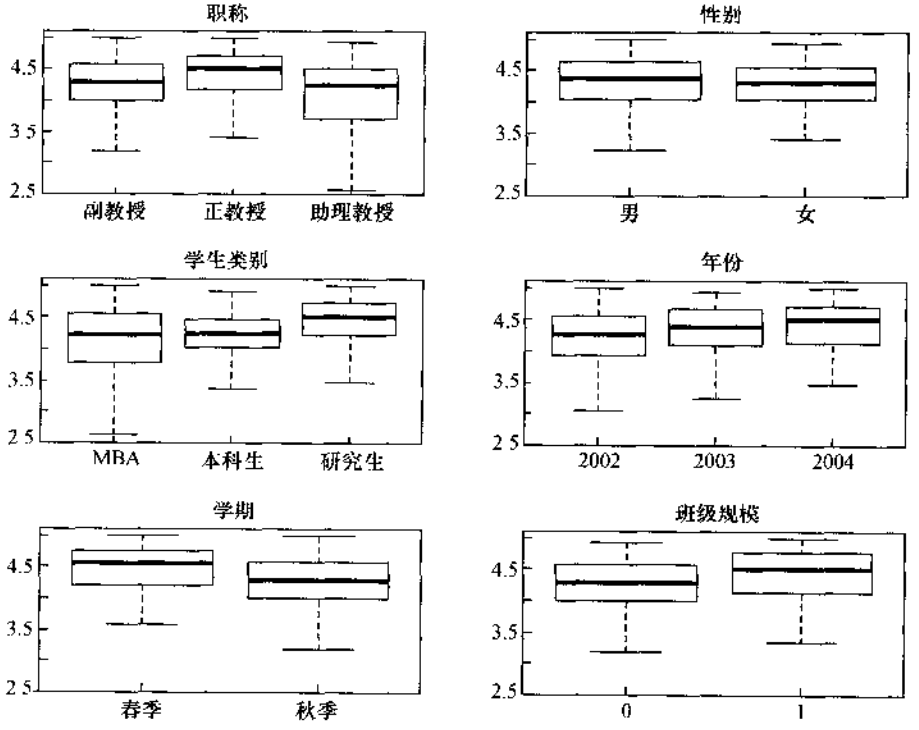


图 3-3　基于盒状图的描述性分析

从图 3-3 我们可以发现以下重要的规律：

- 教员职称确实能够影响教学评估成绩。随着教员职称的提高(从助理教授到副教授再到正教授)，平均教学评估成绩(以中位数计)依次增高。这在一定程度上反映出积累的教学科研经验对教学评估成绩的影响。
- 教员性别对教学评估成绩影响甚微。
- 不同的学生类别对教学评估成绩影响很大。普通研究生给出的平均教学评估成绩明显高于本科生和 MBA，而本科生和 MBA 之间差异不大。
- 随着时间的推移，北京大学光华管理学院的教学评估成绩稳步提高。
- 秋季学期(即每学年的第一个学期)的教学评估成绩低于春季学期。

• 小于 20 人的班级的教学评估成绩明显高于大于 20 人的班级。

以上的描述性分析使我们对数据有了初步的认识，并形成了一些初步的观点。在以下的几节中，我们将通过协方差模型对它们予以严格的分析和检验。

第三节　单因素可加模型

为了简单起见，我们首先考虑单因素协方差模型。具体地说，我们只考虑一个离散型协变量（例如，班级规模是否小于 20 人）以及一个连续型协变量（例如，班级中的具体人数）。为了方便起见，我们将班级规模是否小于 20 人称为“班级规模”，而将具体人数称为“学生人数”。为了更好地理解协方差分析，我们首先考虑一个简单的问题：如果我们只考虑学生人数这一个协变量，那么我们应该用什么统计模型呢？答案似乎很简单，由于我们的因变量（教学评估成绩）以及唯一的协变量（学生人数）都是连续的，所以可以考虑以下线性回归模型：

$$\text{教学评估成绩} = \alpha + \beta \times \text{学生人数} + \varepsilon$$

其中，截距项 α 可以认为是在学生人数为零时的平均教学评估成绩。当然，这只是一个理论上的取值，以方便人们的理解与讨论。那么 β 代表什么呢？β 代表了学生人数对教学评估成绩的影响。更具体地说，β 说明从平均水平来看，学生人数每增加 1 人，教学评估成绩便会改变 β 个单位。由于我们的常识是班级规模越大，教学成绩越差，因此我们预期 β 是一个负数。请注意，该模型还隐含着一个假设，那就是：不管是大班（班级规模大于 20）还是小班（班级规模小于 20），教学评估成绩与学生人数之间都服从相同的等式关系。这又意味着什么样的实际后果呢？一个直接的后果是该模型认为不管班级规模的大小，学生人数的单位变化所带来的教学评估成绩的变化是一样的。

显然，认为班级规模的大小对教学评估成绩没有影响是不合理的。因此，我们需要对上面的普通线性回归模型予以改进，使得班级规模以及学生人数可以同时发生作用。那么，我们应该通过什么形式引入“班级规模”这个离散型变量呢？从直观上考虑，班级规模无非是一个分类变量。它将我们的 340 门课程分成了两组。一组是所有班级规模小于等于 20 人的课程（group = 1），另外一组是所有班级规模大于 20 人的课程（group = 0）。对于一个给定的分组（如 group = 1），我们完全可以假设教学评估成绩和学生人数之间服从一个普通线性模型（如前面定义）。这说明什么呢？说明我们可以根据班级规模的不同，定义不同的教学评估成绩对学生人数的回归直线。这样的话，班级规模和学生人数的影响都得到了考虑。那么，更具体地，班级规模是如何影响教学评估成绩对学生

人数的回归直线的呢？我们知道，一条普通的直线可以由两个不同的参数来决定，即截距(α)和斜率(β)。因此，如果班级规模的不同可以影响到截距或者斜率中的任何一个，那么班级规模的作用就得到了体现。

为了简单起见，我们首先考虑班级规模对截距的影响。具体地说，我们将前面所介绍的简单线性回归模型修改如下：

$$\text{教学评估成绩} = \alpha_0 + \alpha_1 \times \text{group} + \beta \times \text{学生人数} + \varepsilon$$

请比较该模型同前面定义的模型有什么异同。请注意，班级规模(group)是一个取值为0或者1的哑变量，因此，上述模型等价于下面两个回归模型：

$$\begin{cases} \text{教学评估成绩} = (\alpha_0 + \alpha_1) + \beta \times \text{学生人数} + \varepsilon, & \text{如果 group} = 1 \\ \text{教学评估成绩} = \alpha_0 + \beta \times \text{学生人数} + \varepsilon, & \text{如果 group} = 0 \end{cases}$$

由此可以很清楚地看到班级规模的实质作用就是根据group取值的不同，定义了两条不同的回归直线。这两条回归直线有不同的地方，也有相同的地方。不同的地方是它们的截距项。对于小规模的班级(group=1)，截距项为$(\alpha_0 + \alpha_1)$；而对于大规模班级(group=0)，截距项为α_0。那么，它们的差异$(\alpha_0 + \alpha_1) - \alpha_0 = \alpha_1$就是班级规模的效应。如果$\alpha_1 = 0$，那么不同的班级规模所定义的回归直线是一致的。这说明班级规模的不同，不会带来不同的教学评估成绩。因此，班级规模对教学评估成绩无影响。当然，如果$\alpha_1 \neq 0$，那么班级规模对教学评估成绩就有影响。刚才提到，班级规模所定义的两条回归直线也有相同的地方，那么相同的地方是什么呢？那就是斜率。因此，本模型仍然认为不管班级规模的大小，学生人数的单位变化所带来的教学评估成绩的变化是一样的。当然，这是一件值得商榷的事情，但是我们先暂时作此假定，并在R环境中拟合如下：

```
> lm1=lm(score~as.factor(group)+size)
> summary(lm1)

Call:
lm(formula = score ~ as.factor(group) + size)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6527 -0.2262  0.0876  0.3597  0.7417

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.2367448  0.7666886  63.530   <2e-16 ***
as.factor(group)1 0.1850426  0.0714341   2.590    0.010 *
size             -0.0005185  0.0011724  -0.442    0.659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4667 on 337 degrees of freedom
Multiple R-squared: 0.03847,    Adjusted R-squared: 0.03276
F-statistic: 6.742 on 2 and 337 DF,  p-value: 0.001346
```

如何解读上面的报表呢？首先要搞明白截距项的含义。请注意，group 是一个离散型变量，而且只有两个取值：0 和 1。但是，上面的报表中没有出现关于 group = 0 的估计。这说明，截距项所代表的是 group = 0 的回归直线的截距项。因此，我们可以根据以上的估计，构造两条回归直线如下：

$$\begin{cases} \text{教学评估成绩} = (4.237 + 0.185) - 0.001 \times \text{学生人数} + \varepsilon, & \text{如果 group} = 1 \\ \text{教学评估成绩} = 4.237 - 0.001 \times \text{学生人数} + \varepsilon, & \text{如果 group} = 0 \end{cases}$$

我们可以看到，关于 group = 1 的估计高达 0.185，而且其 P 值很小（0.010），在 0.05 的水平下高度显著。因此，我们可以确信小规模班级（group = 1）的教学评估成绩确实要高于大规模班级（group = 0）。另外，我们还可以看到，关于学生人数（size）的估计非常小（-0.001），而且高度不显著。这说明在同等的班级规模下，学生人数对教学评估成绩没有显著的影响。为了更加直观地看到我们的拟合效果，我们作两条回归直线，如图 3-4 所示。

```
> plot(size,score)
> points(size,lm1$fitted,col=2)
```

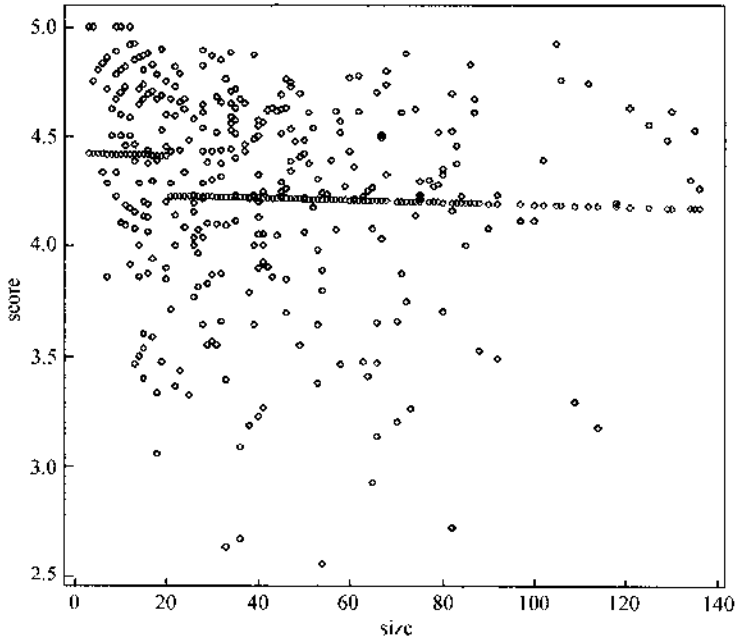


图 3-4　不同班级规模的回归直线图

从图 3-4 可以看到两条平行的直线，它们分别为不同班级规模的回归直线。很显然，这样的拟合效果并不理想。特别是对于小规模班级，教学评估成绩有一个很明显的下降趋势，但是我们所拟合的回归直线却近乎于一条水平的直

线。该模型除了拟合效果不理想以外，在实际应用中的解释能力也很差。具体地讲，如果一个班级的规模为 20 人（group = 1），那么该模型预测其教学评估成绩约为 4.237 + 0.185 = 4.422。但是，如果一个班级的规模为 21 人（group = 0），那么该模型预测其教学评估成绩约为 4.237。请注意，仅仅一人之差却引起了将近 0.2 分的巨大差异，因此，我们需要进一步改进我们的模型。

第四节　单因素交互作用模型

前一节讨论了如何将离散型解释变量（班级规模）的作用表现在回归直线的截距项上。对于本案例而言，我们发现这样分析的效果并不理想。因此，我们进一步探讨如何将班级规模的作用表现在回归直线的斜率上。具体地说，我们可以考虑下面的协方差分析模型：

$$\text{教学评估成绩} = \alpha_0 + \alpha_1 \times \text{group} + \beta_0 \times \text{学生人数} + \beta_1 \times \text{group} \times \text{学生人数} + \varepsilon$$

请比较该模型同前一节所定义的模型有什么异同。请注意，班级规模（group）是一个取值为 0 或者 1 的哑变量。因此，上述模型等价于下面两个回归模型：

$$\begin{cases} \text{教学评估成绩} = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) \times \text{学生人数} + \varepsilon, & \text{如果 group} = 1 \\ \text{教学评估成绩} = \alpha_0 + \beta_0 \times \text{学生人数} + \varepsilon, & \text{如果 group} = 0 \end{cases}$$

由此可以很清楚地看到，班级规模的实质作用就是根据 group 取值的不同，定义了两条非常不同的回归直线，这两条回归直线不仅截距不同，斜率也不同。具体地说，对于小规模的班级（group = 1），截距项为 $(\alpha_0 + \alpha_1)$ 而斜率为 $(\beta_0 + \beta_1)$；对于大规模的班级（group = 0），截距项为 α_0 而斜率为 β_0。两条回归直线截距项的差异为 $(\alpha_0 + \alpha_1) - \alpha_0 = \alpha_1$，这就是班级规模效应在截距项上的反映，而两条回归直线斜率的差异为 $(\beta_0 + \beta_1) - \beta_0 = \beta_1$，这就是班级规模效应在斜率上的反映。如果 $\alpha_1 = \beta_1 = 0$，那么不同的班级规模所定义的回归直线是相同的。这说明，不同的班级规模不会带来不同的教学评估成绩。因此，班级规模对教学评估成绩无影响。当然，如果 $\alpha_1 \neq 0$ 或者 $\beta_1 \neq 0$，那么班级规模对教学评估成绩就有影响。该模型可以在 R 环境中拟合如下：

```
> lm2=lm(score~as.factor(group)*size)
> library(car)
> Anova(lm2,type="III")
Anova Table (Type III tests)

Response: score
                          Sum Sq  Df   F value    Pr(>F)
(Intercept)               870.98   1 4083.1961 < 2.2e-16 ***
as.factor(group)            3.64   1   17.0620 4.571e-05 ***
size                    0.002311   1    0.0108 0.9171621
as.factor(group):size       2.37   1   11.1218 0.0009485 ***
Residuals                  71.67 336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中我们可以看到，班级规模（group）的主效应高度显著，其 P 值为 4.571e－05，在 0.001 的水平下高度显著。因此，我们再一次确定了班级规模的重要性。还值得注意的是，学生人数（size）的 P 值很大（约为 0.9172），这说明学生人数不重要吗？答案是否定的。请注意，班级规模（group）和学生人数（size）的交互作用的 P 值很小，大约为 0.00095，在 0.001 的水平下高度显著。因此，我们知道班级规模（group）和学生人数（size）的交互作用是很显著的。只要该交互作用显著，或者学生人数（size）的主效应显著，我们就可以认为学生人数（size）这个因素是很重要的。下面我们再具体分析各个因素对教学评估成绩的影响的大小以及形式。在 R 中可以具体实现如下：

```
> summary(lm2)

Call:
lm(formula = score ~ as.factor(group) * size)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65404 -0.23660  0.07534  0.34725  0.72012

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.2165679  0.0659871  63.900  < 2e-16 ***
as.factor(group)1         0.6923756  0.1676202   4.131 4.57e-05 ***
size                     -0.0001209  0.0011613   0.104 0.917162
as.factor(group)1:size -0.0377247  0.0113119  -3.335 0.000948 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4619 on 336 degrees of freedom
Multiple R-squared: 0.06928,    Adjusted R-squared: 0.06097
F-statistic: 8.337 on 3 and 336 DF,  p-value: 2.314e-05
```

如何解读这张报表呢？同前面类似，我们首先研究离散型变量班级规模（group）。该变量有两个水平（0 和 1），但是在上面的报表中只出现了水平 1。因此，我们可以知道，截距项以及学生人数（size）的主效应是在估计大规模班级（group = 0）的回归直线。也就是说，对于学生人数超过 20 人的班级，教学评估成绩可以通过下面的线性回归模型表示：

$$\text{教学评估成绩} = 4.2166 - 0.0001 \times \text{学生人数} + \varepsilon$$

由于学生人数（size）的主效应的估计量（-0.0001）太小、而且不显著（P 值 = 0.917），因此，我们可以对大规模班级（group = 0）得出以下结论：① 平均教学评估成绩约为 4.2166；② 学生人数的变化对教学评估成绩影响甚微。

上面讨论了大规模班级（group = 0）的回归直线，那么小规模班级（group = 1）的回归直线会是什么样子呢？首先可以看到的是，对于班级规模（group）的主效应和班级规模与学生人数的交互作用（group * size）的统计检验都是高度显著的，它们的 P 值分别为 0.0000 和 0.0009。这说明班级规模（group）是一个很重要的影响因素，它的不同取值不仅会影响到回归直线的截距，而且还会影响到其斜率。再结合报表中的具体估计，我们可以得到关于小规模班级的回归直线如下：

$$\begin{aligned}\text{教学评估成绩} &= (4.2166 + 0.6924) - (0.0001 + 0.0377) \\ &\quad \times \text{学生人数} + \varepsilon \\ &= 4.9090 - 0.0378 \times \text{学生人数} + \varepsilon\end{aligned}$$

这说明，如果一个教员讲授一门有 0 个学生的课程（当然这是不可能的），那么他的教学评估成绩可以高达 4.909。然后，随着学生人数的增加，其教学评估成绩以 0.0378 分/人的速度下降。这说明，对于小规模班级（group = 1），学生人数（size）的影响是很显著的。最后，为了便于理解，我们作该模型所拟合的两条回归直线，如图 3-5 所示，

```
> plot(size,score)
> points(size,lm2$fitted,col=2)
```

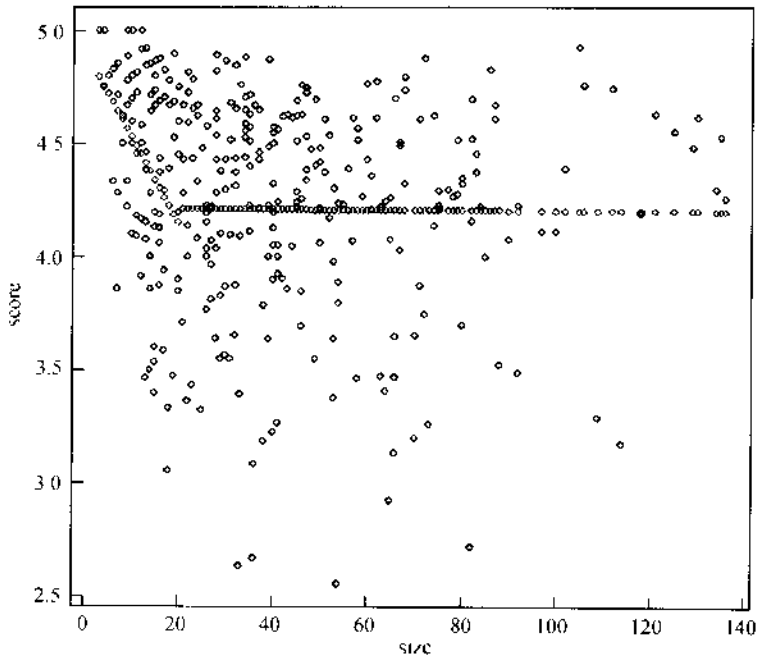


图 3-5　带交互作用的回归直线图

从图 3-5 我们可以清楚地看到，以学生人数 20 为界，我们拟合了两条不同的回归直线。其中，学生人数小于 20 的回归直线截距很高且下降迅速，而学生人数大于 20 的回归直线基本保持水平。

第五节　多因素协方差分析

为了简单起见，前面几节只讨论了一个离散型变量（班级规模）和一个连续型变量（学生人数）之间的关系。而我们前面提到，本案例还涉及许多其他的重要因素，包括教员职称（助理教授、副教授、正教授）、教员性别（男、女）、学生类别（MBA、本科生、研究生）、年份（2002、2003、2004）以及学期（秋季、春季）。因此，本节将讨论如何通过一个协方差分析模型对这些因素予以综合分析。具体地说，我们考虑如下统计模型：

$$\text{教学评估成绩} = \text{教员职称} + \text{教员性别} + \text{学生类别} + \text{年份} + \text{学期} + \text{班级规模} \times \text{学生人数} + \varepsilon$$

其中，ε 是无法被我们考虑的这些客观因素解释的教学评估成绩。换句话说，和

原始的教学评估成绩相比，ε 是剔除了教员职称、教员性别、学生类别、年份、学期、班级规模以及学生人数这些客观因素影响后的教学评估成绩。因此，虽然 ε 不可能是一个完美的教学评估成绩，但是应该比原始教学评估成绩能够更好地反映教员的努力程度与授课效果。因此，ε 和原始教学评估成绩相比，应该是一个更好的评估指标。而要准确估计 ε，前提就是要能够准确地量化其他各个影响因素的作用。在 *R* 中，我们可以实现如下：

```
> lm3.1=lm(score~as.factor(title)+as.factor(gender)+as.factor(student)
+ +as.factor(year)+as.factor(semester)+as.factor(group)*size)
> Anova(lm3.1,type="III")
Anova Table (Type III tests)

Response: score
                          Sum Sq  Df   F value    Pr(>F)
(Intercept)              252.177   1 1351.7607 < 2.2e-16 ***
as.factor(title)           1.938   2    5.1954 0.0060072 **
as.factor(gender)          0.047   1    0.2520 0.6160271
as.factor(student)         3.243   2    8.6916 0.0002099 ***
as.factor(year)            3.511   2    9.4109 0.0001061 ***
as.factor(semester)        0.012   1    0.0639 0.8005250
as.factor(group)           2.605   1   13.9612 0.0002201 ***
size                       0.106   1    0.5670 0.4519841
as.factor(group):size      2.010   1   10.7749 0.0011397 **
Residuals                 61.190 328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中可以发现：① 教员职称（title）、学生类别（student）、年份（year）、班级规模（group）和学生人数（size）都是重要的影响因素；② 教员性别（gender）和学期（semester）影响其微。剔除不显著的影响因素（教员性别和学期），重新拟合分析如下：

```
> lm3.2=lm(score~as.factor(title)+as.factor(student)+as.factor(year)+as.factor(group)*size)
> Anova(lm3.2,type="III")
Anova Table (Type III tests)

Response: score
                          Sum Sq  Df   F value    Pr(>F)
(Intercept)              416.91   1 2246.5544 < 2.2e-16 ***
as.factor(title)           3.57   2    9.6151 8.737e-05 ***
as.factor(student)         3.24   2    8.7373 0.0002007 ***
as.factor(year)            3.57   2    9.6100 8.779e-05 ***
as.factor(group)           2.57   1   13.8368 0.0002343 ***
size                       0.09   1    0.4669 0.4949033
as.factor(group):size      2.00   1   10.7536 0.0011518 **
Residuals                 61.24 330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中可以看到，新模型中所有因素都在 0.10 的水平下显著。值得注意的是，学生人数（size）也是一个显著的影响因素，因为它和班级规模（group）的交互作用高度显著（*P* 值 =0.001152）。下面，我们再具体地分析各个参数估计。

```
> summary(lm3,2)

Call:
lm(formula = score ~ as.factor(title) + as.factor(student) +
    as.factor(year) + as.factor(group) * size)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67461 -0.25094  0.08589  0.31584  0.80545

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.9160820  0.0826215  47.398  < 2e-16 ***
as.factor(title)正教授          0.1748254  0.0524794   3.331 0.000962 ***
as.factor(title)助理教授       -0.0962026  0.0694582  -1.385 0.166975
as.factor(student)本科生        0.0953513  0.0671288   1.420 0.156428
as.factor(student)研究生        0.2415691  0.0577866   4.180 3.73e-05 ***
as.factor(year)2003             0.1470044  0.0589738   2.493 0.013167 *
as.factor(year)2004             0.2414840  0.0567421   4.256 2.72e-05 ***
as.factor(group)1               0.5961725  0.1602708   3.720 0.000234 ***
size                            0.0007578  0.0011090   0.683 0.494903
as.factor(group)1:size         -0.0349458  0.0106566  -3.279 0.001152 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4308 on 330 degrees of freedom
Multiple R-squared: 0.2047,     Adjusted R-squared: 0.1831
F-statistic:  9.44 on 9 and 330 DF,  p-value: 8.422e-13
```

如何解读这张复杂的报表呢？和前面一样，我们首先要搞清楚截距项的含义。这里我们共涉及四个离散型因素：教员职称(title)、学生类别(student)、年份(year)和班级规模(group)。然后，我们可以在上面的报表中寻找每个因素的哪个水平消失了，答案如下：

- 教员职称：副教授
- 学生类别：MBA
- 年份：2002 年
- 班级规模：人数大于 20 人的班级

这说明，我们的标准课程类型为 2002 年副教授讲授的 MBA 课程，而且学生人数超过 20 人。此标准课程对应于一条标准的关于学生人数的回归直线。上表中的截距项(Intercept = 3.9160820)就是该回归直线的截距项，而学生人数的主效应(size = 0.0007578)就是该回归直线的斜率。如果我们关心的是一门 2002 年副教授讲授的 MBA 课程，而且学生人数不超过 20 人，那么它的回归直线应该是什么样的呢？答案是：

$$\text{教学评估成绩} = 3.916 + (0.001 - 0.035) \times \text{学生人数} + \varepsilon$$
$$= 3.916 - 0.034 \times \text{学生人数} + \varepsilon$$

再考虑一个更加复杂的例子：如果我们感兴趣的课程是 2003 年某助理教授讲授的本科生课程，而且学生人数不超过 20 人，那么其回归直线应该为：

$$\text{教学评估成绩} = (3.916 - 0.096 + 0.095 + 0.147)$$

$$+ (0.001 - 0.035) \times \text{学生人数} + \varepsilon$$
$$= 4.062 - 0.034 \times \text{学生人数} + \varepsilon$$

根据上述报表中的估计结果，我们可以获得以下重要结论：

- 教员职称显著影响教学评估成绩。随着教员职称的提高（从助理教授到副教授再到正教授），平均教学评估成绩（以中位数计）依次增高。
- 不同的学生类别对教学评估成绩影响很大。普通研究生给出的平均教学评估成绩明显高于本科生和 MBA，而本科生和 MBA 之间差异不大。
- 随着时间的推移，北京大学光华管理学院的教学评估成绩稳步提高。
- 小于 20 人的班级的教学评估成绩明显高于大于 20 人的班级。
- 学生人数对大规模班级影响甚微，但是对小规模班级影响显著。

最后，为了确保模型分析结果的可靠性，我们诊断如下（参见图 3-6）：

```
> par(mfrow=c(2,2))
> plot(lm3.3,which=c(1:4))
> par(mfrow=c(1,1))
```

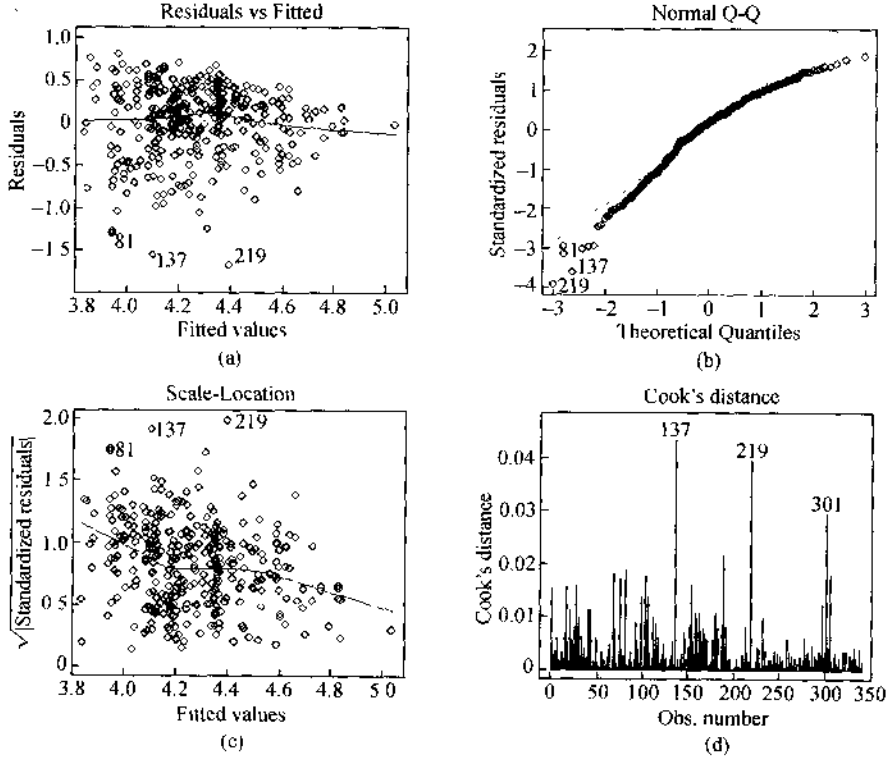


图 3-6 模型诊断图

由图 3-6 可以看出，该模型分析结果虽然不是非常完美（如 QQ 图），但是可以接受。

第六节　模型选择与预测

在前一节中，我们确定了几个能够显著影响教学评估成绩的因素，包括教员职称（title）、学生类别（student）、年份（year）、班级规模（group）和学生人数（size）。具体地说，我们是通过假设检验的方法，然后在 0.10 的显著性水平下发现它们能够显著地影响因变量。这样的方法有它的优点，也有它的缺点。它的主要缺点就是显著性水平的选择问题。对于一个实际问题，我们到底应该选择 0.10 的显著性水平，还是 0.05 的显著性水平？这很难决定。此外，在给定的显著性水平下所选取的模型在预测精度上是否是最优的，也无从知晓。因此，同第一章类似，我们考虑下面两种不同的模型选择标准：

$$\mathrm{AIC} = n\left\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\right\} + 2 \times (df + 1)$$

$$\mathrm{BIC} = n\left\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\right\} + \log(n) \times (df + 1)$$

其中，RSS 是该模型产生的残差平方和，df 是该模型的自由度，即自由参数的个数。如果我们考虑的是全模型，那么每个因素各自消耗多少自由度呢？详细分析如下：

- 截距项：消耗 1 个自由度
- 教员职称：3 个水平，消耗 2 个自由度
- 性别：2 个水平，消耗 1 个自由度
- 学生类别：3 个水平，消耗 2 个自由度
- 年份：3 个水平，消耗 2 个自由度
- 学期：2 个水平，消耗 1 个自由度
- 班级规模（主效应）：2 个水平，消耗 1 个自由度
- 学生人数：连续型变量，消耗 1 个自由度
- 班级规模 * 学生人数：2 个水平，消耗 1 个自由度

把所有因素消耗的自由度相加求和，可以得到全模型总共消耗了 $df = 12$ 个自由度。然后，我们对全模型作方差分析如下：

```
> Anova(lm3.1,type="III");
Anova Table (Type III tests)

Response: score
                        Sum Sq  Df   F value    Pr(>F)
(Intercept)            252.177   1 1351.7607 < 2.2e-16 ***
as.factor(title)         1.938   2    5.1954 0.0060072 **
as.factor(gender)        0.047   1    0.2520 0.6160271
as.factor(student)       3.243   2    8.6916 0.0002099 ***
as.factor(year)          3.511   2    9.4109 0.0001061 ***
as.factor(semester)      0.012   1    0.0639 0.8005250
as.factor(group)         2.605   1   13.9612 0.0002201 ***
size                     0.106   1    0.5670 0.4519841
as.factor(group):size    2.010   1   10.7749 0.0011397 **
Residuals               61.190 328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中我们可以看到，该模型所产生的残差平方和（*RSS*）为 61.190。因此，其相应的 AIC 和 BIC 的值分别为：

$$\begin{aligned}\text{AIC} &= 340\left\{\log\left(\frac{61.190}{340}\right)+1+\log(2\pi)\right\}+2\times(12+1)\\ &= 407.79\end{aligned}$$

$$\begin{aligned}\text{BIC} &= 340\left\{\log\left(\frac{61.190}{340}\right)+1+\log(2\pi)\right\}+\log(340)\times(12+1)\\ &= 457.57\end{aligned}$$

在 *R* 环境中，可以自动计算如下：

```
> AIC(lm3.1);
[1] 407.7906
> AIC(lm3.1,k=log(length(score)));
[1] 457.5669
```

在前一节中，我们通过假设检验的方法，在 0.10 的显著性水平下研究了如下模型：

```
> Anova(lm3.2,type="III")
Anova Table (Type III tests)

Response: score
                      Sum Sq  Df   F value    Pr(>F)
(Intercept)           416.91   1 2246.5544 < 2.2e-16 ***
as.factor(title)        3.57   2    9.6151 8.737e-05 ***
as.factor(student)      3.24   2    8.7373 0.0002007 ***
as.factor(year)         3.57   2    9.6100 8.779e-05 ***
as.factor(group)        2.57   1   13.8368 0.0002343 ***
size                    0.09   1    0.4669 0.4949033
as.factor(group):size   2.00   1   10.7536 0.0011518 **
Residuals              61.24 330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

其相应的 AIC 和 BIC 的值分别为：

```
> AIC(lm3.2)
[1] 404.069
> AIC(lm3.2,k=log(length(score)))
[1] 446.1874
```

两者分别小于全模型的 407.79 和 457.57。由此可见，不论根据 AIC 标准，还是 BIC 标准，我们都认为简化后的模型优于全模型。当然，这仅仅比较了两个模型，我们还可以根据 AIC 标准，对更多的模型进行比较如下：

```
> lm.aic=step(lm3.1,trace=F)
> Anova(lm.aic,type="III")
Anova Table (Type III tests)

Response: score
                      Sum Sq  Df   F value    Pr(>F)
(Intercept)           416.91   1 2246.5544 < 2.2e-16 ***
as.factor(title)        3.57   2    9.6151 8.737e-05 ***
as.factor(student)      3.24   2    8.7373 0.0002007 ***
as.factor(year)         3.57   2    9.6100 8.779e-05 ***
as.factor(group)        2.57   1   13.8368 0.0002343 ***
size                    0.09   1    0.4669 0.4949033
as.factor(group):size   2.00   1   10.7536 0.0011518 **
Residuals              61.24 330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

该选择结果是 *R* 自动搜索并比较了许多（但不是全部）各种各样的线性模型后获得的最优结果（以 AIC 为标准）。它与前面通过假设检验方法所得到的结果相同，即我们认为能够显著影响教学评估成绩的因素为教员职称（title）、学生类别（student）、年份（year）、班级规模（group）和学生人数（size）。类似地，我们也可以根据 BIC 标准，对更多的模型进行比较如下：

```
> lm.bic=step(lm3.1,k=log(length(score)),trace=F)
> Anova(lm.bic,type="III")
Anova Table (Type III tests)

Response: score
                       Sum Sq  Df   F value    Pr(>F)
(Intercept)            416.91   1 2246.5544 < 2.2e-16 ***
as.factor(title)         3.57   2    9.6151 8.737e-05 ***
as.factor(student)       3.24   2    8.7373 0.0002007 ***
as.factor(year)          3.57   2    9.6100 8.779e-05 ***
as.factor(group)         2.57   1   13.8368 0.0002343 ***
size                     0.09   1    0.4669 0.4949033
as.factor(group):size    2.00   1   10.7536 0.0011518 **
Residuals               61.24 330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

我们又获得了与前面相同的结果。因此，我们从不同角度，采用不同方法的分析结果反复证明了能够显著影响教学评估成绩的因素共五个，即教员职称（title）、学生类别（student）、年份（year）、班级规模（group）和学生人数（size）。假设我们下个学期将会开设以下五门课程：

```
> a0=read.csv("D:/Practical Business Data Analysis/case/CH3/new.csv",header=T)
> a0$group=1*(a0$size<=20)
> a0
     title gender student year semester size group
1   副教授      女     MBA 2002      秋季  114     0
2   副教授      男  研究生 2004      秋季   15     1
3 助理教授      女  本科生 2004      秋季   38     0
4   正教授      男  研究生 2002      春季   40     0
5   正教授      男  研究生 2003      春季   21     0
```

我们可以对其教学评估成绩预测如下：

```
> score.hat=predict(lm.aic,a0)
> a0$score.hat=score.hat
> a0
     title gender student year semester size group score.hat
1   副教授      女     MBA 2002      秋季  114     0  4.002468
2   副教授      男  研究生 2004      秋季   15     1  4.482488
3 助理教授      女  本科生 2004      秋季   38     0  4.185510
4   正教授      男  研究生 2002      春季   40     0  4.362787
5   正教授      男  研究生 2003      春季   21     0  4.495394
```

第七节　更科学的绩效评估

作为本章的最后一节，我们将简单演示一下，如何利用合理的回归模型对原始的教学评估成绩予以调整，使之更加科学合理。假设我们将采用的模型就是上一节中通过 AIC 以及 BIC 挑选出来的模型，即：

```
summary(lm(formula = score ~ as.factor(title) + as.factor(student) +
    as.factor(year) + as.factor(group) + size + as.factor(group):size))

Call:
lm(formula = score ~ as.factor(title) + as.factor(student) +
    as.factor(year) + as.factor(group) + size + as.factor(group):size)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67461 -0.25094  0.08589  0.31584  0.80545

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.9160820  0.0826215  47.398  < 2e-16 ***
as.factor(title)正教授      0.1748254  0.0524794   3.331 0.000962 ***
as.factor(title)助理教授    -0.0962026  0.0694562  -1.385 0.166975
as.factor(student)本科生    0.0953513  0.0671268   1.420 0.156428
as.factor(student)研究生    0.2415691  0.0577886   4.180 3.73e-05 ***
as.factor(year)2003        0.1470044  0.0589738   2.493 0.013167 *
as.factor(year)2004        0.2414840  0.0567421   4.256 2.72e-05 ***
as.factor(group)1          0.5961725  0.1602708   3.720 0.000234 ***
size                       0.0007578  0.0011090   0.683 0.494903
as.factor(group)1:size    -0.0349458  0.0106566  -3.279 0.001152 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4308 on 330 degrees of freedom
Multiple R-squared: 0.2047,     Adjusted R-squared: 0.1831
F-statistic:  9.44 on 9 and 330 DF,  p-value: 8.422e-13
```

然后，假设某位女性副教授在 2002 年秋季讲授了一门 114 人的 MBA 课程，最终的教学评估成绩为 3.175 分。而另外一位男性副教授在 2002 年秋季讲授了一门 92 人的研究生课程，最终的教学评估成绩为 3.489 分。那么，到底哪一位教授的教学成绩更好？首先，我们分析第一位教授。如果从平均水平来看，我们的模型认为她的预期教学评估成绩应该为 $3.9160820 + 0.0007578 \times 114 = 4.002471$，很遗憾，她的实际教学评估成绩 3.175 低于 4.002471，差距为 -0.827471。而该差异反映了剔除所有重要的但非教员主观所能够控制的因素后的实际教学评估成绩。因此，可以认为这是一个较原始教学评估成绩更加合理的成绩。我们将其称为调整后的教学评估成绩。类似地，根据我们的模

型，可以对第二位教授的预期教学评估成绩预测如下：3.9160820 + 0.2414840 + 0.0007578 × 92 = 4.227284。而该教授的实际成绩 3.489 低于该预期水平，差距为 −0.738284。因此，他的调整后的教学评估成绩为 −0.738284。同前一位女性教授的 −0.827471 相比，还稍微好一点，但都低于其应有的预期水平。细心的读者可以发现，以上的计算过程实际上就是在分离我们模型中的残差（ε），在 R 中我们可以轻松计算如下：

```
> a$adj.score=lm.aic$residuals
> a[c(1:10),]
   title gender student year semester size score  adj.score
1  副教授      女      MBA 2002      秋季  114 3.175 -0.8274681
2  副教授      女      MBA 2002      秋季   88 3.523 -0.4597660
3  副教授      女      MBA 2003      秋季   83 4.458  0.3320184
4  副教授      女      MBA 2002      秋季   66 3.470 -0.4960950
5  副教授      女      MBA 2003      秋季   46 4.630  0.5320560
6  副教授      女      MBA 2004      秋季   45 4.511  0.3193342
7  副教授      女      MBA 2002      秋季   38 3.184 -0.7608774
8  副教授      女      MBA 2002      秋季   31 3.548 -0.3915730
9  副教授      女      MBA 2003      秋季   30 4.433  0.3471804
10 副教授      女      MBA 2004      秋季   30 4.300  0.1197008
```

基于调整后的教学评估成绩（adj. score），我们可以对不同课程的教学效果重新排序。在新的排序结果的基础上，我们可以将调整后的教学评估成绩转换成我们习惯的单位或进制（如百分制），从而更加科学有效地为教学管理服务。

第八节　简单分析报告

教学评估数据分析报告

内容提要　本报告对北京大学光华管理学院的教学评估数据进行分析，找出影响最终教学评估成绩的因素，并量化了这些影响因素的相对重要性。从我们的分析结果发现，影响最终教学评估成绩的主要因素有教员职称、学生类别、年份、班级规模和学生人数。本报告的分析结果可以为教师的教学评估提供一个客观有效的绩效评估标准，从而更加科学有效地为教学管理服务。

一、研究目的

在大专院校的教学管理中，教学评估是一种重要的衡量教员教学成绩的手段。如果该评估手段非常准确，那么我们就可以通过简单地比较两门课的教学评估成绩来比较两个教员的教学绩效。因此，如何客观有效地对教员的教学进行评估，是一件非常重要且基本的工作。本报告试图通过对北京大学

光华管理学院教学评估数据的分析，建立一个计量经济学模型，以此来找出影响最终教学评估成绩的因素，并根据数据分析的结果，提出一个合理的绩效考核标准。

二、数据来源和相关说明

本报告所使用的数据来自于北京大学光华管理学院的教学评估记录，共有340条有效记录，其中每一条记录都对应于2002年至2004年这三年间，在北京大学光华管理学院开设的某一门课程。因变量是我们所关心的课程的最终评估得分。另外，我们的数据包括以下解释性变量：教员职称（助理教授、副教授、正教授）、教员性别（男、女）、学生类别（MBA、本科生、研究生）、年份（2002、2003、2004）、学期（秋季、春季）以及学生人数。值得注意的是，在我们所考虑的解释性变量中，学生人数（即班级中学生的数目）是一个具有数值意义的变量，可以简单地看做连续型变量，而其他的所有解释性变量都是离散型变量。

三、描述性分析

在所有解释性变量中，学生人数是唯一的连续型变量。经验告诉我们，学生人数是一个非常重要的因素，我们利用散点图（即图3-7）来寻找它们之间的关系。

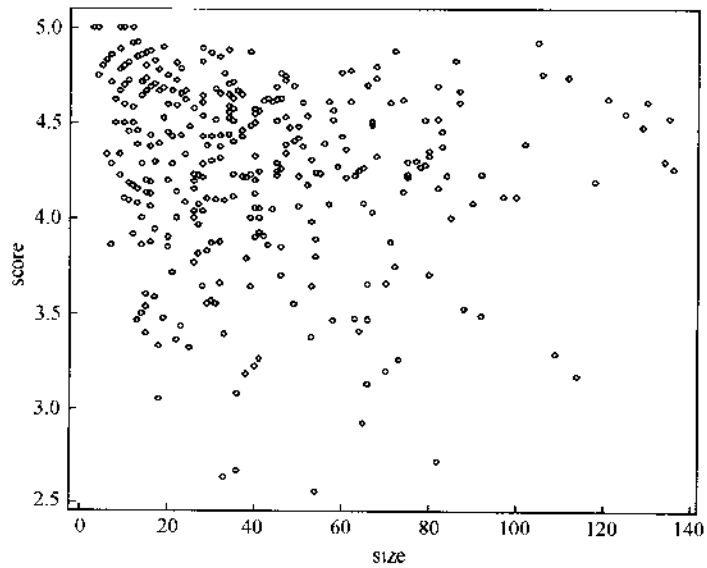


图3-7　最终评估得分与学生人数的散点图

通过散点图可以发现，最终评估得分与学生人数呈现杂乱无章的关系，并没有明显的统计规律，但这很可能是噪音太大的缘故。为了准确地找出最终评估得分与学生人数之间的关系，我们考虑将学生人数进行离散化，即对学生人数进行分组。在综合考虑各组的样本量和相互关系之后，我们发现 20 是一个有意义的分界值。因此，我们定义哑变量班级规模如下：如果学生人数小于等于 20，那么班级规模取值为 1，否则取值为 0。

为了从直观上获得对各个离散型变量与因变量之间关系的初步认识，我们利用盒状图对数据进行简单的描述性分析，得到图 3-8。

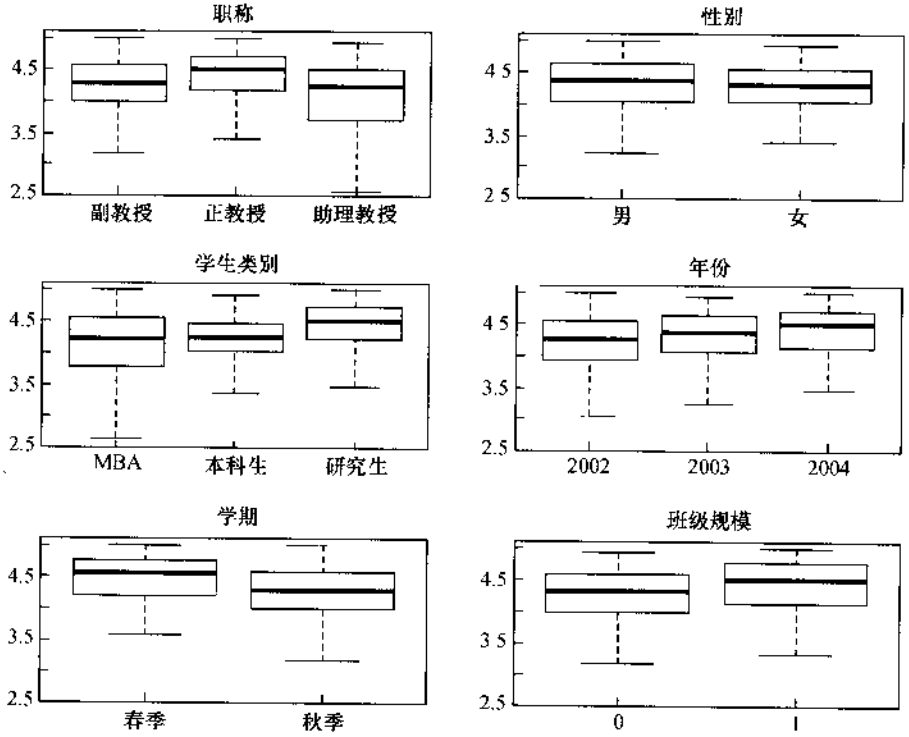


图 3-8　基于盒状图的描述性分析

从图 3-8 中我们可以得到如下直观的印象：

- 教员职称确实能够影响教学评估成绩。随着教员职称的提高（从助理教授到副教授再到正教授），平均教学评估成绩（以中位数计）依次增高。这在一定程度上反映出积累的教学科研经验对教学评估成绩的影响。
- 教员不同性别的两组之间差别很小，表明性别对数学评估成绩影响甚微。

• 不同的学生类别对教学评估成绩影响很大。研究生给出的平均教学评估成绩明显高于本科生和 MBA，而本科生和 MBA 之间差异不大。

• 随着时间的推移，各组的中位数在依次提高，表明北京大学光华管理学院的教学质量在稳步提高。

• 秋季学期（即每学年的第一个学期）的教学评估成绩低于春季学期。

• 小于 20 人的班级的教学评估成绩明显高于大于 20 人的班级。

四、数据建模

1. 全模型分析

通过上一节的分析我们可以发现，自变量与因变量之间确实存在相关性。我们用多因素协方差分析的方法建立模型，来寻找自变量和因变量之间的关系。从先前的分析中发现，最终评估得分与班级规模之间有显著的相关性，但是与学生人数之间的关系并不显著，因此我们考虑班级规模与学生人数之间的交互作用。我们单独对其进行的检验也证明该交互作用是显著的。而对其他的变量，我们考虑可加模型。利用全部数据对此全模型进行估计，我们得到模型估计结果（如表 3-1 所示）。

表 3-1　全模型方差分析

变量名	*P* 值
教员职称	<0.0001
教员性别	0.7565
学生类别	<0.0001
年份	0.0002
学期	0.6963
班级规模	0.0680
学生人数	0.6690
班级规模 * 学生人数	0.0011
残差项标准差 0.4319	模型 *F* 检验 *P* 值 <0.0001
判决系数（R-square）0.2054	调整的判决系数（R-square）0.1787

从表 3-1 中我们可以看到，模型 *F* 检验的 *P* 值非常小，表明该模型是显著的，即自变量和因变量之间确实存在一定的关系。另外，未调整的判决系数（R-square）为 20.54%，调整后的判决系数（adjusted R-square）为 17.87%，这都表明该模型对自变量和因变量之间的关系有一定的解释能力。通过对各个自变量所对应的 *t* 检验的 *P* 值的考察，在 0.10 的显著水平下，我们可以断定：① 教员职称、学生类别、年份、班级规模和学生人数都是重要的影响因素；② 教

员性别和学期对最终评估得分影响甚微。以上五个重要的影响因素也大致与人们的直觉相一致。教员职称在一定程度上代表了教员的教学能力，教学能力高的教员自然能得到较高的教学评估成绩；MBA 和本科生对教员的期望值一般都较高；随着学院的成长，教员各方面经验的积累在教学上也有显著的体现，这就导致最终教学评估得分逐年递增；班级规模是一个很重要的影响因素，小规模的班级总是比大规模的班级更容易教授，而且教员也能有更多的精力来关注每位同学。

2. 模型选择

我们很容易从以上全模型的分析结果中发现，有五个自变量非常重要，对最终评估得分有显著的影响，但是我们不能排除其他变量也有预测能力的可能。因此，我们用两种最为常用的选择变量的方法，即 AIC 和 BIC，来选择最具有预测能力的模型。通过计算，我们发现 AIC 和 BIC 选择了同样的模型，而且正好是包括上述五个重要影响因素的模型。因此，我们可以认为该模型是一个科学合理的模型，称为最优模型。对该模型的参数估计结果如表 3-2 所示。

表 3-2　最优模型参数估计结果

变量名	系数估计值	标准差	*P* 值
截距项	3.9161	0.0826	<0.0001
教员职称：正教授	0.1748	0.0525	0.0010
教员职称：助理教授	-0.0962	0.0695	0.1670
学生类别：本科生	0.0954	0.0671	0.1564
学生类别：研究生	0.2416	0.0578	<0.0001
年份：2003	0.1470	0.0590	0.0132
年份：2004	0.2415	0.0567	<0.0001
班级规模：1	0.5962	0.1603	0.0002
学生人数	0.0008	0.0011	0.4949
班级规模：1 * 学生人数	-0.0349	0.0107	0.0012

为了确保模型分析结果的可靠性，我们对模型的独立性、正态性以及同方差假定进行检验，发现这些假定基本满足。根据表 3-2 中的估计结果，我们可以获得以下重要结论：

- 教员职称显著影响教学评估成绩。随着教员职称的提高（从助理教授到副教授再到正教授），平均教学评估成绩也依次提高。
- 不同的学生类别对教学评估成绩影响很大。普通研究生给出的平均教学评估成绩明显高于本科生和 MBA，而本科生和 MBA 之间差异不大。
- 随着时间的推移，北京大学光华管理学院的教学评估成绩稳步提高。

- 小于 20 人的班级的教学评估成绩明显高于大于 20 人的班级。
- 学生人数对大规模班级影响甚微，但是对小规模班级影响显著。

五、结论及建议

根据以上分析结果，我们知道有五个因素对教学评估成绩有显著的影响，这为教学管理提供了重要的参考。管理部门可以根据这些影响因素更加合理地安排课程计划，如对未来的最终教学评估成绩进行预测，并根据预测结果来合理安排老师的课程。另外，根据最优模型，我们可以得到一个较为客观合理的教学绩效评估标准。因为模型中的因素解释了诸如教员职称、授课年份等客观因素带来的影响，所以我们可以分离出模型的残差，把它作为调整后的教学评估成绩并对不同课程的教学效果重新排序。在新的排序结果的基础上，将调整后的教学评估成绩转换成我们习惯的单位或进制（如百分制），从而更加科学有效地为教学管理服务。

[讨论总结]

本章以北京大学光华管理学院教学评估为例，系统演示并讲解了协方差分析。通过对本章的学习，读者应该能够了解：什么时候可以使用协方差分析，以及如何使用。由于协方差是普通线性模型和方差分析的自然结合，因此，无论是对 *R* 语言学习还是统计理论，都没有特别新的地方。因此，读者可以将本章作为一个很好的复习。对相关统计学理论渴望深入了解的读者可以参阅 Rao (1973)、Draper and Smith(1981)，还有 Milliken and Johnson(2002b)。

附录　程序及注释

```
rm(list=ls())                                                    # 清空当前工作空间
a=read.csv("D:/Practical Business Data Analysis/case/CH3/teaching.csv",header=T)
                                                                 # 读入 csv 格式的数据,并赋值给 a
attach(a)                                                        # 将 a 中各变量加入工作空间
a[c(1:5),]                                                       # 展示 a 的前 5 行数据
plot(size,score)                                                 # 画出 size 与 score 的散点图
boxplot(score ~ ceiling(size/20))                                # 画出 score 与分组的 size 的盒状图
table(ceiling(size/20))                                          # 计算分组的 size 的频数
group=1*(size <=20)                                              # 根据 size 是否大于 20 生成 0、1 变量
par(mfrow=c(3,2))                                                # 设置画图模式为 3x2
boxplot(score ~ title,main="职称")                               # 画出 score 与 title 的盒状图
boxplot(score ~ gender,main="性别")                              # 画出 score 与 gender 的盒状图
boxplot(score ~ student,main="学生类别")                         # 画出 score 与 student 的盒状图
boxplot(score ~ year,main="年份")                                # 画出 score 与 year 的盒状图
boxplot(score ~ semester,main="学期")                            # 画出 score 与 semester 的盒状图
boxplot(score ~ group,main="班级规模")                           # 画出 score 与 group 的盒状图
par(mfrow=c(1,1))                                                # 设置画图模式,还原成 1x1
lm1=lm(score ~ as.factor(group) + size)                          # 用解释性变量 group 和 size 拟合线性模型
summary(lm1)                                                     # 显示模型 lm1 的各方面细节,包括参数估计值、P 值等
plot(size,score)                                                 # 画出 size 与 score 的散点图
```

```
points(size,lm1$fitted,col=2)                    # 在散点图上用红色画出 size 与 lm1$fitted 的数据点
lm2=lm(score ~ as.factor(group)* size)           # 拟合带交互作用的线性模型
library(car)                                     # 载入程序包 car
Anova(lm2,type="III")                            # 对模型 lm2 作三型方差分析
summary(lm2)                                     # 显示模型 lm2 的各方面细节,包括参数估计值、P 值等
plot(size,score)                                 # 画出 size 与 score 的散点图
points(size,lm2$fitted,col=2)                    # 在散点图上用红色画出 size 与 lm2$fitted 的数据点
lm3.1=lm(score ~ as.factor(title) + as.factor(gender) + as.factor(student) + as.factor(year) + as.factor(semester) + as.factor(group)* size)
                                                 # 全模型
Anova(lm3.1,type="III")                          # 对模型 lm3.1 作三型方差分析
lm3.2=lm(score ~ as.factor(title) + as.factor(student) + as.factor(year) + as.factor(group)* size)
                                                 # 删除全模型中不显著的变量,重新拟合
Anova(lm3.2,type="III")                          # 对模型 lm3.2 作三型方差分析
summary(lm3.2)                                   # 显示模型 lm3.2 的各方面细节,包括参数估计值、P 值等
par(mfrow=c(2,2))                                # 设置画图模式为 2x2 的格式
plot(lm3.2,which=c(1:4))                         # 画出 lm3.2 中模型检验的 4 张图,包括残差图、QQ 图和 Cook 距离图
par(mfrow=c(1,1))                                # 设置画图模式,还原成 1x1
Anova(lm3.1,type="III")                          # 对模型 lm3.1 作三型方差分析
AIC(lm3.1)                                       # 计算模型 lm3.1 的 AIC 值
AIC(lm3.1,k=log(length(score)))                  # 计算模型 lm3.1 的 BIC 值
Anova(lm3.2,type="III")                          # 对模型 lm3.2 作三型方差分析
AIC(lm3.2)                                       # 计算模型 lm3.2 的 AIC 值
AIC(lm3.2,k=log(length(score)))                  # 计算模型 lm3.2 的 BIC 值
lm.aic=step(lm3.1,trace=F)                       # 根据 AIC 准则从 lm3.1 中选出最优模型
Anova(lm.aic,type="III")                         # 对模型 lm.aic 作三型方差分析
```

```
lm.bic=step(lm3.1,k=log(length(score)),trace=F)                    # 根据 BIC 准则从 lm3.1 中选出最优模型
Anova(lm.bic,type="III")                                           # 对模型 lm.bic 作三型方差分析
a0=read.csv("D:/Practical Business Data Analysis/case/CH3/new.csv",header=T)
                                                                   # 读入用于预测的数据,并赋值给 a0
a0$group=1*(a0$size<=20)                                           # 根据 size 是否大于 20 生成新的 0、1 变量,并赋值给 a0$group
a0                                                                 # 展示数据 a0
score.hat=predict(lm.aic,a0)                                       # 利用 lm.aic 对 a0 进行预测
a0$score.hat=score.hat                                             # 将预测值赋给 a0 的变量 score.hat
a0                                                                 # 展示数据 a0,此时包括预测值
summary(lm.aic)                                                    # 显示模型 lm.aic 的各方面细节,包括参数估计值、P 值等
a$adj.score=lm.aic$residuals                                       # 将模型 lm.aic 中的残差赋值给 a 中的变量 adj.score
a[c(1:10),]                                                        # 展示数据 a0 的前 10 行,此时包括 adj.score
```

第四章
0-1变量的回归模型

- 案例介绍
- 基本描述
- 单变量逻辑回归
- 参数估计与统计推断
- 多变量逻辑回归
- 模型选择
- 预测与评估
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过股票特殊处理(ST)的实际案例,详细介绍 logit 和 probit 回归这两种重要的统计回归模型。它们主要处理的是因变量为 0-1 型数据的情形。通过对本章的学习,我们希望读者能够了解:(1) 什么情况下使用 logit 或 probit 回归;(2) 0-1 变量回归分析的基本统计学理论;(3) 相关理论在统计学软件 *R* 中的应用;(4) 相应的统计分析报告的撰写。本章所涉及的新统计学概念有 0-1 变量、logit 回归、probit 回归、极大似然估计等。

第一节 案例介绍

前面三章所涉及的数据类型有一个共同之处,那就是不管自变量的类型如何,因变量必须是连续的(如净资产收益率、商品房价格以及教学评估成绩)。但是,现实生活中有很多因变量不是连续型的情况。例如,银行希望通过公司的财务信息预测其破产的可能性,从而决定是否应该发放贷款,那么我们的因变量将是破产与否。该因变量代表着事件的两种可能的结果(破产、不破产),而没有任何数量意义,更不可能是连续型变量。保险公司希望通过驾驶员的驾驶记录预测其来年出险的可能性,并以此确定其相应的保险费用。在这种情况下,我们感兴趣的因变量是出险与否,这又是一个事件的两种可能结果(出险、不出险)。信用卡经理希望通过会员的消费记录预测其是否会购买某项新推出的产品。在这种情况下,我们关心的因变量是购买与否,这同样也是一个事件的两种可能结果(购买、不购买)。

由这些不同的实例,我们可以看到,它们的因变量都没有任何数值意义,都代表着某个事件的两种可能结果。我们可以简单地定义其中任何一种可能结果为 1(如破产、出险、购买),而定义另外一种可能结果为 0(如不破产、不出险、不购买)。值得注意的是,这里的 0 和 1 只是符号,没有任何数值意义。理论上讲,我们完全可以将它们分别定义为 A 和 B,或者甲和乙。但是,为了数学讨论的方便,我们一般定义为 0 和 1。对于这样的数据,我们同样希望能够给出回归模型。具体地说,我们希望知道哪些因素能够影响我们的 0-1 因变量?影响的程度如何?给定一些解释性变量(如会员的历史购买记录),我们能否预测因变量的未来取值?本章将以我国股票市场的特殊处理(special treatment, ST)政策

作为案例，全面讨论这一问题。

我们首先对特殊处理政策作一简要介绍。特殊处理政策是我国股票市场一项特有的、旨在保护投资者利益的政策。根据相关规定，如果某上市公司出现财务状况或其他状况异常，以至于投资者难以判断公司前景，并且投资者权益可能受到损害时，中国证券监督管理委员会将考虑对该公司股票的交易实行特殊处理。为了起到警示投资者的作用，该股票的名称前将冠以“ST”字样。根据相关规定，有可能导致特殊处理的典型原因是“最近两个会计年度的审计结果显示的净利润均为负值”。一旦某股票被特殊处理，那么该股票报价的日涨跌幅就会被人为地限制在5%以内。此外，如果该公司在下一个会计年度还不能取得正盈利，那么其股票将面临退市的风险。显然，股票的特殊处理会给上市公司和投资者带来巨大的经济损失。因此，如何利用公开的财务报表信息预测公司的ST状态（ST=1：被特殊处理；ST=0：没有被特殊处理）就成了人们关注的热点问题。

接下来我们将考虑，有哪些公开的财务指标可能与公司是否被特殊处理相关？本案例考虑了以下财务指标：

- ARA：应收账款与总资产的比例，用于衡量盈利质量。
- ASSET：对数变换后的资产规模，用于反映公司规模。
- ATO：资产周转率，用于度量资产利用效率。
- GROWTH：销售收入增长率，用于反映公司的成长潜力。
- LEV：负债资产比率，用于反映债务状况。
- ROA：资产收益率，用于度量盈利能力。
- SHARE：最大股东的持股比例，用于反映股权结构。

这样一个指标体系的设计是不是很完美？肯定不是。但是，该指标体系大体上能够全面地反映一个公司最重要的一些方面，从而足以为案例演示服务。

最后值得一提的是，我们的数据共包含1 430个完整的观测。其中，684个观测来自1999年，即解释性变量来自1999年，我们用这部分数据来建立模型，剩下的746个观测来自2000年，我们用这部分数据检验模型的预测效果。我们的因变量是什么呢？如果我们的解释性变量来自1999年，那么因变量ST就反映该公司在三年以后（即2002年）是否被宣布ST。类似地，如果我们的解释性变量来自2000年，那么因变量ST就反映该公司在三年以后（即2003年）是否被宣布ST。

第二节　基本描述

按照惯例，我们首先对数据作简要描述以获得初步的认识，并明确下一步的分析。我们首先读入数据，并将 1999 年和 2000 年的数据分离如下：

```
> a=read.csv("D:/Practical Business Data Analysis/case/CH4/st.csv",header=T)
> a1=a[a$year==1999,-1]
> a2=a[a$year==2000,-1]
> a1[c(1:5),]
         ARA     ASSET     ATO      GROWTH       LEV          ROA SHARE ST
1 0.19230963 19.85605 0.0052 -0.9507273 0.4458801 0.087709802 26.89  0
2 0.22011996 20.91086 0.0056 -0.9426563 0.3986864 0.016820383 39.62  0
3 0.32529169 19.35262 0.0166 -0.9374404 0.3033481 0.042468332 26.46  0
4 0.02572868 21.43893 0.0028 -0.8529953 0.7582502 0.018151630 60.16  0
5 0.53359089 21.61334 0.2552 -0.8167039 0.7268753 0.004146607 54.24  1
```

从输出结果的第一行我们可以看到，某公司在 1999 年应收账款与总资产的比例约为 0.19，对数总资产为 19.86，资产周转率为 0.52%，销售收入增长率为 -0.95，债务资产比为 0.45，资产收益率为 0.088，第一大股东持股比例为 26.89%。该公司在三年后（即 2002 年）没有被特殊处理（ST = 0）。而从上面数据的第五行我们可以看到，还有一家公司在 1999 年应收账款与总资产的比例约为 0.53，对数总资产为 21.61，资产周转率为 25.52%，销售收入增长率为 -0.82，债务资产比为 0.73，资产收益率为 0.004，第一大股东持股比例为 54.24%。该公司在三年后（即 2002 年）不幸被特殊处理（ST = 1）了。

由于我们的因变量是一个离散的 0-1 变量，因此传统的散点图无法有效地表示因变量（ST 与否）同各个解释性变量的相互关系。而此时，盒状图却非常有效。我们首先对应收账款与总资产比例（ARA）分析如下（参见图 4-1）：

```
> boxplot(ARA~ST,data=a1,main="ARA")
```

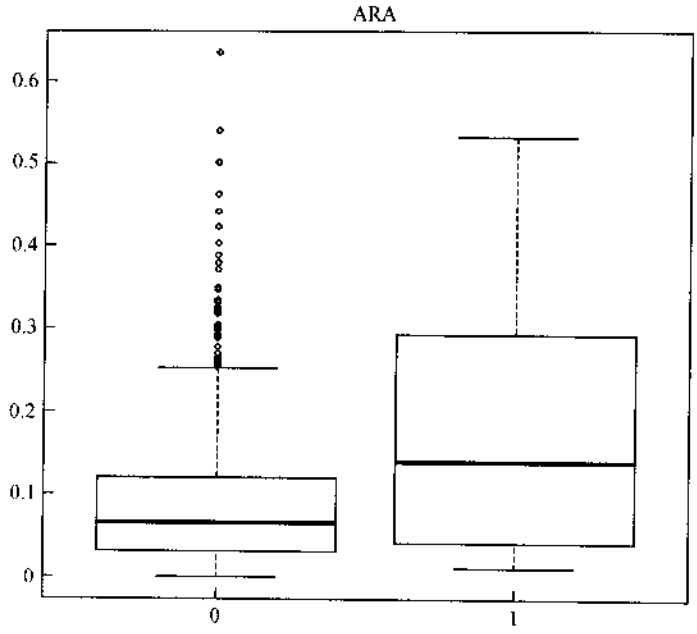


图 4-1　应收账款与总资产比例的盒状图

从图 4-1 可以发现一个重要的规律，那就是被特殊处理的那组样本（ST = 1）所反映出来的平均 ARA 值（以中位数计）要明显地高于没有被特殊处理的那组样本（ST = 0）。因此，我们可以猜测较高的应收账款与总资产比例很可能产生较大的被特殊处理的可能性。下面，我们对其他六个解释性变量作类似的分析（参见图 4-2）：

```
> par(mfrow=c(3,2))
> boxplot(ASSET~ST,data=a1,main="ASSET")
> boxplot(ATO~ST,data=a1,main="ATO")
> boxplot(GROWTH~ST,data=a1,main="GROWTH")
> boxplot(LEV~ST,data=a1,main="LEV")
> boxplot(ROA~ST,data=a1,main="ROA")
> boxplot(SHARE~ST,data=a1,main="SHARE")
> par(mfrow=c(1,1))
```

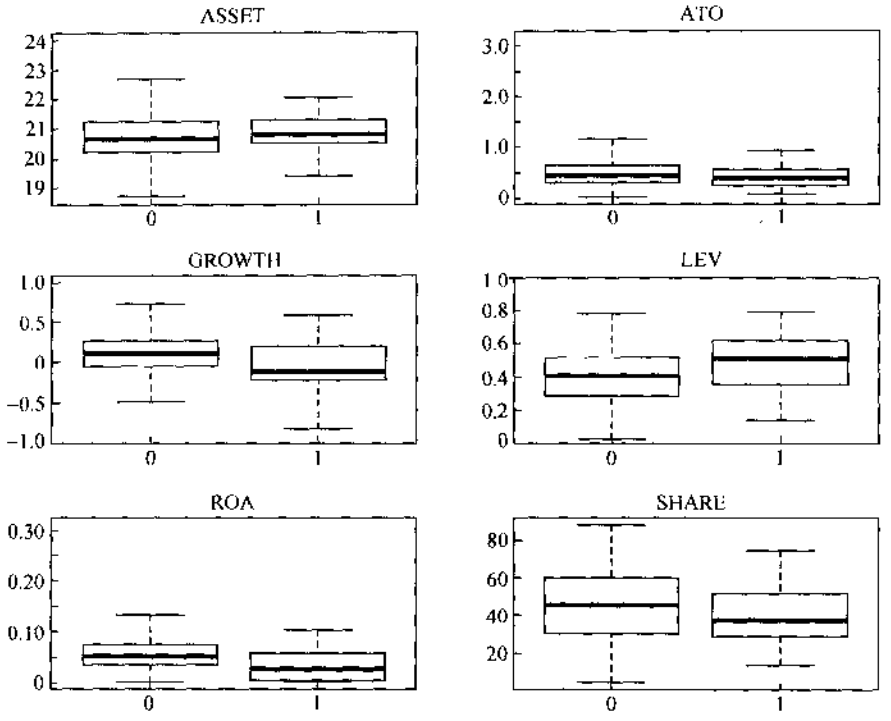


图 4-2 盒状图

从图 4-2 我们可以得到以下重要结论：

- 公司规模(ASSET)同被特殊处理与否似乎没有太强的关系。
- 公司的资产周转率(ATO)同被特殊处理与否没有明显的关系。
- 被特殊处理的公司的平均销售增长率(GROWTH,以中位数计)明显低于没有被特殊处理的公司。
- 被特殊处理的公司的负债水平(LEV,以中位数计)明显高于没有被特殊处理的公司。
- 被特殊处理的公司的盈利能力(ROA,以中位数计)明显低于没有被特殊处理的公司。
- 被特殊处理的公司的第一大股东持股比例(SHARE,以中位数计)明显低于没有被特殊处理的公司。

以上都是对数据进行初步的描述性分析。对于所得到的结论：第一，没有控制其他因素的影响；第二，没有经过严格的统计检验。而这些问题将是我们下面章节所要研究的重要内容。

第三节　单变量逻辑回归

在我们正式介绍逻辑回归模型之前，我们首先需要回答一个问题，那就是：为什么我们需要逻辑回归？不可以使用我们前面三章详细讨论的普通线性模型吗？下面我们就来看一下，如果我们用普通线性模型拟合一个0-1变量会有什么现象发生。为了简便起见，我们暂时只考虑一个解释性变量（如负债资产比LEV）。那么，普通线性模型会作如下假设：

$$ST = \alpha + \beta \times LEV + \varepsilon$$

请注意，这是一个充满矛盾的等式。一方面，等号的右边是一个连续型的实数，理论上讲，它可以取正负无穷之间的任意实数。而另一方面，等号的左边是一个取值必须为0或者1的整数。在实际数据中，左右两边几乎永远不可能相等。也正是由于这样的矛盾，线性模型无法用来预测我们所感兴趣的0-1变量。因此，需要一种特殊的回归模型来处理0-1变量的回归问题。

我们可以看到，之所以普通线性模型不能够用来处理0-1型因变量主要是因为0-1变量不连续。这提示我们，只要我们能够将这个0-1型因变量转换成一个连续型因变量，那么普通线性回归模型的很多概念以及技术手段就可以直接套用。那么，怎样把0-1型因变量（ST与否）同一个连续型变量联系起来呢？

我们假设，对于每一个上市公司都有一个度量其被ST的可能性大小的综合指标。为了简便起见，我们称此指标为“ST可能性”。那么，什么样的公司会被ST呢？我们再假设，一旦某公司的ST可能性指标大于某一阈值，那么该公司就会被ST。请注意，我们这里的ST可能性指标只是一个概念，在现实中并不存在。但是，我们可以合理地推测，如果两个公司的运营状况非常类似，那么它们的ST可能性指标也应该非常近似。这说明什么呢？说明我们可以非常合理地假设这个看不见、摸不着的ST可能性指标是一个连续型的指标，而且还可以假设它的取值范围为正负无穷之间。既然我们假设一个公司的ST与否完全由这个ST可能性指标确定，因此，与其直接拟合ST与否同财务指标的关系，不如先探讨一下这个ST可能性指标同财务指标的关系。请注意，ST可能性指标是一个取值任意的连续型变量。此时，我们完全可以采用以下的普通线性模型：

$$Z = \alpha + \beta \times LEV + \varepsilon$$

其中，Z就代表了这个ST可能性指标。如果上帝告诉我们Z的具体取值，那么我们就可以直接使用第一章中详细讲述的普通线性回归分析的方法来建立上面这个模型。但是，现实中我们并不知道Z的具体取值，那么上面这个线性模

型对我们有什么用处呢？根据这个线性模型以及一个给定的债务水平(LEV)，我们可以判断某公司被 ST 的可能性为：

$$\begin{aligned} P(\mathrm{ST}=1) &= P(z>c) = P(\alpha+\beta\times\mathrm{LEV}+\varepsilon>c) \\ &= P\{-\varepsilon<(\alpha-c)+\beta\times\mathrm{LEV}\} \\ &= F_\varepsilon(\beta_0+\beta_1\times\mathrm{LEV}) \end{aligned}$$

其中，c 就是前面所提到的阈值，$\beta_0=\alpha-c$，$\beta_1=\beta$，而 $F_\varepsilon(t)=P(-\varepsilon<t)$ 是 $-\varepsilon$ 的分布函数。如果我们可以对 $F_\varepsilon(t)$ 的具体函数形式予以合理的假设(即假设 $-\varepsilon$ 的分布)，那么我们就获得了一个关于 0-1 变量的回归模型，即：

$$P(\mathrm{ST}=1)=F_\varepsilon(\beta_0+\beta_1\times\mathrm{LEV})$$

请注意，这个新设定的模型中没有任何地方涉及那个看不见、摸不着的 ST 可能性指标 Z，因此我们可以对该模型中的参数予以估计。其具体的估计方法我们将在下一节中详细讨论。在这里，我们首先需要回答一个问题：$F_\varepsilon(t)$ 的具体函数形式应该如何假设才合理？

理论上讲，关于 $F_\varepsilon(t)$ 具体函数形式的任何假设都是错的，但也都是对的。这是什么意思呢？我们可以负责任地说，所有统计模型都是对数据产生机制的一种近似而不是准确的刻画。因此，无论你对 $F_\varepsilon(t)$ 的具体函数形式作任何假设，该假设都不可能完全反映真实情形，因此都是错的。那为什么又都是对的呢？如果我们不对 $F_\varepsilon(t)$ 的具体函数形式作任何假设，那么我们就彻底失去了研究并预测 ST 的能力。因此，任何大致合理的关于 $F_\varepsilon(t)$ 具体函数形式的假设都有可能为我们的实践提供很有意义的指导。从这个意义上来说，任何关于 $F_\varepsilon(t)$ 具体函数形式的假设，大体上也都是对的。这就应了著名统计学家 G. Box 的一句名言："Every model is wrong, but some are useful."到目前为止，我们还是没有回答大家感兴趣的问题：$F_\varepsilon(t)$ 的具体函数形式到底应该如何假设？

既然所有的关于 $F_\varepsilon(t)$ 具体函数形式的假设都是错的，但又都可能是有用的，那么我们就应该挑选那些"方便"的假设。什么样的假设我们会认为"方便"呢？第一个答案：正态假设。也就是说，假设 $F_\varepsilon(t)$ 是一个标准正态分布函数。为什么我们认为这个假设方便呢？因为这个假设等同于假设 ε 服从标准正态分布，这是第一章普通线性回归模型中讲到的一个非常常见的标准假设。如果我们假设 ε 服从标准正态分布，那么相应的 0-1 变量回归就变成了普通线性回归模型的一个非常自然的推广，理论上非常具有吸引力。如果我们作此假设，相应的统计模型就变为：

$$P(\mathrm{ST}=1)=\Phi(\beta_0+\beta_1\times\mathrm{LEV})$$

其中，$\Phi(t)$ 代表了标准正态随机变量的分布函数。该模型被称为 probit 模型。大家可以看到，probit 模型是一个理论上具有吸引力，而且实际中表现良好的统

计模型，非常有用并且已经在各种统计软件中得以实现。但是，在几十年前，当计算机技术还没有普及的时候，估计该模型却非常困难。其主要原因就是标准正态分布函数 $\Phi(t)$ 没有显式解。因此，人们假定 $F_c(t)$ 的具体函数形式为下面这种容易计算的函数形式：

$$F_c(t) = \frac{\exp(t)}{1 + \exp(t)}$$

这就是逻辑分布，而相应的 0-1 变量回归模型就是逻辑回归模型（logit regression）。更详细地说，该模型隐含着下面的等式：

$$P(\mathrm{ST} = 1) = \frac{\exp(\beta_0 + \beta_1 \times \mathrm{LEV})}{1 + \exp(\beta_0 + \beta_1 \times \mathrm{LEV})}$$

那么，在实际应用中，到底应该采用 probit 回归还是 logit 回归？答案是：不确定。对于某些实际数据，有可能 probit 回归的预测精度高；而对于另外一些数据，有可能 logit 回归的预测精度高。但是，对于绝大多数数据，两种回归方法所产生的结果非常相似。因此，任何一种都非常有用。请注意，probit 回归以及 logit 回归还可以写成：

$$\Phi^{-1}\{P(\mathrm{ST} = 1)\} = \beta_0 + \beta_1 \times \mathrm{LEV}$$

$$\mathrm{logit}\{P(\mathrm{ST} = 1)\} = \log\left(\frac{P(\mathrm{ST} = 1)}{1 - P(\mathrm{ST} = 1)}\right) = \beta_0 + \beta_1 \times \mathrm{LEV}$$

其中，$\Phi^{-1}\{t\}$ 被称为 probit 变换，而 $\mathrm{logit}\{t\} = \log\{t/(1-t)\}$ 被称为 logit 变换。

第四节　参数估计与统计推断

接下来，我们要回答几个重要的理论问题。那就是：对于 0-1 变量回归模型，我们应该如何作参数估计？如何作统计推断？由于 probit 回归和 logit 回归的估计方法以及推断方法非常类似，因此，我们为简单起见，只对 logit 回归作详细讨论。如果上帝告诉我们那个看不见、摸不着的 ST 可能性指标 Z，那么我们就可以通过第一章中的最小二乘法来估计参数的取值，并进而作统计推断。但是，现实中 Z 的取值是未知的，因此我们必须考虑用其他方法来估计我们感兴趣的参数（即 β_0 和 β_1）。为了严格起见，假设 $(\mathrm{ST}_i, \mathrm{LEV}_i)$ 是来自于第 i 个 $(i = 1, 2, \cdots, n)$ 样本的观测。那么，这些样本的似然函数为：

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left\{\frac{\exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}{1 + \exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}\right\}^{\mathrm{ST}_i} \times \left\{\frac{1}{1 + \exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}\right\}^{1 - \mathrm{ST}_i}$$

理论上讲，合理的参数估计应该能够产生较大的似然函数 $L(\beta_0, \beta_1)$ 的值。因此，我们可以通过极大化 $L(\beta_0, \beta_1)$ 或者以下的对数似然函数：

$$\log\{L(\beta_0,\beta_1)\} = \sum_{i=1}^{n} \mathrm{ST}_i \times \log\left\{\frac{\exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}{1 + \exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}\right\} + \sum_{i=1}^{n} (1 - \mathrm{ST}_i) \times \left\{\frac{1}{1 + \exp(\beta_0 + \beta_1 \times \mathrm{LEV}_i)}\right\}$$

来获得参数估计。我们称此估计为极大似然估计(maximum likelihood estimate, MLE),并记为$(\hat{\beta}_0,\hat{\beta}_1)$。我们可以回忆一下,在普通线性回归模型中,如果$(\hat{\beta}_0,\hat{\beta}_1)$是最小二乘估计,$(\hat{\beta}_0,\hat{\beta}_1)$就会服从正态分布。那么,在 logit 回归模型中,$(\hat{\beta}_0,\hat{\beta}_1)$的具体分布是什么呢?很遗憾,我们不知道。但是,我们知道只要样本量足够大,根据中心极限定理,$(\hat{\beta}_0,\hat{\beta}_1)$会近似地服从正态分布。具体如下:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma(\hat{\beta}_j)} \sim N(0,1) \quad (j = 0,1)$$

这说明,只要我们能够准确地估计$\hat{\beta}_j$的标准差$\sqrt{\mathrm{var}(\hat{\beta}_j)} = \sigma(\hat{\beta}_j)$,记为$\hat{\sigma}(\hat{\beta}_j)$,那么我们就可以构造如下的检验统计量:

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

在原假设$\beta_j = 0$成立的情况下,该统计量T_j近似地服从标准正态分布。因此,对于一个给定的显著性水平(如 0.05),我们就可以根据T_j的绝对值是否大于$z_{0.975}$来决定是否拒绝原假设。

上面所介绍的检验方法叫做 Z 检验或者 t 检验。这两种检验方法的缺点是只能对一个因素(即 LEV)的显著性作检验,而不能够同时检验多个因素的显著性。如果 LEV 代表了好几个解释性变量,因而β_1是一个向量,那么我们应该如何检验β_1的各个分量是否同时为零呢?这就引出了另一种检验方法——似然比检验。似然比检验的思想非常简单:首先,我们考虑原假设$\beta_1 = 0$,在这种情况下,我们可以计算极大对数似然函数的取值为$\max_{\beta_0} \log\{L(\beta_0,\beta_1 = 0)\}$。然后,我们考虑如果允许$\beta_1 \neq 0$,那么极大对数似然函数的取值$\max_{(\beta_0,\beta_1)} \log\{L(\beta_0,\beta_1)\}$会改变多少。直观上考虑,如果确实$\beta_1 = 0$(即原假设成立),那么无论在原假设下还是在对立假设下极大化对数似然函数,其差别应该不大。否则,我们就可以断定$\beta_1 \neq 0$。因此,我们构造似然比检验(likelihood ratio test)统计量如下:

$$\lambda = -2 \times \left(\max_{\beta_0} \log\{L(\beta_0,\beta_1 = 0)\} - \max_{(\beta_0,\beta_1)} \log\{L(\beta_0,\beta_1)\}\right)$$

$$= (-2 \times \max_{\beta_0} \log\{L(\beta_0, \beta_1 = 0)\}) - (-2 \times \max_{(\beta_0,\beta_1)} \log\{L(\beta_0, \beta_1)\})$$

如果 β_1 是一个长度为 d 的向量(即 β_1 代表了 d 个不同的自由参数),那么经典的统计理论告诉我们,λ 近似服从一个自由度为 d 的 χ^2 分布,前提是原假设 $\beta_1 = 0$成立而且样本量足够大。请注意,上面的表达式同线性模型(更具体地说,是方差分析模型)中的残差分析非常类似。换句话说,$(-2 \times \max_{\beta_0} \log\{L(\beta_0, \beta_1 = 0)\})$很像是总平方和($SST$),而$(-2 \times \max_{(\beta_0,\beta_1)} \log\{L(\beta_0, \beta_1)\})$很像是残差平方和($RSS$)。这说明,对逻辑回归模型我们也可以作方差分析,而借助的工具就是$(-2 \times \log\{L(\beta_0, \beta_1)\})$,我们称为 deviance。

第五节　多变量逻辑回归

在前面几节中,为了讲解的方便,我们主要考虑的是单变量(LEV)逻辑回归模型。但是,大家可以从案例介绍以及描述性分析中看到,本案例所涉及的解释性变量远远大于一个。回忆一下,本案例所涉及的解释性变量包括:应收账款与总资产的比例(ARA)、对数变换后的资产规模(ASSET)、资产周转率(ATO)、销售收入增长率(GROWTH)、负债资产比率(LEV)、资产收益率(ROA)和最大股东的持股比率(SHARE)。如果我们希望能够将所有的解释性变量放在同一个逻辑回归模型下研究,那么其形式应该为:

$$\begin{aligned}\text{logit}\{P(\text{ST} = 1)\} = {} & \beta_0 + \beta_1 \times \text{ARA} + \beta_2 \times \text{ASSET} \\ & + \beta_3 \times \text{ATO} + \beta_4 \times \text{GROWTH} \\ & + \beta_5 \times \text{LEV} + \beta_6 \times \text{ROA} + \beta_7 \times \text{SHARE}\end{aligned}$$

与单变量逻辑回归模型类似,我们可以通过极大似然法获得参数估计,也可以通过 Z 检验来检验各个因素的显著性,还可以通过似然比检验来检验一组因素的整体显著性(如模型的整体显著性)。

在 R 中具体分析如下:我们首先考虑模型的整体显著性,也就是说,我们希望知道在我们所考虑的所有解释性变量中,是否至少有一个和因变量(ST 与否)显著相关。因此,我们的原假设是:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

我们可以通过似然比检验来达到目的,这意味着我们要比较以下两个模型:

- 空模型:$\text{logit}\{P(\text{ST} = 1)\} = \beta_0$
- 全模型:

$$\begin{aligned}\text{logit}\{P(\text{ST}=1)\} = {} & \beta_0 + \beta_1 \times \text{ARA} + \beta_2 \times \text{ASSET} \\ & + \beta_3 \times \text{ATO} + \beta_4 \times \text{GROWTH} \\ & + \beta_5 \times \text{LEV} + \beta_6 \times \text{ROA} + \beta_7 \times \text{SHARE}\end{aligned}$$

在 R 中，我们对这两个模型作方差分析如下：

```
> glm0.a=glm(ST~1,family=binomial(link=logit),data=a1)
> glm1.a=glm(ST~ARA+ASSET+ATO+GROWTH+LEV+ROA+SHARE,
+ family=binomial(link=logit),data=a1)
> anova(glm0.a,glm1.a)
Analysis of Deviance Table

Model 1: ST ~ 1
Model 2: ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE
  Resid. Df Resid. Dev  Df Deviance
1       683     282.071
2       676     251.506   7   30.565
```

可以看到，广义似然比检验的统计量为两个模型之差，即 30.565。在原假设下，它应该服从一个卡方分布，自由度为两个模型的自由参数之差，即 8 - 1 = 7。因此，我们可以计算模型的整体显著性水平如下：

```
> 1-pchisq(30.565,7)
[1] 7.477255e-05
```

这说明，该模型的整体高度显著，也就意味着我们所考虑的七个解释性变量中，至少有一个与企业是否被 ST 显著相关。如果我们改用 probit 回归，那么结果如下：

```
> glm0.b=glm(ST~1,family=binomial(link=probit),data=a1)
> glm1.b=glm(ST~ARA+ASSET+ATO+GROWTH+LEV+ROA+SHARE,
+ family=binomial(link=probit),data=a1)
> anova(glm0.b,glm1.b)
Analysis of Deviance Table

Model 1: ST ~ 1
Model 2: ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE
  Resid. Df Resid. Dev  Df Deviance
1       683     282.071
2       676     250.369   7   31.702
> 1-pchisq(31.702,7)
[1] 4.612104e-05
```

可以看到，该结果同逻辑回归的结果几乎一样。上面的分析告诉我们，至少有一个财务指标对企业 ST 与否具有一定的预测能力。但是，到底是哪一个财务指标我们并不知道，因此，需要作方差分析如下：

```
> library(car)
> Anova(glm1.a,type="III")
Anova Table (Type III tests)

Response: ST
        LR Chisq Df Pr(>Chisq)
ARA      10.1064  1   0.001478 **
ASSET     1.1999  1   0.273343
ATO       0.6517  1   0.419506
GROWTH    2.2259  1   0.135717
LEV       3.8706  1   0.049138 *
ROA       0.0105  1   0.918196
SHARE     1.0032  1   0.316543
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由此可以知道,应收账款的状况(ARA)以及债务水平情况(LEV)是两个影响特殊处理的重要因素。具体的参数估计为:

```
> summary(glm1.a)

Call:
glm(formula = ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,
    family = binomial(link = logit), data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4165  -0.3354  -0.2536  -0.1959   3.0778

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.86924    4.63586  -1.913  0.05573 .
ARA          4.87974    1.49245   3.270  0.00108 **
ASSET        0.24660    0.22409   1.100  0.27115
ATO         -0.50738    0.65744  -0.772  0.44026
GROWTH      -0.83335    0.56706  -1.470  0.14167
LEV          2.35415    1.20138   1.960  0.05005 .
ROA         -0.63661    6.22354  -0.102  0.91853
SHARE       -0.01111    0.01115  -0.997  0.31891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从这张表中,我们可以得到以下重要结论:

- 应收账款与总资产的比例(ARA)和被ST与否高度相关,其Z检验在0.01的显著性水平下高度显著(P值=0.00108)。并且可以看到,ARA的系数估计值的符号为正,这说明应收账款占总资产的比例越高,被ST的可能性越大。
- 债务资产比率(LEV)和企业被ST与否高度相关,其Z检验在0.10的显

著性水平下高度显著(P 值 $=0.05005$)。同样可以看到,LEV 的系数估计值的符号为正,这说明债务占总资产的比例越高,该企业被 ST 的可能性越大。

- 没有证据证明其他因素对预测 ST 与否有重要作用。

类似地,我们可以对 probit 回归分析如下:

```
> Anova(glm1.b,type="III")
Anova Table (Type III tests)

Response: ST
        LR Chisq Df Pr(>Chisq)
ARA      11.5938  1  0.0006617 ***
ASSET     1.5855  1  0.2079637
ATO       0.6221  1  0.4302550
GROWTH    3.0442  1  0.0810253 .
LEV       4.1680  1  0.0411942 *
ROA       0.0708  1  0.7901788
SHARE     0.9271  1  0.3356237
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(glm1.b)

Call:
glm(formula = ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,
    family = binomial(link = probit), data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3339  -0.3364  -0.2536  -0.1818   3.1031

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.954642   2.222463  -2.229 0.025791 *
ARA          2.646791   0.787440   3.361 0.000776 ***
ASSET        0.135393   0.107405   1.261 0.207457
ATO         -0.232909   0.297494  -0.783 0.433684
GROWTH      -0.473252   0.281559  -1.681 0.092796 .
LEV          1.195192   0.585527   2.041 0.041229 *
ROA          0.670231   2.788393   0.240 0.810048
SHARE       -0.004984   0.005225  -0.954 0.340200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到,基本结论同 logit 回归非常类似。唯一不同的是,在 0.10 的显著性水平下,probit 回归还发现销售收入增长率(GROWTH)对该企业被 ST 与否有显著影响。由于 GROWTH 的系数估计值的符号为负,因此我们可以认为销售收入增长越慢的公司,被 ST 的可能性越大。下面我们在 0.10 的显著性水平下剔除不显著的因素,重新拟合 probit 回归模型如下:

```
> glm2.b=glm(ST~ARA+GROWTH+LEV,family=binomial(link=probit),data=a1)
> summary(glm2.b)

Call:
glm(formula = ST ~ ARA + GROWTH + LEV, family = binomial(link = probit),
    data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3133  -0.3339  -0.2583  -0.2008   3.2517

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4468     0.2618  -9.345  < 2e-16 ***
ARA           2.7331     0.7491   3.648 0.000264 ***
GROWTH       -0.4713     0.2695  -1.749 0.080339 .
LEV           1.2059     0.5205   2.317 0.020510 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到，基本结论没有改变。同样的模型，用 logit 回归拟合的结果为：

```
> glm2.a=glm(ST~ARA+GROWTH+LEV,family=binomial(link=logit),data=a1)
> summary(glm2.a)

Call:
glm(formula = ST ~ ARA + GROWTH + LEV, family = binomial(link = logit),
    data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4020  -0.3323  -0.2647  -0.2118   3.1063

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6022     0.5646  -8.152 3.58e-16 ***
ARA           5.1301     1.4341   3.577 0.000347 ***
GROWTH       -0.9061     0.5471  -1.656 0.097708 .
LEV           2.5501     1.0778   2.366 0.017986 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到，该结果和 probit 回归的结果几乎一样。

第六节　模型选择

和普通线性回归一样，过分复杂的模型会降低预测精度，并且使结果复杂难懂。但是，过分简化的模型会遗漏重要的预测变量，降低预测精度。因此，有必要建立一个尽量简单但又具有良好预测能力的模型。假设检验的方法（如前

一节所介绍)不失为一种简单有效的方法。但是,它受到人为设定的显著性水平的限制,因而非常主观。因此,我们希望能够给逻辑回归模型以及 probit 回归模型也制定出一套类似于 AIC 和 BIC 的模型选择标准。在第四节中,我们曾经提到,0-1 变量回归的 deviance 能够起到一个类似于残差平方和的作用。而普通线性模型的 AIC 和 BIC 就是基于残差平方和而定义的。因此,这提示我们,对于 0-1 回归模型,我们可以定义 AIC 和 BIC 如下:

$$\text{AIC} = -2 \times \log\{L(\beta_0,\beta_1)\} + 2 \times df$$

$$\text{BIC} = -2 \times \log\{L(\beta_0,\beta_1)\} + \log(n) \times df$$

其中,n 代表了样本量,而 df 代表了自由度(即自由参数的个数)。下面,我们以逻辑回归的全模型以及空模型为例,具体演示 AIC 和 BIC 的计算方法。首先计算两个模型的 deviance 如下:

```
> deviance(glm0.a)
[1] 282.0707
> deviance(glm1.a)
[1] 251.5057
```

然后,再计算各自的自由度。对于空模型,我们只有截距项一个自由参数,因此 $df=1$;而对于全模型,我们有七个解释性变量,再加上一个截距项,所以总共有八个自由参数,即 $df=8$。因此,它们的 AIC 取值分别为:

$$\text{空模型}:282.0707 + 2 \times 1 = 284.0707$$

$$\text{全模型}:251.5057 + 2 \times 8 = 267.5057$$

由于我们的样本量为 684,因此,相应的 BIC 取值分别为:

$$\text{空模型}:282.0707 + \log(684) \times 1 = 288.5987$$

$$\text{全模型}:251.5057 + \log(684) \times 8 = 303.7293$$

因为无论是 AIC 还是 BIC 都是取值越小越好,因此,如果我们采用 AIC,我们会认为全模型优于空模型。但是,如果我们采用 BIC,我们就会认为空模型优于全模型。在 *R* 中,以上的计算过程可以自动完成如下:

```
> AIC(glm0.a)
[1] 284.0707
> AIC(glm1.a)
[1] 267.5057
> AIC(glm0.a,k=log(length(a1[,1])))
[1] 288.5987
> AIC(glm1.a,k=log(length(a1[,1])))
[1] 303.7293
```

当然,目前我们仅仅比较了两个不同的模型。而我们有七个解释性变量,因此一共有 $2^7=128$ 个不同的模型。理论上讲,我们应该对这 128 个模型逐一研究,并选择最优的模型。但是,这样做的缺点就是需要我们自己编写程序。

而在 *R* 中,我们可以自动地、尽量多地根据 AIC 搜索最优模型如下:

```
> logit.aic=step(glm1.a,trace=0)
> summary(logit.aic)

Call:
glm(formula = ST ~ ARA + GROWTH + LEV, family = binomial(link = logit),
    data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4020  -0.3323  -0.2647  -0.2118   3.1063

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6022     0.5646  -8.152 3.58e-16 ***
ARA           5.1301     1.4341   3.577 0.000347 ***
GROWTH       -0.9061     0.5471  -1.656 0.097708 .
LEV           2.5501     1.0778   2.366 0.017986 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

也可以自动地、尽量多地根据 BIC 搜索最优模型如下：

```
> n=length(a1[,1])
> logit.bic=step(glm1.a,k=log(n),trace=0)
> summary(logit.bic)

Call:
glm(formula = ST ~ ARA, family = binomial(link = logit), data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3227  -0.3195  -0.2708  -0.2412   2.6962

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6834     0.2722 -13.531  < 2e-16 ***
ARA           6.3316     1.3132   4.821 1.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

对于这个案例，AIC 和 BIC 得出了不一样的结论。AIC 认为共有三个因素同 ST 与否相关，它们分别为应收账款与总资产的比例（ARA）、销售收入增长率（GROWTH）以及债务资产比率（LEV）。但是，BIC 认为只有应收账款与总资产的比例（ARA）是重要的。由于在学术界关于 AIC 好还是 BIC 好至今没有定论，因此我们暂时无法判断到底哪一个模型更为可靠。但是，我们不妨将此结果理解成：AIC 所选择的三个因素都很重要，而其中的应收账款与总资产的比例（ARA）格外重要。笔者自己非常有限的经验是：BIC 选择的模型更简单，而 AIC 选择的模型的预测精度似乎更好。

我们也可以对 probit 回归模型作类似的模型选择，结果如下：

```
> probit.aic=step(glm1.b,trace=0)
> summary(probit.aic)

Call:
glm(formula = ST ~ ARA + GROWTH + LEV, family = binomial(link = probit),
    data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3133  -0.3339  -0.2583  -0.2008   3.2517

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4468     0.2618  -9.345  < 2e-16 ***
ARA           2.7331     0.7491   3.648 0.000264 ***
GROWTH       -0.4713     0.2695  -1.749 0.080339 .
LEV           1.2059     0.5205   2.317 0.020510 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> probit.bic=step(glm1.b,k=log(n),trace=0)
> summary(probit.bic)

Call:
glm(formula = ST ~ ARA, family = binomial(link = probit), data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1960  -0.3240  -0.2691  -0.2349   2.7221

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0050     0.1251 -16.031   <2e-16 ***
ARA           3.2021     0.7015   4.565    5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

可以看到，probit 回归的模型选择结果同 logit 回归完全一致。

第七节　预测与评估

一旦建立了 logit 回归模型（或者 probit 模型），我们就可以对新的观测予以预测。以全模型为例，假设我们有一家新的公司，那么可以对其未来被 ST 的概率估算如下：

$$F_e(\beta_0 + \beta_1 \times \mathrm{ARA} + \beta_2 \times \mathrm{ASSET} + \beta_3 \times \mathrm{ATO} + \beta_4 \times \mathrm{GROWTH} + \beta_5 \times \mathrm{LEV} + \beta_6 \times \mathrm{ROA} + \beta_7 \times \mathrm{SHARE})$$

如果是 logit 回归模型，那么 $F_e(t) = \exp(t)/[1+\exp(t)]$。如果是 probit 回归模型，那么 $F_e(t) = \Phi(t)$。下面我们考虑一个具体的例子：假设我们有这

样一家公司，它的各项财务指标分别为：ARA = 0.1923，ASSET = 19.8561，ATO = 0.0052，GROWTH = -0.9507，LEV = 0.4459，ROA = 0.0877，SHARE = 26.89。另外，logit 回归全模型的各项参数估计如下所示：

```
> summary(glm1.a)

Call:
glm(formula = ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,
    family = binomial(link = logit), data = a1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4165  -0.3354  -0.2536  -0.1959   3.0778

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.86924    4.63586  -1.913  0.05573 .
ARA          4.87974    1.49245   3.270  0.00108 **
ASSET        0.24660    0.22409   1.100  0.27115
ATO         -0.50738    0.65744  -0.772  0.44026
GROWTH      -0.83335    0.56706  -1.470  0.14167
LEV          2.35415    1.20138   1.960  0.05005 .
ROA         -0.63661    6.22354  -0.102  0.91853
SHARE       -0.01111    0.01115  -0.997  0.31891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

因此，我们首先计算线性组合如下：

$$\begin{aligned}&-8.86924 + 4.87974 \times 0.1923 + 0.24660 \times 19.8561\\&\quad - 0.50738 \times 0.0052 + 0.83335 \times 0.9507\\&\quad + 2.35415 \times 0.4459 - 0.63661 \times 0.0877\\&\quad - 0.01111 \times 26.89 = -1.5496\end{aligned}$$

然后，再计算被 ST 的概率为：

$$P(\mathrm{ST}=1) = \frac{\exp(-1.5496)}{1+\exp(-1.5496)} = 0.175$$

最后，我们就可以根据这个概率来预测该公司是否会被 ST。例如，我们可以考虑以 0.50 为界，凡是 $P(\mathrm{ST}=1)>0.50$ 的公司都预测为 ST = 1，否则，预测 ST = 0。因此，对于这个例子来说，我们应该预测 ST = 0（因为 $0.175<0.50$），但是，0.50 是不是一个最好的阈值呢？那可不一定。对于这个问题，我们下面还会仔细地讨论。我们先暂时以 0.50 为界，对本章开头所留下的检验样本作预测并总结如下：

```
> p=predict(glm1.a,a2)
> p=exp(p)/(1+exp(p))
> a2$ST.pred=1*(p>0.5)
> table(a2[,c(8,9)])
   ST.pred
ST     0
  0 699
  1  47
```

从这张报表中可以看到以下重要事实：

- 训练样本共包含 699 + 47 = 746 个样本。
- 训练样本中共有 699 家非 ST 公司，它们被成功地预测为 ST = 0。
- 训练样本中共有 47 家 ST 公司，它们被错误地预测为 ST = 0。

综合这些数字，可以简单地判断我们的预测精度为 699/746 = 93.7%。这是一个非常高的精确度！但是，事实真的如此乐观吗？至少有一点是显然的。精度高达 93.7% 的预测结果中，没有正确预测一家 ST 公司！这显然不令人满意。这说明单纯利用总体的预测精度来评判一种预测方法是有问题的。更具体地说，对于 0-1 变量，我们有可能犯两种不同的错误（就像假设检验的第一类错误与第二类错误）。这两种错误定义如下：

- 把真实的 ST 公司预测为 ST = 0。
- 把真实的非 ST 公司预测为 ST = 1。

对于这个案例，我们最关心的是找出那些 ST 公司。因此，我们可以通过以下两个指标间接地度量以上两种错误的大小：

- True Positive Rate（TPR）：把真实的 ST 公司正确地预测为 ST = 1 的概率。
- False Positive Rate（FPR）：把真实的非 ST 公司错误地预测为 ST = 1 的概率。

TPR 就像是假设检验中的功效（power），越大越好；而 FPR 就像是假设检验中的显著性大小（size），越小越好。那么，刚才以 0.50 为阈值的预测方法有什么问题呢？我们可以看到，它的 FPR 值为 0（非常好），但是它的 TPR 值也是 0（非常差）。因此，这不可能是一种很好的预测方法。那么，如果我们把阈值改为 0，那又会怎样呢？

```
> a2$ST.pred=1*(p>0)
> table(a2[,c(8,9)])
   ST.pred
ST     1
  0 699
  1  47
```

这次我们看到,所有的 ST 公司都被成功地预测到了。因此,TPF 为 100%(非常好)。但是,我们也把所有的非 ST 公司错误地预测为 ST = 1。因此,FPR 也是 100%(非常差)。由此我们可以看到,TPR 和 FPR 是鱼与熊掌不可兼得的。一个合理的预测方案应该取得某种平衡。此外,如果一种预测方法比另外一种预测方法好,那么第一种方法就应该在任何一个 FPR 值下,都有更高的 TPR 取值。我们注意到,在训练样本中大约有 5% 的样本是 ST 公司,因此,我们使用 0.05 作为阈值重新预测如下:

```
> a2$ST.pred=1*(p>0.05)
> table(a2[,c(9,9)])
   ST.pred
ST    0   1
  0 504 195
  1  18  29
```

对于这样一个预测结果,它的 TPF 值为 29/(18 + 29) = 61.7%,而 FPR 值为 195/(504 + 195) = 27.9%。这显然是一个比前面两种方案更加合理的方案。但是,这是不是最好的方案呢?这就不好说了。这与研究者或者用户愿意如何平衡 TPR 和 FPR 的取值,以及不同方案带来的风险和收益有关。作为数据分析者,我们的任务是提供全面的分析与完整的报表。也就是说,我们要为用户提供在众多不同的 FPR 下的 TPR 取值(如图 4-3 所示)。在 *R* 中可以通过编程计算如下:

```
> ngrids=100
> TPR=rep(0,ngrids)
> FPR=rep(0,ngrids)
> for(i in 1:ngrids){
+ p0=i/ngrids;
+ ST.true=a2$ST
+ ST.pred=1*(p>p0)
+ TPR[i]=sum(ST.pred*ST.true)/sum(ST.true)
+ FPR[i]=sum(ST.pred*(1-ST.true))/sum(1-ST.true)
+ }
> plot(FPR,TPR,type="l",col=2)
> points(c(0,1),c(0,1),type="l",lty=2)
```

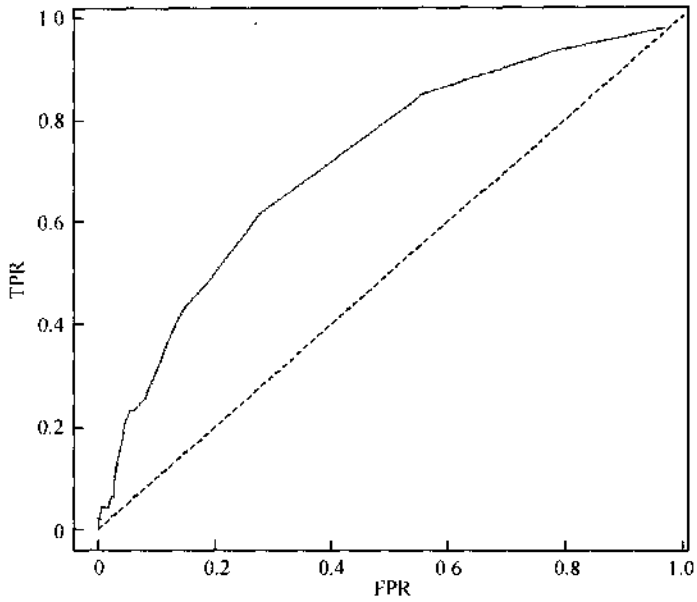


图 4-3 ROC 曲线图

有了这样一张图，我们就可以很简单也很直观地判断，如果我们要获得一定的 TPR 精度，我们需要牺牲多少 FPR 值。在统计文献中，我们通常称这条关于 TPR 和 FPR 的曲线为 ROC(receiver operating characteristics)曲线。值得注意的是，ROC 曲线同对角线(虚线)相比，永远是向上突起的。这说明 TPR 的取值必须高于 FPR 的取值，否则，这种预测方法是错误的。

下面我们就用这种方法对以下六种不同的预测模型的外样本预测精度予以比较。这六个模型分别为：

- logit 回归的全模型、logit 回归的 AIC 模型、logit 回归的 BIC 模型
- probit 回归的全模型、probit 回归的 AIC 模型、probit 回归的 BIC 模型

在 *R* 中通过编程具体实现如下(参见图 4-4)：

```
> p=matrix(0,length(a2[,1]),6)
> p[,1]=predict(glm1.a,a2)
> p[,2]=predict(logit.aic,a2)
> p[,3]=predict(logit.bic,a2)
> p[,c(1:3)]=exp(p[,c(1:3)])/(1+exp(p[,c(1:3)]))
>
> p[,4]=predict(glm1.b,a2)
> p[,5]=predict(probit.aic,a2)
> p[,6]=predict(probit.bic,a2)
> p[,c(4:6)]=pnorm(p[,c(4:6)])
>
> plot(c(0,1),c(0,1),type="l",main="FPR vs. TPR",xlab="FPR",ylab="TPR")
> FPR=rep(0,ngrids)
> TPR=rep(0,ngrids)
> for(k in 1:6){
+ prob=p[,k]
+ for(i in 1:ngrids){
+ p0=i/ngrids
+ ST.hat=1*(prob>p0)
+ FPR[i]=sum((1-ST.true)*ST.hat)/sum(1-ST.true)
+ TPR[i]=sum(ST.true*ST.hat)/sum(ST.true)
+ }
+ points(FPR,TPR,type="b",col=k,lty=k,pch=k)
+ }
> legend(0.6,0.5,c("LOGIT FULL MODEL","LOGIT AIC MODEL",
+ "LOGIT BIC MODEL","PROBIT FULL MODEL","PROBIT AIC MODEL",
+ "PROBIT BIC MODEL"),lty=c(1:6),col=c(1:6),pch=c(1:6))
```

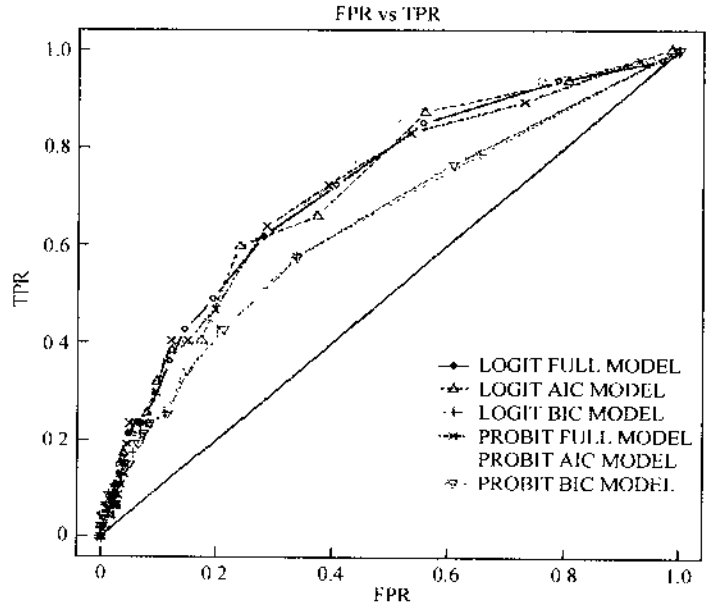


图 4-4　六种模型的 ROC 曲线图

从图 4-4 可以明显地看到，两个 BIC 模型的预测精度较差，这是因为它们的 ROC 曲线基本上都落在了其他 ROC 曲线的下方。而其他的 ROC 曲线非常相似，这说明它们的预测能力相当。因此，综合而言，两个 AIC 模型既有相对满意的预测精度，又有较为简单的模型形式，值得推荐。

第八节　简单分析报告

上市公司 ST 状态预测分析

内容提要　特殊处理(ST)政策是我国股票市场一项特有的、旨在保护投资者利益的政策，因而对公司未来 ST 状态的预测具有重要意义。本报告利用公开的财务数据对公司未来的 ST 状态进行预测分析。我们的分析结果表明有三个财务指标与 ST 状态相关，它们分别为应收账款与总资产的比例(ARA)、销售收入增长率(GROWTH)以及债务资产比率(LEV)，其中应收账款与总资产的比例(ARA)格外重要。基于本报告的分析结果和模型，投资者可以利用上市公司的公开财务指标对公司未来的 ST 状态进行合理的预测，进而进行科学的风险分析和投资；管理者可以更有效地监控公司的财务状态，通过合理的管理规划来降低公司被宣布 ST 的风险。

一、研究目的

特殊处理政策是我国股票市场一项特有的、旨在保护投资者利益的政策。根据相关规定，如果某上市公司出现财务状况或其他状况异常，以至于投资者难以判断公司前景，并且投资者权益可能受到损害时，中国证券监督管理委员会将考虑对该公司股票的交易实行特殊处理(special treatment，ST)。一旦某股票被特殊处理，那么该股票报价的日涨跌幅就会被人为地限制在 5% 以内。此外，如果该公司在下一个会计年度还不能取得正盈利，那么其股票将面临退市的风险。显然，股票的特殊处理可能给上市公司及其投资者带来巨大的经济损失，因此，如何利用公开的财务报表信息预测公司的 ST 状态(ST = 1：被特殊处理；ST = 0：没有被特殊处理)就成了人们关注的热点问题。本报告试图对中国股市的数据予以分析，找出能够合理预测上市公司未来 ST 状态的方法，并根据分析结果提出有意义的结论和建议。

二、数据来源和相关说明

我们需要有效地利用上市公司的历史财务数据，对其未来的 ST 状态，即本分析报告中的因变量，予以合理的预测。为此，我们选取了下列财务指标作为本次分析的解释性变量：

- 应收账款/总资产(ARA)：该财务指标反映了公司的盈利质量。
- 对数变换后的资产总计(ASSET)：简称资产总计，用来反映公司规模。
- 资产周转率(ATO)：该财务指标综合评价了企业全部资产的利用效率。
- 主营业务收入增长率(GROWTH)：该指标反映了公司的成长潜力。
- 债务资本比率(LEV)：该财务指标反映了公司的基本债务状况。
- 资产收益率(ROA)：该财务指标用于度量盈利能力。
- 最大股东的持股比例(SHARE)：该财务指标用于反映股权结构。

我们的数据共包含 1 430 个完整的观测。其中，684 个观测来自于 1999 年，即解释性变量来自于 1999 年，我们用这部分数据来建立模型。剩下的 746 个观测来自于 2000 年，我们用这部分数据检验我们的模型的预测效果。需要注意的是，我们的因变量 ST 反映的是公司在三年以后是否被宣布 ST。

三、描述性分析

为了获得对数据的整体概念，并考虑到因变量取离散值的特点，我们利用盒状图对训练数据进行简单的描述性分析，得到图 4-5。

从图 4-5 我们可以得到如下直观的印象：

- 被特殊处理的公司(ST = 1)的平均 ARA 值(以中位数计)要明显地高于没有被特殊处理的公司(ST = 0)。
- 公司规模(ASSET)同被特殊处理与否没有太强的关系。
- 公司的资产周转率(ATO)同被特殊处理与否没有明显的关系。
- 被特殊处理的公司的平均销售增长率(GROWTH，以中位数计)明显低于没有被特殊处理的公司。
- 被特殊处理的公司的负债水平(LEV，以中位数计)明显高于没有被特殊处理的公司。

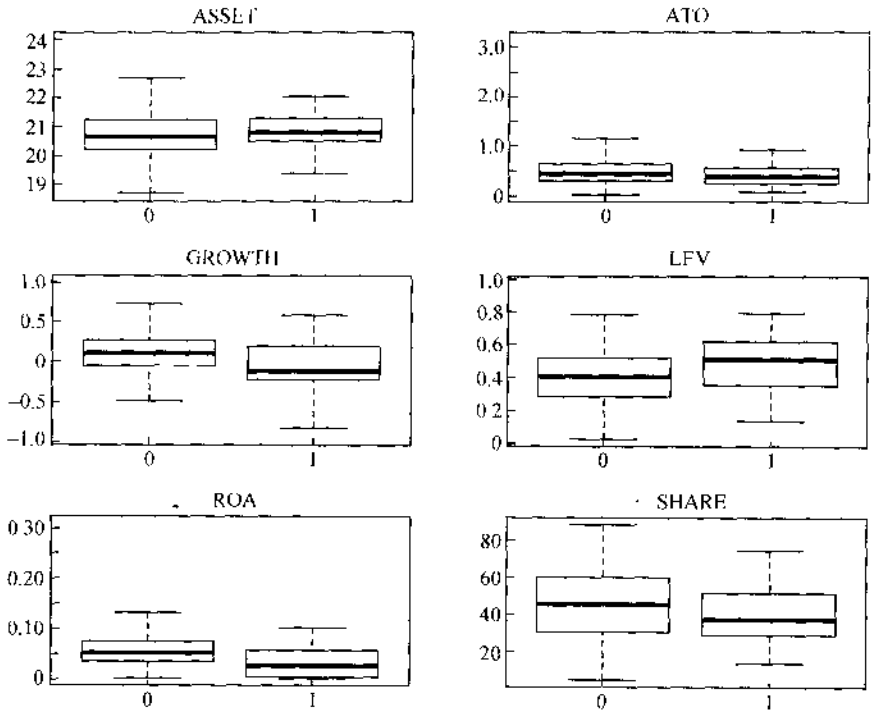


图 4-5 盒状图

- 被特殊处理的公司的盈利能力(ROA,以中位数计)明显低于没有被特殊处理的公司。
- 被特殊处理的公司的第一大股东持股比例(SHARE,以中位数计)明显低于没有被特殊处理的公司。

四、数据建模

1. 全模型分析

根据因变量取离散值的特点,我们采用 probit 回归和 logit 回归的方法建立模型。针对训练数据,我们首先用 logit 回归的方法对包含全部自变量的全模型进行估计,得到估计结果如表 4-1 所示。

表 4-1 全模型(logit 回归)

变量名	系数估计值	标准差	P 值
截距	-8.869	4.639	0.056
ARA	**4.880**	**1.492**	**0.001**
ASSET	0.246	0.224	0.271
ATO	-0.507	0.657	0.440
GROWTH	-0.8335	0.567	0.141
LEV	**2.354**	**1.201**	**0.050**
ROA	-0.637	6.224	0.919
SHARE	-0.011	0.011	0.319
模型 F 检验 P 值 <0.001			

从表 4-1 中我们可以看到,模型 F 检验的 P 值非常小(P 值 <0.001),表明该模型是显著的,即至少有一个财务指标对企业 ST 与否具有一定的预测能力。进一步通过对各个自变量对应的 Z 检验的 P 值的考察,我们可以得到以下重要结论:

- 应收账款与总资产的比例(ARA)和被 ST 与否高度相关,其 Z 检验在 0.01 的显著性水平下高度显著(P 值 =0.001)。并且可以看到,ARA 的系数估计值的符号为正,这说明应收账款占总资产的比例越高,被 ST 的可能性越大。
- 债务资产比率(LEV)和企业被 ST 与否高度相关,其 Z 检验在 0.10 的显著性水平下高度显著(P 值 =0.050)。同样可以看到,LEV 的系数估计值的符号为正,表明债务占总资产的比例越高,该企业被 ST 的可能性越大。
- 没有证据证明其他财务指标对预测 ST 与否有重要作用。

类似地,我们用 probit 回归分析训练数据,得到回归结果如表 4-2 所示。

表 4-2 全模型(probit 回归)

变量名	系数估计值	标准差	P 值
截距	-4.955	2.222	0.026
ARA	**2.647**	**0.787**	**0.001**
ASSET	0.135	0.107	0.207
ATO	-0.232	0.297	0.433
GROWTH	**-0.473**	**0.282**	**0.093**
LEV	**1.195**	**0.586**	**0.041**
ROA	0.670	2.788	0.810
SHARE	-0.005	0.005	0.340
模型 F 检验 P 值 <0.001			

从表 4-2 中我们可以发现，基本结论同 logit 回归非常相似。唯一不同的是，在 0.10 的显著性水平下，probit 回归还发现销售收入增长率（GROWTH）对该企业被 ST 与否有显著影响。由于 GROWTH 的系数估计值的符号为负，我们可以认为销售收入增长越慢的公司，被 ST 的可能性越大。

2. 模型选择

以上的分析结果告诉我们，我们所选取的财务指标确实对公司未来的 ST 状态有一定的预测能力。但是全模型过于复杂，而且其中还有部分变量是不显著的。为得到一个尽量简单同时又具有良好预测能力的模型，我们采用 AIC 和 BIC 的模型选择标准来选择一个最佳的模型。

对 logit 回归模型，我们用 AIC 方法选出的模型及其估计结果如表 4-3 所示。

表 4-3 AIC（logit 回归）

变量名	系数估计值	标准差	*P* 值
截距	-4.602	0.545	<0.001
ARA	5.130	1.434	<0.001
GROWTH	-0.906	0.547	0.098
LEV	2.550	1.078	0.012

对应于 BIC 方法选出的模型及其估计结果如表 4-4 所示。

表 4-4 BIC（logit 回归）

变量名	系数估计值	标准差	*P* 值
截距	-3.683	0.272	<0.001
ARA	6.331	1.313	<0.001

从表 4-3 及表 4-4 中我们可以发现，AIC 和 BIC 得出了不一样的结论。AIC 认为共有三个因素同 ST 与否相关，它们分别是应收账款与总资产的比例（ARA）、销售收入增长率（GROWTH）以及债务资产比率（LEV）。但是，BIC 认为只有应收账款与总资产的比例（ARA）是重要的。从一个较为保守的角度，我们可以得到如下结论：AIC 所选择的三个财务指标都很重要，而其中的应收账款与总资产的比例（ARA）格外重要。通过对 probit 回归方法作类似的模型选择，我们发现两种回归方法的模型选择结果完全一致。因此，我们认为上述的模型选择结果是可靠的。

3. 模型预测与评估

建立模型的一个重要目的就是预测，因此我们对前面分析所得到的六种不

同预测模型的预测精度予以比较。这六个模型分别为：

- logit 回归全模型、logit 回归的 AIC 模型、logit 回归的 BIC 模型
- probit 回归全模型、probit 回归的 AIC 模型、probit 回归的 BIC 模型

具体来说，我们利用单独的检验数据来对公司未来的 ST 状态进行预测，并用对应的预测结果来衡量模型的预测精度。对此类离散变量的预测，我们一般用以下两个指标间接地度量预测精度：

- True Positive Rate(TPR)：把真实的 ST 公司正确地预测为 ST = 1 的概率。
- False Positive Rate(FPR)：把真实的非 ST 公司错误地预测为 ST = 1 的概率。

对于 logit 回归分析，我们需要先给定一个阈值才能进行预测。而不同的阈值所对应的预测结果和精度也是不一样的。另外，由于高 TPR 值总是对应着高 FPR 值，我们只能根据各自的风险偏好来选择合适的 TPR 和 FPR 的取值。因此，我们用 ROC 曲线来进行全面的度量，如图 4-6 所示。

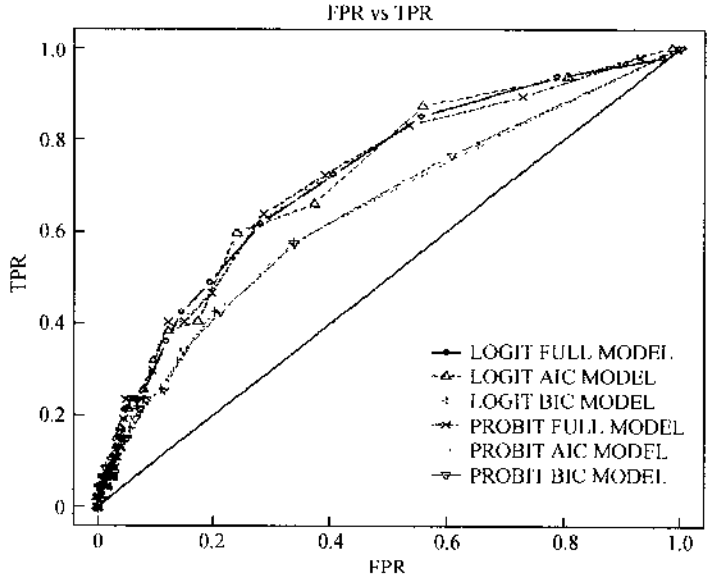


图 4-6　六种模型的 ROC 曲线图

图 4-6 中的 ROC 曲线与对角线相比，永远是向上突起的。这表明我们的预测模型是有效的。从图中可以明显地看到，两个 BIC 模型的预测精度较差，这是因为它们的 ROC 曲线基本上都落在了其他 ROC 曲线的下方。而其他的 ROC 曲线非常相似，这说明它们的预测能力相当。因此，综合而言，两个 AIC 模

型既有相对满意的预测精度，又有较为简单的模型形式，是最为理想的模型。

五、结论及建议

从上述分析结果可知，上市公司的财务信息确实对公司未来的 ST 状态有一定的预测能力。具体来说，从保守、谨慎的角度来看，应收账款与总资产的比例（ARA）、销售收入增长率（GROWTH）以及债务资产比率（LEV）这三个财务指标，尤其是第一个财务指标，对于预测公司未来的 ST 状态非常重要。

进一步地，应收账款占总资产的比例越高，债务占总资产的比例越高，销售收入增长越慢的公司被 ST 的可能性越大。这就告诉公司的管理者，对这三个财务指标要格外地关注，尤其是应收账款与总资产的比例（ARA）。通过对这些财务指标的合理控制，能够有效地降低公司被宣布 ST 的风险。另外，投资者和管理者也可以利用本分析报告所得到的模型，结合公司公开的财务指标对其未来的 ST 状态进行合理的预测，进而进行科学的投资和管理规划。

[讨论总结]

本章以股票 ST 预测为例，系统演示并讲解了 logit 以及 probit 两种 0-1 回归模型。通过对本章的学习，读者应该能够了解：什么时候可以使用 0-1 回归模型，以及如何使用。在 R 语言学习方面，读者应该掌握并区分用于分析线性模型以及广义线性模型的命令。在统计理论方面，读者应该掌握以下概念：0-1 变量、logit 回归、probit 回归、极大似然估计等。对相关统计学理论渴望深入了解的读者可以参阅 McCullagh and Nelder(1999)。

附录　程序及注释

```
a=read.csv("D:/Practical Business Data Analysis/case/CH4/st.csv",header=T)
                                                    # 读入 csv 格式的数据,并赋值给 a
a1=a[a$year==1999,-1]                               # 取出 year 为 1999 的数据,并删除第一列,赋值给 a1
a2=a[a$year==2000,-1]                               # 取出 year 为 2000 的数据,并删除第一列,赋值给 a2
a1[c(1:5),]                                         # 展示 a1 的前 5 行数据
boxplot(ARA ~ ST,data=a1,main="ARA")                # 画出 ARA 与 ST 的盒状图
par(mfrow=c(3,2))                                   # 设置画图模式为 3x2 的格式
boxplot(ASSET ~ ST,data=a1,main="ASSET")            # 画出 ASSET 与 ST 的盒状图
boxplot(ATO ~ ST,data=a1,main="ATO")                # 画出 ATO 与 ST 的盒状图
boxplot(GROWTH ~ ST,data=a1,main="GROWTH")          # 画出 GROWTH 与 ST 的盒状图
boxplot(LEV ~ ST,data=a1,main="LEV")                # 画出 LEV 与 ST 的盒状图
boxplot(ROA ~ ST,data=a1,main="ROA")                # 画出 ROA 与 ST 的盒状图
boxplot(SHARE ~ ST,data=a1,main="SHARE")            # 画出 SHARE 与 ST 的盒状图
par(mfrow=c(1,1))                                   # 设置画图模式为 1x1 的格式
glm0.a=glm(ST ~ 1,family=binomial(link=logit),data=a1)   # 拟合 logit 回归,不使用任何变量的空模型
glm1.a=glm(ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,family=binomial(link=logit),data=a1)
                                                    # logit 回归全模型
anova(glm0.a,glm1.a)                                # 计算 glm0.a 与 glm1.a 的 deviance
1 - pchisq(30.565,7)                                # 计算模型显著性检验的 P 值
glm0.b=glm(ST ~ 1,family=binomial(link=probit),data=a1)  # 拟合 porbit 回归,不使用任何变量的空模型
glm1.b=glm(ST ~ ARA + ASSET + ATO + GROWTH + LEV + ROA + SHARE,family=binomial(link=probit),data=a1)
```

```
                                                                    # porbit 回归全模型
anova(glm0.b,glm1.b)                                                # 计算 glm0.b 与 glm1.b 的 deviance
1 - pchisq(31.702,7)                                                # 计算模型显著性检验的 P 值
library(car)                                                        # 载入程序包 car
Anova(glm1.a,type="III")                                            # 对模型 glm1.a 作三型方差分析
summary(glm1.a)                                                     # 显示模型 glm1.a 的各方面细节,包括参数估计值、P 值等
Anova(glm1.b,type="III")                                            # 对模型 glm1.b 作三型方差分析
summary(glm1.b)                                                     # 显示模型 glm1.b 的各方面细节,包括参数估计值、P 值等
glm2.b=glm(ST ~ ARA + GROWTH + LEV,family=binomial(link=probit),data=a1)
                                                                    # 拟合简化后的 probit 回归模型
summary(glm2.b)                                                     # 显示模型 glm2.b 的各方面细节,包括参数估计值、P 值等
glm2.a=glm(ST ~ ARA + GROWTH + LEV,family=binomial(link=logit),data=a1)
                                                                    # 拟合简化后的 logit 回归模型
summary(glm2.a)                                                     # 显示模型 glm2.a 的各方面细节,包括参数估计值、P 值等
deviance(glm0.a)                                                    # 计算模型 glm0.a 的 deviance
deviance(glm1.a)                                                    # 计算模型 glm1.a 的 deviance
AIC(glm0.a)                                                         # 计算模型 glm0.a 的 AIC 取值
AIC(glm1.a)                                                         # 计算模型 glm1.a 的 AIC 取值
AIC(glm0.a,k=log(length(a1[,1])))                                   # 计算模型 glm0.a 的 BIC 取值
AIC(glm1.a,k=log(length(a1[,1])))                                   # 计算模型 glm1.a 的 BIC 取值
logit.aic=step(glm1.a,trace=0)                                      # 根据 AIC 准则选择最优模型,并赋值给 logit.aic
summary(logit.aic)                                                  # 显示模型 logit.aic 的各方面细节,包括参数估计值、P 值等
n=length(a1[,1])                                                    # 样本大小
logit.bic=step(glm1.a,k=log(n),trace=0)                             # 根据 BIC 准则选择最优模型,并赋值给 logit.bic
summary(logit.bic)                                                  # 显示模型 logit.bic 的各方面细节,包括参数估计值、P 值等
probit.aic=step(glm1.b,trace=0)                                     # 根据 AIC 准则选择最优模型,并赋值给 probit.aic
```

```
summary(probit.aic)
probit.bic=step(glm1.b,k=log(n),trace=0)
summary(probit.bic)
summary(glm1.a)
p=predict(glm1.a,a2)
p=exp(p)/(1+exp(p))
a2$ST.pred=1*(p>0.5)
table(a2[,c(8,9)])
a2$ST.pred=1*(p>0)
table(a2[,c(8,9)])
a2$ST.pred=1*(p>0.05)
table(a2[,c(8,9)])
ngrids=100
TPR=rep(0,ngrids)
FPR=rep(0,ngrids)
for(i in 1:ngrids){
  p0=i/ngrids;
  ST.true=a2$ST
  ST.pred=1*(p>p0)
  TPR[i]=sum(ST.pred*ST.true)/sum(ST.true)
  FPR[i]=sum(ST.pred*(1-ST.true))/sum(1-ST.true)
}
plot(FPR,TPR,type="l",col=2)
points(c(0,1),c(0,1),type="l",lty=2)
p=matrix(0,length(a2[,1]),6)
p[,1]=predict(glm1.a,a2)
```

```
# 显示模型 probit.aic 的各方面细节,包括参数估计值、P 值等
# 根据 BIC 准则选择最优模型,并赋值给 probit.bic
# 显示模型 probit.bic 的各方面细节,包括参数估计值、P 值等
# 显示模型 glm1.a 的各方面细节,包括参数估计值、P 值等
# 利用模型 glm1.a 对数据 a2 进行预测
# 计算预测得到的概率
# 以 0.5 为阈值生成预测值
# 计算预测值与真实值的 2 维频数表
# 以 0 为阈值生成预测值
# 计算预测值与真实值的 2 维频数表
# 以 0.05 为阈值生成预测值
# 计算预测值与真实值的 2 维频数表
# 设置格点数为 100
# 为 TPR(true positive ratio)赋初值
# 为 FPR(false positive ratio)赋初值

# 选取阈值 p0
# 从 a2 中取出真实值并赋值给 ST.true
# 以 0.05 为阈值生成预测值
# 计算 TPR
# 计算 FPR

# 画出 FPR 与 TPR 的散点图,即 ROC 曲线
# 添加对角线
# 生成矩阵,用于存储各模型的预测值
# 利用模型 glm1.a 对数据 a2 进行预测
```

```
p[,2]=predict(logit.aic,a2)                                    # 利用模型 logit.aic 对数据 a2 进行预测
p[,3]=predict(logit.bic,a2)                                    # 利用模型 logit.bic 对数据 a2 进行预测
p[,c(1:3)]=exp(p[,c(1:3)])/(1+exp(p[,c(1:3)]))                 # 计算预测得到的概率
p[,4]=predict(glm1.b,a2)                                       # 利用模型 glm1.b 对数据 a2 进行预测
p[,5]=predict(probit.aic,a2)                                   # 利用模型 probit.aic 对数据 a2 进行预测
p[,6]=predict(probit.bic,a2)                                   # 利用模型 probit.bic 对数据 a2 进行预测
p[,c(4:6)]=pnorm(p[,c(4:6)])                                   # 计算预测得到的概率
plot(c(0,1),c(0,1),type="l",main="FPR vs. TPR",xlab="FPR",ylab="TPR")
                                                               # 画图,生成基本框架

FPR=rep(0,ngrids)                                              # 为 FPR 赋初值
TPR=rep(0,ngrids)                                              # 为 TPR 赋初值
for(k in 1:6){
  prob=p[,k]                                                   # 取出 p 中第 k 列的值,即第 k 个模型的预测概率
  for(i in 1:ngrids){
    p0=i/ngrids                                                # 选取阈值
    ST.hat=1*(prob>p0)                                         # 根据阈值生成预测值
    FPR[i]=sum((1-ST.true)*ST.hat)/sum(1-ST.true)              # 计算 FPR
    TPR[i]=sum(ST.true*ST.hat)/sum(ST.true)                    # 计算 TPR
  }
  points(FPR,TPR,type="b",col=k,lty=k,pch=k)                   # 向图上添加第 k 个模型的 TPR 与 FPR 的散点图
}
legend(0.6,0.5,c("LOGIT FULL MODEL","LOGIT AIC MODEL",
"LOGIT BIC MODEL","PROBIT FULL MODEL","PROBIT AIC MODEL",
"PROBIT BIC MODEL"),lty=c(1:6),col=c(1:6),pch=c(1:6))
                                                               # 为 6 个模型添加标示,区分 6 个模型
```

第五章
定序回归

- 案例介绍
- 描述性分析
- 定序回归模型
- 参数估计与统计推断
- 多变量逻辑回归
- 模型选择
- 预测与评估
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过一个研究消费者偏好的实际案例，详细介绍 logit 以及 probit 回归这两种重要的定序回归模型。它们主要处理的是因变量为定序数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用 logit 或者 probit 定序回归；(2) 定序回归分析的基本统计学理论；(3) 相关理论在统计学软件 *R* 中的应用；(4) 相应的统计分析报告的撰写。本章所涉及的新统计学概念如下：定序数据、logit 定序回归、probit 定序回归。

第一节　案例介绍

在这一章中，我们将介绍另外一种非连续型变量的回归方法。这种新的数据类型就是定序数据(ordinal data)。什么是定序数据？很多读者都会有这样的经历：一位市场调查者拿着问卷和礼品要你回答他的问题，例如：你有多么喜欢农夫山泉？有三个答案：① 不喜欢；② 无所谓；③ 喜欢。在数据分析中，这三个答案往往被简单地记录为：1、2、3。现在的问题是：1、2、3 代表了什么含义？我们作一个简单的比较：

- 情形 1：1 = 1 岁；2 = 2 岁；3 = 3 岁
- 情形 2：1 = 红色；2 = 黄色；3 = 蓝色
- 情形 3：1 = 不喜欢；2 = 无所谓；3 = 喜欢

请注意以上三种不同的情形，其数据的代码都是 1、2、3。第一种情形下，代码代表了小孩的年龄(1 岁、2 岁、3 岁)。请注意，这里的 1、2、3 是有具体的数值意义的。什么叫数值意义？一个简单的判断标准就是可以对其作代数运算。例如，2 岁 − 1 岁 = 1 岁，这说明 2 岁的孩子比 1 岁的孩子大 1 岁；而 3 岁 − 2 岁 = 1 岁，这说明 3 岁孩子和 2 岁孩子的年龄差异等于 2 岁孩子和 1 岁孩子的年龄差异。这是我们的代码 1、2、3 所代表的第一种数据类型。那么，我们可不可以对情形 2 中的 1、2、3 作代数运算呢？当然不可以，否则我们会有：

$$3 - 2 = \text{蓝色} - \text{黄色} = 2 - 1 = \text{黄色} - \text{红色}$$

所以第二种数据类型和第一种很不一样，尽管在数据库中它们很可能都被记录为 1、2、3。因此，情形 2 中的数据类型是没有任何数值意义的。类似地，我们可以知道情形 3 中的数据类型(本章所关心的数据类型)也没有任何数值意义。

因此，我们不能对它作代数运算，否则我们会有：

3 - 2 = 喜欢 - 无所谓 = 2 - 1 = 无所谓 - 不喜欢

这个等式不一定成立。例如，对于绝大多数消费者来说，他们不大可能说："我不喜欢农夫山泉。"但是他们很可能觉得无所谓或者喜欢。因此，喜欢和无所谓之间所表现出来的差异应该远远小于无所谓和不喜欢之间的差异。因此，正确表达式应为：

3 - 2 = 喜欢 - 无所谓 ≠ 2 - 1 = 无所谓 - 不喜欢

那么，情形 2 和情形 3 的数据类型又有什么不一样呢？对于情形 2 中的数据，我们完全可以打乱顺序而不会造成任何混乱。例如，我们可以重新定义：2 = 红色；3 = 黄色；1 = 蓝色。那么，我们可不可以对情形 3 作同样的重新定义呢？例如，3 = 不喜欢；1 = 无所谓；2 = 喜欢。这显然是不可以的，因为这会造成数据分析的混乱。但是，我们完全可以定义：-1 = 不喜欢；2.5 = 无所谓；3.2 = 喜欢。这是没有问题的，因为数据的顺序得到了保护。当然，为了方便起见，我们从来不会这样定义。由此可见，情形 3 中的数据有两个特征：① 没有数值意义；② 有顺序意义。因此，我们将其称为定序数据。而本章的重点就是要讨论如果因变量是定序数据，我们应该如何分析。由于在市场调研中存在着大量的定序数据，因此我们用一个关于手机的市场调研的实际案例来详细讲解定序回归（ordinal regression）的方法。

目前国内手机销售市场竞争日趋激烈，为了确立在市场中的相对优势地位，开发新功能是企业常常采用的手段。因为与开发新产品相比，在现有产品的基础上增加新功能不失为一种既快速又有效的方式。当企业决定为现有产品增加新功能时，往往面对众多选择。例如，下一个升级换代产品是增加数码相机功能还是增加收音机或者 MP3 播放器功能？根据消费者的行为特征，消费者在进行购买决策时愿意支付的价格主要依赖于对产品的偏好，而不是企业的生产成本。因此，研究消费者对新功能的偏好就有着特别重要的意义。在过去对手机功能偏好的研究中，往往假设增加某一新功能所带来的影响同该手机的其他现有功能无关，这显然是一个值得怀疑的假设。常识告诉我们，同样的功能在低端手机和高端手机上的结果是不一样的，甚至可能有巨大差异。因此，在我们的案例中，我们通过对手机的某些功能在不同的功能组合中所起的作用来进行更为精确的分析。

我们的数据来源是对北京大学光华管理学院的 MBA 学生和高级经理培训班的学员的调查。首先，我们选取商务手机厂商较多考虑的六个功能，然后加上品牌（这里我们只涉及诺基亚、摩托罗拉、三星和波导四个品牌）共七个要素，构成我们要研究的影响消费者偏好的要素，包括手机品牌、数码相机功能、收看

电视功能、手写笔、电话本多条记录、MP3 和游戏数目。我们将这七个要素按不同方式组成 12 个产品组合，然后再针对不同的组合进行偏好调查。对每个组合，调查对象根据其偏好程度用 5 分量表进行打分（1 = 根本不喜欢；2 = 比较不喜欢；3 = 一般喜欢；4 = 比较喜欢；5 = 非常喜欢）。而我们关心的问题是哪些因素在影响着消费者对手机的偏好程度以及影响程度如何？由此可见，本案例所关心的因变量（偏好程度）是一个典型的定序数据。基于我们的调查结果并作适当的数据清理，最后共获得来自 148 个调查对象的 1 451 个有效观测值。为了演示的方便，我们将本案例中所涉及的变量名罗列如表 5-1 所示。

表 5-1　变量说明

变量类型	变量含义	变量名	变量水平
因变量	对该产品的偏好程度	score	1 = 根本不喜欢；2 = 比较不喜欢；3 = 一般喜欢；4 = 比较喜欢；5 = 非常喜欢
自变量	手机品牌	W1	共四种（诺基亚、摩托罗拉、三星和波导）
	有无数码相机	W2	共两种（有、无）
	能否收看电视	W3	共两种（能、不能）
	有无手写笔	W4	共两种（有、无）
	电话本能否多条记录	W5	共两种（能、不能）
	有无 MP3	W6	共两种（有、无）
	游戏数目	W7	连续型

由此可见，本案例共涉及六个离散型协变量和一个连续型协变量。根据这七个自变量的不同取值，我们构造 12 种不同的功能组合，如表 5-2 所示。

表 5-2　手机功能组合

品牌	数码相机	能否收看电视	手写笔	电话本能否多条记录	MP3	游戏数目
诺基亚	无	不能	无	能	有	3
	有	不能	有	不能	有	5
	无	能	有	不能	无	7
波导	有	能	无	能	无	3
	无	不能	无	不能	有	5
	有	不能	有	能	有	7
摩托罗拉	无	能	有	能	无	3
	有	能	无	不能	无	5
	无	不能	无	能	无	7
三星	有	不能	有	不能	无	3
	无	能	有	不能	有	5
	有	能	无	能	有	7

从第一张表中可以看到，因变量是离散型的变量，而且是定序变量。而从第二张表中可以看到，自变量是各种组合中所涉及的七个因素，既有连续型数据，也有离散型数据。

这一章将上一章的离散型因变量的变量水平数由两个扩展为多个，这时对因变量的分析和处理以及对自变量的回归过程也将更加复杂。因为这里不仅仅是分类的问题，而且还涉及顺序的问题。对定序变量的回归也是广义线性回归的一种，我们也可以像上一章那样将非线性的概率问题最终转化为线性问题来加以解决。接下来，我们将在上一章的基础上继续介绍因变量为定序变量的 logit 回归。

第二节　描述性分析

按照惯例，我们首先尝试对数据予以描述性分析，以获得对数据的初步认识，形成待检验的结论，并指导下一步的数据分析。我们首先读入数据并展示如下：

```
> rm(list=ls())
> a=read.csv("D:/Practical Business Data Analysis/case/CH5/cellphone.csv")
> attach(a)
> a[c(1:5),]
  score    W1 W2 W3 W4 W5 W6 W7
1     3 Nokia  0  0  0  1  1  3
2     4 Nokia  1  0  1  0  1  5
3     4 Nokia  0  1  1  0  0  7
4     4  Bird  1  1  0  1  0  3
5     3  Bird  0  0  0  0  1  5
```

从上面数据的第一行，我们可以看到，一位被调查者对某一款手机产品作了评估。该手机品牌为诺基亚，无数码相机，不能收看电视，没有手写笔，但是其电话本支持多条记录，有 MP3 功能，游戏的数目为 3。对于这样一款手机，该调查者对其一般喜欢（score = 3）。从上面数据的第二行，我们可以看到，另外一位被调查者对另外一款手机产品也作了评估。该手机品牌还是诺基亚，有数码相机，不能收看电视，有手写笔，其电话本不支持多条记录，有 MP3 功能，其游戏的数目为 5。对于这样一款手机，该调查者对其比较喜欢（score = 4）。

我们先简单描述一下消费者打分（score）和不同品牌之间的关系。在 *R* 语言中，我们可以做列联表（contingency table）如下：

```
> xtabs(~score+W1)
     W1
score Bird Motorola Nokia Samsung
    1   37       24    34      26
    2   80       64    53      66
    3   98      138   132     133
    4  109      108   116      96
    5   28       30    35      44
```

从中我们可以看到，在所有得分为 1 或者 2 的品牌中，频数最高的是波导（Bird），其频数分别为 37 和 80；在得分为 3 的品牌中，摩托罗拉（Motorola）频数最高，为 138；在得分为 4 的品牌中，诺基亚（Nokia）频数最高，为 116；在得分为 5 的品牌中，三星（Samsung）频数最高，为 44。由此可见，国产品牌在与国际品牌的对比中具有较大的劣势，而在国际品牌中，摩托罗拉偏向大众化路线，而诺基亚偏向中高端市场。下面我们再对消费者打分（score）和手机有无数码相机（W1）功能之间的关系予以分析，如图 5-1 所示。

```
> plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Digital Camera")
> points(c(1:5),tapply(W2,score,mean),type="b")
```

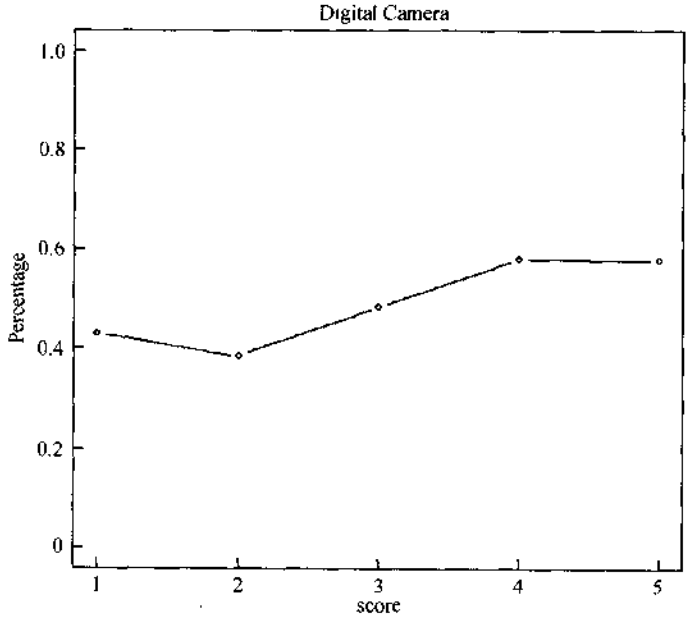


图 5-1　消费者打分和有无数码相机之间关系图

从总体上来讲，我们可以看到一个明显的上升趋势。具体地说，得分越高的手机，具有数码相机功能的比率越高，特别是在比较不喜欢（score = 2）到比较喜欢（score = 4）之间。这从一个侧面说明，有无数码相机功能在当时是一个界定人们对其打分是否高于平均水平的重要属性。我们再对其他几个定性因素

作类似的分析(如图 5-2 所示):

```
  par(mfrow=c(2,2))
> plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Television")
> points(tapply(W3,score,mean),type="b")
> plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Hand Written Pad")
> points(tapply(W4,score,mean),type="b")
  plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Multiple Entry Phonebook")
  points(tapply(W5,score,mean),type="b")
> plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="MP3")
> points(tapply(W6,score,mean),type="b")
> par(mfrow=c(1,1))
```

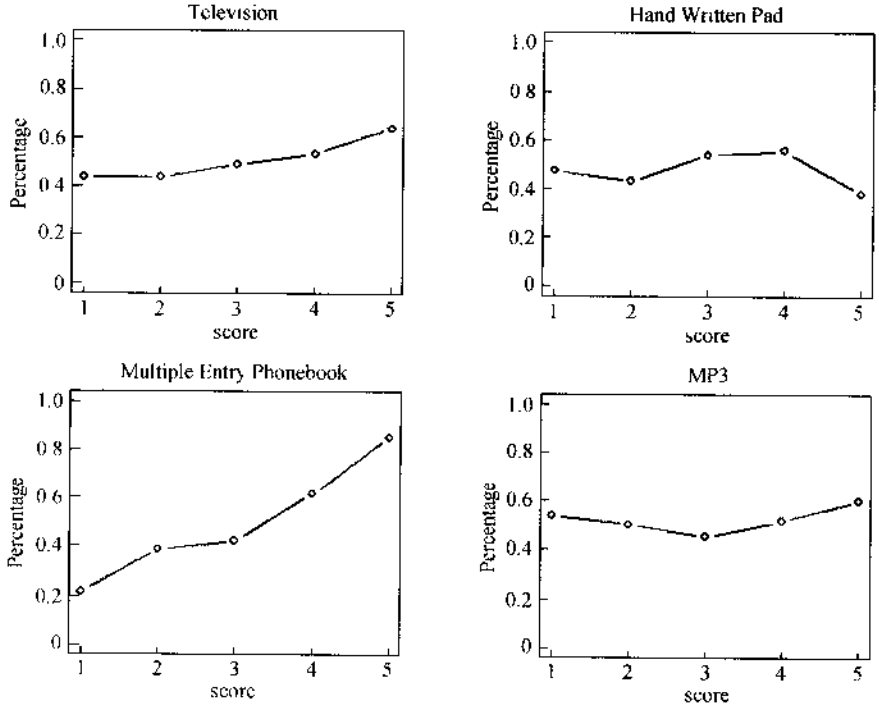


图 5-2　消费者打分和其他功能之间关系图

大家可以看到,能否收看电视(*W3*)以及电话本能否支持多条记录(*W5*)同消费者打分(score)高度正相关。而有无手写笔(*W4*)和能否支持 MP3 功能(*W6*)在我们的样本中似乎并没有受到很大的青睐。最后,我们再简单地描述一下消费者打分(score)和游戏的个数(*W7*)之间的相互关系。我们可以考虑做列联表如下:

```
> xtabs(~score+W7)
      W7
score   3   5   7
    1  25  67  29
    2  71 113  79
    3 169 187 145
    4 155 111 163
    5  64   9  64
```

从中很难看出明显的趋势。虽然我们不能就此断定游戏的个数同人们对手机的打分没有关系，但是我们可以猜测即使相关，那么关系也不会很大。

第三节　定序回归模型

在我们详细介绍定序回归模型以前，我们首先需要回答一个问题，那就是：为什么不可以用普通线性模型，例如（如果只考虑游戏的个数）：

$$\text{score} = \beta_0 + \beta_7 \times W7 + \varepsilon$$

请注意，这显然是不可以的。因为等号的右边是一个具有数值意义的实数（$\beta_0 + \beta_7 \times W7 + \varepsilon$），而等号的左边是一个只有顺序意义的变量（score）。所以，它们当然不可能被放在同一个等式中。那么，我们可不可以用 0-1 变量回归模型呢？很遗憾，也不可以。因为 0-1 变量模型要求因变量有且仅有两个不同的取值（0-1），但是本案例所涉及的因变量有五个不同的取值（1—5）。因此，0-1 变量回归方法也不适用。所以，我们必须建立一个特殊的专门用于定序数据的回归模型，这就是下面要讲的定序回归模型。

定序回归模型是怎么产生的呢？设想一下，如果我们是本案例的被调查者，我们会如何打分呢？我们会首先评价某产品组合，获得初步印象，形成观点，再给出具体得分。我们的这个心理过程可以被简单地想象成这样一个过程：首先根据该产品的特征，形成一个对于该产品的“喜好程度”（preference），并记为 Z。可以想象，如果有两款手机特征非常相似，那么我们对它们的喜好程度也应该非常相似。这说明 Z 是连续的，而且我们还可以假设它的取值范围任意。但遗憾的是，喜好程度是一个看不见、模不着的隐变量（latent variable）。那么，这个隐含的喜好程度是如何形成消费者打分（score）的呢？我们再假设在人们的心理活动中有一定的判断标准，或者叫做阈值（记为 c_k）。当隐含的喜好程度落在某两个相邻的阈值之间时，我们就会给出一定的消费者打分。具体地说：

$$\text{score} = \begin{cases} 1 & \text{如果 } Z < c_1 \\ 2 & \text{如果 } c_1 \leqslant Z < c_2 \\ 3 & \text{如果 } c_2 \leqslant Z < c_3 \\ 4 & \text{如果 } c_3 \leqslant Z < c_4 \\ 5 & \text{如果 } c_4 \leqslant Z \end{cases}$$

下面再考虑，解释性变量（$W1-W7$）是如何影响消费者打分（score）的呢？我们可以非常合理地假设，所有解释性变量都是通过影响喜好程度（Z）来影响消费者打分（score）的。请注意，Z 是一个取值任意的连续型变量，因此我们完全可以用普通线性回归模型来刻画 Z 和解释性变量（如 $W7$）之间的关系，即：

$$Z = \beta_0 + \beta_7 \times W7 + \varepsilon$$

如果上帝告诉我们 Z 的具体取值，那么第一章中详细介绍的建立普通线性回归模型的方法就可以直接用来建立上面这个模型。但是，现实中我们并不知道 Z 的具体取值。那么，上面这个线性模型对我们有什么用处呢？根据这个模型以及游戏数目（$W7$），我们可以判断消费者打分不超过 $k(1 \leqslant k \leqslant 4)$ 的可能性为：

$$\begin{aligned} P(\text{score} \leqslant k) &= P(z \leqslant c_k) = P(\beta_0 + \beta_7 \times W7 + \varepsilon \leqslant c_k) \\ &= P\{\varepsilon \leqslant (c_k - \beta_0) - \beta_7 \times W7\} \\ &= F_\varepsilon(\alpha_k - \beta_7 \times W7) \end{aligned}$$

其中，c_k 就是前面提到的阈值，$\alpha_k = c_k - \beta_0$，而 $F_\varepsilon(t) = P(\varepsilon < t)$ 是 ε 的分布函数。如果我们可以对 $F_\varepsilon(t)$ 的具体函数形式予以合理的假设（即假设 ε 的分布），那么我们就获得了一个关于定序变量的回归模型，即：

$$P(\text{score} \leqslant k) = F_\varepsilon(\alpha_k - \beta_7 \times W7)$$

请注意，这个新的模型设定中没有任何地方涉及那个看不见、摸不着的消费者喜好程度 Z，因此可以进行数据估计，其具体的估计方法我们将在下一节中详细讨论。在这里，我们首先需要注意到该模型形式同 0-1 变量回归模型形式的异同。由于定序数据可能的取值个数大于 2（如本案例为 5），因此我们有好几个不同的截距项 α_k。以本案例为例，我们的隐变量总共有五种不同的取值可能（score = 1,2,3,4,5），因此，我们总共有四个不同的截距项 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$。而且它们之间也有顺序：$\alpha_1 \leqslant \alpha_2 \leqslant \alpha_3 \leqslant \alpha_4$。但是，我们的斜率却只有一个，即 β_7。对于很多实际问题，我们最关心的是解释性变量和因变量之间的关系（即 β_7），而对截距项（$\alpha_1, \alpha_2, \alpha_3, \alpha_4$）的兴趣不大。下面，我们再回答一个问题，那就是：$F_\varepsilon(t)$ 的具体函数形式应该如何假设才合理？同 0-1 变量回归一样，有两个

“方便”的选择，那就是标准正态分布与逻辑分布。它们分别对应于 probit 定序回归和 logit 定序回归。如果我们假设 $F_\varepsilon(t)$ 是标准正态分布函数，那么模型形式为：

$$P(\text{score} \leqslant k) = \Phi(\alpha_k - \beta_7 \times W7)$$

相应地，如果我们假设 $F_\varepsilon(t)$ 是逻辑分布函数，那么模型形式为：

$$P(\text{score} \leqslant k) = \frac{\exp(\alpha_k - \beta_7 \times W7)}{1 + \exp(\alpha_k - \beta_7 \times W7)}$$

这两种模型还可以表述为：

$$\Phi^{-1}\{P(\text{score} \leqslant k)\} = \alpha_k - \beta_7 \times W7$$

$$\text{logit}\{P(\text{score} \leqslant k)\} = \alpha_k - \beta_7 \times W7$$

同 0-1 变量回归问题一样，现实中到底 probit 定序回归好还是 logit 定序回归好，至今还没有定论。但是，可以肯定的是，它们都是非常有用的统计方法，而且结果往往极其相似。

第四节　参数估计与统计推断

下面我们回答几个重要的理论问题，那就是：对于定序变量回归模型，我们应该如何作参数估计？如何作统计推断？由于 probit 定序回归和 logit 定序回归的估计方法以及推断方法非常类似，因此，为简单起见，我们只对 probit 定序回归作详细的讨论。如果上帝告诉我们那个看不见、摸不着的喜好程度 Z，那么我们就可以使用第一章中介绍的最小二乘法来估计参数的取值，并进而作统计推断。但是，现实中 Z 的取值是未知的，因此我们必须考虑使用其他方法来估计我们感兴趣的参数（即 α_k 和 β_7）。为了严格起见，假设 $(\text{score}_i, W7_i)$ 是来自于第 i 个（$i = 1, 2, \cdots, n$）样本的观测。那么，该样本取值为 k 的概率为：

$$P(\text{score} = k) = f_k(W7_i)$$

$$= \begin{cases} \Phi(\alpha_1 - \beta_7 \times W7_i) & \text{如果 } k = 1 \\ \Phi(\alpha_k - \beta_7 \times W7_i) - \Phi(\alpha_{k-1} - \beta_7 \times W7_i) & \text{如果 } 1 < k < 5 \\ 1 - \Phi(\alpha_4 - \beta_7 \times W7_i) & \text{如果 } k = 5 \end{cases}$$

由此，我们可以计算样本的似然函数为：

$$L(\beta_0, \beta_7) = \prod_{i=1}^{n} \prod_{k=1}^{5} \{f_k(W7_i)\}^{I(\text{score}_i = k)}$$

其中，$\beta_0 = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$，$I$ 代表示性函数。理论上讲，合理的参数估计应该能够产生较大的似然函数 $L(\beta_0, \beta_7)$ 取值。因此，我们可以通过极大化 $L(\beta_0, \beta_7)$

或者以下对数似然函数：

$$\log L(\beta_0,\beta_7) = \sum_{i=1}^{n}\sum_{k=1}^{5} I\{\text{score}_i = k\} \times \log\{f_k(W7_i)\}$$

来获得参数的估计值。我们称此估计为极大似然估计，并记为$(\hat{\beta}_0,\hat{\beta}_7)$。同0-1变量回归模型类似，定序回归模型中$(\hat{\beta}_0,\hat{\beta}_7)$的具体分布是不知道的。但是，我们知道，如果样本量足够大，根据中心极限定理，$(\hat{\beta}_0,\hat{\beta}_7)$将近似地服从正态分布。具体如下：

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma(\hat{\beta}_j)} \sim N(0,1) \quad (j = 0,7)$$

这说明，只要我们能够准确地估计$\hat{\beta}_j$的标准差$\sqrt{\text{var}(\hat{\beta}_j)} = \sigma(\hat{\beta}_j)$，记为$\hat{\sigma}(\hat{\beta}_j)$，那么我们就可以构造如下的检验统计量：

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

在原假设$\beta_j = 0$成立的情况下，该统计量T_j近似地服从标准正态分布。因此，对于一个给定的显著性水平（如0.05），我们可以根据T_j的绝对值是否大于$z_{0.975}$来决定是否拒绝原假设。上面所介绍的检验方法叫做Z检验或者t检验。这两种检验方法的缺点是只能对一个因素（即$W7$）的显著性作检验，而不能够同时检验多个因素的显著性。例如，$W7$代表好几个解释性变量，因此β_7是一个向量。那么，我们应该如何检验β_7的各个分量是否同时为零呢？为了解决这个问题，我们可以模仿0-1变量回归时的情形，构造似然比检验的检验统计量如下：

$$\lambda = -2 \times \left(\max_{\beta_0} \log\{L(\beta_0,\beta_7 = 0)\} - \max_{(\beta_0,\beta_7)} \log\{L(\beta_0,\beta_7)\}\right)$$

$$= \left(-2 \times \max_{\beta_0} \log\{L(\beta_0,\beta_7 = 0)\}\right) - \left(-2 \times \max_{(\beta_0,\beta_7)} \log\{L(\beta_0,\beta_7)\}\right)$$

如果β_7是一个长度为d的向量（即β_7代表了d个不同的自由参数），那么经典的统计理论告诉我们，只要原假设$\beta_7 = 0$成立，而且样本量足够大，λ近似服从一个自由度为d的χ^2分布。请注意，上面的表达式同线性模型（更具体地讲，是方差分析模型）中的残差分析非常类似。换句话说，$\left(-2 \times \max_{\beta_0} \log\{L(\beta_0,\beta_7 = 0)\}\right)$很像是总平方和（$SST$），而$\left(-2 \times \max_{(\beta_0,\beta_7)} \log\{L(\beta_0,\beta_7)\}\right)$很像是残差平方和（$RSS$）。这说明，对逻辑回归模型我们也可以作方差分析，而借助的工具就是$(-2 \times \log\{L(\beta_0,\beta_7)\})$，我们称之为deviance。

第五节　多变量逻辑回归

在前面几节中，为了讲解的方便，我们主要考虑的是单变量（W7）逻辑回归模型。但是，大家可以从案例介绍以及描述性分析中看到，本案例所涉及的解释性变量远远大于一个。回忆一下，本案例所涉及的解释性变量包括：手机品牌（W1）、有无数码相机（W2）、能否收看电视（W3）、有无手写笔（W4）、电话本能否多条记录（W5）、有无 MP3（W6）以及游戏数目（W7）。如果我们希望将所有的解释性变量放在同一个逻辑回归模型下研究，那么其形式应为：

$$\begin{aligned}\Phi^{-1}\{P(\text{score} \leqslant k)\} &= \beta_0 - \beta_{11} \times I\{W1 = \text{Motorola}\} - \beta_{12} \times I\{W1 = \text{Nokia}\} \\ &\quad - \beta_{13} \times I\{W1 = \text{Samsung}\} - \beta_2 \times W2 - \beta_3 \times W3 - \beta_4 \times W4 \\ &\quad - \beta_5 \times W5 - \beta_6 \times W6 - \beta_7 \times W7\end{aligned}$$

类似于单变量定序回归模型，我们可以通过极大似然法获得参数估计，也可以通过 Z 检验来检验各个因素的显著性，还可以通过似然比检验来检验一组因素的整体显著性（如模型的整体显著性）。

在 R 中具体实现如下：我们首先考虑模型的整体显著性，也就是说，我们希望知道在我们所考虑的所有解释性变量中，是否至少有一个解释性变量和因变量（对该产品的偏好程度）显著相关。因此，我们的原假设是：

$$H_0: \beta_{11} = \beta_{12} = \beta_{13} = \beta_2 = \cdots = \beta_7 = 0$$

我们可以通过似然比检验达到该目的。这意味着我们要比较以下两个模型：

- 空模型：$\Phi^{-1}\{P(\text{score} \leqslant k)\} = \beta_0$
- 全模型：

$$\begin{aligned}\Phi^{-1}\{P(\text{score} \leqslant k)\} &= \beta_0 - \beta_{11} \times I\{W1 = \text{Motorola}\} - \beta_{12} \times I\{W1 = \text{Nokia}\} \\ &\quad - \beta_{13} \times I\{W1 = \text{Samsung}\} - \beta_2 \times W2 - \beta_3 \times W3 - \beta_4 \times W4 \\ &\quad - \beta_5 \times W5 - \beta_6 \times W6 - \beta_7 \times W7\end{aligned}$$

在 R 中，我们对这两个模型作方差分析如下：

```
> library(MASS)
> probit0=polr(as.factor(score)~1,method="probit",Hess=T)
> probit1=polr(as.factor(score)~W1+W2+W3+W4+W5+W6+W7,method="probit",Hess=T)
> anova(probit0,probit1)
Likelihood ratio tests of ordinal regression models

Response: as.factor(score)
                              Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1                                 1      1447   4257.205
2 W1 + W2 + W3 + W4 + W5 + W6 + W7      1438   3984.573 1 vs 2     9 272.6325       0
```

可以看到，广义似然比检验统计量为两个模型的 deviance 之差，即 272.6325。在原假设下，它应该服从一个自由度为 9 的卡方分布。其相应的 P 值几乎为 0，即高度显著。这说明，本模型整体高度显著，也就意味着我们所考虑的七个解释性变量中，至少有一个对 score 有显著的影响。这里我们要解释一下，为什么自由度为 9。将各个因素所消耗的自由度罗列如下：

- W1：定性变量，四个水平，共消耗 3 个自由度。
- W2 – W6：0-1 变量，每个消耗 1 个自由度，共消耗 5 个自由度。
- W7：数值型变量，消耗 1 个自由度。

全模型和空模型相比，一共多消耗了 9 个自由度。因此，我们的似然比检验的自由度为 9。如果我们改用 logit 定序回归，那么结果如下：

```
> logit0=polr(as.factor(score)~1,method="logistic",Hess=T)
> logit1=polr(as.factor(score)~W1+W2+W3+W4+W5+W6+W7,method="logistic",Hess=T)
> anova(logit0,logit1)
Likelihood ratio tests of ordinal regression models

Response: as.factor(score)
                              Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1                                 1      1447   4257.206
2 W1 + W2 + W3 + W4 + W5 + W6 + W7      1438   3983.182 1 vs 2     9 274.0233       0
```

可见，基本结论同 probit 定序回归一致，没有明显的改变。上面的分析告诉我们，至少有一个解释性变量对消费者的喜好程度具有一定的解释能力。但是，到底是哪一个我们并不知道，因此，需要分析各个因素的显著性：

```
> library(car)
> Anova(probit1,type="III")
Anova Table (Type III tests)

Response: as.factor(score)
   LR Chisq Df Pr(>Chisq)
W1   31.743  3  5.929e-07 ***
W2   48.462  1  3.367e-12 ***
W3   25.632  1  3.724e-07 ***
W4   16.184  1  5.749e-05 ***
W5  197.371  1  < 2.2e-16 ***
W6    7.540  1   0.006034 **
W7   -0.991  1   1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从中可以看到，在0.05的显著性水平下，游戏的个数（W7）是唯一一个不显著的影响因素。相关参数的具体估计如下：

```
> summary(probit)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6 +
    W7, Hess = T, method = "probit")

Coefficients:
                Value Std. Error    t value
W1Motorola  0.28047780 0.09607723  2.9192952
W1Nokia     0.48908026 0.08529173  5.7342052
W1Samsung   0.27652427 0.08612732  3.2106451
W2          0.39144950 0.05983700  6.5419310
W3          0.31159442 0.06187711  5.0356979
W4          0.25499404 0.06175564  4.1290810
W5          0.90093745 0.06362277 14.1606126
W6          0.20204838 0.07563474  2.6713701
W7         -0.01373407 0.01761078 -0.7798672

Intercepts:
    Value   Std. Error t value
1|2 -0.3036  0.1313     -2.3125
2|3  0.5214  0.1295      4.0247
3|4  1.5312  0.1331     11.5057
4|5  2.6876  0.1418     18.9589

Residual Deviance: 3984.573
AIC: 4010.573
```

请注意，标准正态分布的90%分位数为1.28。因此，我们从这张表中可以得到以下重要结论：

- 手机品牌（W1）：总共有四个水平，其中Bird没有出现，因此我们知道Bird是手机品牌这个因素的基准水平。下面我们看到，关于W1的其他任何水平的t值绝对值都大于1.28，这说明所有的相关检验都在0.05的水平下显著（请注意：双边检验）。而且我们还看到，所有的参数估计都是正的，如W1Nokia=0.489。这说明，在手机其他特征（即W2－W7）一样的情况下，消费者对诺基亚的偏好程度显著高于其他的品牌。这可以认为是诺基亚品牌机制的某种体现。但是，其他三个品牌之间的差异在统计上是否显著，我们就不得而知了。
- 除了游戏数目（W7）以外，其他因素都能够显著地影响消费者的喜好程度，尤其是电话本能否多条记录（W5）的影响最大（系数为0.90）。这可能同我们的调查对象主要是商务人士有关。
- 游戏数目（W7）对消费者的喜好程度影响不大。

类似地，我们可以对逻辑定序回归分析如下：

```
> Anova(logit1,type="III")
Anova Table (Type III tests)

Response: as.factor(score)
   LR Chisq Df Pr(>Chisq)
W1   33.134  3  3.018e-07 ***
W2   49.853  1  1.657e-12 ***
W3   27.223  1  1.813e-07 ***
W4   17.574  1  2.763e-05 ***
W5  198.761  1  < 2.2e-16 ***
W6    8.931  1   0.002804 **
W7    0.399  1   0.527499
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> summary(logit1)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6 +
    W7, Hess = T, method = "logistic")

Coefficients:
                Value Std. Error    t value
W1Motorola  0.48742350 0.16660698  2.9255888
W1Nokia     0.84501125 0.14834545  5.6962398
W1Samsung   0.46014425 0.14893757  3.0895110
W2          0.72610326 0.10351353  7.0145732
W3          0.55392611 0.10649928  5.2012379
W4          0.44891842 0.10730550  4.1835546
W5          1.53404456 0.11162693 13.7426025
W6          0.39004329 0.13069129  2.9844628
W7          0.01947124 0.03081032 -0.6319715

Intercepts:
    Value   Std. Error t value
1|2 -0.5244  0.2279     -2.3010
2|3  0.9844  0.2234      4.4072
3|4  2.6680  0.2321     11.4992
4|5  4.6717  0.2531     18.4573

Residual Deviance: 3983.182
AIC: 4009.182
```

可以看到,基本结论同 probit 定序回归非常相似。剔除不显著因素(游戏数目)后,重新拟合 probit 定序回归模型如下:

```
> probit2=polr(as.factor(score)~W1+W2+W3+W4+W5+W6,method="probit",Hess=T)
> summary(probit2)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "probit")

Coefficients:
             Value Std. Error   t value
W1Motorola 0.2709333 0.09528916  2.843276
W1Nokia    0.4895783 0.08528683  5.740374
W1Samsung  0.2793647 0.08606067  3.246137
W2         0.3897138 0.05979540  6.517454
W3         0.3074535 0.06164743  4.987289
W4         0.2534522 0.06172173  4.106369
W5         0.9026449 0.06358570 14.195721
W6         0.1868247 0.07307104  2.556754

Intercepts:
    Value   Std. Error t value
1|2 -0.2466  0.1090    -2.2619
2|3  0.5781  0.1072     5.3923
3|4  1.5876  0.1118    14.2039
4|5  2.7438  0.1221    22.4737

Residual Deviance: 3985.181
AIC: 4009.181
```

可以看到，基本结论没有改变。同样的模型，用 logit 定序回归拟合的结果为：

```
> logit2=polr(as.factor(score)~W1+W2+W3+W4+W5+W6,method="logistic",Hess=T)
> summary(logit2)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "logistic")

Coefficients:
             Value Std. Error   t value
W1Motorola 0.4733601  0.1650770  2.867510
W1Nokia    0.8449309  0.1483251  5.696480
W1Samsung  0.4646191  0.1487723  3.123021
W2         0.7242448  0.1034572  7.000427
W3         0.5492865  0.1062482  5.169843
W4         0.4475002  0.1072765  4.171467
W5         1.5354934  0.1115902 13.760107
W6         0.3684779  0.1261348  2.921302

Intercepts:
    Value   Std. Error t value
1|2 -0.4431  0.1881    -2.3550
2|3  1.0656  0.1828     5.8297
3|4  2.7483  0.1944    14.1342
4|5  4.7528  0.2186    21.7406

Residual Deviance: 3983.582
AIC: 4007.582
```

可以看到，该结果和 probit 定序回归的结果几乎一样。

第六节　模型选择

和普通线性回归一样，过分复杂的模型会降低参数估计的精度，并且使结果复杂难懂。但是，过分简化的模型会遗漏重要的预测变量，更会影响我们的统计推断结果。因此，有必要建立一个尽量简单但又比较准确的模型。因此，我们在本节中给定序回归模型也发展出一套类似 AIC 和 BIC 的模型选择标准。在第四节中我们曾经提到，定序变量回归的 deviance 能够起到类似于残差平方和的作用，这一点和 0-1 变量回归非常类似。这提示我们，对于定序回归模型，我们可以定义 AIC 和 BIC 如下：

$$\text{AIC} = \text{deviance} + 2 \times df$$

$$\text{BIC} = \text{deviance} + \log(n) \times df$$

其中，n 代表了样本量，而 df 代表了自由度（即自由参数的个数）。下面我们以逻辑定序回归的全模型以及空模型为例，具体地演示 AIC 和 BIC 的计算方法。首先计算各个模型的自由度：对于空模型，我们有四个截距项，因此有 4 个自由参数，即 $df = 4$；对于全模型，我们共有七个解释性变量，在第五节中，我们发现这七个解释性变量共消耗了 9 个自由度，再加上 4 个截距项，所以总共有 13 个自由参数，即 $df = 13$。因此，它们的 AIC 取值分别为：

$$\text{空模型：}4\,257.205 + 2 \times 4 = 4\,265.205$$

$$\text{全模型：}3\,984.573 + 2 \times 13 = 4\,010.573$$

其中，deviance 的取值来自于参数估计的报表，如 summary(probit1)。注意到我们的样本量为 1 451，因此，相应的 BIC 取值分别为：

$$\text{空模型：}4\,257.205 + \log(1\,451) \times 4 = 4\,286.325$$

$$\text{全模型：}3\,984.573 + \log(1\,451) \times 13 = 4\,079.213$$

因为无论 AIC 还是 BIC，都是取值越小越好，因此，无论采用 AIC 还是 BIC，我们都会认为全模型优于空模型。当然，目前我们仅仅比较了两个不同的模型。因为我们有七个解释性变量，因此我们一共有 $2^7 = 128$ 个不同的模型。理论上讲，我们应该对这 128 个模型逐一研究，并选择最优的模型。但是，这样做的缺点就是需要我们自己编写程序。而在 R 中，可以自动地且尽量多地根据 AIC 搜索最优模型如下：

```
> logit.aic=step(logit1,trace=F)
> summary(logit.aic)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "logistic")

Coefficients:
              Value Std. Error   t value
W1Motorola 0.4733601  0.1650770  2.867510
W1Nokia    0.8449309  0.1483251  5.696480
W1Samsung  0.4646191  0.1487723  3.123021
W2         0.7242448  0.1034572  7.000427
W3         0.5492865  0.1062482  5.169843
W4         0.4475002  0.1072765  4.171467
W5         1.5354934  0.1115902 13.760107
W6         0.3684779  0.1261348  2.921302

Intercepts:
    Value   Std. Error t value
1|2 -0.4431  0.1881     -2.3550
2|3  1.0656  0.1828      5.8297
3|4  2.7483  0.1944     14.1342
4|5  4.7528  0.2186     21.7406

Residual Deviance: 3983.582
AIC: 4007.582
```

也可以自动地且尽量多地根据 BIC 搜索最优模型如下：

```
> logit.bic=step(logit1,trace=F,k=log(length(A[,1])))
> summary(logit.bic)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "logistic")

Coefficients:
              Value Std. Error   t value
W1Motorola 0.4733601  0.1650770  2.867510
W1Nokia    0.8449309  0.1483251  5.696480
W1Samsung  0.4646191  0.1487723  3.123021
W2         0.7242448  0.1034572  7.000427
W3         0.5492865  0.1062482  5.169843
W4         0.4475002  0.1072765  4.171467
W5         1.5354934  0.1115902 13.760107
W6         0.3684779  0.1261348  2.921302

Intercepts:
    Value   Std. Error t value
1|2 -0.4431  0.1881     -2.3550
2|3  1.0656  0.1828      5.8297
3|4  2.7483  0.1944     14.1342
4|5  4.7528  0.2186     21.7406

Residual Deviance: 3983.582
AIC: 4007.582
```

对于这个案例，AIC 和 BIC 得到相同的结论。它们都认为，除了游戏数目（W7）以外，其他的解释性变量都同消费者的喜好程度显著相关。我们也可以对 probit 定序回归模型作类似的模型选择，AIC 的选择结果如下：

```
> probit.aic=step(probit1,trace=F)
> summary(probit.aic)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "probit")

Coefficients:
              Value Std. Error   t value
W1Motorola 0.2709333 0.09528916  2.843276
W1Nokia    0.4895783 0.08528683  5.740374
W1Samsung  0.2793647 0.08606067  3.246137
W2         0.3897138 0.05979540  6.517454
W3         0.3074535 0.06164743  4.987289
W4         0.2534522 0.06172173  4.106369
W5         0.9026449 0.06358570 14.195721
W6         0.1868247 0.07307104  2.556754

Intercepts:
    Value   Std. Error t value
1|2 -0.2466  0.1090     -2.2619
2|3  0.5781  0.1072      5.3923
3|4  1.5876  0.1118     14.2039
4|5  2.7438  0.1221     22.4737

Residual Deviance: 3985.181
AIC: 4009.181
```

而 BIC 的选择结果如下：

```
> probit.bic=step(probit1,trace=F,k=log(length(a1,1)))
> summary(probit.bic)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5, Hess = T,
    method = "probit")

Coefficients:
              Value Std. Error   t value
W1Motorola 0.1540565 0.08361360  1.842481
W1Nokia    0.4939658 0.08526346  5.793406
W1Samsung  0.3060794 0.08539158  3.584422
W2         0.3658825 0.05905394  6.195734
W3         0.2612849 0.05892215  4.434408
W4         0.2341965 0.06123580  3.824502
W5         0.9187621 0.06326261 14.522987

Intercepts:
    Value   Std. Error t value
1|2 -0.3956  0.0921     -4.2939
2|3  0.4294  0.0901      4.7659
3|4  1.4361  0.0947     15.1610
4|5  2.5872  0.1054     24.5480

Residual Deviance: 3991.721
AIC: 4013.721
```

可以看到，probit 定序回归的模型选择结果同 logit 定序回归完全一致。

第七节　预测与评估

一旦建立了 logit 定序回归模型（或者 probit 模型），我们就可以对新的观测予以预测。具体地说，如果我们有一款新的手机，那么我们就可以对未来消费者对该手机的评价予以预测。假设我们有另外一个独立的检验样本，例如：

```
> a0=read.csv("D:/Practical Business Data Analysis/case/CH5/new.csv")
> a0[c(1:5),]
  score    W1 W2 W3 W4 W5 W6 W7
1     3 Nokia  0  0  0  1  1  3
2     4 Nokia  1  0  1  0  1  5
3     4 Nokia  0  1  1  0  0  7
4     4  Bird  1  1  0  1  0  3
5     3  Bird  0  0  0  0  1  5
```

我们以第一个数据为例，也就是说，我们考虑这样一款手机：它的品牌为诺基亚，没有数码相机，不能收看电视，没有手写笔，电话本支持多条记录，有 MP3 以及三个游戏。现在的问题是：消费者对这款手机的评价如何？请注意，消费者对手机的评价有五种不同的可能（1 = 根本不喜欢；2 = 比较不喜欢；3 = 一般喜欢；4 = 比较喜欢；5 = 非常喜欢）。因此，我们需要计算各种可能的概率。为此，我们首先计算 AIC 模型的参数如下：

```
> summary(logit.olr)
Call:
polr(formula = as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,
    Hess = T, method = "logistic")

Coefficients:
              Value Std. Error   t value
W1Motorola 0.4733601  0.1650770  2.867510
W1Nokia    0.8449309  0.1483251  5.696480
W1Samsung  0.4646191  0.1487723  3.123021
W2         0.7242448  0.1034572  7.000427
W3         0.5492865  0.1062482  5.169843
W4         0.4475002  0.1072765  4.171467
W5         1.5354934  0.1115902 13.760107
W6         0.3684779  0.1261348  2.921302

Intercepts:
    Value   Std. Error t value
1|2 -0.4431  0.1881     -2.3550
2|3  1.0656  0.1828      5.8297
3|4  2.7483  0.1944     14.1342
4|5  4.7528  0.2186     21.7406

Residual Deviance: 3983.582
AIC: 4007.582
```

然后，计算线性组合如下：

$$0.8449309 + 0.7242448 \times 0 + 0.5492865 \times 0 + 0.4475002 \times 0 + 1.5354934 \times 1 + 0.3684779 \times 1 = 2.748902$$

计算 score≤1 的概率为：

$$P(\text{score} \leqslant 1) = \frac{\exp(-0.4431 - 2.748902)}{1 + \exp(-0.4431 - 2.748902)} = 0.0395$$

计算 score≤2 的概率为：

$$P(\text{score} \leqslant 2) = \frac{\exp(1.0656 - 2.748902)}{1 + \exp(1.0656 - 2.748902)} = 0.1566587$$

因此，score = 2 的概率为 0.1566587 - 0.0395 ≈ 0.12。类似地，我们可以推算 score = 3,4,5 的概率分别为 0.34,0.37,0.12。因此，最有可能的 score 取值为 4，其次是 3。我们可以从数据中看到，其真实的打分就是 3。以上烦琐的计算过程可以通过 *R* 自动实现如下：

```
> a0$score.hat=predict(probit.aic,a0)
> a0[c(1:5),]
  score    W1 W2 W3 W4 W5 W6 W7 score.hat
1     3 Nokia  0  0  0  1  1  3         4
2     4 Nokia  1  0  1  0  1  5         3
3     4 Nokia  0  1  1  0  0  7         3
4     4  Bird  1  1  0  1  0  3         4
5     3  Bird  0  0  0  0  1  5         1
```

从中可以看到，我们所展示的五个数据中，只有第五个的预测偏差较大（真实值为 3，而预测值为 1），其他的预测偏差都在 1 个单位以内。我们对整个的数据预测情况描述如下：

```
> table(a0[,2:1,9:])
     score.hat
score  1  2  3  4  5
    1  2  0 12  5  0
    2  3  0 10  9  0
    3  3  0 15 10  0
    4  0  0 13 10  0
    5  0  0  1  7  0
```

我们看到有 15 个数据真实得分为 3，预测结果也是 3。类似地，有 13 个数据真实得分为 4，而预测结果为 3。我们总共有 2 + 15 + 10 = 27 个数据获得了完全准确的预测，有 3 + 13 + 7 + 10 + 10 = 43 个数据的预测误差为 1 个单位，只有 5 + 12 + 9 + 1 + 3 = 30 个数据的预测误差较大，超过了 1 个单位，这部分数据占整个检验样本的 30/100 = 30%。因此，预测结果良好。

第八节　简单分析报告

手机功能对消费者偏好的影响分析

内容提要　研究消费者对新功能的偏好，对于手机厂商来说有着特别重要的意义。本报告利用消费者对手机的评分数据，对手机功能在不同功能组合中所起到的作用进行精确的数量化分析。我们的分析结果发现，手机品牌、数码相机功能、收看电视功能、手写笔、电话本多条记录、MP3 这些因素会显著地影响消费者的喜好程度，尤其是电话本能否多条记录的影响最大。根据本报告的模型及结论，厂商可以选择合理的市场策略来开发手机的新功能。

一、研究目的

目前国内手机销售市场竞争日趋激烈，为了确立在市场中的相对优势地位，开发新功能是企业常常采用的手段。因为与开发新产品相比，在现有产品的基础上增加新功能不失为一种既快速又有效的方式。当企业决定为现有产品增加新功能时，往往面对众多选择。根据消费者的行为特征，消费者在进行购买决策时愿意支付的价格主要依赖于对产品的偏好，而不是企业的生产成本。因此，研究消费者对新功能的偏好对于手机厂商就有着特别重要的意义。本分析报告对手机功能在不同功能组合中所起到的作用进行精确的分析，建立合理的统计模型，并根据分析结果提出合理的建议。

二、数据来源和相关说明

本报告的数据来源是对北京大学光华管理学院的 MBA 学生和高级经理培训班的学员的调查。首先，我们选取商务手机厂商较多考虑的六个功能，然后加上品牌（这里我们只涉及诺基亚、摩托罗拉、三星和波导四个品牌）共七个要素，构成我们要研究的影响消费者偏好的要素。具体来说，包括手机品牌、数码相机功能、收看电视功能、手写笔、电话本多条记录、MP3 和游戏数目。我们将这七个要素按不同方式组成 12 个产品组合，再针对不同的组合进行偏好调查。对每个组合，调查对象根据其偏好程度用 5 分量表进行打分（1 = 根本不喜欢；2 = 比较不喜欢；3 = 一般喜欢；4 = 比较喜欢；5 = 非常喜欢）。基于我们的调查结果并作适当的数据清理，最后共获得来自 148 个调查对象的 1 451 个有效观测值。各个变量的具体取值见表 5-3，不同的功能组合见表 5-4。

表 5-3 变量说明

变量类型	变量含义	变量名	变量水平
因变量	对该产品的偏好程度	score	1 = 根本不喜欢;2 = 比较不喜欢;3 = 一般喜欢;4 = 比较喜欢;5 = 非常喜欢
自变量	手机品牌	W1	共四种(诺基亚、摩托罗拉、三星和波导)
	有无数码相机	W2	共两种(有、无)
	能否收看电视	W3	共两种(能、不能)
	有无手写笔	W4	共两种(有、无)
	电话本能否多条记录	W5	共两种(能、不能)
	有无 MP3	W6	共两种(有、无)
	游戏数目	W7	连续型

表 5-4 手机功能组合

品牌	数码相机	能否收看电视	手写笔	电话本能否多条记录	MP3	游戏数目
诺基亚	无	不能	无	能	有	3
	有	不能	有	不能	有	5
	无	能	有	不能	无	7
波导	有	能	无	能	无	3
	无	不能	无	不能	有	5
	有	不能	有	能	有	7
摩托罗拉	无	能	有	能	无	3
	有	能	无	不能	无	5
	无	不能	无	能	无	7
三星	有	不能	有	不能	无	3
	无	能	有	不能	有	5
	有	能	无	能	有	7

从表 5-3 和表 5-4 中可以看到,因变量是离散型的变量,而且是定序变量。而自变量是各种组合中所涉及的七个因素,既有连续型数据,也有离散型数据。

三、描述性分析

为了获得对数据的整体认识,并注意到因变量为定序变量的特点,我们首先利用列联表考察消费者打分和不同品牌之间的关系,如表 5-5 所示。

表 5-5　消费者打分和不同品牌的列联表

消费者打分	波导	摩托罗拉	诺基亚	三星
1	37	24	34	26
2	80	64	53	66
3	98	138	132	133
4	109	108	116	96
5	28	30	35	44

从中我们可以看到，在所有得分为 1 或 2 的品牌中，频数最高的是波导（Bird），其频数分别为 37 和 80；在得分为 3 的品牌中，摩托罗拉（Motorola）频数最高，为 138；在得分为 4 的品牌中，诺基亚（Nokia）频数最高，为 116；在得分为 5 的品牌中，三星（Samsung）频数最高，为 44。由此可见，国产品牌在与国际品牌的对比中具有较大的劣势，而在国际品牌中，摩托罗拉偏向大众化路线，而诺基亚偏向中高端市场。

我们再对消费者打分和手机有无数码相机功能之间的关系予以简单的分析，如图 5-3 所示。

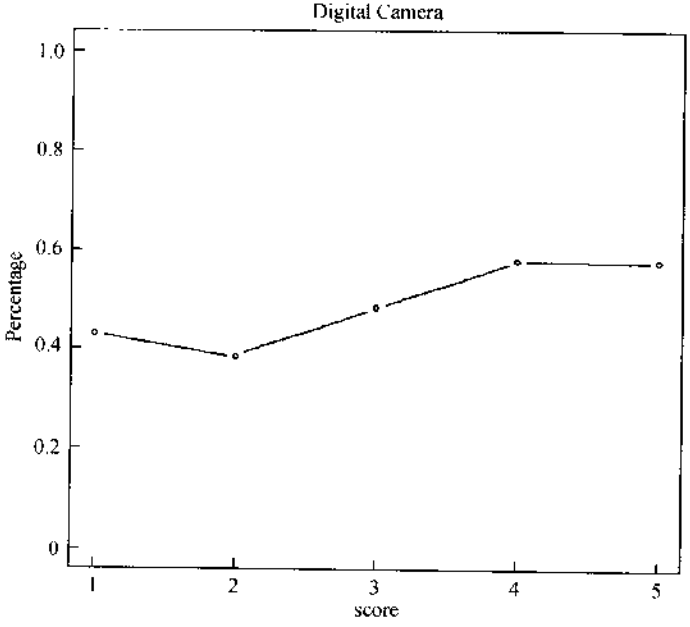


图 5-3　消费者打分和有无数码相机之间关系图

从总体上来讲，我们可以看到一个明显的上升趋势。具体地说，得分越高的手机，具有数码相机功能的比率越高，特别是在比较不喜欢（score = 2）到比较喜欢（score = 4）之间。这从一个侧面说明，有无数码相机功能在当时是一个界

定人们对其打分是否高于平均水平的重要属性。

对其他几个定性因素，我们作类似的分析，如图 5-4 所示。

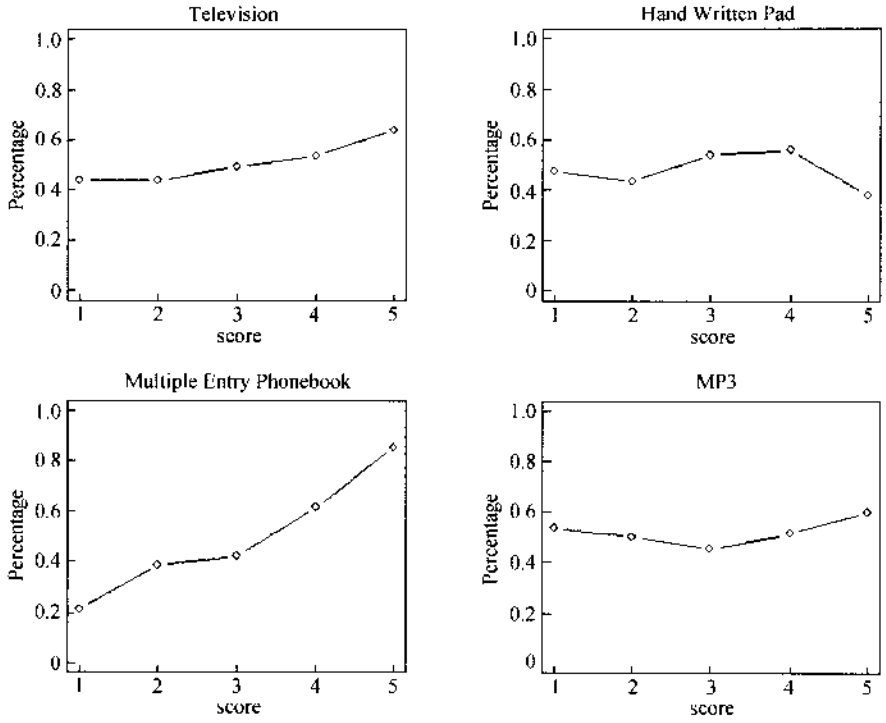


图 5-4　消费者打分和其他功能之间关系图

从中可以看到，能否收看电视（*W3*）以及电话本是否支持多条记录（*W5*）同消费者打分（score）高度正相关。而有无手写笔（*W4*）和是否支持 MP3 功能（*W6*）在我们的样本中似乎并没有受到很大的青睐。

最后，我们考察消费者打分（score）和游戏个数（*W7*）之间的相互关系，得到列联表 5-6。

表 5-6　消费者打分和游戏个数的列联表

消费者打分	游戏个数		
	3	5	7
1	25	67	29
2	71	113	79
3	169	187	145
4	155	111	163
5	64	9	64

从表 5-6 中很难看出明显的趋势，因此，游戏个数同消费者打分之间没有明显的关系。

四、数据建模

1. 全模型分析

考虑到因变量为定序变量的特点，我们采用 probit 定序回归和 logit 定序回归的方法建立模型。我们首先用 probit 定序回归的方法对包含全部自变量的全模型进行卡方检验，如表 5-7 所示。

表 5-7　全模型卡方检验（probit 定序回归）

变量名	卡方统计量	自由度	*P* 值
手机品牌	31.743	3	<0.001
有无数码相机	48.462	1	<0.001
能否收看电视	25.832	1	<0.001
有无手写笔	16.184	1	<0.001
电话本能否多条记录	197.371	1	<0.001
有无 MP3	7.540	1	0.006
游戏数目	-0.991	1	1.000

相应的参数估计结果如表 5-8 所示。

表 5-8　全模型（probit 定序回归）

变量名	系数估计值	标准差	*Z* 值
截距：1|2	-0.3036	0.1313	-2.3125
截距：2|3	0.5214	0.1295	4.0247
截距：3|4	1.5312	0.1331	11.5057
截距：4|5	2.6876	0.1418	18.9589
手机品牌：摩托罗拉	0.2805	0.0961	2.9193
手机品牌：诺基亚	0.4891	0.0853	5.7342
手机品牌：三星	0.2765	0.0861	3.2106
有无数码相机	0.3914	0.0598	6.5419
能否收看电视	0.3116	0.0619	5.0357
有无手写笔	0.2550	0.0618	4.1291
电话本能否多条记录	0.9010	0.0636	14.1606
有无 MP3	0.2020	0.0756	2.6714
游戏数目	-0.0137	0.0176	-0.7799
模型显著性检验 *P* 值 <0.001			

从表 5-8 中我们可以看到，广义似然比的模型显著性检验的 *P* 值非常小（*P*

值 <0.001），表明该模型是显著的，即我们所考虑的七个解释性变量中，至少有一个对消费者打分有显著的影响。进一步通过对各个自变量所对应的 Z 检验的考察，我们可以得到以下重要结论：

• 手机品牌对消费者打分有显著的影响。具体地说，在四个品牌中消费者对波导（Bird）的偏好程度显著低于其他品牌。而且，消费者对于诺基亚的偏好程度最高。这可以认为是不同品牌机制的某种体现。但是，除波导以外的三个品牌之间的差异在统计上是否显著，我们就不得而知了。

• 除了游戏数目以外，其他因素都能够显著地影响消费者的喜好程度，尤其是电话本能否多条记录的影响最大，它的系数为 0.90。而且需要注意的是，这些有显著影响的功能的系数都为正数，也就是说，这些功能增加了消费者的喜好程度。

• 结合我们的调查对象主要是商务人士的特点，我们可以知道，游戏数目对消费者（尤其对商务人士）的喜好程度影响不大。

类似地，我们用 logit 定序回归的方法得出了和 probit 定序回归非常相似的结论。具体地说，全模型卡方检验结果如表 5-9 所示。

表 5-9　全模型卡方检验（logit 定序回归）

变量名	卡方统计量	自由度	P 值
手机品牌	33.134	3	<0.001
有无数码相机	49.853	1	<0.001
能否收看电视	27.223	1	<0.001
有无手写笔	17.574	1	<0.001
电话本能否多条记录	198.761	1	<0.001
有无 MP3	8.931	1	0.003
游戏数目	0.399	1	0.527

从中我们发现，手机品牌、有无数码相机、能否收看电视、有无手写笔、电话本能否支持多条记录，还有有无 MP3 都是影响消费者对手机偏好程度的重要因素。而游戏数目是唯一一个对消费者偏好程度无显著影响的手机属性。对该逻辑回归模型的参数估计如表 5-10 所示。

表 5-10　全模型（logit 定序回归）

变量名	系数估计值	标准差	Z 值
截距：1|2	−0.5244	0.2279	−2.3010
截距：2|3	0.9844	0.2234	4.4072
截距：3|4	2.6680	0.2321	11.4952

（续表）

变量名	系数估计值	标准差	Z 值
截距:415	4.6717	0.2351	18.4573
手机品牌:摩托罗拉	0.4874	0.1666	2.9256
手机品牌:诺基亚	0.8450	0.1483	5.6962
手机品牌:三星	0.4601	0.1489	3.0895
有无数码相机	0.7261	0.1035	7.0145
能否收看电视	0.5539	0.1065	5.2012
有无手写笔	0.4489	0.1073	4.1836
电话本能否多条记录	1.5340	0.1116	13.7426
有无 MP3	0.3900	0.1307	2.9845
游戏数目	-0.0195	0.0308	-0.6319
模型显著性检验 P 值 <0.001			

2. 模型选择

以上的分析结果已经告诉我们，我们所选取的手机功能确实对消费者打分有一定的解释能力。但是全模型过于复杂，而且其中还有部分变量是不显著的。为了得到一个尽量简单同时又具有良好预测能力的模型，我们采用 AIC 和 BIC 的模型选择标准来选择一个最优的模型。对 logit 定序回归模型，我们用 AIC 方法选出的模型及其估计结果如表 5-11 所示。

表 5-11　AIC、BIC（logit 定序回归）

变量名	系数估计值	标准差	Z 值
截距:112	-0.4431	0.1881	-2.3550
截距:213	1.0656	0.1828	5.8297
截距:314	2.7483	0.1944	14.1342
截距:415	4.7528	0.2186	21.7406
手机品牌:摩托罗拉	0.4734	0.1651	2.8675
手机品牌:诺基亚	0.8449	0.1483	5.6965
手机品牌:三星	0.4646	0.1487	3.1230
有无数码相机	0.7242	0.1035	7.0004
能否收看电视	0.5493	0.1062	5.1698
有无手写笔	0.4475	0.1073	4.1715
电话本能否多条记录	1.5355	0.1116	13.7601
有无 MP3	0.3685	0.1261	2.9213
模型显著性检验 P 值 <0.001			

采用 BIC 方法得到了和 AIC 相同的结论。另外值得注意的是，该模型就是从全模型中剔除不显著的变量后得到的模型，而且两个模型的系数估计值和全

模型显示出同样的规律，即同样的符号和同样的大小顺序。因此，我们可以认为该结论是可靠的。具体地说，所有这些模型都认为，除了游戏数目以外，其他的解释性变量都同消费者的喜好程度显著相关。

对 probit 定序回归模型作类似的模型选择，我们发现结果同 logit 定序回归完全一致，而且 AIC 和 BIC 也选择了同样的模型。这进一步验证了我们所得到的结论的可靠性，因而我们有充分的理由相信，这六个因素确实显著地影响消费者对于手机的喜好程度。

3. 模型预测与评估

建立模型的一个重要目的就是作预测。我们用 logit 定序回归中 AIC 所选出的模型对新的观测予以预测。对每一个消费者打分(score = 1，…，5)，我们可以计算其各自的概率，然后选取概率最大的打分作为最终的预测。我们对全部的检验数据进行预测，结果如表 5-12 所示。

表 5-12　模型预测结果

真实的消费者打分	预测的消费者打分				
	1	2	3	4	5
1	2	0	12	5	0
2	3	0	10	9	0
3	3	0	15	10	0
4	0	0	13	10	0
5	0	0	1	7	0

我们看到有 15 个数据真实得分为 3，预测结果也是 3。类似地，有 13 个数据真实得分为 4，而预测结果为 3。我们总共有 2 + 15 + 10 = 27 个数据获得了完全准确的预测，有 3 + 13 + 7 + 10 + 10 = 43 个数据的预测误差为 1 个单位，只有 5 + 12 + 9 + 1 + 3 = 30 个数据的预测误差较大，超过了 1 个单位，这部分数据占整个检验样本的 30/100 = 30%。这表明该模型的预测结果还是比较令人满意的。

五、结论及建议

从上述分析结果可知，不同的手机功能确实对消费者的喜好程度有一定的解释和预测能力。具体来说，手机品牌、数码相机功能、收看电视功能、手写笔、电话本多条记录、MP3 这六个因素会显著地影响消费者的喜好程度，尤其是电话本能否多条记录的影响最大。根据本报告的模型及结论，厂商可以选择合理的市场策略来开发手机的新功能。例如，由于手机中的游戏数目对于消费者的喜好程度没有显著的影响，因此手机厂商在开发新功能时，同等成本下就不应

该添加游戏，而应该考虑添加其他的功能。另外，手机厂商还可以根据本报告中的模型对消费者对于不同的手机功能组合的喜好程度进行预测，进而制定合理的定价及广告宣传等市场策略。需要特别指出的是，由于本报告所使用的数据主要来自于商务人士，因此，所得出的结论也主要适用于商务人士。

[讨论总结]

本章以关于消费者偏好的市场调研数据为例，系统演示并讲解了 logit 以及 probit 两种定序回归模型。通过对本章的学习，读者应该能够了解：什么时候可以使用定序回归模型，以及如何使用。在 *R* 语言学习方面，读者应该掌握相关的广义线性模型的命令。在统计理论方面，读者应该掌握以下概念：定序数据、logit 定序回归、probit 定序回归等。对相关统计学理论渴望深入了解的读者可以参阅 McCullagh and Nelder(1999)。

附录　程序及注释

```
rm(list=ls())                                          # 清空当前工作空间
a=read.csv("D:/Practical Business Data Analysis/case/CH5/ceilphone.csv")
                                                       # 读入 csv 格式的数据,作为训练数据
attach(a)                                              # 将数据 a 中的各个变量放入工作空间,便于直接调用
a[c(1:5),]                                             # 显示数据的前 5 行
xtabs( ~ score + W1)                                   # 根据 score 和 W1 的取值生成列联表
plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Digital Camera")
                                                       # 生成画图的框架,给出 x、y 轴的标签和标题
points(c(1:5),tapply(W2,score,mean),type="b")          # 画出根据 score 的取值计算的 W2 的均值
par(mfrow=c(2,2))                                      # 生成 2x2 的图形
plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Television")
                                                       # 生成画图的框架,给出 x、y 轴的标签和标题
points(tapply(W3,score,mean),type="b")                 # 画出根据 score 的取值计算的 W3 的均值
plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Hand Written Pad")
                                                       # 生成画图的框架,给出 x、y 轴的标签和标题
points(tapply(W4,score,mean),type="b")                 # 画出根据 score 的取值计算的 W4 的均值
plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="Multiple Entry Phonebook")
                                                       # 生成画图的框架,给出 x、y 轴的标签和标题
points(tapply(W5,score,mean),type="b")                 # 画出根据 score 的取值计算的 W5 的均值
plot(c(1,5),c(0,1),type="n",xlab="score",ylab="Percentage",main="MP3")
```

```
                                                    # 生成画图的框架,给出 x,y 轴的标签和标题
points(tapply(W6.score,mean),type="b")              # 画出根据 score 的取值计算的 W6 的均值
par(mfrow=c(1,1))                                   # 设置画图的格式成 1x1 的形式
xtabs( ~ score + W7)                                # 根据 score 和 W7 的取值生成列联表
library(MASS)                                       # 载入程序包 MASS
probit0=polr(as.factor(score) ~ 1,method="probit",Hess=T)   # 拟合 probit 定序回归,不用任何解释性变量的空模型
probit1=polr(as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6 + W7,method="probit",Hess=T)
                                                    # probit 定序回归全模型
anova(probit0,probit1)                              # 对模型 probit0 和 probit1 进行方差分析,检验模型的显著性
logit0=polr(as.factor(score) ~ 1,method="logistic",Hess=T)   # 拟合 logit 定序回归,不用任何解释性变量的空模型
logit1=polr(as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6 + W7,method="logistic",Hess=T)
                                                    # logit 定序回归全模型
anova(logit0,logit1)                                # 对模型 logit0 和 logit1 进行方差分析,检验模型的显著性
library(car)                                        # 载入程序包 car
Anova(probit1,type="III")                           # 对模型 probit1 作三型方差分析
summary(probit1)                                    # 显示模型 probit1 的各方面细节,包括参数估计值、P 值等
Anova(logit1,type="III")                            # 对模型 logit1 作三型方差分析
summary(logit1)                                     # 显示模型 logit1 的各方面细节,包括参数估计值、P 值等
probit2=polr(as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,method="probit",Hess=T)
                                                    # 拟合 probit 定序回归,利用变量 W1 - W6
summary(probit2)                                    # 显示模型 probit2 的各方面细节,包括参数估计值、P 值等
logit2=polr(as.factor(score) ~ W1 + W2 + W3 + W4 + W5 + W6,method="logistic",Hess=T)
                                                    # 拟合 logit 定序回归,利用变量 W1 - W6
summary(logit2)                                     # 显示模型 logit2 的各方面细节,包括参数估计值、P 值等
logit.aic=step(logit1,trace=F)                      # 根据 AIC 准则从全模型 logit1 中选出最优子模型 logit.aic
```

```
summary(logit.aic)                                                    # 显示模型 logit.aic 的各方面细节,包括参数估计值、P 值等
logit.bic=step(logit1,trace=F,k=log(length(a[,1])))                   # 根据 BIC 准则从全模型 logit1 中选出最优子模型 logit.bic
summary(logit.bic)                                                    # 显示模型 logit.bic 的各方面细节,包括参数估计值、P 值等
probit.aic=step(probit1,trace=F)                                      # 根据 AIC 准则从全模型 probit1 中选出最优子模型 probit.aic
summary(probit.aic)                                                   # 显示模型 probit.aic 的各方面细节,包括参数估计值、P 值等
probit.bic=step(probit1,trace=F,k=log(length(a[,1])))                 # 根据 BIC 准则从全模型 probit1 中选出最优子模型 probit.bic
summary(probit.bic)                                                   # 显示模型 probit.bic 的各方面细节,包括参数估计值、P 值等
a0=read.csv("D:/Practical Business Data Analysis/case/CH5/new.csv")   # 读入 csv 格式的数据,用于检验
a0[c(1:5),]                                                           # 显示数据 a0 的前 5 行
summary(logit.aic)                                                    # 显示模型 logit.aic 的各方面细节,包括参数估计值、P 值等
p=predict(probit.aic,a0,type="p")                                     # 利用模型 probit.aic 预测数据取值为各水平(不同 score)的概率
a0$score.hat=predict(probit.aic,a0)                                   # 利用模型 probit.aic 对数据 a0 进行预测,将结果存入 score.hat
a0[c(1:5),]                                                           # 显示数据 a0 的前 5 行
table(a0[,c(1,9)])                                                    # 根据预测值和真实值生成列联表,展示预测精度
```

第六章
泊松回归

- 案例介绍
- 数据描述
- 泊松回归
- 参数估计与统计推断
- 模型选择与预测
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过一个客户关系管理的实际案例，详细介绍泊松回归这种重要的计数回归模型。它主要处理的是因变量为计数数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用泊松回归；(2) 泊松回归分析的基本统计学理论；(3) 相关理论在统计学软件 *R* 中的应用；(4) 相应的统计分析报告的撰写。本章所涉及的新统计学概念有计数数据、泊松回归。

第一节　案例介绍

前一章我们讨论了定序数据，例如：1 = 不喜欢；2 = 无所谓；3 = 喜欢。我们还讨论了该类数据的特征，那就是：无数值意义但有顺序特征。下面我们将这种数据类型和另外一种数据类型比较一下。具体地说，我们考虑某位顾客在某月内光顾超市的次数，可能是 1 次、2 次、3 次等。请注意，这个记录顾客光顾次数的 1、2、3 同记录消费者喜好程度的 1、2、3 有没有差别呢？显然，前者是有数值意义的。也就是说，2 次和 1 次的差异严格等于 3 次和 2 次的差异。那么，我们能否用普通线性回归的方法来研究顾客光顾超市的次数呢？答案是否定的。原因是，虽然顾客光顾超市的次数是一个具有数值意义的变量，但是它不是连续的。简单地说，没有哪个顾客在一个月内可以光顾超市 0.25 次。尽管有可能该顾客在一段很长的时间内“平均”每月光顾超市 0.25 次，但当具体到某个月的时候，他不可能光顾了该超市 0.25 次。请注意，顾客光顾超市的次数是一个衡量顾客活跃程度、度量客户价值的重要指标。因此，超市经理非常希望知道什么样的顾客光顾超市会频繁一些，而什么样的顾客较少光顾超市。本章将以某超市的会员数据为例，详细讲述如何分析此类数据。由于我们所感兴趣的数据是一个关于“次数”的记录，因此，我们称此类数据为计数数据。

具体地说，我们的数据来源于我国北方某城市处于垄断地位的一家超市，数据包含了该超市一部分会员的详细消费记录。我们以某年某月为基准月份（第 0 月），因此，可以将前一个月记为第 -1 月，以此类推。我们的因变量是一个会员在基准月份光顾该超市的次数，因此是一个典型的计数数据。超市经理感兴趣的问题是能否从这些会员前三个月的消费记录中找出什么规律，以便于

判断超市的众多会员中哪些人在这个月还会光顾超市以及大约会光顾多少次。为此，我们整理了每一个会员前三个月的每月光顾次数以及每月的消费金额。如果客户在某月没有光顾超市，那么将他的消费金额记为0。这是不是一种最好的处理方法呢？不大可能。但是为了简单起见，我们暂时接受这种数据处理方法。最后，我们得到的数据共包含了3995个有效样本，而对于相关变量的详细解释如表6-1所示。

表6-1 变量说明

变量名称	作用	实际意义
freq0	因变量	第0月光顾超市的频数
freq1	自变量	第-1月光顾超市的频数
freq2	自变量	第-2月光顾超市的频数
freq3	自变量	第-3月光顾超市的频数
exp1	自变量	第-1月的消费金额
exp2	自变量	第-2月的消费金额
exp3	自变量	第-3月的消费金额

第二节 数据描述

按照惯例，我们首先对数据予以描述性分析，以获得对数据的初步认识，形成待检验的结论并指导我们进行下一步的数据分析。我们首先读入数据并展示如下：

```
> rm(list=ls())
> a=read.csv("D:/Practical Business Data Analysis/case/CH6/crm.csv")
> attach(a)
> a[c(1:5),]
  exp3 exp2 exp1 freq3 freq2 freq1 freq0
1 45.4  0.0    0     1     0     0     0
2 79.6  9.8    0     2     2     0     3
3  0.0  0.0    0     0     0     0     0
4  0.0  0.0    0     0     0     0     0
5  2.1  0.0    0     1     0     0     2
```

从以上数据的第一行我们可以看到，有这样一名超市会员，他在第-3月光顾超市1次并消费45.4元。除此以外，该顾客没有再次光顾该超市。因此，不奇怪他在第0月也没有光顾超市。从以上数据的第二行我们可以看到，有另外一名超市会员，他分别在第-2月和第-3月光顾超市2次，消费金额分别为79.6元和9.8元。因此，这是一个相对活跃的客户，该客户在第0月光顾超市3次。我们再对数据简要描述如下：

```
> Mean=sapply(a[,1:6],mean)
> Min=sapply(a[,1:6],min)
> Median=sapply(a[,1:6],median)
> Max=sapply(a[,1:6],max)
> SD=sapply(a[,1:6],sd)
> round(cbind(Mean,Min,Median,Max,SD),3)
        Mean Min Median   Max     SD
exp3  16.483   0      0 184.9 24.919
exp2   9.241   0      0 159.1 16.621
exp1   7.917   0      0 151.0 16.151
freq3  1.161   0      0  14.0  1.713
freq2  1.177   0      0  12.0  1.738
freq1  0.844   0      0  13.0  1.403
```

从中可以看到，按平均水平来说（以算术平均计），每个月每个会员光顾该超市大约 1 次，消费金额为 5—20 元。但这只是一个最基本的描述，我们无法从中看出各个解释性变量同因变量之间的关系。下面，我们首先通过盒状图（如图 6-1 所示）对第 -1 月的光顾频数（freq1）简要分析如下：

```
> boxplot(freq1~freq0,xlab="freq0",ylab="freq1")
```

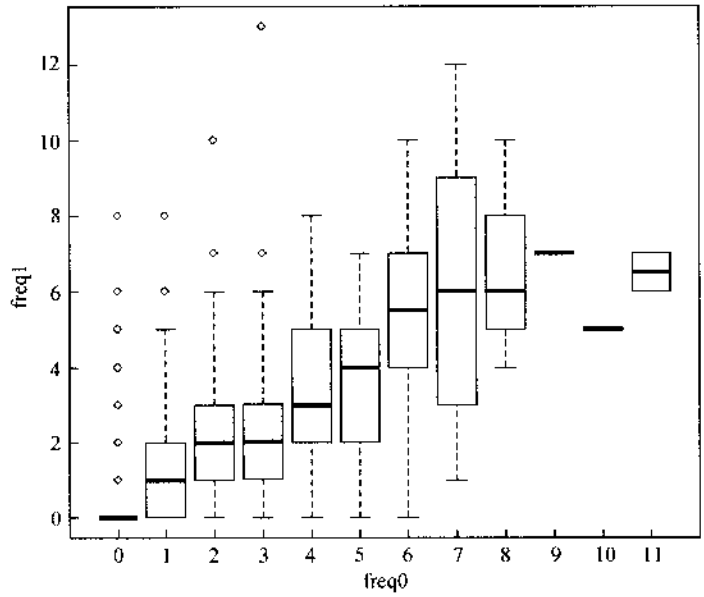


图 6-1　第 -1 月光顾频数盒状图

从图 6-1 可以看到一个明显的趋势，那就是：第 0 月光顾频数高的顾客，其第 -1 月光顾频数（以中位数计）也较高。这暗示我们，消费者在第 -1 月的光顾次数可能是一个对其当月的光顾次数有显著影响的重要变量。

类似地，我们通过盒状图（如图 6-2 所示）对第 -1 月的消费金额（exp1）分

析如下：

```
> boxplot(exp1~freq0,xlab="freq0",ylab="exp1")
```

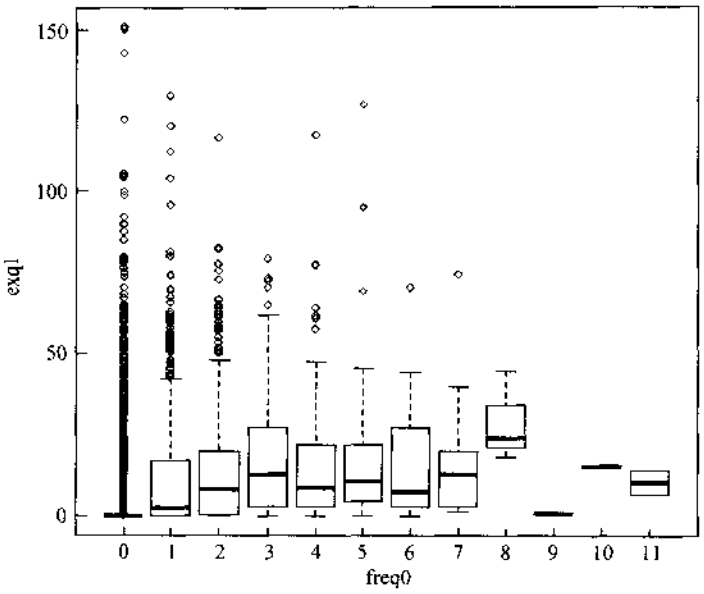


图 6-2　第 -1 月消费金额盒状图

从图 6-1 可以大概看到一个趋势，那就是：第 0 月光顾频数高的顾客，其第 -1 月消费金额（以中位数计）也较高。这暗示我们，消费者在第 -1 月的消费金额也可能是一个对第 0 月光顾次数具有预测能力的变量。但是，由于其变化趋势没有第 -1 月的光顾频数那样明显，因此，我们可以猜测第 -1 月消费金额的预测能力有限。为了便于比较，我们对所有解释性变量作类似的分析并展示如下：

```
> par(mfrow=c(2,3))
> boxplot(freq1~freq0,xlab="freq0",ylab="freq1",main="第-1月")
> boxplot(freq2~freq0,xlab="freq0",ylab="freq2",main="第-2月")
> boxplot(freq3~freq0,xlab="freq0",ylab="freq3",main="第-3月")
> boxplot(exp1~freq0,xlab="freq0",ylab="exp1",main="第-1月")
> boxplot(exp2~freq0,xlab="freq0",ylab="exp2",main="第-2月")
> boxplot(exp3~freq0,xlab="freq0",ylab="exp3",main="第-3月")
> par(mfrow=c(1,1))
```

从图 6-3 可以看到，第 0 月光顾频数高的顾客在第 -2 月和第 -3 月的光顾频数也较高，但这种趋势在第 0 月的光顾频数较高时并不稳定。这表明，第 -2 月和第 -3 月的光顾频数可能对第 0 月的光顾频数有一定的预测能力，但预测能力很有限。第 0 月光顾频数高的顾客在第 -2 月和第 -3 月的消费金额（以中位数计）也较高。第 -2 月的趋势和第 -1 月类似，而第 -3 月的趋势要强于

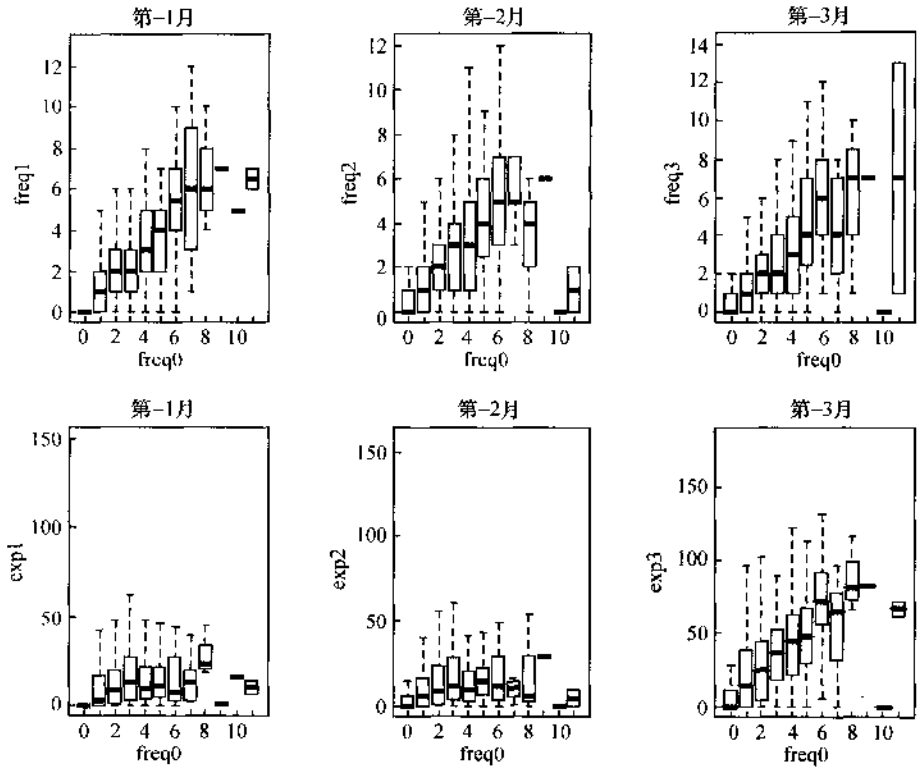


图 6-3 所有解释性变量盒状图

第-1月和第-2月,但这种趋势在第0月的光顾频数较高时也不稳定。这也表明,第-2月和第-3月的消费金额可能对第0月的光顾频数有一定的预测能力,但其预测能力也比较弱。

第三节 泊松回归

和以前一样,在正式介绍泊松回归之前,我们首先要说明为什么需要泊松回归。也就是说,为什么我们前面所讲的线性回归、0-1变量回归以及定序回归都不能够分析计数数据。0-1变量回归和定序回归不能够分析计数数据是因为这两种回归模型处理的都是没有数值意义的数据。但是,计数数据确实是有数值意义的,例如:3次和2次之间差1次,这个差距同2次和1次之间的差距是一样的。因此,我们需要一种能够处理具有数值意义数据的回归模型。那么,

为什么线性回归模型不可以呢？如果我们采用普通的线性回归模型，那么模型形式应该如下：

$$\text{freq0} = \beta_0 + \beta_1 \times \text{freq1} + \beta_2 \times \text{freq2} + \beta_3 \times \text{freq3} + \beta_4 \times \text{exp1} + \beta_5 \times \text{exp2} + \beta_6 \times \text{exp3} + \varepsilon$$

请注意，虽然等号左右两边都是具有数值意义的实数，但是等号的右边可以是小数（如0.12），而等号的左边却是非负的整数。因此，线性模型不适用。那么，我们应该怎样回归计数数据呢？

首先考虑一个高度简化的情形，那就是：先暂时不考虑协变量的影响，单纯考虑计数数据应该如何拟合。换句话说，有什么样的统计分布可以描述顾客每月光顾超市的次数？请注意，顾客光顾超市的次数和他两次光顾之间的时间间隔是高度相关的。如果一个客户两次光顾之间的时间间隔很短，那么他每月的光顾次数自然就高。反过来，如果一个客户两次光顾之间的时间间隔很长，那么他每月的光顾次数自然就低。因此，从统计学上描述每月顾客光顾超市的次数等同于描述该顾客两次光顾之间的时间间隔。假设该顾客的行为是非常稳定的，那么，我们对他的行为能够做出什么其他合理的假设呢？

第一，由于该顾客的“行为稳定”，我们可以假设他在任何相等的时间间隔内的平均光顾次数是一样的。通俗地讲，他每年4月份的平均光顾次数不会高于每年5月份的平均光顾次数。

第二，任给的两次等待时间是互相独立的。换句话说，该顾客前一次等待时间的长短不会影响到他的下一次等待时间。

第三，给定一个时间起点并经过了一段时间以后，我们发现该顾客还没有光顾超市。那么，会不会是因为我们已经等待了一段时间，所以我们就预期该顾客会很快到来呢？答案是否定的。换句话说，无论超市已经等待了多长时间，只要顾客没有光顾，那么超市预期的未来等待时间不会有任何改变。

以上三个假设都是很强的，有合理的一面，但显然也有牵强的一面。而这正是我们研究计数数据的一个很好的出发点。下面，我们就从这三个假设出发，寻找满足这三个假设的统计分布。奇妙的是，我们可以证明同时满足这三个假设的统计分布只有一个，那就是泊松分布。根据泊松分布，我们知道顾客某月光顾超市 $k(k \geq 0)$ 次的概率为：

$$P(\text{freq0} = k) = \frac{\lambda^k}{k!}\exp\{-\lambda\}$$

其中，$\lambda = E(\text{freq0})$ 是顾客平均每月光顾超市的次数。请注意，虽然某月光顾超市的实际次数 k 是整数，但是平均每月光顾超市的次数 λ 却完全可以是小数（当然 λ 是正的）。

大家可以比较一下，如果我们想要确定一个正态分布，我们必须知道它的均值与方差，总共两个参数。但是，泊松分布的概率密度函数告诉我们，要想确定一个泊松分布，我们只需要知道 λ 一个参数就可以了。这进一步告诉我们，与其研究某些协变量同泊松型因变量的关系，不如研究协变量同 λ 之间的关系。而且，同原始因变量 freq0 相比，λ 更容易建模，因为 λ 是连续的。那么，我们可不可以直接考虑线性模型呢？例如：

$$\lambda(x) = \beta_0 + \beta_1 \times \text{freq1} + \beta_2 \times \text{freq2} + \beta_3 \times \text{freq3} + \beta_4 \times \text{exp1} + \beta_5 \times \text{exp2} + \beta_6 \times \text{exp3} + \varepsilon$$

其中，x 代表了所有协变量的信息。请注意，虽然等号两边都可以是小数，但是等号的右边可以是负数，而等号的左边却必须是正数。因此，简单的线性模型不能满足要求。另一方面，这也提示我们如果能够对 $\lambda(x)$ 进行某种变换，使得变换后的 $\lambda(x)$ 可以取任意值，那么线性模型就适用了。什么样的变换能够把正实数变成普通实数呢？当然是对数变换，即：

$$\log\{\lambda(x)\} = \beta_0 + \beta_1 \times \text{freq1} + \beta_2 \times \text{freq2} + \beta_3 \times \text{freq3} + \beta_4 \times \text{exp1} + \beta_5 \times \text{exp2} + \beta_6 \times \text{exp3} + \varepsilon$$

这就是我们本章要详细讨论的泊松回归模型。

第四节　参数估计与统计推断

下面我们再讨论如何对泊松回归的参数予以估计。和以前一样，如果上帝告诉我们 $\log\{\lambda(x)\}$ 的具体取值，那么我们就可以通过第一章中的最小二乘法来估计参数的取值，并进而作统计推断。但是，现实中 $\log\{\lambda(x)\}$ 的取值是未知的，因此我们必须考虑用其他方法来估计我们感兴趣的参数。为了严格起见，假设 (k_i, x_i) 是来自于第 i 个 $(i = 1, \cdots, n)$ 样本的观测。其中，k_i 是因变量（如 freq0），而 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ 是相应的 p 个协变量（如 freq1、freq2、freq3、exp1、exp2、exp3）。根据假定的模型，我们可以观测到该样本的概率为：

$$\frac{\lambda(x_i)^{k_i}}{k_i!}\exp\{-\lambda(x_i)\}$$

其中，$\log\{\lambda(x_i)\} = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \cdots + \beta_p \times x_{ip}$。因此，我们可以计算整个样本的似然函数如下：

$$L(\beta_0, \beta) = \prod_{i=1}^{n} \frac{\lambda(x_i)^{k_i}}{k_i!}\exp\{-\lambda(x_i)\}$$

其中，$\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$。理论上讲，合理的参数估计应该能够产生较大的似

然函数 $L(\beta)$ 取值。因此，我们可以通过极大化 $\log\{L(\beta_0,\beta)\}$，即：

$$\log\{L(\beta_0,\beta)\} = \sum_{i=1}^{n}[k_i\log\{\lambda(x_i)\} - \log(k_i!) - \lambda(x_i)]$$

来获得参数估计。我们称此估计为极大似然估计，并记为 $\hat{\beta}_0$ 和 $\hat{\beta} = (\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p)'$。同 0-1 变量回归模型类似，泊松回归模型中 $\hat{\beta}$ 的真实分布是不知道的。但是，我们知道如果样本量足够大，根据中心极限定理，$\hat{\beta}$ 将近似地服从正态分布。具体如下：

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma(\hat{\beta}_j)} \sim N(0,1) \quad (j = 0,\cdots,p)$$

这说明，只要我们能够准确地估计 $\hat{\beta}_j$ 的标准差 $\sqrt{\mathrm{var}(\hat{\beta}_j)} = \sigma(\hat{\beta}_j)$，那么我们就可以构造如下检验统计量：

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

在原假设 $\beta_j = 0$ 成立的情况下，该统计量 T_j 近似地服从标准正态分布。因此，对于一个给定的显著性水平（如 0.05），我们就可以根据 T_j 的绝对值是否大于 $z_{0.975}$ 来决定是否拒绝原假设。上面所介绍的检验方法只能分别对各个因素的显著性进行检验，而不能同时检验很多因素的显著性，如模型的整体显著性（$\beta = 0$）。为了解决这个问题，我们可以考虑使用似然比检验。其检验统计量如下：

$$\begin{aligned}\lambda &= -2\times(\max_{\beta_0}\log\{L(\beta_0,\beta=0)\} - \max_{(\beta_0,\beta)}\log\{L(\beta_0,\beta)\}) \\ &= (-2\times\max_{\beta_0}\log\{L(\beta_0,\beta=0)\}) - (-2\times\max_{(\beta_0,\beta)}\log\{L(\beta_0,\beta)\})\end{aligned}$$

在这里，β 是一个长度为 p 的向量。经典的统计理论告诉我们，只要原假设 $\beta = 0$ 成立，而且样本量足够大，则 λ 近似服从一个自由度为 p 的 χ^2 分布。和 0-1 变量回归以及定序回归一样，我们称 $(-2\times\log\{L(\beta_0,\beta)\})$ 为 deviance。在 R 中，我们可以具体实现如下：

```
> pos0=glm(freq0~1,family=poisson())
> pos1=glm(freq0~freq1+freq2+freq3+exp1+exp2+exp3,family=poisson())
> anova(pos0,pos1)
Analysis of Deviance Table

Model 1: freq0 ~ 1
Model 2: freq0 ~ freq1 + freq2 + freq3 + exp1 + exp2 + exp3
  Resid. Df Resid. Dev   Df Deviance
1      3994     6755.7
2      3988     4609.7    6   2146.0
```

可以看到，似然比检验统计量为两个模型的 deviance 之差，即 2 146.0。在原假设成立的情况下，它应该服从一个自由度为 6 的卡方分布。因此，我们可以计算模型的整体显著性水平如下：

```
> 1-pchisq(2146.0,df=6)
[1] 0
```

这说明，该模型整体高度显著，也就意味着我们所考虑的六个解释性变量中，至少有一个同因变量显著相关。但是，到底是哪一个解释性变量同因变量相关以及相关程度如何，我们就无从知晓了。因此，我们再作方差分析如下：

```
> library(car)
> Anova(pos1,type="III")
Anova Table (Type III tests)

Response: freq0
      LR Chisq Df Pr(>Chisq)
freq1  181.967  1  < 2.2e-16 ***
freq2   46.473  1  9.291e-12 ***
freq3   71.680  1  < 2.2e-16 ***
exp1    51.487  1  7.206e-13 ***
exp2    25.669  1  4.053e-07 ***
exp3    12.839  1  0.0003395 ***
```

从中可知，所有因素都高度显著。我们再对其具体的参数估计分析如下：

```
> summary(pos1)

Call:
glm(formula = freq0 ~ freq1 + freq2 + freq3 + exp1 + exp2 + exp3,
    family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4063  -0.9075  -0.8005   0.4613   5.5257

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.1382020  0.0305244 -37.288  < 2e-16 ***
freq1        0.1791854  0.0131714  13.604  < 2e-16 ***
freq2        0.0733194  0.0106376   6.892 5.48e-12 ***
freq3        0.0914541  0.0104425   8.758  < 2e-16 ***
exp1         0.0068873  0.0009058   7.604 2.88e-14 ***
exp2         0.0051245  0.0009771   5.245 1.56e-07 ***
exp3         0.0032021  0.0008781   3.647 0.000266 ***
---
```

从上面的报表中，我们可以得到以下三个重要的结论：

第一，六个解释性变量都和因变量高度正相关。这说明，无论是较高的历史光顾频率（无论哪一个月），还是较高的历史消费金额，都预示着更高的未来光顾频率。

第二，这六个解释性变量的影响力很不一样。消费金额的系数非常小，说明它们对未来光顾频率的预测能力弱于历史光顾频率的预测能力。

第三,同样是历史光顾频率,但最近一个月的光顾频率(freq1)的系数明显高于另外两个月(freq2 和 freq3),类似的现象也存在于消费金额中。这说明,消费者最近一个月的行为对未来的预测能力最强。

第五节　模型选择与预测

虽然上面的分析表明所有的解释性变量都高度显著,但是为了理论的完整性,我们还是要详细演示一下如何对泊松回归作变量选择。同 0-1 变量回归以及定序回归非常类似,我们可以定义 AIC 和 BIC 如下:

$$\text{AIC} = \text{deviance} + 2 \times df$$

$$\text{BIC} = \text{deviance} + \log(n) \times df$$

其中,n 代表了样本量,而 df 代表了自由度(即自由参数的个数)。*R* 可以自动地、尽量多地根据 AIC 搜索最优模型如下:

```
> pos.aic=step(pos1,trace=F)
> summary(pos.aic)

Call:
glm(formula = freq0 ~ freq1 + freq2 + freq3 + exp1 + exp2 + exp3,
    family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4063  -0.9075  -0.8005   0.4613   5.5257

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.1382020  0.0305244 -37.288  < 2e-16 ***
freq1        0.1791854  0.0131714  13.604  < 2e-16 ***
freq2        0.0733194  0.0106376   6.892 5.48e-12 ***
freq3        0.0914541  0.0104425   8.758  < 2e-16 ***
exp1         0.0068873  0.0009058   7.604 2.88e-14 ***
exp2         0.0051245  0.0009771   5.245 1.56e-07 ***
exp3         0.0032021  0.0008781   3.647 0.000266 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R 也可以自动地、尽量多地根据 BIC 搜索最优模型如下:

```
> pos.bic=step(pos1,trace=F,k=log(length(a[,1])))
> summary(pos.bic)

Call:
glm(formula = freq0 ~ freq1 + freq2 + freq3 + exp1 + exp2 + exp3,
    family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4063  -0.9075  -0.8005   0.4613   5.5257

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.1382020  0.0305244 -37.288  < 2e-16 ***
freq1        0.1791854  0.0131714  13.604  < 2e-16 ***
freq2        0.0733194  0.0106376   6.892 5.48e-12 ***
freq3        0.0914541  0.0104425   8.758  < 2e-16 ***
exp1         0.0068873  0.0009058   7.604 2.88e-14 ***
exp2         0.0051245  0.0009771   5.245 1.56e-07 ***
exp3         0.0032021  0.0008781   3.647 0.000266 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

对于本案例，AIC 和 BIC 得到了相同的结论。它们都认为全模型就是最优的模型。假设这时候我们有一个新的观测值 $x_0 = (x_{01}, x_{01}, \cdots, x_{0p})'$。那么，我们就可以对该顾客的平均光顾次数估计如下：

$$\lambda(x_i) = \exp(\beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \cdots + \beta_p \times x_{ip})$$

并以此作为对真实光顾频数的估计。举例说明，假设我们有如下检验样本：

```
> a0=read.csv("D:/Practical Business Data Analysis/data/CH6/new.csv")
> a0[c(1:5),]
  exp3 exp2 exp1 freq3 freq2 freq1 freq0
1  0.9  0.0  0.0     1     0     0     0
2  0.0  0.0  0.0     0     0     0     0
3  0.0  0.0  1.5     0     0     1     2
4  0.0 26.2  8.6     0     3     3     3
5 13.6  5.3 26.1     8     4     1     2
```

以第五个样本为例。这位顾客在过去三个月的光顾频数分别为 1、4、8。在这三个月中，他每月的平均消费金额分别为 26.1 元、5.3 元、13.6 元。那么，我们根据全模型的参数估计，可以计算出以下的线性组合：

$$-1.1382020 + 0.1791854 \times 1 + 0.0733194 \times 4 + 0.0914541 \times 8$$
$$0.0068873 \times 26.1 + 0.0051245 \times 5.3 + 0.0032021 \times 13.6 = 0.3163$$

由此，我们可以大概估计该客户的平均光顾频率为 $\exp(0.3163) = 1.37$，而该客户的真实光顾频率为 2。因此，对该客户的预测精度良好。上面烦琐的计算过程可以在 *R* 中自动实现如下：

```
> a0=read.csv("D:/Practical Business Data Analysis/case/CH6/new.csv")
> a0$lam=exp.predict(pos.bic,a0))
> a0[c(1:5),]
  exp3 exp2 exp1 freq3 freq2 freq1 freq0       lam
1  0.9  0.0  0.0     1     0     0     0 0.3520939
2  0.0  0.0  0.0     0     0     0     0 0.3203946
3  0.0  0.0  1.5     0     0     1     2 0.3872497
4  0.0 26.2  8.6     0     3     3     3 0.8292830
5 13.6  5.3 26.1     8     4     1     2 1.3721265
```

如果我们用 k_i^{true} 和 k_i^{pred} 来分别代表来自第 i 个检验样本的真实以及预测的光顾频率，那么我们可以粗略地对预测精度评估如下：

$$\text{绝对预测误差} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(k_i^{\text{true}} - k_i^{\text{pred}})^2}$$

其中，m 是检验样本的个数。在 R 中可以实现如下：

```
> sqrt(mean((a0$freq0-a0$lam)^2))
[1] 1.344686
```

这也从某个角度度量了我们的预测误差。

第六节　简单分析报告

超市会员光顾频数预测分析报告

内容提要　研究消费者光顾超市的频数对于超市的管理者有着特别重要的意义。本报告利用消费者光顾超市的历史数据，对其未来的光顾频数进行了预测分析。我们的分析结果表明，消费者在前三个月的光顾频数和消费金额都影响着其未来的光顾频数，其中前三个月的光顾频数对未来的光顾频数的影响要大于消费金额，尤其是消费者最近一个月的行为对未来的预测能力最强。根据本报告的模型及结论，超市的管理者可以更好地衡量顾客的活跃程度，度量客户的价值，从而制定更为科学、合理的经营管理策略。

一、研究目的

在超市的经营管理中，消费者光顾超市的次数是一个衡量消费者活跃程度、度量客户价值的重要指标。因此，合理地预测消费者的光顾次数，并且知道什么样的消费者光顾超市会频繁一些，而什么样的消费者较少光顾超市，有助于超市的管理者更好地制定相应的经营管理策略。本报告试图利用超市会员的历史消费数据，找出影响消费者未来光顾频率的因素，并建立合理的经济计量学模型对其未来光顾频率进行预测，最后根据分析结果提出合理的建议。

二、数据来源和相关说明

本报告的数据来源于我国北方某城市处于垄断地位的一家超市，数据包含了该超市一部分会员的详细消费记录。为了便于分析，我们以某年某月为基准月份（第0月），将前一个月记为第-1月，以此类推。在数据中，我们感兴趣的因变量是每一个会员在基准月份光顾该超市的次数，对应的自变量是每一个会员在前三个月中每月的光顾次数以及每月的消费金额。特别地，如果该会员在某月没有光顾超市，那么将他的消费金额记为0。具体地说，本报告所使用的数据共包含了3995个有效样本，而对于相关变量的详细解释如表6-2所示。

表6-2 变量说明

变量名称	作用	实际意义
freq0	因变量	第0月光顾超市的频数
freq1	自变量	第-1月光顾超市的频数
freq2	自变量	第-2月光顾超市的频数
freq3	自变量	第-3月光顾超市的频数
exp1	自变量	第-1月的消费金额
exp2	自变量	第-2月的消费金额
exp3	自变量	第-3月的消费金额

三、描述性分析

为了获得对数据的整体认识，我们首先对数据作简单的描述性分析，得到的结果如表6-3所示。从中可以看到，按平均水平来说（以算术平均计），每个月每位会员光顾该超市大约1次，消费金额为5—20元。

表6-3 简单描述统计

变量名	均值	最小值	中位数	最大值	标准差
exp3	16.483	0	0	184.9	24.919
exp2	9.241	0	0	159.1	16.621
exp1	7.917	0	0	151.0	16.151
freq3	1.161	0	0	14.0	1.713
freq2	1.177	0	0	12.0	1.738
freq1	0.844	0	0	13.0	1.403

考虑到因变量为计数变量的特点，我们利用盒状图考察因变量与自变量之间的关系，得到图6-4。

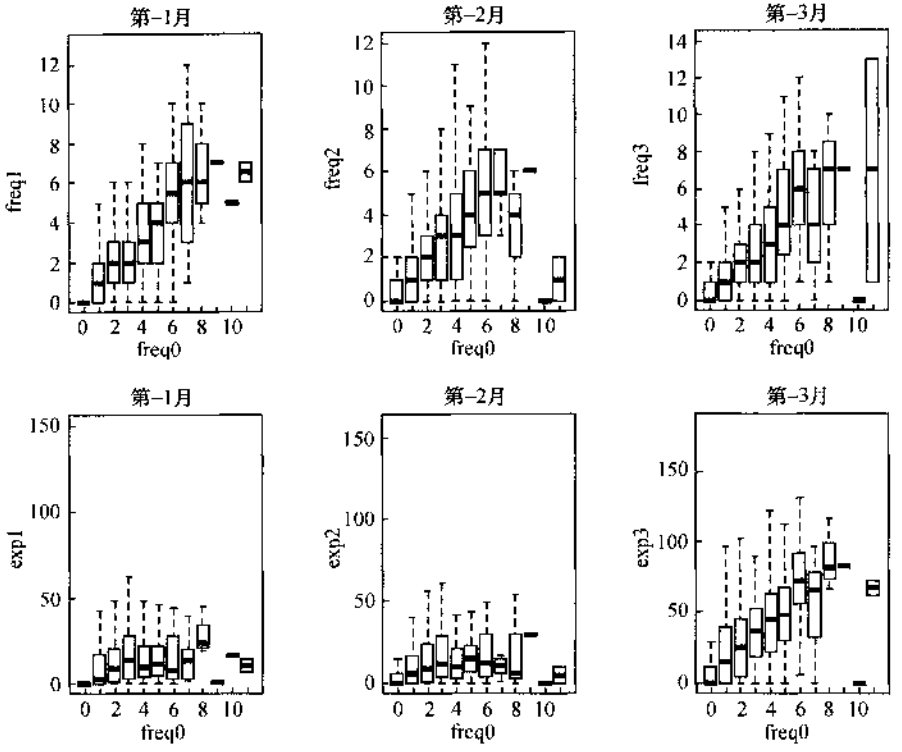


图 6-4　所有解释性变量盒状图

从图 6-4 我们可以得到如下的初步结论：

• 第 0 月光顾频数高的顾客，其第 -1 月光顾频数（以中位数计）也较高。这表明消费者在第 -1 月的光顾次数可能是一个对其当月的光顾次数有显著影响的重要变量。

• 第 0 月光顾频数高的顾客，其第 -1 月消费金额（以中位数计）也较高。这表明消费者在第 -1 月的消费金额也可能是一个对第 0 月光顾次数具有预测能力的变量。但是，其变化趋势没有第 -1 月的光顾频数那样明显，因此，我们可以猜测第 -1 月消费金额的预测能力有限。

• 第 0 月光顾频数高的顾客在第 -2 月和第 -3 月的光顾频数也较高，但这种趋势在第 0 月的光顾频数较高时并不稳定。这表明，第 -2 月和第 -3 月的光顾频数可能有一定的预测能力，但预测能力很有限。

• 第 0 月光顾频数高的顾客在第 -2 月和第 -3 月的消费金额（以中位数计）也较高。第 -2 月的趋势和第 -1 月类似，而第 -3 月的趋势要强于第 -1 月和第 -2 月，但这种趋势在第 0 月的光顾频数较高时也不稳定。这也表明，第

-2 月和第 -3 月的消费金额可能有一定的预测能力，但预测能力也比较弱。

• 第 -2 月和第 -3 月的光顾频数也较高，但这种趋势在第 0 月的光顾频数较高时并不稳定。这表明，第 -2 月和第 -3 月的光顾频数可能有一定的预测能力，但预测能力很有限。

四、数据建模

1. 全模型分析

根据因变量为计数变量的特点，我们采用泊松（Poisson）回归的方法建立模型。首先对包含全部自变量的全模型进行估计，得到的估计结果如表 6-4 所示：

表 6-4 全模型

变量名	系数估计值	标准差	P 值
截距项	-1.138	0.031	<0.001
freq1	0.179	0.013	<0.001
freq2	0.073	0.011	<0.001
freq3	0.091	0.010	<0.001
exp1	0.007	0.001	<0.001
exp2	0.005	0.001	<0.001
exp3	0.003	0.001	<0.001
模型显著性检验 P 值 <0.001			

从表 6-4 中我们可以看到，对模型显著性的广义似然比检验的 P 值非常小（P 值 <0.001），表明该模型是显著的，即我们所考虑的六个解释性变量中，至少有一个与第 0 月消费者的光顾频数显著相关。进一步通过对各自变量所对应的 Z 检验的考察，我们可以得到以下三个结论：

• 六个解释性变量都和因变量高度正相关。这说明，无论是较高的历史光顾频率（无论哪一个月），还是较高的历史消费金额，都预示着更高的未来光顾频率。

• 这六个解释性变量的影响力很不一样。消费金额的系数非常小，这在一定程度上说明，其对未来光顾频率的预测能力弱于历史光顾频率的预测能力。

• 同样是历史光顾频率，最近一个月的光顾频率（freq1）的系数明显高于另外两个月（freq2 和 freq3），类似的现象也存在于消费金额中。这说明，消费者最近一个月的行为对未来的预测能力最强。

2. 模型选择与预测

上一节的分析结果表明，我们所选取的自变量确实对因变量有一定的解释能力。但是全模型过于复杂，为了得到一个尽量简单同时又具有良好的预测能

力的模型，我们采用 AIC 和 BIC 的模型选择标准来选择一个最优的模型。对于本报告所使用的数据，AIC 和 BIC 得到了相同的结论，它们都认为全模型就是最优的模型。这表明，这六个自变量确实都有一定的预测能力，也就是说，前三个月中消费者的购买行为确实影响着当月的购买行为。

根据我们所得到的全模型（同时也是选择的最优模型），我们可以预测消费者当月的光顾频数。通过对检验数据进行预测，并计算预测值与真实值之间的误差，我们得到该模型的绝对预测精度为 1.345。也就是说，其平均的预测误差约为 1.345，这表明该模型有一定的预测能力。

五、结论及建议

从上述的分析结果可知，消费者在前三个月的光顾频数和消费金额都影响着其未来的光顾频数，其中前三个月的光顾频数对未来的光顾频数的影响要大于消费金额。因此，对超市的管理者而言，能够使消费者经常光顾超市对超市的经营具有很重要的意义，而不要太过于关注消费金额。我们还发现消费者最近一个月的行为对未来的预测能力最强，因而超市的管理者要特别关注近期的经营状况，而不是更长期的情形。另外，根据本报告的模型及结论，超市的管理者可以利用会员的历史消费记录，对会员未来的光顾频数进行合理的预测，并根据预测结果来制定相应的营销策略。具体地说，如果预测结果表明消费者未来的光顾频数较小，那么超市就需要通过某些促销活动来增加消费者的光顾次数，进而提高超市的经营业绩。

[讨论总结]

本章以客户关系管理的数据为例，系统演示并讲解了泊松回归这种典型的计数回归模型。通过对本章的学习，读者应该能够了解：什么时候可以使用泊松回归，以及如何使用。在 *R* 语言学习方面，读者应该掌握相关的广义线性模型的命令。在统计理论方面，读者应该掌握计数数据、泊松回归等概念。对相关理论渴望深入了解的读者请参阅 McCullagh and Nelder(1999)。

附录　程序及注释

```
rm(list=ls())                                                    # 清空当前工作空间
a=read.csv("D:/Practical Business Data Analysis/case/CH6/crm.csv")   # 读入 csv 格式的数据,作为训练数据
attach(a)                                                        # 将数据 a 的各个变量加载到工作空间,便于直接调用
a[c(1:5),]                                                       # 显示数据 a 的前 5 行
Mean=sapply(a[,1:6],mean)                                        # 计算数据 a 的第 1 到第 6 列的均值
Min=sapply(a[,1:6],min)                                          # 计算数据 a 的第 1 到第 6 列的最小值
Median=sapply(a[,1:6],median)                                    # 计算数据 a 的第 1 到第 6 列的中位数
Max=sapply(a[,1:6],max)                                          # 计算数据 a 的第 1 到第 6 列的最大值
SD=sapply(a[,1:6],sd)                                            # 计算数据 a 的第 1 到第 6 列的标准差
round(cbind(Mean,Min,Median,Max,SD),3)                           # 将均值、最小值、中位数、最大值、标准差集中在一起展示
boxplot(freq1 ~ freq0,xlab="freq0",ylab="freq1")                 # 根据 freq0 的不同取值来画 freq1 的盒状图
boxplot(exp1 ~ freq0,xlab="freq0",ylab="exp1")                   # 根据 freq0 的不同取值来画 exp1 的盒状图
par(mfrow=c(2,3))                                                # 设置画图模式为 2x3 的格式
boxplot(freq1 ~ freq0,xlab="freq0",ylab="freq1",main="第 - 1 月")  # 根据 freq0 的不同取值来画 freq1 的盒状图
boxplot(freq2 ~ freq0,xlab="freq0",ylab="freq2",main="第 - 2 月")  # 根据 freq0 的不同取值来画 freq2 的盒状图
boxplot(freq3 ~ freq0,xlab="freq0",ylab="freq3",main="第 - 3 月")  # 根据 freq0 的不同取值来画 freq3 的盒状图
boxplot(exp1 ~ freq0,xlab="freq0",ylab="exp1",main="第 - 1 月")    # 根据 freq0 的不同取值来画 exp1 的盒状图
boxplot(exp2 ~ freq0,xlab="freq0",ylab="exp2",main="第 - 2 月")    # 根据 freq0 的不同取值来画 exp2 的盒状图
boxplot(exp3 ~ freq0,xlab="freq0",ylab="exp3",main="第 - 3 月")    # 根据 freq0 的不同取值来画 exp3 的盒状图
par(mfrow=c(1,1))                                                # 设置画图模式为 1x1 的格式
```

```
pos0=glm(freq0 ~ 1,family=poisson())
pos1=glm(freq0 ~ freq1 + freq2 + freq3 + exp1 + exp2 + exp3,family=poisson())
anova(pos0,pos1)

1 - pchisq(2146.0,df=6)
library(car)
Anova(pos1,type="III")
summary(pos1)
pos.aic=step(pos1,trace=F)
summary(pos.aic)
pos.bic=step(pos1,trace=F,k=log(length(a[,1])))
summary(pos.bic)
a0=read.csv("D:/Practical Business Data Analysis/case/CH6/new.csv")
a0[c(1:5),]
a0$lam=exp(predict(pos.bic,a0))

a0[c(1:5),]
sqrt(mean((a0$freq0 - a0$lam)^2))
```

```
# 拟合 poisson 回归模型,不利用任何变量的空模型
# 拟合 possion 回归模型,利用全部变量的全模型
# 为获取不同模型的 deviance,进行模型 pos0 和 pos1 的方差分析
# 计算模型显著性检验的 P 值
# 载入程序包 car
# 对模型 pos1 作三型方差分析
# 显示模型 pos1 的各方面细节,包括参数估计值、P 值等
# 根据 AIC 准则从全模型 pos1 中选出最优子模型 pos.aic
# 显示模型 pos.aic 的各方面细节,包括参数估计值、P 值等
# 根据 BIC 准则从全模型 pos1 中选出最优子模型 pos.bic
# 显示模型 pos.bic 的各方面细节,包括参数估计值、P 值等
# 读入 csv 格式的数据,用于检验
# 显示数据 a0 的前 5 行
# 利用模型 pos.bic 对数据 a0 进行预测,结果存入 a0 的变量 lam 中
# 显示数据 a0 的前 5 行,包括预测值
# 计算绝对预测精度
```

第七章
生存分析模型

- 案例介绍
- 生存函数
- 描述性分析
- 加速死亡模型
- Cox 风险模型
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过一个关于癌症临床数据的实际案例，详细介绍生存分析这种重要的统计工具。它主要处理的是因变量为生存数据的情形。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下使用生存分析；(2) 生存分析的基本统计学理论；(3) 相关理论在统计学软件 *R* 中的应用；(4) 相应的统计分析报告的撰写。本章所涉及的新统计学概念有：生存数据、截断、生存函数、KM 估计、风险函数、加速死亡模型以及 Cox 等比例风险模型。

第一节　案例介绍

从前面的讨论中，大家可能已经注意到了一个现象，那就是：数据类型，特别是因变量的数据类型，在很大程度上决定了相应的统计分析方法！例如，对于连续的因变量，我们可以考虑回归模型、方差分析模型；如果因变量是离散的，我们可以考虑各种各样的广义线性模型（如逻辑回归、泊松回归等）。在这一章中，我们将学习另外一种非常常见却又非常特殊的数据类型，那就是生存数据（survival data）。

生存数据描述的是一个个体的“生存”时间。典型的例子是癌症病人的生存时间，也就是该病人从某一观测时点开始，直到死亡的时间。当然，也不仅仅局限于癌症病人的数据。只要我们能够合理地定义“死亡”，那么就会有生存数据出现。比方说，如果我们定义一个公司的破产为“死亡”，那么公司的生存时间就是该公司从成立到破产所需要的时间。又比方说，我们定义一个品牌的消失为“死亡”，那么品牌的生存时间就是该品牌从上市到消失的时间。为什么生存数据那么特殊呢？以癌症病人为例，他从确诊到死亡的时间是一个典型的连续型数据。如果我们关心的问题是：哪些因素（如治疗方案、性别、年龄等）能够影响这个连续型因变量（即生存时间），那么似乎使用典型的回归分析就应该可以回答我们感兴趣的问题。但是，事情没有那么简单！

考虑癌症病人的生存时间，有的病人生存时间很短（如几个月），但也有的生存时间很长（如许多年）。但是，对于研究者来说，我们不可能永远密切地跟踪每一个病人并确切地知道他的生存时间。因此，我们会有一个人为设定的观测区间（如两年），而这个区间的长度很大程度上是由研究计划、经费支持以及

其他客观条件所决定的。那么，对病人的观测就会有两种情况：一是该病人在两年的观测区间内去世（如 13 个月），那么我们就可以非常确切地知道他的生存时间；二是他没有在两年的观测区间内死亡，因而对于这类病人我们无法知道他确切的生存时间，但是我们可以知道他的生存时间一定大于两年。生存分析的数据特点就来源于此，因而相关统计分析所面临的挑战也来源于此。

为了具体说明，我们考虑一个医学方面的案例。Krall、Uthoff 和 Harley（1975）收集并分析了一个关于某致命疾病的生存数据。该数据共包含了 65 个病人的资料，其中 48 人在研究期间死去，而 17 人活过了最后的研究期限。对于每一个病人，研究者收集并研究了表 7-1 中的变量。

表 7-1　变量说明

变量名称	解释意义
Time	从确诊到死亡的生存时间（单位：月）
VStatus	生存状态（0 = 生存；1 = 死亡）
HGB	确诊时血色素含量
Platelet	确诊时血小板状况（0 = 不正常；1 = 正常）
Age	确诊时年龄（单位：年）
LogWBC	对数变换后白细胞含量
LogPBM	对数变换后骨髓中血浆细胞含量
Protein	确诊时血蛋白含量
SCalc	确诊时血清钙含量

其中，我们感兴趣的因变量是生存时间（Time），而其他的变量都为解释性变量。在 *R* 中可以简单地读入数据如下：

```
> a=read.csv("D:/Practical Business Data Analysis/case/CHS/data.csv",header=T)
> a[1:20,]
    Time VStatus  HGB Platelet Age LogWBC LogPBN Protein SCalc
1   4.00       0 10.2        1  59 4.0453 0.7782      12    10
2   4.00       0 10.0        1  49 3.9590 1.6232       0    13
3   7.00       0 12.4        1  48 3.7993 1.8573       0    10
4   7.00       0 10.2        1  81 3.5911 1.8808       0    11
5   8.00       0  9.9        1  57 3.8325 1.6532       0     8
6  12.00       0 11.6        1  46 3.6435 1.1461       0     7
7  11.00       0 14.0        1  60 3.7324 1.8451       3     9
8  12.00       0  8.8        1  66 3.6388 1.3617       0     9
9  13.00       0  4.9        0  71 3.6435 1.7924       0     9
10 16.00       0 13.0        1  55 3.8573 0.9031       0     9
11 19.00       0 13.0        1  59 3.7709 2.0000       1    10
12 19.00       0 10.8        1  69 3.8808 1.5185       0    10
13 28.00       0  7.3        1  82 3.7482 1.6721       0     9
14 41.00       0 12.8        1  72 3.7243 1.4472       1     9
15 53.00       0 12.0        1  66 3.6128 2.0000       1    11
16 57.00       0 12.5        1  66 3.9685 1.9542       0    11
17 77.00       0 14.0        1  60 3.6812 0.9542       0    12
18  1.25       1  9.4        1  67 3.6628 1.9542      12    10
19  1.25       1 12.0        1  38 3.9868 1.9542      20    18
20  2.00       1  9.8        1  81 3.8751 2.0000       2    15
```

从以上数据中可以看到，我们的病人可以根据其生存状态（VStatus）分为两大类：一类病人是，其生存状态为“死亡”（即 VStatus = 1），如第 20 个病人，他的生存时间确确实实只有 2 个月（因为 Time = 2）；另外一类病人是，其生存状态为“生存”（即 VStatus = 0），如第 1 个病人。这说明什么呢？并非说明该病人被彻底治愈，从此过上了健康的生活。相反，在这种类型的研究中，病人患上的往往是不治之症，在可以预见的未来将会死亡。那么该病人的“生存”是指什么呢？是指在该研究结束之前，没有观测到第 1 个病人的确切死亡时间，因此，我们从数据中仅仅知道他的生存时间要大于 4 个月（因为 Time = 4），但是到底大多少我们是不知道的。对于这种数据，我们称其为右截断（right censored）数据，或者简称为截断数据（censored data）。本章要讨论的所有内容都围绕着该类数据。具体地说，我们要回答两大问题：第一，对于这种数据，应该如何进行描述？第二，对于这种数据，应该如何进行回归？

第二节　生存函数

在传统的描述性分析中，均值占据了一个很重要的位置，它告诉我们一个变量的中心大概位于何处。但是，对于被右截断的生存数据来说，均值是不合理的。因为，大量的截断数据的精确值是不知道的。以我们的数据为例，对于第一个病人，我们只知道他的生存时间大于 4 个月，但是到底几个月确实是不知道的。如果简单地将其生存时间设为 4 个月，会严重地低估真实的平均生存

时间。因此，如何合理有效地描述生存时间，本身就成了一个很有挑战性的难题。这就不可避免地引出了“生存函数”的概念。

具体地说，假设 t_i 是第 i 个个体的真实生存时间，那么它的分布就可以用下面这个简单的生存函数来描述：

$$S(t) = P(t_i > t)$$

从理论上讲，如果我们能够知道该生存函数，我们就可以获得所有关于生存时间 t_i 的有效信息，如均值、中位数、标准差等。在现实中，我们并不知道该函数的具体取值，但是却有可能在一定的范围内获得可靠的估计，因此这将是我们用来描述生存数据的一个主要工具。对于一个给定的 x，我们应该怎样估计 $S(x)$ 呢？如果我们的数据没有被截断，也就是说，对于任意的一个个体，我们都观测到了 t_i 的具体取值，那么 $S(x)$ 的取值可以简单地估计如下：

$$\tilde{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\}$$

但是，对于右截断的生存数据，我们应该怎样估计呢？

在进行详细讨论之前，我们需要定义一些符号。对于给定的个体 i，如果其生存时间没有被截断，我们记为 t_i；如果被截断了，我们记为 t_i^+。这说明该个体的生存时间大于 t_i，但是大多少不知道。在 R 中，可以描述如下：

```
> library(survival)
> a=a[order(a$Time),]
> Surv(a$Time,a$VStatus)
 [1]  1.25   1.25   2.00   2.00   2.00   3.00   4.00+  4.00+  5.00   5.00
[11]  6.00   6.00   6.00   6.00   7.00+  7.00+  7.00   7.00   7.00   8.00+
[21]  9.00  11.00+ 11.00  11.00  11.00  11.00  11.00  12.00+ 12.00+ 13.00+
[31] 13.00  14.00  15.00  16.00+ 16.00  16.00  17.00  17.00  18.00  19.00+
[41] 19.00+ 19.00  19.00  24.00  25.00  26.00  28.00+ 32.00  35.00  37.00
[51] 41.00+ 41.00  41.00  51.00  52.00  53.00+ 54.00  57.00+ 58.00  66.00
[61] 67.00  77.00+ 88.00  89.00  92.00
```

下面，我们再定义两个概念——在险者个数（number of cases at risk）以及事件个数（number of events）。对于一个给定的时间点 t，在险者个数 $r(t)$ 是指样本中确知的存活时间至少为 t 的样本个数。例如，在我们的案例中 $r(0)=65$，即样本容量，这是因为所有的个体都至少存活了 0 个月。类似地，我们知道 $r(1.25)=65$，因为所有的样本都至少存活了 1.25 个月。同样地，我们知道 $r(89)=2$，

那么，什么是事件个数 $d(t)$ 呢？它是指在给定时间点 t 死亡的个体数。由于我们只能够观测到有限个样本的死亡点，所以对一个任意的 t，$d(t)$ 一般取值为零。但是，如果 t 是观测到的某一个死亡点，如 $t=1.25$，那么 $d(1.25)=2$，因为我们知道样本中有两个病人在 1.25 个月的时候死亡。如果 t 取值为一个被截断的生存时间，如 $t=4.0$，那么我们仍然有 $d(t)=0$，因为在该时间点上，我们没有观测到任何死亡事件发生。

对于一个给定的时间点 t，如果我们知道在险者数目 $r(t)$ 以及事件数目 $d(t)$，那么我们就可以估计条件概率：

$$P(t_i > t \mid t_i \geq t) \approx 1 - \frac{d(t)}{r(t)}$$

正如前面所提到的，在大量的时间点上（如在本案例中 $t = 2.1$ 个月）我们都没有观测到任何死亡事件，所以一般来说 $d(t) = 0$，因此估计 $P(t_i > t \mid t_i \geq t) \approx 1$，也就是没有任何分布上的变化。所以，我们只需要关心 $d(t) \neq 0$，也就是那些观测到的并且没有被截断的时间点。我们简单地将此集合记为 E。在我们的案例中，此集合为：

```
> summary(survfit(Surv(a$Time,a$VStatus)))
Call: survfit(formula = Surv(a$Time, a$VStatus))

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
  1.25     65       2   0.9692  0.0214       0.92814        1.000
  2.00     63       3   0.9231  0.0331       0.86052        0.990
  3.00     60       1   0.9077  0.0359       0.83998        0.981
  5.00     57       2   0.8758  0.0411       0.79888        0.960
  6.00     55       4   0.8121  0.0489       0.72171        0.914
  7.00     51       3   0.7644  0.0533       0.66681        0.876
  9.00     45       1   0.7474  0.0547       0.64749        0.863
 11.00     44       5   0.6625  0.0603       0.55429        0.792
 13.00     36       1   0.6441  0.0613       0.53441        0.776
 14.00     34       1   0.6251  0.0624       0.51406        0.760
 15.00     33       1   0.6062  0.0633       0.49398        0.744
 16.00     32       2   0.5683  0.0648       0.45453        0.711
 17.00     29       2   0.5291  0.0660       0.41439        0.676
 18.00     27       1   0.5095  0.0664       0.39470        0.658
 19.00     26       2   0.4703  0.0668       0.35603        0.621
 24.00     22       1   0.4489  0.0671       0.33493        0.602
 25.00     21       1   0.4275  0.0672       0.31417        0.582
 26.00     20       1   0.4062  0.0672       0.29373        0.562
 32.00     18       1   0.3836  0.0671       0.27224        0.541
 35.00     17       1   0.3610  0.0669       0.25115        0.519
 37.00     16       1   0.3385  0.0664       0.23046        0.497
 41.00     15       2   0.2933  0.0647       0.19033        0.452
 51.00     12       1   0.2689  0.0638       0.16890        0.428
 52.00     11       1   0.2445  0.0625       0.14810        0.404
 54.00      9       1   0.2173  0.0612       0.12514        0.377
 58.00      7       1   0.1863  0.0598       0.09927        0.349
 66.00      6       1   0.1552  0.0573       0.07525        0.320
 67.00      5       1   0.1242  0.0536       0.05327        0.289
 88.00      3       1   0.0828  0.0492       0.02583        0.265
 89.00      2       1   0.0414  0.0382       0.00677        0.253
 92.00      1       1   0.0000      NA            NA           NA
```

请注意，这里面没有生存时间为 4 个月的观测，这是因为这种样本都被右截断了。那么对于任给的一个时间点 t，可以作如下分解：

$$S(t) = P(t_0 > t) = \prod_{t_i \leq t, t_i \in E} P(t_0 > t_i \mid t_0 \geq t_i)$$

$$\approx \prod_{t_i \leq t, t_i \in E} \left\{1 - \frac{d(t_i)}{r(t_i)}\right\} = \hat{S}(t_0)$$

等号最右边的就是 Kaplan-Meier 估计（简称 KM 估计）。该估计的方差为：

$$\text{var}\{\hat{S}(t)\} \approx \hat{S}^2(t) \sum_{t_i \leq t, t_i \in E} \frac{d(t_i)}{r(t_i)\{r(t_i) - d(t_i)\}}$$

基于此方差估计量，我们就可以构造生存函数 $\hat{S}(t)$ 的置信区间。在 R 中，我们可以通过 R 中 survfit 函数（如图 7-1 所示）简单描述如下：

```
> plot(survfit(Surv(a$Time,a$Vstatus)))
```

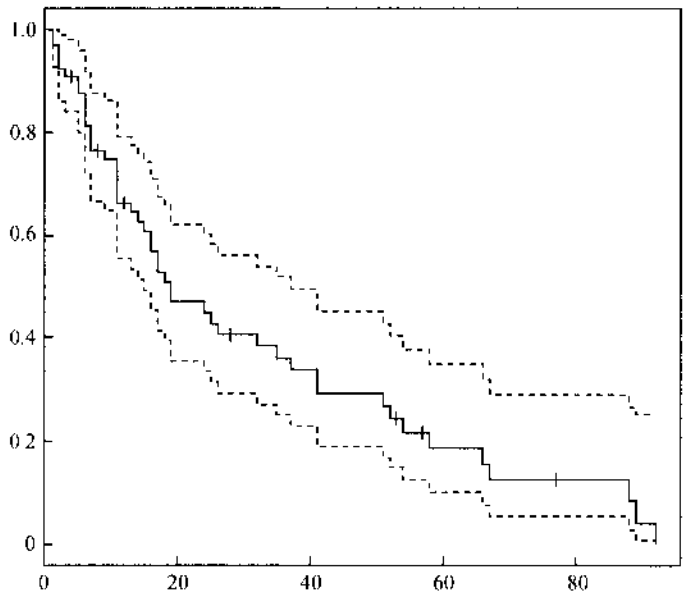


图 7-1　生存函数图

如何分析前面的生存分析报表呢？举例说明：从上表我们可以看到，至少有一个病人的生存期为 24 个月，否则我们不会在以上的输出中看到 24。此外，我们还可以看到，在 $t = 24$ 这个时间点上，只有一个病人死亡，即一个事件（$d(24) = 1$）。同时还可以看到，在我们的样本中，有 22 个病人的生存期大于 24 个月。在这 22 个病人中，有的我们可以知道他的确切生存时间，如有一个病

人的生存期为92个月。但是，也有的病人我们并不知道他的确切生存时间，如有一个病人的生存期为77+。对于这个病人，虽然我们并不知道他的确切生存时间，但是我们知道他的生存时间一定大于24个月。从以上的简单输出中我们可以看到，大约有44.89%的病人，其生存时间大于24个月（即两年）。该估计量的95%的置信区间为(0.33493,0.60200)。我们这里所估计的生存函数，也可以通过图形形象地表示出来。

一般来说，生存曲线越是靠上，说明对于一个给定的时间点，生存时间超过该点的概率越大。因此，生存状况越好。

第三节　描述性分析

有了生存函数，我们就可以对生存数据进行非常详细的描述以及分析。我们首先考虑血色素含量（HGB）对生存函数的影响。由于血色素含量是一个连续变量，因此，我们首先根据血色素含量的平均水平（以中位数计），将所有的样本分成两组。其中一组的血色素水平高于平均水平（HGB > Median），另一组低于平均水平（HGB < Median）。在 *R* 中可以实现如下：

```
> status=1*(a$HGB>median(a$HGB))
> status
 [1] 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 1 1 0 0 1 0 0 0 1 1 1 0 0 0 1
[39] 0 1 1 1 0 1 1 1 0 1 0 1 1 0 0 0 0 1 0 1 1 0 1 1 1 1 1 1
```

由原始数据我们可以验证，HGB的样本中位数为10.2，而第一个样本的血色素水平才9.2，因此，第一个样本取值为0。由于第二个样本的血色素水平为12，因此，第二个样本取值为1。然后，我们可以将两组（即高HGB组以及低HGB组）的生存曲线（如图7-2所示）对比如下：

```
> plot(survfit(Surv(a$Time,a$V8(status)~status),col=c(1,2),
+ lty=c(1,2))
> legend(40,1,c("HGB<Median","HGB>Median"),col=c(1,2),lty=c(1,2))
```

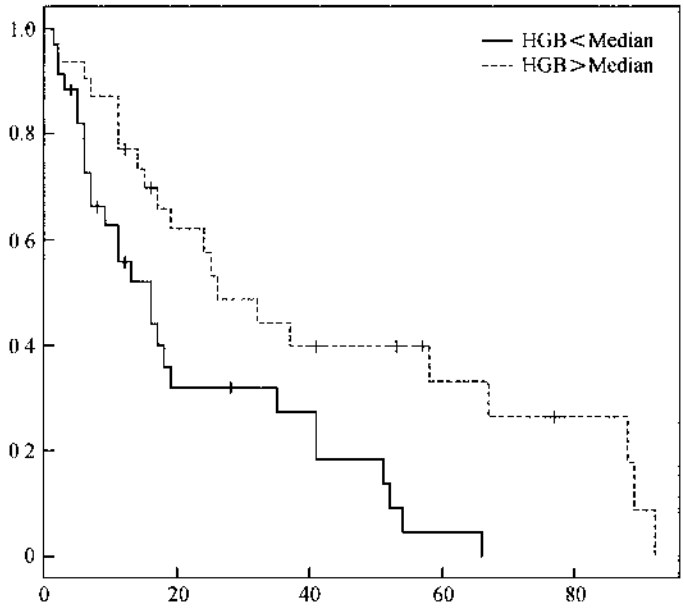


图 7-2　HGB 分组对应的生存函数图

从图 7-2 可以清楚地看到，高 HGB 组的生存曲线（虚线）一直都在低 HGB 组的生存曲线（实线）上方。这说明，对于任意的一个时间点（如 40 个月），高 HGB 组的生存概率都要高于低 HGB 组的生存概率。由此可以初步推断，血色素水平的高低是影响病人生存时间的重要指标。

下面，我们再对血小板水平作类似的分析。由于在原始数据中，血小板水平已经被区分为正常与非正常，因此我们可以直接画图如图 7-3 所示。

```
> plot(survfit(Surv(a$Time,a$VStatus)~a$Platelet),col=c(1,2),
+ lty=c(1,2))
> legend(40,1,c("Abnormal","Normal"),col=c(1,2),lty=c(1,2))
```

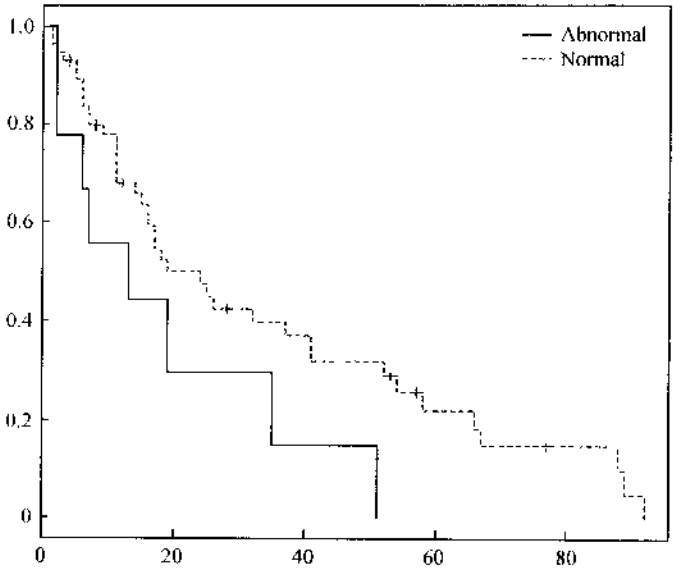


图 7-3　血小板状况对应的生存函数图

从图 7-3 我们可以清楚地看到，具有正常血小板水平的病人的生存概率（虚线）明显地高于具有非正常血小板水平的病人（实线）。

年龄可能是另外一个影响生存概率的重要指标。我们的样本中的年龄分布范围为 38 岁到 82 岁，平均年龄（以中位数计）为 60 岁。因此，我们以 60 岁为界，将病人分成两组，并对比（如图 7-4 所示）。

```
> age.group=1*(a$Age>median(a$Age))
> plot(survfit(Surv(a$Time,a$Status)~age.group),col=c(1,2),lty=c(1,2))
> legend(40,1,c("Age <60","Age>60"),col=c(1,2),lty=c(1,2))
```

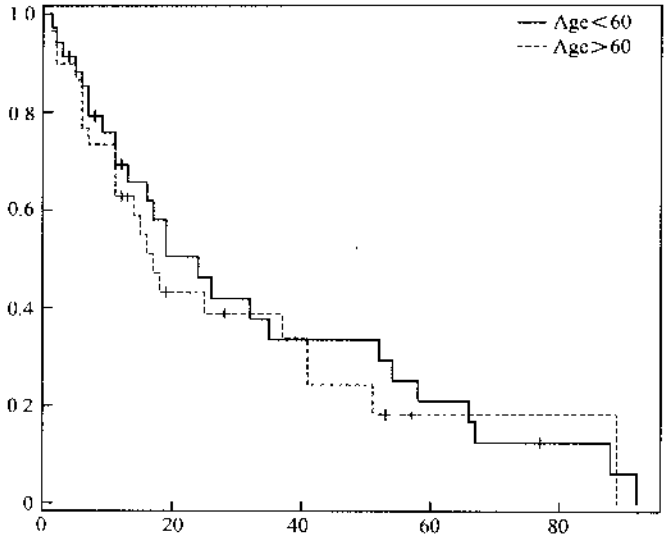


图 7-4　Age 分组对应的生存函数图

从图 7-4 我们可以看到，高年龄组（Age > 60）的生存曲线（虚线）稍微低于低年龄组（Age < 60）的生存曲线（实线），但是差别不大。

最后，我们可以对其他变量也作类似的描述性分析（如图 7-5 所示）。在 *R* 中实现如下：

```
> par(mfrow=c(3,2))
>
> status=1*(a$LogWBC>median(a$LogWBC))
> plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),
+ lty=c(1,2))
> legend(30,1,c("LogWBC<Median","LogWBC>Median"),col=c(1,2),lty=c(1,2))
>
>
> status=1*(a$LogPBM>median(a$LogPBM))
> plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),
+ lty=c(1,2))
> legend(30,1,c("LogPBM<Median","LogPBM>Median"),col=c(1,2),lty=c(1,2))
>
>
> status=1*(a$Protein>median(a$Protein))
> plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),
+ lty=c(1,2))
> legend(30,1,c("Protein<Median","Protein>Median"),col=c(1,2),lty=c(1,2))
>
>
> status=1*(a$SCalc>median(a$SCalc))
> plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),
+ lty=c(1,2))
> legend(30,1,c("SCalc<Median","SCalc>Median"),col=c(1,2),lty=c(1,2))
>
> par(mfrow=c(1,1))
```

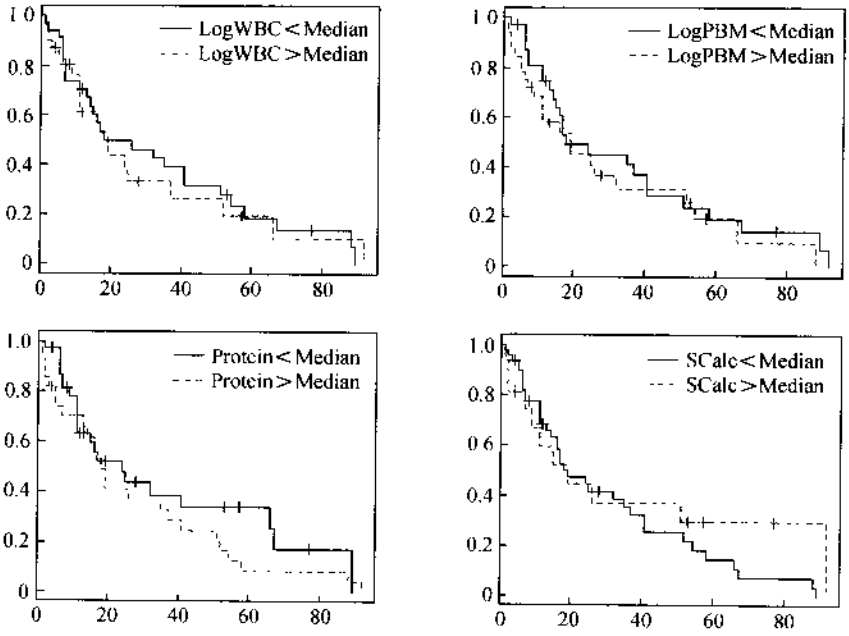


图 7-5 其他变量的生存函数图

从图 7-5 可以看到，白细胞含量（LogWBC）、骨髓中血浆细胞含量（LogPBM）以及血清钙含量（SCalc）的高低都对生存概率影响不大。但是，血蛋白含量（Protein）高的病人的生存概率要低于血蛋白含量低的病人，不过两者差别不大。

第四节 加速死亡模型

结束了对生存数据的详细描述，我们将通过合理的回归模型，进一步研究什么因素在影响生存函数，即如何对生存数据作回归分析。

我们首先考虑一个非常简单的情形：假设我们的观测时间足够长，使得每一个病人的生存时间 t_i 都能够被精确观测到。同时我们还假设，对于第 i 个病人，我们还知道与其生存时间 t_i 相关的解释性变量 $x_i = (x_{i1}, \cdots, x_{ip})'$，那么我们应该怎样量化生存时间 t_i 和解释性变量 x_i 之间的关系呢？首先注意到，生存时间 t_i 是一个永远非负的随机变量，而一般的回归模型要求因变量的可能取值必须在正负无穷之间。因此，有必要对 t_i 作对数变换后，再构造普通的线性模型如下：

$$\log t_i = x_i'\beta + e_i$$

其中，e_i 是服从某种分布的残差项，也可以从概念上认为它是在没有协变量影响（即 $x_i = 0$）的情况下的对数生存时间，记为 $\log t_{i0}$。因此，以上的线性模型可以变形为：

$$\log t_i = x_i'\beta + \log t_{i0} \Rightarrow t_i = t_{i0}\exp(x_i'\beta)$$

这说明，如果没有协变量的影响，该个体的实际生存时间应该为 t_{i0}。但是，由于有了协变量的影响，其生存时间被“加速”为 $t_{i0}\exp(x_i'\beta)$，因此我们称此模型为加速死亡模型。

对于加速死亡模型，我们还有一个重要的问题没有回答，那就是：我们应该对残差 $e_i = \log t_{i0}$ 作什么样的分布假设呢？最自然的假设是正态分布，即：

$$e_i = \log t_{i0} \sim N(0, \sigma^2)$$

但是在生存分析中，人们主要采用的分布假设是 Weibull 分布：

$$e_i \sim \alpha\beta^{\alpha}e_i^{\alpha-1}\exp\left\{-\left(\frac{e_i}{\beta}\right)^{\alpha}\right\}$$

其中，α、β 为未知参数。值得注意的是，Weibull 分布非常灵活。如果我们定义 $\alpha = 1$，那么 Weibull 分布则变化为另外一个重要的生存分布，即指数分布：

$$e_i \sim \beta\exp\{-e_i/\beta\}$$

此外，Weibull 分布还可以变化为 Rayleigh 分布，这里就不再一一赘述了。

一旦我们设定了残差项的分布状况，理论上讲，我们就可以获得相应的似然函数，通过极大化该似然函数获得极大似然估计并进行似然比检验。以 Weibull 分布为例，在 *R* 环境中可以分析如下：

```
> fit=survreg(Surv(Time,VStatus)~HGB+Platelet+Age+LogWBC+LogPBM+Protein+SCalc,data=a)
> summary(fit)

Call:
survreg(formula = Surv(Time, VStatus) ~ HGB + Platelet + Age +
    LogWBC + LogPBM + Protein + SCalc, data = a)
              Value Std. Error      z      p
(Intercept)  7.7142     2.9929  2.577 0.00995
HGB          0.1243     0.0671  1.852 0.06404
Platelet     0.3399     0.4696  0.724 0.46916
Age          0.0027     0.0172  0.156 0.87615
LogWBC      -1.1340     0.6601 -1.718 0.08578
LogPBM      -0.2351     0.4155 -0.566 0.57159
Protein     -0.0106     0.0237 -0.447 0.65524
SCalc       -0.1284     0.1030 -1.246 0.21275
Log(scale)  -0.0939     0.1089 -0.863 0.38820

Scale= 0.91

Weibull distribution
Loglik(model)= -210   Loglik(intercept only)= -215.1
        Chisq= 10.21 on 7 degrees of freedom, p= 0.18
```

首先可以看到，模型整体不显著（P 值 $= 0.18$）。因此，从理论上讲，我们不应该对该模型中的任何系数作解释。为了方便起见，我们假设模型整体显

著。那么,从中可以看到,在10%的显著性水平下,有两个因素能够显著地影响病人的生存时间,即HGB(确诊时血色素含量)以及LogWBC(对数变换后白细胞含量)。具体地说,在确诊时表现为高HGB水平的病人的生存时间会比较长,而表现为高LogWBC水平的病人的生存时间会比较短。我们还可以将残差分布假设为指数分布,重新拟合如下:

```
> fit=survreg(Surv(Time,VStatus)~HGB+Platelet+Age+LogWBC+LogPBM+Protein+SCalc,
+ dist="exponential",data=a)
> summary(fit)

Call:
survreg(formula = Surv(Time, VStatus) ~ HGB + Platelet + Age +
    LogWBC + LogPBM + Protein + SCalc, data = a, dist = "exponential")
              Value Std. Error      z      p
(Intercept)  7.75817     3.2423  2.393 0.0167
HGB          0.13072     0.0727  1.798 0.0721
Platelet     0.33141     0.5105  0.649 0.5162
Age          0.00329     0.0186  0.176 0.8600
LogWBC      -1.15078     0.7118 -1.617 0.1059
LogPBM      -0.22109     0.4498 -0.492 0.6231
Protein     -0.01209     0.0256 -0.472 0.6367
SCalc       -0.13831     0.1085 -1.274 0.2026

Scale fixed at 1

Exponential distribution
Loglik(model)= -210.3   Loglik(intercept only)= -215.1
        Chisq= 9.6 on 7 degrees of freedom, p= 0.21
```

大家可以看到,结论基本一致。类似于其他的模型结构(如逻辑回归),我们也可以作AIC变量选择如下:

```
> fit.aic=step(fit,trace=F)
> summary(fit.aic)

Call:
survreg(formula = Surv(Time, VStatus) ~ HGB + LogWBC + SCalc,
    data = a, dist = "exponential")
              Value Std. Error      z       p
(Intercept)  7.519     2.6441  2.84 0.00446
HGB          0.149     0.0582  2.57 0.01018
LogWBC      -1.068     0.6613 -1.61 0.10638
SCalc       -0.155     0.1027 -1.51 0.13047

Scale fixed at 1

Exponential distribution
Loglik(model)= -210.8   Loglik(intercept only)= -215.1
        Chisq= 8.55 on 3 degrees of freedom, p= 0.036
```

从中可以看到,模型整体的显著性水平得到了极大改善。其P值从0.21变为0.036,在5%的水平下高度显著。因此,我们可以对各个系数估计予以解释。进一步,AIC发现了三个重要的解释性变量,它们分别是HGB、LogWBC和SCalc。如果我们改用BIC,那么模型选择结果如下:

```
> ss=length(a[,1])
> fit.bic=step(fit,trace=F,k=log(ss))
> summary(fit.bic)

Call:
survreg(formula = Surv(Time, VStatus) ~ HGB, data = a, dist = "exponential")
             Value Std. Error     z        p
(Intercept) 2.307     0.5550 4.16 3.23e-05
HGB         0.114     0.0541 2.11 3.50e-02

Scale fixed at 1

Exponential distribution
Loglik(model)= -213   Loglik(intercept only)= -215.1
        Chisq= 4.32 on 1 degrees of freedom, p= 0.038
```

可以看到,BIC 认为 HGB 是唯一的重要解释性变量。

加速死亡模型有它的优点,也有它的缺点。它的优点是简单、容易理解。它将对数变换后的生存时间和普通线性模型联系了起来,因而非常直观,容易被接受。但是,它的缺点是模型的假设太强,它需要对残差分布作严格的假设(如 Weibull 分布)。如果这些假设不成立呢?也许结果是相当稳定的(如本案例),也就是说,无论作何种假设,最终结果都非常相似。但是,结果也可能对假设很敏感,如不同的模型假设产生截然相反的结论。因此,怎样在更弱的假设下回归生存数据就成为人们非常关心的问题。在过去大量的探索研究中,产生了著名的 Cox 等比例风险模型(Cox's proportional hazard model)。

第五节　Cox 风险模型

在讲述 Cox 风险模型以前,首先要明白什么是风险(hazard),它和生存函数(survival function)有什么关系?风险定义如下:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant t_i < t + \Delta t \mid t_i \geqslant t)}{\Delta t} = -\frac{f(t)}{S(t)}$$

其中,$f(t)$是生存函数 $S(t)$的导数,即生存时间 t_i 的密度函数,而 $h(t)$就是人们常说的风险函数。为什么要定义风险函数 $h(t)$呢?直观上讲,给定某个体已经存活了 t 时间,则 $h(t)$从某个侧面度量了该个体"立刻"死亡的可能性。理论上,风险函数与生存函数的关系如下:

$$S(t) = \exp\left\{-\int_0^t h(s)\,\mathrm{d}s\right\}$$

这说明风险函数唯一地决定了生存函数的形式。因此,只要我们能够通过合理的方式,建立生存函数 $h(t)$和解释性变量 t 之间的关系,我们就可以获得一个关于生存数据的回归模型。但是,我们如何建立这样一种关系呢?

著名统计学家 Cox 在他 1972 年的文章里提出了著名的 Cox 等比例风险模型(以下简称 Cox 模型)。该模型具有以下形式:

$$h(t) = h_0(t)\exp\{\beta' x\}$$

其中,$h_0(t)$叫做基准风险函数(baseline hazard function)。直观上讲,它代表了在没有协变量影响(即 $x=0$)的情况下,一个病人的风险函数形式。从某种意义上讲,它的作用有点像前一节的加速死亡模型中的 t_{a0}。在给定协变量 x 的影响下,基准风险函数 $h_0(t)$ 被等比例扩大了 $\exp\{\beta' x\}$ 倍。因此,我们称此模型为等比例风险模型。理论上人们可以验证,通过恰当地选取基准风险函数 $h_0(t)$ 的函数形式,Cox 模型包含了前面所讲的 Weibull 加速死亡模型,即 Weibull 加速死亡模型为 Cox 模型的特例。因此,Cox 模型的形式更加灵活。而且 Cox 模型的一大优点是可以在不对基准风险函数 $h_0(t)$ 的函数形式作任何假设的情况下,获得对回归系数 β 的可靠估计以及相关的统计推断方法。

那么,Cox 模型是怎样对回归系数 β 作简单估计的呢?假设 t_i 是一个观测到的事件时间,也就是说,有那么一个个体,他的生存时间是 t_i,并且没有被截断。对于该个体我们应该怎样分析呢?首先我们知道他至少存活了 t_i 时间,并且"立刻"死亡了。但是,存活时间至少为 t_i 的个体可能并不仅此一个。那么在这么多存活时间大于或等于 t_i 的个体中,该个体死亡的可能性有多大呢?在 Cox 模型下,此概率为:

$$\frac{h_0(t_i)\exp\{\beta' x_i\}}{\sum_{j\in R(t_i)} h_0(t_i)\exp\{\beta' x_j\}} = \frac{\exp\{\beta' x_i\}}{\sum_{j\in R(t_i)}\exp\{\beta' x_j\}}$$

其中,$R(t)$ 是这样一个集合,它包含了所有存活时间至少为 t 的个体。请注意等号的右端,原来的分子和分母中的基准风险函数 $h_0(t_i)$ 相互抵消掉了。因此,最后的概率仅依赖于回归系数 β。将所有这样观测到的事件时间点汇总在一起,就构成了我们的似然函数,又叫做偏似然函数(partial likelihood),形式如下:

$$\prod_{t_i\in E}\frac{\exp\{\beta' x_i\}}{\sum_{j\in R(t_i)}\exp\{\beta' x_j\}}$$

其中,E 是前面定义的事件集合。然后,我们就可以通过极大化该似然函数获得极大似然估计,并进行似然比检验等。在 R 中,可以实现如下:

```
> fit=coxph(Surv(Time,VStatus)~HGB+Platelet+Age+LogWBC+LogPBM+Protein+SCalc,data=a)
> summary(fit)
Call:
coxph(formula = Surv(Time, VStatus) ~ HGB + Platelet + Age +
    LogWBC + LogPBM + Protein + SCalc, data = a)

  n= 65
             coef exp(coef) se(coef)      z     p
HGB       -0.14483     0.865   0.0742 -1.952 0.051
Platelet  -0.29042     0.748   0.5106 -0.569 0.570
Age       -0.00373     0.996   0.0190 -0.196 0.840
LogWBC     0.98530     2.679   0.8017  1.229 0.220
LogPBM     0.21882     1.245   0.4549  0.481 0.630
Protein    0.01260     1.013   0.0260  0.486 0.630
SCalc      0.15111     1.163   0.1115  1.355 0.180
```

从中可以看到，HGB 仍然是显著的（P 值 = 0.051），而 logWBC 不再显著。值得注意的是，Cox 模型所拟合出来的 HGB 的系数是负的，而加速死亡模型的拟合结果是正的。这是为什么呢？加速死亡模型中 HGB 的系数为正说明 HGB 水平高的病人存活时间长，这正好表明此类病人的风险函数小，因此在 Cox 模型中该变量系数的符号为负。最后值得注意的是，该模型中没有截距项，这是因为它已经被吸收到了基准风险函数 $h_0(t)$ 中去了。

类似于加速死亡模型结构，我们也可以对 Cox 模型作变量选择。而 AIC 变量选择方法如下：

```
> ph.aic=step(fit,trace=F)
> summary(ph.aic)
Call:
coxph(formula = Surv(Time, VStatus) ~ HGB + SCalc, data = a)

  n= 65
       coef exp(coef) se(coef)     z     p
HGB   -0.144     0.866   0.0575 -2.50 0.012
SCalc  0.162     1.176   0.1037  1.56 0.120
```

AIC 认为 HGB 和 SCalc 都会显著影响病人的生存风险，因此也就会影响到病人的生存时间。如果我们用 BIC，那么结果如下：

```
> ph.bic=step(fit,trace=F,k=log(ss))
> summary(ph.bic)
Call:
coxph(formula = Surv(Time, VStatus) ~ HGB, data = a)

  n= 65
      coef exp(coef) se(coef)     z     p
HGB -0.126     0.882    0.056 -2.25 0.024
```

同加速死亡模型一样，BIC 认为 HGB 是唯一能够显著影响病人生存的解释性变量。HGB 水平高的病人具有较低的风险，因此生存时间较长。

第六节　简单分析报告

哪些因素影响了骨髓癌患者的生存时间

内容提要　本报告利用生存分析的方法来寻找影响骨髓癌患者生存时间的因素，并建立统计模型来度量其影响程度。我们的分析结果发现，有三个因素能够显著地影响病人的生存时间，即血色素含量、白细胞含量和血清钙含量，第一个因素的影响尤为明显。具体地说，在确诊时表现为高血色素含量的病人的生存时间会更长，高白细胞含量的病人的生存时间会比较短，而高血清钙含量的病人的生存时间会较短。根据我们的分析结果，医生和相关研究者在处理骨髓癌这种疾病时，要特别关注这三个指标，尤其是血色素含量。

一、研究目的

医生在诊断骨髓癌(Multiple Myeloma)这种致命疾病时，常常需要作一个判断，即病人还能够生存多长时间，或者哪些因素影响着病人的生存时间。合理地回答该问题对医生和相关研究者的帮助甚大。基于对该问题的正确回答，医生能够给病人制定更合理的治疗方案，研究者可以对相关课题进行更为精确的研究。本报告试图通过对相关临床数据的分析找出影响患者生存时间的因素，并根据分析结果提出有意义的结论和建议。

二、数据来源和相关说明

本报告所使用的数据来自于 Krall、Uthoff 和 Harley(1975)收集的关于骨髓癌患者的生存数据。该数据共包含了 65 个病人的资料，其中 48 人在研究期间死去，而 17 人活过了最后的研究期限。对于每一个病人，研究者收集并研究了表 7-2 中的变量。

表 7-2 变量说明

变量名称	解释意义
Time	从确诊到死亡的生存时间(单位:月)
VStatus	生存状态(0 = 生存;1 = 死亡)
HGB	确诊时血色素含量
Platelet	确诊时血小板状况(0 = 不正常;1 = 正常)
Age	确诊时年龄(单位:年)
LogWBC	对数变换后白细胞含量
LogPBM	对数变换后骨髓中血浆细胞含量
Protein	确诊时血蛋白含量
SCalc	确诊时血清钙含量

其中,我们感兴趣的因变量是生存时间(Time),而其他的变量都为解释性变量。需要注意的是,生存状态为"死亡"(即 VStatus = 1)时,Time 即为确确实实的生存时间;生存状态为"生存"(即 VStatus = 0)时,意味着在该研究结束之前,没有观测到病人的确切生存时间,因此我们从数据中仅仅知道他的生存时间要大于 Time 的取值,但是到底大多少是不知道的。

三、描述性分析

为了获得对数据的整体概念,并注意到因变量为生存数据的特点,我们利用生存函数来考察生存时间和各个解释性变量之间的关系。我们首先考虑血色素含量(HGB)对生存函数的影响。由于血色素含量是一个连续变量,因此我们根据血色素含量的平均水平(以中位数计),将所有的样本分成高、低两组。然后,我们将两组(即高 HGB 组以及低 HGB 组)的生存曲线进行对比,如图 7-6 所示。

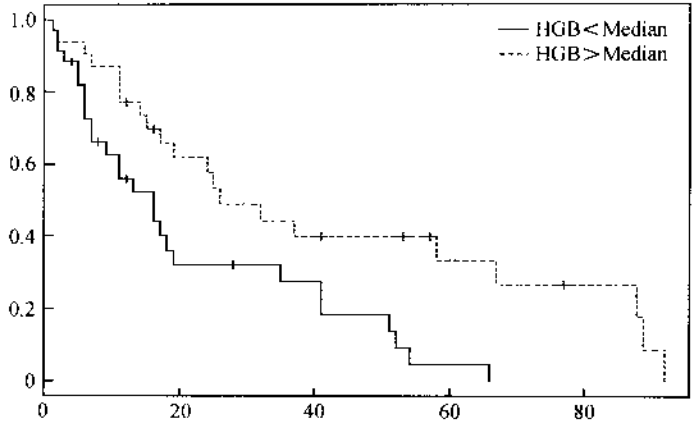


图 7-6　HGB 分组对应的生存函数图

从图 7-6 可以清楚地看到，高 HGB 组的生存曲线（虚线）一直都在低 HGB 组的生存曲线（实线）的上方。这说明，对于任意的一个时间点（如 40 个月），高 HGB 组的生存概率都要高于低 HGB 组的生存概率。由此我们可以得到一个初步的结论，即血色素水平的高低是影响病人生存时间的重要指标。

我们对其他变量也作类似的描述性分析，得到图 7-7。

从图 7-7 中我们可以得到以下初步结论：

- 具有正常血小板水平的病人的生存概率（虚线）明显地高于具有非正常血小板水平的病人（实线）。可见，血小板水平也有可能是影响病人生存时间的重要指标。
- 高年龄组（Age > 60）的生存曲线（虚线）稍微低于低年龄组（Age < 60）的生存曲线（实线），但是这种差别并不大。因而年龄对于病人生存时间的影响还需要进一步的考察。
- 白细胞含量（LogWBC）、骨髓中血浆细胞含量（LogPBM）以及血清钙含量（SCalc）的高低都对生存概率影响不大。
- 血蛋白含量（Protein）高的病人的生存概率要低于血蛋白含量低的病人，但是差别不大。

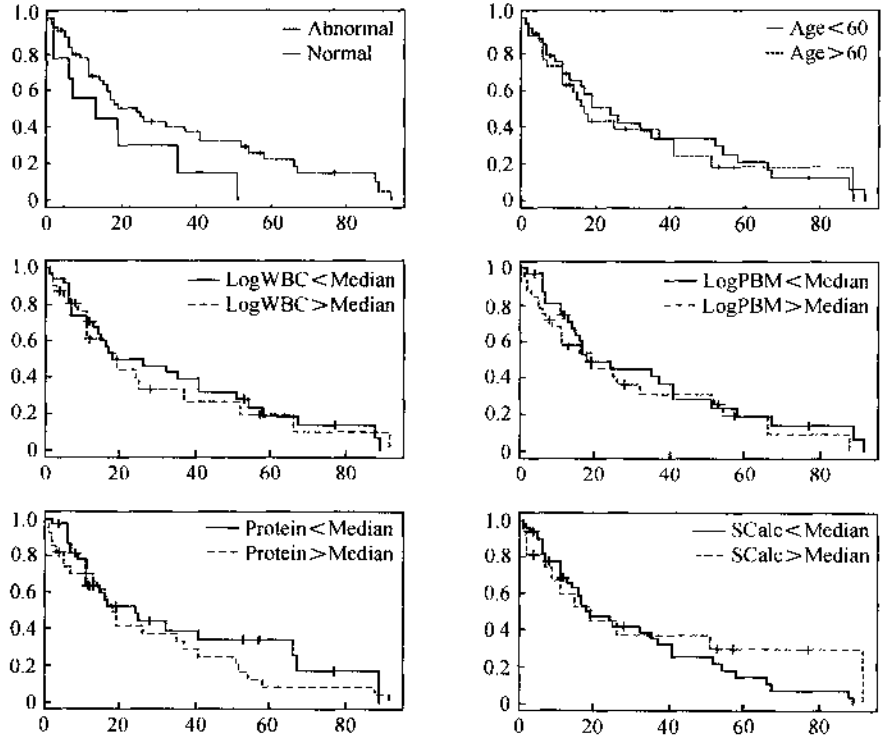


图 7-7　生存函数图

注：从上到下对应自变量依次为：血小板水平、年龄、白细胞含量、骨髓中血浆细胞含量、血蛋白含量、血清钙含量。

四、数据建模

1. 加速死亡模型及模型选择

从上一节的分析中我们可以看到，确实有部分因素对病人的生存时间有影响。为了更准确地找出这些因素并量化其影响程度，我们首先使用生存分析中常用的加速死亡模型进行拟合。假设残差项服从 Weibull 分布，我们得到表 7-3 中的模型估计结果。

表 7-3 加速死亡模型(Weibull 分布)

变量名	系数估计值	标准差	P 值
截距项	7.714	2.993	0.010
HGB	**0.124**	**0.067**	**0.064**
Platelet	0.340	0.470	0.469
Age	0.003	0.017	0.876
LogWBC	**-1.134**	**0.660**	**0.086**
LogPBM	-0.235	0.416	0.572
Protein	-0.012	0.024	0.655
Scale	-0.129	0.100	0.213

从表 7-3 我们可以得到如下结论:

- 在 10% 的显著性水平下,HGB(确诊时血色素含量)能够显著地影响病人的生存时间。具体地说,在确诊时表现为高 HGB 水平的病人的生存时间会更长。
- 在 10% 的显著性水平下,LogWBC(对数变换后白细胞含量)也能够显著地影响病人的生存时间。具体地说,在确诊时表现为高 LogWBC 水平的病人的生存时间会比较短。
- 其他五个因素对于病人的生存时间没有显著的影响。

该结果与我们在描述性分析中所得到的初步结论基本一致。为考察该模型的稳定性,我们假设残差项服从指数分布,再次对加速死亡模型进行拟合,从而得到与 Weibull 分布假设非常类似的结果,如表 7-4 所示。

表 7-4 加速死亡模型(指数分布)

变量名	系数估计值	标准差	P 值
截距项	7.758	3.242	0.017
HGB	**0.131**	**0.073**	**0.072**
Platelet	0.331	0.512	0.516
Age	0.003	0.019	0.860
LogWBC	**-1.151**	**0.712**	**0.106**
LogPBM	-0.221	0.450	0.623
Protein	-0.012	0.026	0.636
SCale	-0.138	0.109	0.202

以上的分析表明,本案例对于不同的分布假设并不敏感,因此我们利用加速死亡模型得到的结论较为可靠。

为获得更为可靠的结果,我们还利用 AIC 信息准则来进行模型选择。从残

差服从指数分布的加速死亡模型中，我们可以得到选择结果如表 7-5 所示。从表中可以看出，AIC 发现了三个重要的解释性变量，它们分别是 HGB（确诊时血色素含量）、LogWBC（对数变换后白细胞含量）和 Scalc（确诊时血清钙含量）。AIC 所选出的模型关于前两个变量的结论与全模型相同，它还认为在确诊时表现为高血清钙含量的病人的生存时间会较短。以上三种模型的结果较为一致，这从另一方面验证了我们利用加速死亡模型得到的结论较为可靠。

表 7-5　AIC 加速死亡模型（指数分布）

变量名	系数估计值	标准差	*P* 值
截距项	7.519	2.644	0.004
HGB	0.149	0.058	0.010
LogWBC	-1.068	0.661	0.106
SCalc	-0.155	0.102	0.130

2. Cox 等比例风险模型及模型选择

为了获得更全面、更可靠的结论，我们考虑使用 Cox 等比例风险模型，通过极大化似然函数的方法得到模型的拟合结果，如表 7-6 所示。

表 7-6　Cox 等比例风险模型

变量名	系数估计值	标准差	*P* 值
HGB	**-0.145**	**0.074**	**0.051**
Platelet	-0.290	0.511	0.570
Age	-0.004	0.019	0.840
LogWBC	0.985	0.802	0.220
LogPBM	0.219	0.455	0.630
Protein	0.013	0.026	0.630
SCalc	0.151	0.112	0.180

从表 7-6 中可以看到，HGB（确诊时血色素含量）仍然是显著的（*P* 值 = 0.051），而且系数为负，表明 HGB 高的病人的风险函数较小，这与加速死亡模型中关于 HGB 的结论相同。这进一步表明，HGB 是一个非常重要的影响因素。但需要注意的是，此时 LogWBC（对数变换后白细胞含量）不再显著，即 Cox 等比例风险模型认为 LogWBC 对病人的生存时间没有显著的影响。

为获得更为可靠的结果，我们还利用 AIC 信息准则来进行模型选择，得到的结果如表 7-7 所示。从表中可以看出，AIC 准则所选出的模型认为除了 HGB（确诊时血色素含量）外，SCalc（确诊时血清钙含量）也是一个重要的变量。在确诊时表现为高血清钙含量的病人的风险函数较高，这与加速死亡模型的结论

一致。

表 7-7 AIC Cox 等比例风险模型

变量名	系数估计值	标准差	P 值
HGB	-0.144	0.054	0.012
SCalc	0.162	0.104	0.120

由此可见，加速死亡模型与 Cox 等比例风险模型给出了略微不同的结论，但从保守而稳妥的角度，我们可以认为 HGB（确诊时血色素含量）、LogWBC（对数变换后白细胞含量）和 SCalc（确诊时血清钙含量）这三个因素都能够显著地影响病人的生存时间，而且第一个因素的影响尤为明显。

五、结论及建议

从上述分析结果可知，HGB（确诊时血色素含量）、LogWBC（对数变换后白细胞含量）和 SCalc（确诊时血清钙含量）这三个因素都能够显著地影响病人的生存时间，而且第一个因素的影响尤为明显。具体地说，在确诊时表现为高血色素含量的病人的生存时间会更长，高白细胞含量的病人的生存时间会比较短，而高血清钙含量的病人的生存时间会较短。因此，医生在诊断病情时，要特别关注这三个指标，而且在治疗时也需要时刻对这三个指标加以合理的控制。同时，这个结论也为相关的研究者指出了一个相对明确的研究方向。

[讨论总结]

本章以一个关于癌症临床数据的实际案例为例，系统演示并讲解了生存分析这种重要的统计学方法。通过对本章的学习，读者应该能够了解：什么时候可以进行生存分析，以及如何做。在 R 语言学习方面，读者应该掌握相关的生存分析的命令。在统计理论方面，读者应该掌握以下概念：生存数据、截断、生存函数、KM 估计、风险函数、加速死亡模型和 Cox 等比例风险模型等。对相关生存分析理论渴望深入了解的读者请参阅 Fleming and Harrington（1991）以及 Kalbfleisch and Prentice（1980）。

附录　程序及注释

```
a=read.csv("D:/Practical Business Data Analysis/case/CH7/data.csv",header=T)
a[1:20,]
ibrary(survival)
a=a[order(a$Time),]
Surv(a$Time,a$VStatus)
a$Time[a$VStatus==1]
summary(survfit(Surv(a$Time,a$VStatus)))
plot(survfit(Surv(a$Time,a$VStatus)))
status=1*(a$HGB > median(a$HGB))
status
plot(survfit(Surv(a$Time,a$VStatus) ~ status),col=c(1,2),lty=c(1,2))
legend(40,1,c("HGB < Median","HGB > Median"),col=c(1,2),lty=c(1,2))
plot(survfit(Surv(a$Time,a$VStatus) ~ a$Platelet),col=c(1,2),lty=c(1,2))

legend(40,1,c("Abnormal","Normal"),col=c(1,2),lty=c(1,2))
age.group=1*(a$Age > median(a$Age))
plot(survfit(Surv(a$Time,a$VStatus) ~ age.group),col=c(1,2),lty=c(1,2))

legend(40,1,c("Age < 60","Age > 60"),col=c(1,2),lty=c(1,2))
par(mfrow=c(2,2))
```

```
# 读入 csv 格式的数据,并赋值给 a
# 展示 a 的前 20 行数据
# 载入程序包 survival
# 根据 Time 的取值进行排序
# 得到生存分析所用时间格式
# 得到没有被截断的时间点
# 计算生存函数的置信区间
# 画出生存函数及其置信区间
# 根据中位数将 HGB 分成两组,分别取值 0 或 1
# 展示上一命令所得到的数据
# 根据 status 的取值画出两组不同颜色和线形的生存函数
# 对不同的生存函数加上标示
# 根据 Platelet 的取值画出两组不同颜色和线形的生存
  函数
# 对不同的生存函数加上标示
# 根据中位数将 Age 分成两组,分别取值 0 或 1
# 根据 age.group 的取值画出两组不同颜色和线形的生存
  函数
# 对不同的生存函数加上标示
# 设置画图模式为 2x2
```

```
status=1*（a$LogWBC > median(a$LogWBC)）                    # 根据中位数将 LogWBC 分成两组,分别取值 0 或 1
plot(survfit(Surv(a$Time,a$VStatus) ~ status),col=c(1,2),lty=c(1,2))   # 根据 status 的取值画出两组不同颜色和线形的生存函数
legend(30,1,c("LogWBC < Median","LogWBC > Median"),col=c(1,2),lty=c(1,2))  # 对不同的生存函数加上标示
status=1*（a$LogPBM > median(a$LogPBM)）                    # 根据中位数将 LogPBM 分成两组,分别取值 0 或 1
plot(survfit(Surv(a$Time,a$VStatus) ~ status),col=c(1,2),lty=c(1,2))   # 根据 status 的取值画出两组不同颜色和线形的生存函数
legend(30,1,c("LogPBM < Median","LogPBM > Median"),col=c(1,2),lty=c(1,2))  # 对不同的生存函数加上标示
status=1*（a$Protein > median(a$Protein)）                   # 根据中位数将 Protein 分成两组,分别取值 0 或 1
plot(survfit(Surv(a$Time,a$VStatus) ~ status),col=c(1,2),lty=c(1,2))   # 根据 status 的取值画出两组不同颜色和线形的生存函数
legend(30,1,c("Protein < Median","Protein > Median"),col=c(1,2),lty=c(1,2))  # 对不同的生存函数加上标示
status=1*（a$SCalc > median(a$SCalc)）                      # 根据中位数将 SCalc 分成两组,分别取值 0 或 1
plot(survfit(Surv(a$Time,a$VStatus) ~ status),col=c(1,2),lty=c(1,2))   # 根据 status 的取值画出两组不同颜色和线形的生存函
                                                            数
legend(30,1,c("SCalc < Median","SCalc > Median"),col=c(1,2),lty=c(1,2))  # 对不同的生存函数加上标志
par(mfrow=c(1,1))                                           # 设置画图模式,还原为 1x1
fit=survreg(Surv(Time,VStatus) ~ HGB + Platelet + Age + LogWBC + LogPBM + Protein + SCalc,data=a)
                                                            # 拟合加速死亡模型
summary(fit)                                                # 显示模型 fit 的各方面细节,包括估计值、标准差等
fit=survreg(Surv(Time,VStatus) ~ HGB + Platelet + Age + LogWBC + LogPBM + Protein + SCalc,dist="exponential",data=a)
                                                            # 加速死亡模型
summary(fit)                                                # 显示模型 fit 的各方面细节,包括估计值、标准差等
fit.aic=step(fit,trace=F)                                   # 根据 AIC 准则选择最优模型
summary(fit.aic)                                            # 显示模型 fit.aic 的各方面细节,包括估计值、标准差等
ss=length(a[,1])                                            # 计算样本大小
fit.bic=step(fit,trace=F,k=log(ss))                         # 根据 BIC 准则选择最优模型
summary(fit.bic)                                            # 显示模型 fit.bic 的各方面细节,包括估计值、标准差等
fit=coxph(Surv(Time,VStatus) ~ HGB + Platelet + Age + LogWBC + LogPBM + Protein + SCalc,data=a)
```

```
summary(fit)
ph.aic=step(fit,trace=F)
summary(ph.aic)
ph.bic=step(fit,trace=F,k=log(ss))
summary(ph.bic)
```

拟合 Cox 等比例风险模型
显示模型 fit 的各方面细节,包括估计值、标准差等
根据 AIC 准则选择最优模型
显示模型 ph.aic 的各方面细节,包括估计值、标准差等
根据 BIC 准则选择最优模型
显示模型 ph.bic 的各方面细节,包括估计值、标准差等

第八章
自回归

- 案例介绍
- 时间序列的平稳性
- 基本描述
- 自相关系数
- 自回归模型及其平稳性
- 模型估计与选择
- 模型诊断
- 模型预测
- 简单分析报告
- 程序及注释

[教学目的]

本章的主要教学目的就是通过一个关于失业率预测的实际案例，详细介绍自回归这种重要的时间序列回归模型。它主要处理的是平稳的时间序列数据。通过对本章的学习，我们希望读者能够了解：(1) 什么情况下采用时间序列分析，什么情况下采用自回归模型；(2) 时间序列以及自回归的基本统计学理论；(3) 相关理论在统计学软件 R 中的应用；(4) 相应的统计分析报告的撰写。本章所涉及的新统计学概念有自相关系数、时间序列的平稳性、自回归模型。

第一节　案例介绍

在前面几章中我们考虑的数据类型都有一个共同特点，那就是：多个个体、单个观测。这是什么意思呢？例如，在第一章的盈利预测案例中，我们同时考虑了很多家公司。如果将一家公司看做一个个体(subject)，那么我们同时考虑了成百上千家公司，即多个个体。由于我们研究的是某一年的具体情况，因此每家公司给我们提供了一年的数据，即单个观测。类似地，我们可以发现第二章研究的房地产数据以及第三章讨论的教学评估数据都有这样的特点。我们称此类数据为横截面数据(cross-sectional data)。与横截面数据相对应的是时间序列数据(time series data)，即一个个体随着时间的推移有多个观测。例如，对于一个国家的 GDP，我们有一个国家，即一个个体；随着时间的推移，我们有该国每年的 GDP 取值，即多个观测。由于不同的横截面数据来自于不同的个体，因此，可以合理假设不同的横截面数据是相互独立的。但是，时间序列数据就不一样了。由于它们都来自于同一个个体，因此，一般来说不同的数据之间有很强的相关性。例如，某个国家当年的 GDP 同前一年的 GDP 高度相关。因此，这给我们提出了一个问题：如何对时间序列数据作回归分析？

在本章中，我们将根据某个国家的失业率数据作具体的演示。我们都知道，失业率是衡量一个国家或地区就业状况最重要的指标之一，也是反映社会稳定性的重要指标之一。因此，分析并把握失业率的变化规律，对于相关机构，特别是政府职能部门，意义重大。我们的数据详细记录了 1990 年 1 月至 2006 年 12 月间，该国各个月份的全国失业率(%)。历史经验表明，失业率的高低受月份的影响很大。因此，不同月份的失业率没有直接的可比性。所以，我们的案例

数据并不是失业率的原始数据，而是经过季节性调整后的失业率。这样一来，不同月份的失业率才具有了一定的可比性。这为我们后面的统计建模打下了良好的基础。那么我们是如何对案例数据进行季节性调整的呢？有兴趣的读者可以参考相关专著，以获得更详细的解释。

对于时间序列数据，人们关心什么问题呢？以本案例为例，我们关心两个问题：第一，失业率的变化规律，如当月的失业率同过去几个月的失业率是否有关系？如果有，是什么样的关系？第二，基于失业率的变化规律作预测。

第二节　时间序列的平稳性

为了方便起见，我们用 $x_t(t=1,\cdots,T)$ 代表来自于 t 时刻的数据，如在本案例中 x_1 代表了第一时刻（即 1990 年 1 月）的失业率。假设我们的数据被存放在“D:\WORKING\teaching\Practical Business Data Analysis\case\CH4”，那么在 *R* 的编程环境中，输入以下语句就可以读入并展示数据如下（参见图 8-1）：

```
> a=read.csv("D:/WORKING/teaching/Practical Business Data Analysis/case/CH4/rate.csv")
> ts.plot(a$rate)
```

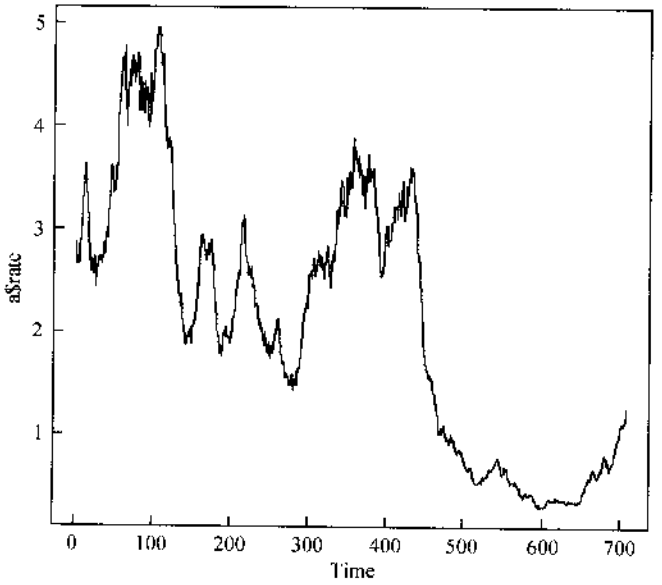


图 8-1　失业率的时间序列图

从图 8-1 我们可以看到，在过去的 27 年（1990—2006）中，该国的失业率基本上保持在 5% 以内，总体平均水平大约为 2%，但是不同时期的失业率平均水平差异巨大。这就给我们提出了一个问题：对于这样的变化巨大的数据，我们能够建立良好的预测模型吗？

要回答这个问题，我们首先需要明白：为什么统计模型具有预测能力？在本案例中，我们拥有 1990—2006 年共 27 年的数据。如果我们希望能够对 2007 年的失业率予以合理预测，我们就必须假设 2007 年失业率的变化规律（不是具体取值）同 1990—2006 年的变化规律有着某种相似性。如果没有这种相似性，1990—2006 年间的数据将完全失去它对 2007 年失业率的借鉴意义。那么，统计上如何定义这种相似性呢？那就是时间序列的平稳性。更严格地说，我们称一个时间序列是平稳的（stationary），如果该序列满足：对于一个任意的整数 $k>0$，随机变量 $(x_{t_0+1}, x_{t_0+2}, \cdots, x_{t_0+k})$ 的分布同时间点 t_0 无关。换句话说，对于任意的 $t_1 \neq t_2$，我们有 $(x_{t_1+1}, x_{t_1+2}, \cdots, x_{t_1+k})$ 和 $(x_{t_2+1}, x_{t_2+2}, \cdots, x_{t_2+k})$ 的统计分布相同（注意：不是取值相同）。为什么时间序列的平稳性对于进行合理预测的意义如此重大呢？举一个简单的例子：首先，我们很容易量化历史数据 $\{x_1, \cdots, x_T\}$ 中相邻观测 (x_t, x_{t+1})，$t=1, \cdots, (T-1)$ 之间的回归关系（如 $E(x_{t+1}|x_t)=0.5x_t$）。如果时间序列是平稳的，那么同样的回归关系也应该存在于 (x_T, x_{T+1}) 中。请注意，x_T 是历史数据，而 x_{T+1} 是需要预测的未来观测。因此，我们可以对 x_{T+1} 预测如下：$\hat{x}_{T+1}=0.5x_T$。由此可见时间序列平稳性的重要意义。而对于平稳的时间序列，如何发现其回归关系将是下面章节讨论的重点。这里，我们首先需要知道：什么样的时间序列是非平稳的？

时间序列的平稳性隐含着一个结论，那就是：该序列的均值和方差都不会随着时间的改变而改变。而上图所显示的时间序列有着明显的上升趋势，因此该时间序列是非平稳的。另外一种典型的非平稳时间序列就是随机游走（random walk）的时间序列，如图 8-2 所示。

随机游走的时间序列有什么特点呢？随机游走的时间序列总体上不会表现出明显的上升或者下降的趋势。但是，该时间序列会在局部表现出上升或者下降的趋势。因此，该时间序列也非平稳。而典型的平稳时间序列应该如图 8-3 所示。

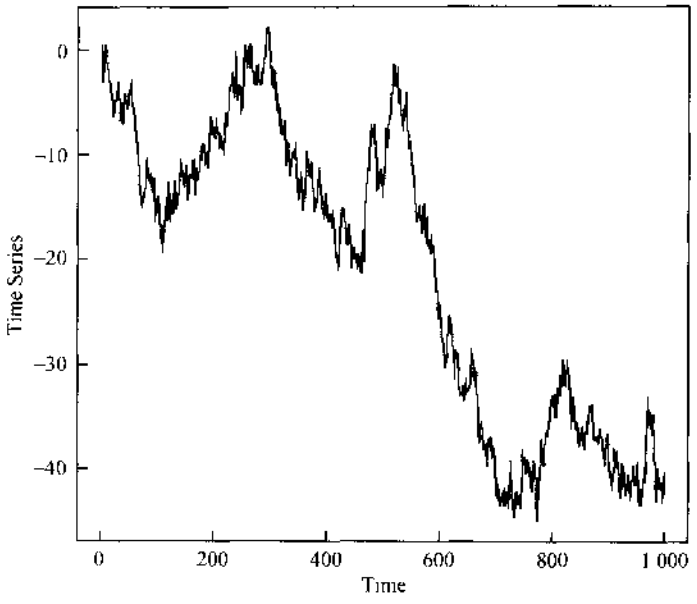


图 8-2　随机游走的时间序列图

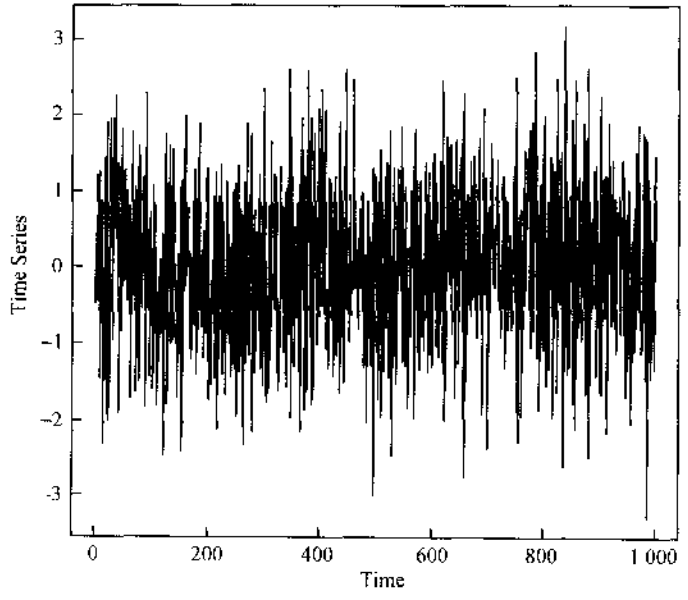


图 8-3　典型的平稳时间序列图

大家可以从图 8-3 看到:没有明显的变化趋势。因此,这样的时间序列才可能是平稳的。此时,再回忆一下我们的失业率数据是一个什么样的情形,如图 8-4 所示。

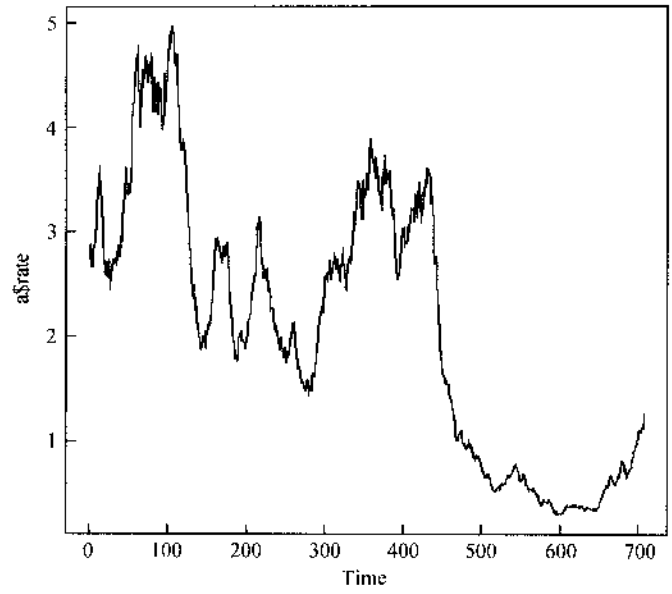


图 8-4　失业率的时间序列图

由此可见,该国的失业率数据显然不是平稳的时间序列数据。但是,我们可以考虑通过适当的变换,把不平稳的时间序列变得平稳。以本案例为例,我们可以定义一个新的时间序列如下:$r_t = \log(x_t/x_{t-1}) = \log(x_t) - \log(x_{t-1})$。简单地说,$r_t$ 定义了每个月份的失业率的对数变化率。而 r_t 的时间序列图如图 8-5 所示。

```
> r=diff(log(a$rate))
> ts.plot(r)
```

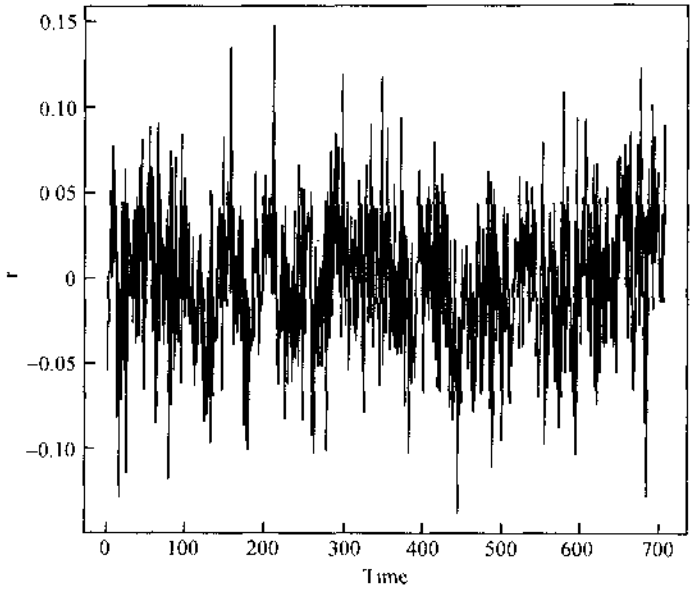


图 8-5　失业率对数变化率的时间序列图

同原来的原始数据(即失业率)相比,新数据(即失业率的对数变化率)的平稳性有了极大的改善。因此,我们将着重分析失业率的对数变化率。

第三节　基本描述

依照惯例,我们将对本案例所涉及的该国失业率的数据予以简单描述,以便为后面的模型分析打下基础。我们首先对原始数据描述如下(参见图 8-6):

```
> boxplot(a$rate)
```

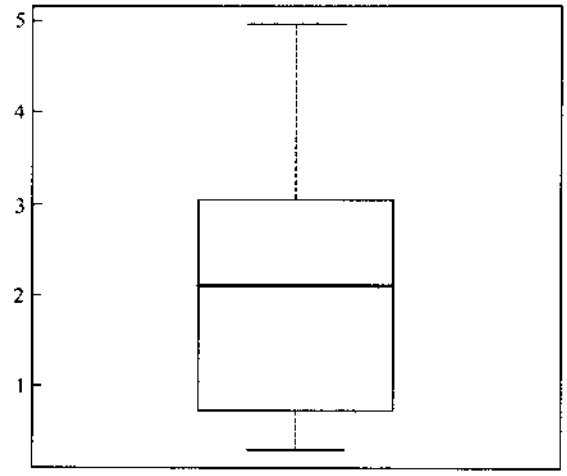


图 8-6　月度平均失业率的盒状图

从图 8-6 可以看到，在过去的 27 年中，该国的平均失业率（以中位数计）大约为 2%，非常低，而且最高也不超过 5%。再对失业率的对数变化率描述如下（参见图 8-7）：

```
> boxplot(r)
```

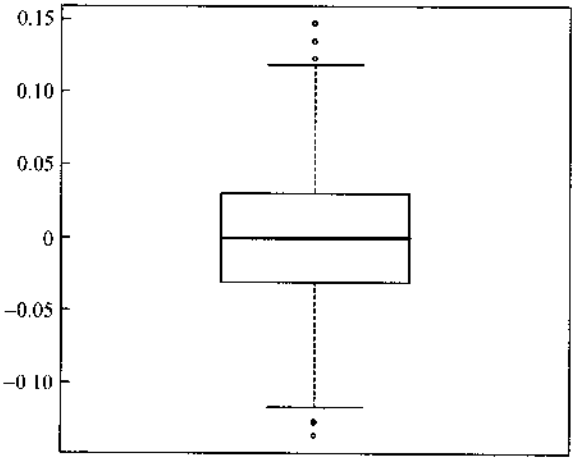


图 8-7　月度平均失业率对数变化率的盒状图

从图 8-7 可以看到，在过去的 27 年中，该国月度平均失业率的对数变化率（以中位数计）大约为 0，但是波动范围较大，最高接近 15%，而最低超过 −10%。

第四节　自相关系数

以上的描述简单地刻画了失业率及其变化率的平均水平以及波动范围。但是，我们前面提到，时间序列的一个重要特征就是其未来数据同历史数据的相关性。我们可以简单地称为自相关性（autocorrelation）。因此，非常有必要对时间序列相关性的强度予以量化和描述。对于横截面数据，我们知道可以通过相关系数（correlation coefficient）来描述两个变量之间的线性相关性。因此，我们也可以通过类似的技术手段描述相邻的时间序列数据的相关性。具体地说，对于一个给定的正整数 $k>0$，我们可以通过 $\rho_k=\mathrm{corr}(x_t,x_{t+k})$ 来简单度量相距为 k 个时间单位的两个时间序列数据的相关性，我们称 ρ_k 为 k 阶自相关系数（ACF）。例如：如果 $k=1$，那么 ρ_1 度量的是相邻两个月的失业率的相关性。如果 ρ_1 的绝对值很大，那么当月失业率和上月失业率的相关性就很强，否则就很弱。如果 ρ_1 是正的，那么上月失业率的上升往往能够带来当月失业率的上升，否则就下降。当然，我们的讨论绝不仅仅局限于 ρ_1（即相邻两个月的数据的相关性），我们可以对任意的 ρ_k 作类似的讨论。在 R 环境中，我们可以对 ρ_k 估计如下（参见图 8-8）：

```
> acf(a$rate)
```

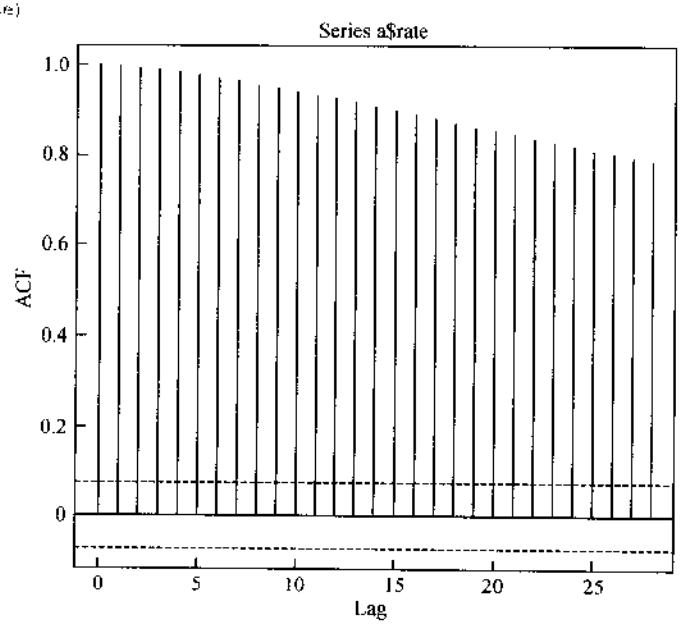


图 8-8　失业率的自相关系数图

请注意图 8-8 中的两条水平虚线，它们代表了一定的显著性水平。如果我们的估计量远远超出了它们所限定的范围，那么该阶的自相关系数高度显著。从图 8-8 可以看到，*R* 自动计算的所有自相关系数（阶数 > 25）都高度显著。这说明，该失业率的自相关性非常强烈而且复杂。如果我们直接对该失业率数据建立后面将要介绍的自回归模型，那么必须要建立一个非常复杂的模型，否则难以充分刻画该数据的时间序列特征。这进一步说明，我们也许应该考虑失业率的对数变化率。为此，我们对失业率的对数变化率描述如下（参见图 8-9）：

```
> acf(r)
```

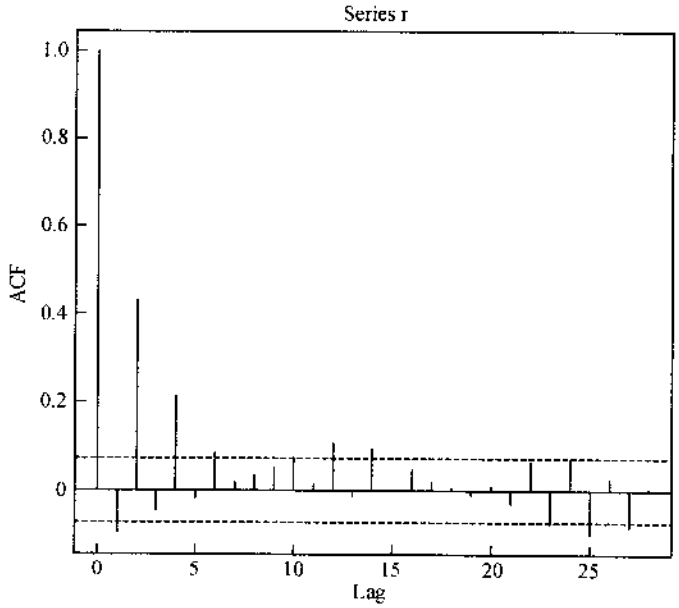


图 8-9　失业率对数变化率的自相关系数图

对比原始数据（即失业率）的自相关图，我们发现对数变化率的自相关系数要简单许多，只有为数不多的几个自相关系数明显地落在了虚线以外。因此，我们相信基于对数变化率的时间序列模型将相对更简单。

第五节　自回归模型及其平稳性

在描述性分析之后，我们将仔细考虑如何对失业率的对数变化率作统计分析，并由此预测未来。我们模仿线性模型，构造自回归模型如下：

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \cdots + \phi_p r_{t-p} + \varepsilon_t$$

其中，$\phi=(\phi_0,\phi_1,\cdots,\phi_p)$是自回归系数，$p$是该自回归模型的阶数，而$\varepsilon_t$是方差为$\sigma^2$的白噪音(white noise)。什么是白噪音？简单地说，如果一个时间序列模型是成功的，那么它所分离出来的残差就不应该还掺杂任何时间序列特征。所有时间序列特征都应该被模型充分利用，以提高预测精度。请注意，不同时刻的时间序列数据(如x_{t_1}和x_{t_2})往往是相关的。但是，不同时刻的白噪音(如ε_{t_1}和ε_{t_2})却是互相独立的。在讨论如何估计自回归系数ϕ以前，我们首先考虑一下什么样的自回归模型是平稳的？

例 1 (随机游走)考虑如下随机游走模型：$r_t=r_{t-1}+\varepsilon_t$，并假设$\mathrm{var}(\varepsilon_t)=1$。由此可得：$\mathrm{var}(r_t)=\mathrm{var}(r_{t-1})+1$。反复迭代该公式，我们可以得到：$\mathrm{var}(r_t)=\mathrm{var}(r_0)+t$。这说明，该序列的方差随着时间的推移越来越大，因此是非平稳的，如图 8-10 所示。

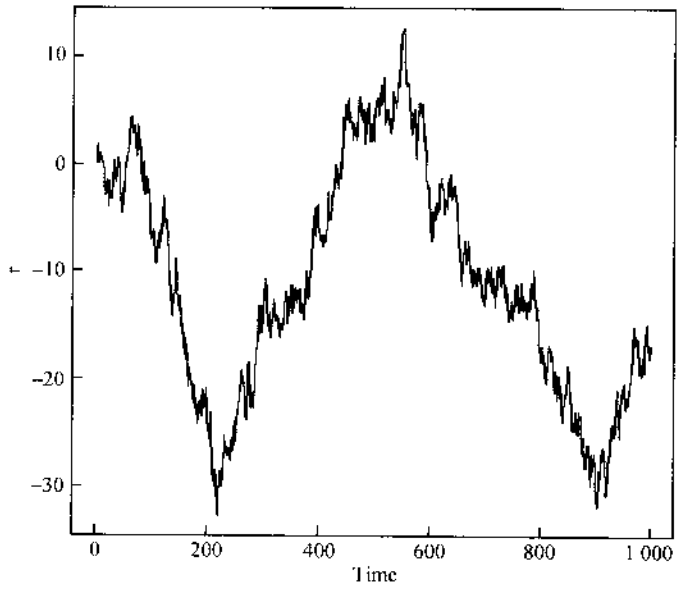


图 8-10　随机游走的时间序列图

图 8-10 是我们随机模拟生成的一个随机游走的时间序列。我们可以同实际的上证指数的随机游走的时间序列(如图 8-11 所示)作一个对比，可见它们有多么相似。这就解释了为什么大量的金融统计模型都是基于随机游走模型的。

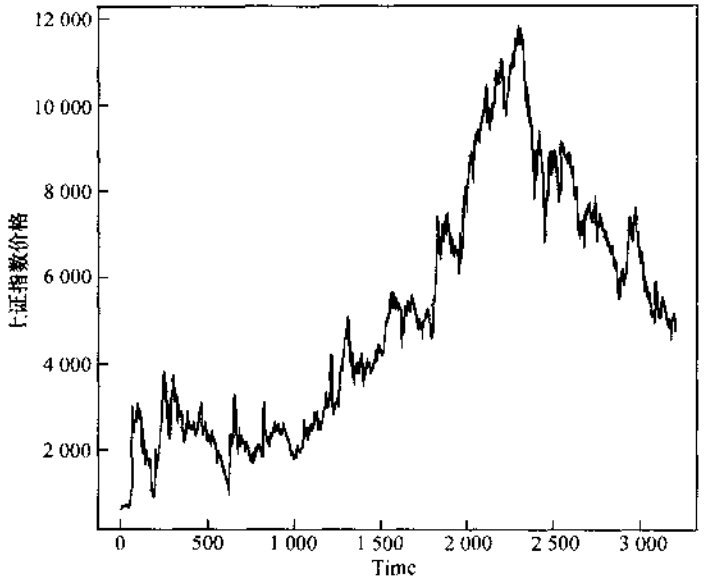


图 8-11　上证指数的时间序列图

例 2　（发散的随机游走）考虑如下随机游走模型：$r_t = \sqrt{2} \times r_{t-1} + \varepsilon_t$，并假设 $\text{var}(\varepsilon_t) = 1$。由此可见：$\text{var}(r_t) = 2 \times \text{var}(r_{t-1}) + 1$。反复迭代该公式，可以得到：

$$\text{var}(r_t) = 1 + 2 + 2^2 + \cdots + 2^t \times \text{var}(r_0)$$

如果我们假设 r_0 是一个常数（即 $\text{var}(r_0) = 0$），那么 $\text{var}(r_t) \approx 2^t$。这说明，该序列的方差随着时间的推移也越来越大，而且发散速度比随机游走还要快，因此是非平稳的，如图 8-12 所示。

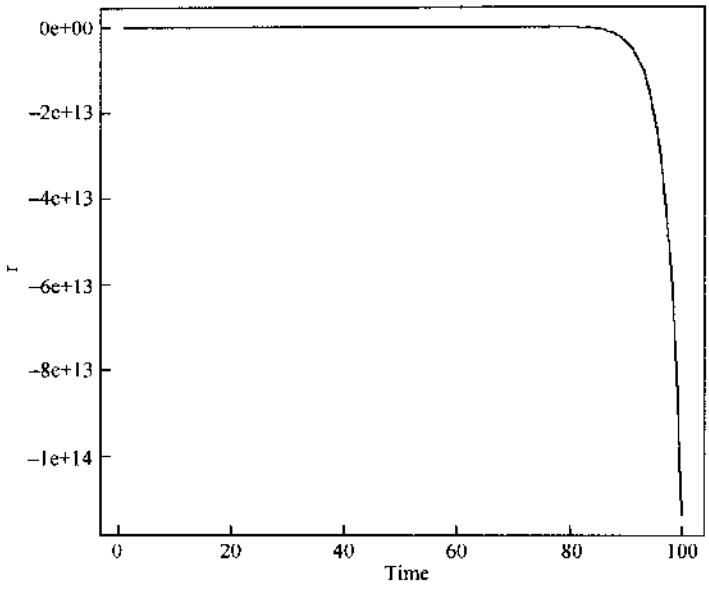


图 8-12　发散的随机游走时间序列图

例 3　(平稳的自回归模型)考虑如下随机游走模型:$r_t = 2^{-1/2} \times r_{t-1} + \varepsilon_t$,并假设 $\mathrm{var}(\varepsilon_t) = 1$,$E(\varepsilon_t) = 0$。由此可得:$\mathrm{var}(r_t) = 0.5 \times \mathrm{var}(r_{t-1}) + 1$。反复迭代该公式,可以得到:

$$\mathrm{var}(r_t) = 1 + 0.5 + 0.5^2 + \cdots + 0.5^t \times \mathrm{var}(r_0)$$

如果我们假设 r_0 是一个常数(即 $\mathrm{var}(r_0) = 0$),那么 $\mathrm{var}(r_t) \approx 2$。这说明,该序列的方差不会随着时间的推移而增加到无穷大,因此,该序列才有可能平稳。我们在该模型的基础上随机模拟生成了一个时间序列,如图 8-13 所示。

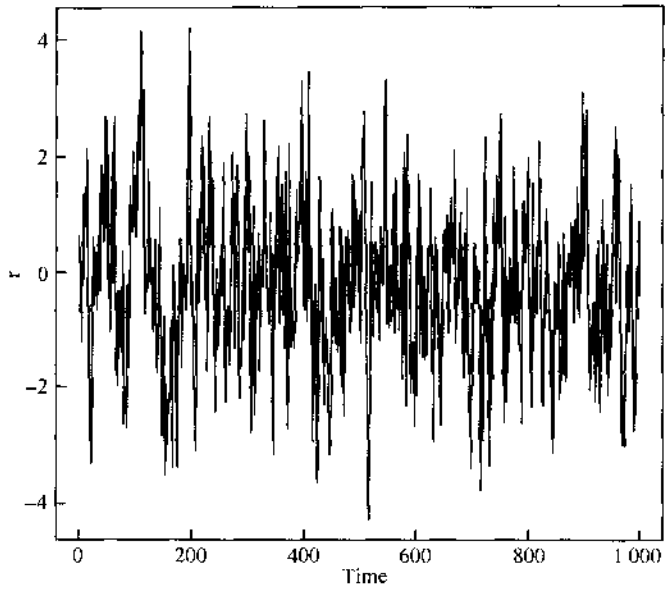


图 8-13　平稳自回归的时间序列图

从上面的几个例子可以发现，自回归模型的平稳与否，同自回归系数的大小高度相关。理论上我们可以证明，对于一阶自回归模型 $r_t = \phi_0 + \phi_1 \times r_{t-1} + \varepsilon_t$，其平稳性的充分必要条件为 $|\phi_1| < 1$。对于一般的自回归模型，其平稳性的充分必要条件为方程：

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

的所有复数根都落在单位圆以外。可见，这是一个非常数学化的条件。简单地说，自回归系数 $\phi = (\phi_0, \phi_1, \cdots, \phi_p)$ 的绝对值越小，该自回归模型越有可能平稳，否则越有可能非平稳。

第六节　模型估计与选择

结束了描述性分析以及平稳性的讨论后，我们将考虑在这样一个平稳的自回归模型假设下，如何获得参数估计。一个简单的办法就是最小二乘法。具体地说，我们可以通过优化以下最小二乘目标函数，即：

$$\hat{\phi} = \arg\min_{\phi} \sum_{t=p+1}^{T} (r_t - \phi_0 - \phi_1 r_{t-1} - \phi_2 r_{t-2} - \cdots - \phi_p r_{t-p})^2$$

来获得关于自回归系数的估计并进行相应的统计检验。所有过程同普通线性

模型非常类似。假设 $p=4$,那么在 R 中可以分析如下:

```
> ar(r,aic=F,order=4)

Call:
ar(x = r, aic = F, order.max = 4)

Coefficients:
      1         2         3         4
-0.0673    0.4123    0.0275    0.0353

Order selected 4  sigma^2 estimated as  0.001586
```

如果该模型是可靠的,我们就可以知道,当月失业率的对数变化率(r_t)同前一个月(r_{t-1})微弱负相关,同再前一个月(r_{t-2})高度正相关,而同 r_{t-3} 和 r_{t-4} 微弱正相关。这提示我们,$p=4$ 是不必要的,也许 $p=2$ 就足够了。那么实际应用中,我们应该如何选择自回归模型的阶数 p 呢?这是另外一种类型的模型选择问题。因此,第一章中所介绍的 AIC 和 BIC 都可以用来选择最优的阶数。如果我们决定采用 AIC,那么可以在 R 中轻松实现如下:

```
> fit=ar(r)
> fit

Call:
ar(x = r)

Coefficients:
      1         2
-0.0557    0.4260

Order selected 2  sigma^2 estimated as  0.001585
```

从中可以看到,根据 AIC 标准,R 认为 $p=2$ 是一个最优的选择。各个模型相应的 AIC 取值如图 8-14 所示。

```
> plot(fit$aic,type="b")
```

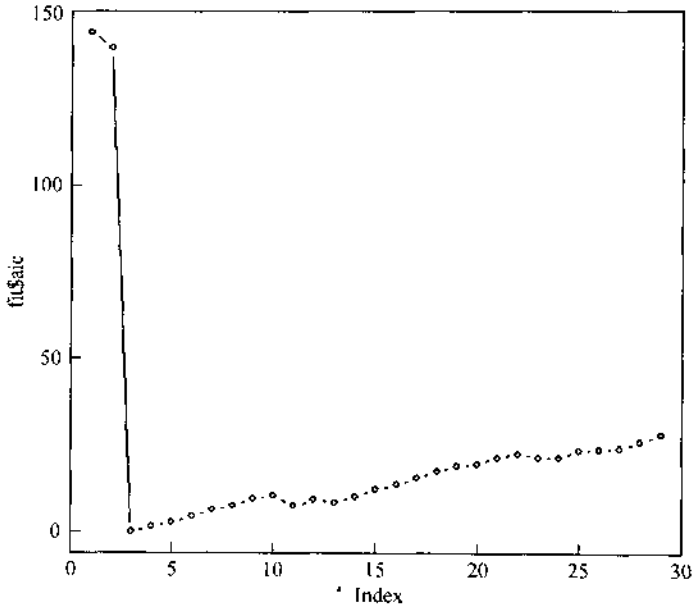


图 8-14　各个模型相应的 AIC 取值

在 $p=2$ 的情况下，拟合结果显示：当月失业率的对数变化率（r_t）同前一个月（r_{t-1}）微弱负相关，但是同再前一个月（r_{t-2}）高度正相关。

第七节　模型诊断

同线性模型一样，良好的自回归模型应该能够经得起模型诊断的检验。而对于一个时间序列模型而言，其模型诊断的主要内容就是判断模型分离出来的白噪音（即残差项）ε_t 是否还具有某些时间序列特征。简单地说，我们可以检查一下残差项 ε_t 的自相关系数。

从图 8-15 可以看到，除了零阶自相关系数 $\rho_0 = \mathrm{corr}(x_t, x_t) = 1$ 以外，所有其他阶的自相关系数都很小，均落在了虚线所限定的范围以内。因此，我们没有证据反对其他阶的自相关系数均为 0。这说明，该模型所分离出来的残差确实非常接近白噪音。从这个角度讲，前一节中 AIC 所挑选的模型经受住了考验。

```
> resid=fit$resid[-c(1:2)]
> acf(resid)
```

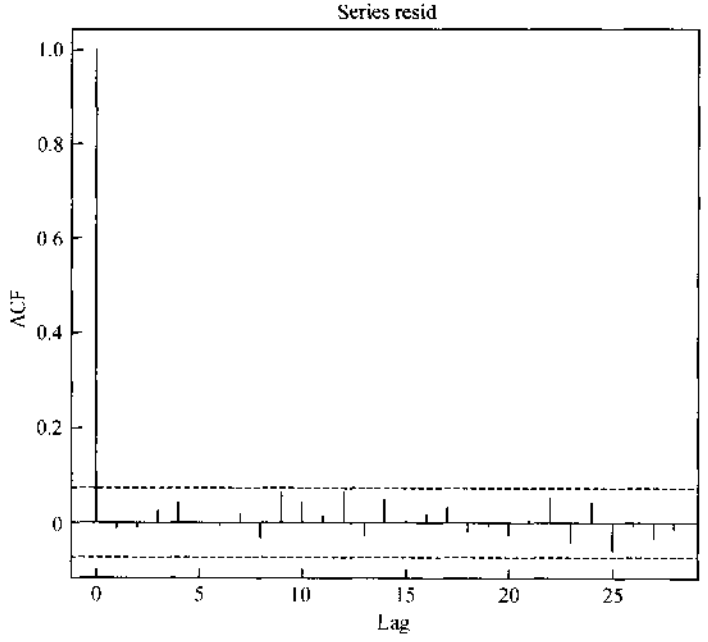


图 8-15　自相关系数图

同普通线性模型一样，我们还关心数据中有没有异常值。因此，我们对所分离出来的残差作时间序列图（如图 8-16 所示）。

```
> ts.plot(resid)
```

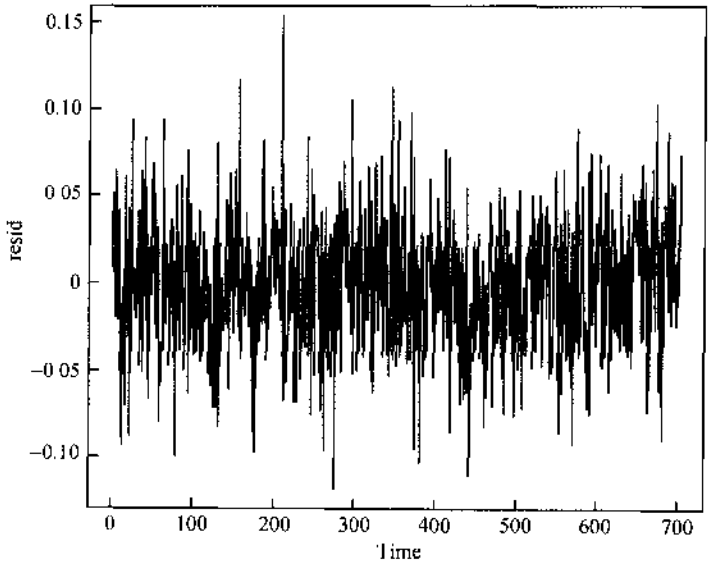


图 8-16　残差时间序列图

从图 8-16 中没有发现明显的异常值。最后，我们对残差的正态性进行检验，作 QQ 图（如图 8-17 所示）。

```
> qqnorm(resid)
```

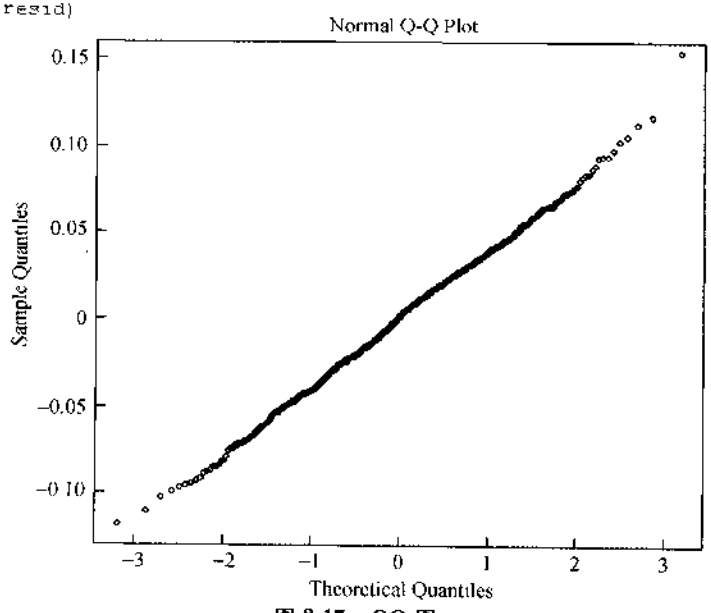


图 8-17　QQ 图

由此可见，残差的分布非常接近正态分布。因此，种种检验结果表明，该模型的拟合效果是比较理想的。

第八节　模型预测

时间序列模型的一个重要应用就是预测。由于失业率的对数变化率数据涵盖了总共707个月份，因此，我们考虑用第1至第600个观测建立二阶自回归模型，然后再用此模型对第601个观测予以预测。首先，我们对数据建立模型如下：

```
> r1=r[c(1:600)]
> r2=r[601]
> fit=ar(r1,aic=F,order=2)
> fit

Call:
ar(x = r1, aic = F, order.max = 2)

Coefficients:
      1         2
-0.0583    0.3992

Order selected 2  sigma^2 estimated as  0.001589
```

然后，我们对第601个月份的失业率的对数变化率预测如下：

```
> predict(fit,nhead=1)$pred
Time Series:
Start = 601
End = 601
Frequency = 1
[1] -0.003972244
```

而第601个月份的失业率对数变化率的真实值为 -0.03801186。单纯比较预测值和真实值难以判断我们的预测是否准确。为了获得更加可靠的判断，我们对第601至第707个月份中的每个月份都用其前600期的数据（即 $\{r_{t_0-1}, r_{t_0-2}, \cdots, r_{t_0-600}\}$）建立模型并预测 r_{t_0}。

```
> r.true=r[c(601:707)]
> r.hat=rep(0,107)
> for(i in 1:107){
+ tmp.r=r[c(i:(i+599))]
+ tmp.fit=ar(tmp.r,aic=F,order=2)
+ r.hat[i]=predict(tmp.fit,nhead=1)$pred
+ }
```

然后，我们定义相对预测误差（relative prediction error，RPE）为：

$$\text{RPE} = \frac{107^{-1}\sum_{t=601}^{707}(r_t^{\text{pred}} - r_t^{\text{true}})^2}{107^{-1}\sum_{t=601}^{707}(r_t^{\text{true}} - r^{\text{true}})^2}$$

其中，$r^{\text{true}} = 107^{-1}\sum_{t=601}^{707} r_t^{\text{true}}$。简单地说，RPE 的分子度量了自回归模型的预测误差，而分母则定义了最优的常数预测误差（即用一个常数预测每一期的失业率的对数变化率）。如果我们认为，常数预测是一种没有统计模型支持的无奈的预测方法，那么我们希望知道自回归模型在此基础上有多大的改进。在 R 中，我们具体计算如下：

```
> mean((r.hat-r.true)^2)/var(r.true)
[1] 0.78828
```

这说明，自回归模型的预测误差是常数预测误差的 78.8%，代表预测精度大约提高了 21.2%。因此，自回归模型的预测效果良好。为了更加直观地比较预测值同真实值之间的关系，我们作时间序列图（如图 8-18 所示）。

```
> ts.plot(r.true,col=4)
> points(c(1:107),r.hat,type="l",col=2,lty=2)
> legend(30,-0.075,c("True Value", "Predicted Value"),col=c(4,2),lty=c(1,2))
```

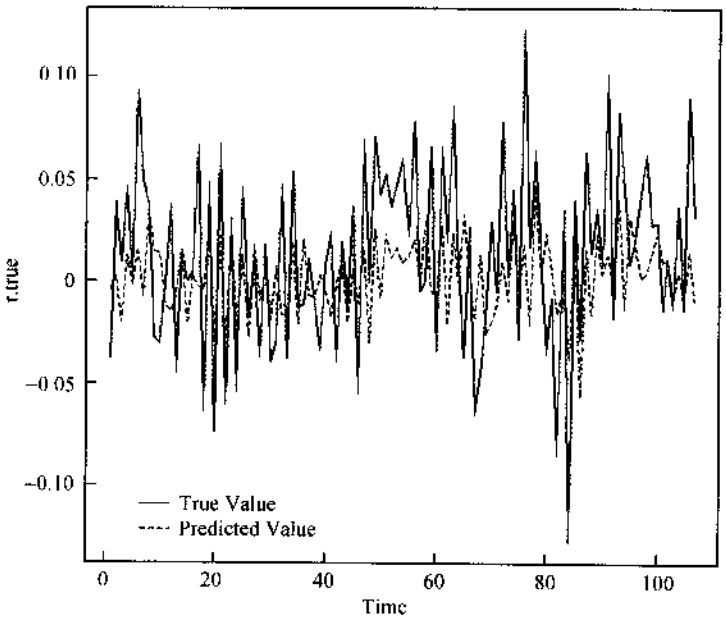


图 8-18　真实值与预测值比较的时间序列图

从图 8-18 可以看出，预测值和真实值之间是比较吻合的，而且在波动方向上表现得非常一致。

第九节　简单分析报告

某国失业率分析报告

内容提要　本报告利用时间序列的自回归方法对某国的失业率数据进行分析。我们发现，该国的失业率数据可以用一个二阶的自回归模型来拟合，并具有较好的预测能力。政府职能部门可以利用本报告所发现的规律来制定和调整相关的政策，而且可以利用报告中的模型对未来的失业率进行合理的预测。

一、研究目的

失业率是衡量一个国家或者地区就业状况最重要的指标之一，也是反映社会稳定性的重要指标之一。因此，分析并把握失业率的变化规律，对于相关机构，特别是政府职能部门，意义非常重大。本报告试图通过对某国失业率数据的分析，建立一个合理的计量经济学模型来回答以下两个问题：第一，失业率的变化是否有规律，如当月的失业率同过去几个月的失业率是否有关系？如果有，是什么样的关系？第二，如何基于失业率的变化规律对以后的失业率作合理的预测。

二、数据来源和相关说明

我们的数据详细记录了 1990 年 1 月至 2006 年 12 月间，该国各个月份的全国失业率(%)。历史经验表明，失业率的高低受月份的影响很大。因此，不同月份的失业率没有直接的可比性。所以，本报告所用数据并不是失业率的原始数据，而是经过季节性调整后的失业率。这样一来，不同月份的失业率具有了一定的可比性。这为我们后面的统计分析建模打下了良好的基础。

三、描述性分析

首先，我们通过时间序列图来获取对失业率的直观印象，如图 8-19 所示。

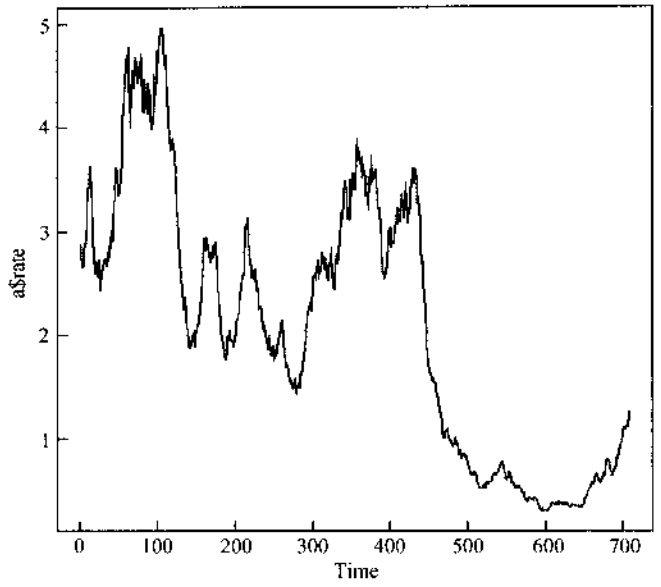


图 8-19　失业率的时间序列图

从图 8-19 我们可以看到，在过去的 27 年（1990—2006）中，该国的失业率基本上保持在 5% 以内，总体平均水平大约为 2%，但是不同时期的失业率平均水平差异巨大。该图也明确地告诉我们，该失业率数据是非平稳的时间序列，对其直接进行统计分析的结果是缺乏预测能力的，因为只有基于平稳的时间序列数据的统计模型才具有良好的预测能力。

因此，我们考虑对失业率的对数变化率进行分析，其时间序列图如图 8-20 所示。

由图 8-20 可知，同原始数据（即失业率）相比，新数据（即失业率的对数变化率）的平稳性有了极大的改善。因此，在以后的分析中我们将着重考虑失业率的对数变化率。

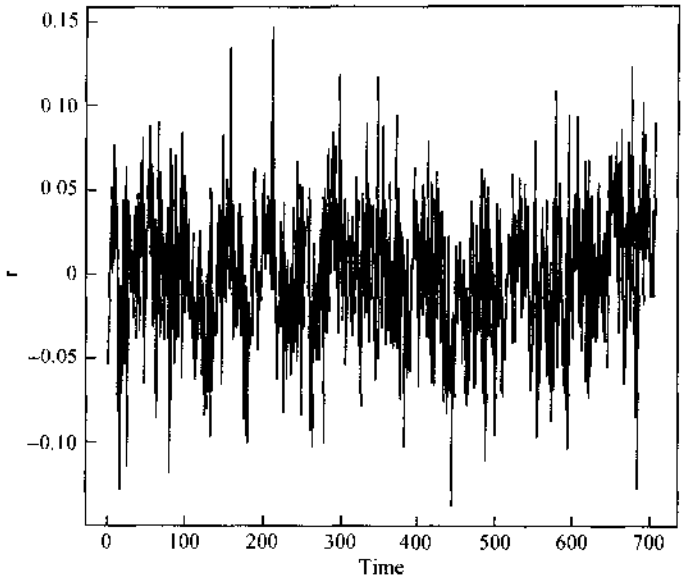


图 8-20　失业率对数变化率的时间序列图

在得到平稳的时间序列数据后，我们可以利用盒状图对失业率以及失业率的对数变化率作简单的描述性分析，如图 8-21 所示。

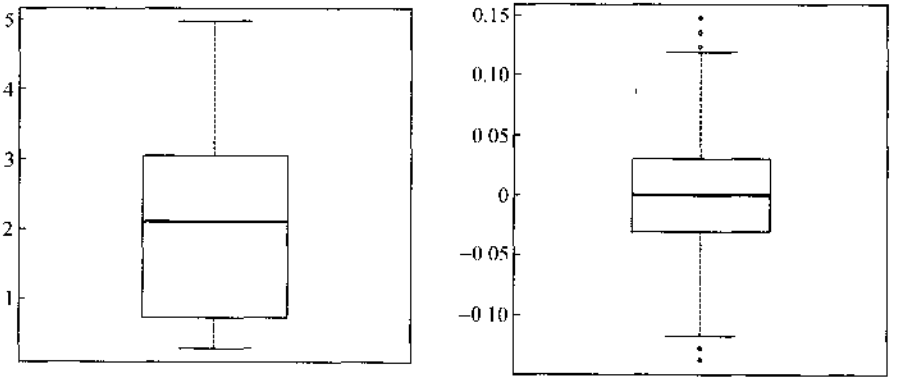


图 8-21　失业率及其对数变化率的盒状图

从盒状图中我们可以看出：对原始的失业率来说，在过去的 27 年中，该国的平均失业率（以中位数计）大约为 2%，处于非常低的水平，而且失业率最高也没有超过 5%；对失业率的对数变化率来说，其月度平均失业率的对数变化率（以中位数计）大约为 0，但是波动范围较大，最高接近 15%，而最低超过 −10%。

四、数据建模

1. 全模型分析

首先，我们观察失业率对数变化率的自相关系数图，如图 8-22 所示。

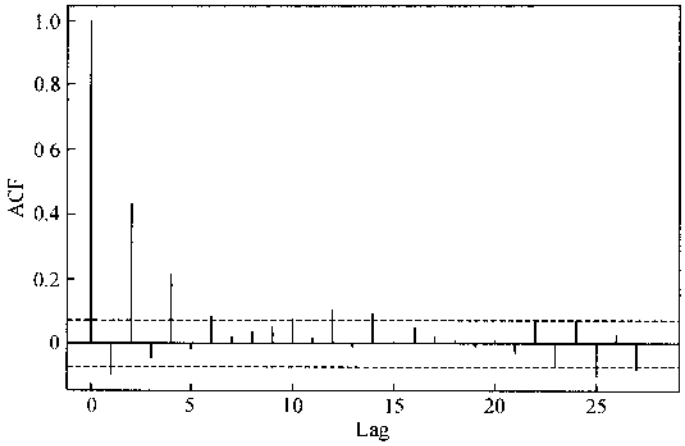


图 8-22　失业率对数变化率的自相关系数图

从图 8-22 中我们可以看出，只有为数不多的几个自相关系数明显地落在了虚线以外，即只有少数几个自相关系数是显著的。具体地说，四阶以后的自相关系数都不显著，因此我们首先考虑四阶的自回归模型。利用最小二乘法，我们可以获得全模型的估计结果，如表 8-1 所示。

表 8-1　全模型

变量名	系数估计值
滞后 1 期	-0.0673
滞后 2 期	0.4123
滞后 3 期	0.0275
滞后 4 期	0.0353

从表 8-1 中我们可以发现，当月失业率的对数变化率同前一个月微弱负相关，同再前一个月高度正相关，而同前三个月以及前四个月微弱正相关。这样的结果表明，影响当前月份失业率对数变化率的主要是再前一个月的失业率的对数变化率，即失业率对数变化率的影响的传递有一个月的滞后期。而且，失业率的对数变化率对后期的影响是递减的，具体来说，约有 41.23% 的对数变化率会传递到间隔一个月的月份中。

2. 模型选择及预测

从以上全模型的分析结果我们很容易发现，第三个和第四个系数非常小，有可能并不需要用到四阶的自回归模型。因此，我们用 AIC 的方法来选择最优的阶数。经过计算我们发现，根据 AIC 标准，二阶自回归模型是一个最优的选择。进一步，我们利用全部数据对二阶自回归模型进行拟合，得到回归系数的估计结果如表 8-2 所示。

表 8-2 AIC

变量名	系数估计值
滞后 1 期	-0.0557
滞后 2 期	0.4260

从表 8-2 我们可以知道，当月失业率的对数变化率同前一个月微弱负相关，而同再前一个月高度正相关。这同我们在全模型中的发现一致，但是该模型比全模型更为简单。

为保证模型的正确性，我们对该模型所分离出来的残差项进行检验，发现残差项非常接近白噪音。其中没有明显的异常值，而且残差分布非常接近正态分布。这些都表明，AIC 所挑选的模型经受住了检验，其拟合效果是比较理想的。

时间序列模型的一个重要应用就是预测。由于失业率的对数变化率数据涵盖了总共 707 个月份，因此，我们考虑用第 1 至第 600 个观测建立二阶自回归模型，然后用此模型对第 601 个观测进行预测。具体地说，我们利用前 600 个数据估计二阶自回归模型的系数，然后利用估计结果预测第 601 个月份的失业率。在得到预测值之后，我们可以与真实值进行比较来判断预测是否准确。为了获得更加可靠的判断，我们对第 601 至第 707 个月份中的每一个月份都用其前 600 期的数据估计一个二阶自回归模型，然后利用估计结果预测当前月份的失业率。最后我们得到相对预测误差为 78.8%，这代表我们的模型的预测精度大约提高了 21.2%，可见预测效果良好。为了更加直观地比较预测值同真实值之间的关系，我们对真实值和预测值作时间序列图，如图 8-23 所示。从图中可以看出，预测值和真实值之间是比较吻合的，而且在波动方向上表现得很一致。

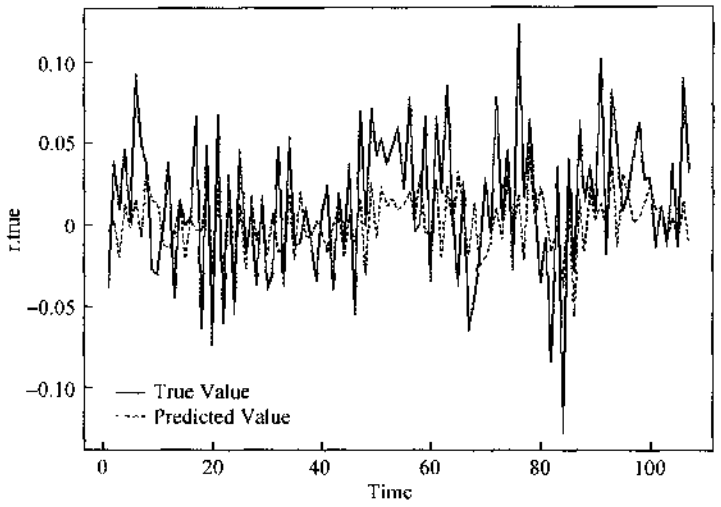


图 8-23　真实值与预测值相比较的时间序列图

五、结论及建议

从以上的分析结果，我们可以看到失业率确实存在一定的变化规律。具体来说，当月失业率的对数变化率同前一个月微弱负相关，而同再前一个月高度正相关。发现这样一个规律能帮助政府的相关职能部门制定更合理的政策。例如，如果发现前两个月的失业率的对数变化率较大，则当月的失业率的对数变化率也就倾向于较大。在具体实施时，利用本报告中所得到的模型可以对当月的失业率有一个合理的预测。根据预测结果，政府相关职能部门可以选择相应的措施，并结合以往的经验来对失业率进行有效调控。具体来说，如果预测结果发现当月的失业率偏高，那么政府就应该考虑采取积极的货币政策或者财政政策来促进就业，降低失业率；如果失业率在正常合理的范围内，则应该更多地考虑采取相对稳健的政策来保持经济的稳定。

[讨论总结]

本章以失业率预测为例，系统演示并讲解了什么是时间序列数据，以及自回归这种处理时间序列数据的有效工具。通过对本章的学习，读者应该能够了解：什么是时间序列数据，什么时候可以使用自回归模型，以及如何使用。在 *R* 语言学习方面，读者应该掌握相关的时间序列分析的命令。在统计理论方面，读者应该掌握以下概念：自相关系数、时间序列的平稳性、自回归模型等。对相关理论渴望深入了解的读者请参阅 Tsay(2005)。

附录　程序及注释

```
a = read.csv("D:/WORKING/teaching/Practical Business Data Analysis/case/CH4/rate.csv")
                                                    # 读入 csv 格式的数据,并赋值给 a
ts.plot(a$rate
                                                    # 画出 rate 的时间序列图
r = diff(log(a$rate))
                                                    # 计算 rate 的对数变化率,并赋值给 r
ts.plot(r)
                                                    # 画出 r 的时间序列图
boxplot(a$rate)
                                                    # 画出 rate 的盒状图
boxplot(r)
                                                    # 画出 r 的盒状图
acf(a$rate)
                                                    # 计算 rate 的自相关系数
acf(r)
                                                    # 计算 r 的自相关系数
ar(r,aic = F,order = 4)
                                                    # 对 r 拟合 4 阶的自回归模型
fit = ar(r)
                                                    # 对 r 拟合自回归模型,并用 AIC 准则选择最优模型
fit
                                                    # 显示模型 fit 的各方面细节,包括估计值等
plot(fit$aic,type = "b")
                                                    # 画出各阶自回归模型的 AIC 取值
resid = fit$resid[ - c(1:2)]
                                                    # 从模型 fit 中取出残差项,赋值给 resid
acf(resid)
                                                    # 计算残差项 resid 的自相关系数
ts.plot(resid)
                                                    # 画出残差项 resid 的时间序列图
qqnorm(resid)
                                                    # 画出残差项 resid 的 QQ 图,检查其正态性
r1 = r[c(1:600)]
                                                    # 取出 r 的前 600 期观测
r2 = r[601]
                                                    # 取出 r 的第 601 期观测
fit = ar(r1,aic = F,order = 2)
                                                    # 利用数据 r1,即前 600 期观测,拟合 2 阶的自回归模型
```

```
fit                                                    # 显示模型 fit 的各方面细节,包括估计值等
predict(fit,nhead = 1)$pred                            # 利用模型 fit 预测下一期的观测值
r.true = r[c(601:707)]                                 # 取出 r 中的第 601 期至第 707 期观测,作为比较的真实值
r.hat = rep(0,107)                                     # 为 r.hat 赋初值 0
for(i in 1:107){                                       # 循环 107 次
  tmp.r = r[c(i:(i + 599))]                            # 取出第 i 期至第 i + 599 期观测,赋值给 tmp.r
  tmp.fit = ar(tmp.r,aic = F,order = 2)                # 利用数据 tmp.r 拟合 2 阶的自回归模型
  r.hat[i] = predict(tmp.fit,nhead = 1)$pred           # 利用模型 tmp.fit 预测下一期的值,并赋值给 r.hat[i]
}
mean((r.hat - r.true)^2)/var(r.true)                   # 计算自回归模型的预测误差相对于常数预测误差的比例
ts.plot(r.true,col = 4)                                # 画出 r.true,即真实值的序列图
points(c(1:107),r.hat,type = "l",col = 2,lty = 2)      # 画出 r.hat,即预测值的序列图
legend(30, -0.075,c("True Value", "Predicted Value"),col = c(4,2),lty = c(1,2))
                                                       # 为两组数据加上标志,便于区分
```

参 考 文 献

[1] Draper, N. R. and Smith, H. (1981), *Applied Regression Analysis*, 2nd Edition, Wiley, New York.

[2] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*, Springer, New York.

[3] Fleming, T. R. and Harrington, D. P. (1991), *Counting Process and Survival Analysis*, Wiley, New York.

[4] Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, Wiley, New York.

[5] McCullagh, P. and Nelder, J. A. (1999), *Generalized Linear Models*, Wiley, New York.

[6] Milliken, G. A. and Johnson, D. E. (2002a), *Analysis of Messy Data I: Designed Experiments*, Chapman & Hall, New York.

[7] Milliken, G. A. and Johnson, D. E. (2002b), *Analysis of Messy Data III: Analysis of Covariance*, Chapman & Hall, New York.

[8] Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd Edition, New York.

[9] Tsay, R. S. (2005), *Analysis of Financial Time Series*, Wiley, New York.

[10] Venables, W. N. and Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus*, Springer, New York.