面向21世纪课程教材
Textbook Series for 21st Century

线性统计模型

线性回归与方差分析

王松桂 陈 敏 陈立萍 编

高等教育出版社
HIGHER EDUCATION PRESS

面向21世纪课程教材
Textbook Series for 21st Century

线性统计模型

线性回归与方差分析

王松桂　陈　敏　陈立萍　编

高等教育出版社
HIGHER EDUCATION PRESS

(京)112 号

内容简介

本书是教育部“高等教育面向 21 世纪教学内容和课程体系改革计划”的研究成果,是面向 21 世纪课程教材.本书主要讲授线性回归模型和方差分析模型,内容包括正态分布、最小二乘估计、岭估计、主成分估计、回归诊断、假设检验与预测、回归方程的选择和方差分析,并对几种具有较复杂误差结构的线性回归模型作了概括介绍.书中前六章选配了适量的习题,附录中对书中用到的矩阵论的重要事实给出了详细证明.

本书可作为高等学校理、工、农、医、经济、管理等专业有关学科的本科生或研究生教科书.

图书在版编目(CIP)数据

线性统计模型:线性回归与方差分析/王松桂编. —北京:高等教育出版社,1999
ISBN 7-04-007605-5

Ⅰ.线… Ⅱ.王… Ⅲ.①数理统计-统计模型②线性回归③方差分析 Ⅳ.0212

中国版本图书馆 CIP 数据核字(1999)第 24402 号

线性统计模型——线性回归与方差分析
王松桂 陈敏 陈立萍 编

出版发行	高等教育出版社		
社　　址	北京市东城区沙滩后街 55 号	邮政编码	100009
电　　话	010—64054588	传　　真	010—64014048
网　　址	http://www.hep.edu.cn		
经　　销	新华书店北京发行所		
排　　版	高等教育出版社照排中心		
印　　刷	国防工业出版社印刷厂		
纸张供应	山东高唐纸业集团总公司		
开　　本	787×960　1/16	版　　次	1999 年 9 月第 1 版
印　　张	15.25	印　　次	1999 年 9 月第 1 次印刷
字　　数	280 000	定　　价	16.40 元

凡购买高等教育出版社图书,如有缺页、倒页、脱页等质量问题,请在所购图书销售部门联系调换。

责任编辑　高尚华
封面设计　张　楠
责任绘图　吴文信
版式设计　马静如
责任校对　胡晓琪
责任印制　杨　明

前　言

线性统计模型是一类很重要的统计模型，它包括了线性回归模型、方差分析模型等应用十分广泛的许多模型，同时线性模型的理论和方法也是学习和研究其它统计方法的基础．正是由于这些原因，线性统计模型不仅已成为统计专业本科生和研究生的必修课，而且也是生物、医学、经济、管理、商业、金融、工程技术以及社会科学等学科本科生和研究生统计课程的重要内容．为了适应各方面的需要，我们在教育部高教司的关心和支持下编写了这本教材．

本书的目的是为需要学习线性统计模型基础知识的各专业的学生提供一本教材，因此，阅读本书所要求的数学方面的预备知识并不多．我们认为，读者如果掌握了工科大学的微积分、线性代数和初等概率统计知识，就可以顺利阅读本书的绝大部分内容．

全书共分七章．前两章是预备性知识．第一章通过例子引进了我们所要研究的模型，第二章介绍了随机向量特别是正态向量的基础知识．接下来的三章，即第三、四、五章对线性回归模型的估计和检验做了系统讨论．第六章介绍了方差分析模型．第七章概要介绍了几类在经济、生物、医学等领域颇为有用的线性回归模型．为了帮助读者理解基本内容，掌握其中的方法，书中给出了一些应用实例．书末有三个附录．附录1给出了矩阵论的一些预备知识．附录2提供了书中部分例子计算机软件如SAS软件、SPSS软件的计算机输出结果，以帮助读者了解如何使用计算机求解线性模型．除第七章外，各章末配有一定数量的习题．

本书第一至第四章由王松桂执笔，第五章和第七章由陈敏执笔，第六章由陈立萍执笔，最后由王松桂统一修改定稿．陈立萍选配了大部分章末的习题．陈立萍和程维虎曾经在北京工业大学应用数学系讲授过本书；王松桂曾多次在北京工业大学工科博士生“数学模型”课上讲授过本书的部分内容．

本书的写作得到国家自然科学基金、北京市自然科学基金和北京市教委科技发展计划支持．另外，1997年12月在苏州大学召开了本书的审稿会．汪仁官、茆诗松、王静龙、陈庆云、傅珏生和高尚华等专家对本书进行了认真评审，提出了许多宝贵意见．这对提高本书质量起了重要作用．趁本书出版之际，我们谨向以上单位和专家表示衷心感谢．

由于作者水平所限，书中的缺点和错误在所难免，恳请同行和广大读者批评指正.

编 者

1999 年 4 月 30 日

目　录

第一章 引 论

线性统计模型是现代统计学中应用最为广泛的模型之一，而且也是其它统计模型研究或应用的基础.之所以如此，其原因主要是

1. 在现实世界中，许多量之间具有线性或近似的线性依赖关系.

2. 在现实世界中，虽然许多量之间的关系是非线性的，但是经过适当的变换，变换过后的新变量之间具有近似的线性关系.

3. 线性关系是数学中最基本的关系，因而比较容易处理.在数学中已经积累了处理线性关系的丰富的理论与方法，为实际应用提供了坚实的理论依据和有效算法.

本章我们通过一些实例引进线性统计模型，使读者对这种模型丰富的实际背景有一定了解，这对后面要引进的一些统计概念和方法的理解将是大有裨益的.

§1.1 线性回归模型

在现实世界中，存在着大量这样的情况：两个变量例如 X 和 Y 有一些依赖关系.由 X 可以部分地决定 Y 的值，但这种决定往往不很确切.常常用来说明这种依赖关系的最简单、直观的例子是体重与身高.若用 X 表示某人的身高，用 Y 表示他的体重.众所周知，一般说来，当 X 大时，Y 也倾向于大，但由 X 不能严格地决定 Y.又如，城市生活用电量 Y 与气温 X 有很大的关系.在夏天气温很高或冬天气温很低时，由于室内空调、冰箱等家用电器的使用，可能用电量就高.相反，在春秋季节气温不高也不低，用电量就可能少.但我们不能由气温 X 准确地决定用电量 Y.类似的例子还很多.变量之间的这种关系称为“相关关系”，回归模型就是研究相关关系的一个有力工具.

在以上诸例中，Y 通常称为因变量或响应变量，X 称为自变量或预报变量.我们可以设想，Y 的值由两部分组成：一部分是由 X 能够决定的部分，它是 X 的函数，记为 $f(X)$.而另一部分则由其它众多未加考虑的因素（包括随机因素）所产生的影响，它被看作随机误差，记为 e.于是我们得到如下模型：

$$Y = f(X) + e. \tag{1.1.1}$$

这里 e 作为随机误差，我们有理由要求它的均值 $E(e)=0$，其中 $E(\cdot)$ 表示随机

变量的均值.

特别,当 $f(X)$是线性函数 $f(X)=\beta_0+\beta_1 X$ 时,我们得到

$$Y = \beta_0 + \beta_1 X + e. \tag{1.1.2}$$

在这个模型中,若忽略掉 e,它就是一个通常的直线方程.因此,我们称(1.1.2)为线性回归模型或线性回归方程.关于"回归"一词的由来,我们留在后面作解释.常数项 β_0 是直线的截距,β_1 是直线的斜率,也称为回归系数.在实际应用中,β_0 和 β_1 皆是未知的,需要通过观测数据来估计.

假设自变量 X 分别取值为 $x_1, x_2, \cdots, x_n$ 时,因变量 Y 对应的观测值分别为 $y_1, y_2, \cdots, y_n$.于是我们有 n 组观测值 (x_i, y_i), $i=1, \cdots, n$.如果 Y 与 X 有回归关系(1.1.2),则这些 (x_i, y_i) 应该满足

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \cdots, n, \tag{1.1.3}$$

这里 e_i 为对应的随机误差.基于(1.1.3),应用适当的统计方法(这将在第三章讨论)可以得到 β_0 和 β_1 的估计值 $\hat{\beta}_0, \hat{\beta}_1$,将它们代入(1.1.2),再略去误差项 e_i 得到

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X, \tag{1.1.4}$$

称之为经验回归直线,也称为经验回归方程.这里"经验"两字表示这个回归直线是基于前面的 n 次观测数据 (x_i, y_i), $i=1, \cdots, n$ 而获得的,在许多文献中,人们往往把"经验"两字省略掉.经过适当的统计检验后,我们可以认为(1.1.4)描述了因变量 Y 与自变量 X 之间的相关关系.

例 1.1.1 肥胖是现代社会人们关注的一个重要问题,那么体重多少才算是肥胖呢?这当然跟每个人的身高有关,于是许多学者应用直线回归方法研究人的体重与身高的关系.假设 X 表示身高(cm),Y 表示体重(kg).我们假设 Y 与 X 之间具有回归关系(1.1.2).在这里误差 e 表示除了身高 X 之外,所有影响体重 Y 的其它因素,例如遗传因素、饮食习惯、体育锻炼多少等.为了估计其中的参数 β_0 和 β_1,研究者测量了很多人的身高 x_i 和他们的体重 y_i, $i=1, \cdots, n$,得到关系(1.1.3).从而应用统计方法可以估计出 β_0 和 β_1.一种研究结果是,若用 $X-150$ 作自变量,则得 $\hat{\beta}_0=50$, $\hat{\beta}_1=0.6$,也就是说我们有经验回归直线.

$$Y = 50 + (X - 150) \times 0.6.$$

我们可以把它改写成如下形式:

$$Y = -40 + 0.6X, \tag{1.1.5}$$

这个经验回归方程在一定程度上描述了体重与身高的相关关系.给定 X 的一个具体值 x_0,我们可以算出对应的 Y 值 $y_0=-40+0.6x_0$.例如某甲身高 $x_0=160$ (cm),代入(1.1.5)可以算出对应 $y_0=56$(kg).我们称 56kg 为身高是 160cm 的

人的体重的预测.这就是说,对于一个身高 160cm 的人,我们预测他的体重大致为 56kg,但实际上,他的体重不可能恰为 56kg.可能比 56kg 多,也可能比 56kg 少.

例 1.1.2 我们知道,一个公司的商品销售量与其广告费有密切关系,一般说来在其它因素(如产品质量等)保持不变的情况下,它用在广告上的费用愈高,它的商品销售量也就会愈多.但这也只是一种相关关系.某公司为了进一步研究这种关系.用 X 表示在某地区的年度广告费,Y 表示年度商品销售量.根据过去一段时间的销售记录(x_i, y_i),$i=1,\cdots,n$,采用线性回归模型(1.1.3),假定计算出$\hat{\beta}_0=1\,608.5$,$\hat{\beta}_1=20.1$,于是得到经验回归直线

$$Y = 1\,608.5 + 20.1X. \tag{1.1.6}$$

这个经验回归直线告诉我们,广告费 X 每增加一个单位,该公司销售收入就增加 20.1 个单位.如果某地区人口增加得很快,那么很可能人口总数也是影响销售量的一个重要因素.若记 X_1 为年度广告费,X_2 为某地区人口总数.我们可以考虑如下含两个自变量的线性回归模型:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e. \tag{1.1.7}$$

同样,根据记录的历史数据,应用适当统计方法可以估计出 β_i,$i=0,1,2$.假定估计出的

$$\hat{\beta}_0 = 320.3, \quad \hat{\beta}_1 = 18.4, \quad \hat{\beta}_2 = 0.2,$$

则我们得到经验回归方程

$$Y = 320.3 + 18.4X_1 + 0.2X_2. \tag{1.1.8}$$

从这个经验回归方程我们可以看出,当广告费 X_1 增加或人口总数 X_2 增加时,商品销售量都增加,且当人口总数保持不变时,广告费每增加 1 个单位,销售量增加 18.4 个单位.而当广告费保持不变,而该地区人口总数每增加一个单位,该公司销售量增加 0.2 个单位.当然,在实际应用中,并不是每个经验回归方程都能描述变量之间的客观存在的真正的关系.关于这一点,将在第四章详细讨论.

在实际问题中,影响因变量的主要因素往往很多,这就需要考虑含多个自变量的回归问题.假设因变量 Y 和 $p-1$ 个自变量 $X_1,\cdots,X_{p-1}$之间有如下关系:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + e, \tag{1.1.9}$$

这是多元线性回归模型,其中 β_0 为常数项,$\beta_1,\cdots,\beta_{p-1}$为回归系数,$e$ 为随机误差.

假设我们对 $Y, X_1,\cdots,X_{p-1}$进行了 n 次观测,得到 n 组观测值

$$x_{i1},\cdots,x_{i,p-1},y_i, \qquad i=1,\cdots,n,$$

它们满足关系式

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1} + e_i, \qquad i = 1,\cdots,n, \quad (1.1.10)$$

这里 e_i 为对应的随机误差.引进矩阵记号

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} \cdots x_{1,p-1} \\ 1 & x_{21} \cdots x_{2,p-1} \\ \vdots & \vdots \quad\quad \vdots \\ 1 & x_{n1} \cdots x_{n,p-1} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix},$$

(1.1.10)就写为如下简洁形式:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad (1.1.11)$$

这里 $\boldsymbol{y}$ 为 $n \times 1$ 的观测向量.$\boldsymbol{X}$ 为 $n \times p$ 已知矩阵,通常称为设计矩阵.对于线性回归模型,术语"设计矩阵"中的"设计"两字并不蕴含任何真正设计的含义,只是习惯用法而已.近年来,有一些学者建议改用"模型矩阵".但就目前来讲,沿用"设计矩阵"者居多.$\boldsymbol{\beta}$ 为未知参数向量.其中 β_0 称为常数项,而 $\beta_1,\cdots,\beta_{p-1}$ 为回归系数.而 $\boldsymbol{e}$ 为 $n \times 1$ 随机误差向量,关于 $\boldsymbol{e}$ 最常用的假设是:

(a) 误差项均值为零,即 $E(e_i)=0, i=1,\cdots,n$.

(b) 误差项具有等方差,即

$$\mathrm{Var}(e_i) = \sigma^2, \qquad i = 1,\cdots,n. \quad (1.1.12)$$

(c) 误差是彼此不相关的,即

$$\mathrm{Cov}(e_i,e_j) = 0, \quad i \neq j, \quad i,j = 1,\cdots,n.$$

通常称以上三条为 Gauss－Markov 假设.模型(1.1.11)和假设(1.1.12)构成了我们以后要讨论的最基本的线性回归模型.

在 Gauss－Markov 假设中,第一条表明误差项不包含任何系统的趋势,因而观测值 y_i 的均值

$$E(y_i) = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1}, \qquad i = 1,\cdots,n.$$

这就是说,观测值 y_i 大于或小于其均值 $E(y_i)$ 的波动完全是一种随机性的,这种随机性来自误差项 e_i.我们知道,一个随机变量的方差刻画了该随机变量取值散布程度的大小,因此假设(b)要求 e_i 等方差,也就是要求不同次的观测 y_i 在其均值附近波动程度是一样的.这个要求有时显得严厉一些.在一些情况下,我们不得不放松为 $\mathrm{Var}(e_i)=\sigma_i^2, i=1,\cdots,n$,这种情况将在 §7.2 讨论.第三条假设等价于要求不同次的观测是不相关的.在实际应用中这个假设比较容易满足.但是在一些实际问题中,误差往往是相关的.这时估计问题比较复杂,本书中不少地方要讨论这种情形.

对于模型(1.1.10),假设 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_{p-1})'$ 为 $\boldsymbol{\beta}$ 的一种估计,将它们代入(1.1.9),并略去其中的误差项 e,得到经验回归方程

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{p-1} X_{p-1}. \tag{1.1.13}$$

和单个回归自变量的情形一样，这个经验回归方程是否真正描述了因变量 Y 与自变量 $X_1, \cdots, X_{p-1}$ 之间的关系，还需要适当的统计检验.

上面我们讨论的都是线性回归模型. 有一些模型虽然是非线性的，但经过适当变换，可以化为线性模型.

例 1.1.3 在经济学中，著名的 Cobb－Douglas 生产函数为

$$Q_t = aL_t^b K_t^c, \tag{1.1.14}$$

这里 Q_t, L_t 和 K_t 分别为 t 年的产值、劳力投入量和资金投入量，a, b 和 c 为参数. 在上式两边取自然对数，得到

$$\ln(Q_t) = \ln a + b\ln(L_t) + c\ln(K_t).$$

若令

$$y_t = \ln(Q_t), x_{t1} = \ln(L_t), x_{t2} = \ln(K_t),$$

$$\beta_0 = \ln a, \beta_1 = b, \beta_2 = c,$$

则再加上误差项，便得到线性关系

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t, \quad t = 1, \cdots, T. \tag{1.1.15}$$

因此我们把非线性模型(1.1.14)化成了线性模型.

例 1.1.4 多项式回归模型

假设因变量 Y 和自变量 X 之间具有关系

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e.$$

这是三次多项式回归模型. 若令 $X_1 = X, X_2 = X^2, X_3 = X^3$，则有

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$

这就是一个线性模型. 从这里我们看到，线性模型(1.1.9)的基本特征是：它关于未知参数 $\beta_1, \beta_2, \cdots, \beta_{p-1}$ 是线性的.

在结束这一节之前，我们解释一下“回归”一词的由来. “回归”英文为“regression”，是由英国著名生物学家兼统计学家高尔顿(Galton)在研究人类遗传问题时提出的. 为了研究父代与子代身高的关系，高尔顿收集了 1 078 对父亲及其一子的身高数据. 用 X 表示父亲身高，Y 表示儿子身高. 单位为英寸(1 英寸为 2.54cm). 将这 1 078 对 (x_i, y_i) 标在直角坐标纸上，他发现散点图大致呈直线状. 也就是说，总的趋势是父亲的身高 X 增加时，儿子的身高 Y 也倾向于增加，这与我们的常识是一致的. 但是，高尔顿对数据的深入分析，发现了一个很有趣的现象——回归效应.

因为这 1 078 个 x_i 值的算术平均值 $\bar{x} = 68$ 英寸，而 1 078 个 y_i 值的平均值为 $\bar{y} = 69$ 英寸，这就是说，子代身高平均增加了 1 英寸. 人们自然会这样推想，若父亲身高为 x，他儿子的平均身高大致应为 $x + 1$，但高尔顿的仔细研究所得

结论与此大相径庭.他发现,当父亲身高为72英寸时(请注意,比平均身高 $\bar{x}=68$ 英寸要高),他们的儿子平均身高仅为71英寸.不但达不到预期的 $72+1=73$ 英寸,反而比父亲身高低了1英寸.反过来,若父亲身高为64英寸(请注意,比平均身高 $\bar{x}=68$ 英寸要矮),他们的儿子平均身高为67英寸,竟比预期的 $64+1=65$ 英寸高出了2英寸.这个现象不是个别的,它反映了一个一般规律:即身高超过平均值 $\bar{x}=68$ 英寸的父亲,他们的儿子的平均身高将低于父亲的平均身高.反之,身高低于平均身高 $\bar{x}=68$ 英寸的父亲,他们的儿子的平均身高将高于父亲的平均身高.高尔顿对这个一般结论的解释是:大自然具有一种约束力,使人类身高的分布在一定时期内相对稳定而不产生两极分化,这就是所谓的回归效应.通过这个例子,高尔顿引进了"回归"一词.用他的数据,可以计算出儿子身高 Y 与父亲身高 X 的经验关系

$$Y=35+\frac{1}{2}X,$$

它代表一条直线,人们也就把这条直线称为回归直线.当然,这个经验回归直线只反映了变量相关关系中具有回归效应的一种特殊情况,对更多的相关关系,并非都是如此.特别是在涉及多个自变量的情况中,回归效应便不复存在.因此将(1.1.9)或(1.1.11)或(1.1.13)分别称为线性回归模型和经验回归方程,并把对应的统计分析称为回归分析,不一定恰当.但"回归"这个名词沿用已久,实无改变之必要与可能.

§1.2 方差分析模型

在上节引进的线性回归模型中,所涉及的自变量一般来说都可以是连续变量,研究的基本目的则是寻求因变量与自变量之间客观存在的依赖关系.而本节所要引进的模型则不同,它的自变量是示性变量,这种变量往往表示某种效应的存在与否,因而只能取0,1两个值.这种模型是比较两个或多个因素效应大小的一种有力工具.因为比较因素效应的统计分析在统计学上叫做方差分析,所以对应地,人们将这种模型称为方差分析模型.在一些文献中,也把这种模型称为试验设计模型,这是因为它所分析的数据往往跟一个预先安排的试验相联系.

例1.2.1 某农业科学研究机构欲比较三种小麦品种的优劣,安排了一种比较试验.为保证试验结果的客观性,他们选择了六块面积相等,土质肥沃程度一样的田地,每一种小麦播种在其中的两块田内,并给予几乎完全相同的田间管理.

这是一个简单的试验,用 y_{ij} 表示种第 i 种小麦的第 j 块田的产量,那么我们可以对 y_{ij} 作如下分解:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1,2,3; j = 1,2. \tag{1.2.1}$$

这里 μ 表示总平均值，α_i 表示第 i 种小麦品种的效应，e_{ij} 是随机误差，它表示所有其它未加控制因素以及各种误差的总效应. 若用矩阵记号，则(1.2.1)变为

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}.$$

记

$$\boldsymbol{y} = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32})',$$
$$\boldsymbol{e} = (e_{11}, e_{12}, e_{21}, e_{22}, e_{31}, e_{32})',$$

它们分别是 6×1 的观测向量和随机误差向量.

$\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3)'$ 为 4×1 未知参数向量，

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

这是 6×4 矩阵，称为设计矩阵. 引进这些矩阵之后，上面的模型就具有形式

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}. \tag{1.2.2}$$

它和上节引进的线性回归模型(1.1.11)形式上完全一样. 不同的是在(1.1.11)中，除第 1 列之外，$\boldsymbol{X}$ 的元素原则上可以取任意连续值，而对现在的模型，设计矩阵 $\boldsymbol{X}$ 的所有元素只能取 0 和 1 两个值. $\boldsymbol{X}$ 的第 2,3 和 4 列分别对应于小麦品种 1,2 和 3. 例如，第 2 列对应于第一种小麦品种. 在这一列中，1 表示在对应的那块田中，播种的是第一种小麦，即效应 α_1 存在. 0 表示在对应的那块田中，播种的不是第一种小麦，其余类推. 我们进行统计分析的目的是比较这三种小麦品种，即比较它们的效应 α_1, α_2 和 α_3 的大小.

这个例子所引进的模型是方差分析模型中最简单的一种，称为单因素方差分析模型，这是因为它只涉及“小麦品种”这一个因素.

例 1.2.2 (续上例)假定小麦产量与所施化学肥料品种有很大关系. 设在六块田中施用了两种化肥. 这时观测值可表示为 y_{ijk}，它为种第 i 种小麦并施了第 j 种化肥的第 k 块田的产量，且可分解为

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk},$$
$$i = 1,2,3; j = 1,2; k = 1,2. \tag{1.2.3}$$

其中 μ 为总平均，α_i 表示第 i 种小麦的效应，β_j 表示第 j 种化肥的效应. 这个问题包含了两个因素：一是小麦品种，二是化肥品种. 因此，这是两个因素的方差分析模型. 假定我们分析的目的仍是比较小麦品种，而不是比较这两种化肥的效力. 在试验设计中，把要比较的这个因素通称为“处理”，而把为消除试验条件差异而引进的“化肥”这个因素通称为“区组”. 这个例子就是所谓的区组设计的一特殊情况. 记

$$\boldsymbol{y} = (y_{111}, y_{112}, y_{121}, y_{122}, y_{211}, y_{212}, y_{221}, y_{222}, y_{311}, y_{312}, y_{321}, y_{322})',$$
$$\boldsymbol{e} = (e_{111}, e_{112}, e_{121}, e_{122}, e_{211}, e_{212}, e_{221}, e_{222}, e_{311}, e_{312}, e_{321}, e_{322})',$$

它们分别是 12×1 的观测向量和随机误差向量. 记 $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2)'$ 它是 6×1 的未知参数向量. 设计矩阵 $\boldsymbol{X}$ 为 12×6 矩阵，具有形式

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix},$$

引进这些矩阵记号之后，我们把两因素方差分析模型也表成了(1.2.2)的形式.

从前面的讨论我们看到，应用矩阵记号，线性回归模型和方差分析模型都具有相同形式. 因此，可以对它们用统一的方法来进行统计分析. 但是它们的用处却有较大不同，线性回归模型本质上用于描述变量之间的依赖关系，方差分析模型主要用于比较效应的大小.

还存在着其它几类线性模型，如协方差分析模型、方差分量模型等. 这些模型的讨论超出了本教材的范围. 感兴趣的读者可以参阅王松桂(1987).

§1.3 应用概述

对回归模型所进行的统计分析，通常称为回归分析. 本节将介绍回归分析的

实际应用.归纳起来主要有以下四个方面.

1. 描述变量之间的关系

从§1.1我们已经看到,有了一组因变量和自变量的数据,我们通过回归分析方法,可以建立一个线性经验回归方程.通过一定的检验,我们认为这个方程描述了因变量和自变量之间的相依关系.在§1.1的讨论中,我们已经举了这方面的两个例子.这里我们要强调的是,我们所建立的刻画变量之间关系的"函数"之所以称为经验回归方程,是因为这个方程建立在现有的一组数据的基础上,它是对现有数据中所包含的变量关系信息的一种归纳.这种归纳当然与所获得的数据有关,因此它不十分准确.当我们观察的数据增加时,经验回归方程随之会有一些改变.一般说来,当观察数据愈多,经验回归方程所归纳的信息也就愈多,因而准确程度会随之提高.

另一方面,当我们通过一组数据,获得了一个经验回归方程后,我们还要考察这个经验回归方程是否真正刻画了因变量与自变量之间客观存在的依赖关系.这是因为,当我们应用线性回归模型对数据进行分析时,面临着模型选择,自变量选择,误差假设适用性等问题.诸多问题有一个处理得不当,都会导致一定的偏差.因此,在实际应用中,建立经验回归方程的过程是一个迭代过程.先选择一个初始模型,基于数据得到经验回归方程后,经过一些统计检验后结合问题专业知识的解释,若认为初始模型不够合理,则对其进行适当修正,或改变估计方法,然后再重新建立经验回归方程.重复上述过程,直到所得到的经验回归方程从诸多角度考察都比较满意为止.

2. 分析变量之间的相互关系

当我们建立了一个比较满意的经验回归方程之后,就可以利用它分析变量之间的关系.用 Y 表示因变量.$X_1,\cdots,X_p$ 表示自变量.设经验回归方程为

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p. \tag{1.3.1}$$

在适当消除了自变量 $X_1,\cdots,X_p$ 所选用计量单位的影响之后,回归系数 β_i 的估计值 $\hat{\beta}_i$ 的大小在一定程度上,反映了变量 X_i 对因变量 Y 的影响大小.当其余自变量保持不变时,X_i 每增加一个单位,因变量 Y 平均"增加" $\hat{\beta}_i$ 个单位.因此,当 $\hat{\beta}_i>0$ 时,Y 与 X_i 是正的相关关系,即 Y 随着 X_i 的增加而增加.相反,当 $\hat{\beta}_i<0$ 时,Y 与 X_i 是负的相关关系,即 Y 随着 X_i 的增加而减少.同时回归系数 $\hat{\beta}_i$ 也刻画了它们相关的程度.这样我们可以根据 $\hat{\beta}_i$ 的符号以及绝对值 $|\hat{\beta}_i|$ 的大小将自变量进行分类:正相关变量、负相关变量或重要自变量和次要自变量.

另一方面,应用一些统计分析方法,我们还可以分析自变量之间存在的相关关系.随着电子计算机的发展与普及,人们愈来愈多地处理含较多自变量的大型