

目 录

第二版前言

第一版前言

1	简单线性回归	(1)
1.1	建立简单回归模型	(3)
1.2	最小二乘估计	(7)
1.3	估计 σ^2	(12)
1.4	最小二乘估计的性质	(13)
1.5	模型的比较: 方差分析	(15)
1.6	测定系数, R^2	(19)
1.7	置信区间和检验	(20)
1.8	残差	(24)
	问题	(28)
2	多元回归	(34)
2.1	在简单回归模型上增加一个自变量	(39)
2.2	回归的矩阵表示	(43)
2.3	方差分析	(50)
2.4	附加变量图	(55)
2.5	通过原点的回归	(58)
	问题	(58)
3	下结论	(68)
3.1	解释参数估计值	(68)
3.2	抽样模型	(74)

3.3	含测量误差的自变量	(79)
	问题	(81)
4	加权最小二乘法, 对拟合失真的检验, 广义 F-检验及置信椭圆	(84)
4.1	广义加权最小二乘法	(84)
4.2	方差已知时对拟合失真的检验	(92)
4.3	方差未知时对拟合失真的检验	(94)
4.4	广义 F 检验	(100)
4.5	联合置信区域	(101)
	问题	(103)
5	诊断 I : 残差及影响	(110)
5.1	残差	(113)
5.2	异常值	(118)
5.3	案例的影响	(122)
	问题	(130)
6	诊断 II : 症状与治疗	(133)
6.1	散点图	(134)
6.2	非常数方差	(139)
6.3	非线性	(146)
6.4	变换响应变量	(153)
6.5	变换自变量	(159)
6.6	正态性假设	(164)
	问题	(168)
7	建立模型 I : 定义新的自变量	(172)
7.1	多项式回归	(172)
7.2	虚拟变量: 二分类的	(177)
7.3	虚拟变量: 多分类的	(186)

7.4	比较回归直线	(187)
7.5	变量的尺度	(194)
7.6	线性变换及主成分	(195)
	问题	(198)
8	建立模型Ⅱ：共线性与变量选择	(205)
8.1	什么是共线性	(205)
8.2	为什么共线性是一个问题	(207)
8.3	共线性的度量	(208)
8.4	变量选择	(212)
8.5	假设和记号	(219)
8.6	根据实际意义选择子集	(220)
8.7	求子集Ⅰ：逐步的方法	(221)
8.8	选择一个子集的准则	(227)
8.9	子集选择Ⅰ：所有可能回归	(230)
	问题	(233)
9	预 测	(240)
9.1	进行预测	(242)
9.2	内插法对外推法	(250)
9.3	附加评注	(252)
	问题	(254)
10	不完全数据	(258)
10.1	随机遗漏	(258)
10.2	通过填入和删除来处理不完全数据	(261)
10.3	正态性假设下的极大似然估计	(264)
10.4	遗漏观测值相关	(264)
10.5	一般推荐	(266)
11	非最小二乘估计	(267)

11.1	稳健回归	(268)
11.2	有偏回归	(270)
12	线性回归的推广	(277)
12.1	非线性回归	(277)
12.2	逻辑斯谛回归	(283)
12.3	广义线性模型	(288)
	问题	(291)
附 录	(293)
1A.1	简单回归模型的形式上的展开	(293)
1A.2	随机变量的均值和方差	(294)
1A.3	最小二乘	(295)
1A.4	最小二乘估计的均值和方差	(296)
1A.5	舍入, 舍入误差及回归计算的精确性	(297)
2A.1	对矩阵和向量的简要介绍	(299)
2A.2	随机向量	(305)
2A.3	最小二乘	(307)
5A.1	相联回归方程	(308)
8A.1	C_p 的由来	(308)
表	(310)
参考文献	(326)
标题索引	(340)

1

简单线性回归

回归被用于研究可以测量的变量之间的关系。线性回归则被用于研究一类特殊的关系，即可用直线或多维时的直线的推广描述的关系。这一技术被用于几乎是所有的研究领域，包括社会科学、物理、生物、商业、科技和人文科学。正如本书例子所示，用线性回归模型拟合的原因因应用而异，而最通常的原因是描述关系并对未来值进行预测。

一般地，回归分析由许多步骤组成。为研究一组变量之间的关系，要收集这些变量在一组单元或案例中的每一个数据。这里研究的回归模型，一个变量起着响应的作用，称为响应变量，而所有其它变量看成是响应的预报因子，称为自变量。我们可以方便地，而且也常是准确地认为，自变量有数据收集者所得的数据值，而把响应变量看作这些自变量的一个函数。除了若干未知参数，对于给定值的自变量，假设模型详细说明了响应变量的行为。模型通常还会指出，由于假设误差项而不能给出准确拟合的某些特征。然后，数据被用于得到未知参数的估计值。尽管存在着多种估计方法，本书中大部分研究的是最小二乘法。这种分析方法称为综合分析，因为其主要目的是将数据聚集在一起，并综合出数据的一个拟合模型。接着，同样重要的回归分析的下一个阶段

是案例分析。这里数据被用于检验拟合模型对被研究的关系是否合适、有用。其结果有可能导致对原先指定的拟合模型的修改。对数据或假设修改以后，回复至综合分析。

本章讨论简单回归，它具有一个自变量和一个响应变量。重点是对一个合适模型的描述，假设的讨论，最小二乘估计，置信区间及检验等过程。

例 1.1 Forbes 数据

在十九世纪四、五十年代，苏格兰物理学家 James D. Forbes，试图通过水的沸点来估计海拔高度。他知道通过气压计测得的大气压可用于得到海拔高度，高度越高，气压越低。在这里讨论的实验中，他研究了气压和沸点之间的关系。由于在 40 年代运输精密的气压计相当困难，这引起了他的研究此问题的兴趣。测量沸点将给旅行者提供一个快速估计高度的方法。

Forbes 在阿尔卑斯山及苏格兰收集数据。选定地点后，他装起仪器，测量气压及沸点。气压单位采用水银柱高度，并根据测量时周围气温与标准气温之间的差异校准气压。沸点用华氏温度表示。我们从他 1857 年的论文中选取了 $n=17$ 个地方的数据，见表 1.1 (Forbes, 1857)。在研究这些数据时，有若干可能引起兴趣的问题，气压及沸点是如何联系的？这种关系是强是弱？我们能否根据温度预测气压？如果能，有效性如何？

Forbes 的理论认为，在观测值范围内，沸点和气压值的对数成一直线。由此，我们取 10 作为对数的底数。事实上统计分析和对数的底数是没有关系的。由于气压的对数值变化不大，最小的为 1.318，而最大的为 1.478，我们将所有气压的对数值乘以 100，如表 1.1 中第 5 列所示。这将在不改变分析的主要性质的同时，避免研究非常小的数字。

着手进行回归分析的一个有效途径是，画一个变量对另一个变量的图。这图称为散点图，它既能用于提示某种关系，也能用于说明这种关系可能是不适当的。散点图可手工在一般作图纸上绘制。X 轴即水平轴，通常留作用于自变量。在 Forbes 的数据中为沸点。Y 轴即垂直轴，通常被用于表示响应变量。在本例中，Y 轴的值为 $100 \times \log(\text{气压})$ 。对 n 对 (x, y) 数据中的每一对，在图上作一个点。大多数回归分析的计算机程序可以作这个图。

Forbes 数据的散点图的总的印象是，这些点基本上，但并不精确地，落在一条直线上。图 1.1 所画的直线将在后面讨论。它指出两个变量之间的关系至少可以初步近似地用一条直线的方程来描述。

在我们学习这一章的过程中，学到的方法都将被用来分析这批数据。

表 1.1 在阿尔卑斯山及苏格兰的 17 个地方沸点°F)
及大气压 (英寸汞柱) 的 Forbes 数据

案例号	沸 点 (°F)	气 压 (英寸汞柱)	log (气压)	100×log (气压)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

1.1 建立简单回归模型

在简单回归中, 两个量, 如 X 和 Y 之间的关系将被研究。首先, 我们希望这一关系可以用一条直线来描述。为使之合理, 我们可能需要变换 X 和 (或者) Y 的尺度, 如我们在 Forbes 数据中所做的那样, 将气压变换成 \log (气压)。在本章中, X 和 Y 的观测值将用带下标的小写字母 (x_i, y_i) 表示, 指 X 和 Y 在研究中

的第 i 个案例。这里给出的是简单回归模型的主要特征。更正式的讨论参见附录 1A. 1。

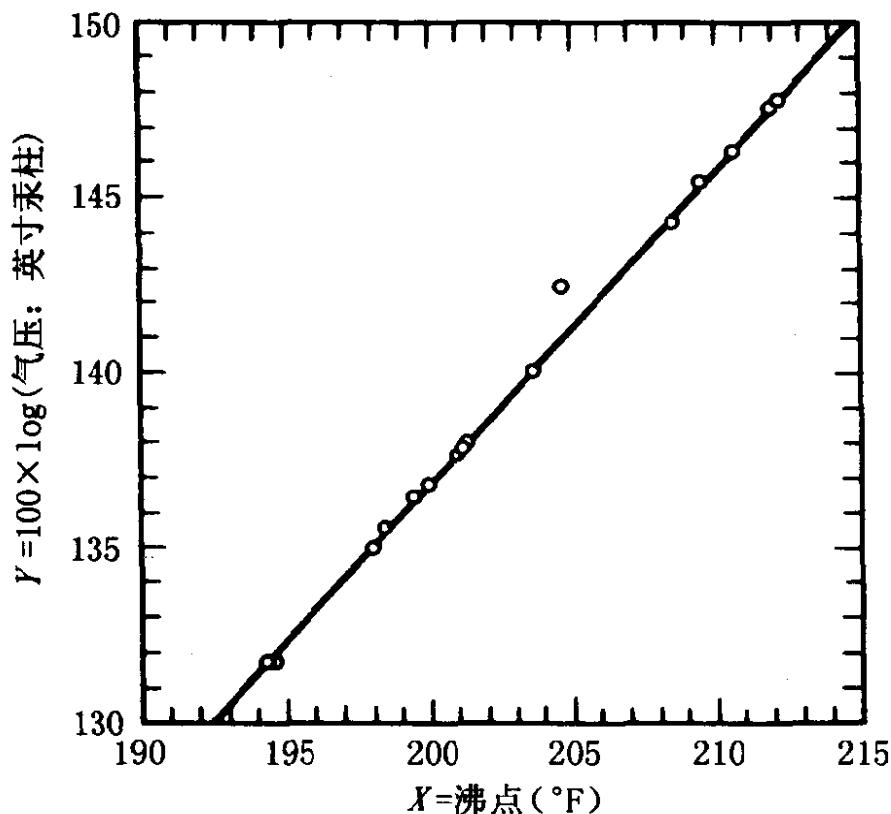


图 1.1 Forbes 数据的散点图

直线方程 关于两个量 X 和 Y 的直线可用方程

$$Y = \beta_0 + \beta_1 X \quad (1.1)$$

表示。在方程 (1.1) 中, β_0 是截距, 它是 X 取 0 时 Y 的值。斜率 β_1 为 X 变化一个单位时 Y 的变化率, 见图 1.2。 β_0 和 β_1 称为参数, 并且因为它们取遍所有可能的值, 它们给出所有可能的直线。在大多数统计模型应用中, 参数是未知的, 并且要通过数据进行估计。

误差 实际数据几乎是从来不会准确地落在一条直线上的。测得的响应变量的值与模型给出的值的差 (对简单回归, 即 Y 的观测值减去 $(\beta_0 + \beta_1 \cdot X)$ 的差) 称为统计误差。这一术语不能同通常所用的近义词“错误”相混淆。模型不能给出一个精确的拟

合，是由统计误差引起的，这些误差可同时含有固定和随机成分。如果给出的模型，如这里的直线，不完全正确，则会引起统计误差中的固定部分。例如， Y 和 X 之间的真实关系如图 1.3 的三次曲线所示，而假设我们不正确地提出一条直线，如虚线所示，来表达这种关系。由于用一条直线，而不是合适的曲线来建立模型，固定误差，有时称为拟合失真误差，是直线与正确曲线之间的垂直距离。在本章标准线性回归理论中，我们假设误差中拟合失真

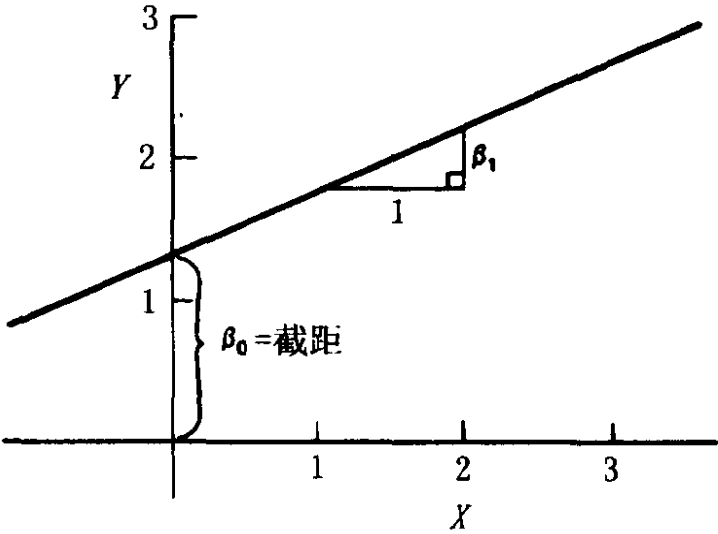


图 1.2 一条直线

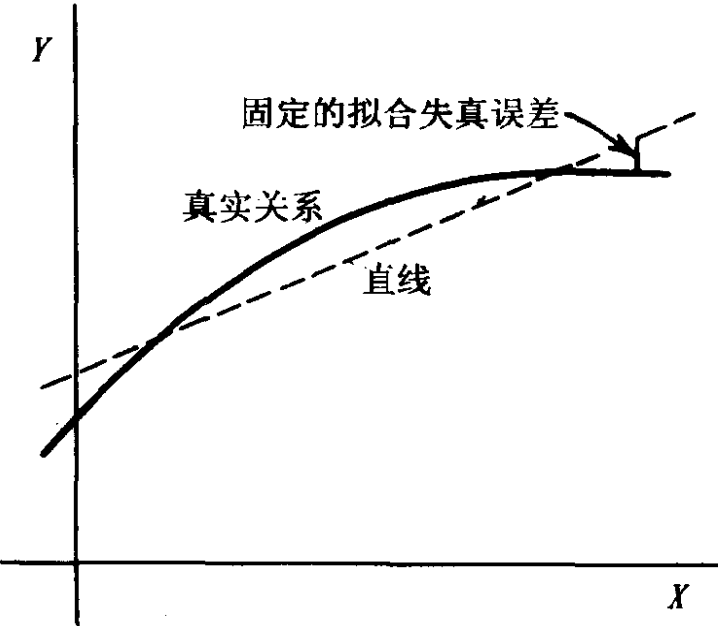


图 1.3 用直线的近似曲线

的成份是可以忽略的。但在实际中，这一假设需要验证。通常使用变换使拟合失真误差变小。

误差中的随机部分可由多种因素引起。测量误差，这里不是 X 而只是 Y 的测量误差，几乎总是存在的，因为几乎没有变量值可被完全精确地测量出来。在模型中没有考虑到的某个变量的作用也能引起误差。例如在 Forbes 实验中，风速可能对大气压产生微小的影响，引起观测值的变化。另外，自然环境的变化也能引起随机误差。

令 e_i 表示第 i 个案例的统计误差， $i=1, 2, \dots, n$ 。假设误差的固定成分可以被忽略。 e_i 的平均值为 0， $E(e_i)=0, i=1, 2, \dots, n$ 。（若不熟悉符号 $E(\quad)$ ， $\text{var}(\quad)$ 及 $\text{cov}(\quad)$ ，可参见附录 1A.2）。另一个方便的假设是，误差互不相关（用协方差符号表示，即 $\text{cov}(e_i, e_j)=0$ ，对所有的 $i \neq j$ 都成立），并且具有共同的、通常是未知的方差 σ^2 ， $\text{var}(e_i)=\sigma^2, i=1, 2, \dots, n$ 。顾名思义，不相关意味着一个误差的值，不依赖于或不决定其它任一误差的值。若用“相互独立”代替“不相关”，将略失一般性。在某些应用中，需要有误差的分布假设。通常的假设是正态分布。它极自然地导致最小二乘估计。在本书中，正态的假设主要用于获得检验和置信陈述。如果认为误差服从其它分布，例如泊松（Poisson）或伽玛（Gamma）分布，最小二乘法以外的其它方法可能会更合适。参见第 12 章，那里有更详细的讨论。

作正态假设，以及刚才关于平均值、方差和协方差的假设，我们可以写成 $e_i \sim NID(0, \sigma^2), i=1, 2, \dots, n$ ，读作 e_i 服从期望值为 0，公共方差为 σ^2 的独立正态分布。在任何实际应用中，本节所作假设必须被检验。这一点将在以后的章节中进行讨论。

简单回归模型 我们已经分别定义了 X 和 Y 为自变量和响应变量。 (x_i, y_i) 是 X, Y 的观测值， $i=1, 2, \dots, n$ 。对于 Forbes 数据，表 1.1 的第 2 列给出了 x_i ，第 5 列给出了 y_i 。例如 $x_3=197.9$ 及 $y_3=142.44$ 。我们又定义 e_i 为第 i 个案例的统计误差， $i=1, 2,$

..., n 。简单线性回归模型叙述为:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, \dots, n) \quad (1.2)$$

其中

$$E(e_i) = 0$$

$$\text{var}(e_i) = \sigma^2$$

$$\text{cov}(e_i, e_j) = 0 \quad (i \neq j)$$

换句话说, 模型表达了, 除了加上 e_i 这个未知随机量外, y_i 可由 x_i 的值通过设定的方程决定。三个量 β_0 , β_1 和 σ^2 是未知的。 e_i 是不可观测的量值, 它被引入模型中, 作为观察值不能完全准确地落在一条直线上的原因。只有 x_i 和 y_i 是观测得到的。这些数据用于估计未知参数, 即 β_0 、 β_1 和 σ^2 的值。

1.2 最小二乘估计

可以用许多方法获得模型中参数的估计值。这里讨论的称为最小二乘法, 根据使一个称为残差平方和的量达到最小来获取参数的估计值。

符号 参数和参数的(统计)估计值的区别, 对于理解和使用统计模型是很重要的。为了以示区别, 参数用小写希腊字母表示, 通常为 α 、 β 、 γ 和 σ , 而参数的估计值通常用在相应希腊字母上面放一顶“帽子”来表示。这样 $\hat{\beta}_1$ (读作 β_1 角) 为 β_1 的估计值。类似地, $\hat{\sigma}^2$ 为 σ^2 的估计值。虽然在通常意义下, e_i 不是参数, 我们使用相同的“角”标记表示观测拟合误差或残差。第 i 个案例的残差, 记作 \hat{e}_i , 由下面的等式给出

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (i = 1, 2, \dots, n) \quad (1.3)$$

把它与关于统计误差的等式相比较

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad (i = 1, 2, \dots, n) \quad (1.4)$$

因为残差是可观测的并用于检验假设, 而统计误差是不可观测的, 所以 e_i 和 \hat{e}_i 的区别是重要的。

可以把“角”标记扩展到用来表示由被估计的回归方程决定的拟合值。这样第 i 个拟合值由下面的方程给出

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (i = 1, 2, \dots, n) \quad (1.5)$$

比较 (1.5) 和 (1.3)，我们可以看到 $\hat{e}_i = y_i - \hat{y}_i$ 。

所有简单回归的最小二乘运算可以只用一些简明的统计量完成，这些统计量由数据计算而得到，它们是样本均值、校正平方和及校正叉积和。为便于参照，所有这些量的定义在表 1.2 中给出。他们被定义为在平方或取叉积之前，从每个变量中减去样本平均值。用于从未校正的平方和及叉积和计算校正平方和及校正叉积和的相应的几个公式也列于表中。用计算器计算未校正的平方和很方便，因为很多计算器有通过数据直接累加运算 $\sum x_i$ 和 $\sum x_i^2$ 。公式 $SXX = \sum x_i^2 - (\sum x_i)^2/n$ 可用于从未校正平方和获得校正平方和。不过，如果使用计算机进行运算，使用未校正平方和可能导致严重的舍入误差，参见附录 1A.5。

表 1.2 还列出常用一元和二元统计量，即样本均值 (\bar{x} , \bar{y})，样本方差 (SD_x^2 , SD_y^2) 及协方差和相关系数的估计 (S_{XY} , r_{XY})。前面所述“角”标记规则指出，这些量应该用不同的符号表示。如 $\hat{\rho}_{XY}$ 可能更好地表示样本相关系数，这里 ρ_{XY} 是总体相关系数。不过，这个不一致性是故意的，因为在很多回归问题中，这些统计量不是总体参数的估计值。例如，在 Forbes 实验中，数据在 17 个被选地点收集。从而沸点的样本方差 $SD_x^2 = 33.17$ ，并不表示任何有意义的总体方差的估计值。类似地，如果总体值 ρ_{XY} 有意义的话， r_{XY} 除了依赖于这一总体值外，很大程度上依赖于采样方法。

然而，这些常用样本统计量常出现，被用于代替校正平方和及叉积和，故给出使用这些量的可选的计算公式。

最小二乘准则 获取估计值的准则是基于拟合误差或残差 $\hat{e}_i = y_i - \hat{y}_i$ ，这里 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 是在 $x = x_i$ 处拟合直线的值。残差给出拟合直线和准确 y 值之间的垂直距离，如图 1.4 所示。很明显，人们可以考虑除垂直误差以外的其它函数来获得选择估计值的准

表 1.2 符号定义*

量	定义及可选形式	说 明
\bar{x}	$\sum x_i/n$	x_i 的样本平均值
\bar{y}	$\sum y_i/n$	y_i 的样本平均值
SXX	$\sum (x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n$ $= \sum x_i^2 - n \cdot (\bar{x})^2$	x_i 的校正平方和
SD_x^2	$SXX/(n-1)$	x_i 的样本方差
SD_x	$\sqrt{SXX/(n-1)}$	x_i 的样本标准差
SYY	$\sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$ $= \sum y_i^2 - n \cdot (\bar{y})^2$	y_i 的校正平方和, 亦称总平方和
SD_y^2	$SYY/(n-1)$	y_i 的样本方差
SD_y	$\sqrt{SYY/(n-1)}$	y_i 的样本标准差
SXY	$\sum (x_i - \bar{x})(y_i - \bar{y})$ $= \sum x_i y_i - (\sum x_i) \cdot (\sum y_i)/n$ $= \sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}$	校正叉积和
S_{XY}	$SXY/(n-1)$	样本协方差
r_{XY}	$SXY/\sqrt{(SXX)(SYY)} = s_{XY}/s_X s_Y$	样本相关系数

* 符号 \sum 是 $\sum_{i=1}^n$ 的缩写, 意为对 i 从1到 n 的所有取值求和。

则, 但因为残差反映了在回归问题中自变量和响应变量固有的非对称性, 因而是一个较好的选择。

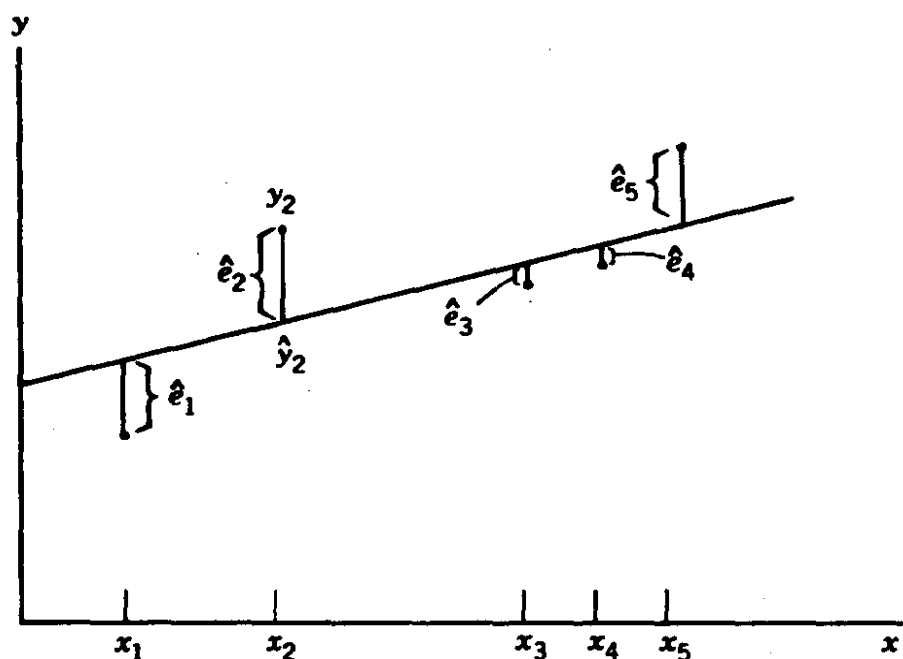


图 1.4 拟合直线的残差

最小二乘估计，即取 β_0 为 $\hat{\beta}_0$ ， β_1 为 $\hat{\beta}_1$ ，使下面的函数取最小值*

$$RSS(\beta_0, \beta_1) = \sum [y_i - (\beta_0 + \beta_1 x_i)]^2 \quad (1.6)$$

取 $\hat{\beta}_0$ ， $\hat{\beta}_1$ 时，称 $RSS(\hat{\beta}_0, \hat{\beta}_1)$ 为残差平方和，或即 RSS 。

最小二乘法是不依赖于任何关于 e_i 的假设的纯数学公式。即使在研究的数据与回归模型并不相符时，也可使用最小二乘估计计算。

最小二乘估计可用许多方法推得，其中的一种见附录 1A.3。它们可用下式表示

$$\begin{aligned} \hat{\beta}_1 &= \frac{SXY}{SXX} = r_{XY} \cdot \frac{SD_Y}{SD_X} = r_{XY} \cdot \left(\frac{SYY}{SXX} \right)^{1/2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \cdot \bar{x} \end{aligned} \quad (1.7)$$

* 在本书此处及其它地方，我们随意使用标记，把一个虽然未知但固定的量，如 β_1 ，看成它仿佛是一个自变量参数。这样，例如 $RSS(\beta_0, \beta_1)$ 是当自变量 β_0 和 β_1 变化时的一个函数。在置信区间讨论中同样随意使用标记。

β_1 的若干形式都是等价的。

有时将简单线性回归模型写成另一种形式是方便的，它略微容易使用。在等式 (1.2) 的右边加上等于零的 $\beta_1\bar{x} - \beta_1\bar{x}$ ，合并各项，我们得到

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \beta_1\bar{x}_i - \beta_1\bar{x} + e_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + e_i \end{aligned}$$

定义 $\alpha = \beta_0 + \beta_1\bar{x}$ (α 不依赖于 i)，把上式写成另一等价形式

$$y_i = \alpha + \beta_1(x_i - \bar{x}) + e_i \quad (i = 1, 2, \dots, n) \quad (1.8)$$

这称为简单回归的样本均值偏差形式。这时最小二乘估计为

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta}_1 \text{ 仍由 (1.7) 给出} \quad (1.9)$$

Forbes 的数据 四个量 SXX , SXY , \bar{x} 和 \bar{y} 是计算最小二乘估计所需要的。它们是

$$\begin{aligned} \bar{x} &= 202.95294118, \quad \bar{y} = 139.60588235 \\ SXX &= 530.78235294, \quad SXY = 475.29570589 \\ SYY &= 427.76281177 \end{aligned} \quad (1.10)$$

量 SYY ，虽然还未用到，但为保持完整性仍然给出。另外，这些计算中每个量的数字位数是过多的，因为原始数据中至多有 4 位有效数字（气压的对数值被四舍五入，如表 1.1。有兴趣的读者可重新完成前述计算）。无论如何，由于有中间运算，它们必须尽量精确，舍入只被用于最后结果。使用给出的计算公式，我们得

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -42.131$$

将 \log （气压）乘以 100 的原因，在这里是很明显的。如果不这样做， β_1 的估计值将小 100 倍，即 0.00895，而这么小的数字会经常导致出错。在简单回归模型均值偏差形式中，斜率的估计值如前面给出的， α 的估计值为 $\hat{\alpha} = \bar{y} = 139.606$ 。

估计的直线作在图 1.1 中，它由下面的任一方程给出

$$\hat{y} = -42.131 + 0.895x$$

$$= 139.606 + 0.895(x - 202.953)$$

如前所述，这一直线对数据的拟合是极好的。

1.3 估计 σ^2

我们想要 σ^2 的估计不依赖于拟合的模型的合适程度。一般，只有具备对于 x 的一组值中的每一个值都有几个 y 值的数据集，或含有 4.2 和 4.3 节所描述的先验信息的情况下，才能得到这样的估计。缺少这些特别的条件， σ^2 的估计是依赖于模型的，因为它是残差平方和 $RSS = \sum \hat{e}_i^2$ 的函数。

由于 σ^2 本质上是 e_i 的平均平方的大小，我们认为其估计值，称为 $\hat{\sigma}^2$ ，是通过将 \hat{e}_i^2 求平均获得的。在 e_i 为不相关的随机变量、期望值为 0、方差均为 σ^2 的假设下， σ^2 的一个无偏估计可以通过用其自由度 (d.f.) 去除 RSS 获得，其中

残差的 d.f. = 案例数 - 模型中参数个数

对于简单回归，残差的 d.f. = $n - 2$ ，故 σ^2 的估计为

$$\hat{\sigma}^2 = \frac{RSS}{n - 2} \quad (1.11)$$

这个量称为残差均方。一般，任何平方和除以其自由度称为均方。

为计算 $\hat{\sigma}^2$ ，读者可自己推得（练习 1.6）

$$RSS = SYY - \frac{(SXY)^2}{SXX} = SYY - \hat{\beta}_1^2 \cdot SXX \quad (1.12)$$

对于 Forbes 数据

$$\begin{aligned} RSS &= 427.76281177 - \frac{(475.29570589)^2}{530.78235294} \\ &= 2.15332 \end{aligned} \quad (1.13)$$

（或取 4 位数字，2.153）且有

$$\hat{\sigma}^2 = \frac{2.15332}{17 - 2} = 0.14355$$

（或取 3 位数字，0.144——后面将用到更精确的数字）。这个量的

平方根, $\hat{\sigma} = \sqrt{0.144} = 0.379$, 通常称为回归标准误。它和变量 Y 具有相同的单位。对于 Forbes 数据, 单位是 $100 \times \log$ (气压)。

若除了前面所作的假设外, 加上 e_i 服从正态分布的假设, 则残差均方为服从自由度为 $n-2$ 的 χ^2 分布随机变量的倍数, 可表示成

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

这一事实用于获得检验统计量的分布, 还用于作出关于 σ^2 的置信推断。其证明可进一步参见关于线性模型的书籍。特别地, 这一事实推出了

$$E(\hat{\sigma}^2) = \sigma^2$$

事实上, 关于无偏性, 正态性的假设并非必要。

1.4 最小二乘估计的性质

最小二乘估计只通过表 1.2 给出的这一组统计量而依赖于数据。这既是一个优点, 它使计算简便, 也是一个缺点, 因为任何两个这些量相同的数据集, 即使直线模型只适合于其中之一而不适合于另一个 (如第 5 章例 5.1), 也会有相同的回归拟合。另外, 估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 对 y_1, \dots, y_n 是线性的。由 (1.8), 在 $x = \bar{x}$ 处的拟合值 $\hat{y} = \bar{y} + \hat{\beta}_1 (\bar{x} - \bar{x}) = \bar{y}$, 故拟合直线必通过点 (\bar{x}, \bar{y}) 。直观上, 这个点为数据的中心。最后, 易知 $\sum \hat{e}_i = 0$, 故正、负残差相互抵消。事实上, $\sum \hat{e}_i = 0$ 是截距 β_0 的拟合的结果。在无截距项出现的模型中, 通常有 $\sum \hat{e}_i \neq 0$ 。

若 e_i 是随机变量, 则 β_0 及 β_1 的估计值亦是随机变量, 因为它们依赖于 y_i , 从而依赖于 e_i 。若所有 e_i 具有为 0 的期望值, $E(e_i) = 0, i = 1, 2, \dots, n$, 并且模型是正确的, 则如附录 1A.4 所示, 最小二乘估计是无偏的,

$$E(\hat{\beta}_0) = \beta_0$$

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

对于估计量的方差，我们现在只考虑特殊情况， $\text{var}(e_i) = \sigma^2$ ，($i = 1, \dots, n$) 且 $\text{cov}(e_i, e_j) = 0$ ， $i \neq j$ 。则由附录 1A.4，

$$\begin{aligned}\text{var}(\hat{\beta}_1) &= \sigma^2 \cdot \frac{1}{SXX} \\ \text{var}(\hat{\beta}_0) &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)\end{aligned}\quad (1.14)$$

在略微简单的样本均值偏差模型中， $\hat{\alpha}$ 的方差由下式给出

$$\text{var}(\hat{\alpha}) = \frac{\sigma^2}{n} \quad (1.15)$$

模型 (1.2) 与 (1.8) 参数化的一个重要区别在于，估计值 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 是相关的，如附录 1A.4 所给出的

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{SXX} \quad (1.16)$$

但估计值 $\hat{\beta}_1$ 和 $\hat{\alpha}$ 是无关的，这使得其它的运算，如预测值的方差相当简单。

基于 e_i 是有共同方差的不相关随机变量的假设，我们可以应用高斯—马尔科夫 (Gauss—Markov) 定理：在这些条件下，最小二乘估计 (它是 y_i 的线性函数) 具有任何线性无偏估计的最小可能方差。这意味着如果相信假设，并对利用线性无偏估计感兴趣，则最小二乘估计正是我们要使用的。

当误差服从正态分布，最小二乘估计可以用一个完全不同的方法加以讨论，因为它此时也是极大似然估计。关于极大似然估计的介绍由 Lindgren 给出 (1976)。

正态性在找出估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的分布时也是有用的。在假设 $e_i \sim NID(0, \sigma^2)$ ， $i = 1, \dots, n$ 下， $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 也是正态分布的，因为它们都是 y_i ，从而是 e_i 的线性函数，具有 (1.14) 和 (1.16) 给出的方差和相关系数。这些结论被用于获得置信区间。

估计方差 在 (1.14) 中用 $\hat{\sigma}^2$ 代替 σ^2 ，可得 $\text{var}(\hat{\beta}_0)$ 和 $\text{var}(\hat{\beta}_1)$ 的估计。我们用符号 $\hat{\text{var}}(\)$ 表示方差的估计。这样

$$\begin{aligned}\hat{\text{var}}(\hat{\beta}_0) &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) \\ \hat{\text{var}}(\hat{\beta}_1) &= \hat{\sigma}^2 \frac{1}{SXX}\end{aligned}\quad (1.17)$$

方差估计的平方根称为标准误，用记号 $\text{se}(\quad)$ 表示。它可以标记为

$$\text{se}(\hat{\beta}_1) = \sqrt{\hat{\text{var}}(\hat{\beta}_1)}$$

1.5 模型的比较：方差分析

方差分析提供了一个方便的方法，以比较对同一个数据集合，两个或多个模型的拟合。虽然所有重要的原理现在就可以阐明，但这里使用的方法在多元回归中也还是非常有用的。

简单回归模型的一个基本情况是，拟合方程为

$$y_i = \beta_0 + e_i \quad (i = 1, 2, \dots, n) \quad (1.18)$$

这一模型断言 y_i 依赖于单个参数 β_0 ，再加上随机误差，但不依赖于 x_i 。拟合这一模型等价于找出平行于 X 轴的一条最佳直线，如图 1.5 所示。最小二乘直线为 $y = \hat{\beta}_0$ ，其中 $\hat{\beta}_0$ 是 β_0 的估计值，它使 $\sum (y_i - \beta_0)^2$ 最小。易知对这一模型

$$\hat{\beta}_0 = \bar{y} \quad (1.19)$$

残差平方和是

$$\sum (y_i - \hat{\beta}_0)^2 = \sum (y_i - \bar{y})^2 = SYY \quad (1.20)$$

残差平方和具有 $n-1$ 的自由度（案例数 n 减去模型中的参数个数 1）。

其次，考虑在 (1.18) 中加上依赖于 x_i 的一项后的简单回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, \dots, n) \quad (1.21)$$

拟合这一模型等价于找出斜率不一定为零的直线，如图 1.5 所示。

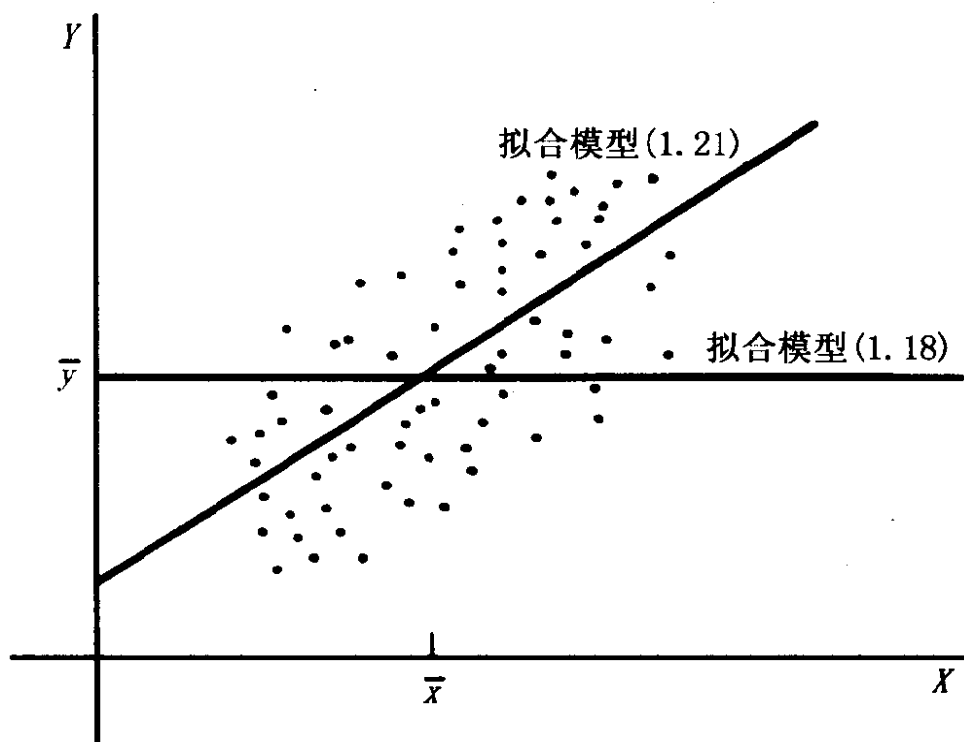


图 1.5 由方差分析比较的两个模型

这一模型的最小二乘估计由 (1.7) 给出。作为一句重要的题外话，我们可以看到在这两个模型下对 β_0 的估计是不同的，如同在两个模型中参数的含义是不同的一样。对于 (1.18)， $\hat{\beta}_0$ 是 y 的均值，而对 (1.21)， $\hat{\beta}_0$ 是当 $x_i=0$ 时的均值。

对于 (1.21)，由 (1.12) 给出的残差平方和是

$$RSS = SY - \frac{(SXY)^2}{SXX} \quad (1.22)$$

正如前面所提到的，RSS 具有 $n-2$ 的自由度。

(1.20) 与 (1.22) 的平方和的区别在于，将 (1.18) 的模型扩充至 (1.21) 的简单回归模型，残差平方和得到了减少。它们的差称为回归平方和，记为 SS_{reg} ，定义为

$$\begin{aligned} SS_{reg} &= SY - RSS \\ &= SY - \left(SY - \frac{(SXY)^2}{SXX} \right) \\ &= \frac{(SXY)^2}{SXX} \end{aligned} \quad (1.23)$$

SS_{reg} 的自由度是模型 (1.18) 的自由度 $n-1$ 与模型 (1.21) 的自由度 $n-2$ 的差, 故 SS_{reg} 的自由度为 $(n-1) - (n-2) = 1$ 。

这些结论总结于方差分析表 (简称 ANOVA) 中, 由表 1.3 给出。在方差分析表中, “来源” 这一列指对平方和来源的描述性标号; 在更复杂的表中, 可能有多个来源, 标号是不唯一的。自由度 (d. f.) 列给出与来源相对应的自由度。下一列给出对应的平方和。将平方和列中的平方和除以对应的自由度得到均方列。如同前面讨论过的, 残差行上的均方即为 $\hat{\sigma}^2$ 。

表 1.3 方差分析

来 源	自由度 (d. f.)	平方和 (SS)	均 方 (MS)	F
X 的回归	1	SS_{reg}	$SS_{\text{reg}}/1$	MS_{reg}/MSE
较大模型的残差	$n-2$	RSS	$RSS/(n-2)$	
总的校正平方和	$n-1$	$SY\bar{Y}$		

对 Forbes 数据的方差分析由表 1.4 给出。这一表中的 $SY\bar{Y}$ 在 (1.10) 中给出, RSS 在 (1.13) 中给出, SS_{reg} 由减法运算得到。

表 1.4 方差分析

来源	d. f.	SS	MS	F
回归	1	425.610	425.610	2955
残差	15	2.153	0.144	
总的	16	427.763		

ANOVA 总是根据某个较大模型计算得出的, 这里指由 (1.21) 给出的模型, 以及根据从完全模型中通过将某些参数置于 0 而得出的较小模型计算而得。较小模型是指由 (1.18) 给出的模型, 它是通过在 (1.21) 中置 $\beta_1 = 0$ 得到的。ANOVA 表中总的平方和这一行对应于具有最少参数的模型的残差平方和。在下一章

中，方差分析被应用于一系列模型，但所参照的一个固定的较大的模型保持不变。

实际应用中，ANOVA 表是通过找到 SYY 和 $SS_{\text{reg}} = SXY^2 / SXX$ ，或由表 1.2 中的一些等价公式计算得到的。 RSS 通过减法得到。

回归的 F -检验 若回归平方和 SS_{reg} 很大，则简单回归模型 $y_i = \beta_0 + \beta_1 x_i + e_i$ 将是对 (1.18) 给定的模型 $y_i = \beta_0 + e_i$ 的一个重要改进。这就是说，在简单回归模型中的附加的参数 β_1 是不等于 0 的，即 Y 事实上与 X 有关。为使这一说法更为严格，我们需要能够判断多大是“大”的。这通过比较回归均方 (SS_{reg} 除以其自由度，对简单回归自由度为 1) 和残差均方 $\hat{\sigma}^2$ 获得。我们称这一比率为 F ：

$$F = \frac{(SYY - RSS)/1}{\hat{\sigma}^2} = \frac{SS_{\text{reg}}/1}{\hat{\sigma}^2} \quad (1.24)$$

很明显， F 是 $SS_{\text{reg}} = SYY - RSS$ 的一个重要标度形式，较大的 SS_{reg} 值将导致较大的 F 值。严格地说，我们考虑检验零假设 (NH) 与备择假设 (AH)

$$\text{NH: } y_i = \beta_0 + e_i \quad (i=1, 2, \dots, n)$$

$$\text{AH: } y_i = \beta_0 + \beta_1 x_i + e_i \quad (i=1, 2, \dots, n) \quad (1.25)$$

若 e_i 为 $NID(0, \sigma^2)$ 分布，则在 NH 下，(1.24) 服从 F 分布。在简单回归中，它有分别对应于 (1.24) 中的分子和分母的自由度，1 和 $n-2$ 。记为 $F \sim F(1, n-2)$ 。 F 分布的分位点可用于给出 F -检验的显著性水平，或 p -值。

对于 Forbes 数据，我们计算得

$$F = \frac{425.610}{0.144} = 2955$$

它的自由度为 (1, 15)。从本书最后的表 B 可见， $F(1, 15)$ 的 0.01 分位点，记为 $F(0.01; 1, 15)$ ，为 8.68。故这个检验的 p

值要比 0.01 小 (得多)。从而给出充足的理由, 拒绝 NH 而接受 AH 。

p -值的解释 在合理的假设下, p -值是在 NH 为真的情况下, 获得一个待计算的统计量的值 (这里是 F 的值) 为观察值那么大或比观察值更大的条件概率。小的 p -值不利于 NH 。观测得到的 p -值依赖于样本量、抽样方案、以及正确的 AH 与 NH 相距多远。大的 β_1 (指绝对值) 将比较小的 β_1 更易导致较小的 p -值。类似地, 因为 F -检验的功效随样本量增大而增大, 故当样本量增加时, p 值一般将变小, 并且 F -检验将检测略弱的备择假设。另外, 在回归中 p 值依赖于数据 X 的样本范围; 若 X 是从较小范围中获得的, 其 p 值相对于 X 是从较大范围中获得的要来得大。

统计显著性 (观测到充分小的 p 值) 与科学显著性 (观测到足够多的有意义的量的效应) 是有重要区别的。对后者的判断通常需要除 p -值外更多的检查。

1.6 测定系数, R^2

将 (1.23) 式的两边除以 SYY , 我们得

$$\frac{SS_{\text{reg}}}{SYY} = 1 - \frac{RSS}{SYY} \quad (1.26)$$

(1.26) 的左边为 Y 的变化中, 可由 X 的回归, 亦即在模型中加入 X 后说明的变化所占的比例。右边为 1 减去剩下的不能由 X 说明的变化的部分。根据能否说明来区分总体变化是很重要的。我们给它一个专有的名字, 称为测定系数记为 R^2 , 其定义为

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = 1 - \frac{RSS}{SYY} \quad (1.27)$$

R^2 很容易根据从方差分析表中获得的量计算得到。它是一个无单位的数, 是表示数据中 x_i 与 y_i 关系强弱的一个数。因为它只依赖于平方和, 能很好地推广到多元回归, 并且易于解释, 所以它在

统计中很常用。对 Forbes 数据,

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = \frac{425.610}{427.763} = 0.995$$

这样, 观察值 $100 \times \log(\text{气压})$ 的变化中, 有 99.5% 可由沸点来说明。

与相关系数的关系 由 (1.27) 及表 1.2, 我们可以写成

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = \frac{(SXY)^2}{(SXX)(SYY)} = r_{xy}^2$$

这样, R^2 就同 X 和 Y 之间的样本相关系数的平方是一样的。

1.7 置信区间和检验

当误差为 $NID(0, \sigma^2)$, 参数估计、拟合值及预测值都将服从正态分布, 这是因为它们都是 y_i , 从而是 e_i 的线性组合。所以置信区间及检验可以以 t 分布为基础。设 $t(\alpha, d)$ 是这样一个值, 它在自由度为 d 的 t 分布的上侧尾部截下 $\alpha/2 \times 100\%$ 的部分。这些值在本书末尾表 A 中给出。

截距 截距的标准误为 $se(\hat{\beta}_0) = \hat{\sigma}(1/n + \bar{x}^2/SXX)^{1/2}$ 。从而截距的一个 $(1-\alpha) \times 100\%$ 的置信区间是下面这个区间中的点 β_0 的集合

$$\hat{\beta}_0 - t(\alpha, n-2)se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t(\alpha, n-2)se(\hat{\beta}_0) \quad (1.28)$$

对于 Forbes 数据

$$se(\hat{\beta}_0) = 0.379(1/17 + (202.95)^2/530.78)^{1/2} = 3.339$$

对于 90% 的置信区间, $t(0.10, 15) = 1.75$, 区间为

$$\begin{aligned} -42.131 - 1.75 \cdot (3.339) &\leq \beta_0 \leq -42.131 + 1.75 \cdot (3.339) \\ -47.974 &\leq \beta_0 \leq -36.288 \end{aligned} \quad (1.29)$$

在这样的区间中, 有 90% 将包含真值。

检验假设

$$\text{NH: } \beta_0 = \beta_0^*, \beta_1 \text{ 任意}$$

$$\text{AH: } \beta_0 \neq \beta_0^*, \beta_1 \text{ 任意}$$

其解可以通过计算 t 统计量的值,

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} \quad (1.30)$$

并将这个比值与自由度为 $n-2$ 的 t 分布相对照而获得。例如, 对 Forbes 数据, 考虑检验假设

$$\text{NH: } \beta_0 = -35, \beta_1 \text{ 任意}$$

$$\text{AH: } \beta_0 \neq -35, \beta_1 \text{ 任意} \quad (1.31)$$

统计量为

$$t = \frac{-42.131 - (-35)}{3.339} = 2.136 \quad (1.32)$$

由于 $t(0.05, 15) = 2.13$, 其中 p -值接近 0.05, 有理由拒绝 NH。当然, 对这些数据, 大多数研究者不会进行这一假设检验, 这里只是用作说明。

斜率 β 的标准误是 $\text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SXX} = 0.0164$ 。斜率的一个 95% 的置信区间是 β_1 的一个集合, 满足

$$0.895 - 2.13 \cdot (0.0164) \leq \beta_1 \leq 0.895 + 2.13 \cdot (0.0164) \\ 0.860 \leq \beta_1 \leq 0.930 \quad (1.33)$$

同截距一样, 因为没有明显的值作为零假设中参数的值, 本例对斜率的假设检验并不令人感兴趣。通常的检验为

$$\text{NH: } \beta_1 = 0, \beta_0 \text{ 任意}$$

$$\text{AH: } \beta_1 \neq 0, \beta_0 \text{ 任意} \quad (1.34)$$

对于给出的数据, $t = (0.895 - 0) / 0.0164 = 54.45$ 。由于 t 很大, 与自由度为 15 的 t 分布进行比较, 在这里显然是不必要的。尽管如此, 我们还是作一比较。对应的 p -值很小, 但这并不令人吃惊。如果 β_1 确实接近 0, Forbes 几乎就不会做这个试验, 即使做了, 也不会将结果发表。

比较假设(1.34)及(1.25), 两者似乎是恒等的。事实上

$$t^2 = \left(\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \right)^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2 / SXX} = \frac{\hat{\beta}_1^2 \cdot SXX}{\hat{\sigma}^2} = F$$

自由度为 d 的一个 t 统计量的平方等于自由度为 $(1, d)$ 的 F 统计量。

预测 拟合的方程常被用于对自变量的给定值获取响应变量的值。这一问题的两个重要方面是预测及拟合值的估计。由于预测更为重要得多，我们先讨论这一问题。

在预测中，我们有一个在估计参数时没有使用过的新的案例。自变量为观测值 x_* ，我们希望得到响应变量 y_* 的值，而它目前还未被测得。由于 y_* 未得到，我们将在观测值 x_* 处使用一个模型去预测它。我们假设用于估计拟合直线的数据与新的案例有关，故对它可以应用拟合模型。在 Forbes 的例子中，我们不指望在飞机或潜水艇中，由水的沸点得到高度或深度的合理的预测，因为其实验条件与 Forbes 的实验条件很可能大不相同。然而，我们可以期望模型在地面上合适的高度都能适用。有了这一附加假设， y_* 的一个预测值，记为 \tilde{y}_* ，它为

$$\tilde{y} = \hat{\beta}_0 + \hat{\beta}_1 x_* \quad (1.35)$$

严格地说， \tilde{y}_* 预测的是，目前尚未测得的 y_* 的期望值。作为结果，这个预测值的变化来自于两个方面：估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的变化，以及 y_* 并不等于其期望值所导致的变化。利用附录 1A.4，可见

$$\text{var}(\tilde{y}_* | x_*)^* = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX} \right] \quad (1.36)$$

* 取平方根，并用 $\hat{\sigma}^2$ 估计 σ^2 ，我们得到在 x_* 处预测值的标准误 (sepred)：

$$\text{sepred}(\tilde{y}_* | x_*) = \hat{\sigma} \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX} \right]^{1/2} \quad (1.37)$$

* 译者注：预测值的方差 $\text{var}(\tilde{y}_* | x_*)$ ，或 $\text{var}(\tilde{y}_*)$ 应理解为 $E(\tilde{y}_* - y_*)^2 = \text{var}(\tilde{y}_* - y_*)$ 。这里， $E(\tilde{y}_* - y_*) = 0$ 。关于预测值的标准误有类似的结论。以下各章节同。

用通常的方法,以基于 t 分布的数作为乘数,可以得到单点预测值的预测区间。在 $x_* = 200$ 处预测 $100 \times \log$ (气压), 预测值为 $\hat{y}_* = -42.13 + 0.895 \cdot (200) = 136.87$, 预测值的标准误

$$\begin{aligned} \text{sepred}(\hat{y}_* | x_* = 200) &= 0.379 \left[1 + \frac{1}{17} + \frac{(200 - 203.0)^2}{530.8} \right]^{1/2} \\ &= 0.393 \end{aligned}$$

这样, y_* 的 99% 的一个预测区间为

$$\begin{aligned} 136.87 - 2.95 \cdot (0.393) &\leq y_* \leq 136.87 + 2.95 \cdot (0.393) \\ 135.71 &\leq y_* \leq 138.03 \end{aligned}$$

图 1.6 是 x_* 在 180 到 220 的范围内, 对 Forbes 数据的最小二乘回归直线图, 以及两条曲线

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t(0.99, 15) \text{sepred}(\hat{y}_* | x_*)$$

给定 x_* , 两条曲线之间的垂直距离为给定 x_* 后, y_* 的 99% 的预测区间。因为曲线朝外突出, 故当 x_* 远离 \bar{x} , 这一区间变宽。这可能在图中不易看到, 因为两条曲线关于回归直线的变化很小。

拟合值 在一些不常见的问题中, 可能需要获得在给定 x 点的 y 的平均值的估计。这只有在除了参数的值被估计出来以外, 还知道使用的模型为正确时才有意义。在 Forbes 例子中, 这就象在一个给定沸点处, 求 $100 \times \log$ (气压) 的“真实”平均值。这个量由拟合值 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 来估计, 其标准误为

$$\text{sefit}(\hat{y} | x) = \hat{\sigma} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right]^{1/2} \quad (1.38)$$

为获取置信区间, 通常计算一个对所有可能的 x 值的联合区间。这也就是, 先计算 β_0 和 β_1 的一个联合置信域, 然后由此计算所有斜率和截距都在该联合置信域中的可能的回归直线的集合。尽管计算联合置信域的方法将在后面 4.5 节中说明, 置信集合可简单地被叙述为所有 y 的集合, 满足

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_1 x) - \text{sefit}(\hat{y} | x) [2F(\alpha; 2, n-2)]^{1/2} &\leq y \\ &\leq (\hat{\beta}_0 + \hat{\beta}_1 x) + \text{sefit}(\hat{y} | x) [2F(\alpha; 2, n-2)]^{1/2} \end{aligned}$$

(对于多元回归, 用 $p'F(a; p', n-p')$ 代替 $2F(a; 2, n-2)$, 这里 p' 是回归方程中 β 的个数)。拟合直线联合带的形状类似于图 1.6。

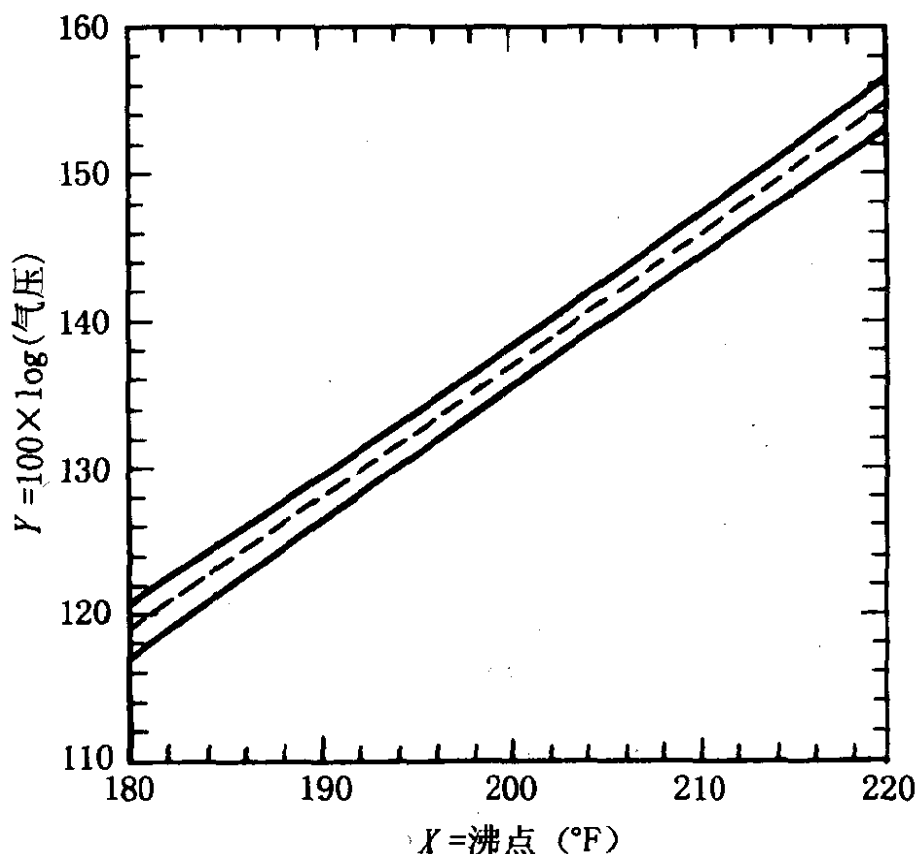


图 1.6 99%预测区间

1.8 残差

残差 $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$ 提供了关于误差项的假设及模型合适度的信息。任何完整的数据分析都要求考察残差。这里我们只给出最粗浅的残差分析的描述, 具体细节留待以后的章节讨论。

残差对其它量构成的图用于找出假设失效处。最简单的图, 尤其在简单回归中, 为 \hat{e}_i 对拟合值 \hat{y}_i 的图。这图的系统特征很有趣。弯曲可能意味着拟合模型并不合适, 建议对数据进行变换。残差

在平均数量上随 \hat{y}_i 增加或减少，可能表示有非常数的残差方差。有几个相对来说比较大的残差可能标志着异常情况，模型有些不适用的案例。另一方面，如果 \hat{e}_i 对 \hat{y}_i 的图不显示系统特征，则我们就没有什么理由怀疑拟合模型对数据并不合适。

表 1.5 Forbes 数据的拟合值及残差

案例号	x_i	y_i	\hat{y}_i	\hat{e}_i
1	194.50	131.79	132.04	-0.25
2	194.30	131.79	131.86	-0.07
3	197.90	135.02	135.08	-0.06
4	198.40	135.55	135.53	0.02
5	199.40	136.46	136.42	0.04
6	199.90	136.83	136.87	-0.04
7	200.90	137.82	137.77	0.05
8	201.10	138.00	137.95	0.05
9	201.40	138.06	138.22	-0.16
10	201.30	138.05	138.13	-0.08
11	203.60	140.04	140.19	-0.15
12	204.60	142.44	141.08	1.36
13	209.50	145.47	145.47	0.00
14	208.60	144.34	144.66	-0.32
15	210.70	146.30	146.54	-0.24
16	211.90	147.54	147.62	-0.08
17	212.20	147.80	147.89	-0.09

Forbes 数据 (结论与概略) Forbes 数据的拟合值 \hat{y}_i 及残差 \hat{e}_i 由表 1.5 给出。它们之间的图见图 1.7。注意，比起 \hat{y}_i ，残差一般较小，并且在图 1.7 中它们并没有提示明显的样式。不过一个残差(案例 12)比其它的大得多。其它残差的绝对值都小于 0.35，而案例 12 的残差约为 1.3。这可能表示关于误差的假设不正确，

或者 σ^2 可能不是常数，或者案例 12 中相应误差有一个大的固定部分。例如，对这一案例，Forbes 可能读错或抄错了他的计算结果，而使数据中的那个数并不代表真实值。Forbes 自己也注意到了这一可能性。由于有大的残差，他在论文中将这对数据标注为“一个明显错误”。

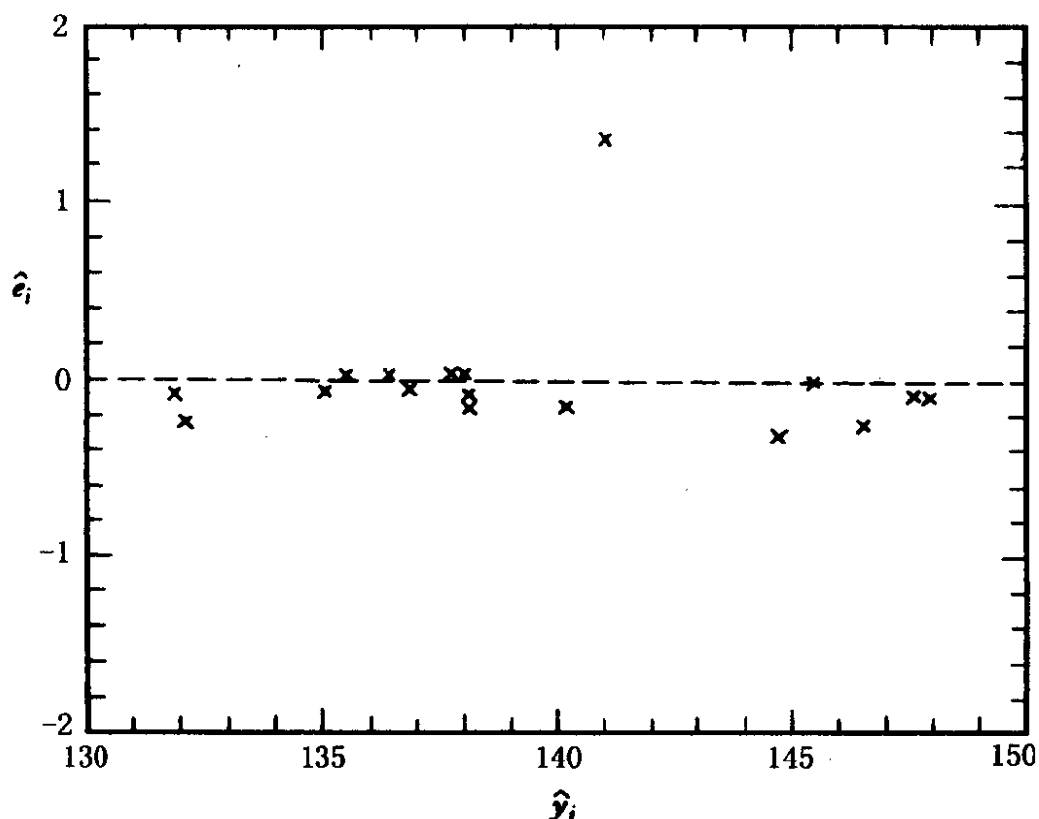


图 1.7 Forbes 的数据的残差图

我们必须面对如何处理可疑案例的问题。对于这类案例，由于缺乏一个严格的程序，在以后的章节中我们将不严格地进行讨论。由于我们关心案例 12 的效用，我们可以重新拟合数据，这次不用案例 12，然后检查参数估计、拟合值、残差方差等的变化。表 1.6 摘要给出对于每个数据集（分别为 17 和 16 个案例）的参数估计、它们的标准误 $\hat{\sigma}$ 及测定系数 R^2 。由表中可以清楚地看到，对于获得参数估计，案例 12 是无关的，因为无论是否包括这一案例，估计基本是相同的。在其它一些回归问题中，删除一个单独的案例可能改变一切。不过，案例 12 在标准误上的影响要显著些：如

果案例 12 被删除, 标准误减小约 3.1 倍, 方差减小 $3.1^2 \cong 10$ 倍。把这一案例包括在一起使用, 不如用 16 个案例可靠。在计算前删除案例 12 获得的残差图图 1.8 显示出, 对余下的 16 个案例无明显的拟合失败。

两个模型被拟合, 一个用 16 个案例, 另一个用 17 个。它们导致略为不同的结论, 尽管两种分析的大部分是吻合的。基于这些数据, 在这两个模型之间无法选择确定的一个, 并且我们也无法决定哪个数据的最小二乘分析是正确的。解决这个问题一个较好的方法是将两者 (一般地为所有的) 都视为一个似乎合理的选择。

表 1.6 Forbes 数据摘要

量	使用所有数据的值	删除案例 12 的值
β_0	-42.131	-41.302
β_1	0.895	0.891
se (β_0)	3.339	1.000
se (β_1)	0.0164	0.00493
$\hat{\sigma}$	0.379	0.113
R^2	0.995	0.999+

总之, Forbes 数据的直线模型可用于描述气压的对数值是水的沸点的一个函数。对于沸点在 180°F 到 200°F , 数据可用一条直线描述, 其方程为 $\hat{y} = -42.131 + 0.895x$, 并且至少对这一范围内的任一 x , 方程给出 $100 \times \log(\text{气压})$ 的一个较好的预测值。数据中的一个案例 (第 12 个) 的拟合程度比其余 16 个案例差得多。删去这一案例后, 估计、预测值等的标准误是包括这一案例的相应结果的 $1/3$ 。

即使直线与数据极好地拟合, 我们也不能认为, $100 \times \log(\text{气}$

压) 对沸点温度的回归是一条直线; 结论必须限制在观测值的范围内。数据并未提供在这一范围以外模型的有效性的信息。

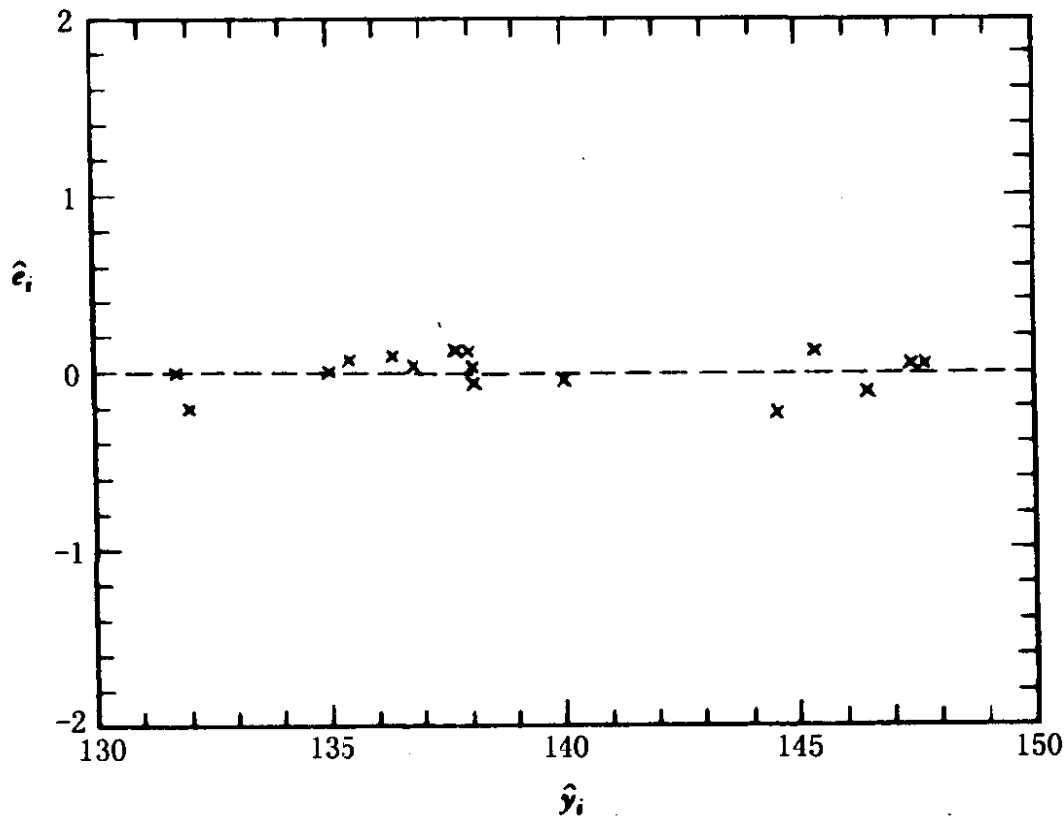


图 1.8 Forbes 数据中删除案例 12 后的残差图

问 题

- 1.1 身高和体重数据, 以下数据给出样本量 $n=10$ 的 18 岁女孩的 X =身高(厘米)和 Y =体重(公斤)。数据取自于问题 2.1 中描述的一个较大的研究。

X	169.6	166.8	157.1	181.1	158.4
Y	71.2	58.2	56.0	64.5	53.0
X	165.5	166.7	156.5	168.1	165.3
Y	52.4	56.8	49.2	55.6	77.8

不使用计算机回答以下问题:

- 1.1.1 作 Y 关于 X 的散点图。根据散点图猜测 Y 对 X 的简单线性回

归模型 $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim NID(0, \sigma^2)$, $i = 1, 2, \dots, 10$ 中的 β_0 , β_1 及 R^2 的值。

- 1.1.2 验证 $\bar{x} = 165.52$, $\bar{y} = 59.47$, $SXX = 472.076$, $SYY = 731.961$ 及 $SXY = 274.786$ 。计算 Y 对 X 回归中斜率和截距的估计值。在散点图中画出拟合直线。
- 1.1.3 估计 σ^2 , 并估计 β_0 和 β_1 的标准误以及 β_0 和 β_1 的协方差。计算假设 $\beta_0 = 0$ 和 $\beta_1 = 0$ 的 t -检验值。找出这些检验的合适的 p 值 (使用双侧检验)。
- 1.1.4 列出回归的方差分析表及 F -检验值。用数字验证: $F = t^2$, 其中 t 是在 1.1.3 中检验 $\beta_1 = 0$ 时计算得到的。
- 1.1.5 计算残差和拟合值。用数字验证: 残差之和为 0。作残差关于拟合值的图。图中是否有明显的异常? 残差有没有明显的样式?
- 1.1.6 解释参数 β_0 和 β_1 的含义。它们的单位是什么? σ 的单位是什么?
- 1.2 Hooker 数据、在 Forbes 关于沸点和气温的论文中, 他也给出了同样两个量的、由 Joseph Hooker 博士收集的数据。与 Forbes 不同, Hooker 在喜马拉雅山的一个更高的高度测得他的数据。以下摘录的 Hooker 的数据给出了 31 对测量值, $TEMP =$ 沸点 (华氏温度) 及 $PRES =$ 校正大气压 (英寸汞柱)。尽管并不必要, 鼓励使用计算机软件包解答下面问题。
 - 1.2.1 作 $PRES$ 关于 $TEMP$ 的散点图。是否有一条直线很密切地与数据匹配? (从回归 $PRES = \beta_0 + \beta_1 TEMP + e$ 所得的残差关于拟合值的图, 在这里是有用的)
 - 1.2.2 作 $100 \times \log(PRES)$ 关于 $TEMP$ 的散点图。与上一问题的散点图相比较, 这一散点图是否更易于用一条直线描述?
 - 1.2.3 对 $100 \times \log(PRES)$ 关于 $TEMP$ 拟合简单回归模型; 即拟合模型 $100 \times \log(PRES) = \beta_0 + \beta_1 TEMP + e$, 并计算相应的主要统计量 (参数估计、检验、方差分析表、 R^2)。在问题 1.2.2 的图中作出拟合直线。作残差 (关于拟合值) 图, 并与 1.2.1 的拟合作比较。
 - 1.2.4 试求 β_0 及 β_1 的 95% 置信区间。
 - 1.2.5 试分别求在 185° 及 212°C 时, $100 \times \log(PRES)$ 的 90% 预测区间。
 - 1.2.6 定量地比较本题分析结果与 Forbes 数据的分析结果。即比较拟合直线、残差估计值、预测区间等等。你得到什么结论? 在第 7 章, 我们将学习对不同组的数据比较回归的检验问题。

TEMP (°F)	PRES (英寸汞柱)	TEMP (°F)	PRES (英寸汞柱)
210.8	29.211	189.5	18.869
210.2	28.559	188.8	18.356
208.4	27.972	188.5	18.507
202.5	24.697	185.7	17.267
200.6	23.726	186.0	17.221
200.1	23.369	185.6	17.062
199.5	23.030	184.1	16.959
197.0	21.892	184.6	16.881
196.4	21.928	184.1	16.817
196.3	21.654	183.2	16.385
195.6	21.605	182.4	16.235
193.4	20.480	181.9	16.106
193.6	20.212	181.9	15.928
191.4	19.758	181.0	15.919
191.1	19.490	180.6	15.376
190.6	19.386		

1.3 奥林匹克记录。以下数据给出到1984年为止的现代奥林匹克运动会中最佳赛跑时间记录（录自1985世界年鉴）：

距离（米）	男		女	
	时间（秒）	年	时间（秒）	年
100	9.9	1968	11.0	1984
200	19.8	1984	21.8	1984
400	43.8	1968	48.8	1984
800	103.0	1984	113.5	1980
1500	212.5	1984	236.6	1980
3000			516.0	1984
5000	785.6	1984		
10000	1658.4	1972		
42195	7761.0	1984	8692.0	1984

对这一问题并不需要计算机。

- 1.3.1 用所有数据作时间关于距离的散点图, 对男子及女子的时间用不同的符号。给出这样一个时间和距离不同取值的图, 你很快会发现, 它既难画, 又难于解释。作为另一个办法, 尝试对时间和距离先作变换后作图, 例如可以是 $\log(\text{时间})$ 关于 $\log(\text{距离})$, 或速度 = 距离/时间关于距离或 $\log(\text{距离})$ 。
- 1.3.2 进行有关的运算, 使得数据中的相互关系明显化。你可能需要对男子和女子分别分析。
- 1.3.3 本问题中的 F -检验和 t -检验是否有关? 为什么? (提示: 这些数据中什么是随机的?)
- 1.4 通过原点的回归。有时候, 可以拟合于一个截距为零的模型。这一模型由下式给出

$$y_i = \beta_1 x_i + e_i \quad (i = 1, 2, \dots, n) \quad (1.39)$$

假设 e_i 相互独立, 有相同的方差 σ^2 , 则这一模型的残差平方和

$$RSS = \sum (y_i - \hat{\beta}_1 x_i)^2$$

- 1.4.1 证明 β_1 的最小二乘估计为 $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$ 。证明 $\hat{\beta}_1$ 是无偏的, 且 $\text{var}(\hat{\beta}_1) = \sigma^2 / \sum x_i^2$ 。找出 $\hat{\sigma}^2$ 的表达式, 它的自由度是多少?
- 1.4.2 导出由 (1.21) 给出的较大模型的方差分析表, 但用的是 (1.39) 的较小模型。证明: 从这一表推得的 F -检验在数值上等于 $B_0^* = 0$ 时 t -检验 (1.30) 的平方。
- 1.4.3 以下数据给出怀俄明州的蛇河河滨的 X = 在 4 月 1 日雪中含水量及 Y = 从 4 月到 7 月的水量 (英寸)。数据来自于 Wilm (1950), $n = 17$ 年 (1919 至 1935)。给出经过原点的回归直线, 并找出 $\hat{\beta}_1$ 及 $\hat{\sigma}^2$ 。求 $\hat{\beta}_1$ 的 95% 置信区间。检验截距为零的假设。

X	Y	X	Y
23.1	10.5	37.9	22.8
32.8	16.7	30.5	14.1
31.8	18.2	25.1	12.9
32.0	17.0	12.4	8.8
30.4	16.3	35.1	17.4
24.0	10.5	31.5	14.9
39.5	23.1	21.1	10.5
24.2	12.4	27.6	16.1
52.5	24.9		

1.4.4 作残差 ($\hat{e}_i = y_i - \hat{\beta}_1 x_i$) 关于拟合值 ($\hat{y}_i = \hat{\beta}_1 x_i$) 的图, 并评价模型的合适程度。在经过原点的回归中, $\sum \hat{e}_i \neq 0$ 。

1.5 尺度不变性。

1.5.1 在 (1.2) 的简单回归模型中, 假设用 cx_i 代替 x_i , 其中 c 为不等于零的常数。这对 $\hat{\beta}_0$ 、 $\hat{\beta}_1$ 、 $\hat{\sigma}^2$ 、 R^2 及 $NH: \beta_1 = 0$ 的 t -检验有何影响?

1.5.2 假设用 dy_i 代替 y_i , $d \neq 0$, 重复 1.5.1。

1.6 利用附录 1A.3, 验证方程 (1.12)。

1.7 利用附录 1A.4, 验证方程 (1.35)。

1.8 假设有模型 (1.2)。验证 \hat{e}_i 和 \hat{y}_i 的样本相关系数为零。什么是 \hat{e}_i 和 y_i 的样本相关系数? 假设在残差图中, 我们作 \hat{e}_i 对 y_i 的图, 以代替 \hat{e}_i 对 \hat{y}_i 的图。评论两者的区别。

1.9 亚马逊河水位。

亚马逊河流域是目前地球上最大的热带雨林, 但是如同其它自然资源一样, 它也由于发展而承受了沉重的压力。20 世纪 70 年代, 公路首先通到亚马逊河上游地区, 引起迅速的人口增长及大规模的森林破坏。由于降雨量和径流量都受到了影响, 这些反过来使这条河流在气候及水文方面发生变化。表 1.7 中的数据给出 1962 年至 1978 年, 在秘鲁的依昆塔的亚马逊河的高、低水位 (单位米)。1962 至 1969 年的数据可以被认为处于控制阶段, 而 1970 至 1978 年为开始发展以后的数据。要求分析这些数据, 以判断亚马逊河上游的森林破坏是否引起亚马逊河流域水平衡的变化。令人感兴趣的是, 这段时间河流的这两个特征值的变化。例如, 如果我们拟合

$$HIGH = \beta_0 + \beta_1 \times YEAR + e$$

则 (1) $\beta_1 = 0$ 指在这段时间内 $HIGH$ (高水位) 无 (线性) 变化, (2) $\beta_1 > 0$ 指 $HIGH$ (高水位) 增加, 即径流量可能增大, (3) $\beta_1 < 0$ 指较小的径流量。更多的讨论参见 Gentry 和 Lopez-Parodi (1980)。

1.9.1 作 $HIGH$ (高水位) 关于年, LOW (低水位) 关于年的散点图。

1.9.2 计算 $HIGH$ (高水位) 关于年, LOW (低水位) 关于年及 $HIGH$ (高水位) 关于 LOW (低水位) 的回归。综述你对这三个回归的结论, 并解释参数含义, 特别是根据问题的实际背景解释斜率的含义。

1.9.3 根据以上数据, 我们能否说森林破坏是引起亚马逊河水变化的原因? 还需要有什么其它信息来说明原因?

表 1.7 亚马逊河数据

年	High (米)	Low (米)	年	High (米)	Low (米)
1962	25.82	18.24	1971	27.36	21.91
1963	25.35	16.50	1972	26.65	22.51
1964	24.29	20.26	1973	27.13	18.81
1965	24.05	20.97	1974	27.49	19.42
1966	24.89	19.43	1975	27.08	19.10
1967	25.35	19.31	1976	27.51	18.80
1968	25.23	20.85	1977	27.54	18.80
1969	25.06	19.54	1978	26.21	17.57
1970	27.13	20.49			

1.10 计算机编程。

运用附录 1A.5 描述的更新的方法来编一个计算机程序，用于计算和存储主要统计量：样本量 n ，样本均值，样本校正平方和及叉积和。编一个程序，允许首先读入 n 个案例的多个变量，可能最多到 10 个变量。计算其主要统计量及其每一对的校正叉积和。我们可以方便地将校正平方和及叉积和存贮于一个二维数组 $T(I, J)$ 中。其中，例如 $T(J, J)$ 是第 j 个变量的校正平方和， $T(I, J)$ 是第 i 个和第 j 个变量的校正叉积和。故 $T(I, J) = T(J, I)$ 。有了这一程序，可以方便地计算所有的回归统计量，如本章中描述的估计和标准误。在其它章节中将有习题以扩充此程序。

2

多元回归

在多元回归里，用几个自变量来描述一个响应变量。对 n 个案例中的每一个案例，收集了响应变量和每个自变量的值。如果响应变量称为 Y ，自变量称为 X_1 、 X_2 、 \cdots 、 X_p (p 是自变量的个数)，那么数据就形成 $n \times (p+1)$ 的排列：

案例号码	数 值					
	Y	X_1	X_2	X_3	\cdots	X_p
1	y_1	x_{11}	x_{12}	x_{13}		x_{1p}
2	y_2	x_{21}	x_{22}	x_{23}		x_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots		\vdots
n	y_n	x_{n1}	x_{n2}	x_{n3}		x_{np}

对简单回归， $p=1$ 。在这个数据的表示中，值 x_{ij} 为第 i 个案例的第 j 个变量的值。一个案例的值出现在一行里，一个变量的所有的值出现在一列里。

在多元回归里，利用观察到的数据估计一个用 p 个自变量的线性函数表示响应变量的方程。模型由一个线性方程

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e \quad (2.1)$$

给出，这里，同前一章一样，诸 β_i 是未知参数， e 是统计误差， Y 是响应变量， X_1 、 X_2 、 \cdots 、 X_p 是自变量。当 $p=2$ 时，(2.1) 式

给出了三维空间 (X_1, X_2, Y) 中二维平面的方程, 如图 2.1 所示。给出了关于 X_j 收集的数据 x_{ij} 和关于 Y 收集的数据 y_i 后, 把 (2.1) 式写成

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \\ (i = 1, 2, \cdots, n) \quad (2.2)$$

在这一章, 我们将关心诸 β_i 的估计, 并对其作解释。大多数结果将通过向量和矩阵给出, 关于向量和矩阵的简要的介绍见附录 2A.1。利用这些符号, 结果显得简单和雅致。否则, 可能会被许多下标搞糊涂。

例 2.1 燃料消耗

表 2.8 的六列列出了 48 个相邻的州的下列数值:

POP = 1971 年人口, 单位: 千人。

TAX = 1972 年汽车燃料税率, 单位: 分/加仑。

$NLIC$ = 1971 年有执照的驾驶员人数, 单位: 千人。

INC = 1972 年每个成人的收入, 单位: 千美元。

$ROAD$ = 1971 年联邦政府援助的主要高速公路的千英里数。

$FUELC$ = 1972 年燃料消耗, 单位: 百万加仑。

这些数据 (除了燃料消耗) 是 Christopher Bingham 从 1974 年美国年鉴上收集的。燃料消耗是在 1974 年世界年鉴上给出的。我们将利用这些数据研究燃料消耗与其它变量的函数关系。特别感兴趣的是税率与燃料消耗之间的评价。虽然其它自变量例如消耗燃料的车辆数可能也是有关的, 但我们只限于表 2.1 的数据, 或它们经过变换后的数据。

在开始分析以前, 看看我们能否把某些变量结合起来或加以转换是有益的。自变量 $NLIC$ 和 $FUELC$ 是在整个州上度量并且随着州的大小而变化的, 同时 INC 是按每个人度量, 因此对州的大小是不敏感的。为了具有可比性, 我们把 $FUELC$ 和 $NLIC$ 除以 POP , 从而转换成每个人的比率。所得的量都在表 2.1 中给出, 它们是 *

* 本书中, X_1, X_2, \cdots, X_p 为一般的自变量名, Y 是一般的响应变量名。然而, 在一个特殊问题中, 所用的变量可以具有其它名称, 例如这个例子中所用的变量名。大多数计算机程序允许赋以 1 至 10 个字符的变量名, 本书后面有这样的例子。

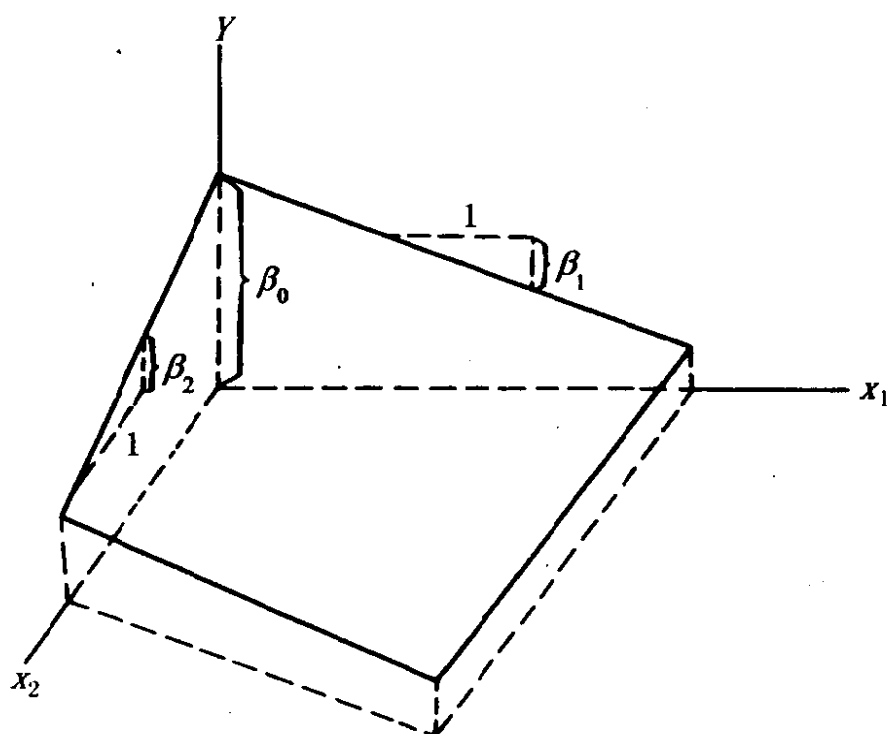


图 2.1 $p=2$ 个自变量时线性回归平面

表 2.1 燃料消耗数据

		X_1		X_3	X_4		X_2		Y
STATE	POP	TAX	NLIC	INC	ROAD	FUELC	DLIC	FUEL	
1ME	1029	9.00	540	3.571	1.976	557	52.5	541	
2NH	771	9.00	441	4.092	1.250	404	57.2	524	
3VT	462	9.00	268	3.865	1.586	259	58.0	561	
4MA	5787	7.50	3060	4.870	2.351	2396	52.9	414	
5RI	968	8.00	527	4.399	431	397	54.4	410	
6CN	3082	10.00	1760	5.342	1.333	1408	57.1	457	
7NY	18366	8.00	8278	5.319	11.868	6312	45.1	344	
8NJ	7367	8.00	4074	5.126	2.138	3439	55.3	467	
9PA	11926	8.00	6312	4.447	8.577	5528	52.9	464	

(续表)

		X_1		X_3	X_4		X_2	Y
STATE	POP	TAX	NLIC	INC	ROAD	FUELC	DLIC	FUEL
10OH	10783	7.00	5948	4.512	8.507	5375	55.2	498
11IN	5291	8.00	2804	4.391	5.939	3068	53.0	580
12IL	11251	7.50	5903	5.126	14.186	5301	52.5	471
13MI	9082	7.00	5213	4.817	6.930	4768	57.4	525
14WI	4520	7.00	2465	4.207	6.580	2294	54.5	508
15MN	3896	7.00	2368	4.332	8.159	2204	60.8	566
16IA	2883	7.00	1689	4.318	10.340	1830	58.6	635
17MO	4753	7.00	2719	4.206	8.508	2865	57.2	603
18ND	632	7.00	341	3.718	4.725	451	54.0	714
19SD	579	7.00	419	4.716	5.915	501	72.4	865
20NE	1525	8.50	1033	4.341	6.010	976	67.7	640
21KS	2258	7.00	1496	4.593	7.834	1466	66.3	649
22DE	565	8.00	340	4.983	602	305	60.2	540
23MD	4056	9.00	2073	4.897	2.449	1883	51.1	464
24VA	4764	9.00	2463	4.258	4.686	2604	51.7	547
25WV	1781	8.50	982	4.574	2.619	819	55.1	460
26NC	5214	9.00	2835	3.721	4.746	2953	54.4	566
27SC	2665	8.00	1460	3.448	5.399	1537	54.8	577
28GA	4720	7.50	2731	3.846	9.061	2979	57.9	631
29FL	7259	8.00	4084	4.188	5.975	4169	56.3	574
30KY	3299	9.00	1626	3.601	4.650	1761	49.3	534
31TN	4031	7.00	2088	3.640	6.905	2301	51.8	571
32AL	3510	7.00	1801	3.333	6.594	1946	51.3	554
33MS	2263	8.00	1309	3.063	6.524	1306	57.8	577
34AR	1978	7.50	1081	3.357	4.121	1242	54.7	628
35LA	3720	8.00	1813	3.528	3.495	1812	48.7	487
36OK	2634	6.58	1657	3.082	7.834	1695	62.9	644
37TX	11649	5.00	6595	4.045	17.782	7451	56.6	640
38MT	719	7.00	421	3.897	6.385	506	58.6	704

(续表)

		X_1		X_3	X_4		X_2	Y
STATE	POP	TAX	NLIC	INC	ROAD	FUELC	DLIC	FUEL
39ID	756	8.50	501	3.635	3.274	490	66.3	648
40WY	345	7.00	232	4.345	3.905	334	67.2	968
41CO	2357	7.00	1475	4.449	4.639	1384	62.6	587
42NM	1065	7.00	600	3.656	3.985	744	56.3	699
43AZ	1945	7.00	1173	4.300	3.635	1230	60.3	632
44UT	1126	7.00	572	3.745	2.611	666	50.8	591
45NV	527	6.00	354	5.215	2.302	412	67.2	782
46WN	3443	9.00	1966	4.476	3.942	1757	57.1	510
47OR	2182	7.00	1360	4.296	4.083	1331	62.3	610
48CA	20468	7.00	12130	5.002	9.794	10730	59.3	524

$X_1 = \text{TAX}$ (分/加仑)

$X_2 = \text{DLIC} = 100 \times \text{NLIC} / \text{POP}$

= 具有驾驶执照的人占的百分比

$X_3 = \text{INC}$, 平均收入 (千美元)

$X_4 = \text{ROAD}$ (千英里)

$Y = \text{FUEL} = 1000 \times \text{FUELC} / \text{POP}$

= 汽车燃料消耗 (每人加仑)

变量 POP 已经被排除在感兴趣的自变量的集合以外, 因为它的效果已经用于构造重新定义的以个人为基础的其它量。然而, 一个比较彻底的分析, 将包括检查 POP 的附加效果。

基本的主要统计量 (样本均值、标准差和五个感兴趣的变量之间的相关系数) 在表 2.2 中给出。为了方便, 所有的变量都被粗略地进位成相同的量级。这样的进位不影响所度量的变量之间的关系。例如, 不论 $DLIC$ 用分数还是用百分数表示, $DLIC$ 和 INC 之间的相关系数都是 0.1571。这样的进位是有用的, 因为所有被估计的系数将有差不多的量级, 并且可以避免利用很小和很大的数。

我们将多次回到这个例子。

表 2.2 基本的主要统计量

变量	<i>n</i>	均值	方差	标准差	最小值	最大值
<i>TAX</i>	48	7.6683	.90396	.95077	5.0000	10.000
<i>DLIC</i>	48	57.033	30.770	5.5470	45.100	72.400
<i>INC</i>	48	4.2418	.32904	.57362	3.0630	5.3420
<i>ROAD</i>	48	5.5654	12.191	3.4915	.43100	17.782
<i>FUEL</i>	48	576.77	12518.	111.89	344.00	968.00

样本相关矩阵					
	<i>TAX</i>	<i>DLIC</i>	<i>INC</i>	<i>ROAD</i>	<i>FUEL</i>
<i>TAX</i>	1.000				
<i>DLIC</i>	-.2880	1.000			
<i>INC</i>	.0127	.1571	1.000		
<i>ROAD</i>	-.5221	-.0641	.0502	1.000	
<i>FUEL</i>	-.4513	.6990	-.2449	.0190	1.000

2.1 在简单回归模型上增加一个自变量

在转到一般多元回归模型前，我们研究在一个简单回归模型上增加一个自变量的问题。特殊地，在燃料消耗数据中，我们将考虑把 *TAX* 增加到 *FUEL* 关于 *DLIC* 的模型中去。最后，我们将得到一个形式为

$$\widehat{FUEL} = \hat{\beta}_0 + \hat{\beta}_1 TAX + \hat{\beta}_2 DLIC \quad (2.3)$$

的方程，这里，例如 $\hat{\beta}_2$ ，象前面一样是参数 β_2 的估计。增加 *TAX* 的主要想法是为了解释未被 *DLIC* 解释的 *FUEL* 的部分。对于多元回归，一个一个地拟合自变量是很重要的，值得仔细研究。

图 2.2 (a) 和 2.2 (b) 给出了 *FUEL* 对 *DLIC* 及 *FUEL* 对 *TAX* 的散点图。这些图显示的不考虑其它自变量的影响时，*FUEL* 与这两个自变量中每一个的关系。如果我们拟合一个 *FUEL*

关于 $DLIC$ 的简单回归模型，得到：

$$\widehat{FUEL} = -227.21 + 14.10DLIC \quad (2.4)$$

$R^2 = (0.6990)^2 = 0.4886$ ，表示 $FUEL$ 的变化的 48.9% 由 $DLIC$ 解释， $DLIC$ 越高， $FUEL$ 消耗就越大。类似地，由 $FUEL$ 关于 TAX 的回归得到拟合模型

$$\widehat{FUEL} = -984.01 - 53.11TAX \quad (2.5)$$

$R^2 = (-0.4513)^2 = 0.2037$ 。所以，如果忽略 $DLIC$ ， TAX 解释了 $FUEL$ 的变化的 20.4%，并且，税率增加 1 便士，相应地每个成人在 $FUEL$ 消耗方面估计减少 53.11 加仑。

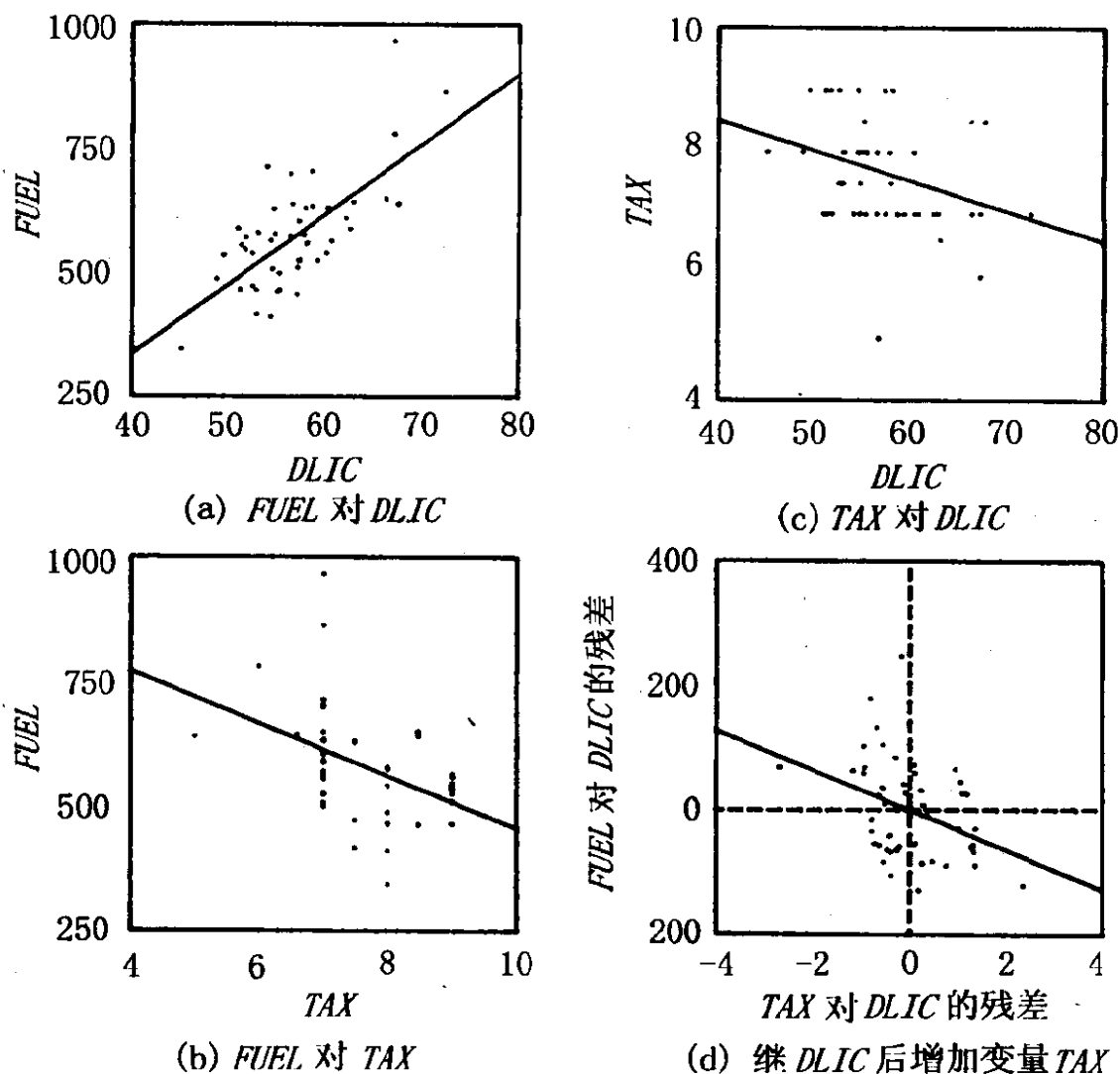


图 2.2

我们的目的是拟合一个模型，它同时用两个自变量来解释 *FUEL*。关于在一个方程中利用 *DLIC* 和 *TAX* 所能解释的 *FUEL* 的变化的比例，我们有什么结论呢？回答是非常少。我们可以说，总共被解释的变化一定超过 48.9%，即分别由每个变量所解释的变化的两个值中大的一個，因为同时知道 *DLIC* 和 *TAX* 显然至少与只知道两者中的一个一样好。只有在两个变量 *DLIC* 和 *TAX* 完全不相关并且测量完全不同的东西时，总的变化才可能相加， $48.9\% + 20.4\% = 69.3\%$ 。如果 *DLIC* 和 *TAX* 是互相相关的，都解释同一个变化，那么总的变化一定小于 69.3%。最后，如果两个变量相互作用，知道两者比只知道其中之一要有多得多的信息量，则总变化也可能超过 69.3%。例如，只根据长度或宽度是难以确定矩形的面积的，但如果在一个模型中同时考虑长度和宽度，则能够准确地确定面积。很明显，正是由于无法根据响应变量分别和每个自变量的关系来预测响应变量和两个自变量的关系，才使得多元回归丰富而复杂。

理解这个问题的关键的一步是作两个自变量 *TAX* 和 *DLIC* 的散点图。这个图由图 2.2 (c) 给出，可以看出两个自变量是相关的，说明具有驾驶员执照的人占的比例高了，税率就会降低。拟合的回归方程为

$$\widehat{TAX} = 10.48 - 0.0494DLIC \quad (2.6)$$

一个有趣的问题是找出在一个已包含 *DLIC* 的模型中增加 *TAX* 后的唯一的影响。因此，我们关心的是 *FUEL* 的未由 *DLIC* 解释的部分与 *TAX* 的未由 *DLIC* 解释的部分的建模问题。在图形上，这要求检查 *FUEL* 关于 *DLIC* 的回归的残差对 *TAX* 关于 *DLIC* 的回归的残差的散点图，或者说 *FUEL* 的未被解释的部分对 *TAX* 的未被解释的部分的散点图。这些从 (2.4) 和 (2.6) 式得到的残差，作在图 2.2 (d) 中。图 2.2 (b) 给出了忽略 *DLIC* 时，*FUEL* 和 *TAX* 的关系，而图 2.2 (d) 考虑了 *DLIC* 的校正作用。如果图 2.2 (d) 给出比图 2.2 (b) 更强的关系，则 *TAX* 和 *DLIC*

相互作用，解释大于 69.3% 的变化，而如果关系减弱，则总共被解释的变化小于 69.3%。这里看到的是后面一种情况。

如果我们对图 2.3(d) 拟合简单回归直线，拟合的直线具有零截距，因为作图用的两个变量的均值均为零，估计的斜率 $\hat{\beta}_1 = -32.08$ 。这恰好是多元回归模型(2.3)中 β_1 的估计。图 2.2(d) 为一个附加变量图的例子，我们将在 2.4 节给出附加变量图的更一般的讨论。

于是，现在有从两个模型得到的 β_1 的估计：

$$\begin{aligned}\hat{\beta}_1 &= -53.11, \text{ 忽略 } DLIC \\ \hat{\beta}_1 &= -32.08, \text{ 计入 } DLIC \text{ 校正}\end{aligned}\quad (2.7)$$

尽管两者都表示每人 *FUEL* 消耗越少，*TAX* 率越高，后一计入 *DLIC* 校正的模型指出这一影响只是如果忽略 *DLIC* 的影响的 60%。在其它回归问题中，同一变量的斜率估计，在不同的模型中，可能会大大不同，符号、数量及显著性都会改变。这自然使拟合的模型的解释复杂化。

就象(2.3)中的 $\hat{\beta}_1$ 为计入 *DLIC* 校正的 *TAX* 的影响的估计， $\hat{\beta}_2$ 为计入 *TAX* 校正的 *DLIC* 的影响的估计。这样，由 *FUEL* 关于 *TAX* 的回归的残差对 *DLIC* 关于 *TAX* 的回归的残差的回归，计算得 $\hat{\beta}_2 = 12.51$ 。在所有多元回归方程中，诸 $\hat{\beta}_j$ 被模型中所有其它变量校正。

为完成(2.3)的估计，我们可以从公式

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^p \hat{\beta}_j \bar{x}_j = 108.97 \quad (2.8)$$

计算 $\hat{\beta}_0$ ，其中样本均值在表 2.2 给出，诸 $\hat{\beta}_j$ 已在上面给出。对两个自变量的拟合方程为：

$$\widehat{FUEL} = 108.97 - 32.08TAX + 12.51DLIC$$

偏相关 在简单回归中，*Y* 与 *X*₁ 的关系可由样本相关系数 r_{YX_1} 度量。类似地，由 *X*₂ 校正的 *X*₁ 与 *Y* 的关系的强弱可由 *Y* 对 *X*₂ 的残差与 *X*₁ 对 *X*₂ 的残差的相关系数概括，如图 2.2d 所示。这

一相关称为偏相关, 记为 $r_{YX_1|X_2}$, 读作由 X_2 校正的 Y 与 X_1 的偏相关。

正交性 如果由 X_2 校正的 Y 关于 X_1 的回归等同于忽略 X_2 时 Y 关于 X_1 的回归, 则两个变量 X_1 与 X_2 是正交的。如果 X_1 与 X_2 的样本相关系数恰好为零, 则出现这个令人愉快的情况。如果 X_1 与 X_2 是正交的, 则每一个变量的影响是明确的, 因此, 通常设计试验, 使它具有正交变量。

2.2 回归的矩阵表示

矩阵符号将使多元回归中大部分结论简单化, 一般, 用黑体字如 X , e 和 β 表示一个向量或矩阵, 向量或矩阵的元素形如 x_{ij} , e_i 及 β_j 。

设 Y 和 e 为 $n \times 1$ 向量, 其元素为 (2.2) 中的 y_i 和 e_i , 例如

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad (2.9)$$

另外, 定义 β 为长度为 $(p+1) \times 1$ 的向量参数, 它包含截距 β_0 ,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad (2.10)$$

再定义 X 为一个 $n \times (p+1)$ 矩阵

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (2.11)$$

为了符号上的方便,我们定义 p' 为 X 的列数。对于有截距项的问题, $p' = p + 1$, 后面我们将看到在无截距的问题中, $p' = p$ 。

矩阵 X 给出自变量的所有的观测值,再在最左边添加一列 1。 X 的第 i 行(及 e 和 Y 的第 i 行)与数据中第 i 个案例的值对应; X 的列对应于不同的自变量。我们称 X 的最左边的一列,即一列 1 为第 0 列,因为这是对应于 β_0 的列。相邻的一列,对应于第一个自变量 X_1 和参数 β_1 ,称为 X 的第一列,依此类推。

利用这些量,多元回归方程 (2.2) 可以写成矩阵的形式

$$Y = X\beta + e \quad (2.12)$$

不熟悉矩阵的读者应使用定义 (2.9) 至 (2.11) 来进行 (2.12) 中指出的乘法和加法运算,并且证明,结果中的第 i 行,与等式 (2.2) 完全相同。

对于燃料消耗数据,矩阵 X 和 Y 的最初和最后的几行为

$$X = \begin{bmatrix} 1 & 9.0 & 52.5 & 3.571 & 1.976 \\ 1 & 9.0 & 57.2 & 4.092 & 1.250 \\ 1 & 9.0 & 58.0 & 3.865 & 1.586 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 9.0 & 57.1 & 4.476 & 3.942 \\ 1 & 7.0 & 62.3 & 4.296 & 4.083 \\ 1 & 7.0 & 59.3 & 5.002 & 9.794 \end{bmatrix}, \quad Y = \begin{bmatrix} 541 \\ 524 \\ 561 \\ \vdots \\ 510 \\ 610 \\ 524 \end{bmatrix}$$

β 是一个参数向量(长度 $p' = 4 + 1 = 5$),当然 β 的值是未知的。误差向量 e 是不可观测的。

e 的方差-协方差矩阵 误差项是由随机变量组成的向量,称为随机向量(如附录 2A.2 一样)。在第 1 章中给出的关于 e_i 的假设用矩阵形式表示为 $E(e) = \mathbf{0}$, $\text{var}(e) = \sigma^2 I_n$, 其中 $\text{var}(e)$ 表示 e 的方差-协方差矩阵, I_n 为 $n \times n$ 单位阵, $\mathbf{0}$ 为 $n \times 1$ 的零向量。如果我们加上每个 e_i 为正态分布的假设,写成

$$e \sim N(\mathbf{0}, \sigma^2 I_n) \quad (2.13)$$

最小二乘估计量 β 的最小二乘估计 $\hat{\beta}$ 是使残差平方和最小

的量。假设令 \mathbf{x}_i^T 为 \mathbf{X} 的第 i 行, $i=1, 2, \dots, n$; 转置是必要的, 因为根据习惯, 所有的向量为列向量。我们选取 $\hat{\beta}$ 使函数

$$RSS(\beta) = \sum (y_i - \mathbf{x}_i^T \beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \quad (2.14)$$

取最小值。

类似于附录 1A.2 中所述的那样, 对矩阵微分, 可以由 (2.14) 得到最小二乘估计量。它们也可以通过利用附录 2A.3 中所述的导致数值稳定的计算方法的论证得到。基本思想是将原来的问题变换为易于解决的一个等价问题。只要 $(\mathbf{X}^T \mathbf{X})^{-1}$ 存在, β 的最小二乘估计 $\hat{\beta}$ 由下式给出:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (2.15)$$

估计量 $\hat{\beta}$ 只依赖于充分统计量 $(\mathbf{X}^T \mathbf{X})$ 和 $(\mathbf{X}^T \mathbf{Y})$, 它们是未校正的平方和矩阵及叉积矩阵。

如同在简单回归中一样, 我们可以导出基于校正平方和及叉积的在数值计算上更优的等价公式。令 $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T$ 为 \mathbf{X} 的样本均值的 $p \times 1$ 向量, 定义 \mathcal{X} 为原始数据减去均值的 $n \times p$ 矩阵; \mathcal{X} 的第 (i, j) 个元素为 $x_{ij} - \bar{x}_j$ 。类似地, 定义 \mathcal{Y} 为一个 $n \times 1$ 向量, 其第 i 个元素为 $y_i - \bar{y}$ 。则校正叉积矩阵 $\mathcal{X}^T \mathcal{X}$, $\mathcal{Y}^T \mathcal{Y}$ 和 $\mathcal{X}^T \mathcal{Y}$ 为

$$\mathcal{X}^T \mathcal{X} = \begin{bmatrix} \sum (x_{i1} - \bar{x}_1)^2 & \cdots & \sum (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) \\ \vdots & \vdots & \vdots \\ \sum (x_{in} - \bar{x}_1)(x_{ip} - \bar{x}_p) & \cdots & \sum (x_{ip} - \bar{x}_p)^2 \end{bmatrix} \quad (2.16)$$

$$\mathcal{X}^T \mathcal{Y} = \begin{bmatrix} \sum (x_{i1} - \bar{x}_1)(y_i - \bar{y}) \\ \vdots \\ \sum (x_{ip} - \bar{x}_p)(y_i - \bar{y}) \end{bmatrix}, \quad \mathcal{Y}^T \mathcal{Y} = [\sum (y_i - \bar{y})^2]$$

一般, 这些矩阵可以用一个 $(p+1) \times (p+1)$ 矩阵 \mathbf{T} 表示, 其中

$$\mathbf{T} = \begin{matrix} & \begin{matrix} p \text{ 列} & 1 \text{ 列} \end{matrix} \\ \begin{pmatrix} \mathcal{X}^T \mathcal{X} & \mathcal{X}^T \mathcal{Y} \\ \mathcal{Y}^T \mathcal{X} & \mathcal{Y}^T \mathcal{Y} \end{pmatrix} & \begin{matrix} p \text{ 行} \\ 1 \text{ 行} \end{matrix} \end{matrix} \quad (2.17)$$

大部分计算机程序输出样本协方差和方差矩阵 S ，它是由对 T 中每个元素除以 $(n-1)$ 得到的， $S = (n-1)^{-1}T$ 。表 2.2 中给出的样本相关矩阵是根据公式 $r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$ 由 S 得到的。由于 S 是对称的，通常只输出下三角部分。

假定我们令

$$\beta^* = (X^T X)^{-1} X^T y$$

则

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \beta^{*T} \bar{X} \\ \hat{\beta} &= \begin{pmatrix} \hat{\beta}_0 \\ \beta^* \end{pmatrix} \end{aligned} \quad (2.18)$$

在本书中，我们按通常习惯，以数值上不稳定的未校正形式来表达公式，因为这种形式求导通常比较容易。不过，校正形式往往更好些，因为它可能可以利用附录 1A.5 所描述的方法直接进行计算。

导出量 一旦计算了 $\hat{\beta}$ ，我们可以定义若干相关的量。拟合值向量 $\hat{Y} = X\hat{\beta}$ ，其第 i 个元素等于 $\hat{y}_i = x_i^T \hat{\beta}$ 。残差问题为 $\hat{e} = Y - \hat{Y}$ ，其第 i 个元素 $\hat{e}_i = y_i - \hat{y}_i = y_i - x_i^T \hat{\beta}$ 。函数 (2.14) 在 $\hat{\beta}$ 处的值称为残差平方和，简记为 RSS ，有

$$RSS = \hat{e}^T \hat{e} = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \quad (2.19)$$

估计量的性质 最小二乘估计的其它性质在附录 2A.3 中导出，这里只作概述。假设 $E(e) = 0$ ， $\text{var}(e) = \sigma^2 I_n$ ，则 $\hat{\beta}$ 是无偏的， $E(\hat{\beta}) = \beta$ ，且

$$\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (2.20)$$

按照 1.3 中的规则，可以得到 σ^2 的一个估计：

$$\hat{\sigma}^2 = \frac{RSS}{n - p'} \quad (2.21)$$

将 $\hat{\beta}$ 的值代入 (2.19) 并进行化简，可以得到 RSS 的若干公式：

$$\begin{aligned} RSS &= Y^T Y - \hat{\beta}^T (X^T X) \hat{\beta} \\ &= Y^T Y - \hat{\beta}^T X^T Y \end{aligned}$$

$$\begin{aligned}
&= \mathcal{Y}^T \mathcal{Y} - \beta^{*T} (\mathcal{X}^T \mathcal{X}) \beta^* \\
&= \mathcal{Y}^T \mathcal{Y} - \beta^T (X^T X) \beta + n \bar{y}^2 \quad (2.22)
\end{aligned}$$

如 1.3 节所述, 如果 e 是正态分布的, 那么 $(n-p') \hat{\sigma}^2/\sigma^2$ 服从 $\chi^2(n-p')$ 分布。

在 (2.20) 中用 $\hat{\sigma}^2$ 代替 σ^2 , 我们得到 $\hat{\beta}$ 的方差估计量 $\hat{\text{var}}(\hat{\beta})$ 为:

$$\hat{\text{var}}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1} \quad (2.23)$$

矩阵形式描述的简单回归 对简单回归, X 和 Y 由

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

给出, 因此

$$(X^T X) = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (2.24)$$

可以证明 $(X^T X)^{-1}$ 为

$$(X^T X)^{-1} = \frac{1}{SXX} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \quad (2.25)$$

故由前面的结论, 有

$$\begin{aligned}
\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} &= (X^T X)^{-1} X^T Y \\
&= \frac{1}{SXX} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \frac{SXY}{SXX} \end{bmatrix}
\end{aligned}$$

并且, 由于 $\sum x_i^2 / (nSXX) = 1/n + \bar{x}^2 / SXX$, 上一章求得的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差和协方差与由 $\sigma^2 (X^T X)^{-1}$ 给出的相同。

在样本均值偏差的形式中, 结论更加简单, 因为

$$\mathcal{X}^T \mathcal{X} = SXX, \quad \mathcal{X}^T \mathcal{Y} = SXY$$

及

$$\hat{\beta}_1 = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y} = \frac{SXY}{SXX}$$

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

燃料消耗数据 (续) 现在我们对燃料消耗数据, 拟合所有 $p=4$ 个自变量的模型。我们用 $FUEL$ 关于 TAX 、 $DLIC$ 、 INC 、 $ROAD$ 表示“拟合模型 $RUEL = \hat{\beta}_0 + \hat{\beta}_1 TAX + \hat{\beta}_2 DLIC + \hat{\beta}_3 INC + \hat{\beta}_4 ROAD$ ”。对于这些数据和这个模型, 矩阵 T [等式 (2.17)] 为

$$T = \begin{bmatrix} 42.48627 & -71.39733 & 0.32465 & -81.46385 & -2256.28833 \\ -71.39733 & 1446.16667 & 23.48977 & -58.37527 & 20388.66667 \\ 0.32465 & 23.48977 & 15.46508 & 4.72193 & -738.62083 \\ -81.46385 & -58.37527 & 4.72193 & 572.95925 & 349.62058 \\ -2256.28833 & 20388.66667 & -738.62083 & 349.62058 & 588366.47917 \end{bmatrix}$$

$(X^T X)^{-1}$ 由表 2.3 给出,

表 2.3 燃料消耗数据的 $(X^T X)^{-1}$

$(\mathcal{X}^T \mathcal{X})^{-1}$ 由位于右下部的 4×4 子矩阵给出

	截距	TAX	$DLIC$	INC	$ROAD$
截距	7.8301941	-.4265133	-.0611076	-.1495090	-.0753492
TAX	-.4265133	.0382636	.0022158	-.0059137	.0057148
$DLIC$	-.0611076	.0022158	.0008411	-.0014500	.0004127
INC	-.1495090	-.0059137	-.0014500	.0674600	-.0015445
$ROAD$	-.0753492	.0057148	.0004127	-.0015445	.0026126

使用本章中基于校正平方和的公式, 估计量 β^* 为

$$\beta^* = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} -34.790149 \\ 13.364494 \\ -66.588752 \\ -2.425889 \end{bmatrix}$$

截距的估计为

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}^{*T} \bar{x} = 377.2911$$

残差平方和为

$$\begin{aligned} RSS &= \mathcal{Y}^T \mathcal{Y} - \hat{\beta}^{*T} (\mathcal{X}^T \mathcal{X}) \hat{\beta}^* \\ &= 588366.48 - 339316.51 = 189049.97 \end{aligned}$$

残差均方为

$$\hat{\sigma}^2 = \frac{RSS}{n - p'} = \frac{189049.97}{48 - (4 + 1)} = 4396.5 \text{ (43d.f.)}$$

由 $\hat{\sigma}^2$ 和 $(X^T X)^{-1}$ 可求得诸 $\hat{\beta}_j$ 的标准误和协方差的估计。例如

$$se(\hat{\beta}_0) = \hat{\sigma} \sqrt{7.83019} = 185.54$$

$$se(\hat{\beta}_4) = \hat{\sigma} \sqrt{0.0026126} = 3.3892$$

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \hat{\sigma}^2 (0.0022158) = 0.1469$$

表 2.4 FUEL 关于 TAX、DLIC、INC、ROAD 回归的计算机程序输出的回归分析简要

变量	估计	标准误	t-值
截距	377.2911	185.5412	2.03
TAX	-34.79015	12.97020	-2.68
DLIC	13.36449	1.922981	6.95
INC	-66.58875	17.22175	-3.87
ROAD	-2.425889	3.389174	-0.72
$\hat{\sigma}^2 = 4396.511$, d.f. = 43, $R^2 = 0.6787$			

在大部分计算机程序中，通常得到的输出比这里给出的要略少一些。表 2.4 的结果是比较典型的期望得到的结果。第一列给出自变量表。第二列给出对应的 $\hat{\beta}_j$ 。第三列给出 $\hat{\sigma}$ 乘以 $(X^T X)^{-1}$ 相应的对角元素的平方根。最后一列 (t-值) 为比率 $\hat{\beta}_j / se(\hat{\beta}_j)$ ，后面将对它作简略的讨论。另外，也给出各种其它的主要统计量，如误差的自由度、 $\hat{\sigma}^2$ 和 (或) $\hat{\sigma}$ 、 R^2 。有时对结果给出了多余的数字

位数，这里就是这样。对这些数据，舍入到至多四位有效数字比较恰当。

预测和拟合值 首先考虑预测的问题。我们已经观察到一个新的 $p' \times 1$ 自变量向量 \mathbf{x}_* ，其响应变量 y_* 尚未得到。这里的问题是利用这些数据预测 y_* 。用在简单回归中同样的方法，预测值为 $\tilde{y}_* = \mathbf{x}_*^T \hat{\boldsymbol{\beta}}$ 。利用附录 2A.2，预测值的标准误 $\text{sepred}(\tilde{y}_* | \mathbf{x}_*)$ 为

$$\text{sepred}(\tilde{y}_* | \mathbf{x}_*) = \hat{\sigma} \sqrt{(1 + \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_*)} \quad (2.26)$$

类似地，对自变量的值 \mathbf{x} 的所有单元的均值的估计为 $\hat{y} = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$ ，它的标准误 $\text{sefit}(\hat{y} | \mathbf{x})$ 为

$$\text{sefit}(\hat{y} | \mathbf{x}) = \hat{\sigma} \sqrt{(\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x})} \quad (2.27)$$

计算上的考虑 最小二乘估计量可以直接用公式 (2.15) 或 (2.18) 进行计算，有关的统计量也可以如本节所述的进行计算。不过，用这种方法常常会引起数值上的问题。计算上的主要问题是求矩阵 $\mathbf{X}^T \mathbf{X}$ 及其逆矩阵。这通常用三种方法中的一种解决。最常用的方法是使用高斯消去法，然后用习题 2.7 中描述的扫描算法完成。这一算法假设等式 (2.17) 的矩阵 T 是精确计算的。扫描算法计算了 $\mathbf{X}^T \mathbf{X}$ 的逆阵。第二种方法由矩阵 T 开始，但是使用了矩阵分解以避免计算 $(\mathbf{X}^T \mathbf{X})^{-1}$ 。这一方法称为 T 的乔勒斯基分解。第三种方法避免计算 T ，直接对原始数据进行所有的计算。后两种方法是密切相关的，它们的某些方面的描述在习题 2.4 中给出。Linpack Users Guide (Dongarra et al., 1979) 给出了进行最小二乘问题计算的好算法。

2.3 方差分析

对于多元回归，方差分析是用于分离变异性及比较包含不同变量集合的模型的一种很有用的技术。在整个的方差分析中，完全模型

$$Y = X\beta + e \quad (2.28)$$

与无 X 变量的模型

$$Y = \beta_0 \mathbf{1} + e \quad (2.29)$$

进行比较, 其中 $\mathbf{1}$ 是一个 $n \times 1$ 的 1 的向量。这两个式子分别与 (1.21) 和 (1.18) 对应。于是, 对模型 (2.29), $\hat{\beta}_0 = \bar{y}$, 且残差平方和为 $SY\bar{Y}$ 。另一方面, 对模型 (2.28), β 的估计由 (2.15) 给出, RSS 由 (2.19) 给出。很明显, 有 $RSS < SY\bar{Y}$, 两者的差

$$SS_{\text{reg}} = SY\bar{Y} - RSS \quad (2.30)$$

对应于由大模型解释而未由小模型解释的 Y 的平方和。 SS_{reg} 的自由度等于 $SY\bar{Y}$ 的 d.f. 值减去 RSS 的 d.f. 值, 它等于 p 。

这些结论概述于下面的方差分析表中:

来 源	方差分析 (整个的)		
	d. f.	SS	MS
关于 x_1, \dots, x_p 的回归	p	SS_{reg}	SS_{reg}/p
残差	$n - p'$	RSS	$RSS/(n - p') = \hat{\sigma}^2$
总的	$n - 1$	$SY\bar{Y}$	

我们可以通过将回归均方与 $\hat{\sigma}^2$ 的比率与 $F(p, n - p')$ 分布比较, 来决定 SS_{reg} 是否充分大, 从而判断关于 X 的回归的重要性。如果计算得到的 F 超过一个适宜的临界值, 则我们可以断言对 X 了解比对 X 不了解提供了一个更好的模型。如果误差为 $NID(0, \sigma^2)$ 且 NH 为真, 则计算得到的比率恰好服从 F 分布。这一 F -检验所检验的假设为:

NH : 应用模型 (2.29), $\beta^* = 0$

AH : 应用模型 (2.28), $\beta^* \neq 0$

测定系数 正如简单回归一样, 比率

$$R^2 = \frac{SY\bar{Y} - RSS}{SY\bar{Y}} = \frac{SS_{\text{reg}}}{SY\bar{Y}} \quad (2.31)$$

给出了由关于诸 X 的回归解释的 Y 的变异性的比例。另外，可以证明值 R^2 是 Y 与诸 X 的复相关系数的平方：它是 Y 与 X 的任一线性函数的相关系数的最大值的平方。

燃料消耗数据整个的方差分析表为：

来源	d. f.	SS	MS	F
回归	4	399316	99829	22.70
残差	43	189050	4387 = $\hat{\sigma}^2$	
总的	47	588366		

由于 $F=22.7$ 大大超过 $F(0.01; 4, 43)=3.79$ ，我们猜想至少 X 的一部分与燃料消耗有关。值 $R^2=399316/588366=0.68$ ，说明响应变量的观测到的变异性中约有 68% 来自通过诸 X 建立的模型。如果没有象这样一类问题的经验，很难确定 68% 是多还是少。

附带说一下，这个例子指出的整个的 F 统计量的计算并非总是令人感兴趣的。通常事先知道变量之间是相关的，因而预计有一个很大的检验统计量的值。我们更为关注的是检验关于某些变量的其它假设。

关于其中一个自变量的假设 在许多问题中，我们关心获得其中一个自变量的有用性的信息。我们是否可以象从四个变量一样，只从例如 $DLIC$ ， $ROAD$ 及 INC 建立模型？这一问题可以更为建议性地表述成：如果 $DLIC$ ， $ROAD$ 和 INC 已知，增加 TAX 的知识是否会有显著的改进？可以使用以下步骤：拟合不包含 TAX 的模型，得到该模型的残差平方和，然后，拟合一个包含 TAX 的模型，求出这个模型的残差平方和。从较小的模型的残差平方和中减去大模型的残差平方和，得到已包含在模型中的变量 ($DLIC$ ， $ROAD$ 和 INC) 调整后的关于 TAX 的回归平方和。这

一计算可以完全按照以下所述进行。首先计算 *FUEL* 关于 *ROAD*, *DLIC* 和 *INC* 的回归。这个模型的残差平方和为 220682 ($\hat{\sigma}^2 = 5015$)。完全模型的残差平方和已给出, 为 189050 ($\hat{\sigma}^2 = 4397$)。继其余变量之后, 关于 *TAX* 的回归平方和为 220682 - 189050 = 31632 (估计的 $\hat{\sigma}^2$ 约减少 7%)。这可以在下面的方差分析表中概述:

来 源	d. f.	SS	MS	F
关于 <i>ROAD</i> , <i>INC</i> , <i>DLIC</i> 的回归	3	367684	122561	
继其余变量后关于 <i>TAX</i> 的回归	1	31632	31632	7.19
残差	43	189050	4397	

关于 *ROAD*, *INC*, *DLIC* 的回归的 *SS* 为 $SS_{\text{reg}}(\text{ROAD}, \text{INC}, \text{DLIC}, \text{TAX}) - SS_{\text{reg}}(\text{继其余后关于 TAX}) = 399316 - 31632 = 367684$ 。均方的比 $F = 31632/4397 = 7.19$ 是用于检验继已包含在模型中的其余变量之后 *TAX* 的有用性的一个统计量; 它丝毫不说明其余变量的有用性。将它与具有自由度为 1 和 $n - p' = 43$ 的 *F* 分布作比较。在这个例子中, *TAX* 为其它变量调整后的一个显著的自变量, 因为 $F(0.01; 1, 43) = 7.26$, 给出 *p*-值接近 0.01。我们称此为部分 *F*-检验。特别, 这个 *F* 检验的假设为

$$\text{NH: } \beta_1 = 0; \beta_0, \beta_2, \beta_3, \beta_4 \text{ 任意}$$

$$\text{AH: } \beta_1 \neq 0; \beta_0, \beta_2, \beta_3, \beta_4 \text{ 任意} \quad (2.32)$$

与 *t* 统计量的关系 检验 *TAX* 的重要性的另一个合理步骤是将系数的估计量除以它的标准误所得的商与自由度为 43 的 *t* 分布比较。可以证明比率 *t* 的平方与刚才计算的 *F* 比为同一数值, 故这两个过程是等同的。因此, *t*-统计量检验关于模型中所有其它变量调整后的变量的重要性的假设, 而不是忽略其它变量。

例如, 由表 2.4, 对 *TAX* 的 *t* 统计量为

$$t = \frac{-34.79015}{12.97020} = -2.68$$

与找到的临界值 $t(43)$ 比较。这个统计量检验的假设由 (2.32) 给出。另外，我们有

$$t^2 = (-2.68)^2 = 7.18$$

考虑舍入误差，它在数值上等同于对这一假设的 F -检验的值。任何一个 β_j 有一个特定值 (β 的所有其余分量为任意) 的 t -检验可以如 1.7 节中所述的进行。

其它的假设检验 我们已经获得了关于被问题中其它变量调整的 TAX 的影响的假设的检验。同样，我们可以得到其余某些变量调整，或不考虑其它变量的 TAX 的影响的检验。一般，这些检验不是等价的：若忽略其它变量，一个变量可能被认为是有用的自变量，而被其它变量调整后，可能被认为是没有用的。另外，单独考虑为无用的一个自变量，当和其它变量一起考虑时，可能会变成重要的。这些检验的结果依赖于在 $X^T X$ 中反映出来的诸 X 间的关系，或通常更清楚地反映在样本相关系数中。因此，在多元回归中，拟合各个 X 的次序的问题是明显的。

序贯方差分析表 通过将 TAX 与另外三个自变量分离， SS_{reg} 被分为两块，一部分为拟合前三个自变量，另一部分为继这三个变量以后拟合 TAX 。通过对每个变量，将回归平方和分块，可以继续这一分块活动。除非所有自变量为正交的，这个划分不是唯一的。例如，我们可以首先拟合 $DLIC$ ，接着由 $DLIC$ 调整后拟合 TAX ，然后由 $DLIC$ 和 TAX 调整后拟合 INC ，最后由其余三个变量调整后拟合 $ROAD$ 。结果在表 2.5 (a) 中给出。另一方面，我们可以按次序 $ROAD$, INC , $DLIC$ 及 TAX 进行拟合，结果见表 2.5 (b)。可以看到，得到的相应的平方和是很不同的，首先拟合 $ROAD$ 时， $ROAD$ 的平方和是 213，而被其它三个变量调整后 $ROAD$ 具有平方和 2252，大了约 10 倍，但与 $\hat{\sigma}^2$ 相比仍不是很大。

表 2.5 具有不同拟合次序的两个方差分析表

	变量	d. f.	SS	MS
(a) 第一次分析				
首先	<i>DLIC</i>	1	287448	287448
然后	<i>TAX</i>	1	40084	40084
然后	<i>INC</i>	1	69532	69532
然后	<i>ROAD</i>	1	2252	2252
	残差	43	189050	4387
(b) 第二次分析				
首先	<i>ROAD</i>	1	213	213
然后	<i>INC</i>	1	35642	35642
然后	<i>DLIC</i>	1	331828	331829
然后	<i>TAX</i>	1	31632	31632
	残差	43	189050	4397

2.4 附加变量图

在简单回归中，响应变量 Y 与自变量 X 的关系用散点图表示。在多元回归中，由于多个自变量之间的关系，情况是复杂的，所以 Y 与任何一个 X 的散点图不能够反映被其它 X 调整后得到的关系。附加变量图是用来显示这种关系的图示方法。

我们已经在图 2.2 (d) 中看到附加变量图的一个例子，它表示 *DLIC* 调整后 *FUEL* 与 *TAX* 之间的关系。一般步骤如下：

1. 考虑模型

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e \quad (2.33)$$

我们感兴趣的是被其它的 X 调整后， X_k 与 Y 之间的关系 ($1 \leq k \leq p$)。

2. 拟合 Y 关于除 X_k 以外的所有的 X 的回归，记下这个回归

的残差。称残差为 $\hat{e}_Y(X_k)$ ，不幸的是记号比较复杂，但是我们必须同时记下响应变量 Y 和未用于计算残差的自变量 X_k 。 $\hat{e}_Y(X_k)$ 是 Y 中未被除了 X_k 以外所有的 X 解释的部分。

3. 拟合 X_k 关于其它 X 的回归，记下所得的残差，称为 \hat{e}_k 。这是 X_k 中未被其它 X 解释的部分。在许多计算机程序中计算和保留残差是容易的。我们所感兴趣的由其它 X 调整后， Y 与 X_k 的关系正是两个残差集合之间的关系。

4. 作 $\hat{e}_Y(X_k)$ 对 \hat{e}_k 的图。这就是附加变量图。图中数量间的强烈的线性关系与被调整后 Y 与 X_k 之间强烈的线性关系相对应。如果图中未显示很强的趋势，则被调整后的关系是弱的。总之，可以类似于简单回归中的散点图对这个图进行解释。

附加变量图的性质 如果由普通的最小二乘法求 $\hat{e}_Y(X_k)$ 关于 \hat{e}_k 的回归，可以发现当截距作为一个变量包含在原始模型 (2.33) 中时，截距恰巧为零。斜率与 (2.33) 中 X_k 的系数一样。这个回归的残差也与 (2.33) 的残差相同。如果将自由度 $n-2$ 改为 $n-p'$ ，则 $\hat{\beta}_k$ 的标准误及残差均方 $\hat{\sigma}^2$ 与拟合 (2.33) 过程中所得的一致。这样，在实际意义上，附加变量图确实概括了被其它的 X 调整后 Y 与 X_k 之间的关系。进一步的细节由 Cook 和 Weisberg (1982, 2.3.2 节) 给出。

图 2.3 对燃料数据四个自变量中的每一个给出了附加变量图。在每个图中，作出了一组残差关于另一组残差的回归直线。从这些图可以看出若干明显的事实。对 $DLIC$ 的图表现出最强的趋势，因为数据与拟合的直线最为匹配。最弱的被调整后的自变量是 $ROAD$ ，因为在图 2.3 (d) 中几乎看不出趋势。附加变量图使分析者可以看出非线性性，以及任何或是决定拟合直线或是远离拟合直线的点。尽管有一些游离的点，但是非线性性或重要的偏离点是不明显的。

偏残差图 作为附加变量图的替换，Ezekiel (1924)，Larsen 和 McCleary (1975) 和 Wood (1975) 提出使用偏残差图，亦称

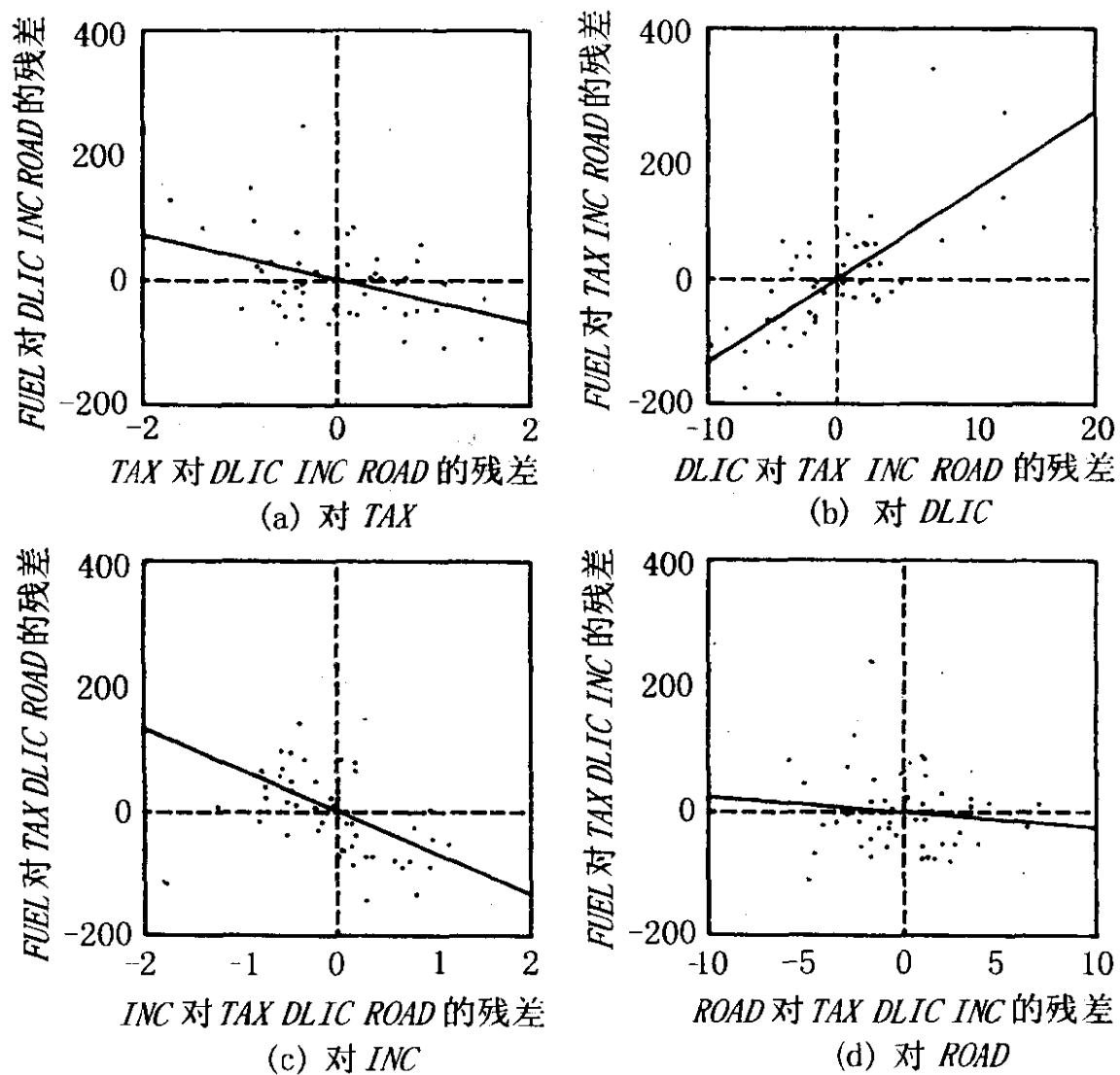


图 2.3 附加变量图

为残差加分量图，即 $\hat{e}_i + \hat{\beta}_k x_{ik}$ 对 x_{ik} 的图，其中 $\hat{e}_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ 是拟合的模型 (2.33) 的残差。这个图也具有等于 $\hat{\beta}_k$ 的斜率，但偏离拟合直线的点的散布与附加变量图不同。附加变量图的优点是在显示被其它 X 调整后 Y 与 X_k 之间的关系时，可以直接解释。另一方面，将 X_k 变换为另一尺度的必要性在偏残差图中似乎得到了更好的反映。这两种图在回归模型中都有其作用。

2.5 通过原点的回归

截距 $\beta_0=0$ 的模型可从某些方面引出。最明显的情况是如果 $x=0$ 必有 $y=0$ 。另外, 如果一个试验, 其一些自变量的和总是等于一个常数, 则拟合一个非零截距将导致秩不足的模型。例如, 混料试验常常要求选择若干化学药品的比例 x_1, \dots, x_p , 对所有的案例, 有 $\sum x_p = 1$ 。对这样一个模型, 只有当诸 x_j 中的一个被从模型中删除后, 才有可能拟合一个截距。去掉截距, 而不是从诸自变量中去掉一个自变量常常使结果的解释简化。

对于通过原点的回归, 可以用我们目前讨论过的同样的方法处理: 定义 X 为 $n \times p$ 矩阵, 它不含一列 1, β 为除 β_0 以外的 $p \times 1$ 向量。经过这些变化后, 模型(2.12)是有效的, 并且如果 $p' = p$ 而不是 $p+1$, 由 (2.12) 得到的所有矩阵结果仍适用。

符号表示上的习惯 为使同样的矩阵结果可应用于通过原点的回归, 故如果模型中包含截距, 令 $p' = p+1$, 而如果回归过原点, 令 $p' = p$, 视 X 为 $n \times p'$ 且 β 为 $p' \times 1$ 。

问 题

2.1 伯克来指导研究。本例的数据摘自伯克来指导研究, 为对一群出生于伯克来, 加利福尼亚州的, 1928 年 1 月至 1929 年 6 月间的男孩和女孩的纵向观测。数据中包含的变量为:

WT_2 = 2 岁时的体重 (kg)

HT_2 = 2 岁时的身高 (cm)

WT_9 = 9 岁时的体重 (kg)

HT_9 = 9 岁时的身高 (cm)

LG_9 = 9 岁时腿部周长 (cm)

ST_9 = 9 岁时的力量的复合测试 (高的值 = 较强壮)

WT_{18} = 18 岁时的体重

HT_{18} = 18 岁时的身高

LG_{18} = 18 岁时腿部周长

ST_{18} = 18 岁时的力量

$SOMA$ = 体型, 七分制, 作为肥胖程度的测试 (1 = 苗条的, 7 = 胖的), 根据 18 岁时的照片决定。

关于 26 个男孩和 32 个女孩的数据分别在表 2.6 和表 2.7 给出。(包含更大的样本量及更多变量的完整的研究, 细节见 Tuddenham 和 Snyder, 1954)

2.1.1 对女孩组求样本相关系数矩阵、样本均值及样本标准差。

2.1.2 对女孩组拟合模型

$$SOMA = \beta_0 + \beta_1 HT_2 + \beta_2 WT_2 + \beta_3 HT_9 + \beta_4 WT_9 + \beta_5 ST_9 + e$$

求 $\hat{\sigma}$ 、 R^2 , 整个的方差分析和 F -检验, 说明 F -检验的结果。计算用于检验模型中每个 β_j 等于零的 t -统计量。明确地陈述每个被检验的假设和结论。

2.1.3 对女孩组拟合模型

$$SOMA = \beta_0 + \beta_3 HT_9 + \beta_4 WT_9 + \beta_5 ST_9 + e$$

并将这个模型与 2.1.2 中拟合的模型比较 (即计算 F -检验)。

2.1.4 对男孩组重复 2.1.1 至 2.1.3。定量地描述对男孩组和女孩组拟合的模型的区别 (正规步骤在第 7 章中研究)。

2.1.5 对男孩组, 把变量 WT_9 增加到模型 $HT_{18} = \beta_0 + \beta_2 WT_2$ 中去, 重复 2.1 节中的推导。

2.2 (矩阵操作) 定义下列矩阵:

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 1 \\ 2 & 1 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$E = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad H = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

2.2.1 求 A^T , B^T , C^T , D^T , E^T 。

2.2.2 求 $A+B$

2.2.3 求 AB 和 BA , AB 是否等于 BA ?

2.2.4 证明 $(AB)^T = B^T A^T$

2.2.5 计算 $C^T C$ 和 CC^T , 它们相等吗?

2.2.6 求 DE^T , $D^T E$ 。

表 2.6 伯克来指导研究:男孩

编号	WT ₂	HT ₂	WT ₉	HT ₉	LG ₉	ST ₉	WT ₁₈	HT ₁₈	LG ₁₈	ST ₁₈	SOMA
201	13.6	90.2	41.5	139.4	31.6	74.0	110.2	179.0	44.1	226.0	7.0
202	12.7	91.4	31.0	144.3	26.0	73.0	79.4	195.1	36.1	252.0	4.0
203	12.6	86.4	30.1	136.5	26.6	64.0	76.3	183.7	36.9	216.0	6.0
204	14.8	87.6	34.1	135.4	28.2	75.0	74.5	178.7	37.3	220.0	3.0
205	12.7	86.7	24.5	128.9	24.2	63.0	55.7	171.5	31.0	200.0	1.5
206	11.9	88.1	29.8	136.0	26.7	77.0	68.2	181.8	37.0	215.0	3.0
207	11.5	82.2	26.0	128.5	26.5	45.0	78.2	172.5	39.1	152.0	6.0
209	13.2	83.8	30.1	133.2	27.6	70.0	66.5	174.6	37.3	189.0	4.0
210	16.9	91.0	37.9	145.6	29.0	61.0	70.5	190.4	33.9	183.0	3.0
211	12.7	87.4	27.0	132.4	26.0	74.0	57.3	173.8	33.3	193.0	3.0
212	11.4	84.2	25.9	133.7	25.8	68.0	50.3	172.6	31.6	202.0	3.0
213	14.2	88.4	31.1	138.3	27.3	59.0	70.8	185.2	36.6	208.0	4.0
214	17.2	87.7	34.6	134.6	30.6	87.0	73.7	178.4	39.2	227.0	3.0
215	13.7	89.6	34.6	139.0	28.9	71.0	75.2	177.6	36.8	204.0	2.5
216	14.2	91.4	43.1	146.0	32.4	98.0	83.1	183.5	38.0	226.0	4.0
217	15.9	90.0	33.2	133.2	28.5	82.0	74.3	178.1	37.8	233.0	2.5
218	14.3	86.4	30.7	133.3	27.3	73.0	72.2	177.0	36.5	237.0	2.0
219	13.3	90.0	31.6	130.3	27.5	68.0	88.6	172.9	40.4	230.0	7.0
221	13.8	91.4	33.4	144.5	27.0	92.0	75.9	188.4	36.5	250.0	1.0
222	11.3	81.3	29.4	125.4	27.7	70.0	64.9	169.4	35.7	236.0	3.0
223	14.3	90.6	30.2	135.8	26.7	70.0	65.6	180.2	35.4	177.0	4.0
224	13.4	92.2	31.1	139.9	27.2	63.0	66.4	189.0	35.3	186.0	4.0
225	12.2	87.1	27.6	136.8	25.8	73.0	59.0	182.4	33.5	199.0	3.0
226	15.9	91.4	32.3	140.6	27.9	69.0	68.1	185.8	34.2	227.0	1.0
227	11.5	89.7	29.0	138.6	24.6	61.0	67.7	180.7	34.3	164.0	4.0
228	14.2	92.2	31.4	140.0	28.2	74.0	68.5	178.7	37.0	219.0	2.0

表 2.7 伯克来指导研究: 女孩

编号	WT ₂	HT ₂	WT ₉	HT ₉	LG ₉	ST ₉	WT ₁₈	HT ₁₈	LG ₁₈	ST ₁₈	SOMA
331	12.6	83.8	33.0	136.5	29.0	57.0	71.2	169.6	38.8	107.0	6.0
334	12.0	86.2	34.2	137.0	27.3	44.0	58.2	166.8	34.3	130.0	5.0
335	10.9	85.1	28.1	129.0	27.4	48.0	56.0	157.1	37.8	101.0	5.0
351	12.7	88.6	27.5	139.4	25.7	68.0	64.5	181.1	34.2	149.0	4.0
352	11.3	83.0	23.9	125.6	24.5	22.0	53.0	158.4	32.4	112.0	5.0
353	11.8	88.9	32.2	137.1	28.2	59.0	52.4	165.6	33.8	136.0	4.0
354	15.4	89.7	29.4	133.6	26.6	58.0	56.8	166.7	32.7	118.0	4.5
355	10.9	81.3	22.0	121.4	24.4	44.0	49.2	156.5	33.5	110.0	4.0
356	13.2	88.7	28.8	133.6	26.5	58.0	55.6	168.1	34.1	104.0	4.5
357	14.3	88.4	38.8	134.1	31.1	57.0	77.8	165.3	39.8	138.0	6.5
358	11.1	85.1	36.0	139.4	28.2	64.0	69.6	163.7	38.6	108.0	5.5
359	13.6	91.4	31.3	138.1	27.6	64.0	56.2	173.7	34.2	134.0	3.5
361	13.5	86.1	33.3	138.4	29.4	73.0	64.9	169.2	36.7	141.0	5.0
362	16.3	94.0	36.2	139.5	28.0	52.0	59.3	170.1	32.8	122.0	4.5
364	10.2	82.2	23.4	129.8	22.6	60.0	49.8	164.2	30.0	128.0	4.0
365	12.6	88.2	33.8	144.8	28.3	107.0	62.6	176.0	35.8	168.0	5.0
366	12.9	87.5	34.5	138.9	30.5	62.0	66.6	170.9	38.8	126.0	5.0
367	13.3	88.6	34.4	140.3	31.2	88.0	65.3	169.2	39.0	142.0	5.0
368	13.4	86.9	38.2	143.8	29.8	78.0	65.9	172.0	35.7	132.0	5.5

(续表)

编号	WT ₂	HT ₂	WT ₉	HT ₉	LG ₉	ST ₉	WT ₁₈	HT ₁₈	LG ₁₈	ST ₁₈	SOMA
369	12.7	86.4	31.7	133.6	27.5	52.0	59.0	163.0	32.7	116.0	5.5
370	12.2	80.9	26.6	123.5	27.2	40.0	47.4	154.5	32.2	112.0	4.0
371	15.4	90.0	34.2	139.9	29.1	71.0	60.4	172.5	35.7	137.0	4.0
372	12.7	94.0	27.7	136.1	26.7	30.0	56.3	175.6	34.0	114.0	3.0
373	13.2	89.7	28.5	135.8	25.5	76.0	61.7	167.2	35.5	122.0	4.5
374	12.4	86.4	30.5	131.9	28.6	59.0	52.4	164.0	34.8	121.0	5.0
376	13.4	86.4	39.0	130.9	29.3	38.0	58.4	161.6	33.0	107.0	6.5
377	10.6	81.8	25.0	126.3	25.0	50.0	52.8	153.6	33.4	140.0	5.0
380	12.7	91.4	29.8	135.5	27.0	57.0	67.4	173.5	34.5	123.0	5.0
382	11.8	88.6	27.0	134.0	26.5	54.0	56.3	166.2	36.2	135.0	4.5
383	13.3	86.4	41.4	138.2	32.5	44.0	82.8	162.8	42.5	125.0	7.0
384	13.2	94.0	41.6	142.0	31.0	56.0	68.1	168.6	38.4	142.0	5.5
385	15.9	89.2	42.4	140.8	32.6	74.0	63.1	169.2	37.9	142.0	5.5

2.2.7 证明 H 是正交的 (即 $HH^T = H^TH = I$)。

2.3 分块矩阵: 一个 $n \times p$ 的矩阵 C 可以按列分块为 $C = (C_1 \ C_2)$, 其中 C_1 为一个 $n \times p_1$ 矩阵, C_2 为一个 $n \times (p - p_1)$ 矩阵, C_1 为 C 的前 p_1 列, C_2 为 C 的后 $p - p_1$ 列。由这一定义

$$\begin{aligned} C^TC &= (C_1 \ C_2)^T(C_1 \ C_2) \\ &= \begin{pmatrix} C_1^T \\ C_2^T \end{pmatrix} (C_1 \ C_2) \\ &= \begin{pmatrix} C_1^TC_1 & C_1^TC_2 \\ C_2^TC_1 & C_2^TC_2 \end{pmatrix} \end{aligned}$$

且 $CC^T = (C_1 \ C_2)(C_1 \ C_2)^T = C_1C_1^T + C_2C_2^T$ 。

2.3.1 通过直接相乘证明, 如果 (C^TC) 是满秩的, 则有

$$(C^TC)^{-1} = \begin{pmatrix} (C_1^TC_1)^{-1} + FE^{-1}F^T & -FE^{-1} \\ -E^{-1}F^T & E^{-1} \end{pmatrix}$$

其中

$$E = C_2^TC_2 - C_2^TC_1(C_1^TC_1)^{-1}C_1^TC_2,$$

及

$$F = (C_1^TC_1)^{-1}C_1^TC_2$$

2.3.2 假定正确的线性模型为 $Y = X\beta + e$, X 为一个 $n \times p'$ 矩阵, 但我们拟合模型 $Y = X\beta + Z\gamma + e$, 且我们并不知道 $\gamma = 0$ 。用 2.3.1 的结果证明, 在后面一个模型中, 对 β 的估计是无偏的, 并求其方差。比较由正确的线性模型得到的 $\hat{\beta}$ 的方差与如果拟合较大的模型所得的 $\hat{\beta}$ 的方差, 并加以评论。求关于 X 和 Z 的条件, 使得对任意一个模型, β 的估计在数值上相等。

2.4 QR 因子分解。假设我们有一个 $n \times p'$ 矩阵 Q 和一个 $p' \times p'$ 上三角矩阵 R , 使 $Q^TQ = I$ 及 $QR = X$ (如附录 2A.3)。

2.4.1 证明 $R^TR = X^TX$ 。

2.4.2 证明如果 $(X^TX)^{-1}$ 存在, 则 $(X^TX)^{-1}X^TY = R^{-1}Q^TY$ 。于是, 为了计算 $\hat{\beta}$, 首先计算 $z = Q^TY$, 然后用回代法解线性方程组 $R\hat{\beta} = z$; 见附录 2A.3。

2.4.3 证明拟合值向量 $\hat{Y} = QQ^TY$ 。这附带地可以证明 $QQ^T = X(X^TX)^{-1}X^T$, 这在第 5 和第 6 章中将是一个重要的矩阵。另外, 由 Y 和 Q 求 \hat{e} 。

2.4.4 对任何向量 c , 拟合值 $c^T\hat{\beta}$ 的方差为 $\sigma^2c^T(X^TX)^{-1}c$ 。如果向量 c 的第 j 个元素等于 1 而其余元素都等于 0, 则 $\sigma^2c(X^TX)^{-1}c = \text{var}(\hat{\beta}_j)$ 。这样, 如果可以对任意 $p' \times 1$ 向量 c 求得 c^T

$(X^T X)^{-1}c$, 我们就可以求出任何拟合值、预测值、系数的线性组合或任一单个系数的方差估计。由 2.4.1, 因为 $X^T X = R^T R$, 如果逆矩阵存在, 则 $(X^T X)^{-1} = (R^T R)^{-1} = R^{-1} R^{-T}$, 其中 $-T$ 表示转置的逆。证明

$$c^T (X^T X)^{-1} c = (R^{-T} c)^T (R^{-T} c)$$

假设我们令 $d = R^{-T} c$ 。则 $c^T (X^T X)^{-1} c = d^T d$ 。为了求 d , 记

$$d = R^{-T} c$$

或

$$R^T d = c$$

用附录 2A.3 中的回代法解 d 。这样避免直接求 $(X^T X)^{-1}$ 或 R 的逆。

2.4.5 假设 $p' = 3$ 且

$$R = \begin{bmatrix} 2 & 4 & 3 \\ 0 & 1 & 5 \\ 0 & 0 & 8 \end{bmatrix}$$

假定 $\sigma^2 = 1$, 求 $\text{var}(\hat{\beta}_0)$ (令 $c = (1, 0, 0)^T$), $\text{var}(\hat{\beta}_1)$ (令 $c = (0, 1, 0)^T$) 及 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的协方差。后者需要将上述结论略加推广。

2.5 计算的例子 考虑线性模型 $Y = X\beta + e$, 其中

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \end{bmatrix}, \quad Y = \begin{bmatrix} 34 \\ 47 \\ 55 \\ 64 \end{bmatrix}$$

2.5.1 计算 $X^T X, X^T Y, Y^T Y$ 。利用规则: 如果 A 是一个 2×2 对称矩阵, 则当 $ac \neq b^2$ 时,

$$A^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix}$$

求 $(X^T X)^{-1}$, $\hat{\beta}$ 和 $\text{var}(\hat{\beta})$ 。求 \hat{Y} 和 \hat{e} 。

2.5.2 定义

$$Q = \begin{bmatrix} -\frac{1}{2} & \frac{3\sqrt{5}}{10} \\ -\frac{1}{2} & \frac{\sqrt{5}}{10} \\ -\frac{1}{2} & -\frac{\sqrt{5}}{10} \\ -\frac{1}{2} & -\frac{3\sqrt{5}}{10} \end{bmatrix}, \quad R = \begin{pmatrix} -2 & -8 \\ 0 & -2\sqrt{5} \end{pmatrix}$$

证明 $Q^T Q = I$ 及 $QR = X$ 。验证 $R^T R = X^T X$ 。求 $z = Q^T Y$ 及 $\hat{Y} = Qz$, 计算 \hat{e} , 用回代法求 β 。

2.6 用适当的方法研究在燃料消耗模型中加入 POP , 并总结你的结论。

2.7 扫描 (Sweep) 算法, 定义 A 为一个 $(p' + 1) \times (p' + 1)$ 矩阵

$$A = \begin{pmatrix} X^T X & X^T Y \\ Y^T X & Y^T Y \end{pmatrix}$$

定义: 我们称元素为 a_{ij} 的矩阵 A 在支点 k 被扫描, 如果它被根据下列规则转化为元素为 b_{ij} 的矩阵 B

$$b_{kk} = \frac{1}{a_{kk}}$$

$$b_{ik} = -\frac{a_{ik}}{a_{kk}} \quad (i \neq k)$$

$$b_{kj} = \frac{a_{kj}}{a_{kk}} \quad (j \neq k)$$

$$b_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \quad (i \neq k, j \neq k)$$

用扫描算法实现矩阵 A 到矩阵 B 的转化, 编程很容易。首先, 用 $b_{kk} = 1/a_{kk}$ 代替 a_{kk} 。然后用 $b_{ik} = -a_{ik}b_{kk}$ 代替第 k 列的其余元素。既不在第 k 行又不在第 k 列的元素用 $b_{ij} = a_{ij} + a_{kj}b_{ik}$ 代替。最后, 第 k 行用 $b_{kj} = a_{kj}b_{kk}$ 代替。扫描算法在 Goodnight (1979) 给出的表述中符号略有不同。这个名称是由 A. Beaton (1964) 给出的, 然而算法在更早已被使用 (Ralston, 1960)。

2.7.1 证明: 如果 X 的第一列是一列 1, 则 A 的第一行为向量

$$(n, n\bar{x}_1, \dots, n\bar{x}_p, n\bar{y})^T = (n, nX^T, n\bar{y})$$

2.7.2 设记号 $B = \text{Sweep } A [i, j, \dots, m]$ 表示 “ B 为在支点 i 扫描矩阵 A , 然后在支点 j 扫描结果矩阵, \dots , 然后在交点 m 扫描结果矩阵所得的结果矩阵”。证明:

$$\text{Sweep } A [1] = \begin{bmatrix} \frac{1}{n} & \bar{x}^T & \bar{y} \\ -\bar{x} & X^T X & X^T Y \\ -\bar{y} & Y^T X & Y^T Y \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & \bar{x}^T & \bar{y} \\ -\bar{x} & T & \\ -\bar{y} & & \end{bmatrix}$$

这样, 除了符号改变外, 在第一个支点扫描矩阵使 A 变为校正叉积矩阵 T , 加上新的第一行及第一列, 它们是 X 和 Y 的均值。一个使用扫描算法的回归计算程序通常首先直接计算 $A[1]$, 可能会运用附录 1A.5 中描述的算法, 因为这会在大多数问题中

提高计算精度。

2.7.3 写一个计算机程序来对回归计算实现扫描算法。(你必须阅读完本题的其余部分后再开始编写程序)。这个程序开始时读入如在习题 2.7.2 中定义的矩阵 Sweep $A[1]$ 。这一矩阵的计算可以使用为习题 1.10 所写的程序。假设我们称这一矩阵为 C ，关于这一程序需要注意两点：

- ① 存储矩阵 C 的两个拷贝，可以称它们为 C 和 CC 。由于计算上的原因可能需要恢复原始矩阵。可以只扫描矩阵 C ， CC 不动，在需要的时候拷贝 CC 至 C 。
- ② 算法中的一个操作是将对角线元素 b_{kk} 变为其逆 $1/b_{kk}$ ，如果 b_{kk} 接近于零，这可能会引起数值上的问题。不考虑舍入误差，若 b_{kk} 恰好为零，则矩阵 B 为不可逆的，唯一的最小二乘估计不存在。由于数字计算机内在的舍入误差，需要有容差检查：如果 $b_{kk} < \text{tol}/SX_kX_k$ ，则认为 b_{kk} 等于零，其中 tol 为根据计算机字长事先确定的数， SX_kX_k 为 X 的第 k 列的值的校正平方和。一般常选 $\text{tol}=0.001$ 。如果容差检查失败，变量 X_k 本质上为对已扫过的支点的变量的线性组合，它必须从模型中删除。

Berk (1977) 指出这一容差检查对保证数值上的精确计算不是充分的，因为加入第 k 个支点可能改变已有的一个支点，使它太大，从而在数值上不稳定。假设已对第 j 个支点进行操作，我们现在考虑第 k 个支点。Berk 建议采用下列步骤。如果 b_{kk} 通过了扫过支点 k 的容差检查，然后检查对应于先前每一次扫过的支点。如果这些支点中的第 j 个太大，超过了 SX_jX_j/tol ，则舍弃第 k 列的支点，然后称 X_k 为已进入等式的列的线性组合。这很可能需要将 CC 拷贝回 C 后重新计算扫描的矩阵，重复扫描的过程直到第 k 个（不包括第 k 个）。

- 2.7.4 不考虑舍入误差，或者在数学上证明或者利用你的程序证明扫描是可逆的：如果 $B = \text{Sweep } A[i]$ 则 $A = \text{Sweep } B[i]$ 。
- 2.7.5 不考虑舍入误差，或者在数学上证明或者利用你的程序证明扫描是可交换的： $\text{Sweep } A[i, j] = \text{Sweep } A[j, i]$ 。
- 2.7.6 不考虑舍入误差，或者在数学上证明或者利用你的程序证明：如果 A 是可逆的，则通过每次以一个为支点扫过 A 可以计算 A^{-1} 。
- 2.7.7 用 $0, 1, \dots, p, p+1$ 标记 A 的支点。第 0 个支点对应于常数列 1 ，而第 $(p+1)$ 个支点对应于响应变量 Y 。证明：

$$\text{Sweep } A[0,1,2,\dots,p] = \begin{pmatrix} (X^T X)^{-1} & \beta \\ -\beta^T & RSS \end{pmatrix}$$

其中，如通常所述的， $\beta = (X^T X)^{-1} X^T Y$ ， RSS 为 Y 关于诸 X 的回归的残差平方和，这个回归包含截距。（如果回归通过原点，则不要在支点 0 扫过）。

2.7.8 描述如何使用扫描算法对 Y 关于诸 X 的任何子集的回归估计参数。

2.7.9 描述一个使用扫描的算法，以求得如表 2.5 的序贯方差分析表。

2.7.10 使用本题的结论，扩充扫描程序，以计算和打印本章所述的多元回归的主要统计量。特别地，程序必须计算 β 、 σ^2 、 R^2 和 β 的分量的标准误。

2.8 1965~1977 年，在旧金山海湾地区，臭氧层超过联邦标准 1 个小时以上的天数，每年约减少 5%，但有很大的未得到解释的波动。波动的一个可能的原因是先前一、二年的天气。例如，冬天的降雨量可能影响夏天的臭氧层。

下列数据来自 Sandberg, Basso 及 Okin (1978)：

$YEAR$ = 臭氧层测量的年份。

$RAIN$ = 在前两个冬季，旧金山海湾地区的平均冬季降雨量，以厘米计。

SF = 在旧金山夏季每小时平均臭氧的最大读数，以块/百万计。

SJ = 在位于海湾南端的 San Jose，夏季每小时平均臭氧的最大读数，以块/百万计。

2.8.1 以 SF 为响应变量，拟合 SF 关于 $YEAR$ 的回归。然后求继 $YEAR$ 后对 $RAIN$ 的附加变量图。 $RAIN$ 是一个有用的自变量吗？

2.8.2 重复上题，但首先拟合 $RAIN$ ，并评述第二次拟合的 $YEAR$ 。

2.8.3 计算两个新变量， $SUE = SF + SJ$ 及 $DIFF = SF - SJ$ 。研究 SUM 关于 $YEAR$ 和 $RAIN$ 以及 $DIFF$ 关于 $YEAR$ 和 $RAIN$ 的回归模型。解释结论。

$YEAR$	$RAIN$	SF	SJ	$YEAR$	$RAIN$	SF	SJ
1965	18.9	4.3	4.2	1972	19.0	3.1	4.6
1966	23.7	4.2	4.8	1973	30.6	3.4	5.1
1967	26.2	4.6	5.3	1974	34.1	3.4	3.7
1968	26.6	4.7	4.8	1975	23.7	2.1	2.7
1969	39.6	4.1	5.5	1976	14.6	2.2	2.1
1970	45.5	4.6	5.6	1977	7.6	2.0	2.5
1971	26.7	3.7	5.4				

3

下 结 论

计算完成以后，必须对结论作出解释。尽管在许多问题中对分析的大致描述基本相似，但所作的结论可能互不相同。一个拟合的模型常常是一个近似，这或者是由于未包括重要的变量，或者是重要变量测量错误，或者是使用的函数形式不完全正确。

3.1 解释参数估计值

参数存在吗？模型 $Y = X\beta + e$ 常常是有用的。选择这一模型并不完全根据任何联系自变量与响应变量的理论，更主要的原因是拟合方便。作为结果， β 可能并不是对一个实在的量的估计。如果数据的收集尽管是对相同的变量，但是在不同范围内，或是按不同的抽样方案，计算得到的 β 可能“估计”另一个量。作为刻划一个过程的未知常数的参数的估计，只有在假设的模型的函数形式基本正确的情况下才有意义。

例 3.1 估计矩形的面积

假设我们有 n 个矩形的一个样本，想由此建立一个模型，其中 $\ln(\text{面积})$ 为 $\ln(\text{长度})$ 的某个函数，或许可以通过简单回归方程

$$\ln(\text{面积}) = \beta_0 + \beta_1 \ln(\text{长度}) + e$$

建立模型，对任意矩形，单单一个 $\ln(\text{长度})$ 不是 $\ln(\text{面积})$ 的一个好的自

变量,因为面积需要由长度和宽度同时确定。但对于一类特殊的矩形总体,只以 $\ln(\text{长度})$ 为自变量建立模型可能效果很好,如果矩形为表的顶端,每个的宽度约为 $0.5 \times \text{长度}$, 则 $\ln(\text{面积}) \cong \ln(\text{长度}) + \ln(0.5 \times \text{长度}) = \ln(0.5) + 2\ln(\text{长度})$ 。线性模型似乎拟合得相当好,该模型截距约为 $\ln(0.5) \cong -0.07$, 斜率约为 2。斜率及截距的值几乎与矩形性质无关,而是取决于抽样的总体。如果对不同的矩形总体抽样,则得到不同的参数。

解释估计值 在例 2.1 的燃料消耗数据中,拟合的模型为

$$\begin{aligned} \widehat{FUEL} = & 377.29 - 34.79TAX + 13.36DLIC \\ & - 66.59INC - 2.43ROAD \end{aligned}$$

通常把估计的系数解释为变化率:把 TAX 率增加 1 分,将降低燃料消耗。如果所有其它因素保持不变,人均燃料消耗下降 34.79 加仑。所有这些是在假设一个自变量可以变化一个单位而不影响其它自变量,并且用现有的数据拟合的模型在自变量如此变化之后仍适用的前提下的,在例子中,数据是观测的,因为自变量的赋值不在分析员控制之下,故无法直接验证如果提高税收,燃料消耗是否会降低。另一方面,我们可以给出一个更保守的解释:据观测,税率较高的州燃料消耗较低。为了作出关于改变税率的影响的结论,必须确实地改变税率,并且已观测到了结果。

估计值的符号 参数估计值的符号指出了自变量与响应变量的关系的方向。在多元回归中,如果自变量是相关的,系数的符号可能会根据模型中其它变量而改变。尽管这在数学上是可能的,并且有时在科学上是合理的,它确实使得解释更为困难。有时可以通过重新将自变量定义为新的较易解释的线性组合来解决这一问题。

例 3.2 伯克来指导研究

来自伯克来指导研究的关于男孩与女孩生长情况的数据在习题 2.1 中给出。在研究这些数据时,假设我们希望对 $n = 32$ 个女孩建立体型 ($SOAM$) 关于在 2 岁, 9 岁和 18 岁时体重 (WT_2, WT_9, WT_{18}) 的回归模型。这四个变量的相关矩阵在表 3.1 给出。正如我们预料的,所有的变量是

正相关的。然而，表 3.2 给出的 $SOMA$ 关于 WT_2 , WT_9 和 WT_{18} 的回归导致了意外的结论：2 岁时越重的女孩在 18 岁时越瘦小（有较小的体型）。这一结果可能是由于自变量间的相关关系。考虑下述变量，以代替以前所用的变量：

WT_2 = 2 岁时的体重

$DW_9 = WT_9 - WT_2$ = 从 2 岁到 9 岁增加的体重

$DW_{18} = WT_{18} - WT_9$ = 从 9 岁到 18 岁增加的体重

由于这三个变量测的都是体重，所以这样对它们组合是合理的。如果变量测量的是不同的量，则对它们进行组合所得到的结论不如原先得到的结论有用。对 $SOMA$ 关于 WT_2 , DW_9 和 DW_{18} 拟合的回归在表 3.3 给出。

表 3.1 伯克来指导研究 女孩体重变量的相关矩阵

WT_2	1.000			
WT_9	.5969	1.0000		
WT_{18}	.3508	.7108	1.0000	
$SOMA$.1234	.6508	.6865	1.0000
	WT_2	WT_9	WT_{18}	$SOMA$

表 3.2 $SOMA$ 关于 WT_2 , WT_9 和 WT_{18} 的回归

变量	系数	标准误	t-值
Intercept	2.03686	1.08218	1.88
WT_2	-.217635	.088180	-2.47
WT_9	.094583	.031871	2.97
WT_{18}	.043269	.018394	2.35
$\hat{\sigma}^2 = 0.332625$, d. f. = 28, $R^2 = 0.610$			

将这一回归与用体重本身拟合的回归进行比较。估计 $\hat{\beta}_0$, $se(\hat{\beta}_0)$, $\hat{\sigma}^2$ 和 R^2 是相同的。事实上，因为三个变量 WT_2 , DW_9 和 DW_{18} 可以由 WT_2 , WT_9 和 WT_{18} 通过线性变换得到，这两组变量包含关于 $SOMA$ 的完全相同的信息。然而 WT_2 的估计的系数依赖于选择哪一组变量。表 3.2 给出 $\hat{\beta}_{WT_2} = -0.22$, $t = -2.47$ ，而表 3.3 中， $\hat{\beta}_{WT_2} = -0.08$, $t = -1.05$ 。在前一种情况

中, WT_2 的作用似乎是重要的。而在后一种情况下则不是。尽管 β_{WT_2} 都为负值, 在后一种情况中我们可以认为 WT_2 的作用是可以忽略的。由此可见, 对一个变量的作用的解释不仅依赖于模型中的其它变量, 还依赖于对这些变量采用什么线性变换。

表 3.3 SOMA 关于 WT_2 、 DW_9 和 DW_{18} 的回归

变量	系数	标准误	t-值
Intercept	2.03686	1.08218	1.88
WT_2	-.079782	.076185	-1.05
WT_9	.137853	.023908	5.77
WT_{18}	.043269	.018394	2.35
$\hat{\sigma}^2 = 0.332625$, d.f. = 28, $R^2 = 0.610$			

以上使用的线性变换不是唯一的, 并且根据情况可能其它的某些变换更好。例如, 另一组变换可能为

$$AVE = (WT_2 + WT_9 + WT_{18}) / 3$$

$$LIN = WT_{18} - WT_2$$

$$QUAD = WT_2 - 2WT_9 + WT_{18}$$

这个变换基于下列事实: WT_2 , WT_9 和 WT_{18} 是按时间排序的, 并且几乎是等间隔的, 假设认为体重测量是等间隔的, AVE , LIN 和 $QUAD$ 分别为体重增加的平均、线性和二次趋势。

秩不足以及参数过多的模型 在上一个例子中, 研究了基本自变量 WT_2 , WT_9 和 WT_{18} 的若干组合。有人很自然地会问: 如果在同一回归模型中使用这些自变量的多于三个的组合, 会发生什么情况? 只要我们使用这些自变量的线性组合, 我们不能使用多于三个, 即被测量的线性无关的量的个数。

为说明这一点, 考虑在 SOMA 关于 WT_2 , DW_9 , DW_{18} 的模型中加入 $QUAD$ 。如在第二章中一样, 我们可以通过研究 SOMA 关于 WT_2 、 DW_9 、 DW_{18} 的回归的残差与 $QUAD$ 关于 WT_2 、 DW_9 、 DW_{18} 的回归的残差来检查这一问题。但因为 $QUAD$ 可以写成其它自变量的线性组合, $QUAD = DW_{18} - DW_9$, 由第二个回归得到

的残差全部恰好为零。这样,被其它三个自变量调整后, $QUAD$ 的斜率系数是不确定的, 我们说四个自变量 WT_2 , DW_9 , DW_{18} 和 $QUAD$ 是线性相关的, 因为任何一个可以由其它三个决定, 可以包含在一个模型中的自变量的最大个数称为模型或数据矩阵 X 的秩。本书中大部分模型是满秩的, 所有自变量线性独立, 然而常常会使用退化模型或自变量线性相关的模型。最简单的例子是单向设计。假设一个单元被分配给三个处理组中的一个, 如果单元在第一组, 则令 $X_1=1$, 其它等于零, 如果单元在第二组, 令 $X_2=1$, 其它为零, 如果单元在第三组, 则令 $X_3=1$, 其它为零。这样, 因为每个单元只在三组中的一组, 所以有 $X_1+X_2+X_3=1$ 。因此, 我们不能拟合模型

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

因为有这些 X 的和等于 1 的那一列, 所以模型为退化的。为拟合一个模型, 我们必须做些其它工作。供选择的有: (1) 对参数加一个限制条件如 $\beta_1 + \beta_2 + \beta_3 = 0$; (2) 从模型中去掉一个 X ; 或者 (3) 去掉截距, 强行使回归过原点。这些选项在某些意义下是等价的, 因为所得到的 R^2 、 $\hat{\sigma}^2$ 、整个的 F -检验统计量的值和预测值是相同的。当然, 在使用参数估计时必须当心, 因为这将和得到一个满秩模型参数化有关。关于不满秩的矩阵和模型的进一步阅读资料见 Searle (1971, 1982)。

怎样好才是最好? (Ehrenberg, 1982) 从残差平方和函数最小化的意义上来说, 最小二乘估计给出了最优拟合直线。然而, 这一最优直线比其它直线好多少呢? 正如我们将看到的, 对许多估计量, 残差平方和函数的值与最小二乘估计给出的非常接近。

只考虑具有最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 残差平方和为 RSS 的简单回归。作为其它估计, 考虑对每个 k 通过

$$\beta_1 = k\hat{\beta}_1 \quad \text{和} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

定义的 β_0 和 β_1 。当 k 变化时, 这些估计给出通过点 (\bar{x}, \bar{y}) , 但斜率任意的拟合直线。这些估计的残差平方和函数 $RSS(\beta_0, \beta_1)$

为

$$\begin{aligned}RSS(\tilde{\beta}_0, \tilde{\beta}_1) &= \sum (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \\&= \sum [(y_i - \bar{y}) - \tilde{\beta}_1(x_i - \bar{x}) - (\tilde{\beta}_0 - \beta_0)(x_i - \bar{x})]^2 \\&= RSS + (\tilde{\beta}_1 - \beta_1)^2 SXX\end{aligned}$$

由于 $RSS = SYY(1 - r^2)$, 其中 r 为 x 与 y 的样本相关系数, 以及 $\beta_1 = SXY/SXX$,

$$RSS(\tilde{\beta}_0, \tilde{\beta}_1) = SYY[(1 - r^2) + r^2(1 - k)^2] \quad (3.1)$$

如果在 (3.1) 两边除以 RSS 并取平方根, 我们得到使用其它估计的回归的标准误与使用最小二乘估计的回归的标准误的比率,

$$\left[\frac{RSS(\tilde{\beta}_0, \tilde{\beta}_1)}{RSS} \right]^{1/2} = \left[1 + \frac{r^2}{1 - r^2} (1 - k)^2 \right]^{1/2}$$

这个比率至少为 1, 因为最小二乘估计使残差平方和函数最小。但 k 能有多大并且多大的 k 能使回归的标准误在其最小值的 5% 以内? 我们令上一表达式的右边为 1.05 并且解出 k , 就能得到答案。我们发现

$$k \leq 1 \pm 0.32 \left[\frac{1 - r^2}{r^2} \right]^{1/2}$$

图 3.1 中阴影面积内的任何 k 值满足这个界。例如, 如果 $r^2 = 0.49$ 在 $0.67\beta_1$ 至 $1.33\beta_1$ 范围内的任何 $\tilde{\beta}_1$ 给出在最小二乘估计的标准误 5% 以内的回归的标准误的估计。如果 r^2 不大, 最小二乘估计不会比整个范围内其它可能的估计好得多。

检验 即使拟合模型是正确的, 误差服从正态分布, 由于数据的非正交性导致可能的检验的多样性, 对检验以及参数的置信断言作出解释是困难的。有时对受其它变量调整的效应的检验是明显需要的, 例如在受其它变量调整使变异性减少后, 对处理效应的评估。在另一些时候, 拟合的次序不明显, 分析者预料将得到模糊的结论。

如果所拟合的模型依赖于数据, 情况更为复杂, 因为检验一

个假想的参数为零未必等价于检验在回归中一个变量的显著性。不过，即使严格的概率解释是不可靠的，通常的检验统计量仍给出了有效的指导。例如，表 3.3 中 WT_2 的 t -值指出 WT_2 对回归没有什么影响，如果模型中没有那个变量，也能得到本质上相同的结果。

以上所述指出使用接受/拒绝规则来确定自变量的显著性是没有用的。在大多数情况下，对显著性唯一真正的检验是重复试验。

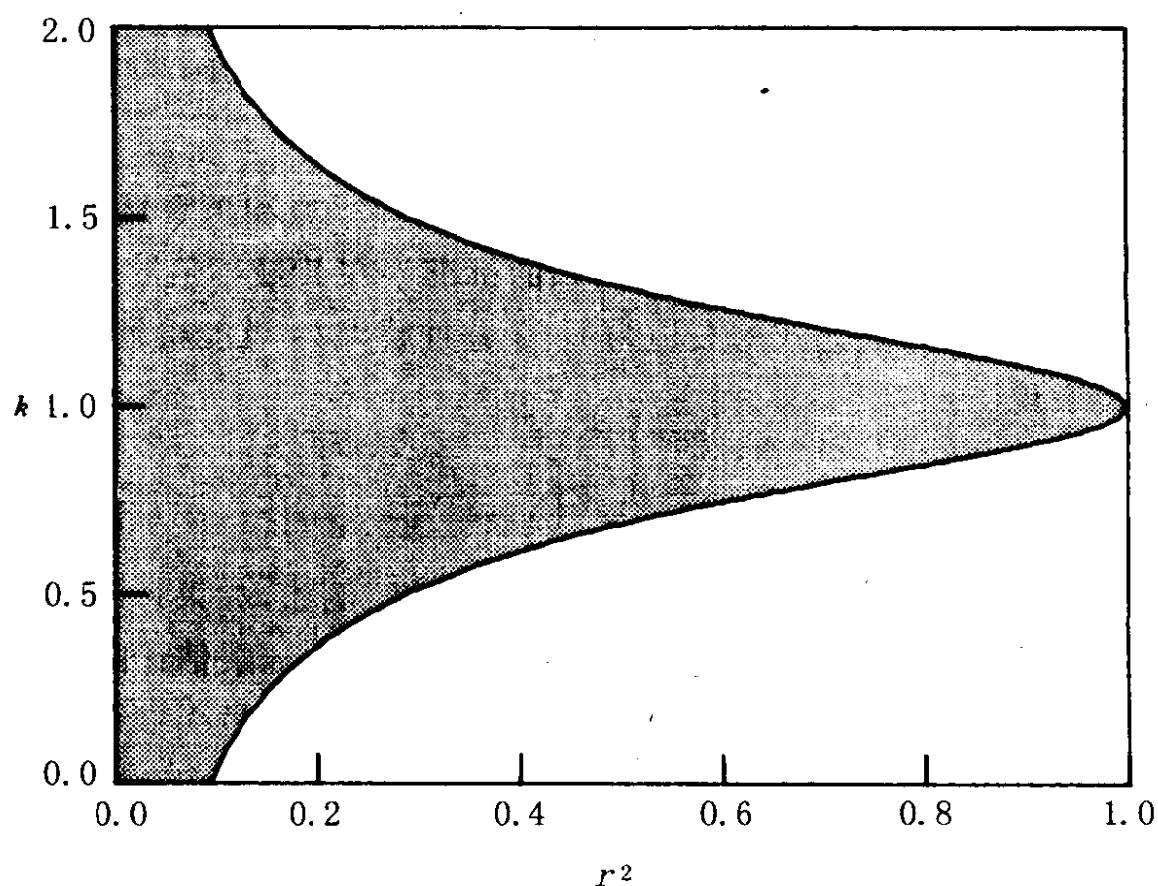


图 3.1 阴影区域给出使 $[RSS(\beta_0, \beta_1) / RSS]^{1/2} \leq 1.05$ 的 k 值

3.2 抽样模型

试验与观测 在回归分析中主要有两种类型的自变量：试验

型与观测型的。前一种类型，自变量的值是在试验者控制之下的，而后一种类型，自变量的值是观测的，不是设置的。例如，考虑一种决定某种谷物产量的因素的假设的研究，试验型变量可能包括使用的化肥的类型的数量、作物的间距以及灌溉量，因为它们中的每一个都可以由试验者按单位即田块指定。观测型自变量可能包括研究中的田块的特征，例如降雨、日照、土壤肥沃程度及气候变化，所有这些不受试验者控制，然而可能对观测的产量造成重要的影响。

某些试验设计，包括使用随机化设计，使得观测因素的影响可以被忽略或用于协方差分析（例如，见Cox, 1958）。用来自被设计的试验的数据，拟合的模型可以得出关于试验因素的影响的有用的结果，这些模型可以用来预测响应变量的未来值。

在另一个极端，不在分析员控制下的纯观测型研究可能只能用来预测，或对在数据中被观测的事件建立模型，就象在燃料消耗的例子中的那样。为了运用观测结果预测未来的值，必须作出关于未来值的状态与现有数据的状态相比较的假设。

从正态总体中抽样 使用最小二乘估计的直觉，大多是基于观测数据是多元正态分布的一个样本的假设，尽管在实际回归问题中，多元正态性的假设几乎总是站不住的，但是探索正态数据的有关结果还是值得的，先假设（从正态总体中）随机抽样，然后移去这一假设。

例 3.3 多元正态性

假设所有观测变量为正态随机变量，每个案例的观测与其它案例的观测是独立的。在一个二元问题中，对第 i 个案例观测到 (x_i, y_i) ，并假设

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right) \quad (i = 1, 2, \dots, n) \quad (3.2)$$

(3.2) 式说明 x_i 和 y_i 是具有均值 μ_X 和 μ_Y ，方差 σ_X^2 和 σ_Y^2 及相关系数 ρ_{XY} 的正态随机变量的实现。现在，假设我们已有观测值 x_i ，考虑 y_i 的条件分布。可以证明（例如见Lindgren, 1976）给定 x_i 时 y_i 的条件分布，记作 $y_i | x_i$ ，是正态的，并且

$$y_i|x_i \sim N(\mu_y + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(x_i - \mu_X), \sigma_Y^2(1 - \rho_{XY}^2)) \quad (i = 1, 2, \dots, n) \quad (3.3)$$

如果我们定义

$$\beta_0 = \mu_Y - \beta_1 \mu_X; \quad \beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}; \quad \sigma^2 = \sigma_Y^2(1 - \rho_{XY}^2) \quad (3.4)$$

则给定 x_i 时 y_i 的条件分布简化为

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad (i = 1, 2, \dots, n) \quad (3.5)$$

这与简单回归模型本质上是相同的。

对已给的随机样本，利用表 1.2 的记号，(3.2) 中五个参数估计为

$$\hat{\mu}_X = \bar{x}, \quad \hat{\sigma}_X^2 = SD_X^2, \quad \hat{\rho}_{XY} = r_{XY} \\ \hat{\mu}_Y = \bar{y}, \quad \hat{\sigma}_Y^2 = SD_Y^2 \quad (3.6)$$

用 (3.6) 得到的估计代替 (3.4) 中的参数，以致 $\hat{\beta}_1 = r_{XY} SD_Y / SD_X$ 等等，就象在第 1 章导出的那样，可以得到 β_0 和 β_1 的估计。（不过， $\hat{\sigma}^2 = [(n-1)/(n-2)] SD_Y^2 (1 - r_{XY}^2)$ 以校正自由度。）

如果对第 i 个案例的观测是 y_i 和一个 $p \times 1$ 向量 x_i ， x_i 不包括常数，多元正态性用符号表示为

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{bmatrix} \sigma_y^2 & \sum_{XY}^T \\ \sum_{XY} & \sum_{XX} \end{bmatrix} \right]$$

其中 Σ_{XX} 是诸 X 间的一个 $p \times p$ 的方差协方差矩阵， Σ_{XY} 是诸 X 与 Y 的一个 $p \times 1$ 的协方差向量。则给定 x_i 时 y_i 的条件分布是

$$y_i|x_i \sim N((\mu_Y - \beta^{*T} \mu_X) + \beta^{*T} X_i, \sigma^2) \quad (3.7)$$

并且，如果 \mathcal{R}^2 为总体复相关系数，

$$\beta^* = \Sigma_{XX}^{-1} \Sigma_{XY}; \quad \sigma^2 = \sigma_Y^2 - \Sigma_{XY}^T \Sigma_{XX}^{-1} \Sigma_{XY} = \sigma^2(1 - \mathcal{R}^2)$$

β^* 和 σ^2 的公式与它们的最小二乘估计公式的不同仅仅是以估计代替参数，用 $n^{-1}(\mathcal{X}^T \mathcal{X})$ 估计 Σ_{XX} 用 $n^{-1}(\mathcal{X}^T \mathcal{Y})$ 估计 Σ_{XY} 。

例 3.4 总体的非随机抽样（或如何得到更大的 R^2 ）

读者可能已经注意到 (3.3) 或 (3.7) 中的条件分布不依赖于随机抽样，只依赖于正态分布。因此，只要对于变量，多元正态性是一个合理的模型，给定其它变量时，一个变量的条件分布可使用线性回归模型。然而，如果不使用随机抽样，某些常用的主要统计量，包括 R^2 ，就失去意义了。这可由人为制造的数据来说明。

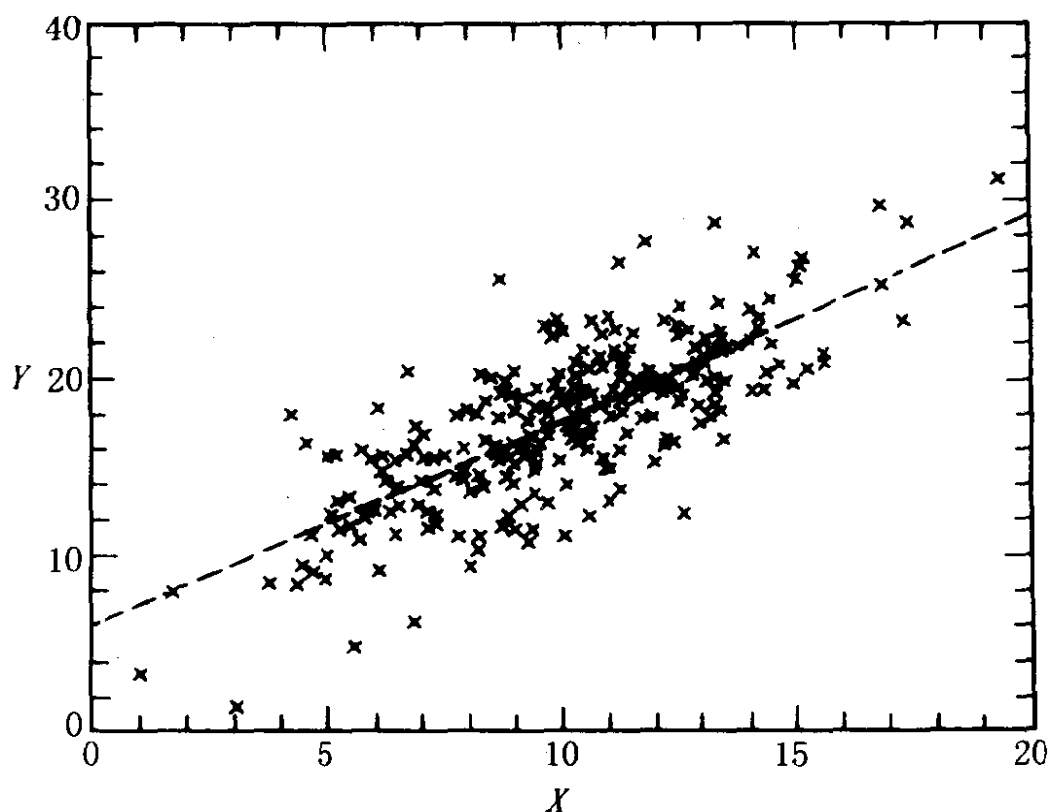


图 3.2 一个二元正态样本, $n=250$

图 3.2 给出计算机生成的 $n=250$ 对 (x_i, y_i) 的二元伪随机样本, 使每对 (x_i, y_i) 似乎都是独立地取自

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N \left(\begin{pmatrix} 10.0 \\ 17.5 \end{pmatrix}, \begin{pmatrix} 9.0000 & 11.2510 \\ 11.2510 & 23.0625 \end{pmatrix} \right) \quad (i=1, 2, \dots, 250) \quad (3.8)$$

由 (3.3), $y_i | x_i$ 的条件分布为

$$y_i | x_i \sim N(5 + 1.25x_i, 9) \quad (i=1, 2, \dots, 250)$$

在图中, 点群通常为椭圆的, 正如正态数据的特征那样。拟合的回归直线在表 3.4 给出, 作在图 3.2 中, 它与真实直线 $y=5+1.25x$ 相差不远。计算得到的 $R^2=0.584$, 接近于真实的 $\rho_{xy}^2=0.610$ 。因为采用了随机抽样, 所有常用的主要统计量和检验都是有用的。

现在, 考虑同一个点集, 但根据它们的值 x_i 选择 n 个案例。令 $S_x^2 = \sum (x_i - \bar{x})^2 / n$ 为实际在样本中的 X 的值的方差; 因为 X 的值为选择的, 而不是抽样的, S_x^2 不是 σ_x^2 的估计。试验者常常可以选择案例, 使 S_x^2 取任何值。在设计的试验中, S_x^2 通常尽可能地大。图 3.3 和 3.4 中图解地说明了两种选

择。这两个图由图 3.2 得到。在图 3.3 中, 只选择了 $x_i \leq 7$ 或 $x_i \geq 13$ 的案例, 而在图 3.4 中 X 的范围限制于 $7 < x_i < 13$ 。这样, 在图 3.3 中, S_x^2 为大的, 而在图 3.4 中, S_x^2 为小的。对这三种数据集 (图 3.2 至图 3.4) 拟合的方程几乎是相同的, $\hat{\sigma}^2$ 基本上为常数, 如表 3.4 给出的。然而, 注意 R^2 有很大的变化。在图 3.3 中, $S_x^2 > \sigma_x^2$, $R^2 = 0.762$ 太大了, 而在图 3.4 中, $S_x^2 < \sigma_x^2$, $R^2 = 0.279$ 太小了。尽管这三个图得到几乎相同的拟合方程, 拟合常用的主要统计量— R^2 —导致很不相同的结论。

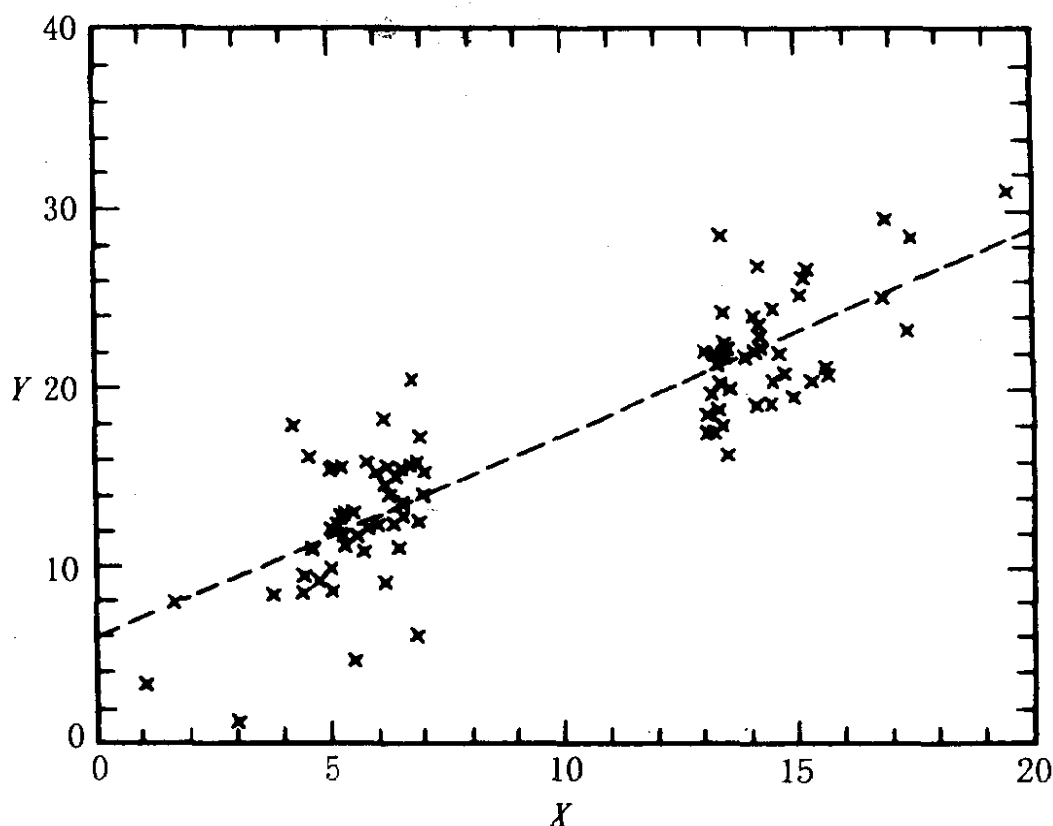


图 3.3 $x_i \leq 7$ 或 $x_i \geq 13$ 的样本

这个原因在三个图中是明显的。在图 3.4 中, 由于 x 的范围狭小, 忽略 X 和 Y 的变化并不比给定 X 后 Y 的变化大很多。因此回归似乎不说明什么—— R^2 是小的。将类似的论证应用于图 3.3, 可指出 R^2 是大的。

这个例子指出即使在分析的数据取自一个多元正态总体的不常见的情况下, 如果从总体中的抽样不是随机的, 对主要统计量如 R^2 的解释可能是完全令人误解的, 因为这个统计量很大程度上受抽样方法的影响。特别, 几个具有不寻常的自变量的值的案例会基本上确定这个统计量的观测值。

表 3.4 回归分析简要

	真值	估 计 值		
		整个样本	$x \leq 7$ 或 $x \geq 13$	$7 < x < 13$
n	∞	250	90	160
β_0	5.000	5.816	5.829	5.376
β_1	1.250	1.167	1.157	1.216
σ^2	9.000	9.010	9.166	9.024
R^2	0.610	0.584	0.762	0.279

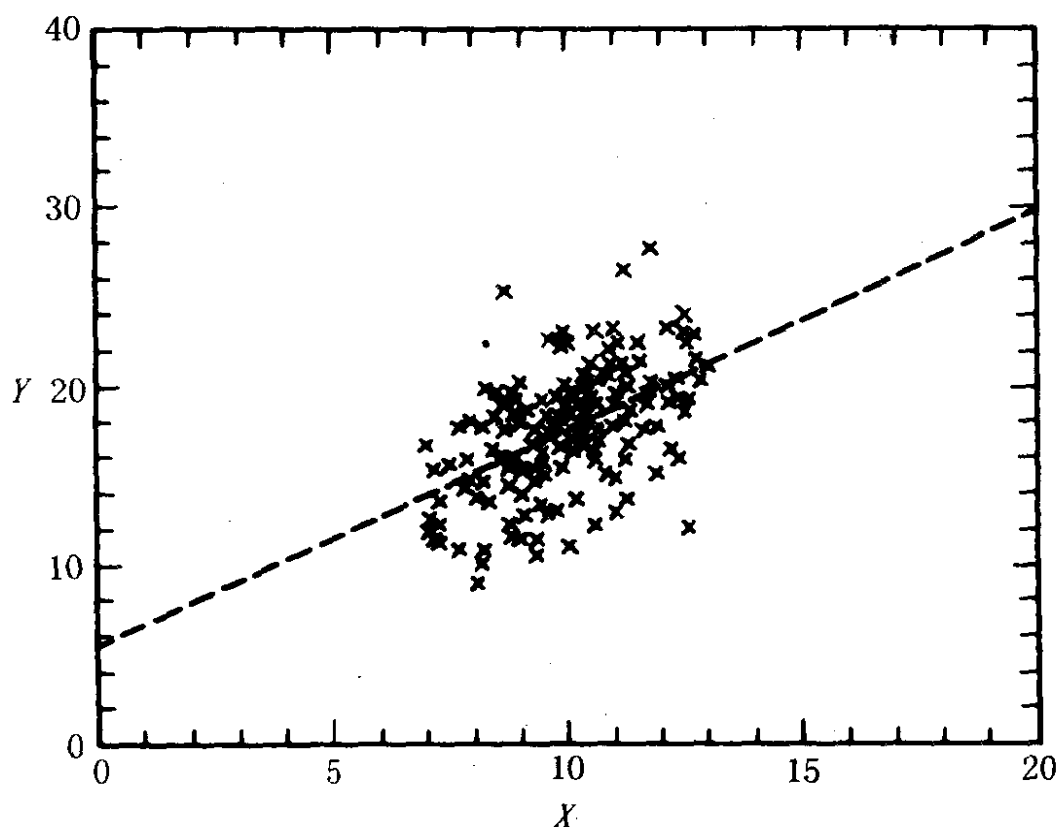


图 3.4 $7 \leq x_i \leq 13$ 的样本

3.3 含测量误差的自变量

最小二乘估计量的合乎需要的性质依赖于假设。本书的一个主要的论题是研究当假设不成立时发生的问题。一个假设是自变

量为没有测量误差的确定的值。当在自变量和响应变量中确实出现误差时，通常的最小二乘准则可能没有什么意义，因为只有响应变量的误差用于决定估计量。允许自变量中含有误差的假设要复杂得多，没有统一的方法可循。

假设 X 为自变量的观测矩阵，其元素可能有测量误差。应该观测得到的“真值”由 \tilde{X} 给出，其中

$$X = \tilde{X} + D$$

D 是误差的 $n \times p'$ 矩阵；通常 D 的第一列都为零，因为 X 的由“1”组成的列是已知无误差的， X 的任何其它已知精确的列在 D 中的对应列也全为零。 D 的其余元素可能表示舍入误差或测量误差，当使用计算机时，舍入误差总是会出现的。我们这里假设 D 的任何行 d_i^T 表示一个案例的误差，它与其它任何案例的误差是独立的。我们假设 $E(d_i) = 0$ ， $\text{var}(d_i)$ 为一个 $p' \times p'$ 的对角矩阵 S ，

$$S = \begin{pmatrix} 0 & & & 0 \\ & s_1^2 & & \\ & & s_2^2 & \\ & & & \ddots \\ 0 & & & & s_p^2 \end{pmatrix}$$

左上角的零与截距的指示变量无误差的假设相对应。

首先假设感兴趣的问题是线性模型

$$Y = \tilde{X}\beta + e \quad (3.9)$$

但 β 的估计必须基于含误差的模型

$$Y = X\beta + e = (\tilde{X} + D)\beta + e \quad (3.10)$$

拟合 (3.10) 的估计量 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 不一定是模型 (3.9) 中 β 的合理的估计量。Hodges 和 Moore (1972) 证明了

$$E(\hat{\beta}) - \beta = - (n - p') (X^T X)^{-1} S \beta \quad (3.11)$$

于是拟合 (3.10) 的估计量 $\hat{\beta}$ 给出了 (3.9) 中的 β 的有偏估计。偏倚可能是大的，或正或负，并且当样本容量增大时不会消失。对简单回归，斜率系数的期望是

$$E(\hat{\beta}_1 | \text{简单回归}) = \beta_1 \left[1 - \frac{s_1^2}{SXX/(n-2)} \right]$$

平均而言, 估计的斜率可能会太小, 这依赖于测量误差 s_1^2 与 $SXX/(n-2)$ 之比。在出现测量误差的情况下, 如果对 (3.9) 感兴趣, 最小二乘回归可能非常使人误解。

幸运的是, 我们并不总是对 (3.9) 感兴趣。Berkson (1950) 考虑了预测这个很重要的问题。如果自变量含有测量误差, 则现在及将来, 我们对拟合 (3.10) 而不是 (3.9) 感兴趣, 因为未来的自变量也不是真值, 而是有测量误差的值。因此, 如果感兴趣的关系是基于观测值而不是非观测的真值, 测量误差在问题中可能并不重要。

在参数估计问题中, 响应变量与非观测的真值之间的关系是令人感兴趣的, 我们处于先前描述的拟合 (3.10) 的情况, 但是却想拟合 (3.9)。对这一更为复杂的问题的合理的探讨, 需要进一步的假设及信息。Madansky (1959) 给出二元问题的一个明确的处理方法。更近期的叙述与书目由 Anderson (1976) 给出。

解决变量误差问题的另一个途径是寻求诊断统计量以确定拟合 (3.10) 是否与拟合 (3.9) 有很大的不同。这些诊断量一般利用数值分析的方法得到; 例如, 我们可以在 $D=0$ 对 β 以泰勒级数展开, 并利用高阶项的大小来判断测量误差的影响; 例如参见 Hodges 和 Moore (1972); Daries 和 Hutton (1975); 以及 Beaton, Rubin 和 Barone (1976)。这些诊断量与第 8 章讨论的自变量的共线性问题是紧密相关的; 目前尚未提出确定测量误差是否可以被忽略的完全合适的方法。

问 题

- 3.1 对伯克来指导研究中的女孩组拟合 *SOMA* 关于 3.1 节中定义的 *AVE*, *LLN* 和 *QOAD* 的回归, 并与 3.1 节的结论比较。

3.2 对二元正态分布,求 $x_i|y_i$ 的条件分布由此你会发现 y 关于 x 的回归的回归直线与 x 关于 y 的回归的回归直线是不同的。在什么条件下两者相同? 对在 (3.8) 给出的二元正态分布,求 $x_i|y_i$ 的条件分布,并在同一图上作两条回归直线 (x 关于 y 及 y 关于 x)。

3.3 表 3.5 中的数据首先由 Longley (1967) 给出,以证明目前使用的回归计算程序的不充分性。七个变量为:

X_1 =GNP 价格紧缩,按百分比计

X_2 =GNP,按百万美元计

X_3 =失业人数,按千计

X_4 =军队大小,按千计

X_5 =14 岁以上非特殊人口,按千计

X_6 =年

Y =得到的总的就业人数,按千计

3.3.1 拟合 Y 关于所有 X 的回归。

3.3.2 这些数据中的自变量有明显的测量误差。对各个自变量,求测量误差的估计 $s_1^2, s_2^2, \dots, s_6^2$ 。(提示:数据本身几乎不包含关于测量误差的信息。一种方法是给出 s_j^2 的下界,假设每个数的最后一位数字有舍入误差。这指出 1947 年 X_2 的“真值”为 234, 288.5 至 234, 289.5 之间的任何数。如果我们假设所有可能的结果等概率地等于真值,则舍入误差在区间 $(-0.5, +0.5)$ 上均匀分布。利用在区间 (a, b) 上均匀分布的随机变量具有方差 $(b-a)^2/12$ 这一事实,我们可以估计 $s_2^2 = [0.5 - (-0.5)]^2/12 = 1/12$ 。这一方法同样适用于除了年以外的其它自变量。这些估计的 s_j^2 可能太小,但它们提供了一个起点。年的舍入误差可对应于一年的长度的变化,闰年要长些,或对应于其它未完全按年分计的量。

3.3.3 为检查误差对本题上节中估计的量的影响,进行如下模拟试验:
(1) 从 Longley 数据开始。(2) 在计算机上,生成在你认为对舍入误差是合理的区间上均匀的随机数。例如,对 X_2 ,你可以生成 $(-0.5, 0.5)$ 上的随机变量。(3) 将随机数加入自变量;不要改变 Y 。(4) 计算 Y 关于修改后的诸 X 的回归,记录系数的估计。(5) 重复步骤 2 至 4 若干次,如 100 次,并总结结论。这一模拟对本问题中系数的估计的舍入误差的影响,给出一个很

好的理解。

表 3.5 Longley 的数据

X_1	X_2	X_3	X_4	X_5	X_6	Y
83.0	234289.	2356.	1590.	107608.	1947.	60323.
88.5	259426.	2325.	1456.	108632.	1948.	61122.
88.2	258054.	3682.	1616.	109773.	1949.	60171.
89.5	284599.	3351.	1650.	110929.	1950.	61187.
96.2	328975.	2099.	3099.	112075.	1951.	63221.
98.1	346999.	1932.	3594.	113270.	1952.	63639.
99.0	365385.	1870.	3547.	115094.	1953.	64989.
100.0	363112.	3578.	3350.	116219.	1954.	63761.
101.2	397469.	2904.	3048.	117388.	1955.	66019.
104.6	419180.	2822.	2857.	118734.	1956.	67857.
108.4	442769.	2936.	2798.	120445.	1957.	68169.
110.8	444546.	4681.	2637.	121950.	1958.	66513.
112.6	482704.	3813.	2552.	123366.	1959.	68655.
114.2	502601.	3931.	2514.	125368.	1960.	69564.
115.7	518173.	4806.	2572.	127852.	1961.	69331.
116.9	554894.	4007.	2827.	130081.	1962.	70551.

4

加权最小二乘法， 对拟合失真的检验， 广义 F -检验及置信椭圆

本章从讨论使用误差方差及协方差的附加信息开始。这些信息有时被用于得到广义最小二乘估计，而不是前面讨论过的普通最小二乘估计。另外，这些信息还可用于观测数据，以检验模型的拟合失真的情况。然后，我们将富有启发性地讨论，当误差服从正态分布时，得到服从 F -分布的检验统计量的一般方法。这些检验用于回归的很多场合，包括前面第一、二章所描述的，以及将在以后章节中描述的那些场合。最后，讨论多于一个参数的作为联合置信域的椭圆。这些依赖于 F -分布的区域，在评估案例对回归估计的影响时将是有用的，这将在第五章中讨论。

4.1 广义加权最小二乘法

在前面的章节中，假设误差的方差是未知的，相等的，并且误差是相互独立的。这些假设通常是不必要的，因为方差的这些明确具体的信息是罕见的。然而，在不少问题中，关于误差方差

的附加信息是可以得到的。这时方差或者是已知的，或者知道是某个数的常数倍数。合并这些信息用于分析的方法并不困难。这将在本节中讨论。

广义最小二乘法 假设我们已知一个正定对称矩阵 Σ 的值，其误差向量 e 的协方差矩阵由 $\text{var}(e) = \sigma^2 \Sigma$ 给出，其中 $\sigma^2 > 0$ ，但不一定是已知的。我们可能合理地推测：在这种情况下， β 的普通的最小二乘估计，尽管仍是无偏的，但不再是最小方差估计，因为它忽略了某些明显有用的信息。严格地说，考虑模型

$$Y = X\beta + e \quad X: n \times p', \text{ 秩为 } p' \quad (4.1a)$$

$$\text{var}(e) = \sigma^2 \Sigma \quad \Sigma \text{ 已知, } \sigma^2 > 0 \text{ 不一定已知} \quad (4.1b)$$

我们继续使用符号 $\hat{\beta}$ 表示对 β 的估计，尽管这一估计是通过广义的、而非普通的最小二乘法得到的。一旦确定了 $\hat{\beta}$ ，残差 \hat{e} 由式子 $\hat{e} = Y - \hat{Y} = Y - X\hat{\beta}$ 给出。估计值 $\hat{\beta}$ 的选取是使下面的广义残差平方和函数取最小值

$$RSS(\beta) = (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) \quad (4.2)$$

粗略地说，使用广义残差平方和是承认某些残差，或拟合误差，比其它的一些更为重要。特别地，和有较大方差的误差相对应的残差，在计算广义残差平方和时是不太重要的。广义最小二乘估计由下式给出

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y \quad (4.3)$$

尽管可以直接求解上述方程，但更方便地，我们可以将 (4.1) 描述的模型转化为可以用普通最小二乘法解决的模型。然后，所有普通最小二乘法的结论可用于广义最小二乘法中。

模型 (4.1) 与普通最小二乘模型的区别仅在于 $\text{var}(e) = \sigma^2 \Sigma$ 。现在假设我们可以得到一个 $n \times n$ 的矩阵 C ，使 C 是对称的，并且 $C^T C = C C^T = \Sigma^{-1}$ (从而 $C^{-1} C^{-T} = \Sigma$)。这样的矩阵 C 称为 Σ^{-1} 的平方根。由附录 2A.2，随机向量 Ce 的方差—协方差矩阵由下式给出：

$$\begin{aligned}
\text{var}(Ce) &= C(\sigma^2 \Sigma)C^T \\
&= \sigma^2 C(C^{-1}C^{-T})C^T \\
&= \sigma^2 CC^{-1}C^{-T}C^T \\
&= \sigma^2 I_n
\end{aligned} \tag{4.4}$$

将方程 (4.1a) 的两边乘以 C , 得

$$CY = CX\beta + Ce \tag{4.5}$$

现在定义 $Z = CY$, $M = CX$ 及 $d = Ce$ 。等式 (4.5) 变为

$$Z = M\beta + d \tag{4.6}$$

这里, 由 (4.4), $\text{var}(d) = \sigma^2 I_n$, 并且 (4.6) 中的 β 与 (4.1) 中的 β 完全一样。模型 (4.6) 可以用普通最小二乘法求解。例如, 由 (2.15), 用 Z 和 M 表示的估计 $\hat{\beta}$ 为

$$\hat{\beta} = (M^T M)^{-1} M^T Z$$

用 X 、 Y 和 C 代替 M 和 Z , 上式变成

$$\begin{aligned}
\hat{\beta} &= [(CX)^T(CX)]^{-1}(CX)^T(CY) \\
&= (X^T C^T CX)^{-1}(X^T C^T CY) \\
&= (X^T \Sigma^{-1} X)^{-1}(X^T \Sigma^{-1} Y)
\end{aligned}$$

即为 (4.3) 给出的估计。

广义最小二乘法的实际步骤为, 首先得到 C , Σ^{-1} 的平方根, 在观测到的数据向量 Y 及矩阵 X 的左边乘以 C , 再用普通最小二乘法解得到的回归问题。在实际求 C 的过程中有些数值上的困难。不过, 在加权最小二乘法的特例中, 从 Σ 计算 C 是简单的。

加权最小二乘法 当误差有不同的方差, 且所有的误差都互不相关时, Σ 是一个对角矩阵, 我们得到一个加权最小二乘问题。方差通常用权 ω_i 表示, $\omega_i > 0$ 且 $\text{var}(e_i) = \sigma^2 / \omega_i$ 。具有较大权的案例有较小的方差, 并且它在回归问题中更为重要。矩阵 Σ 的形式为 $\Sigma = W^{-1}$, 其中

$$W^{-1} = \begin{pmatrix} \frac{1}{w_1} & & & 0 \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_n} \end{pmatrix} \quad (4.7)$$

对这一 $\Sigma (=W^{-1})$, 矩阵 C 很容易求得:

$$C = \begin{pmatrix} \sqrt{w_1} & & & 0 \\ & \sqrt{w_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{w_n} \end{pmatrix} \quad (4.8)$$

并且

$$M = \begin{pmatrix} \sqrt{w_1} & \sqrt{w_1}x_{11} & \cdots & \sqrt{w_1}x_{1p} \\ \sqrt{w_2} & \sqrt{w_2}x_{21} & \cdots & \sqrt{w_2}x_{2p} \\ \vdots & \vdots & & \vdots \\ \sqrt{w_n} & \sqrt{w_n}x_{n1} & \cdots & \sqrt{w_n}x_{np} \end{pmatrix}, \quad Z = \begin{pmatrix} \sqrt{w_1}y_1 \\ \sqrt{w_2}y_2 \\ \vdots \\ \sqrt{w_n}y_n \end{pmatrix} \quad (4.9)$$

它们的列也被乘以案例的权。然后用 M 和 Z 代替 X 和 Y , 可以求解回归问题。(例外: 残差及拟合值仍从 $\hat{e} = Y - \hat{Y}$ 及 $\hat{Y} = X\beta$ 计算, 而不是从含有 M 和 Z 的等价公式计算。)

大多数计算机程序允许在回归中使用案例的权。 w_i 通常表示为一列数据, 使用者仅需说明这一列为权。程序计算 $M^T M$ 及 $M^T Z$ 或它们关于平均值的偏差, 然后类似不加权的最小二乘法进行计算 (见问题 4.8)。

加权最小二乘法的应用 如果第 i 个响应变量是同一变量的 n_i 个观察值的平均, 则 $\text{var}(y_i) = \sigma^2/n_i$, 且 $w_i = n_i$; 如果 y_i 是 n_i 个观察值之和, 则 $\text{var}(y_i) = n_i\sigma^2$, 且 $w_i = 1/n_i$; 如果方差与某个自变

量 x_i 成正比, 则 $\text{var}(y_i) = x_i \sigma^2$, 且 $\omega_i = 1/x_i$ 。

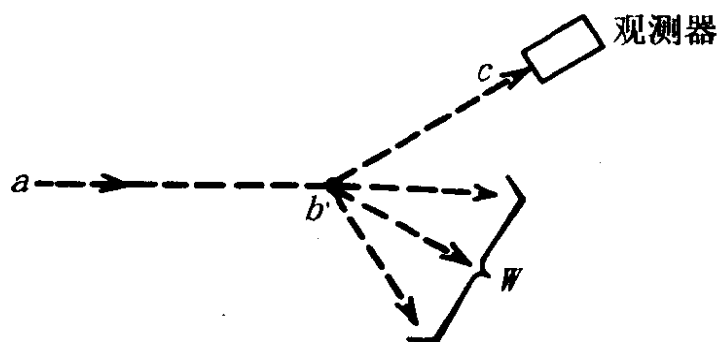
例 4.1 强交互作用

这里叙述的某个试验的目的是研究与质子碰撞中某种基本粒子的相互作用。(Weisberg et al. 1978) 这里研究的粒子包括 π^- 介子及其反粒子 π^+ 。它们是不稳定的, 可以通过高能加速器获得。它们在一组称为强子的粒子组中, 具有与一个电子相同的正的或负的电量。强子通过电磁力相互作用, 使原子聚合在一起。另外, 与电子不同, 它们还通过一种称为强相互作用的力相互作用, 使原子核聚合在一起。尽管现在电磁力已被很好地认识, 强相互作用对物理学家仍多少是个谜, 而这一试验是用于测试关于强相互作用的某种理论。

离散过程可以表述为

$$a + p \longrightarrow c + x \quad (4.10)$$

并且可以用一个图表示:



这个图表示一束 a 粒子瞄准含质子 p 的目标 b 。观测器观测 c 类型的飞出粒子。试验令 a 和 c 分别为各种粒子。这里只考虑 π^- 和 π^+ 。对于一个典型的高能碰撞, a 和 p 都将散裂, 并且通过能量转化为物质的过程, 碎片变成许多强子。试验测量一种特别的 c 型粒子的生成速度, 而不论生成什么其它粒子 (图中标为 W)。

观测的量是散射截面 y (通常表示成 $\Delta\sigma$), 由下式给出

$$y = \frac{N_c}{N_a \rho l} \quad (4.11)$$

其中 N_a 是每秒射入的粒子束数。 ρ 是每单位体积粒子的目标密度, l 是目标长度, N_c 是每秒观测到的 c 粒子数。截面 y 可以度量并能方便地用毫靶 (mb) 计量, $1\text{mb} = 1 \times 10^{-27} \text{cm}^2$ 。

实验中粒子束 a 有各种不同的人射动量值 p_a^{lab} , 以实验室作为参照系统进行计量。一个比 p_a^{lab} 更具有基本理论重要性的量是 s , 它是以质心为参照系

统计量的总能量的平方。由于实验中使用高动量, 粒子 a 几乎以光速进行, 关系

$$s = 2m_p p_a^{\text{lab}} \quad (4.12)$$

是一个很好的近似。 s 的单位是 $(\text{GeV})^2$, 其中 $1\text{GeV} = 1 \times 10^9$ 电子伏, 为一个基本粒子被十亿伏电能加速后达到的能量。动量 p_a^{lab} 及质量 m_p 用 GeV 计量。对一个质子来说, $m_p = 0.938\text{GeV}$ 。

理论物理学家为强相互作用力建立了各种模型, 某些模型预测, 在高能极限 $s \rightarrow \infty$, 截面 y 接近一个常数极限。另外, 这个极限的函数形式被预测为

$$y = \beta_0 + \beta_1 \cdot s^{-1/2} + \text{相当小的项} \quad (4.13)$$

这个理论对 β_0 和 β_1 及它们对 a 和 c 型粒子的依赖性作了量化的预测。从而, 令人感兴趣的是 (1) 对每种可能的 a 和 c , 采用 (4.13) 作为模型, 估计 β_0 和 β_1 ; (2) 评价模型 (4.13) 是否对观测数据给出一个精确描述; 和 (3) 如果 (4.13) 是合适的, 将 β_0 与 β_1 和理论预测值进行比较。

表 4.1 给出的数据综合了 $a=c=\pi^-$ 时的试验结果。在每个 p_a^{lab} , 使用大量的 N_a 粒子, 以使观测值 y 的方差可以精确地从理论思考中获得。这些方差的平方根 (即 $\sigma/\sqrt{w_i}$) 由表 4.1 的第 4 列给出。

表 4.1 物理例子的数据

p_a^{lab} GeV/c	$s^{-1/2}$ GeV/c ⁻¹	y (μb)	$\sigma/\sqrt{w_i} =$ 估计的标准差
4	0.345	367	17
6	0.287	311	9
8	0.251	295	9
10	0.225	268	7
12	0.207	253	7
15	0.186	239	6
20	0.161	220	6
30	0.132	213	6
75	0.084	193	5
150	0.060	192	5

为了符号一致, 令 $x = s^{-1/2}$, $e =$ 较小项, 故 (4.13) 可被写成

$$y_i = \beta_0 + \beta_1 \cdot x_i + e_i \quad (i = 1, 2, \dots, n) \quad (4.14)$$

并且 e_i 相互独立, 如表(4.1)给出的 $\text{var}(e_i) = \sigma^2/w_i$ 。不失一般性, 我们设 $\sigma^2 = 1$ 。故表 4.1 中第 4 列的数值对应于 $1/\sqrt{w_i}$, $i=1, 2, \dots, n$ 。

β_0 和 β_1 的估计值必须通过加权最小二乘法获得。这或者可以通过直接求广义残差平方和的最小值, 其标量形式由下式给出

$$RSS(\beta) = \sum w_i (y_i - X_i^T \beta)^2 \quad (4.15)$$

或者可以通过变换数据尺度, 然后应用普通最小二乘法。表 4.2 列出加权最小二乘法计算结果, 其拟合线作于图 4.1 中。常用统计量, 如 R^2 及 t -检验表示, 拟合模型相当好地与观测数据匹配, 他们的参数估计也很好地被确定了。下一个问题是 (4.13) 是否确实拟合了数据。这个关于模型拟合或拟合失真的问题是下一节讨论的主题。

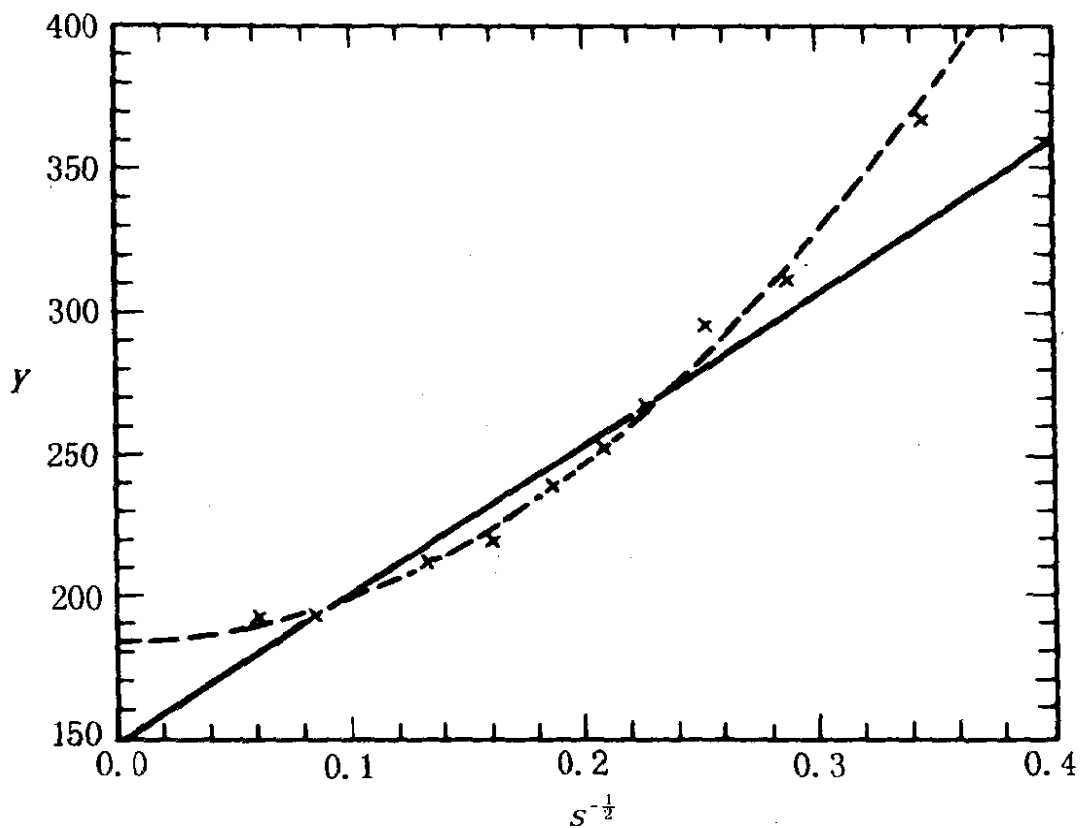
附加评注 许多统计模型, 包括方差分量, 时间序列, 以及一些经济计量模型都认为, Σ 依赖于少数几个参数。这导致迭代重加权最小二乘法, 定义为:

1. 令 $\tilde{\Sigma} = I$, 或其它一些方便的初始值。
2. 在 $\Sigma = \tilde{\Sigma}$ 时估计系数。
3. 由第 2 步得到的残差, 给出 Σ 的一个新的估计 $\tilde{\Sigma}$ 。
4. 重复第 2 与第 3 步, 直到 $\tilde{\Sigma}$ 不再有大的改变。

这个方法可用渐近理论进行证明, 它保证在大样本的情况下, 只要第 3 步的估计方法有好的性质, 估计将具有与 Σ 已知时相同的性质 (见 Carroll, 1982)。在小样本中, 如果误差呈对称分布, 参数估计很可能是无偏的, 但是估计的系数的标准误可能会太小; Freedman 和 Peters (1984) 曾发现在一个问题中低估了 3 倍。在迭代重加权最小二乘法中使用估计的权值, 可能在估计精度上误导研究者。

如果 x 的每个值在数据中出现多次, 则权值有时是可以估计的。对于固定的 x , 响应变量的样本方差给出 $\text{var}(y|x)$ 的一个估计。这些样本方差的逆可以用作权值。这一方法被用于获取在物理数据中的权值, 其中对于 x 的每个值的案例数是极大的。问题

4.7 给出估计权值作为真实权值的另一个例子。本方法的有效性依赖于对 x 的每个值要有一个大的样本量。



实线：拟合简单回归 虚线：拟合二次回归

图 4.1 物理数据的散点图

表 4.2 物理数据的加权最小二乘估计

变量	估计	标准误	t -值	
截距	148.473	8.079	18.38	
斜率	530.835	47.550	11.16	
$\sigma^2=2.744$, d. f. =8, $R^2=0.940$				
方差分析				
来源	d. f.	SS	MS	F
回归	1	341.991	341.991	124.63
残差	8	21.953	2.744	

4.2 方差已知时对拟合失真的检验

一个模型说明了, 自变量和响应变量的关系的形式。当假设的形式正确时, 拟合模型的残差均方 $\hat{\sigma}^2$ 将给出残差方差的一个无偏估计。如果假设的形式不正确, 因为 $\hat{\sigma}^2$ 的大小既依赖于误差, 也依赖于由拟合错误形式而引起的系统偏差, 故估计 $\hat{\sigma}^2$ 将大于 σ^2 。如果 σ^2 已知, 或者如果可以得到 σ^2 的不依赖于模型的估计, 一个关于模型拟合失真的检验, 可以通过比较 $\hat{\sigma}^2$ 和与模型无关的那个值得到。如果 $\hat{\sigma}^2$ 太大, 我们就有了拟合模型不合适的证据。

对例 4.1 的物理数据, 我们试图知道直线模型 (4.13) 或 (4.14) 是否给出了一个合适的描述。如 4.1 节所述, 表 4.1 中第 4 列值的平方的倒数被用作权值, 其中 $\sigma^2=1$, 为一已知值。由表 4.2, $\hat{\sigma}^2=2.744$ 。如果我们判定 $\hat{\sigma}^2=2.744$ 比已知值 $\sigma^2=1$ 大, 则我们得到了这一简单回归模型不合适的证据。为对这一比较赋一个 p 值, 我们使用如下结论:

如果 $e_i \sim NID(0, \sigma^2/w_i)$, $i=1, 2, \dots, n$, 其中 w_i 和 σ^2 已知, 并且线性模型的参数估计使用以 w_i 为权的加权最小二乘法, 若模型是正确的, 则

$$\chi^2 = \frac{RSS}{\sigma^2} = \frac{(n-p')}{\sigma^2} \hat{\sigma}^2 \quad (4.16)$$

服从自由度为 $n-p'$ 的 χ^2 随机变量的分布。和往常一样, RSS 为残差平方和。

对于例子, 由表 4.2,

$$\chi^2 = \frac{21.953}{1} = 21.953$$

从本书最后的表 C 得, $\chi^2(0.01, 8) = 20.09$ 。故检验对应的 p 值小于 0.01。这意味着模型可能不合适。

当这一检验表示拟合失真时, 通常通过变换某些自变量或响

应变量，或在自变量中加上多项式项来改变拟合模型。物理学理论提议采用后一种方法。模型

$$y = \beta_0 + \beta_1 \cdot s^{-1/2} + \beta_2 \cdot (s^{-1/2})^2 + \text{较小项}$$

或

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + e_i \quad (4.17)$$

可能较好地与数据拟合。这一模型要求 y 和 x 的关系是一条二次曲线，而不是一条直线。

表 4.3 物理数据的模型 (4.17)

变量	估计	标准误	t -值
截距	183.831	6.459	28.46
x	0.971	85.369	0.01
x^2	1 597.505	250.586	6.38

$$\hat{\sigma}^2 = 0.461, \text{ d.f. } = 7, R^2 = 0.991$$

方差分析				
来源	d.f.	SS	MS	F
回归	2	360.719	180.359	391.41
残差	7	3.226	0.461	

模型(4.17)是一个有两个自变量 x 和 x^2 的多元回归模型。拟合必须再次使用加权最小二乘法，其权重同前。拟合方程及方差分析由表 4.3 给出。拟合曲线作于图 4.1。曲线与数据非常匹配。我们可以通过计算

$$\chi^2 = \frac{RSS}{\sigma^2} = \frac{3.226}{1} = 3.226$$

来检验这个模型是否拟合失真。将这个值与 $\chi^2(7)$ 的分位点进行比较，得到模型 (4.17) 无拟合失真的结论。

尽管 (4.13) 没有很好地描述数据，而 (4.17) 确实产生了一个合适的拟合。对于强相互作用力模型成功或失败的判断，除了这个数据分析外，还要进行其它射入及射出粒子的数据分析。基

于这进一步的分析, Weisberg et al (1978) 得出, 强相互作用力的理论模型与观测数据一致的结论。

4.3 方差未知时对拟合失真的检验

当 σ^2 未知时, 对拟合失真的检验需要有与模型无关的残差方差的估计。最常见的与模型无关的估计是利用所有具有相同自变量值的案例之间的变化。例如, 考察由表 4.4 给出的人造数据, $n=10$ 。数据的产生是通过首先选择 x_i 的值, 然后计算 $y_i=2.0+0.5x_i+e_i$, $i=1, 2, \dots, 10$, 其中 e_i 取自标准正态随机数表。如果只考虑对应 $x_i=1$ 的 y_i 的值, 我们可以计算 y_i 的平均值 \bar{y} 及自由度为 $3-1=2$ 的标准差 SD , 如表中所示。如果我们假设对所有 x 值, 真正的残差方差都相等, 则通过合并 SD 为一个估计可得到共同方差的一个合并估计。如果 n 是 x 的一个取值处的案例数, 则关于纯误差的平方和, $SS(pe)$, 可以表示为:

$$SS(pe) = \sum (n-1)SD^2 = \sum \sum (y_i - \bar{y})^2 \quad (4.18)$$

其中的和取遍所有案例组。例如, $SS(pe)$ 可以是表 4.4 中第 4 列的所有数字的和,

$$SS(pe) = 0.0243 + 0.0000 + 0.1301 + 2.2041 = 2.3585$$

与 $SS(pe)$ 相关联的是其自由度, $d.f.(pe) = \sum (n-1) = 2+0+1+3=6$ 。方差的合并, 或纯误差估计为 $\hat{\sigma}_{pooled}^2 = SS(pe)/d.f.(pe) = 0.3931$ 。这与如果数据是根据 x 值分组, 通过单向方差分析得到的残差方差是相同的。

方差的纯误差估计没有参考线性回归模型。它利用了例子中, 对每个 x 的残差方差都相等以及所有的观测值都相互独立的假设。现在假定我们对数据拟合一个线性回归模型。方差分析由表 4.5 给出。表 4.5 的残差均方给出 σ^2 的一个估计。这一估计依赖于模型。这样我们有 σ^2 的两个估计。如果后者比前者大得多, 则说明模型是不适当的。

表 4.4 一个假设的例子

X	Y	\bar{y}	$\sum (y_i - \bar{y})^2$	SD	d. f.
1	2.55	2.6233	0.0243	0.1102	2
1	2.75				
1	2.57				
2	2.40	2.4000	0	0	0
3	4.19	4.4450	0.1301	0.3606	1
3	4.70				
4	3.81	4.0325	2.2041	0.8571	3
4	4.87				
4	2.93				
4	4.52				
$SS (pe) = 2.3585, \text{ d. f. (pe)} = 6$					

表 4.5 方差分析

来 源	d. f.	SS	MS	F
回归	1	4.5693	4.5693	
残差	8	4.2166	0.5271	
拟合失真	2	1.8581	0.9291	2.36
纯误差	6	2.3585	0.3931	

表 4.5 的残差平方和可以被分成两部分，即如表 4.4 给出的纯误差平方和以及剩下的部分，称为拟合失真平方和，或 $SS(lof) = RSS - SS(pe) = 4.2166 - 2.3585 = 1.8581$ ，其自由度为 $n - p' - d.f.(pe)$ 。由此我们得到一个 F -检验。这个 F -检验是拟合失真均方与纯误差均方的比值。观测到的 $F = 2.36$ 比 $F(0.05; 2, 6) = 5.14$ 小得多，这意味着模型对这些数据没有拟合失真。

虽然本节中的例都只有单个自变量，但得到一个与模型无关

的 σ^2 的估计的想法是相当一般的。方差的纯误差估计是以所有有相同自变量值的案例的响应变量值之间的平方和为基础的。

例 4.2 苹果树枝

许多种类的树生出两种在生物形态上不同的树枝。有些树枝一年年保持生长,并且对树的大小具有相当影响。它们被称为长枝。在一个生长季节里它们可能会长 15 到 20cm。

另一方面,某些树枝在总长度上很少超过 1cm,称为短枝。它们通常能生出花朵,并从花朵结出果实。使问题更为复杂的是,长枝偶尔在一个生长季节里会变成短枝,反之亦然。树木用于控制长、短枝的机制目前并不清楚。

Bland (1978) 完成了一个对 McIntosh 苹果树长、短枝的差别的描述性研究。使用种于 1933 和 1934 年的健康树木的无性系根茎。他在 1971 年的生长季节(约 106 天)里每隔几天即对长、短枝采样。样枝被假设为在采样日有效的树枝。样枝被从树上剪下,作标记后带到实验室供分析。

在所作的许多测量中, Bland 计算在每一树枝上茎元的个数。长、短枝会因茎元的个数或茎元的平均大小的不同,或因这两者都不同而不同。Bland 数据的一个摘要在表 4.6 给出,其中包括长枝和短枝。现在我们只考虑长枝,而将短枝留给问题部分。

我们的目的是要找到一个能适当描述 $DAY =$ 休眠天数和 $Y =$ 茎元个数之间的关系的方程。由于缺乏这一方程的理论形式,我们首先查看图 4.2。它是平均茎元数关于 DAY 的散点图。图表示的明显的线性关系提示我们拟合一条直线

$$Y = \beta_0 + \beta_1 \cdot DAY + \text{误差} \quad (4.19)$$

如果模型是适当的,我们可以得到在生长季节,观察到的每天的茎元生长率是个常数,这个有趣的结论。

对每个采样日,表 4.6 给出 $n =$ 样本树枝数, $\bar{y} =$ 那一天的平均茎元数, $SD =$ 一日之内的标准差。假设残差方差每天相同,我们可以按两种方法进行回归。首先,由于 $\text{var}(\bar{y}) = \sigma^2/n$, 我们可以计算 \bar{y}_i 关于 DAY 的加权回归,权重为 n 。这由表 4.7 给出。另一种方法是,如果可以得到原始的 189 个数据点,我们可以计算原始数据关于 DAY 的不加权回归。这由表 4.8 给出。这两种方法给出了相同的截距、斜率及回归平方和。关于残差平方和的计算,它们有不同的结果。表 4.8 中的残差平方和是 $SS(\text{pe})$ 与 $SS(\text{lof})$ 的和。例如,两个表中系数的标准误不同,这是因为在表 4.7 中,显示出方差估计为 3.7196,自由度为 20,而在表 4.8 中为 1.7621,自由度为 187。通常情况下,

表 4.6 Bland 的苹果树长、短枝数据

长 枝				短枝			
DAY	<i>n</i>	\bar{y}	<i>SD</i>	DAY	<i>n</i>	\bar{y}	<i>SD</i>
0	5	10.20	0.83	0	5	10.00	0.00
3	5	10.40	0.54	6	5	11.00	0.72
7	5	10.60	0.54	9	5	10.00	0.72
13	6	12.50	0.83	19	11	13.36	1.03
18	5	12.00	1.41	27	7	14.29	0.95
24	4	15.00	0.82	30	8	14.50	1.19
25	6	15.17	0.76	32	8	15.38	0.51
32	5	17.00	0.72	34	5	16.60	0.89
38	7	18.71	0.74	36	6	15.50	0.54
42	9	19.22	0.84	38	7	16.86	1.35
44	10	20.00	1.26	40	4	17.50	0.58
49	19	20.32	1.00	42	3	17.33	1.52
52	14	22.07	1.20	44	8	18.00	0.76
55	11	22.64	1.76	48	22	18.46	0.75
58	9	22.78	0.84	50	7	17.71	0.95
61	14	23.93	1.16	55	24	19.42	0.78
69	10	25.50	0.98	58	15	20.60	0.62
73	12	25.08	1.94	61	12	21.00	0.73
76	9	26.67	1.23	64	15	22.33	0.89
88	7	28.00	1.01	67	10	22.20	0.79
100	10	31.67	1.42	75	14	23.86	1.09
106	7	32.14	2.28	79	12	24.42	1.00
				82	19	24.79	0.52
				85	5	25.00	1.01
				88	27	26.04	0.99
				91	5	26.60	0.54
				94	16	27.12	1.16
				97	12	26.83	0.59
				100	10	28.70	0.47
				106	15	29.33	1.74

特别在模型可疑的情况下,更适宜只用纯误差来估计 σ^2 。这将导致第三个标

准误差集合。

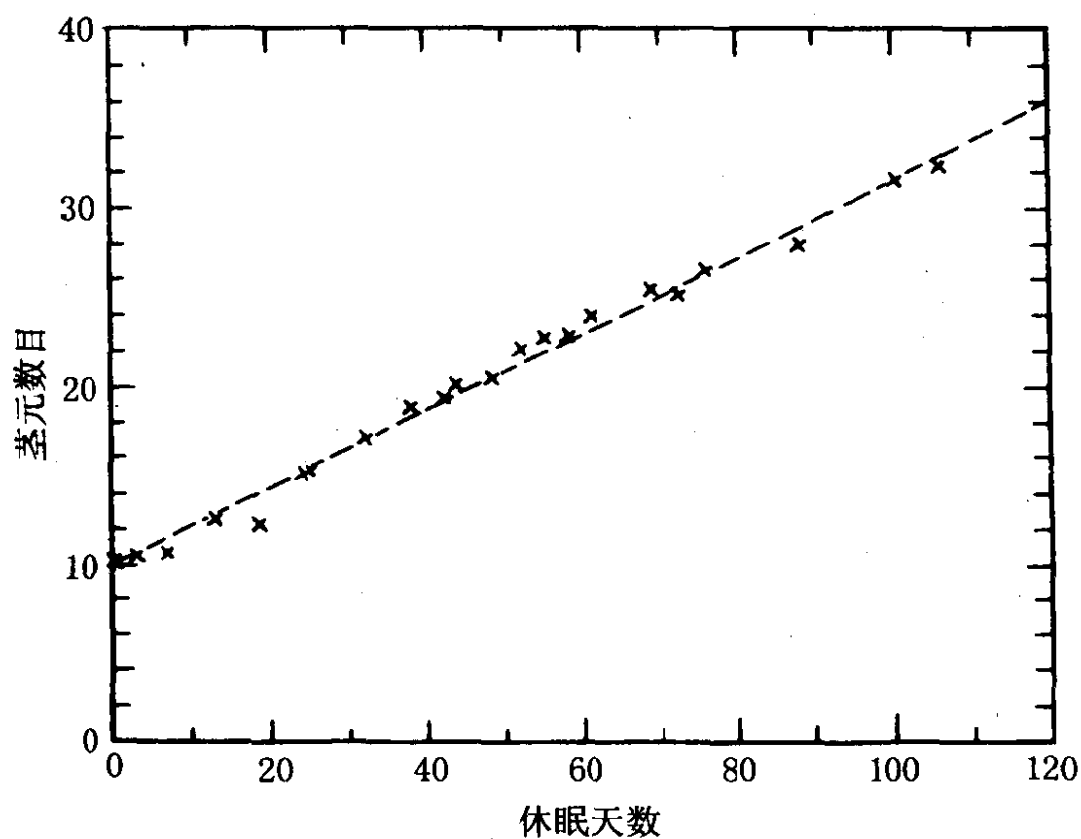


图 4.2 苹果树枝数据的散点图

表 4.7 \bar{y} 关于 DAY 的加权回归

变量	估计	标准误	t-值	
截距	9.9738	.31427	31.74	
DAY	0.2173	.00534	40.71	
$\hat{\sigma}^2=3.7196, \quad \text{d.f.}=20, \quad R^2=0.988$				
方差分析				
来源	d.f.	SS	MS	F
回归	1	6164.28	6164.28	1 657.2
残差	20	74.39	3.72	

表 4.8 y_i 关于 DAY 的不加权回归

变量	估计	标准误	t -值	
截距	9.9738	.21630	46.11	
DAY	0.2173	.00367	59.12	
$\hat{\sigma}^2=1.7621, \quad \text{d.f.}=187, \quad R^2=0.949$				
方差分析				
来源	d. f.	SS	MS	F
回归	1	6164.28	6164.28	3498.35
残差	187	329.50	1.76	
{ 拟合失真 纯误差	{ 20 167	{ 74.39 255.11	{ 3.72 1.53	2.43

利用 (4.18), 可以从表 4.6 直接计算 $SS(\hat{p}e)$, $SS(\hat{p}e) = \sum (n-1)SD^2 = 255.11$, 其自由度为 $\sum (n-1) = 167$ 。对拟合失真 F -检验为 $F = 2.43$ 。由于 $F(0.01; 20, 167) = 1.99$, 这一检验的 p -值小于 0.01。这表示直线模型 (4.19) 是不适当的。不过, 有这么大的自由度的 F -检验是强有力的, 它将测出零假设的很小的偏差。这样, 尽管这一结果在统计上是重要的, 在科学上未必重要。为了描述苹果树枝的生长, 模型 (4.19) 可能是适当的。

附加评注 纯误差检验要求重复观测以得到方差估计。如果没有重复试验或附加信息, 则没有对拟合失真的检验。Daniel 和 Wood (1981) 提出一个对拟合失真检验的近似。他们的想法是利用聚类算法找出几乎重复的案例, 并利用其响应变量的变化来计算拟合失真的检验。检验的有效性依赖于聚类算法的成功。不同的聚合给出不同的检验。目前, 尚缺乏这一方法的经验及支持的理论。Landwehr, Pregibon 和 Shoemaker (1984) 给出这一想法在逻辑斯谛回归中的一个应用。

4.4 广义 F -检验

到目前为止, 我们已多次遇到需要计算一个统计量, 当有零假设及正态性时, 统计量名义上服从 F 分布。 F -检验的理论是相当普遍的。在基本结构上, 一个小模型 (零假设) 与一个大模型 (备择假设) 相比较。小模型可以通过将大模型中的某些参数置为零, 或置为彼此相等, 或置为等于某个特定值而得到。前面遇到的一个例子是检验, 在一个多元回归中拟合了前 $p' - q$ 个自变量后, 剩下的 q 个自变量是否需要。用矩阵形式表述, 矩阵分块 $X = (X_1, X_2)$, 其中 X_1 为 $n \times (p' - q)$, X_2 为 $n \times q$; 分块 $\beta' = (\beta_1', \beta_2')$, 其中 β_1 为 $(p' - q) \times 1$, β_2 为 $q \times 1$ 。故两个假设 NH 和 AH 为

$$\begin{aligned} \text{NH: } Y &= X_1 \cdot \beta_1 + e \\ \text{AH: } Y &= X_1 \cdot \beta_1 + X_2 \cdot \beta_2 + e \end{aligned} \quad (4.20)$$

小模型是通过在大模型中置 $\beta_2 = 0$ 得到的。

为计算 F -检验, 两个模型都要作观测数据的拟合。在 NH 下, 得到残差平方和 RSS_{NH} 及其自由度 d. f. $_{\text{NH}}$ 。类似地, 在备择假设下, 得到 RSS_{AH} 和 d. f. $_{\text{AH}}$ 。显然, 由于备择模型拟合更多的参数, 所以 d. f. $_{\text{NH}} > \text{d. f.}_{\text{AH}}$ 。另外, 由于 AH 的拟合至少不会比 NH 的拟合差, 故 $RSS_{\text{NH}} - RSS_{\text{AH}} \geq 0$ 。与 $F(\text{d. f.}_{\text{NH}} - \text{d. f.}_{\text{AH}}, \text{d. f.}_{\text{AH}})$ 的分位点进行比较, 如果

$$F = \frac{(RSS_{\text{NH}} - RSS_{\text{AH}}) / (\text{d. f.}_{\text{NH}} - \text{d. f.}_{\text{AH}})}{RSS_{\text{AH}} / \text{d. f.}_{\text{AH}}} \quad (4.21)$$

大, 则 F -检验说明, NH 拟合不适当。

非零分布 (4.21) 的分子和分母是相互独立地分布的。在正态性及 AH 下, 每一个的分布如同 σ^2 乘以一个 (非中心) χ^2 变量除以其自由度。特别地, (4.21) 的分子的期望值为

$$E((4.21) \text{ 的分子}) = \sigma^2 \left(1 + \frac{\text{非中心参数}}{q} \right) \quad (4.22)$$

对于 (4.20) 的那个假设, 非中心参数由下式给出

$$\frac{\beta_2^T (X_2^T X_2 - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2) \beta_2}{\sigma^2} \quad (4.23)$$

为帮助理解这一点, 考虑特例, $X_2^T X_2 = I$ 且 $X_1^T X_2 = 0$, 即 X_2 中的变量相互正交, 且对 X_1 中的变量是正交的。则 (4.22) 变为

$$E(\text{分子}) = \sigma^2 + \beta_2^T \beta_2 / q \quad (4.24)$$

对这一特例, 如果 β_2 大的话, (4.21) 中分子的期望值及 F -检验的功效将是大的。在一般情况下, $X_1^T X_2 \neq 0$, 结果将更加复杂, 非中心参数的大小及 F -检验的功效不仅依赖于 β_2 , 也依赖于 X_1 中变量与 X_2 中变量的样本相关系数。如果这些相关系数大的话, 即使 β_2 是大的, F 的功效也可能是小的。

有关 F -检验的更普遍的结论在高等线性模型教材, 如 Seber (1977) 中有描述。

附加评注 当误差项服从正态分布时, 本节导出的 F -检验有许多重要性质。例如, 它们是似然比检验, 并且这种检验的所有性质都适用于 F -检验。由于误差项通常不具有准确的正态性, 从实际角度来看, 最优性的讨论是不必要的。幸运的是, 当误差偏离正态性时, 这些都是“稳健的”, 即估计、检验及置信推断只是有节制地受到偏离正态性的影响。

4.5 联合置信区域

正如单个参数的置信区间是基于 t 分布的, 对多个参数的置信域需要使用 F 分布。这些区域是椭圆形的。

β 的 $(1-\alpha) \times 100\%$ 置信域是向量 β 的集合, 满足

$$\frac{(\beta - \hat{\beta})^T (X^T X) (\beta - \hat{\beta})}{p' \hat{\sigma}^2} \leq F(\alpha; p', n - p') \quad (4.25)$$

通常我们对不包括 β_0 的参数向量 β^* 的置信域感兴趣。利用第二章的表述, β^* 的 $(1-\alpha) \times 100\%$ 区域是点 β^* 的集合, 满足

$$\frac{(\beta^* - \hat{\beta})^T (\mathcal{X}^T \mathcal{X}) (\beta^* - \hat{\beta})}{p \hat{\sigma}^2} \leq F(\alpha; p, n - p') \quad (4.26)$$

区域 (4.25) 是个 p' 维椭圆, 它以 $\hat{\beta}$ 为中心, 而 (4.26) 是个 p 维椭圆, 以 β^* 为中心。

例如, 例 2.1 中, $FUEL$ 关于 $X_1 = TAX$ 和 $X_2 = DLIC$ 的回归中, β_1, β_2 的 95% 的置信域由图 4.3 给出。这一椭圆的中心在 $(-32.075, 12.515)$ 。椭圆的方向 (主轴和副轴的方向) 由 $X^T X$, 或等价地由 X_1 和 X_2 的样本相关系数决定。如果 X_1 和 X_2 是不相关的, 则椭圆的轴将平行于 X_1 和 X_2 轴。

对 β 的任意子集的置信椭圆 与其给出普遍结果, 不如在这里给出可以导出普遍结果的一个特例。假设在燃料消耗问题的数据中, 需要从具有 4 个变量的模型中得到 $(\beta_1, \beta_2)^T$ 的 95% 置信域。令 S 为对应于 X_1 和 X_2 (例子中的 TAX 和 $DLIC$) 的 $(X^T X)^{-1}$ 的一个 2×2 的子矩阵。即, 由表 2.3,

$$S = \begin{bmatrix} 0.0382636 & 0.0022158 \\ 0.0022158 & 0.0008411 \end{bmatrix}$$

则 95% 的置信域为点 $\beta = (\beta_1, \beta_2)^T$ 的集合, 满足

$$\frac{\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \right)^T S^{-1} \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \right)}{2 \hat{\sigma}^2} \leq F(\alpha; 2, n - p') \quad (4.27)$$

其中 $(\hat{\beta}_1, \hat{\beta}_2)$ 由四元模型计算而得, $(\hat{\beta}_1, \hat{\beta}_2) = (-34.79, 13.36)$ 。这一区域由图 4.4 给出。对这一例子, 它与图 4.3 没有太大的不同。在其它问题中, 这些区域可能会相当地不同, 正如参数估计可能因模型而改变一样。

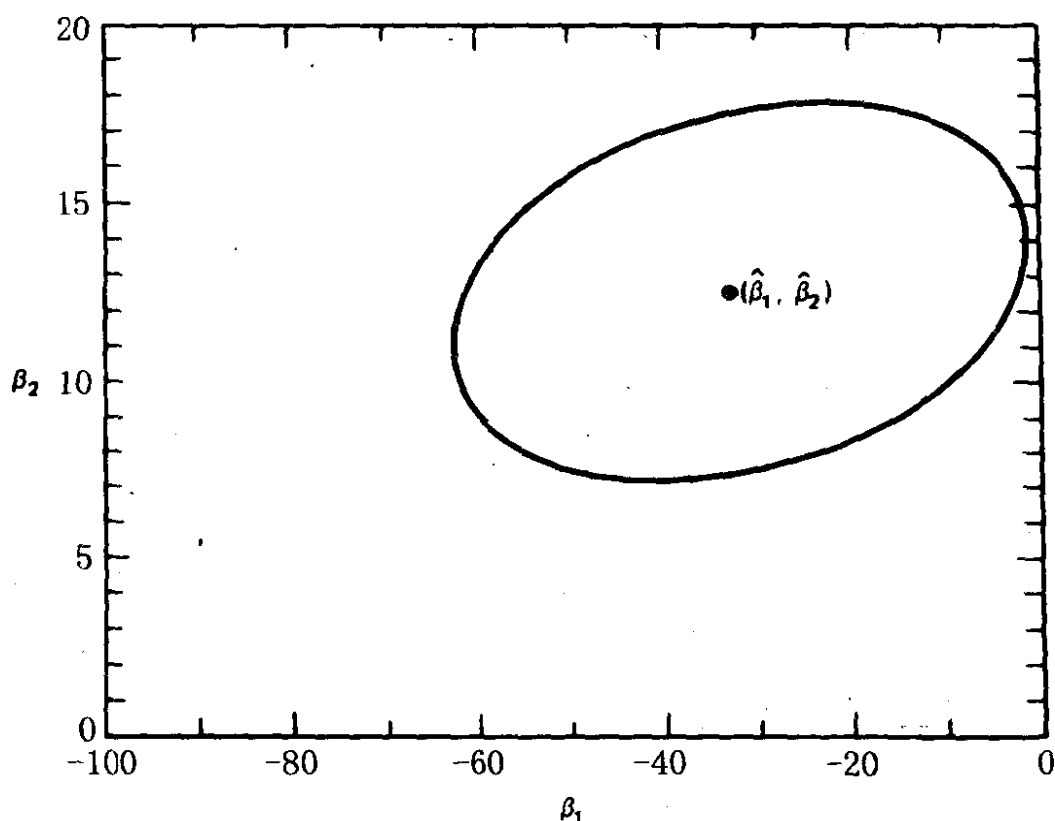


图 4.3 95%联合置信区域

问 题

- 4.1 Galton 的甜豌豆。许多关于回归的想法首先在高尔顿爵士的关于一代到下一代的遗传特征的工作中出现。在一篇“遗传的典型定律”的论文中(于 1877 年 2 月 9 日递交给皇家研究院), Galton 讨论了甜豌豆的一些试验。通过比较由母株产生的甜豌豆与子株产生的甜豌豆,他可以观察到一代到下一代的遗传。Galton 按母株结出的甜豌豆的代表性直径对母株进行分类。对从 0.15 到 0.21 英寸的七种大小类别,他安排了他的九位朋友中的每一个种植从每一类得到的十个植株。不过,有两个没有收获。Galton 数据概要后来由 Karl. Pearson (1930) 发表了(见表 4.9)。Pearson 只给出了子样豌豆的平均直径及标准差。样本量未知。
- 4.1.1 作 $Y = \text{平均子株直径}$ 关于 $X = \text{母株直径}$ 的散点图。
- 4.1.2 假设给出的标准差为总体的值,计算 Y 关于 X 的加权回归。在散点图上作出拟合直线。

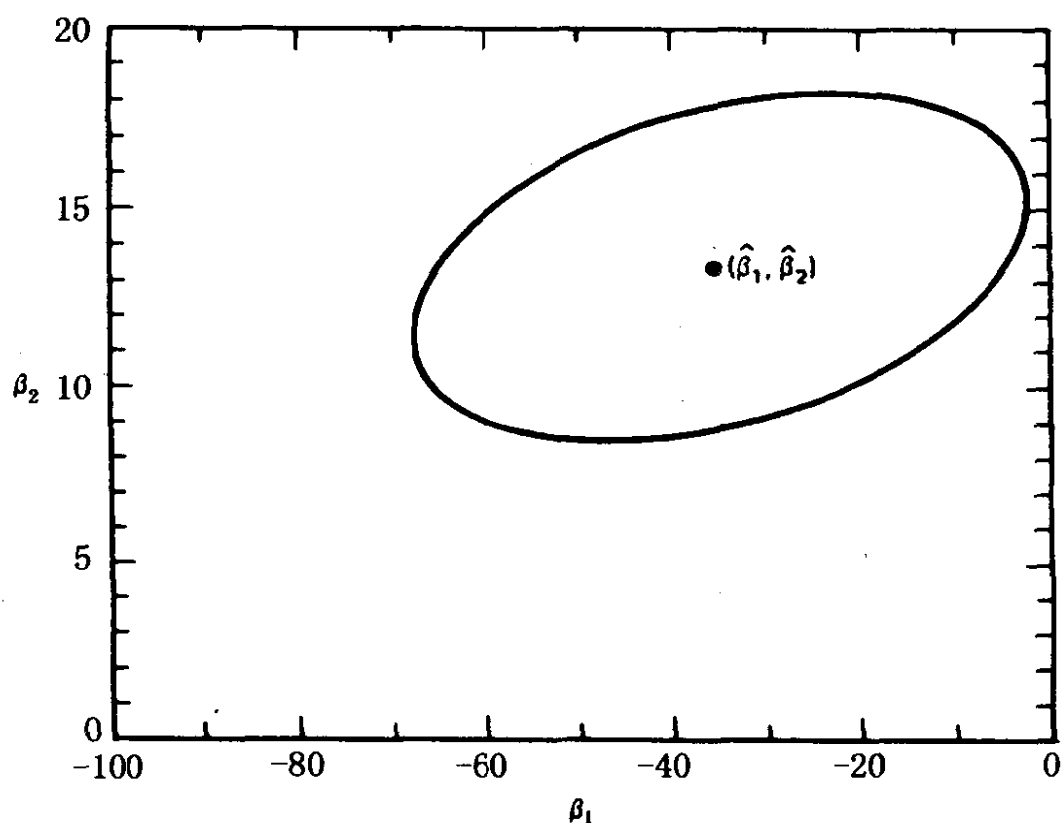


图 4.4 95% 联合条件置信区域

表 4.9 高尔顿数据

母株豌豆直径 (百分之一英寸)	子株豌豆直径 (百分之一英寸)	标准差
21	17.26	1.988
20	17.07	1.938
19	16.37	1.896
18	16.40	2.037
17	16.13	1.654
16	16.17	1.594
15	15.98	1.763

4.1.3 Galton 想要知道母株的特征, 如大小等是否传给了子株。在拟合回归中, 参数值 $\beta_1 = 1$ 对应于完全遗传, 而 $\beta_1 < 1$ 表示子株“恢复”了某些或许可以大略称为“一般祖先类型”(很可能, Galton 在 1885 年用“回归”代替了“恢复”) 检验假设 $\beta_1 = 1$, 取备择假设为 $\beta_1 < 1$ 。

4.1.4 作关于直线模型拟合失真的检验。

表 4.10 强相互作用数据

p_a^{lab} (GeV/c)	$s^{-1/2}$ GeV/c ⁻¹	$y(a p \rightarrow \pi^-)$ (μb)	$y(a p \rightarrow \pi^+)$ (μb)
$a = \pi^-$			
4	0.345	367 ± 17	284 ± 13
6	0.287	311 ± 9	288 ± 9
8	0.251	295 ± 9	304 ± 9
10	0.225	268 ± 7	284 ± 8
12	0.207	253 ± 7	281 ± 7
15	0.186	239 ± 6	276 ± 7
20	0.161	220 ± 6	275 ± 7
24	0.148		257 ± 10
30	0.132	213 ± 6	275 ± 7
75	0.084	193 ± 5	277 ± 7
150	0.060	192 ± 5	281 ± 7
$a = \pi^+$			
4	0.345	133 ± 6	499 ± 18
6	0.287	157 ± 6	464 ± 13
8	0.251	157 ± 5	442 ± 12
10	0.225	154 ± 6	389 ± 11
12	0.207	149 ± 6	388 ± 12
15	0.186	156 ± 10	341 ± 14
20	0.161	173 ± 22	375 ± 22
30	0.132	155 ± 5	331 ± 9
75	0.084	166 ± 5	303 ± 8
150	0.060	160 ± 6	311 ± 9
250	0.046	171 ± 11	306 ± 18

4.1.5 在Galton 的实验中, 他取一株植株所生产的所有豌豆的平均尺寸来划分其母株所属的类。在选择代表植株及培育下一代的种子时, Galton 选择尽可能接近平均尺寸的种子。这样对一个小

植株，特别大的种子被选作为代表，而大的更强壮的植株，则选用相对小的种子来代表。你认为这些试验偏差对（1）截距和斜率的估计，（2）误差估计，有何影响。

4.2 物理数据。表 4.10 给出类似例 4.1，选择 a 和 c 等于 π^+ 和 π^- 的试验结果（来自 Weisberg et al. (1978)）。在 $a=c=\pi^-$ 时的数据也在表 4.1 给出。表 4.10 中土号右边的值为 $\sigma/\sqrt{\omega_i}$ 。对这些数据集中的的一个或多个，拟合模型 (4.13)，并检验拟合失真。若模型拟合失真，尝试用 (4.17) 并再次检验是否拟合失真。不严格地对 a 和 c 的各种组合的拟合模型进行比较。

4.3 苹果树枝、对表 4.6 的短枝数据，用 4.3 节描述的方法进行分析。不严格地比较短枝和长枝的拟合回归。（一定要作 \bar{y} 关于 DAY 的散点图）

4.4 控制图、表 4.11 给出数天里一个工业过程生产的曲柄销的外围直径的数据(Jensen, 1977)。所有生产的曲柄销的直径必须在 0.7425 到 0.7430 英寸之间。表中给出的数据是关于 0.742 英寸的偏差。偏差以 0.00001 英寸为单位。例如，数字 93 表示 $0.742 + 0.00093 = 0.74293$ 英寸。当生产过程被控制，曲柄销的平均尺寸必须（1）落在指定范围的中点的两侧，且（2）不依赖于时间而变化。拟合适当的模型，看生产过程是否处于控制之下，并检验模型是否拟合失真。

4.5 一个 F -检验、设有简单回归

$$NH: y_i = x_i + e_i \quad (\beta_0 = 0, \beta_1 = 1)$$

$$AH: y_i = \beta_0 + \beta_1 \cdot x_i + e_i \quad (i = 1, 2, \dots, n)$$

求出 F -检验的显式公式。

4.6 雪鹅。空中调查的方法常被用于估计加拿大哈得逊湾以西夏季雪鹅的数目。为获得这一估计值，小飞机在这一地区飞行，当发现一群雪鹅时，有经验的人会估计出这群雪鹅的数目。为研究这一计数方法的可靠性，做了一个实验，让一架飞机载着 2 位观测者飞过 $n=45$ 群飞鹅，每个观测者独立地对每个鹅群中雪鹅的数目作出估计。对鹅群也拍了照片，以计算鹅群中的准确雪鹅数。所得的数据由表 4.12 给出 (Cook 和 Jacobsen, 1978)

4.6.1 作 Y =照片上计数关于 X_1 =观测者 1 计数，和关于 X_2 =观测者 2 计数的散点图。这些图是否表示，一个简单回归模型是适当的？为什么？在 Y 关于 X_1 ，或关于 X_2 的简单回归模型中，误差项测量什么？为什么用 Y 关于 X_1 或 X_2 作回归拟合，要比用

X_1 或 X_2 关于 Y 作回归更合适?

表 4.11 曲柄销数据

天	曲柄销直径
1	93, 98, 90, 94, 94
4	93, 100, 88, 85, 89
7	89, 90, 92, 95, 100
10	93, 88, 87, 87, 87
13	88, 86, 91, 89, 86
16	82, 72, 80, 72, 89
19	81, 80, 78, 94, 90
22	90, 92, 82, 77, 89

4.6.2 通过普通最小二乘法计算 Y 关于 X_1 以及 Y 关于 X_2 的回归, 并对每个观测者检验问题 4.5 中的假设。叙述这一假设的意义及检验的结果。哪个观测者更可靠些 (你必须给出“可靠”的定义) 总结你的结论。

4.6.3 拟合回归 $Y^{1/2}$ 关于 $X_1^{1/2}$, $Y^{1/2}$ 关于 $X_2^{1/2}$ 。重复 4.6.2。平方根刻度是用来稳定误差方差。

4.6.4 假设 $\text{var}(e_i) = x_i \sigma^2$, 重复 4.6.2。

作为实验的一个结果, 用目测决定鹅群的总体估计数的方法不再被用, 而采用拍照的方法。

4.7 Jevons 的金币。这个例子中的数据由 Stephen M. Stigler 提供, 是从 W. Stanley Jevons (1868) 的一篇论文的图中导出的。在对铸造货币进行的研究中, Jevons 称量了 274 个英国金镑。这是他从英国曼彻斯特的货币流通中收集的。对每个金币, 他记录了其重量, 直到 0.001 克, 并记录其铸造日期。表 4.13 列出每个年龄段中的平均, 最小及最大重量。年龄段被编号为 1 到 5, 大致对应着货币已生产了多少个十年。一个金币的标准重量为 7.9876 克, 最小的合法重量为 7.9379 克。

4.7.1 令 x = 年编号及 \bar{y} = 平均重量。作 \bar{y} 关于 x 的散点图, 并对常用线性回归模型的适用性作评论。另作 SD 关于 x 的散点图, 并总结这个图给出的信息。

4.7.2 由于每个年龄段的金币数 n 都相当大, 故可以合理地认为, x 年龄段的金币重量的方差由 SD^2 给出很好的近似。从而 $\text{var}(\bar{y})$ 由 SD^2/n 给出。由此得到权重, 并计算 \bar{y} 关于 x 的加权回归。

表 4.12 鹅群大小的估计

照片 计数	观测者 1 计数	观测者 2 计数	照片 计数	观测者 1 计数	观测者 2 计数
56	50	40	119	75	200
38	25	30	165	100	200
25	30	40	152	150	150
48	35	45	205	120	200
38	25	30	409	250	300
22	20	20	342	500	500
22	12	20	200	200	300
42	34	35	73	50	40
34	20	30	123	75	80
14	10	12	150	150	120
30	25	30	70	50	60
9	10	10	90	60	100
18	15	18	110	75	120
25	20	30	95	150	150
62	40	50	57	40	40
26	30	20	43	25	35
88	75	120	55	100	110
56	35	60	325	200	400
11	9	10	114	60	120
66	55	80	83	40	40
42	30	35	91	35	60
30	25	30	56	20	40
90	40	120			

4.7.3 计算线性回归模型的拟合失真的检验。概括所得的结论。

4.7.4 拟合回归是否与一块新的金币的标准重量一致？

4.7.5 对于年龄 $X=1, 2, 3, 4, 5$ 的先前未被采样的金币，估计其重量小于最小合法重量的概率。（提示：在计算中，利用年龄为 x 的金币的残差方差是已知值 SD^2 ，从而预测值将是正态，而不是

t -分布) 估计所有 $x=4$ 的金币轻于最小合法重量的比例。一个令人感兴趣的问题是, 确定金币预测重量等于最小合法重量的年龄 x 。 x 值的一个点估计可以通过设 $y=7.9379$, 然后解关于 x 的拟合回归方程而得到。这个问题称为逆回归, 在 Williams (1959, 第六章) 中有讨论。

表 4.13 金币数据

年龄, x , 以十年计	样本容 量= n	平均重 量= \bar{y}	SD	最小 重量	最大 重量
1	123	7.9725	0.01409	7.900	7.999
2	78	7.9503	0.02272	7.892	7.993
3	32	7.9276	0.03426	7.848	7.984
4	17	7.8962	0.04057	7.827	7.965
5	24	7.8730	0.05353	7.757	7.961

来源于: Stephen M. Stigler。

- 4.8 计算机程序。为在加权最小二乘法中应用扫描算法程序, 仅需要修改计算校正平方和及叉积和矩阵的算法。附录 1A.5 的适当的修改由 West (1979) 给出。令 \bar{x}_m, \bar{y}_m 表示读入 m 个案例后两个变量 X 和 Y 的平均值, $\sum w_i$ 为前 m 个案例权重的和, $SXX_m = \sum w_i (x_i - \bar{x}_m)^2$, $SYY_m = \sum w_i (y_i - \bar{y}_m)^2$, $SXY_m = \sum w_i (x_i - \bar{x}_m)(y_i - \bar{y}_m)$, 其中所有的求和都是对前 m 个案例求和。然后令 $\bar{x}_0 = \bar{y}_0 = SXX_0 = SYY_0 = SXY_0 = 0$ 。我们可以通过下述公式更新对 $(m+1)$ 个案例的估计,

$$SXX_{m+1} = SXX_m + \frac{w_{m+1}}{\sum w_i + w_{m+1}} (\sum w_i) (x_{m+1} - \bar{x}_m)^2$$

$$SYY_{m+1} = SYY_m + \frac{w_{m+1}}{\sum w_i + w_{m+1}} (\sum w_i) (y_{m+1} - \bar{y}_m)^2$$

$$SXY_{m+1} = SXY_m + \frac{w_{m+1}}{\sum w_i + w_{m+1}} (\sum w_i) (x_{m+1} - \bar{x}_m) (y_{m+1} - \bar{y}_m)$$

$$\bar{x}_{m+1} = \bar{x}_m + \frac{w_{m+1}}{\sum w_i + w_{m+1}} (x_{m+1} - \bar{x}_m)$$

$$\bar{y}_{m+1} = \bar{y}_m + \frac{w_{m+1}}{\sum w_i + w_{m+1}} (y_{m+1} - \bar{y}_m)$$

对任意两个变量, 这些公式给出了平方和、叉积和及平均值的更新。修改问题 2.6 所写的扫描程序来做加权最小二乘法。

5

诊断 I：残差及影响

到目前为止的得到估计、检验及其它主要统计量的方法，只是整个回归分析的一部分。在认为模型及假设都正确的情况下，这些方法被用于计算。然而在许多实际问题中，这些假设是令人怀疑的。分析的第二阶段用于检验假设与建立模型，这通常是需要的。在建模的初级阶段汇总数据以生成主要统计量，而这后一阶段需要统计量的检验。这些统计量通常对每个案例取值。作为一类，我们称之为诊断统计量，因为它们用于考察在分析中用到的假设的问题。

诊断所关心的是两个相互有关的问题。首先是模型在多大程度上与观测数据相一致。这里的基本统计量是残差的一个有效用的变换。如果拟合模型没有给出一个合理的残差集合，则模型的某些方面被称为有疑问的。第二个令人感兴趣的问题是，每个案例在估计及综合分析的其它方面的影响。在某些数据集合中，如果一个案例被删除，观测的综合统计量可能有重要改变。这样一个案例被称为影响性的。我们需要检测出这样的案例。我们将研究并使用两个相对熟悉的诊断统计法，称为距离测量和位势或杠杆值。

例 5.1 图的有效性 (Anscombe, 1973)

表 5.1 给出的四组人造数据很好地说明了案例分析的必要性。每个数据集由 11 对点 (x_i, y_i) 组成, 拟合于简单线性回归模型 $y_i = \beta_0 + \beta_1 \cdot x_i + e_i$ 。每个数据集导致一个相同的综合分析, 即

$$\hat{\beta}_0 = 3.0$$

$$\hat{\beta}_1 = 0.5$$

$$\hat{\sigma}^2 = 13.75$$

$$R^2 = 0.667$$

由于每个数据集的综合统计量是相同的, 有人可能会作出对每个数据集, 线性回归模型同等地合适的结论。然而, 最简单的案例分析表示, 这是不对的。这可以通过查看图 5.1 (a), —5.1 (d) 得到。

表 5.1 四个假设的数据集

案例 编号	数据集编号					
	1-3	1	2	3	4	4
	变量					
	X	Y	Y	Y	X	Y
1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

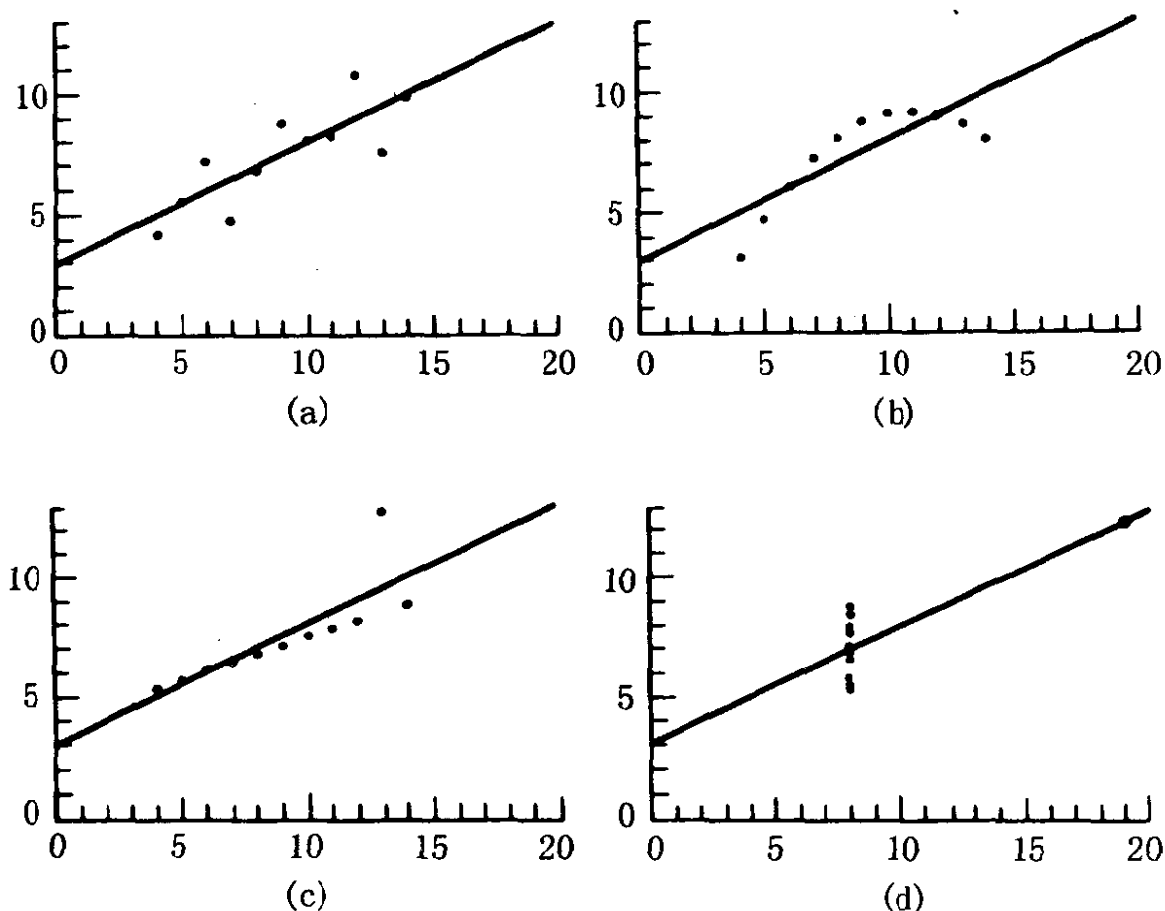


图 5.1 四个假设的数据集。经 Anscombe (1973)

同意后重新生成

第一个数据集合，作于图 5.1 (a)。如果简单线性回归模型合适的话，这就是我们期望看到的数据集合。图 5.1 (b) 给出第二个数据集合。它给出一个不同的结论，即基于简单线性回归的分析是不正确的，而一条光滑的曲线，可能是二次多项式，可以以较小的剩余变异性拟合于数据。

图 5.1 (c) 表示，简单回归的描述对于大部分数据是正确的，但一个案例距离拟合回归直线太远。这称为异常值问题。很可能需要从数据集合中删除那个与其它数据不匹配的案例。回归需要根据剩下的 10 个案例重新拟合。如果这样做的话，拟合方程为 $\hat{y} = 4.0 + 0.346x$ ，与根据 11 个案例得到的相当不同。如果不知道数据的具体内容，我们无法判断一条直线“正确”，而另一条“不正确”。需要了解和陈述两条直线之间的差别。

最后一个集合，作于图 5.1 (d)。它与上述其它三个不同，没有足够的信息来对拟合模型作判断。斜率参数估计值 $\hat{\beta}_1$ 很大程度上由 y_8 的值决定。

如果第8个案例被删除,我们不能估计 β_1 。我们无法相信这样一个综合分析,它对单个案例这般依赖。

为使案例统计量确实有用,我们需要了解它们在模型正确,以及如果可能的话,在模型不正确时候的表现。将案例统计量与它们的预期表现作比较,可以诊断出模型对观测数据的拟合是成功的,还是失败的。重要案例统计量的定义与性质是本章讨论的重点。这里讨论的统计量的使用方法,将在第六章给出。本章内容的更高数学水平的完整的论述,见Cook和Weisberg(1982a,第二和三章)

5.1 残差

为研究残差,我们使用第二章描述的矩阵表示。基本模型为

$$Y = X\beta + e, \quad \text{var}(e) = \sigma^2 I \quad (5.1)$$

其中 X 是已知的满秩矩阵,有 n 行 p' 列,如果模型包含所有分量为1的向量,则 $p' = \text{自变量个数} + 1 = p + 1$;如果所有分量为1的向量不包含在模型中,则 $p' = p$ 。类似地, β 是一个 $p' \times 1$ 的未知向量。向量 e 由未知误差组成。本章通篇假设误差同分布,且互不相关。论述扩充到加权最小二乘法是类似的,详见Cook和Weisberg, 1982a, 附录A.1)。

拟合模型(5.1),我们用 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 估计 β ,对应于观测值 Y 的拟合值 \hat{Y} 为

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= X [(X^T X)^{-1} X^T Y] \\ &= X (X^T X)^{-1} X^T Y \\ &= HY \end{aligned} \quad (5.2)$$

其中 H 是 $n \times n$ 矩阵,定义为*

* 在本书第一版中, V 用于表示这一帽子矩阵,其元素用 v_{ij} 表示。

$$H = X(X^T X)^{-1} X^T \quad (5.3)$$

H 称为帽子矩阵, 因为它将响应变量的观测值向量 Y 变换成响应变量的拟合值向量 \hat{Y} (通常读作 Y -角)。残差向量 \hat{e} 被定义为

$$\begin{aligned} \hat{e} &= Y - \hat{Y} \\ &= Y - X(X^T X)^{-1} X^T Y \\ &= [I - X(X^T X)^{-1} X^T] Y \\ &= [I - H] Y \end{aligned} \quad (5.4)$$

e 和 \hat{e} 的区别 误差 e 是不可观测的随机变量, 假设其均值为零, 且互不相关, 每个具有相同的方差 σ^2 。残差 \hat{e} 是可以画图表示, 或用其它方式研究的可计算的量。根据 (5.4) 及附录 2A.2, 它们的均值与方差为

$$\begin{aligned} E(\hat{e}) &= 0 \\ \text{var}(\hat{e}) &= \sigma^2 (I - H) \end{aligned} \quad (5.5)$$

类似于误差, 残差的均值都为零, 但残差可以有不同的方差, 且它们是相关的。由 (5.4) 可知, 残差是误差的线性组合, 故若误差是正态分布, 则残差亦是正态分布。另外, 如果模型包含截距, 则残差和为零, 即 $e^T \mathbf{1} = 0$ 。用标量形式表示, 第 i 个残差的方差为

$$\text{var}(\hat{e}_i) = \sigma^2 (1 - h_{ii}) \quad (5.6)$$

其中 h_{ii} 是 H 的第 i 个对角元素。诊断过程是基于计算所得的残差, 它被假设与不可观测的误差有着相同的行为。这个假设的效用依赖于帽子矩阵, 这是因为 H 联系了 \hat{e} 和 e , 并给出了 \hat{e} 的方差和协方差。

帽子矩阵 H 为 $n \times n$ 且是对称的。它具有许多易于直接从定义 (5.3) 加以验证的特殊性质。例如, 将 H 左乘 X 仍得 X , $HX = X$ 。类似地, $(I - H)X = 0$ 。性质 $HH = H^2 = H$ 也表示 $H(I - H) = 0$ 。故拟合值 HY 与残差 $(I - H)Y$ 的协方差 $\sigma^2 H(I - H) = 0$ 。 H 称为 X 的列空间上的正交投影算子。它一般是不可逆的, 且有和 X 相同的秩, 通常为 p' 。 H 的第 (i, j) 个元素记为 h_{ij} , 由下式给出

$$h_{ij} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_j = \mathbf{x}_j^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = h_{ji} \quad (5.7)$$

对角线上的元素 h_{ii} 为

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \quad (5.8)$$

在 h_{ij} 之间可以发现很多关系。例如

$$\sum_{i=1}^n h_{ii} = \text{rank}(\mathbf{X}) = p' \quad (5.9)$$

另外, 对有截距的模型, 每个 h_{ii} 被限制不小于 $1/n$ 。 h_{ii} 不大于 $1/r$, r 是 \mathbf{x} 中与 \mathbf{x}_i 相同的行向量的个数。还有, 对于含截距的模型, $\mathbf{H}\mathbf{1}=\mathbf{1}$, 或用标量形式

$$\sum_{i=1}^n h_{ij} = \sum_{j=1}^n h_{ij} = 1 \quad (5.10)$$

由 (5.6) 可见, 具有大的 h_{ii} 值的案例, 其 $\text{var}(\hat{e}_i)$ 取小的值。当 h_{ii} 接近于 1, 这个方差接近于零。对这样一个案例, 无论第 i 个案例的观测值 y_i 是多少, 我们几乎总能得到一个接近 0 的残差。Hoaglin 和 Welsch (1978) 采用 (5.2) 的标量形式

$$\hat{y}_i = \sum_{j=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \quad (5.11)$$

指出了这一点。联系到 (5.10), (5.11) 表示当 h_{ii} 接近 1 时, \hat{y}_i 接近于 y_i 。由此, 我们称 h_{ii} 为第 i 个案例的杠杆。然而, 这个特殊的定义忽略了随机变量 y_i 在确定 \hat{y}_i 时的作用, 所以 h_{ii} 又有一个不太独特的名字。Cook 和 Weisberg (1982a) 对 h_{ii} 使用位势的名字, 以提醒我们在回归中, 如果 h_{ii} 大的话, 第 i 个案例的作用可能也会是大的, 但其重要性是不定的, 依赖于 y_i 。 h_{ii} 和 y_i 的综合作用的度量方法, 在 5.3 节中给出。

我们可以给出使 h_{ii} 大或小的案例特征的几何描述。假设模型中有截距项。再次考虑第二章讨论的叉积矩阵的离差形式。令 $\mathcal{X}^T \mathcal{X}$ 为校正叉积矩阵 (2.16), $\bar{\mathbf{x}}$ 是 p 个自变量样本均值的 $p \times 1$ 向量, 并重新定义 \mathbf{x}_i^T 为 \mathbf{X} 的除去与截距项相对应的数 1 后的第 i 行。则我们能将 h_{ii} 表示成

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathcal{X}^T \mathcal{X})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (5.12)$$

在几何意义上, (5.12) 右边的第二项给出了中心位于 \bar{x} 的椭圆方程。

例如, 再一次考虑首先在第二章中讨论过的燃料消耗的数据。我们只用自变量为 $X_1 = TAX$ 和 $X_2 = DLIC$ 的模型, 由此可以画一个二维图象。 (X_1, X_2) 的数据由图 5.2 中的散点图给出。图中所画的椭圆对应于 h_{ii} 为常量 0.25, 0.20, 0.15, 0.10, 0.05 的椭圆等值线。这样, 任何恰好落在外圈等值线上的点有 $h_{ii} = 0.25$, 而落在最内圈等值线上的点有 $h_{ii} = 0.05$ 。这个距离的定义依赖于数据。在通常的欧几里得距离的意义下, 在椭圆长轴 (主轴) 附近的点, 要比有相同 h_{ii} 值的在副轴附近的点距离 \bar{x} 远得多。

在该例中, 具有最低税值的州——得克萨斯州, 税为 5 分/加仑——具有最大的 h_{ii} , 约为 0.2。虽然在其它数据集合中, h_{ii} 在 0.5 和 1.0 之间的案例并不少见, 但在本例中, 没有一个 h_{ii} 是非常大的。具有大的 h_{ii} 值的案例, 很可能对拟合一个模型最具有影响力。潜在影响不依赖于 Y , 而只依赖于 X 。影响力的综合度量必须同时考虑 Y 和 X 。

马哈拉诺比斯距离 如果去掉 (5.12) 等式右边的 $1/n$ 项, 再将剩余项乘以 $(n-1)$, 所得的量称为从 x_i 列数据中心 \bar{x} 的马哈拉诺比斯距离。马哈拉诺比斯距离在多元分析中应用很广, 特别是在判别分析中。它是单元间相对位置的一个度量。根据单元间的相对位置, 把单元分配到不同的总体中去。在判别分析中使用马哈拉诺比斯距离建立在 X 的行的多元正态性的基础上。我们既不需要也不使用这个多元正态性的假设。

例 5.2 简单回归

现在考虑第一章中的简单回归模型, 其中 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 。矩阵 H 是 $n \times n$ 的。即使简单回归, 其 H 的元素个数 n^2 也是相当大的。它很少被全部计算。然而, 利用 (2.25) 式得到 $(X^T X)^{-1}$, 由此获得单个 h_{ij} 的公式。我们有

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

$$\begin{aligned}
&= [1 \quad x_i] \begin{bmatrix} \frac{\sum x_i^2}{nSXX} & \frac{-\bar{x}}{SXX} \\ \frac{-\bar{x}}{SXX} & \frac{1}{SXX} \end{bmatrix} \begin{bmatrix} 1 \\ x_j \end{bmatrix} \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \quad (5.13)
\end{aligned}$$

在 (5.13) 中令 $j=i$, 我们得到诊断元素 h_{ii} 为

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \quad (5.14)$$

在 (5.14) 中直接求和将验证 (5.9) 的结论, 即 $\sum h_{ii} = p' = 2$ 。在 $x_i = \bar{x}$ 时, h_{ii} 将达到其最小值 $1/n$, 并且随着 x_i 偏离 \bar{x} , h_{ii} 的值将增加。只有在从数据集中删除第 i 个案例使模型中的一个参数不可估计的时候, h_{ii} 才能达到其最大值 1。在例 5.1 的第四组数据中, $h_{88} = 1$, 因为删除第 8 个案例使得斜率的估计变为不可能。

学生化残差 我们已经指出, 只要 h_{ii} 较大, 则 $\text{var}(\hat{e}_i)$ 将较小。故一般地说, 接近 \bar{x} 的案例 x_i 将比远离 \bar{x} 的案例有更大的残差。这是我们不希望的, 因为远离 \bar{x} 处, 尽管残差倾向较小, 但模型很可能会崩溃。这给我们一个提示, 换算 \hat{e}_i , 即将 \hat{e}_i 除以标准差的估计, 可以改进诊断。如果模型正确, 这些换算过的, 或学生化残差将具有相同的方差。我们考虑两个紧密联系的学生化方法, 它们只在 σ^2 的估计的选择上不同。第一个用 $\hat{\sigma}^2$ 来估计 σ^2 , 给出公式

$$r_i = \frac{\hat{e}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \quad (5.15)$$

r_i 称学生化内残差, 这是因为 σ^2 的估计用了包括第 i 个案例在内的全部数据。第二个换算使用当第 i 个案例从回归中删除后得到的 σ^2 的估计。这将在 5.2 节中讨论。

与 \hat{e}_i 不同, 虽然 $E(r_i) = 0, i = 1, 2, \dots, n$, 但 r_i 的和不为零。 r_i 和 \hat{y}_i 轻微相关, 但在实际应用中这一相关是可以忽略的。当模型正确时, 学生化残差的方差都等于 1, 与 σ^2 和 h_{ii} 无关。 r_i 和 r_j 的协方差等于 \hat{e}_i 和 \hat{e}_j 的相关系数,

$$\text{cov}(r_i, r_j) = \frac{-h_{ij}}{(1 - h_{ii})^{1/2} (1 - h_{jj})^{1/2}}$$

在模型及正态性假设下, $r_i^2 / (n - p')$ 的分布如参数为 $\frac{1}{2}$ 和 $(n - p' - 1) / 2$ 的 β_1 随机变量的分布。由于 β_1 变量介于 0 和 1 之间, r_i 介于 $-(n - p')^{1/2}$ 和 $+(n - p')^{1/2}$ 之间。

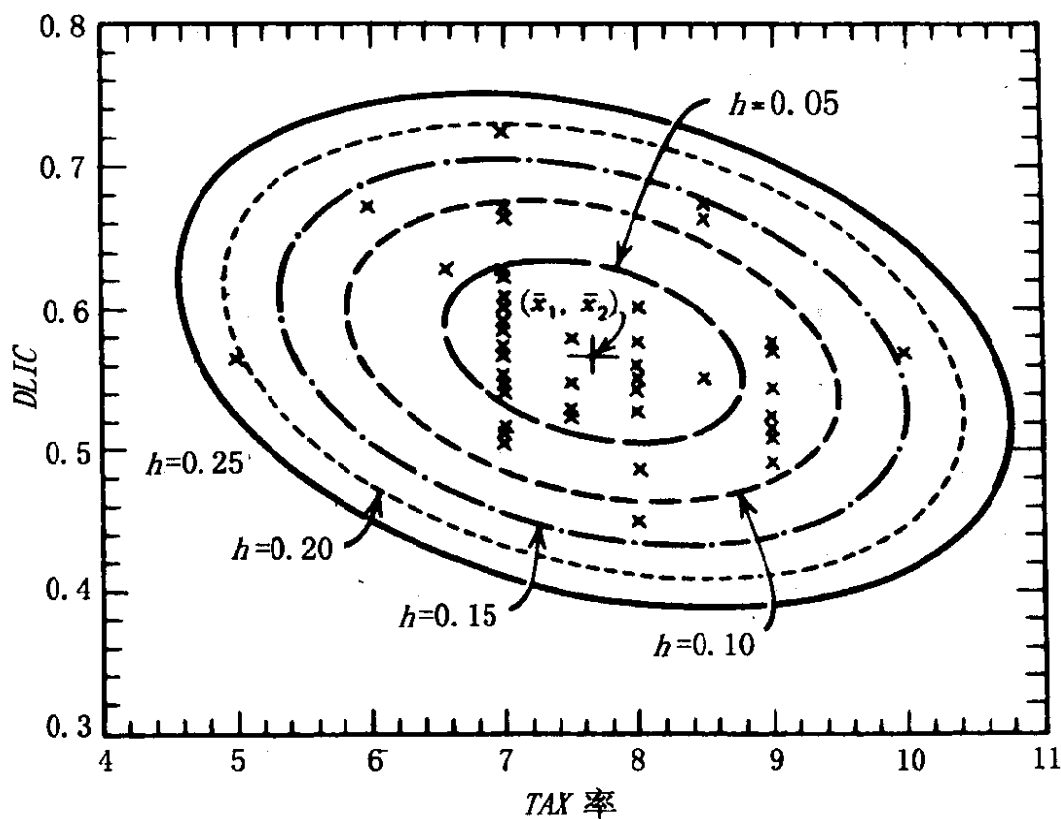


图 5.2 二维平面中常量 h_u 的等高线

5.2 异常值

在回归分析中的一个重要假设是, 使用的模型对所有数据是适当的。在应用中, 有一个或多个案例其观测值似乎与模型不相符, 但模型拟合于大多数数据, 这种情况并不罕见。如图 5.1c 的简单回归问题, 在 y 关于 x 的图中, 可以明显地看到, 大多数观测值落在拟合直线的附近, 但有小部分不是。如果一个案例不遵从某个模型, 但其余数据遵从这个模型, 则该案例称为异常值。案例分析的一个重要功能是区分这样的案例。

仔细定义术语异常值是有益的。为此我们使用一个称为均值漂移异常值模型的显式公式。设第 i 个案例可能是个异常值。我们假设对所有其它案例的模型为

$$y_j = x_j^T \beta + e_j \quad (j \neq i)$$

但对第 i 个案例，模型为

$$y_i = x_i^T \beta + \delta + e_i$$

第 i 个响应变量 y_i 有与 $x_i^T \beta$ 相差 δ 的期望值。因此，我们可以通过检验 $\delta = 0$ 来检验第 i 个案例是否为异常值。

在明确地导出检验之前，某些关于异常值性质的评注可能是有益的。我们将看到，被怀疑是异常值的，是那些有大的 $|e_i|$ 值的案例。并非所有有大的残差的案例都是异常值，这是因为根据模型，大误差 e_i 出现的频率由生成概率分布决定。无论我们采用什么检验程序，必须避免宣称太多的案例为异常值。这导致采用联合检验程序。另外，并非所有的异常值都是坏的。例如，我们可以想象一个地质模型，其中异常值对应着有石油贮藏，或其它有用特征的案例。它与大多数案例不一样。这样，找出异常值成了分析的目的。异常值的识别与一个特定的模型有关。如果模型的形式被修正，作为异常值的个别案例的情况也可能改变。最后，某些异常值可能比其它点在回归估计中有更大的作用。这点将在下节讨论。

异常值检验 假设第 i 个案例可能是个异常值。首先定义一个新的自变量 U ， U 的第 j 个元素， $u_j = 0$ ，当 $j \neq i$ 时，而第 i 个元素 $u_i = 1$ 。然后，简单计算 Y 关于 X 和 U 的回归。 U 的系数的估计是平均漂移 δ 。检验 $\delta = 0$ 的双边假设检验问题的 t 统计量是合适的检验统计量。如果误差是正态分布的，则这个检验服从所谓的“学生” t -分布，自由度为 $n - p' - 1$ 。

现在我们考虑，从不同的角度，但导致同一检验的另一途径。

再次假设第 i 个案例可能是个异常值。我们如下地进行。

1. 从数据中删除第 i 个案例。余下的 $n - 1$ 个案例用于拟合线

性模型。

2. 使用删除后的数据集估计 β 和 σ^2 。记这些估计为 $\hat{\beta}_{(i)}$ 和 $\hat{\sigma}_{(i)}^2$ ，以提醒我们第 i 个案例没有用于估计。 $\hat{\sigma}_{(i)}^2$ 有 $n-p'-1$ 的自由度。

3. 对于被删除的案例，计算其拟合值 $\tilde{y}_i = \mathbf{x}_i^T \hat{\beta}_{(i)}$ 。由于第 i 个案例没有用于估计， y_i 和 \tilde{y}_i 互相独立。 $y_i - \tilde{y}_i$ 的方差为

$$\text{var}(y_i - \tilde{y}_i) = \sigma^2 + \sigma^2 \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i \quad (5.16)$$

其中 $\mathbf{X}_{(i)}$ 由矩阵 \mathbf{X} 删除第 i 行得到。这个方差的估计由 (5.16) 式中的 σ^2 被 $\hat{\sigma}_{(i)}^2$ 代替后得到。

4. 现在，如果 y_i 不是一个异常值，则 $E(y_i - \tilde{y}_i) = 0$ 。假定正态误差，检验假设 $E(y_i - \tilde{y}_i) = 0$ 的“学生” t -检验由下式给出

$$t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{(i)} [1 + \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i]^{1/2}} \quad (5.17)$$

这个检验的自由度为 $n-p'-1$ 。

值得注意的是，不仅这两个途径导致同一检验，而且该检验与上节讨论的学生化内残差密切相关。借助于附录 5A.1， t_i 的计算可化为

$$t_i = r_i \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{1/2} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)} \sqrt{1-h_{ii}}} \quad (5.18)$$

t_i 称为学生化外残差，这是因为 σ^2 的估计没有用第 i 个案例。我们可以看到， r_i 和 t_i 互为单调增函数。

r_i ，或者 t_i 可由很多计算机软件包得到。不幸的是，这两个统计量的名字是不标准的。唯一的方法是查看使用的公式，以确实得到所要的变量。

异常值检验的显著性水平 如果研究者事先怀疑第 i 个案例是个异常值，则可将 t_i 与有适当自由度的中心 t 分布进行比较。通常，试验者对异常值无先验选择。如果我们检验具有最大 t_i 值的案例为异常值，我们事实上进行了 n 次显著性检验，对 n 个案例中的每一个都进行了一次检验。例如，假设没有异常值，且 $n=65$ ，

$p' = 4$ 。一个自由度为 60 的 t 统计量的绝对值超过 2.000 的概率是 0.05。然而，65 个独立 t -检验中最大的超过 2.000 的概率是 0.964。这清楚地表明，需要一个不同的临界值（当然，在我们的问题中，这些检验是相关的，上述计算只是一个说明*）我们用于寻找临界值的技术是基于邦弗伦尼 (Bonferroni) 不等式。 n 个水平为 α 的检验，至少有一个点被错误地认为是异常值的概率不大于 $n\alpha$ 。这项计算是保守的，因为邦弗伦尼不等式仅指出，65 个检验的最大值超过 2.00 的概率不大于 65 (0.05)。然而，这是大于 1 的数。不过，若选取临界值为 t 的 $(\alpha/n) \times 100\%$ 分位点，则给出的显著性水平不大于 $n(\alpha/n) = \alpha$ 。对每个检验，我们选择水平为 $0.05/65 = 0.00077$ ，以给出不大于 65 (0.00077) = 0.05 的总的水平。

在例 1.1 的 Forbes 数据中，由于案例 12 有大的残差，它被怀疑是异常值。为进行异常值检验，我们首先需要计算学生化残差。由 $\hat{e}_i = 1.36$ (表 1.5)， $\hat{\sigma} = 0.379$ (表 1.6) 以及据 (5.14) 计算得的 $h_{12,12} = 0.0639$ ，使用 (5.15) 可以计算得

$$r_{12} = \frac{1.3592}{0.379 \sqrt{1-0.0639}} = 3.7078$$

异常值检验为

$$t_{12} = 3.7078 \left(\frac{17-2-1}{17-2-3.7078^2} \right)^{\frac{1}{2}} = 12.40$$

这个统计量的自由度为 14。由 $p' = 2$ 和 $n = 17$ ，查表 E，得到水平 0.01 的检验的临界值为 4.41。由于 t_i 明显超过这个值，有理由认为这个案例是个异常值。

附加评注 有许多文献谈到处理异常值的方法，其中有两份文献。第一份为 Barnett 和 Lewis (1978) 写的，第二份为 Hawkins (1981) 写的及 Beckman 和 Cook (1983) 写的综述文章。异常值

* Miller (1981) 很好地讨论了这个及其它多重检验的问题。

模型的建立，除了用响应变量均值漂移的方法外，还可用其它的方法，例如用方差漂移的方法。另外，这里讨论的方法是对单个异常值的。如果一个数据集有多于一个的异常值，则这些案例会相互隐蔽，使异常值难于找出。Cook 和 Weisberg (1982a, p. 28) 将这里讨论的均值漂移模型推广到多个案例。Hawkins, Bradu 和 Kass (1984) 给出一个有希望的方法，用以遍查案例的各个子集，寻找远离中心的子集。Cook 和 Prescott (1981) 讨论了异常值检验的邦弗伦尼界限。他们发现对一次一个案例的方法，界限是非常精确的，但对多个案例的方法，精确性要差得多。Butler (1984) 给出用二阶的界限得到 p -值的更精确的计算方法。

这里采用的寻找异常值的方法是识别的方法：目标是发现异常值以用于进一步的研究。另外，我们可以考虑使用这样的统计方法，它能容忍或适应一定比例的坏或远离中心的数据。这就是“稳健的”统计方法得以发展的理由。在第十一章，我们将简要讨论这种统计方法。

由于异常值的特性与其具体内容有关，关于识别出异常值以后该干什么的问题说得很少。在某些问题中，发现异常值就是目标，而进一步的研究只对这些案例感兴趣。在另一些问题中，异常值被认为不是正在研究的活动的代表而被舍弃。还有一些问题中，异常值可以被校正。有时候有必要做两次分析，一次含有可疑异常值，而另一次不含有异常值。

5.3 案例的影响

案例分析的另一个方面是，试图了解每一个案例在拟合模型中的影响力或重要性。一般的想法是，当数据略有扰动时，对分析中的一个特定部分的变化进行研究。残差等统计量用于发现一个模型的问题。与之不同的是，影响分析是在认为模型正确时进行的研究。对给定的模型，研究结论对扰动的稳健性。一个非常

有用和重要的研究数据扰动的方法是，从数据中一次删除一个案例。然后通过比较全部数据的分析与删去一个案例后的数据的分析，研究每一单个案例的作用或影响。删去后在分析中引起大的变化的案例称为是有影响力的。

现在我们将上节所用的某些标记严格化。下标_(i)意味着“删除第 i 个案例后的”。所以，例如 $\hat{\beta}_{(i)}$ 为删除第 i 个案例后 β 的估计值， $X_{(i)}$ 为由 X 删除第 i 行后的 $(n-1) \times p'$ 矩阵，等等。特别地，有

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \quad (5.19)$$

作为一个例子，图 5.3 是燃料消耗数据中， $\hat{\beta}_{1,(i)}$ 关于 $\hat{\beta}_{2,(i)}$ 的图，其中自变量只取 $X_1 = TAX$ 和 $X_2 = DLIC$ 。标记为 Wyoming

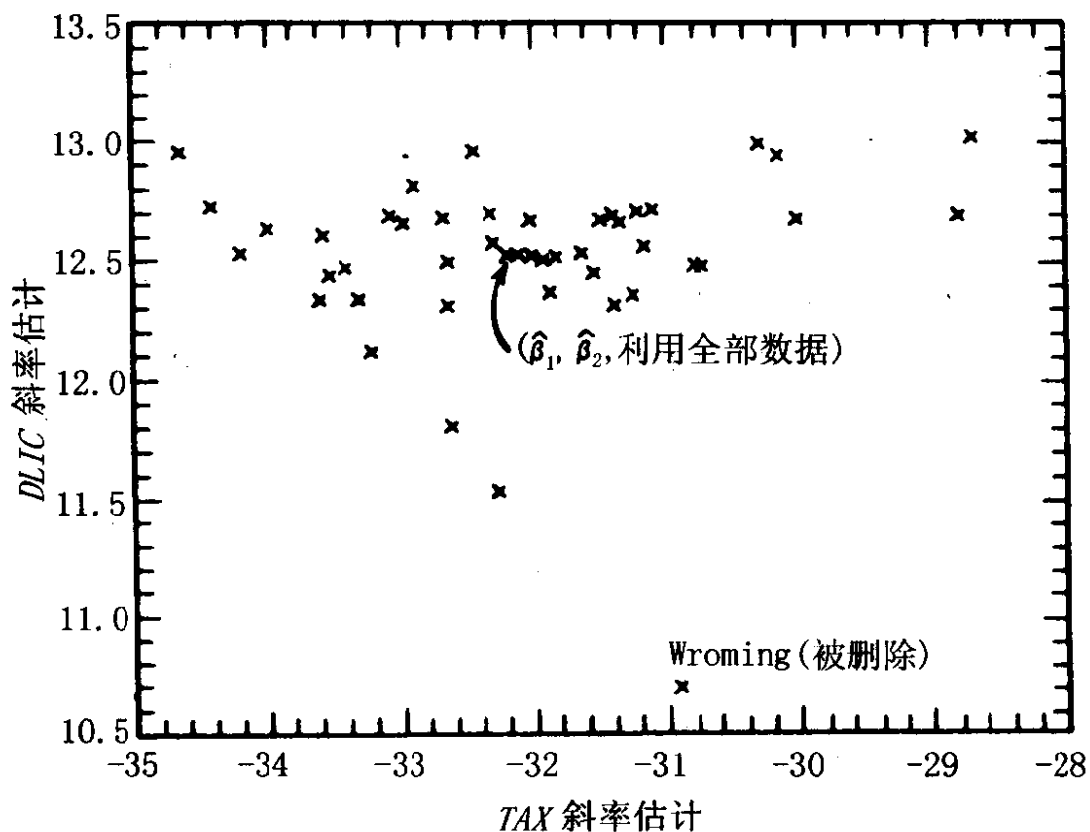


图 5.3 轮流删除每个案例得到的 (β_1, β_2) 的估计

(怀俄明州)的点对于，在数据集中删除怀俄明州后用余下的 47 个案例得到的 TAX 和 $DLIC$ 的斜率的估计。这里，没有给出截距的值，因为为了完全显示删除一个案例的影响，需要一个三维图。

由图 5.3 可以清楚地看到, 删除一个案例将引起估计的改变, 但对大多数案例, 这一改变是小的。使用全部数据的估计为 $(-32.075, 12.515)$ 。如果得克萨斯州被删除, 估计为 $(-33.933, 12.436)$ 。而在怀俄明州被删除后, 估计为 $(-30.933, 10.691)$ 。由图发现, 后一点离原先估计相当远。为了判断估计的变化是否足够大, 以致于已实质性地改变结论, 需要有一个衡量这些点间的距离的方法。

Cook 距离 通过比较 β 和 $\beta_{(i)}$, 我们可以测量影响。由于它们每一个都是 p' 维向量, 所以为了比较, 需要有一个将来自 p' 个元素中每一个的信息综合成单个数值的方法。文献中提出了若干方法, 但大多数方法, 至少对多元线性回归, 导致大致相同的信息。我们所用的方法由 Cook (1977) 提出。我们定义 Cook 距离 D_i 为

$$D_i = \frac{(\beta_{(i)} - \beta)^T (X^T X)^{-1} (\beta_{(i)} - \beta)}{p' \hat{\sigma}^2} \quad (5.20)$$

这一统计量有若干理想特性。首先, D_i 为某个常量的等值线是椭圆。它与置信椭圆具有相同的形状。其次, 等高线可被理解为, 从 $\beta_{(i)}$ 到 β 的距离。第三, 通过线性变换, 更改 X 的列, D_i 的值不变, 即 D_i 不受参数化的影响。最后, 如果我们如通常一样, 定义 $\hat{Y} = \bar{X}\beta$ 及 $\hat{Y}_{(i)} = X\beta_{(i)}$, 则 (5.20) 可被写成

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{p' \hat{\sigma}^2} \quad (5.21)$$

故 D_i 为 \hat{Y} 与 $\hat{Y}_{(i)}$ 间的普通欧几里得距离。 D_i 大的案例对 β 及对拟合值都有实质性影响, 删除它们将导致结论的重大改变。

D_i 的数量 D_i 值大的案例被删除后, 将引起分析的实质性改变。典型地, 具有最大 D_i 值的案例, 或在大的数据集合中, 前几个具有最大 D_i 值的案例, 是我们感兴趣的。通过与置信域的类比, 我们得到测定 D_i 的一个方法。如果 D_i 恰好等于自由度为 p' 和 $n - p'$ 的 F 分布的 $\alpha \times 100\%$ 分位点, 则删除第 i 个案例将把 β 估计

移至基于全部数据集的 $(1-\alpha) \times 100\%$ 置信域的边缘。大多数 F 分布的 50% 分位点接近于 1, 所以 $D_i=1$ 的一个值将把估计移至约 50% 的置信域的边缘。这是一个潜在的重要变化。如果最大的 D_i 明显小于 1, 则删除一个案例不会太多地改变对 β 的估计。为更仔细地研究一个案例的影响, 我们必须删除大的 D_i 的案例, 并重新计算分析, 看它的哪些方面确实地改变了。

计算 D_i 从 Cook 距离的导出, 并不清楚使用这些统计量是否便于计算, 然而, 附录 5A.1 描述的结果能用于计算 D_i 。它只用到一些更为熟悉的量。 D_i 的最简单的形式为

$$D_i = \frac{1}{p'} r_i^2 \left(\frac{h_{ii}}{1-h_{ii}} \right) \quad (5.22)$$

D_i 是第 i 个学生化残差 r_i 的平方与 h_{ii} 的一个单调函数的乘积。如果 p' 固定, D_i 的大小由两个不同的方面决定: 一个是 r_i 的大小, 另一个是位势 h_{ii} 。 r_i 是反映模型在第 i 个案例处拟合失真的随机变量, 而 h_{ii} 反映 x_i 相对于 \bar{x} 的位置。 D_i 的值大可能是由于 r_i 大, 或 h_{ii} 大, 或两者都大。

表 5.2 给出燃料消耗数据中三个州的案例统计量。这里, 仍只用 TAX 和 $DLIC$ 作为自变量。为从 \hat{e}_i 和 h_{ii} 计算 r_i 和 D_i , 我们还需知道 $\hat{\sigma}^2 = 5\,796.31$ 。对怀俄明州, 我们有

$$r_{40} = \frac{242.6}{\sqrt{5\,796.31 (1-0.0930)}} = 3.3453$$

$$D_{40} = \frac{(3.3453)^2}{3} \left(\frac{0.0930}{1-0.0930} \right) = 0.3826$$

具有最大位势的得克萨斯州有相对小的测量影响, 这是因为其学生化残差 r_i 很小。类似地, 怀俄明州是相对有影响力的, 因为其 r_i 取较大的值。看起来怀俄明州的人消耗的燃料比由 $DLIC$ 和 TAX 预计的更多。不过, 在这个例子中没有一个案例, 单独对系数的估计起相当大的影响。

一个完整的分析需要对每个案例, 考虑 D_i , r_i 和 h_{ii} , 或它们

的等价函数。

表 5.2 燃料消耗的数据中选择
的案例的统计量

州	e_i	h_{ii}	r_i	D_i
18ND	153.8	.0444	2.0659	.0661
37TX	-16.94	.2067	-0.2497	.0054
40WY	242.6	.0930	3.3453	.3826

例 5.3 老鼠数据

有人进行了一项试验以研究某一种药物在老鼠肝中聚集的量。随机选取 19 只老鼠，称过体重后施以少量乙醚麻醉剂，并服下这种药的口服剂量。由于人们认为，大的肝脏比小的肝脏会吸收较多的药物，一个动物服下的准确剂量近似为每千克体重服 40mg 药物（已知肝脏重量与体重密切相关）。在给定的一段时间以后，解剖每只老鼠，然后称肝脏重量，确定药物在肝脏中的百分比。

试验假设为，对于确定剂量的方法，肝脏中药物含量的百分比（ Y ）与体重（ X_1 ），肝重（ X_2 ）及相应的剂量（ X_3 ）没有什么关系。

数据及样本相关系数由表 5.3 和 5.4 给出。正如所料想的，响应变量与这些独立变量之间的样本相关系数都较小，而且响应变量关于任意一个独立变量的简单回归都是不显著的。这如表 5.5 所示，所有的 t -值都小于 1。然而 Y 关于 X_1 , X_2 和 X_3 的回归给出不同的而且矛盾的结论。两个独立变量， X_1 和 X_3 有显著的 t -检验值， $p < 0.05$ 都满足。这表明，把 X_1 和 X_3 结合在一起，可有效地预测 Y 。如果从模型中拿掉 X_2 ，出现相同的现象。到目前为止，分析仅只基于综合统计量。它可能导致结论：把剂量和老鼠体重结合在一起，它与响应变量有关。然而，剂量（近似）为老鼠体重的一个倍数，故至少以一阶近似，老鼠体重和剂量是在测量同一事物！

我们求助于案例分析，试图解决这一似是而非的问题。在表 5.6，列出模型中 Y 关于 X_1 , X_2 , X_3 的残差及有关的统计量。关于这一似是而非的情况，残差（ \hat{e}_i 或 r_i ）没有表出任何不寻常的特征或原因。例如， $|r_i|$ 都是小于 2 的，不存在明显的趋势或规律。然而， D_i 立即指出一个可能的原因。案例 3 的 $D_3 = 0.93$ ，但是其它案例的 D_i 都不大于 0.27。这表示案例 3 可能具有足够大的影响，能单独影响拟合，导致异常。值 $h_{33} = 0.85$ 表示，这一案

例的问题在于向量 X_3 与其它的不同。

表 5.3 田鼠数据

X_1 = 体重	X_2 = 肝重	X_3 = 剂量	Y
176	6.5	.88	.42
176	9.5	.88	.25
190	9.0	1.00	.56
176	8.9	.88	.23
200	7.2	1.00	.23
167	8.9	.83	.32
188	8.0	.94	.37
195	10.0	.98	.41
176	8.0	.88	.33
165	7.9	.84	.38
158	6.9	.80	.27
148	7.3	.74	.36
149	5.2	.75	.21
163	8.4	.81	.28
170	7.2	.85	.34
186	6.8	.94	.28
146	7.3	.73	.30
181	9.0	.90	.37
149	6.4	.75	.46

表 5.4 样本相关系数——田鼠数据

X_1 = 体重 (g)	1.000			
X_2 = 肝重 (g)	0.500	1.000		
X_3 = 相关剂量	0.990	0.490	1.000	
Y	0.151	0.203	0.228	1.000
	体重	肝重	剂量	Y

表 5.5 Y 关于各个自变量的回归 (括号中为 t -值)

系数	模 型 包 括			
	X_1	X_2	X_3	(X_1, X_2, X_3)
截距	0.196 (0.89)	0.220 (1.64)	0.133 (0.63)	0.266 (1.37)
β_1 (田鼠重量)	0.0008 (0.63)			-0.0212 (-2.66)
β_2 (肝重)		0.0147 (0.86)		0.0143 (0.83)
β_3 (剂量)			0.235 (0.96)	4.178 (2.74)

表 5.6 Y 关于 X_1 、 X_2 、 X_3 的所有 $n=19$ 个案例的残差

案例编号	y_i	\hat{e}_i	r_i	h_{ii}	D_i
1	.42	.124	1.77	.178	.17
2	.25	-.089	-1.27	.179	.09
3	.56	.024	.81	.851	.93
4	.23	-.101	-1.38	.108	.06
5	.23	-.068	-1.12	.392	.20
6	.32	.007	.10	.161	.00
7	.37	.057	.79	.137	.02
8	.41	.050	.74	.254	.05
9	.33	.012	.16	.067	.00
10	.38	-.003	-.04	.120	.00
11	.27	-.080	-.11	.120	.04
12	.36	.042	.60	.172	.02
13	.21	-.098	-1.54	.316	.27
14	.28	-.027	-.38	.131	.01
15	.34	.032	.43	.076	.00
16	.28	-.059	-.86	.217	.05
17	.30	-.018	-.26	.195	.00
18	.37	.061	.85	.149	.03
19	.46	.135	1.92	.178	.20

我们建议，删除案例 3 并重新计算回归。这些计算由表 5.7 给出。似是而非的问题被解决了，起初分析中发现的关系可被归因于第 3 个案例。

表 5.7 案例 3 删除后的回归

变 量	估 计	标准误	t-值
截距	0.3114	0.2051	1.52
X_1	-0.0078	0.0187	-0.42
X_2	0.0090	0.0187	0.48
X_3	1.4849	3.7173	0.40
$\hat{\sigma}^2 = 0.00612 \quad R^2 = 0.0211 \quad \text{d.f.} = 14$			

仔细的分析者现在应该准确了解，为什么第 3 个案例有这么大的影响。查看数据会发现，这个重 190 克的老鼠，接受的剂量为 1.000。这比根据规则它应该接受的要多（例如，重 195 克的 8 号老鼠接受的剂量为 0.98）。在第一次分析中，发生这一似是而非的问题，可能有若干原因：（1）案例 3 记录的剂量或体重有错，故这一案例应在研究中删除；或（2）在由 18 个点（不包括案例 3）确定的区域以外，第二次分析所得的回归，其拟合是不佳的。关于这一次试验，它有许多含意。加入一个明显地在不同的条件下选取的数据，导致了不同的结论。故剂量和老鼠重量的组合的选择，可以是随机的，对他们的其它组合可能是有关系的。这表示需要收集另外的数据，其剂量不是根据体重的一个常数比例，而是根据其它的规则决定的。

影响的其它度量 Cook 距离是通过类比于估计向量 $\hat{\beta}$ 的置信椭圆得到的。对影响的一个更一般的处理需要使用更基本的原理。Cook 和 Weisberg (1982a, 5.2 节) 通过在删除一个案例后，查看对数似然函数的变化，定义了关于影响的一类度量方法。当兴趣集中于对 β 的估计时，这与线性模型中的 Cook 距离是等价的。如果把兴趣集中于对 β 和 σ^2 两者的估计，则产生了一个不同的度量方法。在将影响的思想应用于其它问题中时，似然分析的方法同样是有用的。

依照 Cook 关于 D_i 的原来的推导，得到了其它一些关于影响

的独特的度量方法。例如 Belsley, Kun 和 Welsch (1980) 提出一个非常类似的统计量, 称为 DFFITS_i, 其定义如下:

$$(\text{DFFITS}_i)^2 = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X) (\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\sigma}_{(i)}^2} \quad (5.23)$$

这个统计量与 D_i 的差别在于, 这里没有标量因子 p' 以及用 $\hat{\sigma}_{(i)}^2$ 代替 $\hat{\sigma}^2$ 。Atkinson (1982) 提出, 在作图的过程中使用 (5.23) 的多重形式, 他称为“修改的 Cook 统计量”。通过比较 (Cook 和 Weisberg, 1982a, 第 4 章) 显示出, 一般来说, 所有这些影响的度量给出基本上相同的信息。

关于异常值问题, 那些有影响力的案例可能相互隐藏, 使得在一次检验中不能被找出。在某些问题中, 多重案例法可能是理想的。进一步讨论参见 Cook 和 Weisberg (1982a, 3.6 节) 及 Gray 和 Ling (1984)。

问 题

5.1 在一个 $n=54$, $p'=5$ 的回归问题中, 计算得 $\sigma=2.0$, 以及 4 个案例的统计量, 如下所示:

对这 4 个案例中的每一个, 计算 r_i , D_i 及 t_i 。对每一个都进行异常值检验。对每一个在分析中的影响作定量表述。

\hat{e}_i	h_{ij}
0.6325	0.9000
1.732	0.7500
9.000	0.2500
10.295	0.0185

5.2 在 2 个自变量的燃料消耗的例子中, 具有最大 r_i 的州是怀俄明州。

5.2.1 检验怀俄明州是一个异常值 (见表 5.2)。

5.2.2 在前面的分析中没有包括阿拉斯加和夏威夷这两个州。下表给出这两个州的 4 个自变量及因变量的数据。在第二章中由其它

48 个州拟合的模型是否适用于阿拉斯加和夏威夷？如果这两个州中的一个或两个都被包括进去，拟合的模型会有大的变化吗？

	<i>TAX</i>	<i>DLIC</i>	<i>INC</i>	<i>ROAD</i>	<i>FUEL</i>
Alaska	8.0	45.2	5.162	3.246	551
Hawaii	5.0	64.8	4.995	0.602	345

5.3 利用附录 2A.3 和问题 2.4 定义的 QR 分解，证明

$$H = QQ^T$$

从而，如果 q_i^T 是 Q 的第 i 行，则

$$h_{ii} = q_i^T q_i \quad h_{ij} = q_i^T q_j$$

这样，如果计算了 X 的 QR 分解，可以很容易地得到 h_{ii} 和 h_{ij} 。

5.4 令 U 为一个 $n \times 1$ 向量，第 1 个元素为 1，其它元素为 0。计算 U 关于一个 $n \times p'$ 满秩矩阵 X 的回归。通常，令 $H = X(X^T X)^{-1} X^T$ 。它是在 X 的列向量空间上的正交投影算子，其元素为 h_{ij} 。

5.4.1 证明： U 关于 X 的回归，其拟合值向量的元素为 h_{1j} , $j=1, 2, \dots, n$ 。

5.4.2 证明：残差向量的第一个元素为 $1-h_{11}$ ，其它元素为 $-h_{1j}$, $j>1$ 。

5.4.3 用 h_{ij} 表示学生化内残差。用这一方法，利用任何计算残差的回归程序，可以得到 h_{ij} 的值。

5.5 如果 $AB=BA=0$ ，则称这二个 $n \times n$ 矩阵 A 和 B ，互相正交。证明： H 和 $I-H$ 是正交的。由此证明：有截距的回归模型，残差 $\hat{e} = (I-H)Y$ 与拟合值 $\hat{Y} = HY$ 之间的样本相关系数恰好为 0。或等价地， \hat{e} 关于 \hat{Y} 的回归，其斜率为 0。

5.6 矩阵 $(X_{(i)}^T X_{(i)})$ 可被写为 $(X_{(i)}^T X_{(i)}) = X^T X - X_i X_i^T$ ，其中 x_i^T 为 X 的第 i 行。由此证明，(5A.1) 成立。

5.7 去掉第 i 个案例后估计 β ， $y_i - x_i^T \hat{\beta}_{(i)}$ 是第 i 个案例的残差。由 (5A.1) 证明

$$y_i - x_i^T \hat{\beta}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}}.$$

这个量被称为预测或 PRESS 残差。它将在第 8、第 9 章中用到。

5.8 用 (5A.1) 验证 (5.22)。

- 5.9 假设兴趣在 β^* , 而非 β , 其中 β^* 为不包含截距的参数向量。基于 (4.26), 类似于 Cook 的 D_i , 定义一个距离量度 D_i^* 。证明 (Cook, 1979)

$$D_i^* = \frac{r_i^2}{p} \left(\frac{h_{ii} - 1/n}{1 - h_{ii}} \right)$$

- 5.10 利用 (5A.1) 证明 (5.18), 即 r_i 和 t_i 的关系。另外, 证明 5.2 节中描述的检验异常值的两个途径, 导出相同的统计量 t_i 。提示: 你将需要证明

$$\mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i = \frac{h_{ii}}{1 - h_{ii}}$$

- 5.11 证明 (5A.4)。

6

诊断Ⅱ：症状与治疗

在第五章中定义的统计量，残差 \hat{e}_i ，学生化残差 r_i 和 t_i ，位势 h_{ii} 及 Cook 距离 D_i ，是我们研究关于线性模型所作的假设的基础。使用它们的诊断方法为本章讨论的主题。

Box (1980) 给出了拟合统计模型的一个有用模式，图 6.1 给

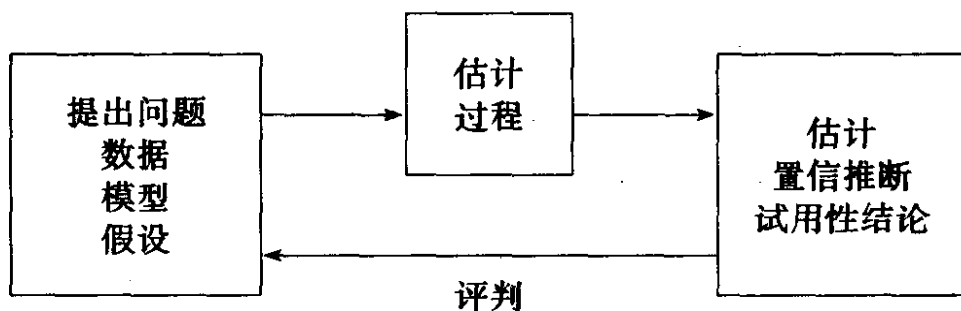


图 6.1 建模模式 (Box, 1980 年后)

出了其修改后的形式的轮廓。左边的正方形框表示：提出一个有兴趣的问题以供研究，选择一个模型，作假设并收集数据。图中上面的箭头表示在得到估计的过程中，统计量的更为传统的作用。估计是在正方形框表达的内容为正确的前提下进行的。通常使用如极大似计那样的一般的方法。所得的结果为拟合的模型，检验及推断等等位于右边正方形框中的内容。

诊断方法对应于底下的箭头，标记为评判。在分析的这个阶段，我们在右边正方形框内拟合模型的条件下，给出关于左边正方形框内假设的信息。所使用的方法自然地基于类似残差的量，因为这些量含有拟合失真的信息。经常地，诊断会提出对模型或假设的修改的建议，并再次进入图中迭代循环。

可以提出设计有用的诊断方法的若干准则（Weisberg, 1983）。首先，必须至少是近似地知道一个诊断过程的表现，包括在正确模型以及在有一个假设被改变的模型下的表现。这表示诊断方法必须为特定目的而设计。一般化的多种用途的方法，如残差对拟合值图，其效果较差。一类重要的诊断方法是基于模型扩展（Cox, 1977），即拟合一个比原先考虑的要大的线性模型。它有附加参数。当关于模型的一个假设正确时，附加参数取特定的值。然后，由一个关于附加参数的检验，得到一个诊断方法。在本章中，我们将多次使用模型扩展。其次，诊断必须易于计算。对线性回归，许多诊断只使用基本统计量及通常适用于残差的回归计算。第三，诊断必须用图表示，或者用与图等价的量表示。2.4节的附加变量图常是有用的。最后，诊断必须向分析者提出治疗行为的建议。最后一个要求经常只能近似地满足。在我们考虑特定的模型时，将看到这一点。

6.1 散点图

最常见的诊断是散点图。它或者是数据的散点图，或者是残差等导出的统计量的散点图。其次是矩阵图（Chambers, Cleveland, Kleiner 和 Tukey, 1983）。图 6.2 为燃料消耗数据的矩阵图。这个图是用于检验所有的二维图，即基于从多维数据集合中取出的一对对变量而得到的所有的二维图。在有 5 个变量的燃料消耗问题中，有 10 对变量。如果认为 *FUEL* 关于 *INC* 的图不同于 *INC* 关于 *FUEL* 的图，则可能有 20 个图。这些图在图 6.2 中有规

律地排列着。所有在同一行中的图有相同的 Y 轴变量；所有在同一列中的图有相同的 X 轴变量。例如，第一行的最后一个图是 TAX 关于 $FUEL$ ，最后一行的第二个图是 $FUEL$ 关于 TAX ，最后一行的最后一个图是 $FUEL$ 关于 $ROAD$ 。图 6.2 中，在这些图

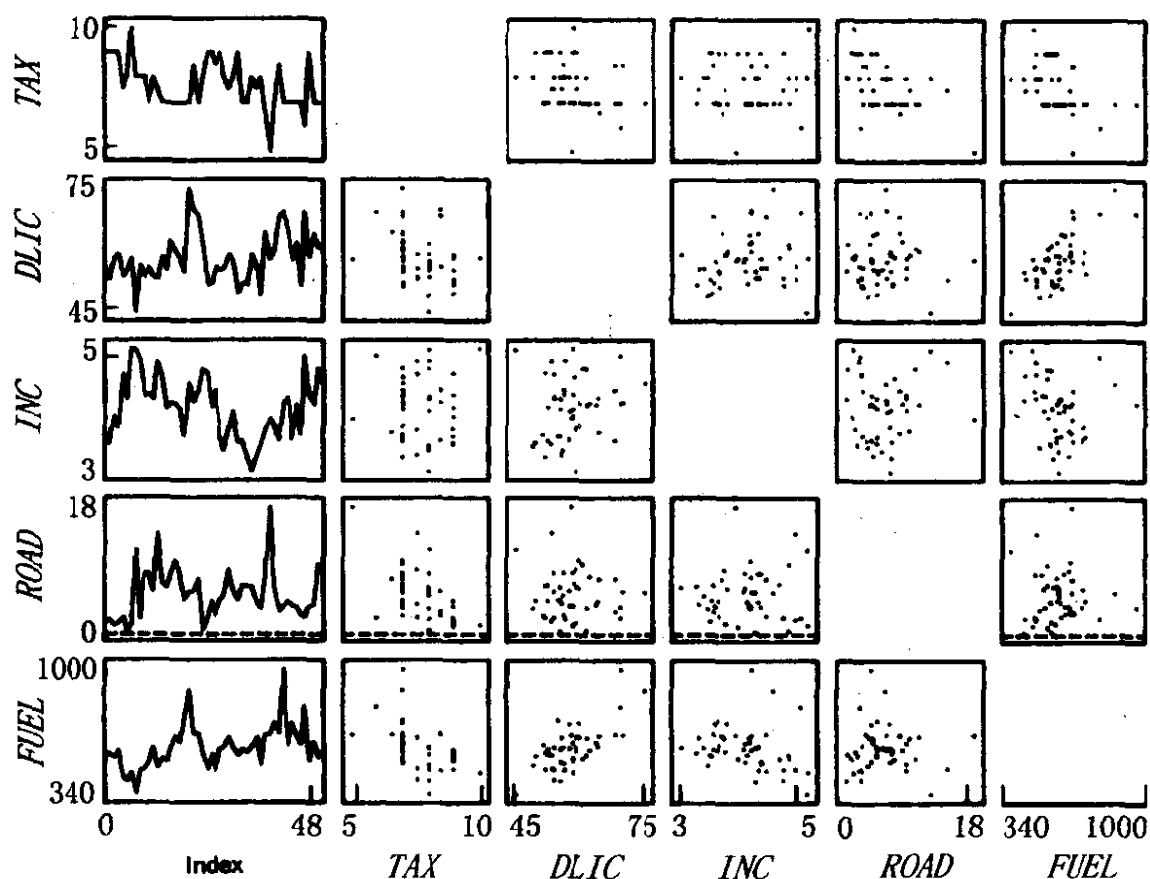


图 6.2 燃料消耗数据的矩阵图

的左边有一列索引图。其 Y 轴上为对应的变量， X 轴上为数据序列号。索引图可用于代替数据列表。在小的问题中，这是容易做到的。虽然在矩阵图中单个图是很小的，但保持着足够的分辨性，可以表示二维关系的主要特性。总的来说，这个图可以看作是相关矩阵的图表示。图 6.2 显示 $FUEL$ 和 $DLIC$ 是相当密切联系的，而在 $FUEL$ 和 INC 的关系中，看起来有三个异常点。

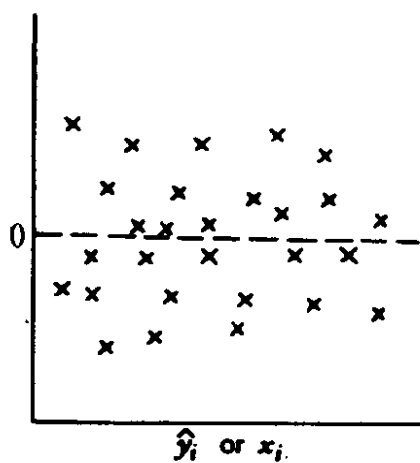
残差图 最早使用的单个诊断方法是残差，或学生化残差关

于拟合值，或一个自变量的图。由这个图的形状，许多问题都能潜在地被诊断，所以它成了一种“万能图”式的诊断。由于 r_i 或 t_i 和 \hat{y}_i 几乎是不相关的 (\hat{e}_i 和 \hat{y}_i 是精确地不相关的)，在作残差关于拟合值的图时，我们总可以得到一个斜率接近于 0，而且只要模型中有截距，在 0 周围散布的图。图 6.3 给出 8 个理想化的残差关于拟合值的图。第一个图是在设定的模型正确时，我们所期望的图。它是一个零图。这个图是一群无规律的点。非常数方差可以由图 6.3 (b) 至 6.3 (d) 中任一个指出。图 6.3 (b) 象一个朝右的话筒，它表示当 X -轴上的量增加时，方差随之增加。这种情况经常在当一个固有为正的响应变量在一个大范围内，如从 0 附近到成千变化时发生。这是因为大的值通常比小的值有更大的变化余地。图 6.3 (c) 朝左的话筒表示方差随 X 轴上的量的增加而减少。图 6.3 (d) 的双弓在响应变量被限制于一个最小值与一个最大值，如从百分比的 0 到 100 之间时发生。大的和小的百分比通常比接近 50% 的百分比的变化要小。

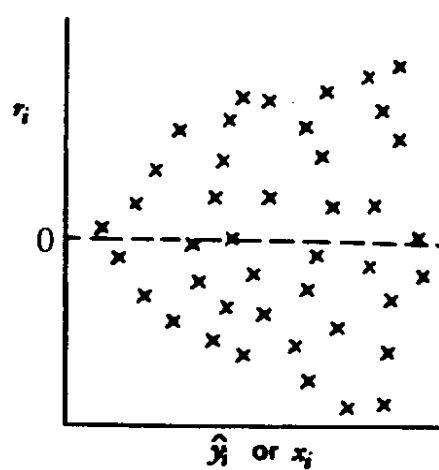
图 6.3 (e) 和 6.3 (f) 表示，回归是 X 轴上量的非线性函数。这是因为线性趋势排除后，留下了非线性曲线趋势。这通常需要对数据，或响应变量或自变量进行变换，或使用非线性模型。图 6.3 (g) 和 6.3 (h) 被认为非常数方差和非线性两个症状的综合。

这些图中远离 y 轴上零点的孤立的点可能是异常值。如果用 r_i 或 t_i 作图，每 100 个作图的值中，约有 5 个超过 ± 2 ，约有 1 个超过 ± 3 。

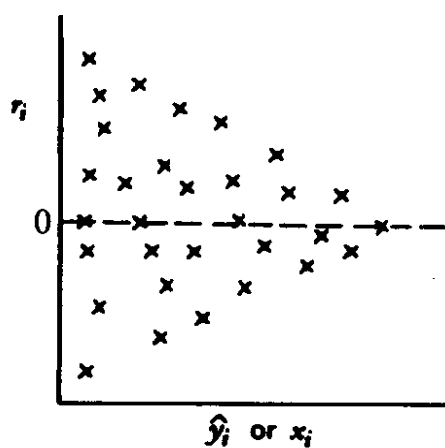
遗憾的是，这些理想化的图掩盖了非常重要的一点，在实际数据集合中，事情的真实状况这么清楚是少有的。考虑图 6.4 的残差图。产生该图的数据来自于 Cook 和 Weisberg (1982a) 的例 2.3.4；亦可见 Cook 和 Weisberg (1982b)。这是一个 r_i 关于 \hat{y}_i 的图。由于着眼点的不同，前面所描述的几个或全部症状都可在这个图中找到。右上角的点可能是一个异常值，也可能表示一个朝



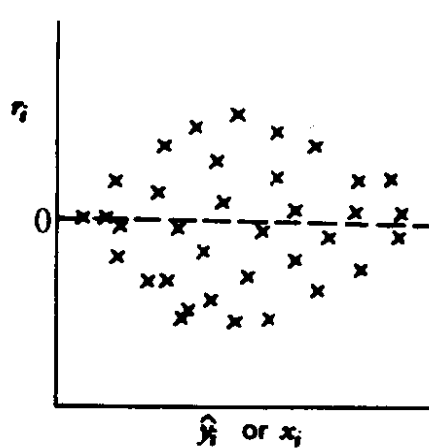
(a) 零图



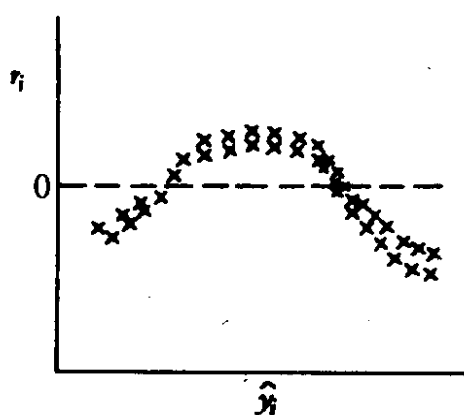
(b) 朝右话筒



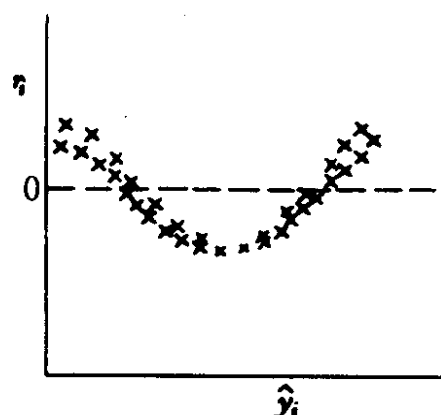
(c) 朝左话筒



(d) 双凸弓形

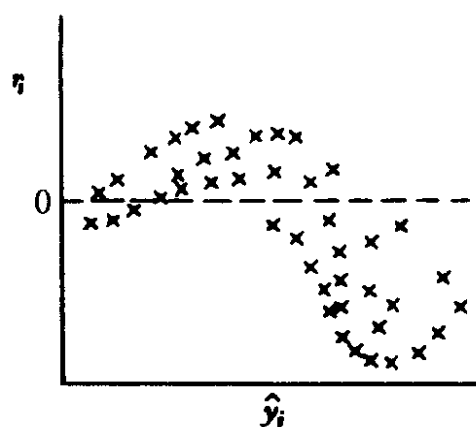


(e) 非线性

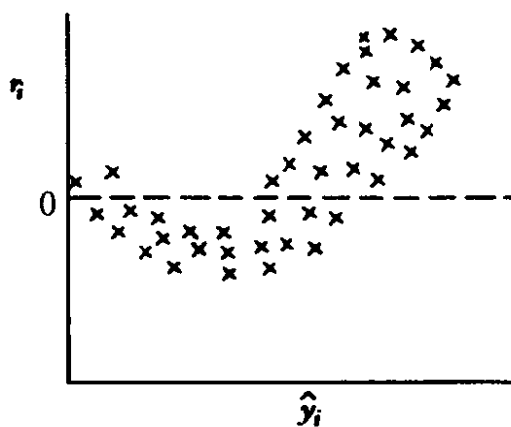


(f) 非线性

图 6.3 残差图



(g) 非线性和非常数方差



(h) 非线性和非常数方差

图 6.3 残差图 (续)

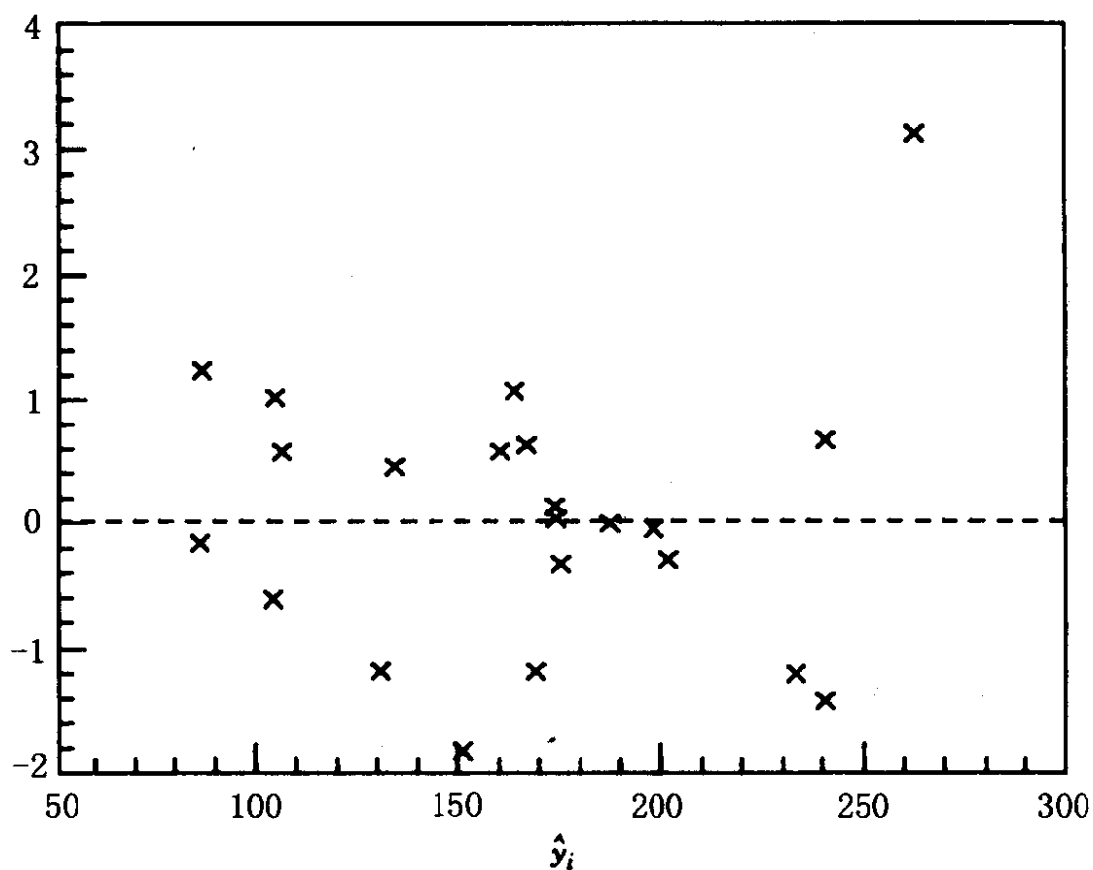


图 6.4 一个模棱两可的残差图

右的话筒。这一形状还使人联想到图 6.3 (h)，表示非线性或可能是非常数方差。总之，这个图不是很有用。

关于自变量或其它量的图 如同与拟合值一样, r_i 与每个自变量也几乎是不相关的, 故关于自变量的图与关于拟合值的图有相同的值和限制。如果模型的欠缺同其它量如案例编号或时间有关, 则关于这些量的图可能也是有用的。用残差或其它统计量的索引图代替统计量的简单列表也常是有用的。这是因为从一个图中往往比从一系列数据中, 能更快地找到大的值。

6.2 非常数方差

首先研究的假设是, $\text{var}(e_i) = \sigma^2$ 对数据的所有案例都成立。在许多问题中, 由于方差可能和响应变量, 和一个或多个自变量, 也可能和其它因素, 如时间或物理次序等有关, 所以这一假设是有疑问的。

如果诊断出非常数方差, 但准确的方差未知, 我们可以考虑两种处理办法。首先, 可以使用加权最小二乘法, 权值根据经验选定。有时可以从理论上证明, 权值为单个自变量的简单函数, 例如 $\text{var}(e_i) = \sigma^2 x_{1i}$, 其中 $x_{1i} > 0$ 。如果可以得到有重复的大样本, 则组内方差可用于提供近似的权值。然而, 一般地不推荐使用将经验权值取为由普通最小二乘法得到的 \hat{y}_i 或 \hat{e}_i 的函数, 除非用非标准方法来估计方差。例子参见 Holt 和 Scott (1981)。

第二种方法是通过方差稳定化变换, 对响应变量 Y 进行变换。对 $\text{var}(e_i)$ 与一个恒取正值的响应变量之间的几乎所有的关系, 都可以得到一个合适的方差稳定化变换 (技术细节参见 Scheffe' 1959, 第 9 章)。在表 6.1 中列出了常用的方差稳定化变换。当方差随着响应变量增加或减少时, $Y^{1/2}$, $\log(Y)$ 和 $1/Y$ 都是合适的变换, 但每一个变换都比前者更严厉。平方根变换是相对温和的。通常, 在计数误差的模型中, 如果 y_i 服从泊松分布, 则平方根变换是最合适的。对数变换是最常用的。它与对数的底是没有关系的。当误差标准差是响应变量的一个百分比, $[\text{var}(e_i)]^{1/2}$

$\propto E(y_i)$, 例如响应变量的 $\pm 10\%$, 而非 ± 10 个单位时, 对数变换是合适的。

表 6.1 常用的方差稳定化变换

变 换	使用条件	评 注
\sqrt{Y}	$\text{var}(e_i) \propto E(Y_i)$	其理论基础为泊松分布的计数。
$\sqrt{Y} + \sqrt{Y+1}$	同上	用于某些 Y_i 为 0 或很小的情况; 这称为 Freeman-Tukey(1950)变换。
$\log Y$	$\text{var}(e_i) \propto [E(Y_i)]^2$	这个变换很常见; 在 Y 的值域很大, 例如从 1 到数千时, 它是一个很好的选择; 所有的 Y_i 必须严格为正
$\log(Y+1)$	同上	用于某些案例 $Y_i=0$ 的情况
$1/Y$	$\text{var}(e_i) \propto [E(Y_i)]^4$	适用于响应变量值都集中在零附近, 但是在明显下降的数中, 确实会出现大的响应变量值; 例如, 如果响应变量是一种治疗或一种药物的潜伏或反应时间, 某些病人可能迅速反应, 而少数病人要化长得多的时间; 逆变换将每个反应的时间转化为反应的速度, 即每个单位时间的反应; 所有的 Y_i 必须是正的。
$1/(Y+1)$	同上	用于某些案例 $Y_i=0$ 的情况
$\sin^{-1}(\sqrt{Y})$	$\text{var}(e_i) \propto E(Y_i)(1-E(Y_i))$	用于二项比例($0 \leq Y_i \leq 1$)

如果响应变量是直到某一事件发生的时间，如完成一项任务的时间，或直到治愈的时间，则常使用倒数或逆变换。它将每个事件的时间转化为每单位时间的比率。有时为避免一个很小的数，可以将变换了的量乘以一个常数。比率能提供一个自然度量尺度。

如果响应变量有 0 或负值时，由于变换，例如对数变换是没有意义的，出现了技术性问题。如表所示，解决办法之一是，采用变换 $[Y + (\text{一个小的常数})]$ ，常数可取为 1。在数据中 Y 都较小时，这可能不是一个好办法，但当 Y 偶而较大时，这是一条合理的权宜之计。

非常数方差的诊断 Cook 和 Weisberg (1983) 提出了一个诊断程序，用于检验非常数方差的假设。其基本思想是将常数方差假设转化为一个可检验的参数假设。在方差不是常数的时候，我们将设定它的形式。

现在假设 $\text{var}(e_i)$ 依赖于一个未知向量参数 λ 和一个已知向量 z_i 。对每个 i ， z_i 都可能不同。例如，如果 $z_i = y_i$ ，则方差依赖于响应变量。类似地， z_i 可能是自变量向量 x_i 。这时，残差方差依赖于所有的自变量。 z_i 也可能是自变量或其它量的子集，例如时间或研究单位的空间次序等等。给定 z_i ，我们假设

$$\text{var}(e_i) = \sigma^2[\exp(\lambda^T z_i)] \quad (6.1)$$

这一复杂的形式表示：(1) 对所有 z_i ， $\text{var}(e_i) > 0$ ；(2) 方差仅通过 $\lambda^T z_i$ 依赖于 z_i 和 λ ；(3) 关于 z_i 的每一个分量， $\text{var}(e_i)$ 是单调的，或者上升，或者下降；(4) 如果 $\lambda = 0$ ，则对所有的 i 有 $\text{var}(e_i) = \sigma^2$ 。Chen (1983) 得到的结论认为，只要四个条件满足，这里描述的检验对 (6.1) 所用的准确函数形式不是非常敏感。

给定 e_i 为独立且正态分布时，使用标准回归软件， $\lambda = 0$ 的得分 (score) 检验的计算是特别的容易。检验按下面的步骤执行：

1. 计算模型中 Y 关于所有 X 的回归，并保存残差 \hat{e}_i 。
2. 计算比例平方残差 u_i ，其定义为 $u_i = \hat{e}_i^2 / \tilde{\sigma}^2$ ，其中 $\tilde{\sigma}^2 = \sum \hat{e}_i^2 / n$ 是 σ^2 的极大似然估计。它与通常 σ^2 的估计的不同只是除数

为 n 而不是 $n-p'$ 。

3. 计算 u_i 关于 z_i 的包含截距的回归。计算这个回归的 SS_{reg} 。不包含截距, 如果每个 z_i 有 q 个分量, 则 SS_{reg} 的自由度为 q 。如果方差是响应变量 y_i 的函数, 则计算 u_i 关于拟合值 \hat{y}_i 的回归, 这一回归的 SS_{reg} 有 1 个自由度, 其中 \hat{y}_i 是 y 关于 x 的回归的拟合值。

4. 计算得分检验, $S=SS_{\text{reg}}/2$ 。在假设 $\lambda=0$ 下, S 的渐近分布为 $\chi^2(q)$ 。该检验的 p -值可以通过将 S 与 $\chi^2(q)$ 分布比较得到。如果 $\lambda \neq 0$, 则 S 的值将会大。大的 S 值给出拒绝常数方差假设的证据。

如果 $q=1$, 方差依赖于单个变量 z_i , 检验的图形的等价形式是, 作 r_i^2 关于 $(1-h_{ii})z_i$ 的图, 其中 r_i 与 h_{ii} 都是从 Y 关于 X 的回归计算中得到的。对许多问题, 作一个关于 z_i 的图更简单, 并不会丢失太多的信息。图中的楔形形状表示非常数方差。不过仍然需要计算统计量, 这是因为如果 X 轴上点的密度不均匀, 图中的楔形会变得模糊。如果 $q>1$, 可以作 r_i^2 关于 $(1-h_{ii})\hat{\lambda}^T z_i$ 的图, 其中 $\hat{\lambda}^T z_i$ 是在第三步, 回归中获得的拟合值。

例 6.1 雪鹅

在第四章的问题 4.6, 研究了在加拿大哈得逊湾地区雪鹅的 Y =照片计数值与 X =观察者计数值之间的关系。用第一个观测者, 我们现在可以来考察 Y 对 X 的简单回归模型的常数残差方差问题。

表 6.2 (a) 给出 Y 关于 X_1 的回归分析简要。虽然其中没有给出关于模型拟合失真的信息, 这些主要回归量指出 Y 与 X_1 有相当牢固的联系。表 6.2 (b) 是 u_i 关于 X_1 的回归的方差分析表。非常数方差的得分检验为 $S = \frac{1}{2}SS_{\text{reg}} = \left(\frac{1}{2}\right)162.83 = 81.41$ 。将 S 与自由度为 1 的 χ^2 分布进行比较, 则有一个极小的 p -值。常数残差方差的假设是站不住脚的。非常数方差几乎为肯定的。我们现在必须处理这个问题。问题 4.6.3 和 4.6.4 给出了两个处理程序。

表 6.2

(a) 第一个观测者所得的雪鹅数据的回归分析简要			
变量	估计	标准误	t-值
截距	26.65	8.61	3.09
X	0.88	0.08	11.37
$\sigma^2 = 1971.87$, $R^2 = 0.750$, d. f. = 43			
(b) 对雪鹅数据 u_i 关于 X_1 的方差分析			
来源	d. f.	SS	MS
回归	1	162.8264	162.8264
残差	43	137.8813	3.206541

例 6.2 嗅探器数据 (John Rice)

当向罐子压入汽油时, 烃气被赶出罐子, 溢漏到空气中。为减少这一明显的空气污染源, 安装一个设备以回收气体。在检验这些气体的回收系统中, 溢出的气体量是不能测量的, 但是“嗅探器”可以测出是否有气体溢漏。另外, 被回收的量是可以测量的。为估计系统的效率, 必须使用某种方法以估计溢出的气体的总量。为此, 在实验室进行一系列可精心控制的试验能测出溢出的气体的量。和建模有关的四个变量如下。

X_1 = 罐子的初始温度 ($^{\circ}\text{F}$)

X_2 = 注入汽油的温度 ($^{\circ}\text{F}$)

X_3 = 罐中的最初气压 (psi)

X_4 = 注入汽油的气体压力 (psi)

在试验中, 这些条件是变化的。溢出烃气的量 Y 以克为单位。32 个回合的试验数据列于表 6.3。

我们将研究非常数方差的可能性。图 6.5 (a) 是通常的 Y 关于 X_1, X_2, X_3, X_4 回归的 r_i 关于 \hat{y}_i 的残差图。由于这图还很不完善, 它没有显示出考虑非常数方差的假设的必要性。表 6.4 给出若干个非常数方差得分检验的结果。在计算时, 每一个选择不同的 z_i 。对于给出的 z_i 的各种选择, 每一个这样的检验仅是 u_i 关于 z_i 的回归平方和的一半。显然, 虽然非常数方差不是 X_1 或 X_4 各自的函数, 但它看起来可能是 X_1 与 X_4 的联合函数。取 $z = (X_1, X_4)$, 则 $S = 9.28$ 。与 $\chi^2(2)$ 相比较, 得 p -值 = 0.01。图 6.5 (b) 和 6.5

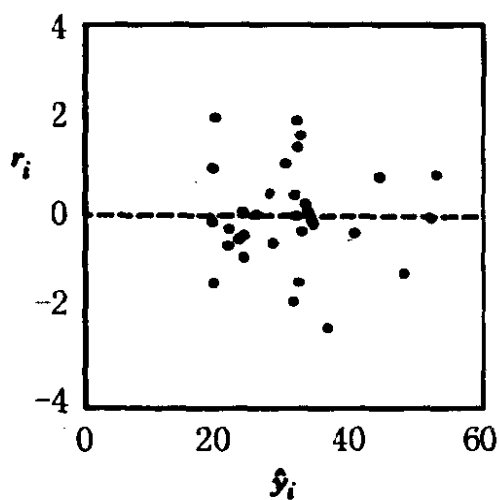
(c) 分别是 $z = \hat{y}_i$ 与 $z = X_1$ 时的非常数方差图。这两个图中都没有楔形形状特征。图 6.5 (d) 给出 $z = (X_1, X_4)$ 的图。X 轴上的值为 u_i 关于 X_1, X_4 回归的拟合值的 $(1 - h_{ii})$ 倍。在这个图上, 楔形形状是明显的。

表 6.3 烃气数据

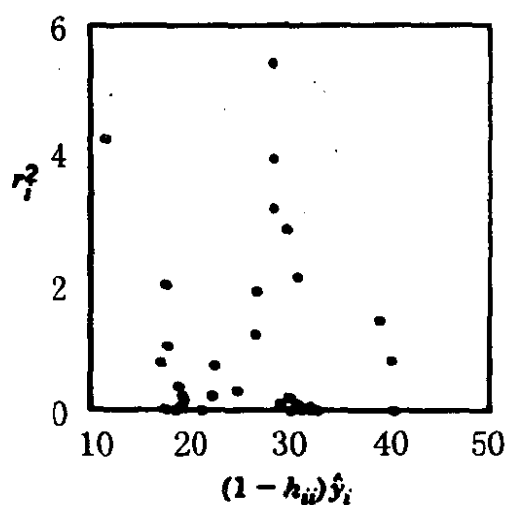
X_1	X_2	X_3	X_4	Y	X_1	X_2	X_3	X_4	Y
33.	53.	3.32	3.42	29.	90.	64.	7.32	6.70	40.
31.	36.	3.10	3.26	24.	90.	60.	7.32	7.20	46.
33.	51.	3.18	3.18	26.	92.	92.	7.45	7.45	55.
37.	51.	3.39	3.08	22.	91.	92.	7.27	7.26	52.
36.	54.	3.20	3.41	27.	61.	62.	3.91	4.08	29.
35.	35.	3.03	3.03	21.	59.	42.	3.75	3.45	22.
59.	56.	4.78	4.57	33.	88.	65.	6.48	5.80	31.
60.	60.	4.72	4.72	34.	91.	89.	6.70	6.60	45.
59.	60.	4.60	4.41	32.	63.	62.	4.30	4.30	37.
60.	606	4.53	4.53	34.	60.	61.	4.02	4.10	37.
34.	35.	2.90	2.95	20.	60.	62.	4.02	3.89	33.
60.	59.	4.40	4.36	36.	59.	62.	3.98	4.02	27.
60.	62.	4.31	4.42	34.	59.	62.	4.39	4.53	34.
60.	36.	4.27	3.94	23.	37.	35.	2.75	2.64	19.
62.	38.	4.41	3.49	24.	35.	35.	2.59	2.59	16.
62.	61.	4.39	4.39	32.	37.	37.	2.73	2.59	22.

表 6.4 得分检验, 烃气数据

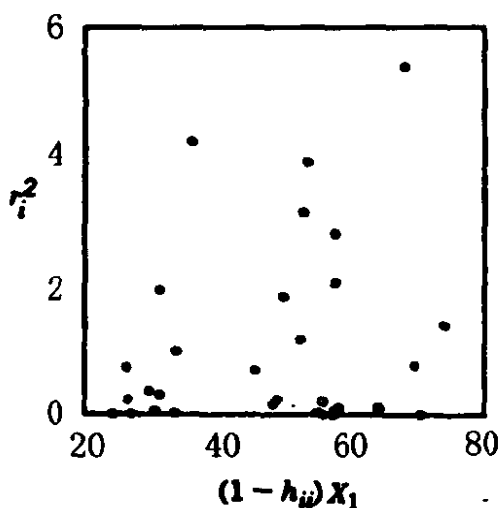
对 z_i 的选择	d. f.	$n=32$		$n=125$	
		S	p -值	S	p -值
X_1	1	1.40	.24	9.46	.002
X_4	1	0.01	.92	5.00	.019
X_1, X_4	2	9.28	.01	11.76	.003
X_1, X_2, X_3, X_4	4	10.30	.04	13.76	.008
拟合值	1	.00	.99	9.70	.002



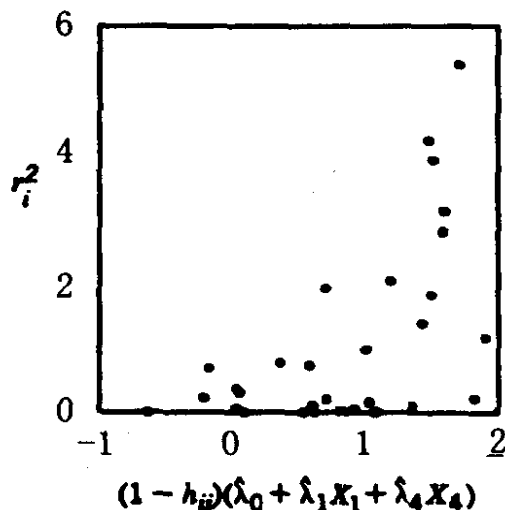
(a) 残差图



(b) r_i^2 对 $(1-h_{ii}) \hat{y}_i$



(c) r_i^2 对 $(1-h_{ii}) \hat{x}_1$



(d) r_i^2 对 $(1-h_{ii}) \hat{\lambda}_0 + \hat{\lambda}_1 X_1 + \hat{\lambda}_4 X_4$

图 6.5 烃气数据

产生这些数据的试验共有 125 个案例，其中在表 6.3 中列出的 32 个案例在本书的第一版中被作为家庭作业。我们希望对完整的数据集，拟合非常数方差模型，从而作进一步的研究，研究方差是否仍依赖于 X_1 和 X_4 两者的联合。表 6.4 的后两列给出了这个检验。使用更多的数据，从而是更有力的检验，从而可以明显地看出，非常数方差是关于单个 X_1 或 X_4 的，还是关于两者的联合，或者是拟合值的一个函数。

图 6.6 是对所有 125 个案例的 r_i^2 关于 $(1-h_{ii}) X_1$ 的图。和图的两侧的

点集相比较,在中央的点集有更多取较大值的点,所以图中看不出楔形形状。但这只是一个假象,这是因为在中央的点集里有较多的点,单是这一点便能导致在中央有较明显的变动。具有1个自由度的检验统计量 $S=9.46$ 校准了本图。我们可以看到,作为一个好的近似, X_1 只取3个值,而对这些值中的每一个,其方差是不同的。增大样本容量似乎使明显的非常数方差得到简化:残差方差随 X_1 增加。

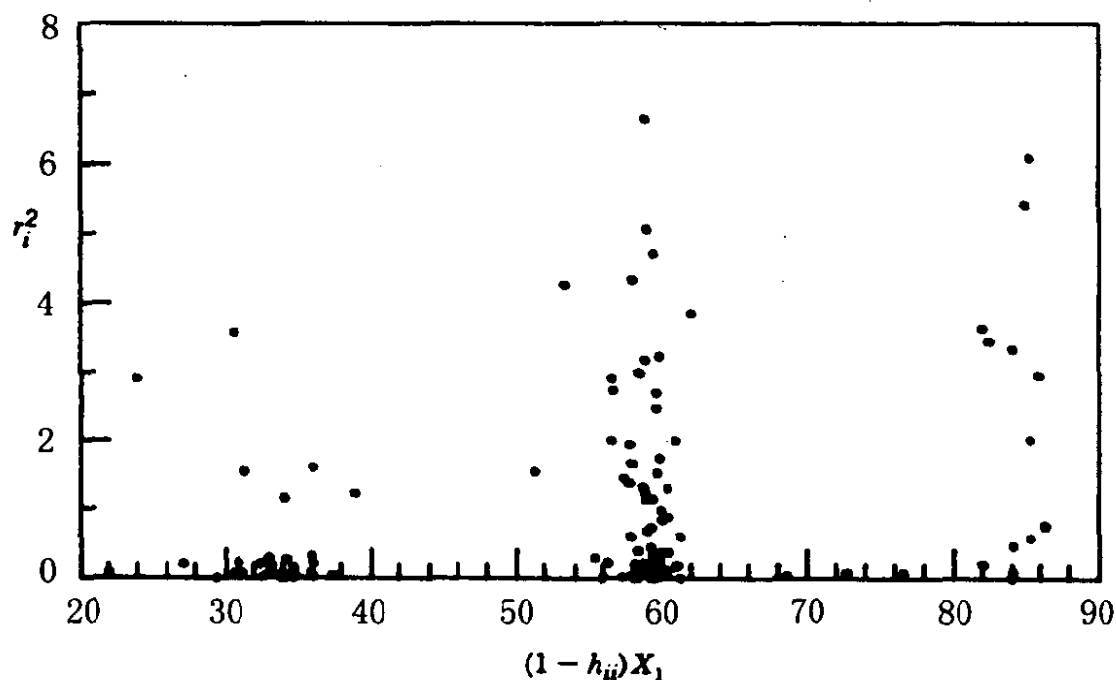


图 6.6 烃气数据 ($n=125$)

附加评注 有某些程序,可以更方便地计算得分检验,步骤如下:(1)由 Y 关于 X 的回归计算残差 \hat{e}_i ;令 $\hat{\sigma}^2$ 为这一回归的通常的残差均方。(2)对感兴趣的 z ,计算 \hat{e}_i^2 关于 z 的回归,并令 $SS_{\text{reg}}(z)$ 为所得的回归平方和。(3)计算 $S = \frac{1}{2} SS_{\text{reg}}(z) / [(n - p') \hat{\sigma}^2 / n]^2$ 。

6.3 非线性

尽管并非所有的响应变量与一组自变量之间的关系都是线性的,但是线性模型有着比一开始就明显表现出来的更广泛的应用。

尽管在自变量的整个范围内，函数关系可能是非线性的，但在一个限定的区域内，线性近似可能是一个合适的模型。不同的限定区域可能要求不同的近似线性模型。这样的模型在自变量的限制区域内有参考作用，而在有关区域之外，它们可能是没有意义的。

另外，有时可以找到对数据的合适变换，它使得一个非线性模型可以用一个线性模型来近似。例如，响应变量与单个自变量之间的真实关系由指数曲线给出，

$$Y = \alpha X^\beta$$

对于固定 $\alpha=1$ 和取不同值的 β ，这一族中的某些成员在图 6.7 给出。这一形式通过求对数而线性化，

$$\log Y = \log \alpha + \beta \log X \quad (6.2)$$

$\log Y$ 关于 $\log X$ 的回归是线性的。然而，我们必须注意误差。将形式为“响应变量的百分之 k ”的乘法型误差与 (6.2) 式结合，则有

$$Y = (\alpha X^\beta)(e)$$

取对数将模型变为有附加误差的线性模型

$$\log Y = \log(\alpha) + \beta \log(X) + \log(e)$$

如果 $\log(e)$ 的期望值为 0，有常数方差，则可以使用通常的分析方法。

表 6.5 列出常用的可线性化的形式，及达到线性化所需要作的变换。如果其中任一形式在理论背景下是合理的，则要使用给出的变换。当然，变换并不总是从一个理论模型中选出的，但是用数据可以给出一个合理的形式。

并非所有函数都可线性化，在某些情况中也不希望线性化。例如，两个指数的和

$$Y = \alpha_1 e^{\beta_1 X_1} + \alpha_2 e^{\beta_2 X_2}$$

是不可线性化的。逻辑斯谛函数

$$Y = \frac{e^{-(\alpha + \beta X)}}{1 + e^{-(\alpha + \beta X)}} \quad (0 \leq Y \leq 1)$$

可以通过将 Y 变换为 $\ln[Y/(1-Y)]$ 而被线性化，但产生的误差可

能没有为 0 的均值和常数方差。处理这些模型最好用非线性或广义线性模型，这在第 12 章中讨论。

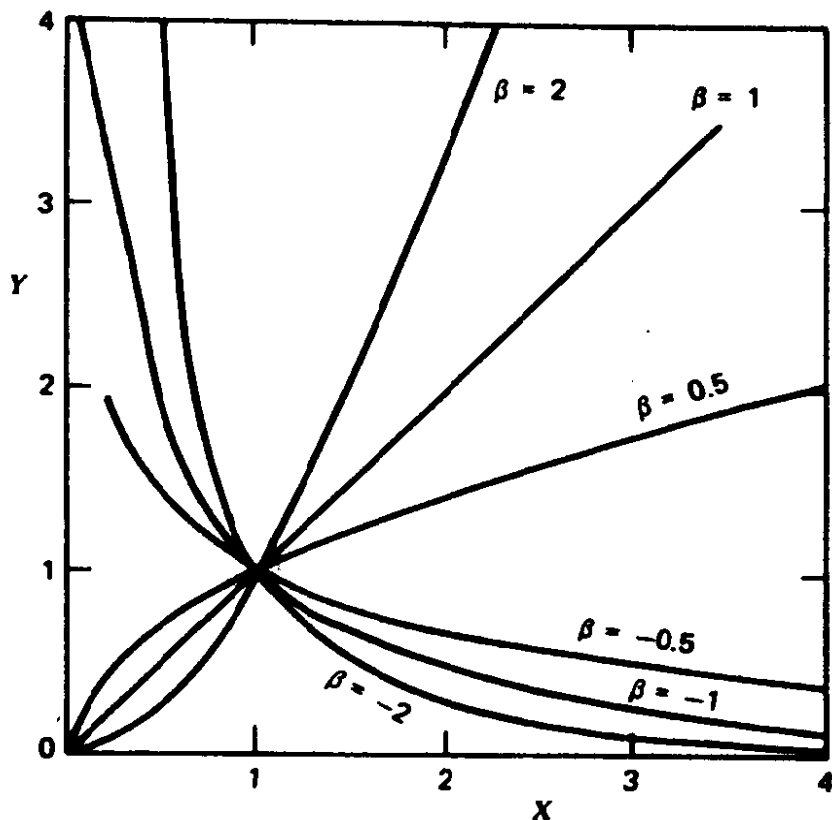


图 6.7 一族指数曲线, $Y = X^\beta$

表 6.5 线性化变换

变换		简单回归形式	多元回归形式
$\log Y$	$\log X$	$Y = \alpha X^\beta$	$Y = \alpha X_1^{\beta_1} X_2^{\beta_2} \cdots X_p^{\beta_p}$
$\log Y$	X	$Y = \alpha e^{\beta X}$	$Y = \alpha e^{\sum \beta_j X_j}$
Y	$\log X$	$Y = \alpha + \beta (\log X)$	$Y = \alpha + \sum \beta_j \log (X_j)$
$\frac{1}{Y}$	$\frac{1}{X}$	$Y = \frac{X}{\alpha X + \beta}$	$Y = \frac{1}{\alpha + \sum (\beta_j / X_j)}$
$\frac{1}{Y}$	X	$Y = \frac{1}{\alpha + \beta X}$	$Y = \frac{1}{\alpha + \sum \beta_j X_j}$
Y	$\frac{1}{X}$	$Y = \alpha + \beta \left(\frac{1}{X} \right)$	$Y = \alpha + \sum \beta_j \left(\frac{1}{X_j} \right)$

非线性性症状 作图的方法有助于找出非线性性。残差关于拟合值或关于自变量的图的曲线化趋势表示可能需要变换。在多元回归中，附加变量图或与其紧密相联系的偏残差图（两者在 2.4 节中有描述）比通常的残差图在诊断非线性性时更为有用。曲线化趋势再次表示需要进行变换。

如果在自变量的每个值重复进行观测，则 4.3 节中描述的拟合失真的 F -检验可用在这里，以提供非线性性的症状。然而， F 统计量仅起导向作用，它不是一个检验，这是因为探索过程可能使 F -检验的分布假设变为无效。

治疗 变换，或增加多项式，或者，可能的话，增加原自变量的叉积项可能是对非线性性的疗法。要求增加多项式的症状常类似于要求进行变换的症状。唯有经验可以帮助决定，在类似的问题中使用两者中的哪一个。

例 6.3 脑重和体重

表 6.6 给出 62 种哺乳动物的平均脑重与体重。我们考虑将脑重作为体重的函数的建模问题。这些数据取自一个更大型的、用于其它目的的研究 (Allison 和 Cicchetti, 1976)。

首先作脑重（克）关于体重（千克）的图，见图 6.8 (a)。它立即显示出需要某些变换。图中大多数点都挤在左下角，只有小部分散落在别处。由于这两个变量的变化都很大，显然应该选用对数变换。这如同假设正确的函数关系的形式为：脑重 = a_0 (体重) a_1 。在研究物体的部分之间的关系时，这样的模型是非常有效的。讨论见 Sprent (1972) 或 Gould (1966)。重量以 10 为底的对数由表 6.6 给出。取对数后的散点图由图 6.8 (b) 给出。它显示出，在取对数后，存在着强线性关系。由拟合回归得到的残差的散点图在图 6.8 (c) 给出。拟合直线很好地对应于观测数据。表 6.7 给出回归分析简要。

在这个例子里，使用的变换似乎为了达到两个重要目标：线性性与常数方差。在许多问题中，对其中一个问题合适的变换，对另一个问题可能并不合适。由图 6.8 (c) 可见，最大残差 r_i 为 2.848，它在案例人类处达到。由 (5.8) 计算得 t_i 为 3.04。这一统计量可以用于检验人类是否为异常值。如果在察看数据前即对人类是否为一特别的种类感兴趣的话，则正则 t -表，而非邦弗伦尼 t -表，可以用于得到显著性水平。由表 A，显著性水平约为 0.005，

故人类被认为是脑重太大，以至于与数据的模型不一致。

表 6.6 62 种哺乳动物的平均脑重与体重

	体重 (kg)	脑重 (g)	log (体重)	log (脑重)
1. 北极狐	3.385	44.500	0.530	1.648
2. 泉猴	0.480	15.500	-0.319	1.190
3. 山狸	1.350	8.100	0.130	0.908
4. 母牛	465.000	423.000	2.667	2.626
5. 灰狼	36.330	119.500	1.560	2.077
6. 山羊	27.660	115.000	1.442	2.061
7. 牦麝鹿	14.830	98.200	1.171	1.992
8. 天竺鼠	1.040	5.500	0.017	0.740
9. 长尾灰獾猴	4.190	58.000	0.622	1.763
10. 栗鼠	0.425	6.400	-0.372	0.806
11. 松鼠	0.101	4.000	-0.996	0.602
12. 北极松鼠	0.920	5.700	-0.036	0.756
13. 非洲巨袋鼠	1.000	6.600	-0.000	0.820
14. 短尾	0.005	0.140	2.301	-0.854
15. 星鼻鼹鼠	0.060	1.000	-1.222	-0.000
16. 犰狳	3.500	10.800	0.544	1.033
17. 树蹄兔	2.000	12.300	0.301	1.090
18. 美洲负鼠	1.700	6.300	0.230	0.799
19. 亚洲象	2 547.000	4 603.000	3.406	3.663
20. 大棕蝙蝠	0.023	0.300	-1.638	-0.523
21. 驴	187.100	419.000	2.272	2.622
22. 马	521.000	655.000	2.717	2.816
23. 豪猪	0.785	3.500	-0.105	0.544
24. 帕特斯猴	10.000	115.000	1.000	2.061
25. 猫	3.300	25.600	0.519	1.408
26. 狢	0.200	5.000	-0.699	0.699
27. 香猫	1.410	17.500	0.149	1.243

(续表)

	体重 (kg)	脑重 (g)	log (体重)	log (脑重)
28. 长颈鹿	529.000	680.000	2.723	2.833
29. 大猩猩	207.000	406.000	2.316	2.609
30. 灰海豹	85.000	325.000	1.929	2.512
31. 灰蹄兔	0.750	12.300	-0.125	1.090
32. 人类	62.000	1 320.000	1.792	3.121
33. 非洲象	6 654.000	5 712.000	3.823	3.757
34. 水鼩	3.500	3.900	0.544	0.591
35. 罗猴	6.800	179.000	0.833	2.253
36. 大袋鼠	35.000	56.000	1.544	1.748
37. 黄腹土拨鼠	4.050	17.000	0.607	1.230
38. 金仓鼠	0.120	1.000	-0.921	0.000
39. 老鼠	0.023	0.400	-1.638	-0.398
40. 小棕鼠	0.010	0.250	-2.000	-0.602
41. 蜂猴	1.400	12.500	0.146	1.097
42. 霍加披	250.000	490.000	2.398	2.690
43. 兔子	2.500	12.100	0.398	1.083
44. 绵羊	55.500	175.000	1.744	2.243
45. 美洲豹	100.000	157.000	2.000	2.196
46. 黑猩猩	52.160	440.000	1.717	2.643
47. 狒狒	10.550	179.500	1.023	2.254
48. 沙漠豪猪	0.550	2.400	-0.260	0.380
49. 巨犰狳	60.000	81.000	1.778	1.908
50. 岩蹄兔	3.600	21.000	0.556	1.322
51. 浣熊	4.288	39.200	0.632	1.593
52. 田鼠	0.280	1.900	-0.553	0.279
53. 东部美洲鼯鼠	0.075	1.200	-1.125	0.079
54. 鼯鼠	0.122	3.000	-0.914	0.477
55. 麝鼩	0.048	0.330	-1.319	-0.481
56. 猪	192.000	180.000	2.283	2.255
57. 针鼹	3.000	25.000	0.477	1.398

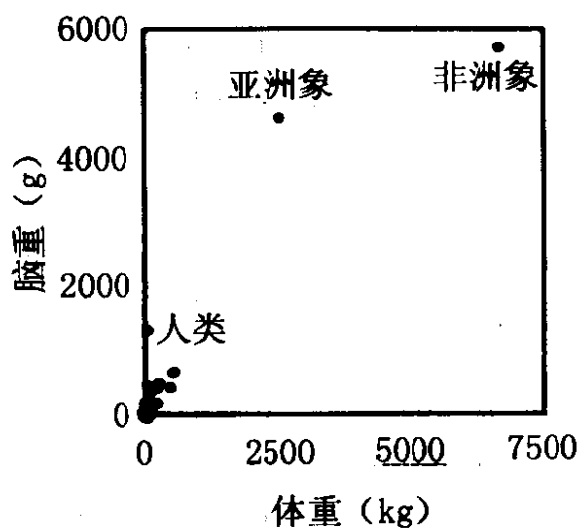
(续表)

	体重 (kg)	脑重 (g)	log (体重)	log (脑重)
58. 巴西豹	160.000	169.000	2.204	2.228
59. 无尾猬	0.900	2.600	-0.046	0.415
60. 袋貂科动物	1.620	11.400	0.210	1.057
61. 树鼯	0.104	2.500	-0.983	0.398
62. 红狐狸	4.235	50.400	0.627	1.702

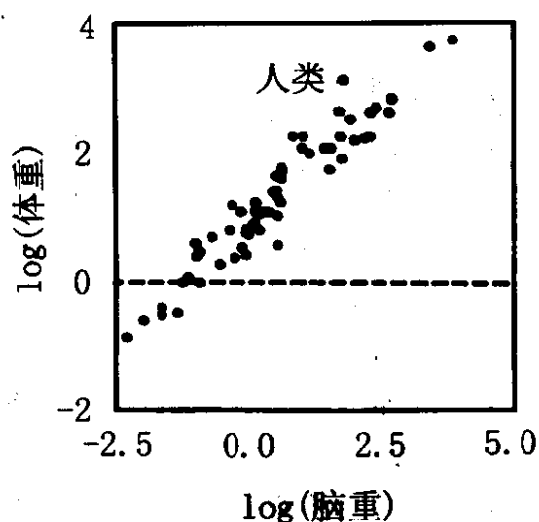
表 6.7 回归分析简要: 脑重/体重数据

	估计	标准误	t-值
截距	0.927	0.0417	22.23
斜率	0.752	0.0285	26.41

$R^2=0.92$, $\hat{\sigma}=0.0909$, d.f. = 60

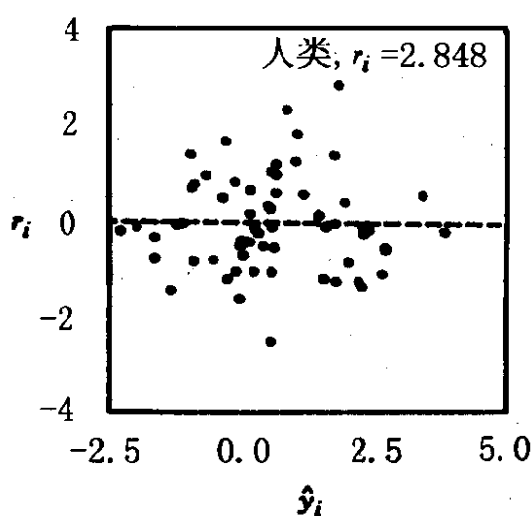


(a) 散点图



(b) 取对数后的散点图

图 6.8 脑重数据



(c) 残差图

图 6.8 脑重数据 (续)

6.4 变换响应变量

到现在为止, 讨论过的选择变换的方法, 或者使用变量间关系的特定知识, 或者依靠不严格的图形的帮助。在本节及下一节, 我们将用模型扩展的思想, 寻找选择变换的客观方法。尽管这些方法通常需要若干很强的假设, 它们在许多问题中还是很有用的。我们已经看到, 变换是对非常数方差或非线性的一个疗法。另外, 变换后回归中的误差比未变换时的误差更接近正态分布。尺度的变换使得我们可以用一个较简单的模型来描述一组数据。同一个变换并不是总能同时满足这四个目标的, 有时需要有所折衷。

Box 和 Cox (1964) 对选择变换的问题给出了一个系统化的处理方法。他们扩展模型, 将寻找变换的问题转化为估计一个参数的问题。我们现在假设每一个 $y_i > 0, i = 1, 2, \dots, n$ 。

假设 Y^λ 是一个 $n \times 1$ 向量, 其第 i 个元素为 y_i^λ 。如果 $\lambda = 0$, 我们令 $y_i^\lambda = \ln(y_i)$ 。如果 $\lambda = 1$, 则数据未被变换; 否则 λ 决定了一个幂变换。常用值, 如倒数, $\lambda = -1$; 对数, $\lambda = 0$; 平方根和立方根, λ 分别等于 $1/2$ 和 $1/3$, 都包括在这一变换族中。Box 和 Cox

建议检验模型

$$Y^\lambda = X\beta + e, \quad \text{var}(e) = \sigma^2 I \quad (6.3)$$

给出了数据及这一模型，我们可以同时估计 $(\beta, \sigma^2, \lambda)$ ，从而确定一个估计变换。与本书前面迁到的大部分模型不同，(6.3) 不是一个线性模型，故对 λ 的估计需要除最小二乘法以外的其它估计技术。我们将考虑估计 λ 的两种方法。第一种方法为 Box 和 Cox 提出的极大似然类型估计。第二种方法由 Atkinson (1973, 1981) 提出。他用一个近似线性模型代替 (6.3)，从而可以用最小二乘计算来估计 λ 。

Box 和 Cox 提出的问题是，如何选择 λ ，使误差尽可能接近于一个正态样本。为此，他们的方法给出了朝正态化方向的变换。我们在此不需牵涉到该方法的技术难点。这方面的讨论可参见 Hernandez 和 Johnson (1980) 或 Cook 和 Weisberg (1982a, 2.4 节)。

首先假设我们已知 λ 。由此可立即求得 β 的估计并计算残差平方和。它们是

$$\hat{\beta} = (X^T X)^{-1} X^T Y^\lambda \quad (6.4)$$

$$RSS_\lambda = (Y^\lambda)^T (I - H) Y^\lambda \quad (6.5)$$

在一个一般的线性回归程序中，通过 Y^λ 关于自变量的回归，可以求得上述值。

由于 λ 实际上是未知的，(6.5) 可以在 λ 的一个合理范围内进行计算。范围约指 -2 到 $+2$ 。如果 λ 在这一范围之外，这一方法的有用性是值得怀疑的。为比较 λ 的各个值，我们不能直接比较残差平方和，这是因为对每个 λ ，残差平方和的单位是不同的。这可以用两种等价的方法来处理。两种方法都需要进行尺度变换，以给出每个 λ 的可比值。一种这样的尺度是对数似然函数。使用 (6.5)，在 $\lambda \neq 0$ 时，对数似然函数为

$$L(\lambda) = n \ln(|\lambda|) - \frac{n}{2} \ln(RSS_\lambda) + n(\lambda - 1) \ln(GM(y)) \quad (6.6a)$$

其中 $GM(y)$ 是 y 的几何平均, $GM(y) = (\prod y_i)^{1/n}$ 。如果 $\lambda = 0$, 则 $L(0)$ 的计算如下

$$L(0) = -\frac{n}{2} \ln(RSS_0) - n \ln(GM(y)) \quad (6.6b)$$

使 $L(\lambda)$ 取最大值的 λ 值为估计值, 即 $\hat{\lambda}$ 。Box 和 Cox 建议作 $L(\lambda)$ 关于 λ 的图, 并从图中读出最大值。我们通常将 $\hat{\lambda}$ 舍入至邻近的一个常用值, 如 -1 , $-1/2$, 0 等等。这对于决定一个变换是足够精确的了。

如果我们使用一个更复杂的幂变换族, 则可能有一个更简单的方法来进行这些计算。令 Z^λ 是一个 $n \times 1$ 向量, 其第 i 个元素被定义为

$$z_i^\lambda = \begin{cases} \frac{y_i^\lambda - 1}{\lambda [GM(y)]^{\lambda-1}} & \lambda \neq 0 \\ GM(y) \ln(y_i) & \lambda = 0 \end{cases} \quad (6.7)$$

其中, 与前面一样, $GM(y)$ 为几何平均。与 y_i^λ 不同, z_i^λ 在 $\lambda=0$ 处是连续的。如果我们拟合模型

$$Z^\lambda = X\beta + e \quad (6.8)$$

并计算 Z^λ 关于自变量的回归, 则对于每一个 λ , 残差平方和 $RSS_\lambda(Z)$ 的尺度都是相同的。从而对不同的 λ 值, 这些值可以进行比较。虽然这一结论并不显而易见, 但其证明并不困难, 例如, 在 Cook 和 Weisberg (1982a, 2.4 节) 就有证明。 λ 可以被选择为使 $RSS_\lambda(Z)$ 最小的值。或者, 可以对任意的 λ , 计算对数似然函数

$$L(\lambda) = -\frac{n}{2} \ln[RSS_\lambda(Z)] \quad (6.9)$$

这将给出与 (6.6a) 和 (6.6b) 相同的值。

例 6.4 罗马式教堂的规模

在例 6.3 中, 我们已经看到了在哺乳动物的脑重与体重之间的非常强的关系。对身体的其它部分也可以找到类似的关系。Gould (1973) 考虑生物中的形状“规律”对其它物体的适用性。为研究这一点, 他选择了一个“头脑简单”的例子: 中世纪教堂。它们被建成各种规模和形状, 但服务于同一目

的。由于作为建筑材料的石料的限制，我们可以设想，教堂的各种测量之间的关系是很强的。表 6.8 列出了 1066 年英国被威廉征服后的 25 个罗马式教堂的周长（百米）和面积（百平方米）。数据从 Clapham (1934) 给出的地面布局计划中测得，由 S. J. Gould 提供。因为很少有教堂是矩形的，故这两个测量值之间的关系并不明显（亦可见例 3.1）。图 6.9a 的散点图指出，可能是非线性关系，且方差随周长增加。基于从窗口进入的光线要有穿透相对较厚的墙的需要，Gould 认为，对这两个量都要进行变换。这里我们忽略他的理论推导，选择面积为响应变量，并用 Box 和 Cox 的方法去选择响应变量所要进行的变换。

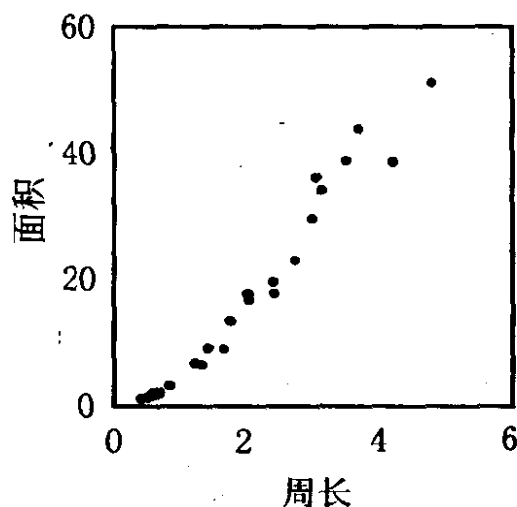
表 6.8 25 个罗马式教堂的周长与面积

教堂	周长 (百米)	面积 (百平方米)	教堂	周长 (百米)	面积 (百平方米)
St. Albans	3.48	38.83	Byland	3.14	34.27
Durham	3.69	43.92	Roche	2.04	17.61
Blyth	1.43	9.14	Carmel	1.77	13.37
Binham	2.05	16.66	Bengeo	0.59	2.04
Gloucester	3.05	36.16	Copford	0.69	2.22
Norwich	4.19	38.66	Kempley	0.50	1.46
Leominster	2.43	17.74	Birkin	0.69	1.92
Southwell	2.40	19.46	Hales	0.63	1.86
Chertsey	2.72	23.00	Moccas	0.58	1.69
Hereford	2.99	29.75	Peterchurch	0.86	3.31
Canterbury	4.78	51.19	Little Tey	0.41	1.13
Lindesfarne	1.33	6.60	Melbourne	1.23	6.74
Tintern	1.67	9.04			

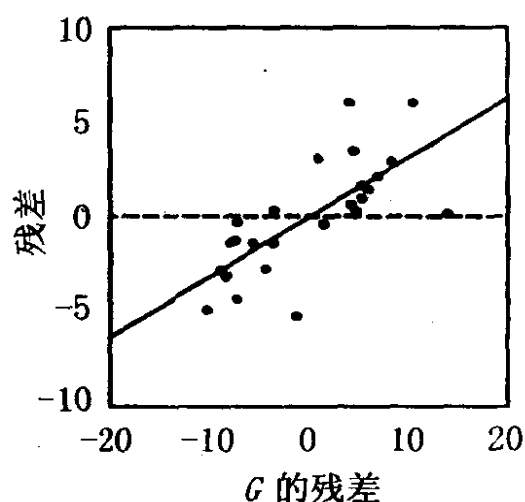
来源：S. J. Gould.

为得到 $\hat{\lambda}$ ，我们对 -2 到 $+2$ 区间的一段范围内的 λ 值计算 (6.7) 的值。为做到这一点，取值 -1 ， -0.5 ， 0 ， 0.5 和 1 通常就足够了。对每个 λ 值，我们计算残差平方和 $RSS_{\lambda}(Z)$ 。例如，当 $\lambda=0$ ，我们得到 $RSS_0(Z) = 377.3043$ ，而当 $\lambda=0.5$ ，我们有 $RSS_{0.5}(Z) = 116.2636$ ，这个值要小得多。据 (6.9)，我们有 $L(0) = -74.16$ ，而 $L(0.5) = -59.45$ 。在 $\lambda \in (0, 1)$ 区间的完整曲线由图 6.9c 给出。由于 $L(\lambda)$ 是单峰的，曲线在这个区间外是下降的。

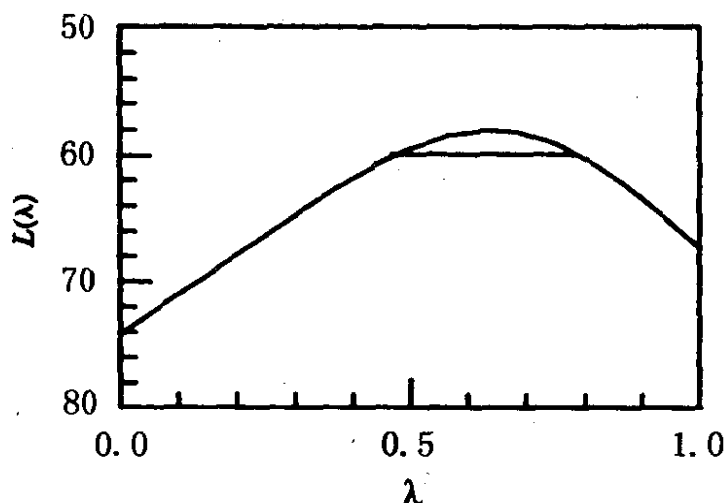
$L(\lambda)$ 的最大值约在 $\hat{\lambda}=0.63$ 处取得, $L(0.63)=-58.05$ 。很明显, 在 $\lambda=0.5$ 与 $\lambda=0.63$, $L(\lambda)$ 的值几乎是相同的, 故用 λ 的这两个值中任一个, 效果是差不多的。我们可以得到 λ 的 $(1-\alpha) \times 100\%$ 的近似置信区间, 即所有满足 $L(\lambda) > L(\hat{\lambda}) - (1/2)\chi^2(\alpha; 1)$ 的 λ 的集合。例如, 一个 95% 的区间包括满足 $L(\lambda) > -58.05 - \frac{1}{2}(3.84) = -59.97$ 的 λ 。这个集合对应于图中用水平线标出的集合, $0.45 \leq \lambda \leq 0.80$ 。我们一定会选择面积的平方根变换, 即因为它接近 $\hat{\lambda}$, 也因为面积的平方根与周长的单位是相同的。



(a) 面积对周长



(c) Atkinson 的得分方法的附加变量图



(b) $L(\lambda)$ 对 λ

图 6.9 罗马式教堂数据

Atkinson 的得分方法 虽然 Box 和 Cox 的选择变换的方法在近些年中被广泛应用, 但它需要有相当大的计算量, 并且要有

某些专门的软件包来作对数似然函数图。Atkinson (1973) 提出了在选择 λ 的不为 1 的一个值时, 需要进行的快速检验。而后 (1981), 他提议使用这一方法给出一个快速估计及一个图形诊断。

Atkinson 的方法是用一个可以用线性计算进行分析的模型代替 (6.8)。这一方法可以通过在值 $\lambda=1$ 处, 用泰勒展开式的前两项近似 Z^λ 来导出。

$$\begin{aligned} Z^\lambda &\cong Z^1 + (\lambda - 1) \left. \frac{\partial Z^\lambda}{\partial \lambda} \right|_{\lambda=1} \\ &= Z^1 + (\lambda - 1) G \end{aligned} \quad (6.10)$$

其中 G 的第 i 个元素 g_i 是通过将 Z^λ 的第 i 个元素对 λ 求导, 并求该导数在 $\lambda=1$ 处的值得到。

$$g_i = y_i \{ \ln[y_i/GM(y)] - 1 \} + \{ \ln[GM(y)] + 1 \} \quad (6.11)$$

(6.11) 右边第二个括号内的项不依赖于 i , 可以被忽略。将 (6.10) 代入 (6.8), 我们近似地得到

$$Z^1 \cong X\beta + (1 - \lambda)G + e \quad (6.12)$$

由于 $Z^1 = Y - 1$, 我们可以用 Y 代替 Z^1 。令 $\phi = 1 - \lambda$, 则有

$$Y = X\beta + \phi G + e \quad (6.13)$$

这样, $\phi=0$ 的检验和 $\lambda=1$ 的检验是近似一样的。 $\phi=0$ 的检验是由 Y 关于 X 和 G 的回归得到的通常的 t 统计量。Atkinson 称它为 t_D 统计量。把它与标准正态分布相比较, 以得到近似的 P -值。一个双边的检验是合适的。对 λ 的一个快速估计也可从这个回归中得到。如果 $\tilde{\phi}$ 是从回归得到的 ϕ 的估计, 则 λ 的估计 $\tilde{\lambda}$ 为 $1 - \tilde{\phi}$ 。这一检验的诊断图为将 G 添加到 Y 关于 X 的回归模型后的附加变量图。

对于教堂的例子, 面积关于周长和 G 的回归给出 $\tilde{\phi} = 0.32$, $t_D = 4.87$ 。由于对应的 p -值很小, 故我们有理由要求对响应变量进行变换。变换的快速估计为 $1 - 0.32 = 0.68$, 与似然类型的结果相一致。在大多数问题中, 我们不能期望有这么高的一致性。在什么条件下它们会不一致是未知的。图 6.9 (c) 是有了周长之后再

添加 G 的附加变量图。这个图中的总体线性趋势显示出对面积进行变换是有用的。再次建议使用平方根变换。下一节我们将回到这一例子。

附加评注 Box 和 Cox 方法只在响应变量严格为正时是可用的。如果出现零值或负值，常用的方法是，首先在响应变量上加上一个常数，然后再应用 Box 和 Cox 的方法。不幸的是，数据几乎不提供什么信息来帮助选择附加常数，参见 Atkinson (1983)。Carroll (1980) 提出了一个确定 λ 值的“稳健”的方法，见 Cook 和 Wang (1983)。

最近有一份令人感兴趣的文献探讨了由数估计了 λ 之后所作的推断的解释。Bickel 和 Doksum (1981) 认为，这一方法可能不一定有用，因为没有关于尺度 λ 的条件，置信断言将是不精确的，并且难以使用。其他有些人认为，这一观点实际上并不重要，并且一旦计算了 $\hat{\lambda}$ ，则所有进一步的分析是在给出 $\hat{\lambda}$ 的条件下进行的，参见 Hinkley 和 Runger (1984)，以及在这篇论文后面的对这一争论问题的讨论。

6.5 变换自变量

在变换自变量时我们要区分两种情况。首先，响应变量在自变量的范围内达到最大或最小值。达到极值的点通常是我们感兴趣的。据此，常常是拟合自变量的乘方，如 X_1^2 ， X_1^3 ， X_1X_2 等等，对响应变量建立模型。这是多项式回归，它将在下一章中讨论。另外，响应变量可能随自变量单调上升或下降，但不是按常数速率。如果要用幂变换，指数值在区间 $(-2, +2)$ 中是合适的。这种类型的变换是本节的主题。

通过模型扩展，可以得到对自变量变换的检验，所需变换的快速估计以及一个诊断图。这里给出的方法是在本书第一版中给出方法的简介。

假设我们有一个含 p 个自变量的多元回归问题。我们考虑变换它们中的一个, 如 X_1 。如果我们限制在幂变换的范围内, 则线性模型

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + e \quad (6.14)$$

可被扩展为

$$Y = \beta_0 + \beta_1 X_1^{\alpha_1} + \sum_{j=2}^p \beta_j X_j + e \quad (6.15)$$

其中, 和前面一样, 如果 $\alpha_1 = 0$, 我们取 $X_1^{\alpha_1} = \ln(X_1)$ 。这一非线性模型的参数可以用非线性最小二乘法来估计。这里我们通过 $\alpha_1 = 1$ 的泰勒级数来展开 $X_1^{\alpha_1}$, 从而用线性模型来近似 (6.15)。

$$\begin{aligned} X_1^{\alpha_1} &\cong X_1 + (\alpha_1 - 1) \frac{\partial}{\partial \alpha_1} X_1^{\alpha_1} \Big|_{\alpha_1=1} \\ &\cong X_1 + (\alpha_1 - 1) X_1 \ln(X_1) \end{aligned} \quad (6.16)$$

将 (6.16) 代入 (6.15), 得

$$\begin{aligned} Y &\cong \beta_0 + \beta_1 [X_1 + (\alpha_1 - 1) X_1 \ln(X_1)] + \sum_{j=2}^p \beta_j X_j + e \\ &= \beta_0 + \sum_{j=1}^p \beta_j X_j + \beta_1 (\alpha_1 - 1) X_1 \ln(X_1) + e \\ &= \beta_0 + \sum_{j=1}^p \beta_j X_j + \eta X_1 \ln(X_1) + e \end{aligned} \quad (6.17)$$

其中 $\eta = \beta_1 (\alpha_1 - 1)$ 。 $\eta = 0$ 对应于 $\beta_1 = 0$ 或 $\alpha_1 = 1$ 。粗略地说, 这是一个是否需要变换的检验。和通常一样, 我们从 Y 关于自变量和 $X_1 \ln(X_1)$ 的回归估计 η 。常用的 $\eta = 0$ 的 t -检验, 其自由度为 $(n - p' - 1)$, 它是是否需要变换的一个合适的检验。在模型中的其它自变量之后的, $X_1 \ln(X_1)$ 的附加变量图是它的一个诊断图。对 α_1 的一个快速估计是解关于 α_1 的方程, $\eta = \beta_1 (\alpha_1 - 1)$,

$$\hat{\alpha}_1 = \frac{\hat{\eta}}{\hat{\beta}_1} + 1 \quad (6.18)$$

其中 $\hat{\eta}$ 由(6.17) 估计得到, 而 β_1 是由(6.14) 估计得到的, 而不是由(6.17)。

我们对例 6.4 的教堂数据应用这一方法, 并通过拟合

$$\sqrt{(\text{面积})} = \beta_0 + \beta_1(\text{周长}) + e \quad (6.19)$$

来检验变换周长的需要。拟合模型如表 6.9 (a) 所示。 β_1 的 t -值为 26.20。很明显, 斜率等于零是不合适的。在斜率近似为零的时候, 上述确定变换的方法是没有用处的。拟合下列模型如表 6.9 (b) 所示,

$$\sqrt{(\text{面积})} = \beta_0 + \beta_1(\text{周长}) + \eta(\text{周长})[\ln(\text{周长})] + e \quad (6.20)$$

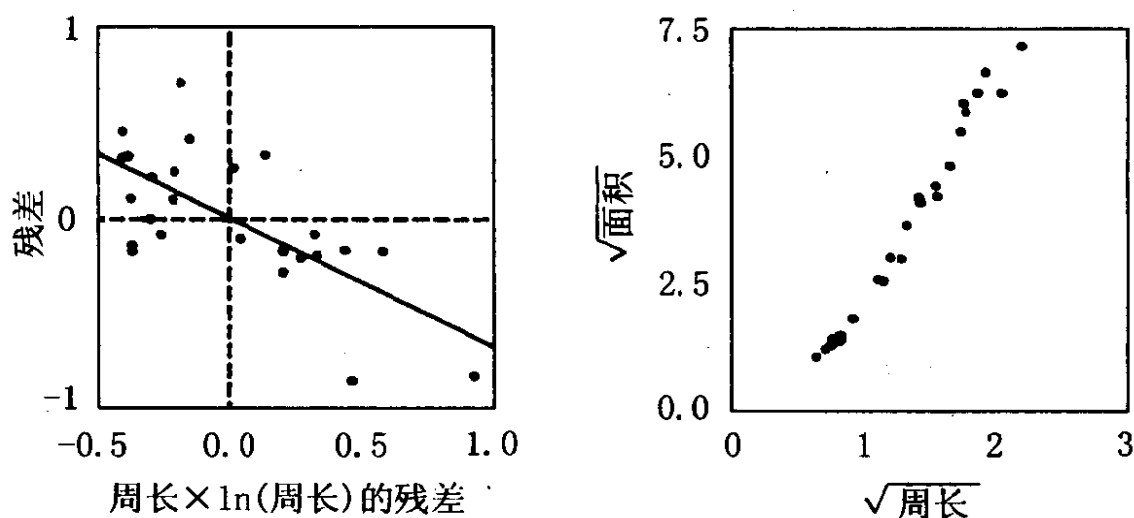
在这一回归中, 我们仅对 η 的估计、其标准差及 t -值感兴趣。 $\eta=0$ 的检验具有值 $t=-4.45$, 与 $t(22)$ 相比较, 我们有理由认为, 需要变换周长。建议的变换为

$$\hat{\alpha}_1 = \frac{-0.6726}{1.5437} + 1 = 0.56 \quad (6.21)$$

再舍入至一个常用的常数因子, 我们取周长的平方根。周长 $\times \ln(\text{周长})$ 的附加变量图, 如图 6.10a 所示。由其大致的线性趋势可知, 需要进行变换。

表 6.9 变换周长, 教堂数据

(a)			
	估计	标准误	t -值
截距	0.5998	0.1374	4.36
周长	1.5437	0.0589	26.20
$\hat{\sigma}^2=0.1344, R^2=9676, \text{d.f.}=23$			
(b)			
	估计	标准误	t -值
截距	-0.5036	0.2679	-1.88
周长	2.6966	0.2625	10.27
附加变量	-0.6727	0.1510	-4.45
$\hat{\sigma}^2=0.0739, R^2=9830, \text{d.f.}=22$			



(a) 周长 $\times \ln$ (周长) 的附加变量图

(b) $\sqrt{\text{面积}}$ 对 $\sqrt{\text{周长}}$

图 6.10

(面积) $^{1/2}$ 关于 (周长) $^{1/2}$ 的散点图由图 6.10b 给出。虽然这个图是相当直的，但小型教堂对整体趋势似乎有系统偏差。这里的一个重要教训是，自动选择变换的方法有时并不能给出一个完全令人满意的结果。对这些数据的进一步讨论，见 Gould (1968)。

附加评注 这里导出的方法部分由 Box 和 Tidwell (1962) 给出。他们给出一个递归过程，用以估计 α_1 。该递归过程是通过重复一个单步过程而实现的。这个单步估计等价于一个得分检验。其思想和前面描述的对常数方差的检验及 Atkinson 检验是一样的。在多元回归问题中，可能会对变换多个自变量感兴趣。我们建议用序贯方法，逐个研究每个自变量。

在许多问题中，这个对一个自变量确定一个变换的方法可能会失效。这是因为从数据中只能得到很少的信息来区分，比如说， Y 关于 X 的回归以及 Y 关于 $X^{1/2}$ 的回归。另外，当限定用 $(-2, +2)$ 内的幂变换时，可能会出现荒谬的结论，如 $\hat{\alpha} = -74$ 或 $+18$ 。

考虑对一个自变量，如 X_1 的变换的选择。如果 (6.14) 中 β_1 的 t -统计量不大，则 (6.18) 将不合适地给出 $\hat{\alpha}_1$ 。类似地，如果 X_1

的最大值与 X_1 的最小值的比小于 10, 则 X_1 与 $X_1 \ln(X_1)$ 两个量会高度相关, 这样 $\hat{\eta}$, 从而 $\hat{\alpha}$ 也将被很不合适地给出。如果这一比率大于 10, 在开始任何分析之前, 很可能应把 X_1 转换为对数, 使我们得到共线性的自变量。

在响应变量与自变量的变换之间, 有明显的相互作用。虽然存在同时变换的方法 (见 Cook 和 Weisberg, 1982a, 例 2.4.5), 但对大多数问题并不需要使用这种方法。假设所有数据严格为正, 我们推荐以下的程序。首先, 对最大值/最小值的比大于 10 的任何一个自变量, 采用对数变换。其次, 用 Box 和 Cox, 或者 Atkinson 的方法变换响应变量。最后, 对任何具有大的 t -值的自变量, 考虑使用本节概略描述的 Box-Tidwell 方法检验自变量。

如果我们将这些想法用于教堂数据, 由于最大 (周长) / 最小 (周长) 的比约为 10, 我们将 $\ln(\text{周长})$, 而不是周长看作一个自变量。如果我们使用 Box 和 Cox 的方法来选择面积的一个尺度, 我们得到如图 6.11a 所示的似然曲线, 其中 $\hat{\lambda} = 0.02$, $RSS(Z) = 34.81$, 明显地显示出要对面积进行对数变换。图 6.11 (b) 所示的 $\ln(\text{面积})$ 关于 $\ln(\text{周长})$ 的图是令人满意的。

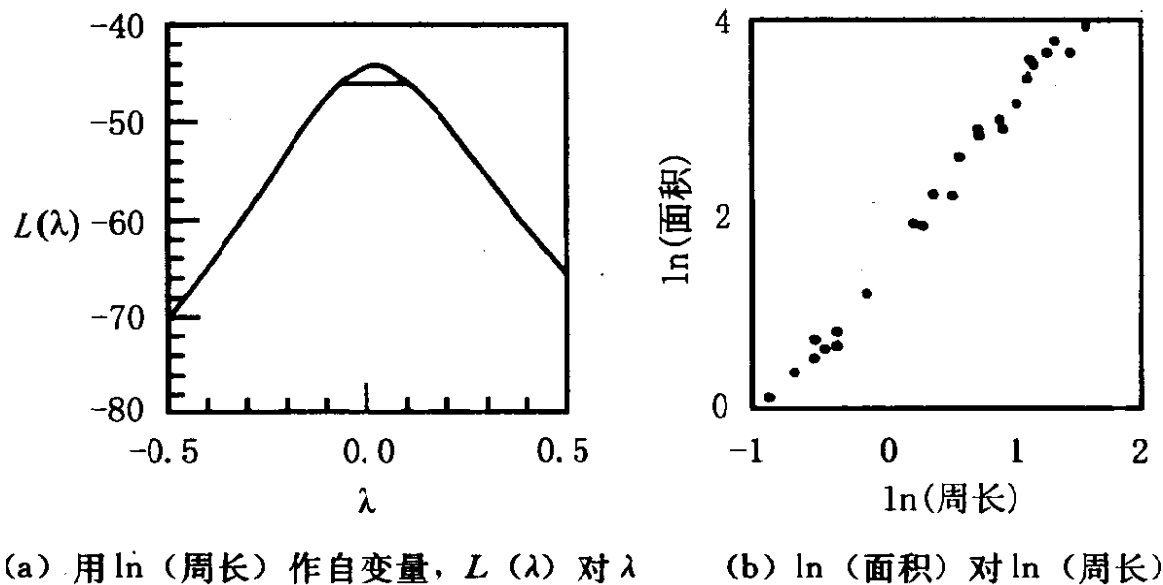


图 6.11

6.6 正态性假设

在回归分析中，通常的分布假设是，误差服从正态分布。这假定被用于给出 F -检验， t -检验及置信推断。在小样本中，很难通过检验残差来诊断误差的非正态性。假设我们有一个线性模型 $Y = X\beta + e$ ，其中 $\text{var}(e) = \sigma^2 I$ ，我们希望检验正态性假设。当然， e 是不可观测的，故正态性检验必须基于残差 \hat{e} 。由 (5.3)，用 $X\beta + e$ 代入 Y ，因为 $(I - H)X = 0$ ，故

$$\begin{aligned}\hat{e} &= (I - H)Y \\ &= (I - H)(X\beta + e) \\ &= (I - H)X\beta + (I - H)e \\ &= (I - H)e\end{aligned}$$

记 H 的第 (i, j) 个元素为 h_{ij} ，则最后一个等式的标量形式是

$$\hat{e}_i = e_i - \left(\sum_{j=1}^n h_{ij} e_j \right) \quad (6.22)$$

这样， \hat{e}_i 等于 e_i 减去一个包括 e_i 在内的所有 e_j 的加权和。如果误差的自由度 $n - p'$ 较小，且某些 h_{ij} 较大，则 (6.22) 括号内的项可能在决定 \hat{e}_i 的分布时，比 e_i 更为重要。根据中心极限定理，即使 e_i 不是正态的，这个和也将如正态的。这样，至少在小样本中，任何关于残差的非正态性检验不会是很理想的。Gnanadesikan (1977) 称之为残差的超正态性。

对固定的 p' ，随着 n 的增加， h_{ij} 趋于零。(6.22) 中的 e_i 项将起决定作用，这是因为和式有相对小的方差。这样，对大样本应用于残差的通常方法可以与应用于误差本身的同一方法给出同样多的信息。

概率图 我们选择的研究非正态性的方法为正态概率或 rankit 图 (由 Wilk 和 Gnanadesikan, 1968 及 Gnanadesikan 1977 给出了对概率图的一般处理)。假设我们有一个容量为 n 的样本

z_1, z_2, \dots, z_n , 我们希望检验如下的假设: 这些 z_i 是从正态分布得到的同类样本, 均值 μ 与方差 σ^2 未知。一个有效的处理方法如下:





1. 将样本排序得到 $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$, 它们称为样本的次序统计量。

2. 现在考虑一个样本容量为 n 的, 均值为 0 并具有单位方差的正态样本。我们重复从标准正态总体中取容量为 n 的样本, 令 $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ 为其次序统计量的均值。 $u_{(i)}$ 称为正态次序统计量的期望值, 或称 rankit。rankit 常被列表表示, 如表 D, 或者可以使用计算机方便地近似得到*。

3. 如果这些 z_i 是正态的, 则

$$E(z_{(i)}) = \mu + \sigma \cdot u_{(i)}$$

故 $z_{(i)}$ 关于 $u_{(i)}$ 的回归将是一条直线。如果样本不是正态的, 则 $Z_{(i)}$ 关于 $u_{(i)}$ 的 rankit 图将不会近似于一条直线。

作为 rankit 图的一个例子, 由计算机产生一个伪随机样本, $n = 17$, 样本取自 $N(3, 4)$ 。使用表 D 的 rankit, 作如图 6.12 (a) 的 rankit 图。这个图近似地, 虽然不是精确地为一条直线。作为对比, 表 6.12 (b) 为取自 $(0, 1)$ 区间上的均匀分布的 $n = 17$ 的伪随机数的 rankit 图。后一个图在图的两端显示出扁平形状。这表示在样本中有太多的相当远离均值的点, 以致于我们不能认为它是取自正态分布的一个样本。除了本图看到的  , 其它非正态分布的常见形状包括  ,  和  。第一个表示有太多的极端值, 其余的两个分别表示负的和正的偏度。

总之, 判断一个 rankit 图是否指出一个样本的行为与正态分

* 使用一个子程序计算累积正态分布函数 $\Phi(t) = \int_{-\infty}^t (2\pi)^{-\frac{1}{2}} \exp(-x^2/2) dx$ 及其逆 $\Phi^{-1}(p)$, 样本容量为 n 时, $u_{(i)}$ 的一个近似是 $u_{(i)} \doteq \Phi^{-1} \times \left[\left(i - \frac{3}{8} \right) / \left(n - \frac{1}{4} \right) \right]$; 见 Blom (1958) 或 Weisberg 和 Bingham (1975), 或 Royston (1982b)。

布相同,需要有有经验的观测者。Daniel 和 Wood (1981) 及 Daniel (1976) 给出若干个“训练图”,以帮助分析者学习解释 rankit 图。

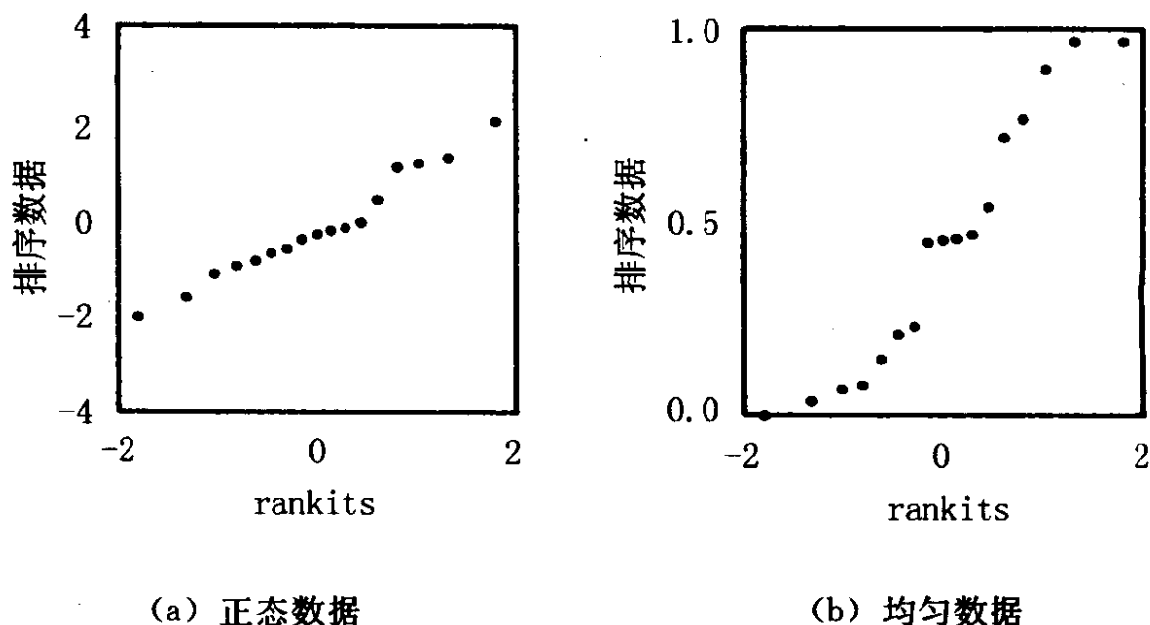


图 6.12 两个 rankit 图

6.4 节描述的 Box-Cox 方法常被作为向正态性接近的变换。 $\hat{\lambda}$ 是 λ 的使残差尽量接近于正态分布的值。Hernandez 和 Johnson (1980) 指出,即使“尽量接近”也可能不是很接近,诸如 rankit 图等的诊断检验应该在变换后进行。

附加评注 许多统计量被提出用于检验一个样本的正态性。其中一个相当有效的是 Shapiro 和 Wilk (1965) 的 W 统计量。它是 rankit 图中 $z_{(i)}$ 和 $u_{(i)}$ 的相关系数的平方。如果 W 太小,则拒绝正态性假设。Royston (1982abc) 给出检验及得到 p -值的细节及计算程序。

对于小样本的残差,提议用次序的 $r_{(i)}$ 或 $t_{(i)}$ 代替 $\hat{e}_{(i)}$, 但没有什么理由证明这三种选择中的一个会比其它的好。Atkinson (1981) 给出一个可用于小样本的近似显著性检验的计算方法。方法如下: (1) 固定样本容量 n 及自变量矩阵 X 。(2) 产出 $m=19$ 个 $n \times 1$ 的向量, 记为 A_1, \dots, A_{19} , 使每个向量的所有元素皆为标准正态伪随机数。(3) 计算每个 A_k 关于 X 的回归, 保存残差。

记第 K 个残差集合为 E_K 。(4) 对每个向量 E_K ，按由小到大的次序排列其中的元素。(5) 从这些 E_K 向量的 19 个最小值中，只保留最小和最大的。在 19 个其次小的值中，保留最小和最大的，…，在 19 个最大值中，保留最小和最大的。这些是对每个次序统计量的 $19/20 \times 100\% = 95\%$ 的置信区间估计。(6) 画这些最小、最大值和观测值 $\hat{e}_{(i)}/\hat{\sigma}$ 关于 $u_{(i)}$ 的图。被观测的残差应大部分落在这一模拟的“信封”里面。这一方法可通过使用学生化残差，变化 m 或使用不同的标准分布而被改变。Atkinson 使用学生化残差的绝对值，以及半正态分布。

关于教堂数据，最后得出的模型是对数模型。其 r_i 的 rankit 图，以及模拟的信封，如图 6.13 所示。我们没有理由怀疑数据的正态性。

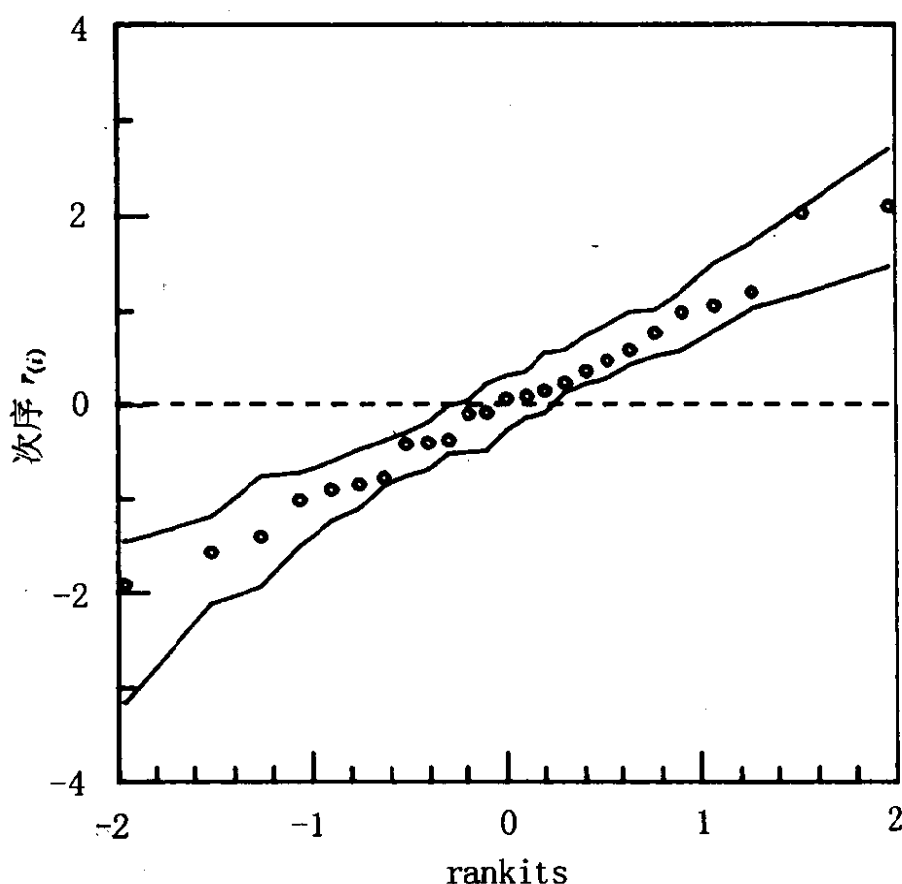


图 6.13 教堂数据的 rankit 图及模拟“信封”

相关误差 另一个关于误差的分布假设是，它们是不相关的。

这意味着一个案例的误差值不依赖于其它任何案例的误差值。在某些问题中会违反这一假设，特别是当案例按时间或空间的次序排序，且相邻案例相互影响时。

用于找出相关误差的诊断通常只能用在特定的环境下。例如，如果案例在时间上是等间隔的，则 Durbin-Watson 统计量 (Durbin 和 Watson, 1950, 1951, 1971) 适用于检验相邻案例的相关。一般地，诊断相关误差是很困难的，最好的诊断方法来自产生数据的过程中的仔细考虑。

问 题

- 6.1 对问题 1.2 的 Hooker 数据，使用 Box 和 Cox，以及 Atkinson 方法，在 *PRES* 关于 *TEMP* 的回归中，试确定对 *PRES* 的合适的变换。求 $\hat{\lambda}$, $\hat{\lambda}$, 得分检验和附加变量图。总结你得到的结果。
- 6.2 对线性模型假定的适用性，考查问题 3.3 的 Longley 数据。
- 6.3 以下数据是从一项关于被溶解的硫对液态铜的表面张力的影响的研究中取得的 (Baes 和 Killogg, 1953)

X=硫的重量 (%)	Y=表面张力的下降 (dynes/cm), 两次重复试验	
0.034	301	316
0.093	430	422
0.30	593	586
0.40	630	618
0.61	656	642
0.83	740	714

- 6.3.1 试求对 *X* 和 *Y* 的变换，使变换尺度后的回归为线性的。
- 6.3.2 假设 *X* 已被变换为 $\ln(X)$ ，对 *Y* 的哪一种变换将给出更好的结果，*Y* 还是 $\ln(Y)$? (Sclove, 1972)
- 6.4 以下 (假设的) 数据给出 $n=62$ 次试验中各种以速度 *X* (英里/小时) 行

驶的汽车及刹车距离 Y (英尺) (Ezekiel 和 Fox, 1959)

X	Y	X	Y
4	4	20	48
5	2, 4, 8, 8	21	39, 42, 55
7	7, 7	24	56
8	8, 9, 11, 13	25	33, 48, 56, 59
9	5, 5, 13	26	39, 41
10	8, 14, 17	27	57, 78
12	11, 19, 21	28	64, 84
13	15, 18, 27	29	54, 68
14	14, 16	30	60, 67, 101
15	16	31	77
16	14, 19, 34	35	85, 107
17	22, 29	36	79
18	29, 34, 47	39	138
19	30	40	110, 134

- 6.4.1 作 Y 关于 X 的散点图。拟合简单回归模型并作 r_i 关于 \hat{y}_i 的图。有什么明显的问题? 对拟合的模型计算拟合失真的 F -检验 (4.3 节)
- 6.4.2 使用 6.5 节的方法, 求对 X 的变换。为了知道需要还是不需要变换, 必须画附加变量图。
- 6.4.3 用 Atkinson 得分方法检验: 需要变换响应变量 Y 而不需要变换 X 。另外用 Box 和 Cox 的似然方法, 求估计的变换。作合适的图形上的概述。
- 6.4.4 基于理论研究, Hold (1960) 认为, 模型 $Y = \beta_1 x + \beta_2 x^2 + e$, 方差 $\text{var}(e) = \sigma^2 x^2$, 对于这一类型的数据是合适的。比较这一模型与问题 6.4.2 中得到的模型。关于 0 到 40mph 范围内的 X ,

对每个模型画预测 Y 的曲线图，并从数量上进行比较。另外，对给定的 X 值，计算预测值的方差。然后，对外推值 $X=60\text{mph}$ 和每一个模型，计算预测值及预测的标准误。这一比较将说明，如同问题 6.4.2 中得到的那种经验关系，在某些场合，如内推值的预测，是足够适用的，但在其它场合，如外推，将给出错误的结论。(Draper 和 Hunter, 1969)。

- 6.4.5 拟合问题 6.4.4 给出的 Hald 模型，但为常数方差， $\text{var}(e) = \sigma^2$ 。用诊断非常数方差的得分检验，试求下列检验问题的解：原假设为常数方差，而备择假设为方差随 x 及 x^2 增加。作图形上的概述。

- 6.5 表 6.10 的数据取自于一项对 1977 年用种植紫花苜蓿的农业用地的地租的变化情况的研究。数据包括：

Y = 种植紫花苜蓿土地的每英亩的平均地租。

X_1 = 所有可耕土地的每英亩的平均地租。

X_2 = 牛奶场奶牛密度 (头数/平方英里)。

X_3 = 用于牧场的耕地比例。

$X_4 = 1$ ，种植紫花苜蓿如果要求撒石灰；0，不撒石灰。

紫花苜蓿是一种高蛋白植物，适于喂养奶牛场奶牛。通常认为在一个奶牛场奶牛密度大的地方种植紫花苜蓿的地租，相对种植其它农作物的地租要高，而在要求撒石灰的地方，地租会低一些，这是因为撒石灰意味着额外的费用。

用目前所学的所有技术来研究这些数据，以了解地租结构。总结你的结论。

表 6.10 地租数据

Y	X_1	X_2	X_3	X_4	Y	X_1	X_2	X_3	X_4
18.38	15.50	17.25	0.24	0	51.79	56.00	14.25	0.15	1
20.00	22.29	18.51	0.20	1	96.67	71.41	21.37	0.05	0
11.50	12.36	11.13	0.12	0	50.83	65.00	13.24	0.08	1
25.00	31.84	5.54	0.12	1	34.33	36.28	5.85	0.10	1
52.50	83.90	5.44	0.04	0	48.75	59.88	32.99	0.21	0
82.50	72.25	20.37	0.05	1	25.80	23.62	28.89	0.24	1
25.00	27.14	31.20	0.27	0	20.00	24.20	6.29	0.06	1
30.67	40.41	4.29	0.10	1	16.00	17.09	33.34	0.66	0

(续表)

Y	X ₁	X ₂	X ₃	X ₄	Y	X ₁	X ₂	X ₃	X ₄
12.00	12.42	8.69	0.41	0	48.67	44.56	16.70	0.15	1
61.25	69.42	6.63	0.04	1	20.78	34.46	4.20	0.03	1
60.00	48.46	27.40	0.12	0	32.50	31.55	23.47	0.19	1
57.50	69.00	31.23	0.08	0	19.00	26.94	8.28	0.10	1
31.00	26.09	28.50	0.21	1	51.50	58.71	7.40	0.04	1
60.00	62.83	29.98	0.17	0	49.17	65.74	7.71	0.02	1
72.50	77.06	13.59	0.05	0	85.00	69.05	46.18	0.22	1
60.33	58.83	45.46	0.16	0	58.75	57.54	14.98	0.11	1
49.75	59.48	35.90	0.32	0	19.33	21.73	6.58	0.06	0
8.50	9.00	8.89	0.08	0	5.00	6.17	13.68	0.18	0
36.50	20.64	23.81	0.24	0	65.00	51.00	50.50	0.24	0
60.00	81.40	4.54	0.05	1	20.00	18.25	16.12	0.32	0
16.25	18.92	29.62	0.72	0	62.50	69.88	31.48	0.07	0
50.00	50.32	21.36	0.19	1	35.00	26.68	58.60	0.23	0
11.50	21.33	1.53	0.10	1	99.17	75.73	35.43	0.05	0
35.00	46.85	5.42	0.08	1	40.25	41.77	4.53	0.08	1
75.00	65.94	22.10	0.09	0	39.17	48.50	6.82	0.08	1
31.56	38.68	14.55	0.17	1	37.50	21.89	43.70	0.36	0
48.50	51.19	7.59	0.13	1	26.25	38.33	2.83	0.04	1
77.50	59.42	49.86	0.13	0	52.14	53.95	42.54	0.25	0
21.67	24.64	11.46	0.21	1	22.50	17.17	24.16	0.36	0
19.75	26.94	2.48	0.10	1	90.00	82.00	7.89	0.03	1
56.00	46.20	31.62	0.26	0	28.00	40.60	3.27	0.02	1
25.00	26.86	53.73	0.43	0	50.00	53.89	53.16	0.24	0
40.00	20.00	40.18	0.56	0	24.50	54.17	5.57	0.06	1
56.67	62.52	15.89	0.05	0					

7

建立模型 I：定义新的自变量

7.1 多项式回归

当一个自变量与一个响应变量 Y 之间的关系是平滑的，但不是一条直线时，如果有对 X 及 Y 的变换，使变换后有直线关系，则常可以使用线性模型。此外，因为任何光滑的函数可用足够高阶的多项式来近似，我们可以通过附加项来扩展模型。附加项是自变量的幂，所得到的模型为

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + e \quad (7.1)$$

其中 d 是多项式的阶；如果 $d=2$ ，模型是二次的， $d=3$ 为三次的，等等。在 (7.1) 中，我们假设误差是独立的，有常数方差 σ^2 ，或者 $\text{var}(e_i) = \sigma^2/w_i$ ，其中 $w_i > 0$ 是已知的。当方差不是一个常数时，在拟合一个多项式之前，需要有方差稳定化变换。多项式模型一般用作近似，几乎从来不用于代表一个物理模型。

可以用最小二乘法分析模型 (7.1)。定义 d 个新变量， Z_1, Z_2, \dots, Z_d ，它们是， $Z_1 = X, Z_2 = X^2, \dots, Z_d = X^d$ 。则 (7.1) 被写成

$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_d Z_d + e \quad (7.2)$$

这样，最小二乘法程序原则上可以计算出 β 的估计以及 Y 关于这些 Z_i 的回归的其它常用回归统计量。然而，如果 d 是大的，如 3 或更大，则可能产生严重的数值问题，并且直接拟合 (7.2) 可能是不可靠的。通过中心化变换，使 $Z_k = (X - \bar{X})^k$, $k=1, \dots, d$ ，可以保持数值精度。Seber (1977, 第八章) 研究了更好的办法。

一个二次回归的例子已由例 4.2 的物理数据给出。在那里，拟合失真的检验指出，直线模型对数据是不合适的，而对二次模型的拟合失真的检验指出，这一模型对数据是合适的。在拟合失真的检验不能得到，或不适用的时候，二次模型

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \text{误差} \quad (7.3)$$

与简单线性回归模型

$$Y = \beta_0 + \beta_1 X + \text{误差} \quad (7.4)$$

的比较，通常是基于 (7.3) 中对 $\beta_2=0$ 的 t -检验。事实上，选取 d 值的一个策略是，连续向模型中添加项，直到最高阶项的 t -检验是不显著的；或者，可以使用一个删除策略，它固定 d 的最大值，每次从模型中删去一项，即从最高阶项开始，直到剩下的最高阶项有一个显著的 t -值。Kennedy 和 Bancroft (1971) 提议，对这一过程使用的显著性水平约为 0.10。不过，在绝大多数多项式回归的使用中，只要考虑 $d=1$ 或 $d=2$ 就足够了。对于更大的 d 值，拟合的多项式曲线变为有波动的。它对观测数据中的波动给出一个更好的，更加密切的拟合。这样的曲线不是对变量间的关系的整体形状建立模型，而是对随机波动建立模型。

有多个自变量的多项式 扩展成多个变量的多项式是容易的。每增加多项式的一个项，便是增加了一个新的自变量。这也使模型有可能依赖于多个自变量的叉积项。例如，如下形式的模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \text{误差} \quad (7.5)$$

它有两个自变量。为了了解项 $\beta_{12} X_1 X_2$ 的作用，首先假设 β_{12} 为已

知的, 且 $\beta_{12} = 0$ 。如果 X_1 变为 $X_1 + \delta$, 则响应变量 Y 将由 (7.5) 变为 Y' ,

$$Y' = \beta_0 + \beta_1(X_1 + \delta) + \beta_{11}(X_1 + \delta)^2 + \beta_2 \cdot X_2 + \beta_{22} \cdot X_2^2 \quad (7.6)$$

Y 的变化为

$$\Delta Y = Y' - Y = \beta_1 \delta + \beta_{11}(2X_1 \delta + \delta^2) \quad (7.7)$$

故 ΔY 依赖于 X_1 , 但不依赖于 X_2 。现在, 若 $\beta_{12} \neq 0$, 且 X_1 变为 $X_1 + \delta$, 则 Y 的变化为

$$\Delta Y = \beta_1 \delta + \beta_{11}(2X_1 \delta + \delta^2) + \beta_{12} \delta X_2 \quad (7.8)$$

改变 X_1 而引起的 Y 的变化依赖于 X_1 和 X_2 。这样, 如果 $\beta_{12} \neq 0$, 则建立了 X_1 与 X_2 交互作用的模型。

响应曲面 用于估计一个多项式参数的试验设计, 可能是用于求诸 X 的一个组合, 以得到 Y 的最大或最小值。它称为响应曲面设计。关于它们的讨论由 Box 和 Wilson (1951), John (1971), Myers (1971) 及 Box, Hunter 和 Hunter (1978) 给出。

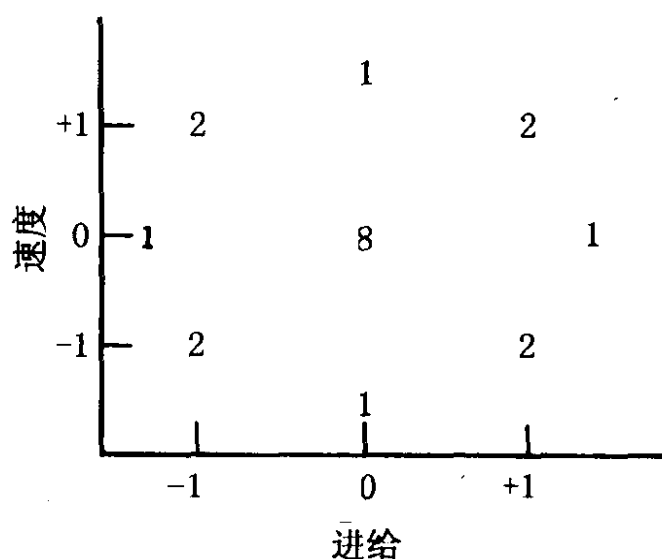
例 7.1 车床试验

在车床上进行一项试验, 其目的是了解切割工具的材料在切割钢材时的情况。表 7.1 的数据为试验结果的特征值。完全随机地做了 20 次试验。它包括两个因子, 切割速度 $speed$ (英尺/分钟) 和进给率 $Feed$ (千分之一英寸/转)。为方便起见, 两个因子的水平都被编码, 且被中心化, 即有 $S = (\text{速度} - 900) / 300$ 及 $F = (\text{进给率} - 13) / 6$ 。响应变量为 $Y = \text{工具寿命 (分钟)}$ 。图 7.1 是 S 关于 F 的散点图。图上的数字表示在每种试验配置下的重复试验的次数。这些点的排列称为中心复合设计。在拟合多项式时它是有用的。它允许考虑交互作用。由于某些试验条件, 特别是中心点被重复, 可以得到纯误差来估计模型的拟合失真的情况。

对数据的观察可以发现, 工具寿命的波动是很大的。在速率很高时, 少到只有一分钟, 而在低的 (F, S) 组合时, 多至一个小时。另外, 工具寿命越长, 其波动越大。这些观测现象表明, 需要对寿命进行变换, 可能是取对数。如果将 Box 和 Cox 的方法应用于由

$$Y = \beta_0 + \beta_1 S + \beta_2 F + \beta_3 S^2 + \beta_4 F^2 + \beta_5 S \times F + e$$

给出的完全二次模型, 则可以确认对 Y 取对数的有用性。我们取以 10 为底



(图中的数字为在给定的速度与进给下的重复试验次数)

图 7.1 S 对 F 的散点图

的对数，使用该完全二次模型

$$\log(Y) = \beta_0 + \beta_1 S + \beta_2 F + \beta_3 S^2 + \beta_4 F^2 + \beta_5 S \times F + e \quad (7.9)$$

对模型 (7.9) 的回归分析简要由表 7.2 给出。我们看到，除了 SF ，交互作用项，所有的系数都有相当大的 t -值。拟合失真的 F -检验值为 $F=4.09$ ，自由度为 (3, 11)， P -值约为 0.04。如果从模型中除去 SF ，我们得到 $F=3.26$ ，自由度 (4, 11)， P -值约为 0.05。9 个试验条件下的拟合值及观测所得的平均 \log (寿命) 值如图 7.2 所示。我们可以看到，除了点 $(\pm\sqrt{2}, 0)$ 以外，其它点的拟合值与平均值匹配得很好。在 $(+\sqrt{2}, 0)$ ，单个观测值为 0.4 分钟，其拟合值为 -0.210，对应的工具寿命有 $10^{-0.210}=0.6$ 分钟。如果从数据中删去这一案例，并重新拟合回归，则拟合失真的 F -检验值 $F=2.59$ ，自由度 (2, 11)，给出 P -值约为 0.12。如果进给轴上的另一个异常点 $(-\sqrt{2}, 0)$ 也被删除，并重新拟合模型。则拟合失真的 F -检验值可被进一步减至 1.25，自由度 (1, 11)。这样，模型拟合失真的原因是，在相当“极端”的进给率下，寿命无法仅由速度和进给率来说明。对试验的其它部分，拟合可以被认为是合适的。

表 7.1 关于车床的一个试验

速度	进给	寿命	速度	进给	寿命
-1	-1	54.5	$-\sqrt{2}$	0	20.1
-1	-1	66.0	$\sqrt{2}$	0	2.9
1	-1	11.8	0	0	3.8
1	-1	14.0	0	0	2.2
-1	1	5.2	0	0	3.2
-1	1	3.0	0	0	4.0
1	1	0.8	0	0	2.8
1	1	0.5	0	0	3.2
0	$-\sqrt{2}$	86.5	0	0	4.0
0	$\sqrt{2}$	0.4	0	0	3.5

来源: M. R. Delozier.

表 7.2 完全二次模型的 log (寿命) 的回归

变量	估计	标准误	t-值
截距	0.5160	0.0456	11.31
S	-0.6901	0.0373	-18.52
F	-0.3432	0.0373	-9.21
S ²	0.1251	0.0437	2.86
F ²	0.1818	0.0437	4.16
S×F	-0.0316	0.0456	-0.69

$\hat{\sigma}^2 = 0.1291$, d. f. = 14, $R^2 = 9706$

方差分析

来源	d. f.	SS	MS	F
回归	5	7.6003	1.5201	
拟合不佳	3	0.1228	0.0409	4.09
纯误差	11	0.1104	0.0100	

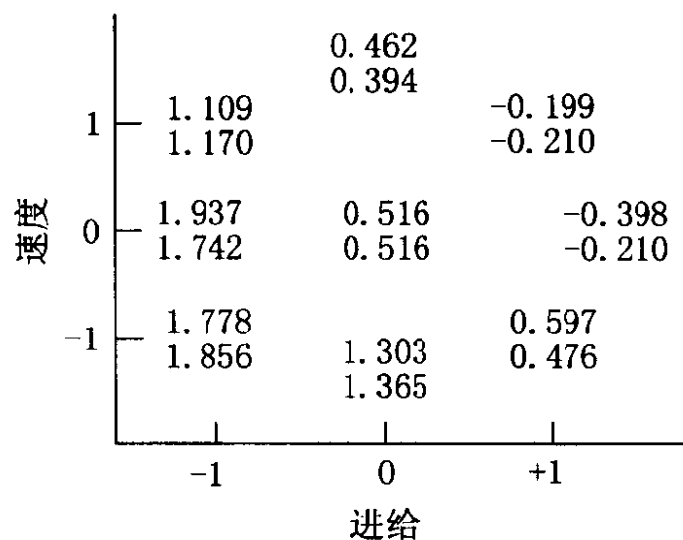


图 7.2 在试验设计的九种组合下的 \log (寿命) 的均值
(上一个数) 及 \log (寿命) 的拟合值 (下一个数)

7.2 虚拟变量：二分类的

虚拟变量，又称指示变量。在回归分析中包含分类自变量时，它将被使用。有许多分成两类的变量，例如男性或女性，处理或未处理，有病或无病。虚拟变量通常假设只取两个值 0 和 1。对一个给定的案例，0 和 1 分别表示哪一类是正确的。用 0 表示，还是用 1 表示，一般是任意的。

例 7.2 估计 (人工降雨的) 云的催化的效应

一个重要的问题是，判断用于增加降雨量的云的催化是成功的。还是失败的。云的催化的实验结果一般是各式各样的。有时催化的效应是增加降雨量，有时是减少降雨量，而有时观察不到效应。可以推测，这些各式各样的效应至少一部分是因为不能直接控制或测量试验中的潜在的变化。此外，催化的效应可能是相对小的，但小的变化却是重要的。

表 7.3 给出的数据取自佛罗里达地区的积云试验 (FACE)。数据收集于 1975 年 (Woodley et al. 1977)。试验情况简述如下，在佛罗里达的 Coral Gables 的东北部建立了一个约 3000 平方英里的固定目标区域。在 1975 年的夏天，每天都要判断当天是否适宜做云的催化的试验。决定在某一天做催化

试验,主要是根据一个适宜标准 S , 这里 S 是一个依赖于降雨的数学模型的计算值。 $S \geq 1.5$ 的那一天可以做试验。在 1975 年选中了 24 天。在选中的每一天,通过抛硬币来决定是否进行催化试验。结果是 12 天进行催化,12 天不进行。在催化的日子里,从小飞机向云里射入碘化银。总计,每个试验的日子里,要测量以下的量:

A = 行动 (1 = 催化, 0 = 不催化)。

D = 从试验的第一天起算的天数 (June 16, 1975 = 0)。

S = 催化的适宜性。

C = 试验地区的云的覆盖率,用佛罗里达的 Coral Gables 的雷达测量。

P = 前湿度,在催化前一小时的降雨量,以 10^7 立方米为单位。

E = (雷达的) 回波行为分类, 1 或 2, 为云的类型的一个度量。

Y = 在催化后或没有催化后的降雨量 (10^7 立方米)。

我们研究的问题是,估计由于催化引起降雨的效应。另一个令人感兴趣的,但在此不在任何细节上进行考虑的问题是,通过其它可测量的变量建立降雨的模型。在试验中使用随机化是因为降雨可能受到研究中没有包括的许多其它因素的影响。我们希望在整个试验过程中,这些其它因素能够平衡,而且不论催化,还是没有催化,这些其它因素产生相同的效果。有了这一点,尽管降雨与其它自变量之间的关系很可能要复杂得多,我们可以考虑降雨的一个线性模型。如通常一样,我们希望拟合的线性模型给出对事物真实状况的一个有效近似。

我们必须仔细考虑由虚拟变量 A 引起的效应的可能具有的性质。这一效果,如果它存在的话,可以至少在两个不同的方面证实它自己。首先,降雨的效应是可加的。其次,对其它变量的任何一个组合,云的催化的效应是增加或减少降雨量的某个固定的量,且不依赖于任何其它变量。如果这点成立,我们具有以下形式的模型。

$$Y = \beta_0 + \beta_1 A + (D, S, C, E \text{ 和 } P \text{ 的一个函数}) + e$$

在等式中,其它自变量的函数没有准确地表示出来,因为它们可能相互作用,或需要进行变换,可能需要多项式项,或从模型中最好删去某些自变量。如果这一模型对云的催化试验是合适的,则催化的效应将是平均降雨量的一个变动,改变量是 β_1 立方米 ($\times 10^7$), 且不受其它自变量的影响。

关于有单个附加变量或协变量的问题,模型的意义可从图 7.3 中的图形看出。在这个图中, X 轴是协变量的值, Y 轴是响应变量的值。图中分别作出 $A=0$ 和 $A=1$ 的回归直线。在可加处理效应模型,两条直线是平行的。

另一个较易理解的催化效应,对除虚拟变量 A 以外的其它自变量的不

表 7.3 云的催化数据

案例	A	D	S	C	P	$\log(P)$	E	SA	CA	PA	$\log(P)A$	EA	Y	$\log(Y)$
1	0	0	1.75	13.40	0.274	-0.56225	2	0	0	0	0	0	12.85	1.10890
2	1	1	2.70	37.90	1.267	0.10278	1	2.70	37.90	1.267	0.10278	1	5.52	0.74194
3	1	3	4.10	3.90	0.198	-0.70333	2	4.10	3.90	0.198	-0.70333	2	6.29	0.79865
4	0	4	2.35	5.30	0.526	-0.27901	1	0	0	0	0	0	6.11	0.78604
5	1	6	4.25	7.10	0.250	-0.60206	1	4.25	7.10	0.250	-0.60206	1	2.45	0.38917
6	0	9	1.60	6.90	0.018	-1.74473	2	0	0	0	0	0	3.61	0.55751
7	0	18	1.30	4.60	0.307	-0.51286	1	0	0	0	0	0	0.47	-0.32790
8	0	25	3.35	4.90	0.194	-0.71220	1	0	0	0	0	0	4.56	0.65896
9	0	27	2.85	12.10	0.751	-0.12436	1	0	0	0	0	0	6.35	0.80277
10	1	28	2.20	5.20	0.084	-1.07572	1	2.20	5.20	0.084	-1.07572	1	5.06	0.70415
11	1	29	4.40	4.10	0.236	-0.62709	1	4.40	4.10	0.236	-0.62709	1	2.76	0.44091
12	1	32	3.10	2.80	0.214	-0.66959	1	3.10	2.80	0.214	-0.66959	1	4.05	0.60746
13	0	33	3.95	6.80	0.796	-0.09909	1	0	0	0	0	0	5.74	0.75891

(续表)

案例	A	D	S	C	P	log(P)	E	SA	CA	PA	log(P)A	EA	Y	log(Y)
14	1	35	2.90	3.00	0.124	-0.90658	1	2.90	3.00	0.124	-0.90658	1	4.84	0.68485
15	1	38	2.05	7.00	0.144	-0.84164	1	2.05	7.00	0.144	-0.84164	1	11.86	1.07408
16	0	39	4.00	11.30	0.398	-0.40012	1	0	0	0	0	0	4.45	0.64836
17	0	53	3.35	4.20	0.237	-0.62525	2	0	0	0	0	0	3.66	0.56348
18	1	55	3.70	3.30	0.960	-0.01773	1	3.70	3.30	0.960	-0.01773	1	4.22	0.62531
19	0	56	3.80	2.20	0.230	-0.63827	1	0	0	0	0	0	1.16	0.06446
20	1	59	3.40	6.50	0.142	-0.84771	2	3.40	6.50	0.142	-0.84771	2	5.45	0.73640
21	1	65	3.15	3.10	0.073	-1.13668	1	3.15	3.10	0.073	-1.13668	1	2.02	0.30535
22	0	68	3.15	2.60	0.136	-0.86646	1	0	0	0	0	0	0.82	-0.08619
23	1	82	4.01	8.30	0.123	-0.91009	1	4.01	8.30	0.123	-0.91009	1	1.09	0.03743
24	0	83	4.65	7.40	0.168	-0.77469	1	0	0	0	0	0	0.28	-0.55284

同组合，其催化效应也是不同的。例如，以下结论可能为真：如果可催化标准 S 较小，则催化有小的效应，但如果可催化标准 S 较大，则催化的效应是大的。由图 7.4 可见，对应于催化处理和无催化处理的两条回归直线是不平行的。当两条曲线平行时，处理效应被清楚地定义为两条曲线之间的距离。这是理想的情况。在非平行的情况，处理的效应依赖于协变量的值，可能对协变量的某些值为正，而另一些为负。

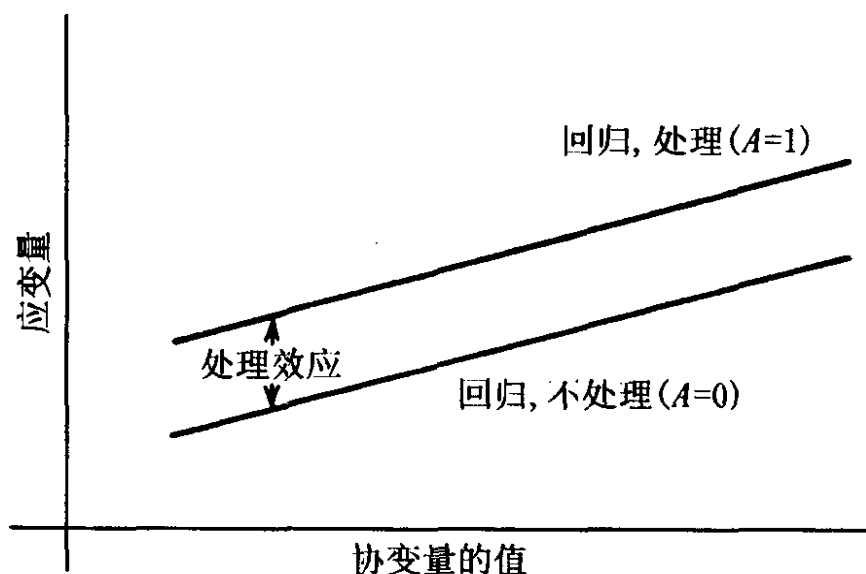


图 7.3 可加处理效应

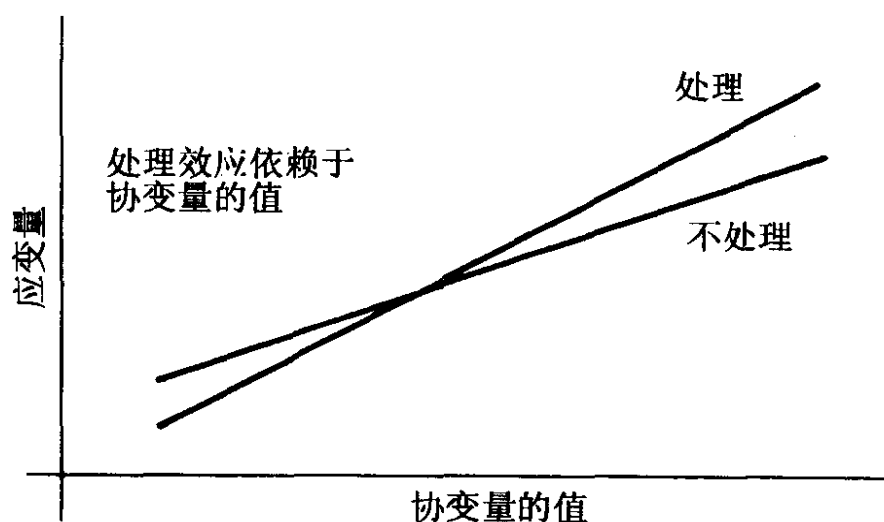


图 7.4 非可加处理效应

非平行回归的模型比简单可加处理效应模型要复杂。我们可以用两

个几乎等价的方法表示这一点。首先，我们可以分别对催化 ($A=1$) 与没有催化 ($A=0$) 的日子，建立模型方程。用这一方法，我们还能得到误差方差的两个估计，每组一个。如果可以假设每组的误差方差相同，一个可行的且更简单的方法，是对所有数据写出单个的方程。这通过定义自变量的一个新的集合完成。集合为催化指示 (A) 与其它变量 (S, C, E 和 P) 的乘积。这四个新变量为 $SA=S \times A$, $CA=C \times A$, $EA=E \times A$ 及 $PA=P \times A$ 。它们在表 7.3 中给出。例如， SA 的值当 $A=1$ 时为 S ，当 $A=0$ 时为 0。完整的模型为

$$Y = \beta_0 + \beta_1 A + \beta_2 D + \beta_3 S + \beta_4 C + \beta_5 E + \beta_6 P \\ + \beta_7 SA + \beta_8 CA + \beta_9 EA + \beta_{10} PA + e \quad (7.10)$$

如果处理的效应不依赖于 S, C, E 或 P ，则 $\beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$ 这时我们说处理效应是可加的。如果根本没有处理效应，则 $\beta_1 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$ 。在非可加处理效应可疑时的一个分析中，(7.10) 是一个合理的首选模型。

在分析这些数据时，下一个要关心的是，得到变量的合适的变换。由于 P 和 Y 是体积的尺寸，两者的变化超过 10 的一个指数幂，故对它们进行对数或立方根变换，可能是合适的。用 $\log(Y)$ 作为响应变量，具有排除降雨量的负的预测值的附加作用。这提示我们用下面的模型代替 (7.10)，

$$\log(Y) = \beta_0 + \beta_1 A + \beta_2 D + \beta_3 S + \beta_4 C + \beta_5 E + \beta_6 (\log P) \\ + \beta_7 SA + \beta_8 CA + \beta_9 EA + \beta_{10} (\log P) A + e \quad (7.11)$$

在这一模型中，不是以立方米，而是以 $\log(10^7 \text{ 立方米})$ 为单位，测量云的催化效应。这对于判断是否存在一个影响，同样是有用的。

t_i , D_i 及 h_{ii} 的索引图如图 7.5 所示。因为 $D_2=1.51$, $t_7=-4.03$ 和 $t_{24}=-3.62$ ，故我们对三个案例感兴趣；案例 2, 7 和 24。由于 D_2 的值较大，故删除案例 2 可能会在拟合模型中引起重大变化。由表 7.3 我们看到，那天云的覆盖率达到一个相当大的值，为 37.9%。原始论文的作者将它归类为一个“被干扰的日子”。案例 2 与其它数据明显不同，在分析之前删除它可能是一个合理的步骤。不过，这不是一般规则。删除一个有影响力的案例，只应在有理由相信这一案例确实与数据中其余案例不同时才能进行。案例 7 和 24 两者都有非常大的 t_i 值。对异常值检验，它们的 P -值都接近 0.05。参考原始数据，我们看到这两天都没有进行催化试验，其降雨量的观测值都很小。由与它们关联的 h_{ii} 可见，没有一个是特别不寻常的。由 D_i 看出，它们对系数的估计，至少对模型 (7.11) 系数的估计，也不是相当有影响的。关于案例 7 和 24 的进一步信息，可从非常数方差的得分检验得到。当数据包括所有 24 个案例并使用模型 (7.11) 时，方差作为拟合值的一个函数的非常数方差的

检验为 $S=5.38$ ，其自由度为 1。当删除案例 2, 7 和 24 后，我们得到 $S=0.66$ 。由于案例 7 和 24 被判为可能的异常值，并且与常数方差还是非常数方差的问题明显有关，我们在进一步分析前必须将它们删除。一个更彻底的方法要求包含这些点的重复分析。见 Cook 和 Weisberg (1980, 1982a)。

在继续分析之前，我们删除案例 2, 7 和 24，留下 21 个案例。由拟合 (7.11) 所得的残差作在图 7.6 中，其回归分析简要由表 7.4 给出。图及主要统计量表示对假设无明显违反。表 7.4 中小的 t -值指出，拟合模型可以通过删除某些变量加以改进。使用将在下一章讨论的变量的选择的方法，我们得到表 7.5 给出的模型的主要统计量。拟合模型为

$$\begin{aligned} \hat{\log(Y)} = & 0.492 + 1.294A - 0.007D + 0.022C + 0.399(\log(P)) \\ & + 0.301E - 0.326SA \end{aligned} \quad (7.12)$$

SA 的系数是非零的，催化的效应依赖于 S 。从而我们处于图 7.4 所示的情况之下。为研究这一依赖关系，倘若在一个非催化日进行催化，变化为

$$\begin{aligned} \Delta[\hat{\log(Y)}] &= [(\hat{\log(Y)}), \text{若 } A=1] - [(\hat{\log(Y)}), \text{若 } A=0] \\ &= \hat{\beta}_1 + \hat{\beta}_7(SA) \\ &= 1.294 - 0.326S \end{aligned} \quad (7.13)$$

表 7.4 主要回归统计量，模型 (7.11)，3 个案例被删除

变量	估计	标准误	t -值
截距	.4171	.3995	1.04
A	1.4263	.5100	2.80
D	-.0063	.0017	-3.72
S	.0060	.0085	.07
C	.0301	.0150	2.00
$\log(P)$.3414	.1458	2.34
E	.2652	.1353	1.96
SA	-.3334	.1073	-3.11
CA	-.0229	.0282	-.81
$\log(P)A$.0730	.2236	.33
EA	.0500	.1783	.28
$\hat{\sigma}^2=0.0194$, d.f. =10, $R^2=0.90$			

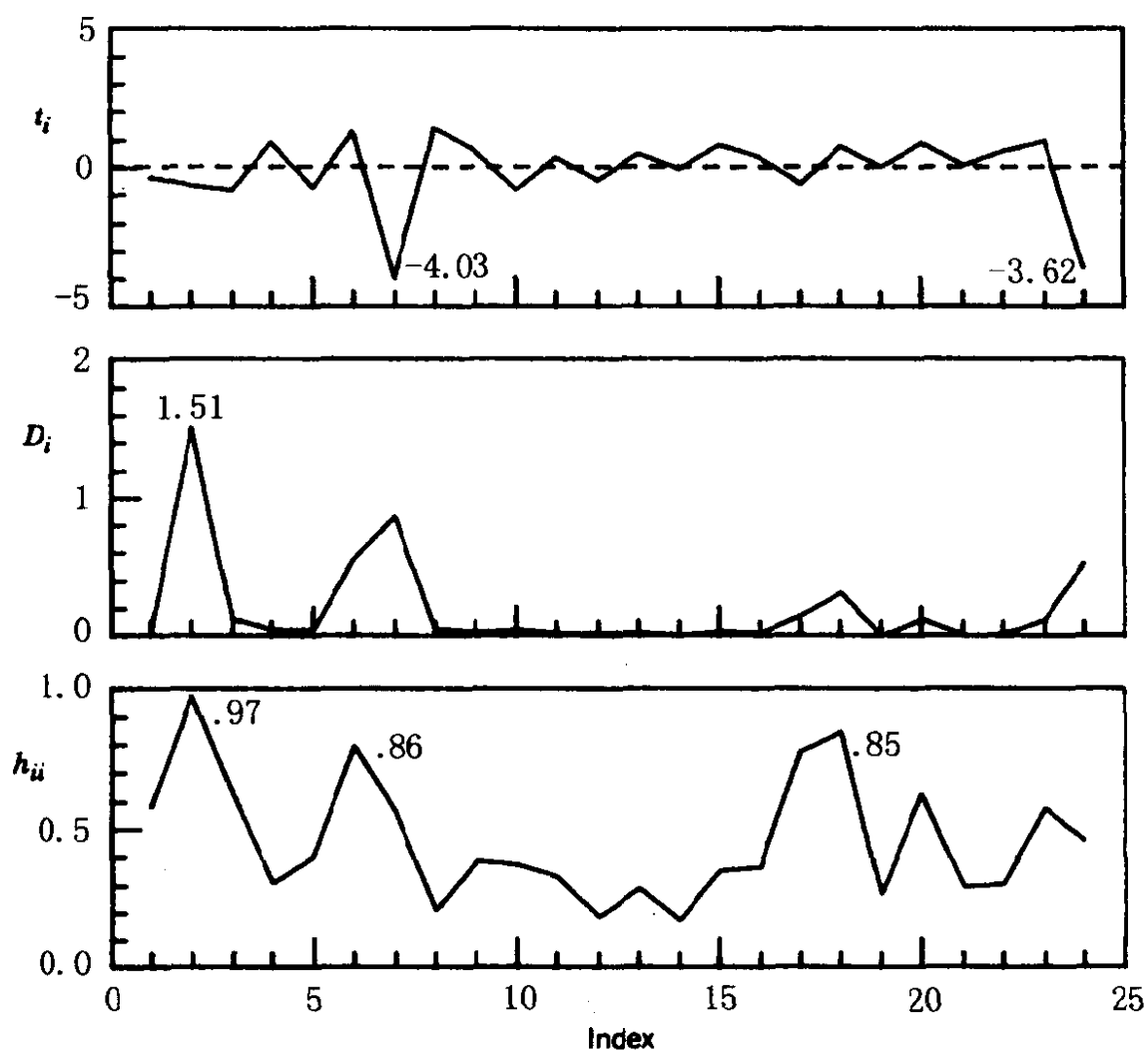


图 7.5 云催化数据的诊断统计量

表 7.5 $\log(Y)$ 的最终模型, 3 个案例被删除

变量	估计	标准误	t -值
截距	0.4925	0.1291	3.81
A	1.2944	0.1901	6.81
D	-0.0066	0.0013	-5.26
C	0.0219	0.0097	2.26
$\log(P)$	0.3990	0.0830	4.80
E	0.3010	0.0735	4.09
SA	-0.3263	0.0520	-6.28
$\hat{\sigma}^2 = .0149$, d.f. = 14, $R^2 = .89$			

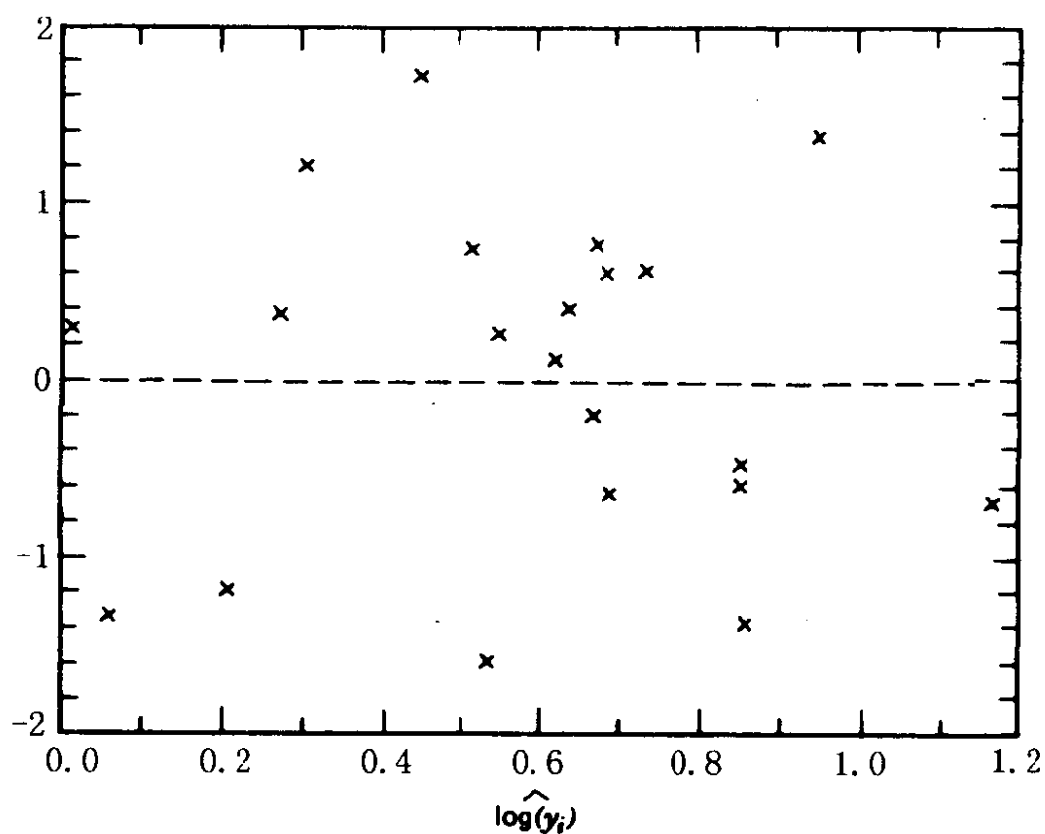


图 7.6 残差图，完整模型，案例 2，7 及 24 被删除

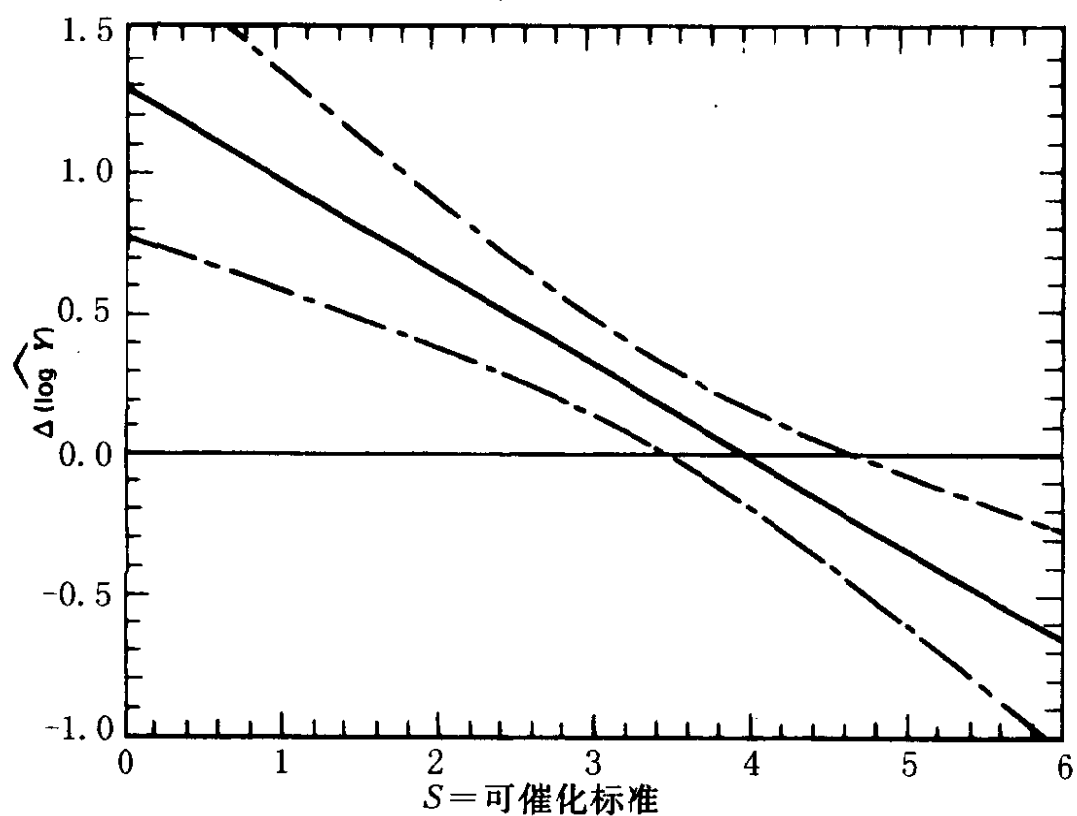


图 7.7 $\log(\text{降雨量})$ 的增加量作为可催化性的一个函数

图 7.7 是 $\log(\text{降雨量})$ 的拟合值的变化作为可催化标准 S 的一个函数的图。我们有如下有趣的结论：当催化的适宜性增加时，降雨量的增加量看来在下降，并且在 $S > 4$ 时变为负值。图 7.7 中两条曲线间的垂直距离给出 $\Delta(\log(Y))$ 的 95% 的置信区间。为计算置信区间，我们首先要得到 $\Delta(\log(Y))$ 的方差

$$\text{var}(\Delta(\log(Y))) = \text{var}(\hat{\beta}_1) + S^2 \text{var}(\hat{\beta}_7) + 2S \text{cov}(\hat{\beta}_1, \hat{\beta}_7) \quad (7.14)$$

(7.14) 中这三项的估计方法在第 2 章中已有讨论。类似于 (1.38)，可以得到 95% 的置信限。

Kerr (1982) 报告了一项后继试验的结果。后继试验称为 FACE-2，进行于 1978 年。这个试验是严格随机化的，双重未知试验：直到试验结束，没有一个参加者知道哪天是催化的，哪天不催化。当最后完成分析时，FACE-2 中关于催化成功或失败的结论依赖于一个有非常大的降雨量的非催化日。如果删除那一天，催化被判断为有效；如果包括进来，则看不到催化的明显效应。

7.3 虚拟变量：多分类的

如果一个分类自变量有多于两个的类，则可能需要多个虚拟变量。例如，假设云的催化试验由三种试验条件组成：未催化的，用碘化银催化以及用干冰催化。我们可以考虑使用一个变量，0 表示未催化，1 表示用碘化银，2 表示用干冰。但这表示，三种处理是有顺序的，并且从未催化到碘化银催化的变化效应，与从碘化银到干冰的变化效应是相同的。在大多数问题中，类有序及相邻类之间等间距的假设都没有得到证实。因此，我们必须定义两个虚拟变量，

$A_1 = 1$ ，如果没有催化；0，否则

$A_2 = 1$ ，如果用碘化银；0，否则

某天使用干冰可由 $A_1 = A_2 = 0$ 唯一给出，故所有三个处理条件都可以用这两个虚拟变量表示。在有了这两个变量之后，只要模型中有截距，第三个变量

$A_3=1$ ，如果用干冰；0，否则就是多余的了。如果我们只考虑另外一个变量，如 S ，则可加的平行回归模型为

$$y = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_s S + e \quad (7.15)$$

对所有组，我们假设斜率 β_s 是相同的，则这三组的回归是平行的。这三组的截距分别为：用干冰时的 β_0 ，未催化时的 $\beta_0 + \beta_1$ ，及用碘化银时的 $\beta_0 + \beta_2$ 。关于 $\beta_1 = \beta_2 = 0$ (β_0 及 β_s 任意) 的 F —检验就是三种处理条件下的截距都是相等的检验。 A_1, A_2, A_3 中的任意两个都可用来建立模型，它们只是改变所得到的系数的含义，但导致相同的 F —检验。或者，我们可以不考虑截距，拟合模型

$$y = \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_s S + e \quad (7.16)$$

如此参数化的好处是，(7.16) 拟合中的每一个 $\hat{\beta}_j$ 就是那组的截距的估计。由于大部分的计算机程序可以直接产生 $\hat{\beta}_j$ 的标准误，故可以避免用公式 (2.26) 来得到它们的值。

模型 (7.15) 和 (7.16) 都对应于一个平行的回归，或可加处理效应模型。这两个模型都可以通过加入 $S \times A_j$ 项而加以推广。这也就是允许每一组中协变量的斜率可以是不同的。由于组容量可能较小，所以在有许多虚拟变量的问题中，考虑非平行回归而增加这种类型的交互作用，顿时就不适用了。分析者被迫对大多数的分类变量，假设平行回归模型。其所得到的回归分析，也只是和平行回归的假设，具有一样的可靠性。

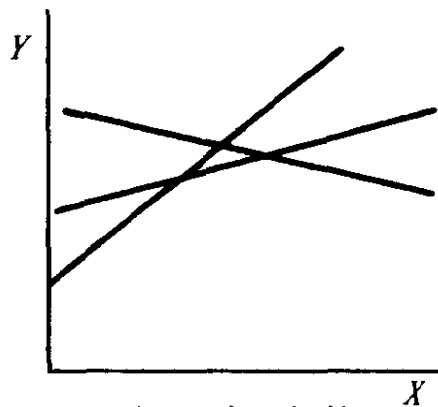
7.4 比较回归直线

为简单起见，考虑 m 组案例的简单回归。例如，第 k 组有 n_k 个案例，正确的模型为

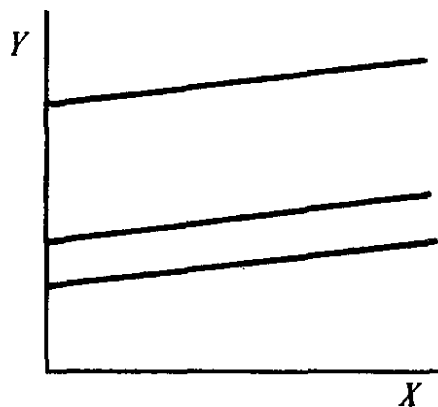
$$Y = \beta_{0k} + \beta_{1k} X + \text{误差} \quad (k = 1, 2, \dots, m)$$

基于观测数据，比较这 m 条回归直线，常是我们感兴趣的。我们区分四种不同情况：

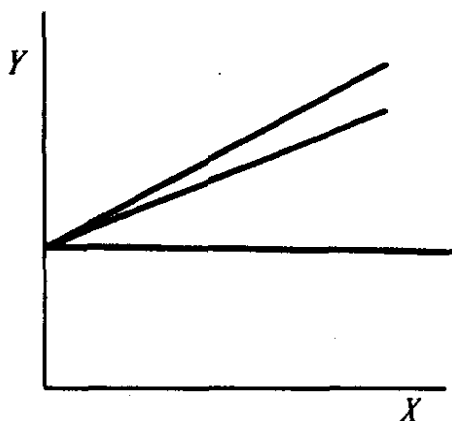
模型 1: 最一般化的。如果所有参数是不同的, 我们有如图 7.8 (a) 所示的情况。



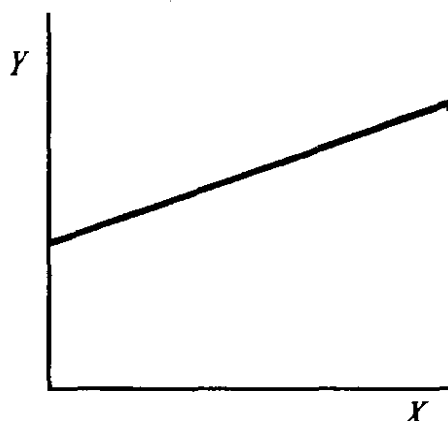
(a) 最为一般的



(b) 平行的



(c) 共点的



(d) 重合的

图 7.8 比较回归的四个模型

模型 2: 平行回归。在这个模型中 $\beta_{11} = \beta_{12} = \dots = \beta_{1m}$, 但截距是任意的。这些是导致虚拟变量的可加模型的情况。如 7.3 节所述, 图 7.8 (b) 所表示的。

模型 3: 共点回归。在这个模型中的所有截距都相等, $\beta_{01} = \dots = \beta_{0m}$, 但斜率是任意的, 如图 7.8 (c) 所示。

模型 4: 重合回归直线。这里所有的直线是相同的。 $\beta_{01} = \dots = \beta_{0m}$ 且 $\beta_{11} = \dots = \beta_{1m}$ 。这是最严格的模型, 如图 7.8 (d) 所示。

我们在检验模型 4 或 2 的可行性时,通常对用一个不同的,不是非常严格的模型作为备择假设感兴趣。这些检验的形式,从 4.4 节给出的广义 F -检验的公式立即可以得到。该方法通过下一个例子能加以最好的表述。

例 7.3 双胞胎数据

表 7.6 的数据给出同卵双生的双胞胎的 IQ (智商) 分数,其中一个在寄养别人孩子的家庭里长大 (Y),另一个由生父母抚养 (X)。数据原来为 Burt (1966) 所用。为举例的目的,我们根据生父母的社会地位将案例分为三组。

本例中, $m=3$, $n_1=7$, $n_2=6$, $n_3=14$, 和 $n=\sum n_k=27$ 。数据画在图 7.9 中。该图显示出,回归直线很可能是一样的。对于检验问题,我们将在家抚养的双胞胎的 IQ (X) 作为标准,和在寄养别人孩子的家庭长大的双胞胎的 IQ (Y) 作比较,研究分开抚养的效应。除了对不同社会阶层的不同直线作比较外,我们还对斜率估计的准确值,和它是否不为 1 感兴趣。但在这里不讨论这一点。

表 7.6 比较回归直线的双胞胎数据

案例编号	Y	X	G_1	G_2	G_3	Z_1	Z_2	Z_3
1	82	82	1	0	0	82	0	0
2	80	90	1	0	0	90	0	0
3	88	91	1	0	0	91	0	0
4	108	115	1	0	0	115	0	0
5	116	115	1	0	0	115	0	0
6	117	129	1	0	0	129	0	0
7	132	131	1	0	0	131	0	0
8	71	78	0	1	0	0	78	0
9	75	79	0	1	0	0	79	0
10	93	82	0	1	0	0	82	0
11	95	97	0	1	0	0	97	0

(续表)

案例编号	Y	X	G ₁	G ₂	G ₃	Z ₁	Z ₂	Z ₃
12	88	100	0	1	0	0	100	0
13	111	107	0	1	0	0	107	0
14	63	68	0	0	1	0	0	68
15	77	73	0	0	1	0	0	73
16	86	81	0	0	1	0	0	81
17	83	85	0	0	1	0	0	85
18	93	87	0	0	1	0	0	87
19	97	87	0	0	1	0	0	87
20	87	93	0	0	1	0	0	93
21	94	94	0	0	1	0	0	94
22	96	95	0	0	1	0	0	95
23	112	97	0	0	1	0	0	97
24	113	97	0	0	1	0	0	97
25	106	103	0	0	1	0	0	103
26	107	106	0	0	1	0	0	106
27	98	111	0	0	1	0	0	111

除了 X 和 Y 以外, 还给出六个变量, 即三个虚拟变量 G_1 , G_2 和 G_3 , 用于表示社会阶层 ($G_1=1$ 为最高阶层的, $G_2=1$ 为中等的, $G_3=1$ 为最低的), 以及三个附加变量 $Z_1=G_1X$, $Z_2=G_2X$ 和 $Z_3=G_3X$ 。这些变量将有助于在拟合四个模型时比较回归直线。

模型 1: 为拟合模型 1, 分别进行回归计算, 以得到每一组的参数估计。其 RSS , 记为 RSS_1 , 等于从每个回归得到的 RSS 的和。这个 RSS 的自由度为 $df_1 = \sum (n_k - p')$ (对简单回归, $p' = 2$)。或者等价地, 我们可以拟合模型

$$Y = \beta_{01}G_1 + \beta_{02}G_2 + \beta_{03}G_3 + \beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3 + e$$

其中 G 和 Z 由表 7.6 给出 (这一模型中没有总的截距)。其拟合的结果由表 7.7 给出。

模型 2: 为拟合平行回归, 使用模型

$$Y = \beta_{01}G_1 + \beta_{02}G_2 + \beta_{03}G_3 + \beta_1X + \text{残差}$$

由此获得每一个截距及共同斜率 β_1 的估计。这一拟合的 RSS , 记为 RSS_2 , 有 $df_2 = n - m - p$ 的自由度 (对简单回归, $p = 1$)。这一拟合的主要统计量由表 7.7 给出。

模型 3: 这一模型需要拟合

$$Y = \beta_0 + \beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3 + \text{误差}$$

这一模型的 RSS , 记为 RSS_3 , 有自由度 $df_3 = n - mp - 1$ 。 RSS_3 的值如表 7.7 所示。

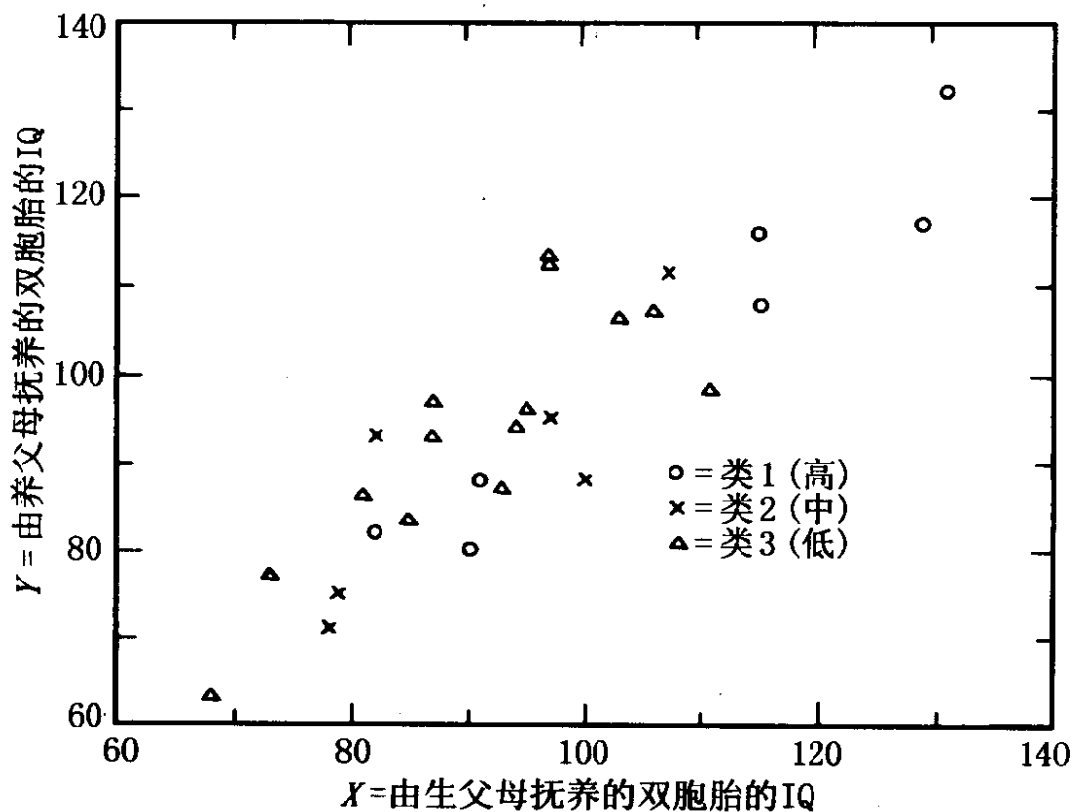


图 7.9 双胞胎数据

模型 4: 这一模型假设共同的回归直线, 故将数据合并在一起, 然后拟合模型

$$Y = \beta_0 + \beta_1X + \text{误差}$$

从而计算得出估计。这一模型的 RSS , RSS_4 , 有自由度 $df_4 = n - p'$ 。 RSS_4 的值如表 7.7 所示。

表 7.7 双胞胎数据的计算

估计 (t -值)				
变量	模型 1	模型 2	模型 3	模型 4
截距	—	—	2.56(0.24)	9.21(0.99)
X	—	0.97(9.03)		0.90(9.36)
G_1	-1.87(-0.11)	-0.61(-0.05)		
G_2	0.82(0.03)	1.43(0.14)		
G_3	7.20(0.43)	5.62(0.56)		
Z_1	0.98(5.99)		0.94(9.44)	
Z_2	0.97(3.40)		0.95(7.93)	
Z_3	0.95(5.21)		1.00(8.57)	
RSS	1317	1318	1326	1494
d.f.	21	23	23	25

大部分关于不同回归直线的斜率与截距的检验,将使用一般模型(模型 1)作为备择模型。为检验模型 2, 3 和 4, 通常的 F -检验由下式给出

$$F_l = \frac{(RSS_l - RSS_1) / (df_l - df_1)}{RSS_1 / df_1} \quad (l=2, 3, 4)$$

自由度为 $(df_l - df_1, df_1)$ (7.17)

如果假设给出与备择模型同样好的一个模型, 则 F 将较小。如果与一般模型相比, 模型是不合适的, 则 F 将较大(与 $F(df_l - df_1, df_1)$ 分布的分位点进行比较)。

对于双胞胎数据, F -统计量为

$$F_2 = \frac{(1318 - 1317) / 2}{1317 / 21} = 0.01$$

$$F_3 = \frac{(1326 - 1317) / 2}{1317 / 21} = 0.08$$

$$F_4 = \frac{(1494 - 1317) / 2}{1317 / 21} = 0.71$$

由于所有的 F 值都比 1 小得多, 与对应的 F 分布的分位点进行比较是不必要的。限制最多的模型, 模型 4, 与限制最少的模型, 模型 1 一样的好。 F_2

$=0.01$ 这一小值可以作为仔细的分析者的一个标志，这是因为在假设的模型下，观测到这么小的一个 F -统计量的可能性极小（如果所有假设成立，在这一问题中 F_2 比其观测到的值小的概率为 0.005）。观测数据与理论的吻合比随机理论表达所能平均达到的结果要好得多。

协方差分析 相对于模型 2，检验模型 4，通常称为协方差分析。自变量称为协变量，它被要求在各个组中具有相同的效应。组间的区别被要求为一个可加处理效应。把这两个要求合在一起，给出了以下假定：调整后的备择假设具有相等的斜率，但可能有不同的截距（截距间的距离为“处理效应”）。

附加评注 尽管本节给出的例子使用简单回归，但完全一样的方法也可用于比较有多个自变量的组。我们常设计计算机程序，使组的比较更容易。这些程序看起来使用三种方法。最简单的方法是允许用户在加权最小二乘法中用值 0 和 1 定义权值。给定权值为 0 的案例不用于计算。把它与设立虚拟变量与交互作用的能力结合在一起，能进行模型 1 至 4 中所有需要的计算。一个较老的方法，在程序 BMDP1V (Dixon, 1983) 中作为例子，为自动比较各组计算了某些检验量，但没有多用户控制。比较各组的最好的方法是 GLIM 程序所采用的。该程序由皇家统计协会公布。这一程序有易于理解的语言，用来识别因子，基本上为分类变量，以及协变量，基本上为连续变量，从而使拟合模型 1 至 4 特别容易。细节参见 McCullagh 和 Nelder (1983, 第 3 章)。

在比较各组时，最常见的问题可能是在简单回归中对两个组检验斜率是否相等。由于这一 F -检验的分子有 1 个自由度，故它等价于一个 t -检验。令 $\hat{\beta}_j$, $\hat{\sigma}_j^2$, n_j 和 SXX_j 分别为斜率的估计，残差均方，样本容量及在第 j 组 ($j=1, 2$) X 的校正平方和。则 σ^2 的一个合并估计为

$$\hat{\sigma}^2 = \left(\frac{(n_1 - 2)\hat{\sigma}_1^2 + (n_2 - 2)\hat{\sigma}_2^2}{n_1 + n_2 - 4} \right) \quad (7.18)$$

且斜率相等的 t -检验为

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\hat{\sigma}(1/SXX_1 + 1/SXX_2)^{1/2}} \quad (7.19)$$

其自由度为 $n_1 + n_2 - 4$ 。这个 t -统计量的平方在数值上等于相应的 F 统计量。

对共点回归模型，7.4 节的模型 3 可以容易地推广到回归直线在任何固定一点 $X=c$ 共点的情况。假设在例 7.3 的双胞胎数据中，我们希望检验在 $c=100$ ，IQ 的一般水准处是否共点。共点回归模型指出，直线在这一点重合，但在其它地方可能不同。这一模型可以按如下步骤去拟合：(1) 用 $X-100$ 代替 X ；(2) 如同该例，定义 G 和 Z ，但在定义 Z 时，使用重新定义的 X ；(3) 完全如文中所述，拟合模型 3。使用原来的 X 或重新定义的 X ，模型 1，2 和 4 都给出相同的 RSS 。这一模型在另一个方面的推广是，允许回归直线在某个任意的未知点共点，故交点必须由数据来估计。这一问题更加困难，Saw (1966) 进行了讨论。

7.5 变量的尺度

在有截距的模型中，最小二乘回归有个非常好的性质，称为位置与尺度不变性：如果数据中的任何变量通过增加一个常数，或乘以一个常数而改变尺度，则在所得的回归中，估计将以一种可以预料的方式改变；而尺度无关量，如 R^2 ， F 以及 t -检验将不受影响。在这一节中，我们将概要叙述，改变变量尺度的效果。令 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 和 $\hat{\sigma}^2$ 分别为变换尺度前得到的最小二乘估计。我们只给出变换尺度对这些量，以及 R^2 和检验统计量等的影响。由这些结果，可以容易地推导出它对标准误的估计的影响。

变换尺度 如果用 $(X_j - c_j) / d_j$ 代替变量 X_j ，则在得到的回归中， $\hat{\beta}_j$ 变为 $d_j \hat{\beta}_j$ ， $\hat{\beta}_0$ 变为 $\hat{\beta}_0 + c_j \hat{\beta}_j$ ，而 $\hat{\sigma}^2$ ， F -检验， t -检验及 R^2 未受影响。

变换响应变量的尺度 如果用 $(Y - f) / g$ 代替 Y ，则 $\hat{\beta}_0$ 被 $(\hat{\beta}_0 - f) / g$ 代替，每个 $\hat{\beta}_j$ ($j=1, \dots, p$) 被 $\hat{\beta}_j / g$ 代替，且在方差分析中的所有平方和及 $\hat{\sigma}^2$ 都被 g^2 除。不过， F 和 t -检验不受影响。

变换尺度通常通过减去样本均值并除以样本标准差，使变量标准化。这样，对 $j=1, \dots, p$, X_j 被

$$\frac{X_j - \bar{x}_j}{SD_j}$$

代替，而 Y 被

$$\frac{Y - \bar{y}}{SD_Y}$$

代替。在数据标准化后，截距的估计恰为零，而 β_j 的估计 $\hat{\beta}_j$ 变为

$$\hat{\beta}_j \text{ (标准化的)} = \frac{SD_j}{SD_Y} \cdot \hat{\beta}_j \quad (j=1, 2, \dots, p) \quad (7.20)$$

某些研究者对不同的自变量，比较它们标准化系数的估计。在这一逻辑下，有较大的标准化系数的自变量更为重要。不幸的是，由于标准化变换依赖于数据中变量的取值范围，所以这一推理是有缺陷的。例如，如果两个分析员对同一自变量收集数据，一个在小范围内收集数据，而另一个在较大的范围内收集，他们可能得到关于标准化系数的相对量值的完全不同的结论。另外，当所用的模型只是更为复杂的关系的近似时，我们过分相信了系数的估计。

7.6 线性变换及主成分

线性回归在位置/尺度变换下的不变性是回归在自变量的线性变换下的不变性的一个特例。在一个线性变换中，模型中的 p 个自变量被至多 p 个的，它们的（线性无关）线性组合所代替。我们已经见到一个线性变换的例子，这就是例 3.1 的“伯克来指导研究”的讨论。在那个例子中，三个原始变量 WT_2 , WT_9 及 WT_{18} 被 WT_2 , $WT_9 - WT_2$ 及 $WT_{18} - WT_9$ 所代替。而对这三者的回归，似乎比关于原始数据的回归，对所得的信息给出了更好的解释。

我们遇到的线性变换的另一个例子是在附录 2A.3 的 QR 因子分解，其中列正交矩阵 Q 为 X 的列的线性组合。由于 Q 有正交列，最小二乘计算非常简单。使用 Q 之后得到的结论向原始数据

的变换, 可以通过数值稳定的方法完成。

总之, 对一个 $n \times p'$ 的矩阵 X 的一个 (非奇异) 线性变换, 就是求得一个 $p' \times p'$ 矩阵 U (秩为 p'), 使变换后变量 Z 由下式给出

$$Z = XU \quad (7.21)$$

在前面所述的指导研究的例子中, U 为 4×4 , 且由下式给出

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (7.22)$$

由 $Z = XU$ 的直接相乘, 我们看到 Z 的第一列为 X 的第一列 (即元素全为 1 的列), Z 的第二列为 WT_2 , 第三列为 $WT_9 - WT_2$, 最后一列为 $WT_{18} - WT_9$ 。

如果线性模型为 $Y = X\beta + e$, 且 U 可逆, 定义 $\alpha = U^{-1}\beta$, 则

$$\begin{aligned} Y &= X\beta + e \\ &= X(UU^{-1})\beta + e \\ &= Z\alpha + e \end{aligned} \quad (7.23)$$

α 的最小二乘估计为 $\hat{\alpha} = (Z^T Z)^{-1} Z^T Y$, β 的最小二乘估计为 $\hat{\beta} = U\hat{\alpha}$ 。

主成分 通过选择合选的 U , 我们可以得到一个变换数据矩阵 Z , 它具有某些想要的性质。其最重要的一个例子是求得一个 U , 使 $Z^T Z = U^T (X^T X) U = D$, 其中 D 是对角线元素 $\lambda_1, \lambda_2, \dots, \lambda_p$ 皆为正的对角矩阵。我们假设这些 λ 是有序的, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 。 λ_p 仅在 $X^T X$ 为奇异时取零。我们可以证明以下事实。

1. 矩阵 U 是一个正交矩阵, $U^T U = U U^T = I$, 并且当所有的 λ 皆不同时, U 是唯一的。

2. U 的列是矩阵 $X^T X$ 的特征向量, 这些 λ 称为特征值。 $Z = XU$ 的列称为主成分。

3. 数据的尺度不同时, 特征值及特征向量也可能不同。这样, 基于 $X^T X$, $\mathcal{X}^T \mathcal{X}$ 及样本相关矩阵的计算可能导致不同的特征值和特征向量。当对数据进行尺度变换, 给出了不同样式的叉积矩

阵后，对大部分其它回归计算，结果没有本质上的变化。一般地，我们使用 $\mathcal{X}^T \mathcal{X}$ ，或更常用的样本相关矩阵的特征值与特征向量。

Stewart (1974) 及 Seber (1977) 讨论了特征值与特征向量的计算方法。在 IMSL 和 Eispack 库中都有 Fortran 子程序。后者的手册，Smith et al. (1976)，列出程序源代码。许多统计软件包，如 Minitab (Ryan, Joiner 及 Ryan, 1985) 计算特征值，特征向量及主成分，并允许它们用于其它计算。

例 7.4 伯克来指导研究

伯克来指导研究中的男孩，只使用变量 WT_2 , WT_9 及 WT_{18} 。 $\mathcal{X}^T \mathcal{X}$ 的特征向量矩阵及特征值为

$$U = \begin{bmatrix} 0.0354 & -0.3551 & 0.9341 \\ 0.2754 & -0.8951 & -0.3507 \\ 0.9607 & 0.2697 & 0.0662 \end{bmatrix}$$

$$(\lambda_1, \lambda_2, \lambda_3) = (3604.0, 246.2, 34.30)$$

三个主成分 Z_1 , Z_2 和 Z_3 定义为

$$Z_1 = 0.0354WT_2 + 0.2754WT_9 + 0.9607WT_{18}$$

$$Z_2 = -0.3551WT_2 - 0.8951WT_9 + 0.2697WT_{18}$$

$$Z_3 = 0.9341WT_2 - 0.3507WT_9 + 0.0662WT_{18}$$

我们可以证明主成分是如此定义的，它使：如果拟合模型

$$Y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + e$$

其中 $\text{var}(e) = \sigma^2$ ，则估计 $\hat{\alpha}_1$ 比 X 的其它任何的线性组合（其模为 1），的系数估计有更小的方差， $\hat{\alpha}_1$ 的方差为 σ^2/λ_1 。类似地，在所有与 Z_1 正交的 X 的线性组合中， Z_2 的系数的估计 $\hat{\alpha}_2$ ，其方差 σ^2/λ_2 小于其它任何线性组合。依此类推至所有其它主成分。在这个例子中， $\text{var}(\hat{\alpha}_1) = \sigma^2/3604 = 2.77 \times 10^{-4} \sigma^2$ ，而 $\text{var}(\hat{\alpha}_3) = \sigma^2/34.30 = 2.9 \times 10^{-2} \sigma^2$ 。在数据中可以得到的关于 α_1 的信息比关于 α_3 的要多得多。

如果研究的主成分不是基于 $\mathcal{X}^T \mathcal{X}$ ，而是从样本相关矩阵对它们进行计算，则可能得到完全不同的结论。例如，变换矩阵 U^* 及特征值 λ_1^* , λ_2^* , λ_3^* 为

$$U^* = \begin{bmatrix} 0.4925 & -0.7809 & -0.3843 \\ 0.6648 & 0.0525 & 0.7452 \\ 0.5618 & 0.6224 & -0.5450 \end{bmatrix}$$

$$(\lambda_1^*, \lambda_2^*, \lambda_3^*) = (2.028, 0.7890, 0.1829)$$

则近似地, 通过舍入系数, 得 3 个主成分为

$$Z_1 = 0.4925WT_2 + 0.6648WT_9 + 0.5618WT_{18}$$

$$\cong \frac{1}{2} (WT_2 + WT_9 + WT_{18})$$

= 平均重量的度量

$$Z_2 = -0.7809WT_2 + 0.0525WT_9 + 0.6224WT_{18}$$

$$\cong 0.7 (-WT_2 + WT_{18})$$

= 从 2 岁到 18 岁的线性重量增益

$$Z_3 = -0.3843WT_2 + 0.7452WT_9 - 0.5450WT_{18}$$

$$\cong 0.4 (-WT_2 + 2WT_9 - WT_{18})$$

= 二次重量增益

这样在相关形式中, 主成分有非常简单的解释, 而对基于 $\mathcal{X}^T \mathcal{X}$ 的运算, 无法得到简单的解释。那些运算表示, 第一主成分—— X 的组合, 近似等于 $0.3[WT_9 + 3(WT_{18})]$, 其系数被最为精确地估计。在这些结果中 WT_{18} 的重要性是由于 WT_{18} 的样本方差在 WT_2, WT_9 和 WT_{18} 的方差中为最大, 而较大的方差给出 WT_{18} 在主成分中更高的权值。在相关形式中, 消除了方差之间的差别, 所有三个变量以看起来有效的方法被使用。

多个变量的主成分 对任何变量集合都可以求得主成分, 例如在“伯克来指导研究”中 SOMA 的那些自变量。然而, 得到的变量可能是不相类似的变量, 如高度、重量及力量等的线性组合。新变量可能没有意义。用于指导研究的一个可选方法是, 和其它变量分开求出高度变量的主成分, 和其它变量分开求出重量变量的主成分等等, 将这些变换的自变量用于进一步的分析。

问 题

7.1 作为拟合模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \text{误差}$$

的另一个选择, 考虑拟合模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 - X_2) + \text{误差}$$

讨论这两个模型的区别。什么情况下, 第二个模型是合适的? 什么时候第一个更好? (提示: 在每个模型中, 如果 X_1 变为 $X_1 + \delta_1$, 和/或 X_2 变

为 $X_2 + \delta_2$, 响应变量将有何反应?)

- 7.2 比较 Fordes 数据 (例 1.1) 与 Hooker 数据 (问题 1.2) 的回归直线。
- 7.3 在问题 2.1 的“伯克来指导研究”数据中, 考虑由收集到的 2 岁至 9 岁的数据对 $HT18$ 建模。对男孩和女孩拟合相同的模型, 并施行合适的检验来比较拟合回归平面。
- 7.4 对“伯克来指导研究”数据中的女孩, 求三个重量变量的主成分。对文中给出的男孩进行同样的计算。如果可能的话, 阐述主成分的含义。
- 7.5 对物理数据 (问题 4.2), 严格比较问题 4.2 给出的回归直线。
- 7.6 对例 4.2 的苹果树枝数据, 进行完整的数据分析。
- 7.7 表 7.8 中的数据给出车床上切割钢材的工具寿命的另一项试验的结果。试验有三个因子: 速度 (英尺/分钟), 进给率 (千分之一英寸) 及钻头半径 (亦为千分之一英寸)。响应变量是工具寿命 (分钟)。这项试验也是完全随机的, 但它被安排在半径有三个水平及进给和速度分别有二个水平的全因子设计上。每个组合 (进给、速度、半径) 重复试验三次。使用到目前为止所学的任何模型, 求出可以合适地给出工具寿命的估计的一个模型。这里, 工具寿命的估计是试验中三个因子的函数。

表 7.8 车床的第二个试验

速度	进给	半径	工具寿命 (3 次重复试验)		
750	5	1	100.7	60.0	75.9
750	5	4	25.0	17.5	20.5
750	5	7	14.9	18.0	17.0
750	10	1	39.5	42.0	47.0
750	10	4	15.1	16.6	19.3
750	10	7	11.0	14.0	14.5
950	5	1	35.3	17.7	29.0
950	5	4	17.0	12.5	13.8
950	5	7	7.0	9.0	11.5
950	10	1	18.3	17.0	13.1
950	10	4	10.0	8.0	9.0
950	10	7	9.5	8.0	8.6

来源: M. R. DeLozier, Kennametal, Inc., Latrobe, Pennsylvania.

- 7.8 使用 Box 和 Cox 方法确定在云的催化数据 (例 7.2) 中响应变量的合适

的尺度，并将分析结果与文中给出的结果作定量比较。

- 7.9 哥德式与罗马式教堂 表 7.9 给出中世纪英国教堂的数据， X = 教堂的中殿高度及 Y = 总长度，它们都以英尺为单位。另外，教堂可以根据其建筑风格进行分类，或者为罗马式的，或者为后来的哥德式的。某些教堂同时具有哥德式与罗马式的部分，且各个部分有不同的高度。这类教堂在数据中出现两次。

使用这些数据回答问题：对这两种建筑风格，响应变量 Y = 长度与 X = 高度的关系是否一样。如果它们不同，略述它们的区别。

表 7.9 英国中世纪教堂的高度及长度

	高度	长度		高度	长度
罗马式			哥特式		
Durham	75	502	York	100	519
Canterbury	80	522	Bath	75	225
Gloucester	68	425	Bristol	52	300
Hereford	64	344	Chichester	62	418
Norwich	83	407	Exeter	68	409
Peterborough	80	451	Gloucester	86	425
St. Albans	70	551	Lichfield	57	370
Winchester	76	530	Lincoln	82	506
Ely	74	547	Norwich	72	407
			Ripon	88	295
			Southwark	55	273
			Wells	67	415
			St. Asaph	45	182
			Winchester	103	530
			Old St. Paul	103	611
			Salisbury	84	473

来源：Stephen Jay Gould。

- 7.10 土地估价 在明尼苏达州，明尼亚波利斯——圣·保罗的大都市地区的评估者受法律约束，对在“绿色英亩”活动中进行过登记的农田，仅根据农田产量对农田进行估价，并且对在附近建有商业中心或工业化停车场的事实不能用来估价。这造成了困难，因为几乎所有的作为设置估价值基础的销售是根据土地的潜在发展，而不是根据农田的价值进行标价的。为了帮助设置估价值，需要有一种对质量相同的土地，给出“相等”估价的方法。

表 7.10 明尼苏达州南部四个郡的土地产量得分 (P) 及
1981 和 1982 平均农田估价 (美元/英亩)

Le Sueur			Sibley			McLeod			Meeker		
价			价			价			价		
P	1981	1982	P	1981	1982	P	1981	1982	P	1981	1982
51	1495	1719	75	1652	1982	71	1524	1752	31	1047	1548
54	1222	1405	78	1544	1865	73	1448	1665	32	814	895
55	1200	1380	81	1536	1843	76	1483	1705	35	1143	1257
57	1254	1442	83	1541	1849	76	1489	1712	35	1263	1389
63	1358	1562	84	1570	1884	77	1489	1712	35	1318	1450
67	1416	1628	85	1554	1865	78	1465	1685	37	995	1095
69	1428	1642	85	1601	1921	79	1427	1641	38	1286	1415
71	1146	1318	87	1587	1904	79	1493	1717	40	1000	1100
75	1347	1549	89	1435	1722	79	1497	1721	44	1036	1140
75	1474	1695	89	1676	2011	79	1455	1673	45	1251	1376
78	1382	1589	90	1599	1919	79	1496	1720	55	1308	1439
78	1392	1601	90	1647	1976	82	1449	1666	56	1059	1165
80	1441	1659	92	1643	1972	83	1481	1703	60	1413	1554
			93	1656	1987	84	1419	1632	68	1309	1440
			94	1592	1919				73	1404	1544
			94	1619	1943				75	1282	1410
									79	1450	1595

来源: Douglas Tiffany。

表 7.10 的数据表示一种可能的均等方法, 它基于一个计算所得的土地产量得分。这是 1 到 100 之间的一个数, 数字越大, 土地越好。分析的单位是一个镇, 大小约为 6 英里×6 英里。对每个有耕地的镇, 记录平均土地产量得分 P 以及 1981 和 1982 年每英亩的平均估价值。数据来自位于明尼亚波利斯的南部和西部的四个郡 (Le Sueur, Meeker, McLeod, Sibley), 其发展压力对土地估价值没有什么影响。

使用这些数据分析土地产量得分是否给出决定估价值的基础。检验郡的差别与年份的差别。总结你的结果。

- 7.11 性别歧视 表 7.11 中的数据为在一个小的 Midwestern 学院里, 所有职员的薪水及其它特征。收集这些数据是为了按法律诉讼程序陈述的一个待裁决的问题: 妇女在薪水上受歧视的问题。数据中所有人获得任职资格, 或有任职记录。临时职员不在统计范围之内。数据来自于

个人资料，由以下量组成：

SX = 性别，代码 1 为女性，0 为男性

RK = 职称，代码 1 为助理教授，2 为副教授，3 为正教授

YR = 居目前职位的年数

DG = 最高学历，代码 1 为博士，0 为硕士

YD = 取得最高学历后的年数

SL = 学年薪水（美元）

- 7.11.1 给出下列假设的检验：“对三种职位中每一个，因居目前职位的年数，最高学历及获得最高学历后的年数的不同，对薪水的调整是相同的”，及备择假设“是不同的”。
- 7.11.2 使用所有的变量，通过适当的诊断证明，需要对响应变量，薪水，实施变换，并给出一个合适的变换。
- 7.11.3 变换响应变量后，检验非常数方差：(a) 作为薪水的一个函数；(b) 作为性别标志的一个函数。
- 7.11.4 检验，对变换后的薪水，在每种职位中，性别的差别是否一样。
- 7.11.5 使用所有自变量，分析这些数据，讨论关于男性与女性职员的薪水区别问题，并将你的结论总结成可以用于法庭的形式。
- 7.11.6 Finkelstein(1980)，在歧视事例中回归使用的讨论中写道，“...[a] 变量反映雇主授予的职位或地位。如果在授予职位或地位时有歧视，则变量可能“受污染”。这样，例如，如果在职员提升更高职位中有歧视，在比较性别之前用职位来调整薪水可能不为法庭所接受。排除职位的影响后，准确拟合在 7.11.5 中得到的模型，总结并比较除去职位的影响后，关于不同性别报酬的推理的差异的推断所得到的结果。

表 7.11 薪水数据

行	SX	RK	YR	DG	YD	SL
1	0	3	25	1	35	36350
2	0	3	13	1	22	35350
3	0	3	10	1	23	28200
4	1	3	7	1	27	26775
5	0	3	19	0	30	33696
6	0	3	16	1	21	28516

(续表)

行	<i>SX</i>	<i>RK</i>	<i>YR</i>	<i>DG</i>	<i>YD</i>	<i>SL</i>
7	1	3	0	0	32	24900
8	0	3	16	1	18	31909
9	0	3	13	0	30	31850
10	0	3	13	0	31	32850
11	0	3	12	1	22	27050
12	0	2	15	1	19	24750
13	0	3	9	1	17	28200
14	0	2	9	0	27	23712
15	0	3	9	1	24	25748
16	0	3	7	1	15	29342
17	0	3	13	1	20	31114
18	0	2	11	0	14	24742
19	0	2	10	0	15	22906
20	0	3	6	0	21	24450
21	0	1	16	0	23	19175
22	0	2	8	0	31	20525
23	0	3	7	1	13	27959
24	1	3	8	1	24	38045
25	0	2	9	1	12	24832
26	0	3	5	1	18	25400
27	0	2	11	1	14	24800
28	1	3	5	1	16	25500
29	0	2	3	0	7	26182
30	0	2	3	0	17	23725
31	1	1	10	0	15	21600
32	0	2	11	0	31	23300
33	0	1	9	0	14	23713
34	1	2	4	0	33	20690

(续表)

行	<i>SX</i>	<i>RK</i>	<i>YR</i>	<i>DG</i>	<i>YD</i>	<i>SL</i>
35	1	2	6	0	29	22450
36	0	2	1	1	9	20850
37	1	1	8	1	14	18304
38	0	1	4	1	4	17095
39	0	1	4	1	5	16700
40	0	1	4	1	4	17600
41	0	1	3	1	4	18075
42	0	1	3	0	11	18000
43	0	2	0	1	7	20999
44	1	1	3	1	3	17250
45	0	1	2	1	3	16500
46	0	1	2	1	1	16094
47	1	1	2	1	6	16150
48	1	1	2	1	2	15350
49	0	1	1	1	1	16244
50	1	1	1	1	1	16686
51	1	1	1	1	1	15000
52	1	1	0	1	2	20300

8

建立模型 II：共线性与变量选择

当自变量彼此相关时，回归模型可能非常令人糊涂。估计的效应会由于模型中的其它自变量而改变数值，甚至符号。故在分析时，了解自变量间关系的影响是很重要的。这一复杂的问题，通常称为共线性或多重共线性问题，是本章讨论的第一个问题。然后我们转向更为一般的问题：对一个模型在众多可能的自变量中进行选择。

8.1 什么是共线性

如果存在某些常数 c_0 ， c_1 和 c_2 ，使得线性等式

$$c_1X_1 + c_2X_2 = c_0 \quad (8.1)$$

对数据中所有案例都成立，则两个自变量 X_1 和 X_2 为精确共线性的。例如， X_1 和 X_2 为两种化学药品的量，且有 $X_1 + X_2 = 50\text{ml}$ 。在试验中，对任意给定的 X_1 值， $X_2 = 50 - X_1$ ，即知道 X_1 就是精确地知道 X_1 和 X_2 两者。精确共线性常常是偶然发生的。例如，以磅为单位的重量和以千克为单位的重量同时包含在模型中，或有一组虚拟变量。

如果等式 (8.1) 近似地对测量数据成立，则有近似共线性。

一个常用但不是完全合适的 X_1 与 X_2 间共线性程度的度量, 是它们样本相关系数的平方, r_{12}^2 。精确共线性对应于 $r_{12}^2=1$; 非共线性对应于 $r_{12}^2=0$ 。当 r_{12}^2 越接近于 1, 近似共线性越强。通常, 我们去掉形容词“近似”, 当 r_{12}^2 较大时, 我们说 X_1 和 X_2 是共线性的。

定义自然地扩展到 $p>2$ 个自变量。称一组自变量 X_1, X_2, \dots, X_p 是共线性的, 如果存在常数 c_0, c_1, \dots, c_p , 使得

$$c_1X_1+c_2X_2+\dots+c_pX_p=c_0 \quad (8.2)$$

近似成立。这表示, 至少有一个 X_k , 可以由其它的决定

$$X_k \cong (c_0 - \sum_{j \neq k} c_j X_j) / c_k \quad (8.3)$$

与两个变量的相关系数的平方相类似, 多于两个变量情况的一个简单的诊断量为, X_k 与其它 X 之间的复相关系数的平方。我们称它为 R_k^2 。这个数是通过 X_k 关于其它 X 的回归计算得到的。如果最大的 R_k^2 接近 1, 我们将试探性地诊断为近似共线性。

附加评注 当一组自变量精确共线性时, 则必须删除一个或多个自变量, 否则不存在系数的唯一的最小二乘估计。在有了其它变量后, 删除的自变量不包含任何信息。故这一做法并未失去什么。当共线性是近似的, 一个通常的做法是, 从模型中删去些自变量, 使丢失的信息为最少。

原先, 当散点图为椭圆, 大致对应于 X 的近似正态性时, 相关系数是共线性的一个有用的度量。但在许多问题中, X 的正态性假设是不可接受的。同时, 相关系数对极端值或非寻常案例是极为敏感的。因此, 它可能是共线性的不合适的诊断。

不幸的是, 要给出改进的度量方法是不容易的。问题多半是因为共线性的不精确定义。我们只要求好的近似的线性关系成立, 但没有说明多“好”或它将怎么被度量。由 (8.2) 出发, 不难构造出一个例子, 使等式几乎是满足的, 但如果这些 X 中的一个乘以一个常数, 相关系数不变, 但找不到 C 能满足同样程度的近似。我们将在 8.3 节再讨论这一问题。

8.2 为什么共线性是一个问题

共线性的自变量将典型地增大系数估计的方差。例如，考虑有两个自变量的回归

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \quad (8.4)$$

并假设 X_1 和 X_2 的样本相关系数为 r_{12} 。定义符号 $SX_j X_j = \sum (X_j - \bar{X}_j)^2$ 。作为练习（问题 8.9），证明

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - r_{12}^2} \right) \left(\frac{1}{SX_j X_j} \right) \quad (j=1, 2) \quad (8.5)$$

在 $r_{12}^2 = 0$ 时， $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差取到最小。当 r_{12}^2 接近 1 时，这些方差被大大地扩大了。例如，若 $r_{12}^2 = 0.95$ ， $SX_1 X_1$ 保持不变， $\hat{\beta}_1$ 的方差为 $r_{12}^2 = 0$ 时的 20 倍。这样，与使用非共线性的自变量的问题相比较，使用共线性自变量可能使得估计出来的变量的系数令人不能接受。

$p > 2$ 时的情况与 $p = 2$ 时类似。我们可以证明（问题 8.10），第 j 个系数的方差为

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \left(\frac{1}{SX_j X_j} \right) \quad (j=1, \dots, p) \quad (8.6)$$

量 $1/(1 - R_j^2)$ 称为第 j 个方差扩大因子，简记为 VIF_j (Marguardt, 1970)。假定这些 X_j 可以取样，使得 $R_j^2 = 0$ 而保持 $SX_j X_j$ 不变，那么 VIF 表示由于自变量间的相关系数，从而由共线性引起的方差的增大。

共线性还会影响预测值的方差，但其效果不太明显。对某些预测值，共线性自变量能精确地比具有相同 $SX_j X_j$ 值的正交自变量有更小的方差。当然，这并不是对所有可能的预测值都是这样的。

例 8.1 尖桩篱笆

Hocking 和 Pendelton (1983) 由图 8.1 所示的“尖桩篱笆”给出了共线

性的非常有效的刻划。这个图表示，与两个共线性自变量 X_1 和 X_2 对应的点的一个可能的结构。对给定的值 X_1 和 X_2 ，每根尖桩的长度给出响应变量 Y 的值。拟合一个回归模型 (8.4)，就象在尖桩上试着平衡一个拟合平面。在垂直于尖桩行的方向上，平面将是不稳定的。如果尖桩恰好是一条直线上，在垂直于尖桩行的方向上，平面的倾斜是任意的。这一“篱笆”两旁的预测将有很大波动。另一方面，沿着“篱笆”的点的预测可能会相当精确。对这些点的预测，共线性不是一个问题。系数的估计大致与沿着坐标轴的预测相类似。如图所示，篱笆与 X_1 和 X_2 轴都约成 45° 角，故我们将对 X_1 和 X_2 的系数估计考虑为不沿篱笆行的预测。它将导致两者的不稳定的估计。

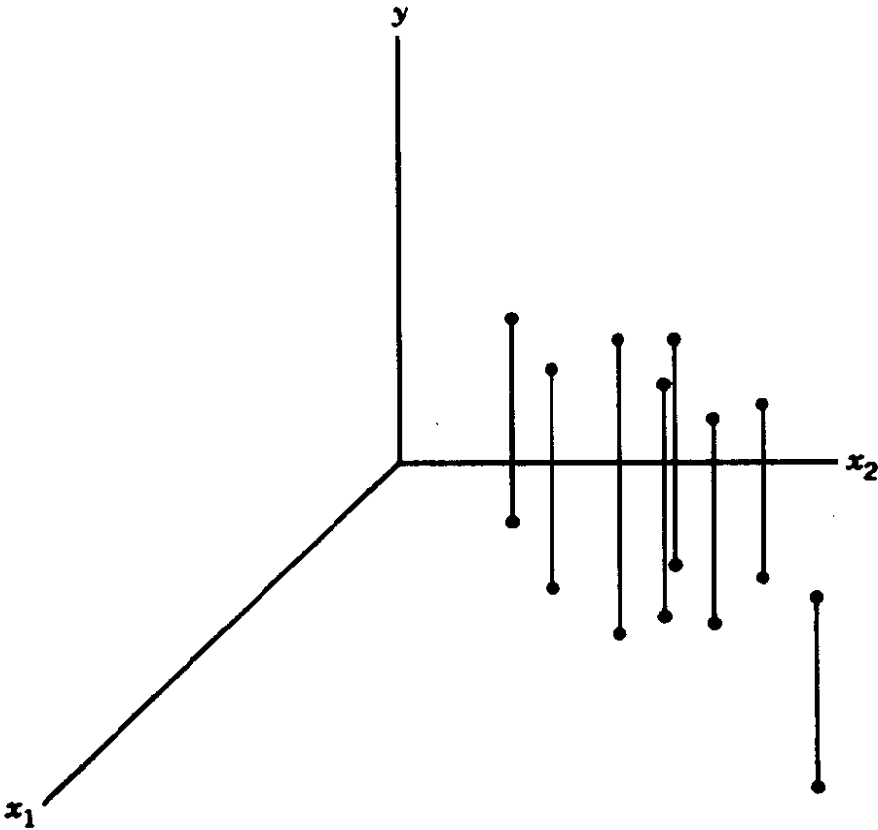


图 8.1 尖桩篱笆 (Hocking 和 Pendleton, 1983)

8.3 共线性的度量

度量一组自变量的共线性的程度，只有当 X 是来自一个多元正态分布的样本的时候是容易的。这时，相关系数和 VIF 是敏感的诊断统计量。另一种方法是直接度量 (8.2) 是否成立。用矩阵

术语, (8.2) 要求有一个单位向量 c , 使 Xc 几乎等于 0。或者, 用向量长度来表达, 对某个 c , $(Xc)^T(Xc) = c^T(X^TX)c$ 充分接近于零。我们可以证明, 对 c 的任何选择, $c^T(X^TX)c$ 大于等于 (X^TX) 的最小特征值, 等号在 c 取为对应于最小特征值的特征向量时成立。相对其它特征值而言, 如果最小特征值充分小的话, 可以判断出共线性。因为最小特征值的大小依赖于 X 的列的尺度, 故判断最小特征值的大小是困难的。如 7.6 节指出的, (X^TX) , $(\mathcal{X}^T\mathcal{X})$, 或样本相关矩阵的特征值可能是完全不同的。通常, $(\mathcal{X}^T\mathcal{X})$ 的特征值和使用优于 (X^TX) , 但 Belsey (1984) 有不同意见。如果 X 的列的度量单位是任意的, 则通常计算相关矩阵的特征值。如果使用 $(\mathcal{X}^T\mathcal{X})$, X 的具有大的样本方差的列将比有小的样本方差的列在确定特征值时更为重要, 并且这些差异在共线性中得到与在特征值中一样强的反映。基于相关矩阵的计算, 可以消除方差的差异的影响。

一个基于特征值的常用量称为条件数 k , 定义为

$$k = (\text{最大特征值} / \text{最小特征值})^{1/2} \quad (8.7)$$

k 大于等于 1。大的 k 值示意共线性。 k 是在分析计算机算法的数值特征时自然地产生的一个量, 但其作为共线性的一个统计量的含义是不太清楚的。已经提出一些规则, 例如 $k \geq 30$ 时认为有共线性, 但在理论上并没有什么证明。

k 的计算一般需要得到一个矩阵的特征值的最大值和最小值, 这个运算对线性回归问题是相当复杂的。Berk (1977) 证明了, k 必定至少和 VIF_j 的最大值的平方根一样大, 这样可用后一个统计量代替 k 而不丢失太多的信息。

许多共线性诊断可以被放入一个通用框架。仿 Cook (1984), 假设我们对估计参数向量 β 的线性组合感兴趣。考虑两个回归问题:

一个是观测的, 模型为

$$Y_1 = X\beta + e_1$$

另一个是理想化的，或标准的问题

$$Y_2 = Z\beta + e_2$$

在每个问题中，误差有相同的分布，并且 β 相同。这两个模型的区别是，标准的是从 Z 的行中，而不是从 X 的行中取数据。设计 Z 表示可能被观测的情况的一个理想化形式。通常，它对应于一个正交设计，因为这样的设计不表现出共线性。令 d 为一个 $p' \times 1$ 向量， $d^T \beta$ 为 β 的元素的一个线性组合。 $d^T \hat{\beta}$ 的方差的增大是由于使用了现实的设计 X ，而非标准设计 Z 。 $d^T \hat{\beta}$ 的方差的增量可以由如下的比率进行测量，

$$A(d|Z) = \frac{\text{var}(d^T \hat{\beta} | \text{现实的})}{\text{var}(d^T \hat{\beta} | \text{标准的})} = \frac{d^T (X^T X)^{-1} d}{d^T (Z^T Z)^{-1} d} \quad (8.8)$$

(8.8) 式大的值表示，至少对选取的 d ，现实设计比理想设计给出大得多的方差。 Z 和 d 的选取必须在共线性诊断可以被有效地解释和使用之前作出。

方差扩大因子和条件数都可以从 (8.8) 导出。选择不同的 d ，但选择相同的标准设计 Z ，使 $Z^T Z$ 和 $X^T X$ 有相同的第一行和列（第一行和列对应于截距的元素全为 1 的列），及相同的主对角线元素。两种设计的区别在于， $X^T X$ 的剩下的非对角线元素可能非 0，而 Z 的选择使得 $Z^T Z$ 的这些元素全为 0。从某种意义上， Z 是最接近于 X 的正交设计。

为得到第 j 个方差扩大因子，则选择 d ，在第 j 个位置为 1，而其它地方为 0。如果我们考虑所有可能的、形如 $(0, d_1^T)^T$ 的 d ，则会产生条件数。这类 d 给出不包括截距的 β 的其它元素的所有可能的线性组合的集合。我们可以证明，当 d 在这一类中变化时，(8.8) 的最大可能值与其最小值的比率恰为基于由 X 导出的相关矩阵的条件数的平方。

例 8.2 一个农业试验

这一例子由 Dennis Cook 提出。一个农业家希望做一个小试验来估计向

土壤加入 X_1 和 X_2 两种化学物质后, 某种作物的产量。一位研究助理被派去做这个试验。这是 2×2 因子设计, $X_1=0$ 或 1kg , $X_2=0$ 或 1kg 。计划的设计为

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

这是一个正交设计, 一点不表现出共线性。三个月之后, 这位研究助理做了这个试验, 并带着增加了的一点信息返回。由于两种化学物质均为 50kg 袋装的, 且由于有一小块地可利用, 故取 $X_1=X_2=48\text{kg}$, 做了第 5 次试验。这样, 现实设计为

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 48 & 48 \end{pmatrix}$$

这个现实设计现在明显是高度共线性的, $r_{12}^2=0.9989$, 且 $VIF_1=VIF_2=1/(1-r_{12}^2) \cong 903$ 。另外, 基于 $X^T X$ 的条件数 $k=67.9$, 基于 $\mathcal{X}^T \mathcal{X}$ 或样本相关矩阵的 k 为 60.1 。共线性的诊断量的值是很大的。在这个问题中, 这些诊断所用的标准设计可以近似表为

$$Z = \begin{pmatrix} 1 & 10 & 10 \\ 1 & -11.25 & -11.25 \\ 1 & -11.25 & 31.25 \\ 1 & 31.25 & -11.25 \\ 1 & 31.25 & 31.25 \end{pmatrix}^*$$

这是一个正交设计。同现实设计, 其中心为点 $(10\text{kg}, 10\text{kg})$ 。列校正平方和**与现实设计相等。标准的共线性诊断都一致得出下述结论: 相对于这一设计,

* 译者注: 原书上

$$Z = \begin{pmatrix} 1 & 10 & 10 \\ 1 & -11.25 & -11.25 \\ 1 & -11.25 & 11.25 \\ 1 & 11.25 & -11.25 \\ 1 & 11.25 & 11.25 \end{pmatrix}$$

** 原书上为列平方和。

观测得到的设计是很差的，共线性是一个问题。

有理由说明，标准设计是不符合题意的。首先，我们不能取观测值 X_1 或 X_2 为负，因为它们都是加入的量。其次，我们没有取 10kg 附近的数据，故以此为中心的标准设计不能满足原来的问题。观测得到的五点试验的真正困难不是共线性，而是我们可能相信的一个事实：尽管一个可加的一次模型对原始设计可能效果很好，但对 X_1 和 X_2 这样大的值它可能不会工作得很好。事实上，如果模型对所有五个点都适用，则给出前四个，再加上一个 X_1 和 X_2 尽可能大的单个点，研究助理将得到最大量的信息。对于实际进行的设计， $\text{var}(\hat{\beta}) \cong \sigma^2/2^*$ ，但如果第 5 个点取在 $X_1 = X_2 = 0.5$ ，则所得的设计将是正交的，且 $\text{var}(\hat{\beta}) = \sigma^2$ ，有二倍多大**。

对这一问题设计一个更为敏感的标准设计，并相对这一设计测量共线性的影响是一个有趣的练习，留给读者作为练习。

8.4 变量选择

从一个模型中删除自变量可以改进一个模型，并减少明显的共线性。考虑下述两个模型之间的选择，即在有 $p=2$ 个自变量的模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \quad (8.9)$$

与从 (8.9) 中删去变量 X_2 得到的有 $p=1$ 个自变量的子集模型

$$Y = \beta_0 + \beta_1 X_1 + e \quad (8.10)$$

之间的选择。假设我们对 β_1 的估计感兴趣。我们假设模型 (8.9) 是 Y 对自变量的依赖性的正确描述，以及误差相互独立并有常数方差。为简单起见，我们又假设 $SX_1X_1 = SX_2X_2 = 1$ 。对模型 (8.9)， β_1 的通常的最小二乘估计是无偏的，并且由 (8.5) 可知

$$\text{var}[\hat{\beta}_1 | \text{完全模型}(8.9)] = \frac{\sigma^2}{1 - r_{12}^2} \quad (8.11)$$

对模型 (8.10) 需要特别注意，因为现在 β_1 有不同的含义。它是

* 译者注：原书上， $\text{var}(\hat{\beta}) \cong \sigma^2/3.25$ 。

** 译者注：原书上，有三倍多大。

忽略 X_2 后, X_1 对 Y 的影响。而在 (8.9) 中的 β_1 为由 X_2 调整后 X_1 对 Y 的影响。因此, 如果两者不同的话, 则从 (8.10) 得到的 β_1 的估计将是有偏的。我们可以证明

$$E(\hat{\beta}_1 | \text{子集模型}) = \beta_1 + r_{12}\beta_2 \quad (8.12)$$

由子集模型得到的 β_1 的估计的偏差为 $\beta_1 - (\beta_1 + r_{12}\beta_2) = -r_{12}\beta_2$, 并且可以证明, 从子集模型得到的 $\hat{\beta}_1$ 的方差为

$$\text{var}(\hat{\beta}_1 | \text{子集模型}) = \sigma^2 \quad (8.13)$$

它不依赖于 r_{12} 。为比较从完全模型和子集模型得到的 β_1 的估计, 我们必须计算, 由子集模型得到的 $\hat{\beta}_1$ 的均方误差, $\text{mse}(\hat{\beta}_1 | \text{子集模型})$, 其中 mse 定义为

$$\text{mse}(\hat{\beta}_1 | \text{子集模型}) = \text{var}(\hat{\beta}_1 | \text{子集模型}) + (\text{偏差})^2$$

从而有

$$\text{mse}(\hat{\beta}_1 | \text{子集模型}) = \sigma^2 + (r_{12}\beta_2)^2 \quad (8.14)$$

比较 (8.11) 和 (8.14), 我们看到只要

$$\frac{|\beta_2|}{\sigma} \leq \frac{1}{\sqrt{1-r_{12}^2}} \quad (8.15)$$

子集模型将更精确地估计 β_1 , 即 $\text{mse}(\hat{\beta}_1 | \text{子集模型}) < \text{var}(\hat{\beta}_1 | \text{完全模型})$ 。结论 (8.15) 如图 8.2 所示。如果 r_{12}^2 接近 1, 子集模型几乎总是比完全模型好, 而对 r_{12}^2 的任何值, 如果 $|\beta_2| < \sigma$, 子集模型更好。这样, 当数据中观测到共线性, 一般从子集模型比从完全模型可以得到对 β_1 的更精确的估计, 除非被删除变量的系数非常大。如果 β 与 σ^2 都已知, 删除 $|\beta_j|/\sigma$ 较小的变量较为理想。由于这些量一般未知, 选择技术必须具有使删除变量的 $|\hat{\beta}_j|/\sigma$ 可能较小的性质。

共线性不是变量选择的唯一原因。在许多问题中, 我们可能会求一个相对较小的自变量集合, 它与完全集合包含几乎相同的信息。进一步的分析则可以针对这一自变量子集, 并很可能得到简化的结果。这样, 选择是许多回归分析但非全部的回归分析的

一个部分。

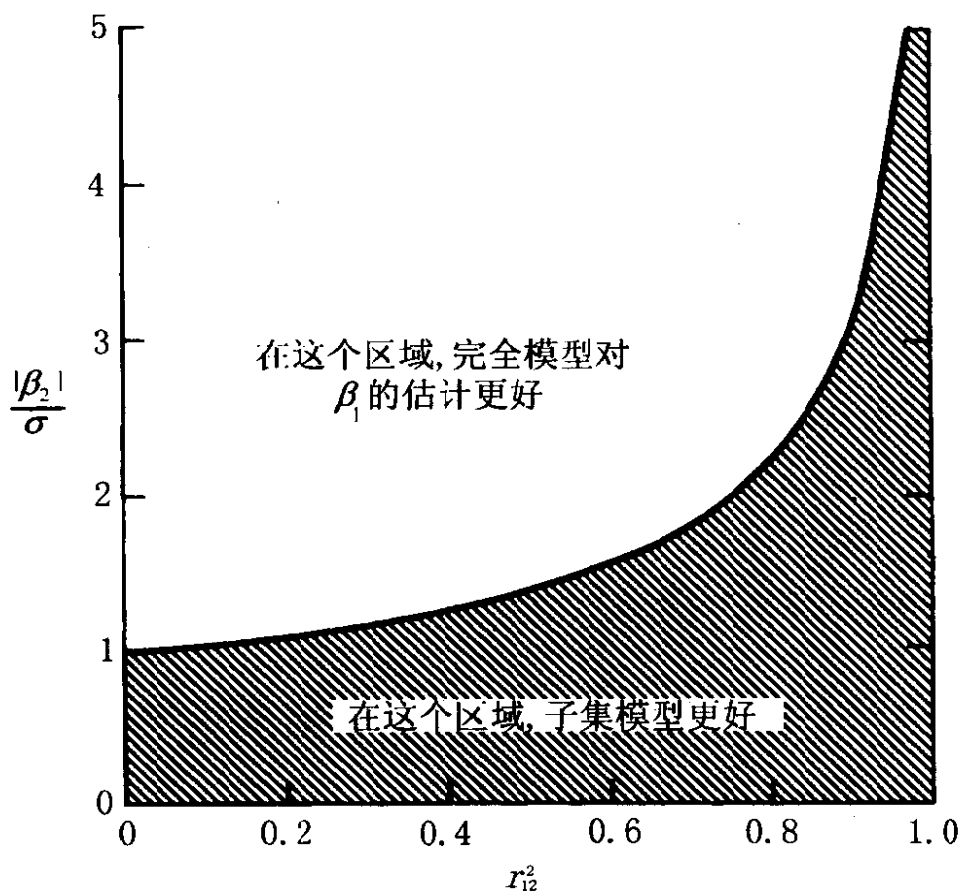


图 8.2 比较一个单自变量模型与一个双自变量模型

例 8.3 高速公路数据

这个数据集合，由表 8.1 给出，取自 Carl Hoffstedt 的国内工程学方面的一篇未发表的硕士论文。它将汽车意外事故率（事故数/百万英里行车）(Y) 与 13 个可能的相关变量相联系。数据包括 1973 年在明尼苏达州的 39 段高速公路。变量按表 8.1 的次序为

- $Y = RATE$ = 1973 年每百万英里行车的故事率
- $X1 = LEN$ = 段的长度数 (英里)
- $X2 = ADT$ = 以千计的平均日流量 (估计)
- $X3 = TRKS$ = 卡车容量在全部容量中的百分比
- $X4 = SLIM$ = 时速限制 (在 1973 年, 小于 55 英里/小时的限制)
- $X5 = LWID$ = 道路宽度 (英尺)
- $X6 = SHLD$ = 道路的外侧路肩宽度 (英尺)
- $X7 = ITG$ = 路段中每英里的快车道类型交换数

表 8.1 高速公路数据

	Y	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃
1	4.58	4.99	69	8	55	12	10	1.20	0	4.60	8	1	0	0
2	2.86	16.11	73	8	60	12	10	1.43	0	4.40	4	1	0	0
3	3.02	9.75	49	10	60	12	10	1.54	0	4.70	4	1	0	0
4	2.29	10.65	61	13	65	12	10	0.94	0	3.80	6	1	0	0
5	1.61	20.01	28	12	70	12	10	0.65	0	2.20	4	1	0	0
6	6.87	5.97	30	6	55	12	10	0.34	1.84	24.80	4	0	1	0
7	3.85	8.57	46	8	55	12	8	0.47	0.70	11.00	4	0	1	0
8	6.12	5.24	25	9	55	12	10	0.38	0.38	18.50	4	0	1	0
9	3.29	15.79	43	12	50	12	4	0.95	1.39	7.50	4	0	1	0
10	5.88	8.26	23	7	50	12	5	0.12	1.21	8.20	4	0	1	0
11	4.20	7.03	23	6	60	12	10	0.29	1.85	5.40	4	0	1	0
12	4.61	13.28	20	9	50	12	2	0.15	1.21	11.20	4	0	1	0
13	4.80	5.40	18	14	50	12	8	0	0.56	15.20	2	0	1	0
14	3.85	2.96	21	8	60	12	10	0.34	0	5.40	4	0	1	0
15	2.69	11.75	27	7	55	12	10	0.26	0.60	7.90	4	0	1	0
16	1.99	8.86	22	9	60	12	10	0.68	0	3.20	4	0	1	0
17	2.01	9.78	19	9	60	12	10	0.20	0.10	11.00	4	0	1	0
18	4.22	5.49	9	11	50	12	6	0.18	0.18	8.90	2	0	1	0
19	2.76	8.63	12	8	55	13	6	0.14	0	12.40	2	0	1	0
20	2.55	20.31	12	7	60	12	10	0.05	0.99	7.80	4	0	1	0
21	1.89	40.09	15	13	55	12	8	0.05	0.12	9.60	4	0	1	0
22	2.34	11.81	8	8	60	12	10	0	0	4.30	2	0	1	0
23	2.83	11.39	5	9	50	12	8	0	0.09	11.10	2	0	1	0
24	1.81	22.00	5	15	60	12	7	0	0	6.80	2	0	1	0
25	9.23	3.58	23	6	40	12	2	0.56	2.51	53.00	4	0	0	1
26	8.60	3.23	13	6	45	12	2	0.31	0.93	17.30	2	0	0	1
27	8.21	7.73	7	8	55	12	8	0.13	0.52	27.30	2	0	0	1
28	2.93	14.41	10	10	55	12	6	0	0.07	18.00	2	0	0	1
29	7.48	11.54	12	7	45	12	3	0.09	0.09	30.20	2	0	0	1
30	2.57	11.10	9	8	60	12	7	0	0	10.30	2	0	0	1
31	5.77	22.09	4	8	45	11	3	0	0.14	18.20	2	0	0	1
32	2.90	9.39	5	10	55	13	1	0	0	12.30	2	0	0	1
33	2.97	19.49	4	13	55	12	4	0	0	7.10	2	0	0	1
34	1.84	21.01	5	12	55	10	8	0	0.10	14.00	2	0	0	1
35	3.78	27.16	2	10	55	12	3	0.04	0.04	11.30	2	0	0	1
36	2.76	14.03	3	8	50	12	4	0.07	0	16.30	2	0	0	1
37	4.27	20.63	1	11	55	11	4	0	0	9.60	2	0	0	1
38	3.05	20.06	3	11	60	12	8	0	0	9.00	2	0	0	0
39	4.12	12.91	1	10	55	12	3	0	0	10.40	2	0	0	0

表 8.2 主要统计量

变量	N	均值	(a) 均值与方差				最小值	最大值
			方差	标准差	最小值	最大值		
RATE	39	3.933	3.944	1.986	1.610	9.230		
LEN	39	12.88	57.91	7.610	2.960	40.09		
ADT	39	19.62	346.4	18.61	1.000	73.00		
TRKS	39	9.333	5.544	2.355	6.000	15.00		
SLIM	39	55.00	34.21	5.849	40.00	70.00		
LWID	39	11.95	0.2078	0.4559	10.00	13.00		
SHLD	39	6.872	9.220	3.036	1.000	10.00		
ITG	39	0.2964	0.1691	0.4112	0	1.540		
SIGS	39	0.4005	0.4012	0.6334	0	2.510		
ACPT	39	12.16	86.83	9.318	2.200	53.00		
LANE	39	3.128	1.852	1.361	2.000	8.000		
FAI	39	0.1282	0.1147	0.3387	0	1.000		
PA	39	0.4872	0.2564	0.5064	0	1.000		
MA	39	0.3333	0.2281	0.4776	0	1.000		

(b) 样本相关矩阵

RATE	1.00
LEN	-0.47 1.00
ADT	-0.03 -0.27 1.00
TRKS	-0.51 0.50 -0.10 1.00
SLIM	-0.68 0.19 0.24 0.30 1.00
LWID	-0.01 -0.31 0.13 -0.15 -0.10 1.00
SHLD	-0.39 -0.10 0.46 0.00 0.69 -0.04 1.00
ITG	-0.02 -0.25 0.90 -0.07 0.24 0.10 0.38 1.00
SIGS	0.56 -0.32 0.15 -0.45 -0.41 0.04 -0.13 -0.07 1.00
ACPT	0.75 -0.24 -0.22 -0.36 -0.68 -0.04 -0.43 -0.20 0.50 1.00
LANE	-0.03 -0.20 0.82 -0.15 -0.26 0.10 0.48 0.70 0.25 -0.21 1.00
FAI	-0.21 -0.03 0.76 0.14 0.46 0.04 0.40 0.81 -0.25 -0.34 0.59 1.00
PA	-0.16 -0.15 -0.03 -0.05 0.04 0.23 0.37 -0.13 0.30 -0.23 0.17 -0.37 1.00
MA	-0.34 0.13 -0.46 0.10 -0.42 -0.28 -0.62 0.36 -0.07 0.51 -0.51 -0.27 -0.69 1.00
RATE	LEN ADT TRKS SLIM LWID SHLD ITG SIGS ACPT LANE FAI PA MA

$X_8 = SIGS$ = 路段中每英里信号交换数

$X_9 = ACPT$ = 路段中每英里进入点数

$X_{10} = LANE$ = 在两个方向上的交通车道总数

$X_{11} = FAI$ = 1, 如果是联邦帮助州际高速公路; 0, 否则

$X_{12} = PA$ = 1, 如果是首要干道高速公路; 0, 否则

$X_{13} = MA$ = 1, 如果是主干道高速公路; 0, 否则

两个高速公路段, 38 和 39 号, 既不是州际高速公路, 不是首要干道, 也不是主干道, 而是被分类归于主要收集数据的高速公路段, 并被编码 $FAI = PA = MA = 0$ 。没有用单独的变量 MC 。否则, 所得的数据矩阵会是精确共线的, 必须删除一个虚拟变量才能得到估计。基本主要统计量由表 8.2 和 8.3 给出。

表 8.3 对完全模型的回归

变量	估计	标准误	t-值
截矩	13.7	6.87	1.99
LEN	-0.065	0.033	-1.94
ADT	-0.004	0.034	-0.12
$TRKS$	-0.100	0.115	-0.87
$SLIM$	-0.124	0.082	-1.52
$LWID$	-0.134	0.598	-0.22
$SHLD$	0.014	0.162	0.09
ITG	-0.475	1.28	-0.37
$SIGS$	0.713	0.525	1.36
$ACPT$	0.067	0.043	1.56
$LANE$	0.027	0.283	0.09
FAI	0.543	1.72	0.31
PA	-1.01	1.11	-0.91
MA	-0.548	0.976	-0.56
$\hat{\sigma}^2 = 1.44$, d.f. = 25, $R^2 = 0.76$, $RSS = 35.89367$			

表 8.3 的一个有趣特征, 在对完全模型的回归分析简要中, 尽管有 $R^2 = 0.76$, 但没有一个系数的 t -值的绝对值超过 2。这样, 尽管把所有的自变量作为一组, 对于预测事故率是有用的, 但由其它自变量的调整, 没有一个自变量是明显很重要的。通常这表示, 可从模型中删除一些自变量, 而不降低拟合度。

对这个数据集, 所有标准共线性诊断都有中等的值。基于相关矩阵的条

件数是 $k = (4.50/0.051)^{1/2} = 9.4$, 而 VIF 的值由 $TRKS$ 的 1.9 变化至 FAI 的 9.1。在最坏的情况, 对这些数据来说, 共线性可以认为不是一个严重的问题。

8.5 假设和记号

对 n 个案例中的每一个, 我们观测了 k 个自变量 X_1, \dots, X_k 和一个响应变量 Y 。在早先的章节中, 自变量的个数为 p 。在这一章中, 我们仍用 p 表示在一个可能包含截矩的选出的子集中, 自变量的个数。如果模型包括一个常数项, 则令 $k' = k + 1$, 如果回归通过原点, $k' = k$ 。

用矩阵形式, 使用所有自变量的完全模型为

$$Y = X\beta + e \quad \begin{cases} X: n \times k' \\ \beta: k' \times 1 \\ e: n \times 1 \quad (\text{var}(e) = \sigma^2 I) \end{cases} \quad (8.16)$$

如果 $n < k'$, 自变量个数多于案例数, 我们不能得到 β 的一个唯一估计。不过, 模型 (8.16) 仍能给出 Y 与 X 之间关系的一个描述。现在可以通过将 X 分成两个矩阵 X_1 和 X_2 来说明子集模型, 其中 X_1 为 $n \times p$, $p \leq n$ (且 $\text{rank}(X_1) = p$), X_2 为 $n \times (k' - p)$ 。 X_1 由子集模型中的变量组成, X_2 由不在子集模型中的变量组成。

与 X 分为 X_1 和 X_2 相对应, 我们将 β 分为 β_1 和 β_2 , 其中 β_1 为 $p \times 1$, β_2 为 $(k' - p) \times 1$ 。完全模型 (8.16) 可以重新写为

$$Y = X_1\beta_1 + X_2\beta_2 + e \quad (8.17)$$

然后, 通过删除项 $X_2\beta_2$ 给出

$$Y = X_1\beta_1 + e^* \quad (8.18)$$

得到一个特定的子集模型。只有当 $\beta_2 = 0$ 时, 两个模型 (8.17) 和 (8.18) 是相同的。总之, 因为在每个模型中参数 β_1 的估计通常是不同的, 而且对参数的解释依赖于使用什么模型, 所以这两个模型是相当不同的。另外, 尽管对 (8.17), 如果 $k' > n$, 则不存在

唯一的估计, 但若我们限制于 $p \leq n$ 的子集, 我们总能得到 (8.8) 中参数的唯一的估计。

所有的选择模型的出发点是认为, 完全模型包含所有正确变换后的有关的自变量, 以及可能是额外的不重要的自变量。选择的目的是删去无关的自变量, 或者在给出其它自变量后, 不是很有用的自变量。

在实际中, 变换和可能是自变量的叉积或其它组合是需要的。分析者必须在有关变换和选择变量, 这两个方面作出决定。作为一个一般途径, 变换的选择及其它问题的诊断可先于子集选择。然而, 在选择以后应审慎地重复进行诊断。

在子集模型中, 我们用 $\beta_1 = (X_1^T X_1)^{-1} X_1^T Y$ 估计 β_1 , 仿佛是用最小二乘法, 并且其它的 X 没有被观测。

8.6 根据实际意义选择子集

选择一个用于模型的变量的子集的简单而最重要的工具是分析者对所研究的实际领域的知识, 以及每个变量的包括其系数的符号与数量的知识。

在高速公路事故数据中, 有 $k=13$ 个潜在自变量, 故有 $2^{13}=8192$ 种可能的子集模型, 包括完全模型和只包含截距的模型。不过, 这 13 个自变量可以被分为几个类型。首先, 变量 11、12 和 13, 即 FAI 、 PA 和 MA , 只是指示变量。放在一起以表示高速公路的类型。取这些变量中的一个而省略另外两个可能是不合理的 (这部分是因为第四种类型的高速公路—主要收集数据的高速公路段是由 $FAI=PA=MA=0$ 表示)。这样我们可以考虑或者包含这三者, 或者全不包含。这些可能有特别的重要性。这是因为道路的类型是由提供财政支持的单位定义的, 高速公路部门使用此经费来维护道路。我们甚至可以将这一问题看作是协方差分析, 在这里我们探讨了由其它自变量调整的高速公路类型的差异的可能

性。

另外,变量 LEN 的处理与其它的不同,因为由高速公路段定义的方法需要将它包含在预测方程中。假设高速公路由“安全路段”和“损坏地点”组成,而大多数事故发生在损坏地点。如果我们在研究中要将高速公路段延长 1 英里,由于我们假设损坏地点是很少的,故不大可能将 1 个损坏地点加入这一延长段中。然而,作为在研究中延长道路的结果,计算得到的响应变量,在这个路段上每百万行车英里的事故数将有一个较小的值。行驶的英里数将会增加,而事故数几乎不变。这样,响应变量和 LEN 必定是负相关的,我们应该只考虑包含 LEN 的模型。

这样,高速公路段的可能子集模型数现在由 8192 减至 512 个包含 LEN 及类型指标的,以及 512 个包含 LEN 但不包含类型指标的。尽管数目仍然很大,但要容易处理些。

重新定义变量 另一种减少自变量个数的重要方法是定义新的变量,它是旧变量的组合。例如,在某些研究中,高度和重量常是两个高度相关的变量,可以用它们的组合来代替,给出单个高度——重量指标变量。又如,两个 IQ 测试的自变量,可以用它们的平均来代替。

8.7 求子集 I: 逐步的方法

在许多问题中,有一个共同点就是使用数据求自变量的子集。为达到这一目的,形成两种基本方法。第一种方法,一般称为逐步回归。它使用一个方便的计算算法,将可能的模型数限制为一个相当小的数。虽然它不对应于选择一个模型的任何特定的准则,但它在实践中被频繁地使用。第二种方法使用对所有自变量的可能子集进行计算的准则统计量。它随着几种计算所有可能回归的快速算法的发展,而逐渐被应用于实践中。我们首先讨论逐步的方法。

逐步的方法给出一种系统化技术，用于检验各种大小的仅少数几个子集。首先选择通向可能模型的一条途径，查看一个子集，然后只是从前面所得的模型中增加或删除变量。

逐步回归有三种基本算法，一般称为前向选择 (FS)，后向消去 (BE)，和逐步 (SW)。在 FS 过程的每一步，都增加自变量；在 BE 中，消去自变量；在 SW 的每一步，可以或者增加，或者消去，或者交换一个“内部”变量和一个“外部”变量。

FS 过程如下进行。由简单回归模型开始。选取一个自变量，它与响应变量的样本相关系数的绝对值最大。这是第一步。其次，我们在模型中增加满足以下三个等价准则的自变量：(1) 由已经在方程中的自变量调整后，它与响应变量的样本偏相关系数的绝对值最大；(2) 与其它任何变量相比，增加这个变量后， R^2 增加得最多；(3) 在模型尚没有包含的变量中，增加的变量具有最大的 t -或 F -统计量值。这样在 FS 中，我们从大小为 1 的一个子集开始，然后在每一步，我们根据准则把一个变量增加到模型中。我们持续一次加入一个变量，直到满足停止规则。可能的停止规则如下：

FS. 1 子集到达一个预先给定的大小 p^* 时停止。

FS. 2 在每一个尚未进入的变量的 F -检验小于某个预先给定的数时停止。这个数记为 $F-IN$ (或者，等价地，如果 t 统计量的绝对值小于 $(F-IN)^{1/2}$ 时停止)。

FS. 3 当加入下一个自变量，会使自变量集合过于接近共线性的时候停止。这称为一个容差检验。它通常与下一个要加入的自变量和已经包含在方程中的自变量的复相关系数的平方有关。Berk (1977) 给出实现这一检验的计算方法的细节 (亦见问题 2.7)。容差检验用于识别某些极值共线性和防止计算中的舍入误差。

对后向消去 (BF) 方法，除了我们是从完全模型开始，并且在每一步消去一个变量外，它和 FS 类似。在方程的所有变量中，要消去的变量具有最小的 t 或 F 值。这等价于消去引起 R^2 最小变

化的变量，或者是由模型中剩余的所有其它变量调整后，与 Y 的偏相关系数的绝对值最小的变量。BE 的停止规则亦与 FS 的相类似：

BE. 1 子集到达预先给定的大小 p^* 时停止。

BE. 2 如果模型中所有变量的 F -检验值都大于某个预先给定的数时停止，这个数记为 $F-OUT$ 。

逐步的算法如在 FS 中的那样开始。在第一步后的每一步，我们考虑四种选择：增加一个变量；消去一个变量，交换两个变量，停止。SW 的规则可以总结为：

SW. 1 如果在当前模型中至少有两个变量，并且一个或多个具有小于 $F-OUT$ 的 F 值，有最小 F 值的变量被从模型中消去。

SW. 2 如果模型中有两个或多个变量，则当消去某个变量后得到的 R^2 值比先前同样数目变量得到的 R^2 要大的时候，具有最小 F 值的变量将被消去。这个现象在 SW 过程中是可能发生的，因为变量在不同的步骤中被增加和消去。

SW. 3 如果模型中有两个或多个变量，它们中的一个将与不在模型中的一个变量交换，如果交换使 R^2 增大。

SW. 4 如果一个变量有如 FS 中的最高的 F -值，只要 F 大于 $F-IN$ ，并且满足容差准则，则该变量加入到模型中去。

按不同次序应用规则，或改变 $F-IN$ 或 $F-OUT$ 的值，得到的这些算法是有差异的。如果 $F-IN$ 非常小，如 0.01，则 FS 的最后一步将一般不让那些不满足容差检验的自变量增加到模型中去。某些分析者用 $F-IN$ 的一个小值，来对自变量排序。早进入的自变量被假设更为重要。对这一实际做法不存在理论证明。Butler (1984) 使用邦佛伦尼不等式给出一个方法。在使用逐步的方法时，它可用来估计在一个方程中加入一个变量的 p -值。它要求在每一

步改变 $F-IN$ 的值。通常, $F-IN$ 和 $F-OUT$ 的值是固定不变的。它们在 2 和 4 之间。

高速公路事故数据 (续) 我们现在对表 8.1 中给出的高速公路事故数据应用各种算法。为便于说明, 我们选择 $F-IN = F-OUT = 2.0$ 。不同的程序, 这些参数的缺省值往往是不同的。首先我们使用 FS 方法。

在 FS 中首先一件工作是求得与响应变量最为高度相关的一个变量。由表 8.2 得到 Y 与 $ACPT$ 的相关系数是最大的, 其值为 0.75, 故它成为模型中的第一个变量。回归由表 8.4 的第一列给出。受 $ACPT$ 调整, $RATE$ 与其它可能的自变量的偏相关系数, 从与 FAI 的最小的 (绝对值) 0.015 变化至与 LEN 的最大的 -0.45 。这样, LEN 第二个被选入模型中, 见表 8.4。用这一方法, 我们依次加入 $SLIN$, $SIGS$ 以及 PA 。下一个考虑的变量是 $TRKS$ 。然而, 我们看到 $TRKS$ 有 $t=0.95$, 小于 $(F-IN)^{1/2}$ 。这样, 在加入 $TRKS$ 之前, 我们在步骤 5 终止。如果我们使用 $F-IN$ 的一个不同的值, 我们可能得到一个不同的模型。例如, 若取 $(F-IN)^{1/2}$ 为 1.5, 而不是 1.44, 则 $SIGS$ 不会被选入。我们最终得到的是一个三个变量的模型。

表 8.4 FS 方 法

在各步骤的估计与 t -值						
变量	1	2	3	4	5	6
截距	1.98 (5.64)	3.19 (6.21)	9.325 (3.56)	8.81 (3.38)	9.94 (3.85)	10.56 (3.96)
$ACPT$	0.160 (6.94)	0.145 (6.71)	0.101 (3.72)	0.089 (3.17)	0.064 (2.12)	0.0628 (2.07)
LEN		-0.079	-0.077	-0.0685	-0.074	-0.0635

(续表)

在各步骤的估计与 t -值						
变量	1	2	3	4	5	6
		(-2.99)	(-3.10)	(-2.72)	(-3.02)	(-2.35)
<i>SLIM</i>			-0.103	-0.096	-0.105	-0.103
			(-2.39)	(-2.26)	(-2.54)	(-2.49)
<i>SIGS</i>				0.485	0.797	0.701
				(1.42)	(2.16)	(1.83)
<i>PA</i>					-0.774	-0.743
					(-1.89)	(-1.80)
<i>TRKS</i>						-0.089
						(-0.95)
d. f	37	36	35	34	33	32
$\hat{\sigma}^2$	1.76	1.45	1.28	1.24	1.16	1.16
R^2	0.56	0.65	0.70	0.72	0.74	0.75

表 8.5 对高速公路数据的 BE

步骤	删去	R^2	R^2 的减少	删去变量后的 F
1		0.7605		
1	<i>LANE</i>	0.7604	0.0001	0.01
2	<i>ADT</i>	0.7604	0.0000	0.01
3	<i>SHLD</i>	0.7603	0.0001	0.01
4	<i>FAI</i>	0.7592	0.0012	0.14
5	<i>LWID</i>	0.7579	0.0012	0.15
6	<i>ITG</i>	0.7562	0.0017	0.22
7	<i>MA</i>	0.7521	0.0041	0.52
8	<i>TRKS</i>	0.7450	0.0070	0.90
9	<i>PA</i>	0.7521	0.0071	3.57

对 BE 算法, 由完全模型开始计算回归。只要 $|t| < (F-OUT)^{1/2}$ (或 $F < F-OUT$), 具有最小的 $|t|$ 或 F 的变量被消去。然后计算余下的变量的回归, 并再次消去具有最小的 $|t|$ 或 F 的变量。重复这一过程, 直到满足停止标准。取 $F-OUT=2.0$, BE 的结果 (或任何逐步的算法) 常在一个如表 8.5 的表中给出。最后的模型与由 FS 得到的相同。

表 8.6 $n=100, k=50$ 的一个模拟样本的结果

方法	P	R^2	总的 F 的 P -值	自变量数, 其 p -值 \leq	
				.25	.05
不作选择	50	0.59	.13	16 (32%)	6 (12%)
$F-IN=2$	16	0.48	$<.001$	16 (100%)	11 (69%)
$F-IN=4$	4	0.46	$<.001$	4 (100%)	4 (100%)

逐步方法的讨论 逐步的方法解释容易, 计算费用低, 并且使用广泛。由逐步回归得到的结论的相对简单性吸引了许多分析者。但是使用逐步的方法要谨慎。用逐步的方法选择模型, 不需要为选择一个模型而优化任何合理的准则函数。自变量的一个显而易见的次序是该方法的人为现象, 并且不必反映实际兴趣的关系。最后, 逐步回归可能严重夸张结论的显著性。

考虑一个模拟例子。使用标准正态随机偏差, 产生了一个有 100 个案例的数据集。响应变量为 Y , 50 个自变量为 X_1, \dots, X_{50} 。故所有的 β_j 皆已知为 0, 并且 Y 与 X 间的真实复相关系数也恰好为零。数据中的所有数字都是独立地取值的。 Y 关于 X_1 至 X_{50} 的回归在表 8.6 的第一行给出。考虑到所有数据为相互独立的随机数, 值 $R^2=0.59$ 看起来大得惊人。整个的 F -检验是一种更易于标准化的尺度, 它对数据给出的 p -值为 0.13。Rencher 和 Pun (1980) 和 Freedman (1983) 报告了类似的模拟, 整个的 p -值从接近于 0 变化到接近于 1。由于 $\beta=0$ 的零假设为真, 这样的变化

范围是本该如此的。在这里报告的模拟中, 50 个自变量中有 16 个其 t 统计量的 p -值小于 0.25, 而 50 个中有 6 个相应的 p -值小于 0.05。表 8.6 的第二行给出用 $F-IN=2$ 的 FS 回归得到的最终模型。只有 16 个自变量被保留, R^2 略下降到 0.48。在结果的显著性方面, 我们能感觉到主要的变化。整个的 F 值现在有一个非常小的 p -值, 小于 0.001。并且在方程中, 16 个自变量中有 11 个其 t 统计量相应的 p -值小于 0.05。除了使用严格的 $F-IN=4$ 外, 第 3 行类似于第 2 行。只使用 4 个自变量, $R^2=0.46$, 除了所有这些, 整个的回归还有非常小的相应的 p -值。

这个例子给出许多教训。首先, 自变量的逐步选择可以对结论的表面上的显著性有重要影响。留在模型中的自变量的系数, 其绝对值一般太大, 并且有太大的 t -或 F -值。其次, 即使响应变量与自变量是不相关的, R^2 也可能较大。Freedman (1983) 证明了, 不作选择, R^2 的期望值为 $k/(n-1)$ 。作了选择, R^2 将是有偏的, 并可能很大。

8.8 选择一个子集的准则

基于准则的子集选择有两部分。首先, 我们必须为比较子集, 选择一个准则统计量。其次, 要有一个有效的计算程序, 用于找到满足准则的最好子集。在这一节, 我们考虑基于预测误差的准则。如果一个子集模型给出的预测在某种意义上更好, 则选择这一子集模型。一种可能性是, 对我们感兴趣的某个未来点或点集, 度量预测的总的或平均的均方误差的估计。

一个子集模型可能产生有偏的预测或拟合值。如果可以减小方差, 偏差还是可以容忍的。作为拟合值的一个综合度量, 我们考虑对每个拟合值的均方误差 $\text{mse}(\hat{y}_i)$ 。对一个给定的 p 个自变量的模型, 定义 J_p 为

$$J_p = \frac{1}{\sigma^2} \sum_{i=1}^n \text{mse}(\hat{y}_i) \quad (8.19)$$

好的子集，其 J_p 应取较小的值。 J_p 的使用强调了观测数据如果一个点被重复，则在那一点的 mse 在 J_p 中将得到较大的权值。如果这些案例是要求对未来值进行预测的那些点的一个随机样本，或者如果我们对数据中的 n 个案例感兴趣，这是合理的。在其它情况 $\text{mse}(\hat{y}_i)$ 的别的函数可能更好。

J_p 的值依赖于一些不可观测的参数，故它必须用数据进行估计。可以形成对 J_p 的若干估计。由 Mallows 提出的，最简单的一个称为 C_p ，其导出见附录 8A.1。 C_p 可以写成下面三种等价形式中的任一种。

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n \quad (8.20)$$

$$= \frac{RSS_p - RSS_{k'}}{\hat{\sigma}^2} + p - (k' - p) \quad (8.21)$$

$$= (k' - p)(F_p - 1) + p \quad (8.22)$$

其中 $\hat{\sigma}^2$ 来自于完全模型， F_p 是假设“所有在子集模型外，但在完全模型内的自变量的系数都为零”的通常的 F 统计量。 C_p 有许多有用的性质。

1. 由 (8.20)， C_p 仅依赖于通常的回归计算，即 RSS_p ， $\hat{\sigma}^2$ ， p 和 n ，并且是易于计算的。这是在所有可能回归的快速算法中使用 C_p 的基础。

2. 由 (8.21)， C_p 度量完全模型和子集模型在拟合误差上的差别。

3. 由 (8.22)， C_p 由两部分组成，一个随机部分 F_p 和一个固定惩罚 p 。减少 F_p 就必定增加变量，受到惩罚。 C_p 兼顾这两者的得失。这是选择一个模型的惩罚方法的一个例子。对线性回归问题， C_p 的使用类似于大部分的惩罚方法。

4. 直接由 (8.22) 得到，对完全模型。 $C_k = k'$ 。

5. 对一个子集模型，如果所有未包括的自变量的系数为零。

则平均而言, 因为 $E(F_p) \cong 1$, 故 $C_p \cong p$ 。当然, C_p 是一个随机变量, 由 (8.22), 其名义上的分布与 F_p 的分布密切相关。

6. 可以通过比较两个子集模型的 C_p 的值, 来比较这两个子集模型。Mallows (1973) 认为, 好的模型应有 $C_p \cong p$ 。由于 C_p 是一个随机变量, 两个模型的 C_p 值可能非常接近, 它们是不易区分的。当然, 任何 $C_p \leq k'$ 的模型, 相应地有 $F_p \leq 2$, 将可能是一个好的子集模型。有某些问题中, 将找到许多这样的模型。

C_p 统计量不是计算类似 J_p 的一个度量的估计的唯一方法。其它重要的可选方法是基于相互证实的思想。将数据分为若干组, 其中最直截了当的方法是将数据大致一分为二。使用一半数据拟合一个模型, 然后对另一半进行拟合, 使用拟合误差平方来估计 J_p 。关于这一方法的细节, 参见 Snee (1977)。

另一个相互证实型的方法与第 5 章中关于诊断量的工作密切相关, 是计算预测残差

$$\hat{e}_{(i)} = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \quad (8.23)$$

并将预测残差平方和 (PRESS)

$$PRESS = \sum \hat{e}_{(i)}^2 \quad (8.24)$$

作为一个准则函数 (Allen, 1974; Geisser 和 Eddy, 1979)。这一统计量的一个美学优点是, 它使用一个不包括案例 i 的估计, 对案例 i 的情况进行预测。它的一个计算上的难点是, 它不能由充分统计量计算得到, 而需要许多其它的工作。使用 PRESS 的一种可能是, 使用一些较易计算的准则, 如 C_p , 得到入选的回归模型, 然后只对这几个模型计算 PRESS。

附加评注 C_p 准则没有谈到具有不同变换的响应变量的两个子集模型的直接比较问题。对这一问题的处理方法的书目见 Pereira (1977)。另一个用于在子集模型中进行选择的准则函数称为校正的 R^2 或 \bar{R}^2 , 定义为

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p} \right) (1 - R^2) \quad (8.25)$$

与 R^2 不同, \bar{R}^2 中有自由度 $(n-1)/(n-p)$ 的校正, 故若向模型中加入一个自变量, \bar{R}^2 不必定增大。Kennard (1971) 指出, \bar{R}^2 与 C_p 紧密有关, 故我们在这一讨论中只有 C_p 。

8.9 子集选择 I: 所有可能回归

对只有几个, 如 $k=8$ 或更少的自变量的问题, 我们不难对所有可能的自变量子集计算一或两个准则的值。对此的算法由 Garside (1971), Schatzoff, Fienberg 和 Tsao (1968) 及 Morgan 和 Tartar (1972) 所描述。这样, 我们可以求得根据准则最有利的几个子集。这少数几个子集可用于进一步的研究, 包括案例分析、问题的解释等等。如果 k 较大, 需要的计算量增长很快, 因此需要有不计算所有可能回归而求得最优的几个子集的方法。对此的算法由 Furnival 和 Wilson (1974), Hocking 和 Leslie (1967), Beale et. al (1967) 以及 Lamotte 和 Hocking (1970) 给出。Furnival 和 Wilson 算法看起来是最常用的。它使用根据已经计算的回归得到的信息去指出尚未计算的回归的准则函数的可能值的界限。这一技巧允许略过大部分回归的运算。算法在 BMDP 系列程序 (BMDP9R, Dixon, 1983) 中实现, 并且也可在 IMSL 库 (IMSL, 1979) 的子程序 RLEAP 中得到。对大至 30 的 k , 求得有最小 C_p 的 5 个子集的费用, 与计算同样规模的逐步回归的费用大致相同。

高速公路数据 (续) 如 8.6 节所指出的, 需要考虑 1024 个模型, 其中 512 个含 LEN 但不包括三个关于高速公路类型的虚拟变量, 512 个含 LEN 及三个虚拟变量。我们所用的方法是对这些模型中最优的几个计算 C_p , 其中 $\hat{\sigma}^2$ 总是由 13 个自变量的完全模型计算得到的。表 8.7 列出有最小 C_p 的 20 个模型, 其中 10 个含虚拟变量, 10 个不含。它们是用 Furnival 和 Wilson 算法求得的, 在一台 CDC Cyber72 计算机上所花的时间少于 1 秒钟。表中列出的有 p = 模型中参数个数, C_p , R^2 , RSS_p 及在子集模型中的

表 8.7 具有最小的 C_p 的 20 个模型

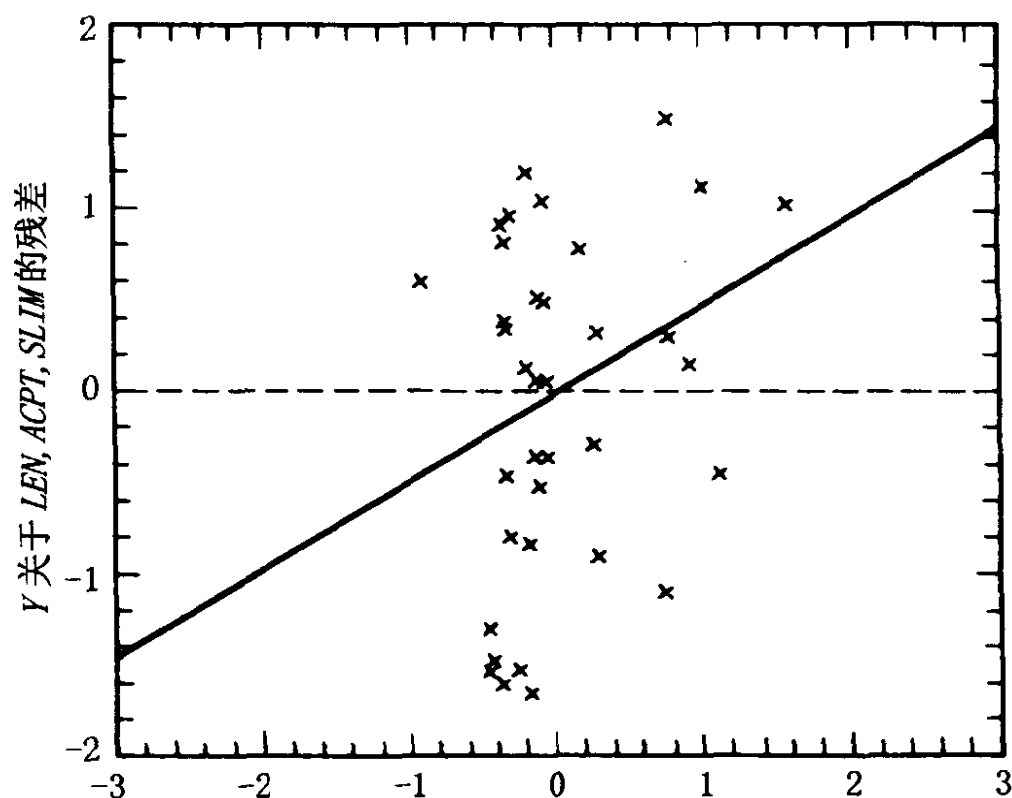
p	C_p	R^2	RSS_p	模型中的自变量					
4	.23	.701	44.8465	LEN	SLIM	ACPT			
5	.48	.718	42.3333	LEN	SLIM	SIGS	ACPT		
5	.56	.717	42.4400	LEN	TRKS	SLIM	ACPT		
5	1.33	.710	43.5449	LEN	SLIM	ACPT	LANE		
5	1.59	.707	43.9135	LEN	ADT	SLIM	ACPT		
5	1.63	.707	43.9754	LEN	SLIM	ITG	ACPT		
5	1.97	.703	44.4587	LEN	SLIM	LWID	ACPT		
6	1.51	.727	40.9282	LEN	TRKS	SLIM	SIGS	ACPT	
6	2.00	.722	41.6406	LEN	TRKS	SLIM	ACPT	LANE	
6	2.00	.722	41.6406	LEN	TRKS	SLIM	ITG	ACPT	
7	4.69	.716	42.6231	LEN	SLIM	ACPT	FAI	PA	MA
7	5.12	.712	43.2450	LEN	SLIM	SIGS	FAI	PA	MA
8	3.32	.748	37.7958	LEN	SLIM	SIGS	ACPT	FAI	PA MA
8	4.68	.735	39.7448	LEN	TRKS	SLIM	ACPT	FAI	PA MA
9	4.44	.756	36.5374	LEN	TRKS	SLIM	SIGS	ACPT	FAI PA MA
9	5.10	.750	37.4670	LEN	SLIM	ITG	SIGS	ACPT	FAI PA MA
9	5.12	.750	37.5032	LEN	SLIM	SHLD	SIGS	ACPT	FAI PA MA
9	5.25	.749	37.6839	LEN	ADT	SLIM	SIGS	ACPT	FAI PA MA
9	5.25	.748	37.6934	LEN	SLIM	LWID	SIGS	ACPT	FAI PA MA
9	5.30	.748	37.7629	LEN	SLIM	SIGS	ACPT	LANE	FAI PA MA

自变量。截距被包含在表中的所有模型中。

所有的最小 C_p 模型都包含被强制加入所有模型的 *LEN* 以及 *SLIM*，同时 *ACPT* 被包含在大多数模型中。加入 *TRKS*，

SIGS, *ITG* 或 *LANE* 中的任一个, 可能有某些有用的影响, 但一般不需要加入它们中的两个或更多。

表 8.7 的检查表明, 没有一个模型是明显最优的, 并且有许多同样好的模型。进一步的分析将明确指出决定采用哪个模型。然而, 除 *LEN* 外, *SLIM* 和 *ACPT* 是显然有用的。另外, 虚拟变量的重要性可以通过 *SLIM*、*ACPT* 和 *LEN* 调整后的这些变量的 *F*-检验的判断得到。因为 $F = [(44.85 - 42.62)(32)] / [(42.62)(3)] = 0.56$, 其自由度为 (3, 32), 所以没有什么证据认为在预测事故率时虚拟变量是重要的。



SIGS 关于 *LEN*, *ACPT*, *SLIM* 的残差
图 8.3 在包含 *LEN* *ACPT* 和 *SLIM* 的
模型中加入 *SIGS* 的影响

求得一个模型的合理的方法是由只使用 *LEN*、*SLIM* 和 *ACPT* 作为自变量的模型开始。然后可以进行诊断分析, 这留给有兴趣的读者。这一模型的回归分析简要见表 8.8。这一模型的一个有趣特征是 *SLIM* 的系数估计的符号——较高的速度限制 (低

于 55 英里/小时的最大限) 与较低的事故率相联系。容易——但不正确地——断言, 较高的速度使得有较低的事故率。事实上, 高速公路部门对高事故率的一个反应是降低速度限制。这样, 高事故率可能引起较低的速度限制, 但反之不成立。

为决定附加自变量是否应加入由表 8.8 总结的模型中, 在 2.4 节中首先引入的附加变量图能给出一个图形化的帮助。作为一个例子, 将 *SIGS* 加入包括 *LEN*、*SLIM* 和 *ACPT* 的模型的影响如图 8.3 所示。图的实线有斜率 0.485。这是如果将 *SIGS* 加入到模型中对 *SIGS* 的估计。由图可见, 如果加入 *SIGS*, 几乎得不到什么系统信息。在四个变量的模型中, *SIGS* 的 *t* 值为 1.42, 就说明了这个问题。这样我们决定, 不把 *SIGS* 加入到模型中。

最终模型 没有最终模型, 只有一组被认为几乎同等有效的可能的模型。如果一个模型用于预测, 我们并不在乎准确的自变量是否被包含在内。在这种情况下, 节约原则——越少越好——经常被使用。在其它问题中, 分析者对一个问题的知识, 被用于获得结论。

表 8.8 主 要 回 归 量

变量	估计	标准误	<i>t</i> -值
截距	9.32	2.617	3.56
<i>LEN</i>	-0.0771	0.0249	-3.10
<i>SLIM</i>	-0.102	0.0429	-2.39
<i>ACPT</i>	0.101	0.0276	3.72
$\hat{\sigma}^2 = 1.281, \text{d.f.} = 35, R^2 = 0.70$			

问 题

8.1 对以下数据应用 BE 和 FS 算法。对所有可能的回归求 C_p , 并比较结果。

什么是“正确的模型”? (Mantel, 1970)

Y	X_1	X_2	X_3
5	1	1004	6.0
6	200	806	7.3
8	-50	1058	11.0
9	909	100	13.0
11	506	505	13.1

- 8.2 在高速公路事故数据中,对拟合的完全模型,估计高速公路类型的改变对事故率的影响。所考虑的类型改变有(1)从 *MC* 到 *MA*; (2) *MA* 到 *PA*; (3) *PA* 到 *FAI*。
- 8.3 对高速公路事故数据进行诊断分析。是否需要诊断? 如果需要, 变换数据并求有最小 C_p 的模型。
- 8.4 对例 7.2 的云的催化数据, 求有最小 C_p 的五个模型。
- 8.5 对“伯克利指导研究”中的男孩 (问题 2.1), 求 *SONA* 作为其它变量的函数的模型。进行完整的分析, 包括诊断分析, 并总结你的结论。
- 8.6 用 Schatzoff et al. (1968) 描述的方法, 写一个对所有可能回归计算 C_p 的程序。
- 8.7 有人进行一项试验, 从五种化学物质的观测值, 来对氧气吸收 (*O2UP*, 以毫克氧气/分钟为单位) 建立模型 (Moore, 1975)。五种化学物质为: 生物氧气需求 (*BOD*), 全部 Kjeldahl (氮气) (*TKN*), 全部固体 (*TS*), 全部易挥发固体 (*TVS*), *TVS* 是 *TS* 的一个组成部分, 以及化学氧气需求 (*COD*)。每一项的度量以毫克/升为单位。在实验室, 废弃物被保持悬浮于水中 220 天。以每日的废弃物为样本收集数据, 在这段时间内, 所有的观测都是对同一样本。我们想要求得一个联系 $\log(O2UP)$ 与其它变量的方程。目标是求出必须被进一步研究的变量, 以达到建立预测方程的最终目的 (天数不能用作自变量)。数据由表 8.9 给出。表 8.10 给出所有可能回归的主要统计量, C_p , R^2 , RSS 。

当被考虑的变量总数较小 (这里, $k=5$) 时, C_p - p 关于 p 的图, 称为 C_p 图, 是 C_p 统计量的一个方便的总述。 C_p 图常被建议为是 C_p 关于 p 的图, 但这里建议的方法, 使解释变得略微容易。好的模型, C_p - p 一般将小于零。

表 8.9 氧气吸收实验数据 (Moore, 1975)

天数	BOD	TKN	TS	TVS	COD	02UP	Log(02UP)
0	1125.	232	7160.	85.9	8905.	36.0	1.5563
7	920.	268.	8804.	86.5	7388.	7.9	0.8976
15	835.	271.	8108.	85.2	5348.	5.6	0.7482
22	1000.	237.	8370.	83.8	8056.	5.2	0.7160
29	1150.	192.	6441.	82.1	6960.	2.0	0.3010
37	990.	202.	5154.	79.2	5690.	2.3	0.3671
44	840.	184.	5896.	81.2	6932.	1.3	0.1139
58	650.	200.	5336.	80.6	5400.	1.3	0.1139
65	640.	180.	5041.	78.4	3177.	0.6	-0.2218
72	583.	165.	5012.	79.3	4461.	0.7	-0.1549
80	570.	151.	4825.	78.7	3901.	1.0	0.0000
86	570.	171.	4391.	78.0	5002.	1.0	0.0000
93	510.	243.	4320.	72.3	4665.	0.8	-0.0969
100	555.	147.	3709.	74.9	4642.	0.6	-0.2218
107	460.	286.	3969.	74.4	4840.	0.4	-0.3979
122	275.	198.	3558.	72.5	4479.	0.7	-0.1549
129	510.	196.	4361.	57.7	4200.	0.6	-0.2218
151	165.	210.	3301.	71.8	3410.	0.4	-0.3979
171	244.	327.	2964.	72.5	3360.	0.3	-0.5229
220	79.	334.	2777.	71.9	2599.	0.9	-0.0458

8.7.1 作 C_p 图。求有最小 C_p 的模型。找出所有的模型,使得没有包括在它里面的变量的 F -值小于2。总结结论。

8.7.2 完成对这些数据的分析,包括一个完整的诊断分析。什么诊断表

示需要将 02UP 变换为对数尺度?

8.8 对一个固定的 p 个参数的子集模型, 试由 (8.21) 求得一个检验, 其原假设 $NH: J_p \leq p$, 备择假设 $AH: J_p > p$ 。

表 8.10 以 $\log(02UP)$ 作为响应变量的所有可能回归

p	C_p	R^2	RSS	Model		
2	6.29	.697	1.5370	TS		
2	6.57	.693	1.5563	COD		
2	13.50	.598	2.0338	BOD		
2	20.33	.505	2.5044	TVS		
2	56.84	.008	5.0219	TKN		
3	1.74	.786	1.0850	TS	COD	
3	5.27	.738	1.3287	TVS	COD	
3	6.87	.716	1.4388	TKN	COD	
3	6.88	.716	1.4397	BOD	TS	
3	7.16	.712	1.4590	TS	TVS	
3	7.33	.710	1.4707	TKN	TS	
3	7.70	.704	1.4963	BOD	COD	
3	9.09	.686	1.5921	BOD	TKN	
3	11.33	.655	1.7462	BOD	TVS	
3	21.36	.518	2.4381	TKN	TVS	
4	2.32*	.805	0.9871	TKN	TS	COD
4	3.42	.790	1.0634	TS	TVS	COD
4	3.44	.790	1.0644	BOD	TS	COD
4	5.66	.760	1.2178	TKN	TVS	COD
4	6.25	.752	1.2582	BOD	TKN	TS
4	6.51	.748	1.2764	BOD	TKN	COD
4	7.15	.739	1.3204	BOD	TVS	COD

(续表)

p	C_p	R^2	RSS	Model			
4	8.15	.726	1.3894	<i>BOD</i>	<i>TS</i>	<i>TVS</i>	
4	8.16	.726	1.3900	<i>TKN</i>	<i>TS</i>	<i>TVS</i>	
4	8.68	.718	1.4257	<i>BOD</i>	<i>TKN</i>	<i>TVS</i>	
5	4.00	.809	0.9653	<i>TKN</i>	<i>TS</i>	<i>TVS</i>	<i>COD</i>
5	4.32	.805	0.9871	<i>BOD</i>	<i>TKN</i>	<i>TS</i>	<i>COD</i>
5	5.07	.795	1.0388	<i>BOD</i>	<i>TS</i>	<i>TVS</i>	<i>COD</i>
5	6.78	.772	1.1565	<i>BOD</i>	<i>TKN</i>	<i>TVS</i>	<i>COD</i>
5	7.70	.759	1.2199	<i>BOD</i>	<i>TKN</i>	<i>TS</i>	<i>TVS</i>
6	6.00	.809	0.9652	<i>BOD</i>	<i>TKN</i>	<i>TS</i>	<i>TVS</i> <i>COD</i>

8.9 证明结论 (8.5)。在回归模型的均值偏差形式中, 使用以下关于一个 2×2 对称矩阵的逆的表达式,

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix}^{-1} = \frac{1}{ab-c^2} \begin{pmatrix} a & -c \\ -c & b \end{pmatrix}$$

可以直接推得结论 (8.5)。

8.10 证明结论 (8.6)。(提示: 为避免繁复的代数运算, 使用问题 2.7 描述的扫描算法)。如果不仔细考虑, 这一问题的解答可能非常长, 且不提供信息。

8.11 在厄瓜多尔沿岸的 Galápagos 群岛是一个极好的实验室, 可用来研究影响不同生物种类的发展与生存的因素。Johnson 和 Raven (1973) 给出了表 8.11 的数据。其中有 29 个不同岛屿的生物种类数以及有关变量。表既给出了物种总数, 也给出了只在那一个岛出现的 (地方特有的) 物种数。

用这些数据求出影响这一差异的因子, 其中的差异用物种数及地方特有的物种数的某个函数来度量。总结你的结论。一个复杂的因素是, 六个非常小的岛屿没有记录海拔高度, 故必须对此作出某些预处理。四种可能的处理是: (1) 求得海拔高度; (2) 从数据中删去这六个小岛; (3) 忽略海拔高度, 不让它作为差异的自变量; (4) 对缺省数据给出可信的替代值。查看大比例的地图, 我们发现它们中没有一个海拔超过 200m 的。

表 8. 11 Calápagos 岛物种数据

岛屿	观测物种		距离 (km)				
	物种数	地 方 特有的 物种数	面积 (km ²)	海拔 (m)	距最近 的岛	距 Santa Cruz 岛	相邻岛 面积 (km ²)
Baltra	58	23	25.09	—	0.6	0.6	1.84
Bartolomé	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamano	2	1	0.05	—	1.9	1.9	903.82
Daphne Major	18	11	0.34	—	8.0	8.0	1.84
Darwin	10	7	2.33	168	34.1	290.2	2.85
Eden	8	4	0.03	—	0.4	0.4	17.95
Enderby	2	2	0.18	112	2.6	50.2	0.10
Espanola	97	26	58.27	198	1.1	88.3	0.57
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32
Gardner *	58	17	0.57	49	1.1	93.1	58.27
Gardner ***	5	4	0.78	227	4.6	62.2	0.21
Genovesa	40	19	17.35	76	47.4	92.2	129.49
Isabela	347	89	4669.32	1707	0.7	28.1	634.49
Marchena	51	23	129.49	343	29.1	85.9	59.56
Onslow	2	2	0.01	25	3.3	45.9	0.10
Pinta	104	37	59.56	777	29.1	119.6	129.49
Pinzon	108	33	17.95	458	10.7	10.7	0.03
Las plazas	12	9	0.23	—	0.5	0.6	25.09

(续表)

岛屿	观测物种		距离 (km)				
	物种数	地 方 特有的 物种数	面积 (km ²)	海拔 (m)	距最近 的岛	距 Santa Cruz 岛	相邻岛 面积 (km ²)
Rabida	70	30	4.89	367	4.4	24.4	572.33
San Cristóbal	280	65	551.62	716	45.2	66.6	0.57
San Salvador	237	81	572.33	906	0.2	19.8	4.89
Santa Cruz	444	95	903.82	864	0.6	0.0	0.52
Santa Fé	62	28	24.08	259	16.5	16.5	0.52
Santa Maria	285	73	170.92	640	2.6	49.2	0.10
Seymour	44	16	1.84	—	0.6	9.6	25.09
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

* Near Espanola.

* * Near Santa Maria.

来源: Johnson and Raven(1973)。

9

预 测

回归的一个最重要的用处是对自变量的给定值预测响应变量的未来值。由于拟合方程被预定能对给定的自变量给出响应变量的期望值，故回归方法对完成上述任务似乎是理想的。然后我们几乎就把估计的预测函数看成变量间的真实关系似地继续进行讨论。例如，假设一个重物从一幢高楼上落下，预计物体在七秒内落下的距离为 $4.9t^2$ 米。只要 t 足够小或楼房足够高，预测将是相当准确的。参数 4.9 米/秒² 为重力加速度的一半，是一个已知常数。由于测量误差或被忽略的因素如摩擦，用这一方程所作的关于距离的预测可能与真实的观测值不能很好地吻合。

如果重力加速度未知，预测值可以通过收集数据估计得到。可能的方法是落下一个物体，测量它在不同的时间内落下的距离。令 d 表示距离， t 表示时间，我们对观测数据可以拟合模型 $d = \gamma t^2$ ，并估计 γ 的未知值。由含有估计的参数的方程所作的预测，将由于估计的不确定性而有所不同，但只要方程的函数形式是正确的，它们仍将是可靠的。

更一般的回归问题，因为很少知道正确的函数形式，比这个要复杂得多。用经验选取的模型会使预测值的可用性有所下降。一般地，我们依据于以下事实，即对限制在一定范围内的自变量，许

多模型的效果几乎一样，所以即使拟合了错误的模型，我们仍可能得到有用的预测。不过，对自变量的某些值，估计可能是没用的，见图 9.1。在物体下落的实验中，如果不知道真实的函数形式，我们可能决定用本书的方法，拟合一个线性模型，也许是用 $\log(t)$ 作为自变量。只要预测时 t 值的范围大致与我们建立模型时 t 值的范围相同，得到的预测值能被预期是相当好的，这是因为在 t 的一个适当的范围内， $d = \gamma t^2$ 可由 $d = \beta_0 + \beta_1 \log(t)$ 来近似。

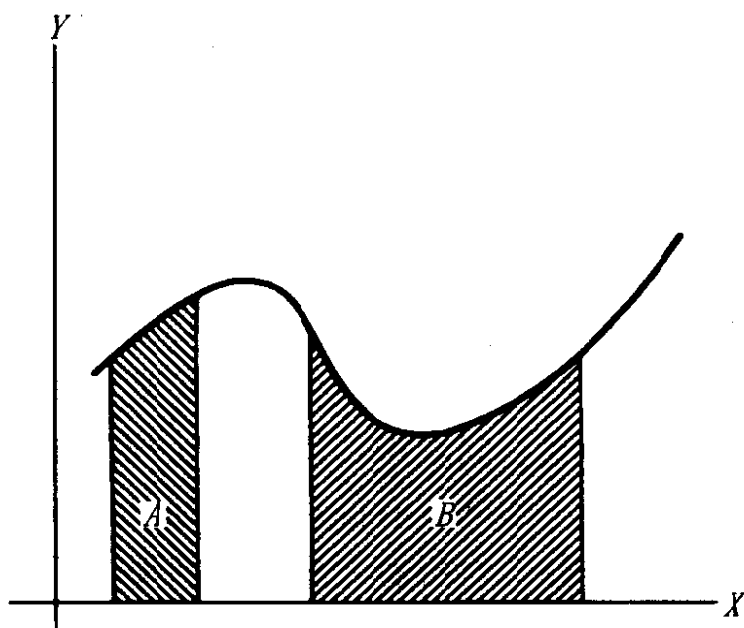


图 9.1 在区域 A，一个线性近似可能是合适的，而对 B，可能二次近似是合理的。在拟合范围以外，任何一个都不是合适的

如果变量间的函数关系是已知的，参数的估计与预测之间的区别被模糊了，并且参数可以有物理含义，如前面讨论的重力加速度。如果函数关系是未知的，则估计与预测之间的区别是有意义的、且有指导性的。我们为了获取预测值而估计参数。参数本身很少有固有的含义。如果我们对一组落体数据拟合 $\hat{d} = \hat{\beta}_0 + \hat{\beta}_1 \log(t)$ ， β_0 和 β_1 是拟合方程的产物，而不是时间与距离的关系。估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 依赖于估计中自变量的范围，所以参数 β_0 和 β_1 代表依赖于数据以及时间与距离关系的两个量，它们是“变量常

数”。不过，如果新的案例是取自与原始数据一样的总体，则基于这些变量常数的预测可能是有用的。

9.1 进行预测

数据从何而来？理想地，用于建立预测函数以及预测未来案例的数据，在某种意义上，是一个界限分明的总体的一个样本。如果是这样的话，从样本到总体的统计推断的一般方法是适用的。但是我们常常不能保证这些条件。例如，在一个化学实验中，可以通过重复具有变化的自变量值的组合的试验来得到许多案例。如果实验是在一天内，或由一位实验员，或在一个化学实验室内完成，则被估计的预测方程不一定适用于不同的日子，或不同的实验员，或不同的实验室的实验结果。我们必须或者假设起因于这些因素的影响是可忽略的，或者从多个实验员或实验室收集数据来估计这些影响。在其它问题中，用于估计预测函数构造的样本只对较大样本中的一小部分具有设计代表性。例如，只用健康老鼠进行的每日饮食研究，对所有老鼠，包括那些不健康的老鼠作预测可能是不合适的。

内插与外推 一般地，一个预测方程只能适用于自变量的某些值。例如，在一个化学实验中变换到极端气温，可能使事实上的结论与根据中等气温下的实验所作的预测完全不同。从而在大多数预测问题中，有一个确定自变量的有效范围的重要问题。在那个范围内的预测称为内插，而在那个范围以外的我们称为外推。一个不幸的事实是，尽管外推值很不可靠，我们却往往对它更感兴趣。

作预测 如本书所描述的，回归模型是一门科学，同样也是一门艺术。没有两个问题能用完全相同的方法处理，因此我们不能轻易地描述出求得一个预测方程的过程。不过一般地，我们可以描述一个三阶段过程：数据收集，模型选择与估计，以及模型

的确认。数据的收集是这一过程的重要部分，因为如果没有好的数据，其它阶段是没有用的。

模型的选择由变量的选择和变换组成，包括使用本书前面所描述的诊断方法。因为使第八章描述的 C_p 最小，就好比选取一个模型，使预测误差尽可能小，所以 C_p 准则使预测的作用明显化。一旦选择了一个函数形式，需要估计那个函数的参数。尽管非最小二乘的其它估计方法有可能是更好的 (Copas, 1983)，但我们将继续使用最小二乘估计，仿佛所选取的模型预先就确定了。这样，我们假设 Y 是将要使用的响应变量，选取的自变量由一个 $n \times p'$ 的矩阵 X 给出。线性模型为

$$Y = X\beta + e \quad (9.1)$$

并且 β 的估计取为 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 。对一个自变量为 x_* 的新的案例，我们希望预测至今未观测的响应变量的值 y_* 。预测值为 $\tilde{y}_* = x_*^T \hat{\beta}$ 。如果线性模型是正确的，预测值为 $E(y_*)$ 的一个无偏点估计。再次假设模型是正确的，则预测的方差将是 $\text{varpred}(\tilde{y}_* | x_*) = \sigma^2 (1 + x_*^T (X^T X)^{-1} x_*) = \sigma^2 (1 + h_*)$ 。由于一般是用模型 (9.1) 给出的残差均方 $\hat{\sigma}^2$ 来估计 σ^2 ，预测的标准误估计为

$$\text{sepred}(\tilde{y}_* | x_*) = \hat{\sigma}^2 \sqrt{1 + h_*} \quad (9.2)$$

如果是通过数据分析来选择模型 (9.1)，则预测的标准误的公式 (9.2) 可能会低估预测误差。首先，模型的选择是使 σ^2 的估计相对地小，故可能低估残差方差。另外，如果函数形式未知，(9.1) 的预测可能是有偏的，需要加上偏差的平方，以得到均方误差去代替 (9.2)。

与参数不同，未来值有一个有用的特征，它是可观测的。我们可以使用一个拟合模型来求得预测值，然后视其预测的好坏给出预测误差的直接估计。这样就将我们带到作预测的第三步：模型的确认。这可以通过多种途径完成，但对每一途径至少有一个隐含的要求：拟合模型必须用来求未用于估计的数据的预测值。我

们在第八章已经遇到关于这方面的两种方法，使用相互证实及 *PRESS*，或重复使用样本的统计量来度量平均预测误差。

在相互证实中，数据被分为两个或多个子集。其中一个子集称为结构集合，用于估计。其它子集称为确认集。由结构集合拟合的模型可以求得在确认集中的案例的预测。它们可以与响应变量的观测值进行比较。一个有用的准则函数是预测的均方误差的平方根。在可以得到大量的数据时，这是一个有吸引力的方法。数据可以随机地或者用某些优化的方法 (Snee, 1977)，分为几个子集。

相互证实的极端情况，是将数据分为 n 个重迭子集，每个子集由 $n-1$ 个案例组成。由 $n-1$ 个案例得到的估计，用于预测一个被删除案例的值。这导致 8.8 节中描述的 *PRESS* 统计量。假设手头的数据是取自未来值总体的一个样本，则平均预测误差的一个敏感的估计为 $(PRESS/n)^{1/2}$ 。

例 9.1 老忠实间歇泉喷发间隔的估计

一个热泉偶尔会变得不稳定，并喷发热水和气到空气中。这时，它称为间歇泉。不同的间歇泉有不同长度的时间和间隔。最著名的间歇泉大概是怀俄明州黄石国家公园中的老忠实间歇泉。老忠实间歇泉的喷发间隔在 30 到 90 分钟之间变化。水柱高度一般超过 35 米，喷发持续 1 到 $5\frac{1}{2}$ 分钟 (Marller, 1969)。由于其规律性及美丽景观，老忠实间歇泉是一个重要的吸引游客的景点。在一个夏天的下午，有数千人观看一次喷发是不足为奇的。

对这一间歇泉下次喷发时间的预测是公园服务部门以及游客都感兴趣的。事实上，下次喷发的时间的预测张贴在间歇泉附近的显著位置。表 9.1 给出的为 y = 到下一次喷发的间隔及 x = 喷发的持续时间，都以分钟计算。它们是从 1978 年 8 月 1 日至 8 月 8 日早晨 8 时至午夜老忠实间歇泉的所有喷发情况*。

* 非常不幸，本书第一版给出的数据是错的。在这一版中已作了更正。这些数据由黄石公园地理学家 Roderick A. Hutchinson 提供。

表 9.1 老忠实间歇泉的喷发
(1978 年 8 月 1 日至 8 月 8 日)

日	y	x	日	y	x	日	y	x	日	y	x
1	78	4.4	2	80	4.3	3	76	4.5	4	75	4.0
1	74	3.9	2	56	1.7	3	82	3.9	4	73	3.7
1	68	4.0	2	80	3.9	3	84	4.3	4	67	3.7
1	76	4.0	2	69	3.7	3	53	2.3	4	68	4.3
1	80	3.5	2	57	3.1	3	86	3.8	4	86	3.6
1	84	4.1	2	90	4.0	3	51	1.9	4	72	3.8
1	50	2.3	2	42	1.8	3	85	4.6	4	75	3.8
1	93	4.7	2	91	4.1	3	45	1.8	4	75	3.8
1	55	1.7	2	51	1.8	3	88	4.7	4	66	2.5
1	76	4.9	2	79	3.2	3	51	1.8	4	84	4.5
1	58	1.7	2	53	1.9	3	80	4.6	4	70	4.1
1	74	4.6	2	82	4.6	3	49	1.9	4	79	3.7
1	75	3.4	2	51	2.0	3	82	3.5	4	60	3.8
									4	86	3.4
5	71	4.0	6	55	1.8	7	81	3.5	8	77	4.2
5	67	2.3	6	75	4.6	7	53	2.0	8	73	4.4
5	81	4.4	6	73	3.5	7	89	4.3	8	70	4.1
5	76	4.1	6	70	4.0	7	44	1.8	8	88	4.1
5	83	4.3	6	83	3.7	7	78	4.1	8	75	4.0
5	76	3.3	6	50	1.7	7	61	1.8	8	83	4.1
5	55	2.0	6	95	4.6	7	73	4.7	8	61	2.7
5	73	4.3	6	51	1.7	7	75	4.2	8	78	4.6
5	56	2.9	6	82	4.0	7	73	3.9	8	61	1.9
5	83	4.6	6	54	1.8	7	76	4.3	8	81	4.5
5	57	1.9	6	83	4.4	7	55	1.8	8	51	2.0
5	71	3.6	6	51	1.9	7	86	4.5	8	80	4.8
5	72	3.7	6	80	4.6	7	48	2.0	8	79	4.1
5	77	3.7	6	78	2.9						

注: x = 持续时间, y = 间隔 (都以分钟计)。

数年来, 公园里的管理者/自然学家用秒表对老忠实间歇泉收集了数据。喷发持续时间被舍入至最接近的 0.1 分钟, 即 6 秒, 而喷发间隔被舍入至分钟。国家公园服务部门用 x 的值预测 y 的未来值。我们将仿效他们的例子进行预测。在 x 和 y 之间不太可能有一种因果关系, 而可能他们都是某个或某组未被观测的其它因素影响的结果。由 x 和 y 之间观测得到的联系建立的模型可能导致有效的预测, 但拟合的关系可能不具有地质意义。

表 9.1 中 107 个案例的 y 关于 x 的散点图由图 9.2 给出。图大致显示出, 短的喷发后跟随着短的间隔, 而长的喷发后有长的间隔。点好象落在右上角和左下角的两个点簇中, 图的大致的线性性建议使用一个简单线性回归模型,

$$y = \beta_0 + \beta_1 x + \text{误差} \tag{9.3}$$

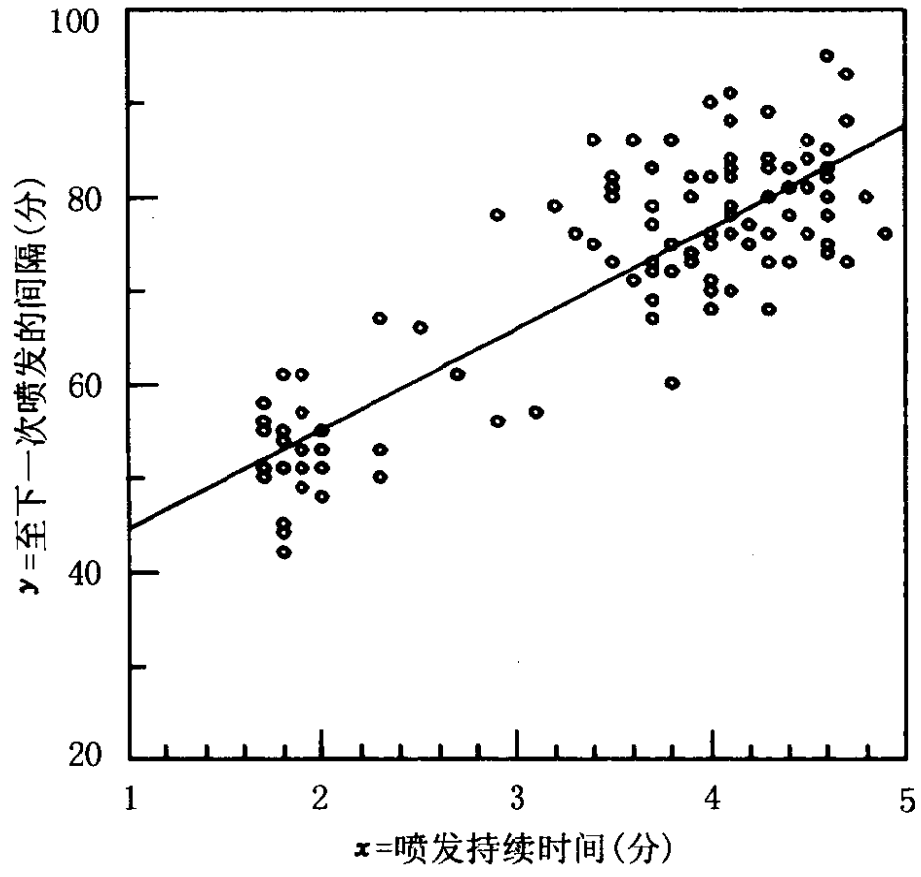


图 9.2 老忠实间歇泉数据

拟合模型为

$$y = 33.8 + 10.7x \tag{9.4}$$

一些主要统计量见表 9.2。由 (9.2), 预测的标准误为

$$\text{sepred } (\hat{y}_* | x_*) = 6.68 \left[1 + \frac{1}{107} + \frac{(x_* - 3.461)^2}{113.83} \right]^{1/2} \quad (9.5)$$

例如, 如果 $x_* = 4.5$ 分钟, $\hat{y}_* = 33.8 + 10.7(4.5) = 82.2$ 分钟, 预测的标准误差为 6.8 分钟。用由 $t(105)$ 分布得到的一个乘数, 可以得到一个 95% 的预测区间

$$82.2 - t(0.05; 105)6.8 \leq y_* \leq 82.2 + t(0.05; 105)6.8$$

$$68.7 \leq y_* \leq 95.7$$

表 9.2 老忠实间歇泉数据, 简单回归的主要统计量

变量	n	均值	标准差	最小值	最大值
x	107	3.4607	1.0363	1.7	4.9
y	107	71.000	12.967	42	95

y 关于 x 的回归

变量	估计	标准误	t -值
截距	33.83	2.26	14.96
x	10.74	0.63	17.15

$$\hat{\sigma} = 6.68, \text{ d. f. } = 105, R^2 = 0.74$$

对这些数据的简单回归模型的确认可以用两种方法完成。首先, 我们可以用 $(PRESS/n)^{1/2}$ 来估计平均预测误差。对这些数据及这一模型的 $(PRESS/n)^{1/2} = (4,808/107)^{1/2} = 6.7$ 分钟。为使用相互证实的方法, 我们需要更多的数据。在原始数据的一年之后, 从 1979 年 8 月 16 日至 8 月 23 日, 老忠实泉的 115 次喷发情况见表 9.3。为使得相互证实方法有意义, 我们必须假设在这两段时间里, 支配 x 和 y 的基本的作用没有变化。故我们能够认为表 9.1 和 9.3 中的数据来自相同的总体。对确认集的每一个 x 值, 运用模型 (9.4), 均方拟合误差的平方根是 $(4029.1/115)^{1/2} = 5.9$ 分钟。认为这两个证实方法是接近一致的, 故在使用 (9.4) 去预测老忠实间歇泉喷发时间时, 我们能合理地预料预测标准误差大约为 6 分钟。

尽管简单模型 (9.4) 提供的预测有 90% 精确到 10 分钟之内, 但仍需要对模型作一些改进。如果按时间次序检验数据点列, 可以看到一个小于 60 分钟的短间隔之后总是跟随着一个长间隔。然而, 长的间隔常跟随在长的间隔

表 9.3 1979 年 8 月老忠实间歇泉的喷发

日	y	x	日	y	x	日	y	x	日	y	x
16	82	4.1	17	91	4.8	18	58	2.2	19	84	4.5
16	80	4.2	17	66	4.1	18	82	4.8	19	72	3.8
16	76	4.5	17	71	4.0	18	77	4.3	19	89	4.3
16	56	1.9	17	75	4.0	18	75	3.8	19	75	4.4
16	82	4.7	17	81	4.4	18	77	4.0	19	57	2.2
16	47	2.0	17	77	4.1	18	77	4.1	19	81	4.8
16	76	4.7	17	74	4.3	18	53	1.8	19	49	1.9
16	61	2.5	17	70	4.0	18	75	4.4	19	87	4.7
16	75	4.3	17	83	3.9	18	78	4.0	19	43	1.8
16	72	4.4	17	53	3.2	18	51	2.2	19	94	4.8
16	74	4.4	17	82	4.5	18	81	5.1	19	45	2.0
16	69	4.3	17	62	2.2	18	52	1.9	19	81	4.4
16	78	4.6	17	73	4.7	18	76	5.0	19	59	2.5
16	52	2.1	17	84	4.6	18	73	4.4	19	82	4.3
20	80	4.4	21	84	3.7	22	72	3.0	23	51	1.7
20	54	1.9	21	58	1.8	22	54	2.1	23	83	4.4
20	75	4.7	21	90	4.7	22	75	4.6	23	76	4.2
20	73	4.3	21	82	4.5	22	74	4.0	23	51	2.2
20	57	2.2	21	71	4.5	22	51	2.2	23	90	4.7
20	80	4.7	21	80	4.8	22	91	5.1	23	71	4.0
20	51	2.3	21	51	2.0	22	60	2.9	23	49	1.8
20	77	4.6	21	80	4.8	22	80	4.3	23	88	4.7
20	66	3.3	21	62	1.9	22	54	2.1	23	52	1.8
20	77	4.2	21	84	4.7	22	80	4.7	23	79	4.5
20	60	2.9	21	51	2.0	22	70	4.5	23	61	2.1
20	86	4.6	21	81	5.1	22	60	1.7	23	81	4.2
20	62	3.3	21	83	4.3	22	86	4.2	23	48	2.1
20	75	4.2	21	84	4.8	22	78	4.3	23	84	5.2
20	67	2.6							23	63	2.0
20	69	4.6									

之后。这表示案例互不相关这一通常的假设可能不成立。将这一附加信息吸收到模型中的一个近似的方法是用 y 和 x 的过去值来预测 y 的当前值。例如，我们可以对 y ，即到下一次喷发的间隔，建立如下的模型

$$y = \beta_0 + \beta_1 x + \beta_2 (\text{上一次喷发的持续时间}) + \beta_3 (\text{上两次喷发的间隔}) + \text{误差} \quad (9.6)$$

这两个附加变量称为滞后变量。

为拟合模型 (9.6)，我们将表 9.1 中的数据削减至 99 个案例。这是因为对每天记录的第一次喷发，滞后的间隔与持续时间是未知的。例如 8 月 1 日的数据为

y	x	y (滞后的)	x (滞后的)
74	3.9	78	4.4
68	4.0	74	3.9
76	4.0	68	4.0
\vdots	\vdots	\vdots	\vdots
74	4.6	58	1.7
75	3.4	74	4.6

y 关于 x ， x (滞后的) 及 y (滞后的) 的回归的综述由表 9.4 给出。所有的自变量有大的 t -值。模型看起来是对只含 x 的模型的一个有限改进，这是因为残差方差有所变小，并且在图 9.2 中案例的丛集，或在残差关于预测值的图中的丛集，在模型 (9.6) 的同意义的图中将会消失。这些图被省略了。为验证这一模型，我们可以再次使用 *PRESS* 以及相互证实的方法。这些方法都能给出一个约 6.2 分钟的平均预测误。关于预测，没有明显的理由说明有必要用更复杂的 (9.6) 代替简单回归模型 (9.4)。

表 9.4 滞后变量的回归

变量	估计	标准误	t -值
截矩	64.15	6.96	9.21
x	88.85	0.80	11.12
y (滞后的)	-0.54	0.10	-5.12
x (滞后的)	4.08	1.17	3.48

$\hat{\sigma} = 6.07$, d. f. = 95, $R^2 = 0.80$

9.2 内插法对外推法

内插法意味着对新的案例作出预测，该案例的自变量与在结构样本中的自变量的值没有太大的差别。对一个简单线性回归模型，一般当自变量在结构样本的观测范围内时为内插；在这一范围以外，我们称这一预测为外推。在老忠实间歇泉数据中，如果 x 非常短（如几秒钟），或非常长（如 10 分钟），则我们不愿对 y 作预测，因为在我们的数据中没有在这些条件下 y 与 x 关系的信息。

图 9.3 说明了，内插法与外推法的区别的有用性。如果预测函数的形式预先未知，则在自变量的观测值范围以外，我们没有这一关系的任何信息。图中所示的任何一条虚线路径都可能是合适的。如果没有进一步的信息，我们不可能知道应该用哪一条。为

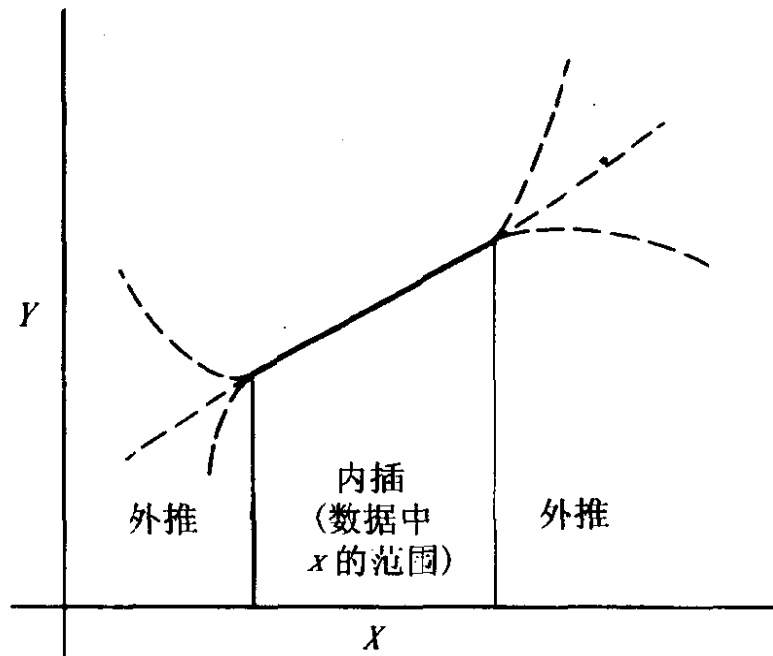


图 9.3 外推法

使内插法有用，只需满足几个假设，主要是新的案例与结构样本中的那些案例的情况要大致相同。对外推法，假设“估计的预测函数与具有在内插范围以外的自变量的案例有关”仍是需要的。例如，降雨一般会增加农作物产量。当降雨量在一个合理范围内时，收集的数据可能给出一个由降雨量给出的产量的一个好的预测函数。不过这一函数对外推值可能是没有用的。根据方程，零降雨量可能导致负的农作物产量，而非常大的降雨量可能预测得到非常大的产量——尽管在某个值之后，再大的降雨量会降低产量。

在多元回归中，很难对预测的有效性定义一个范围。很明显，它依赖于在结构样本中测得的数据。作为一个例子，假设我们考虑老忠实间歇泉的数据，我们只用 x 和 x （滞后的）作为自变量，这样可以得到一个二维图象。图 9.4 给出 x 关于 x （滞后的）的一个散点图。一个预测被称为是一个内插值，如果它的 $(x, x$ （滞

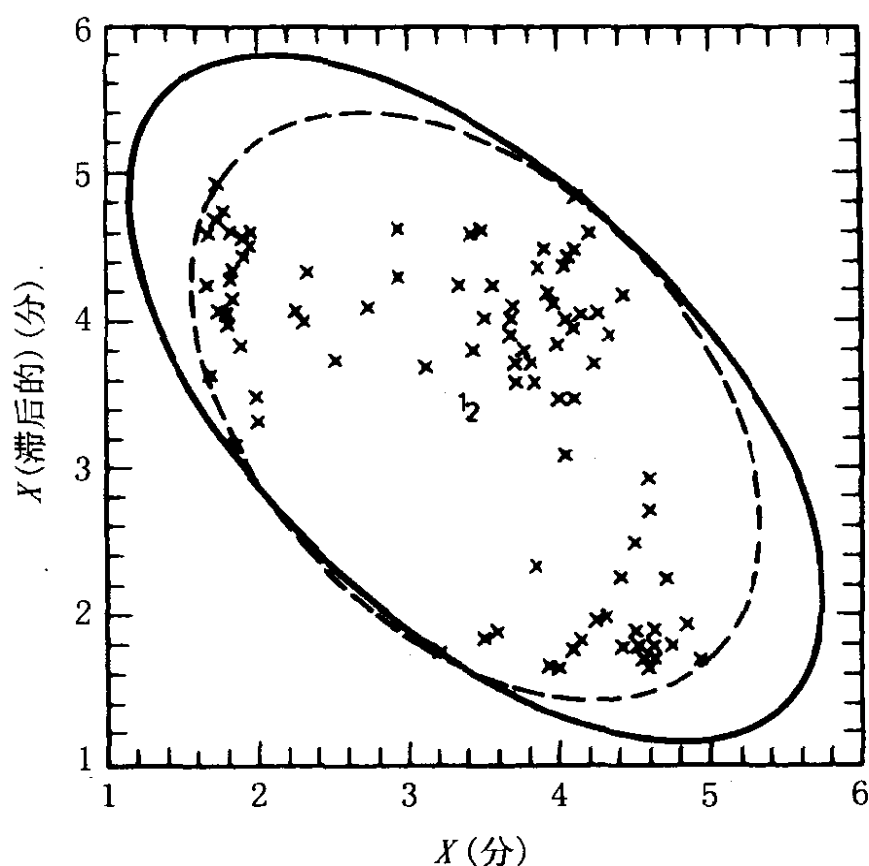


图 9.4 内插区域的近似

后的))与用于建立预测方程的那些相类似。这表示,有效的范围的一个可以接受的定义是,包含所有点的最小封闭图形。我们将考虑这一区域的两个近似,一个易于求得,另一个不太容易。

固定量 $h = \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}$ 的等高线是椭圆。如果 h_{\max} 是结构样本中最大的 h_{ii} , 则所有满足 $\mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_* \leq h_{\max}$ 的 \mathbf{x}_* 的点集是一个包含所有结构样本的椭圆。这一椭圆在图 9.4 中用实线作出。它在图上以点 1 为中心。对在点 \mathbf{x}_* 的预测, 如果 $h_* = \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_* > h_{\max}$, 则 \mathbf{x}_* 不在椭圆内, 预测值可被认为是一个外推值。

在图 9.4 中, 由实线椭圆定义的集合比包含观测数据的最小封闭图形要大得多。为给出一个更好的近似, 可以求包含这些点的最小容量椭圆。这称为最小覆盖椭圆或 *MCE*, 在图中由中心在点 2 的虚线椭圆给出。外推法的一种解释, 是将 *MCE* 外的任何一点称为外推值。Titterington (1978) 给出求 *MCE* 的一个算法。

9.3 附加评注

不同尺度的预测 一个预测方程可能产生一个变换尺度的预测值, 并且有时可能需要变换回原来的单位, 如例 6.1, 其中讨论了哺乳动物脑重与体重之间的关系, 求得方程

$$\log(\text{脑重}) = 0.9271 + 0.7517 \log(\text{体重})$$

它很好地表示了数据。如果某新的物种有平均 10 千克体重, 则预测的 $\log(\text{脑重})$ 为 $0.9271 + 0.7517 \log(10) = 1.6788$ 。假设取了对数后的误差为对称的, 就如在一个正态性假设之下一样, 给定体重为 10kg, 1.678 是 $\log(\text{脑重})$ 的预测分布的均值的估计。这样一个物种的脑重的一个自然的估计是其反对数, $10^{1.6788} = 47.73$ 克。不过, 宁愿说这是中位数的估计, 而不说这是脑重的预测分布的均值的估计。因为, 如果 $\log(\text{脑重})$ 的误差是对称的, 则脑重的误差是不对称的, 中位数能给出非对称分布中心的更好的概述, 所以这是一个合理的一个数值的概述。

简单地用新的尺度重新表达通过计算得到的置信断言的限,从而能方便地将置信及预测区间从一种尺度转换为另一种尺度。在脑重例子中,10 千克物种的 $\log(\text{脑重})$ 的一个 95% 的预测区间由集合 $1.0704 \leq \log(\text{脑重}) \leq 2.2872$ 给出。从而脑重的对应的 95% 的置信区间为 $10^{1.0704} \leq \text{脑重} \leq 10^{2.2872}$, 即在 11.8 和 193.7 克之间。这是一个很大的区间。另外,区间关于脑重的预测值 47.73 克不是对称的,这与预测分布的不对称性相对应。

给定假设,这一得到置信断言的方法总是正确的,真实的置信水平与表述的置信水平相等,并且从而在大多数情况是合适的。但是,在具有最短长度的意义上,区间可能不是最理想的,见 Land (1974)。

在完成任何对尺度的重新表述之前,仔细的分析者必须考虑新的尺度是否比旧的好。在脑重/体重例子中,用对数表述重量可能更为自然,如果按这一尺度,重量与其它感兴趣的量线性相关。

预测与模型选择 如第八章所表述的,在预测问题与模型选择之间有紧密的联系。本章所采用的方式是使用由回归分析得到的标准输出来得到预测函数。但以预测作为目标,也可以导出其它拟合过程。Copas (1983) 对这些预测及建模的可选方法给出一个有趣的非贝叶斯的方法。Picard 和 Cook (1984) 讨论了使用本书的方法来确认一个模型的问题。在更一般的条件下,对这些问题的一个贝叶斯的或预测的观点由 Aitchison 和 Dunsmore (1975), Geisser (1980) 和 Stone (1974) 给出。对预测的贝叶斯方法可能是很丰富的。给定自变量及先验数据和信息,我们可以将注意力集中在未测量的将来值的完整的预测分布上。这比只着眼于点预测能得到更多的信息。例如,如果一个预测模型被用于估计一种疾病的严重性,预测分布提供了估计病人的严重程度超过了某一值需要医生采取行动的的概率的工具。如果被估计的概率为足够高,则可以采取行动。

计算 使用大多数统计软件包可以相当容易地计算 *PRESS*

和相互证实平方和。例如，在 Minitab 中，方程 (8.23) 可以用于从 \hat{e}_i 和 h_{ii} 计算 $\hat{e}_{(i)}$ ，而方程 (8.24) 可用于计算 *PRESS*。其它程序，如 BMDP2R 或 BMDP9R，只要计算残差就自动计算 *PRESS*。

可以由许多方法得到一组新的数据的预测误差。一个有用的方法是将确认集合添加到结构集合，并对加权最小二乘法定义一系列数值，数值 1 用于在结构集合中的每个案例，权值 0 用于在确认集合中的每个案例。大多数程序在计算估计时将忽略权值为 0 的案例，但在计算残差和有关统计量时会用到它们。相互证实平方和由对权值为 0 的案例的残差计算得到。

问 题

9.1 在老忠实间歇泉数据，表 9.1 中，日对日的变化的任何可能影响被忽略。如何把可能的日对日的变化包括到模型中去？在什么情况下，日对日的变化使预测变为不可能？什么情况下预测是可能的，但具有更大的变化？考虑日对日效应，拟合一个模型并总结结论。

9.2 考虑在模型 (9.6) 中加入滞后-2 变量（即上一次喷发前的一次喷发的 x 和 y 的值）。这些变量有用吗？

9.3 (John Rice) 在心脏搭桥手术中，直径为 3 毫米的一根聚四氟乙烯试管（或导管）被通过大腿部分的一根主要静脉或动脉通入心脏。导管可被操纵进入特定区域以提供有关生理学及心脏功能的信息。这一过程有时用于有先天性心脏病的儿童，医生必须猜测导管的合适长度。

在一个对 12 个病人的小研究中，导管的合适长度 Y 是通过查看荧光屏检查（X-光）导管尖是否到达主动脉瓣膜来决定。记录了每个病人的身高 X_1 （英尺）及体重 X_2 （磅）以用来检验它们对预测导管长度（厘米）是否有帮助。下面给出数据。

单独用 X_1 ，单独用 X_2 以及同时使用 X_1 ， X_2 构造预测方程。如果预测的 ± 2 厘米的误差是可以容许的，是否有哪个预测方程是合适的？你必须决定，作为一个概率表述，“ ± 2 厘米”的含义。总结你的结论。

Y	X_1	X_2
37	42.8	40.0
50	63.5	93.5
34	37.5	35.5
36	39.5	30.0
43	45.5	52.0
28	38.5	17.0
37	43.0	38.5
20	22.5	8.5
34	37.0	33.0
30	23.5	9.5
38	33.0	21.0
47	58.0	79.0

9.4 70年代中期开始,法律及其它职业学校的申请入学人数比可接受人数要多许多。以致使在申请入学学生之间的公平选择的方法变得非常重要。应用如下的回归思想。首先对目前在校注册了的学生收集数据。所采用的典型度量为大学期间等级点平均 (X_1),一场标准测试,如法律学校能力测试的得分 (X_2),以及表现的一个度量,如第一年等级点平均 (Y)。然后,估计得到一个形如 $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ 的预测方程(尽管估计可能不能由最小二乘法得到,见 Rubin, 1980)。在申请者集合中可以得到 X_1 和 X_2 ,故对每位申请者可以计算 \hat{Y} 的值,具有大的 \hat{Y} 值的申请者被录取(尽管不是所有的学校单用 \hat{Y} 来决定录取)。

用这一方法讨论问题。特别地,将结构总体与目标总体进行比较。假设对当前学生的 X_1 和 X_2 之间的相关变为负的(如它偶尔所为那样)。解释这是怎么发生的。如果有意义的话,这样一个相关意味着什么?(亦见 Aitken, 1934 和 Lawley, 1943)。

9.5 一幢房子的财产税假设依赖于目前市场上的房价。由于房子实际的销售较少,当财产税设定时,每幢房子的销售价格每年必须进行估计。回归函数有时用于得到一个预测函数(Renshaw, 1958)。

表 9.5 数据为宾夕法尼亚的伊利市, $n=27$ 幢确实被售出的房子的数据(Narula 和 Wellington, 1977)。变量为

X_1 = 当前税 (地方, 学校及郡) $\div 100$ (美元)

X_2 = 浴室数目

X_3 = 空地大小 $\div 1000$ (平方英尺)

X_4 = 起居室大小 $\div 1000$ (平方英尺)

X_5 = 车库数

X_6 = 房间数

X_7 = 卧室数

X_8 = 房子年龄 (年)

X_9 = 壁炉数

Y = 确实销售价格 $\div 1000$ (美元)

用数据估计一个函数, 用于由这些 X 及它们的函数预测 Y 。(在实际中, 用于估计售价的数据集合比这里所用的数据集合要大得多, 并且包括街道地区指标的其它的变量, 如学校质量等。)

表 9.5 房子数据

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
4.9176	1.0	3.4720	.9980	1.0	7	4	42	0	25.9
5.0208	1.0	3.5310	1.5000	2.0	7	4	62	0	29.5
4.5429	1.0	2.2750	1.1750	1.0	6	3	40	0	27.9
4.5573	1.0	4.0500	1.2320	1.0	6	3	54	0	25.9
5.0597	1.0	4.4550	1.1210	1.0	6	3	42	0	29.9
3.8910	1.0	4.4550	.9880	1.0	6	3	56	0	29.9
5.8980	1.0	5.8500	1.2400	1.0	7	3	51	1	30.9
5.6039	1.0	9.5200	1.5010	0	6	3	32	0	28.9
15.4202	2.5	9.8000	3.4200	2.0	10	5	42	1	84.9
14.4598	2.5	12.8000	3.0000	2.0	9	5	14	1	82.9
5.8282	1.0	6.4350	1.2250	2.0	6	3	32	0	35.9
5.3003	1.0	4.9883	1.5520	1.0	6	3	30	0	31.5
6.2712	1.0	5.5200	.9750	1.0	5	2	30	0	31.0
5.9592	1.0	6.6660	1.1210	2.0	6	3	32	0	30.9
5.0500	1.0	5.0000	1.0200	0	5	2	46	1	30.0

(续表)

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
8.2464	1.5	5.1500	1.6640	2.0	8	4	50	0	36.9
6.6969	1.5	6.9020	1.4880	1.5	7	3	22	1	41.9
7.7841	1.5	7.1020	1.3760	1.0	6	3	17	0	40.5
9.0384	1.0	7.8000	1.5000	1.5	7	3	23	0	43.9
5.9894	1.0	5.5200	1.2560	2.0	6	3	40	1	37.5
7.5422	1.5	4.0000	1.6900	1.0	6	3	22	0	37.9
8.7951	1.5	9.8900	1.8200	2.0	8	4	50	1	44.5
6.0931	1.5	6.7265	1.6520	1.0	6	3	44	0	37.9
8.3607	1.5	9.1500	1.7770	2.0	8	4	48	1	38.9
8.1400	1.0	8.0000	1.5040	2.0	7	3	3	0	36.9
9.1416	1.5	7.3262	1.8310	1.5	8	4	31	0	45.8
12.0000	1.5	5.0000	1.2000	2.0	6	3	30	1	41.0

来源: Narula and Wellington (1977)。

注: 表的原始数据中, 一个案例被意外地给出二次, 这里将这个案例删除了。

10

不完全数据

在许多数据集合中,某些案例的某些变量可能没有记录下来。事实上,在大型研究中,完整的数据往往是例外,而不是惯例。由于分析的标准方法只能直接应用于完整数据集,因此需要有附加的技巧。所需要的最常用的技巧为修改数据集,或者是删除部分被观测的案例或变量,或者是对未被观测的值填入猜测值。然后进行通常的分析。在必要时进行调整,以适应数据中所作的修改。此外,也有对不完全数据不填入或删除的分析方法。使用这些方法一般依赖于强的假设。如我们将会看到的,这两个一般的方法没有一个是完全令人满意的。

关于不完全数据问题的统计文献是很多的。这里给出的处理不是综合性的。关于这一题目的书目提要由 Afifi 和 Elashoff (1966) 给出。本章所引用的是更近期的论文。

10.1 随机遗漏

分析不完全数据的大多数方法使用假设:值没有被观测的原因与所研究的关系是无关的。例如,数据没有被观测是因为遗失测试的试管或遗失编码纸。这一般能满足假设。在与这些类似的

条件下,忽略引起观测值失败的原因的分析是可以的。另一方面,如果没有观测值的原因依赖于将被观测得到的值,则对数据的分析必须包括对数据没有观测到的原因建立模型。

Rubin (1976) 对两种类型的不完全数据作了精确的区分。对这一讨论,如果未被观测的一个值不依赖于将被观测得到的值,则一个不完全数据集有随机遗漏 (*MAR*) 的值。决定一个 *MAR* 的假设,对一个特定的数据集是否合适,是分析不完全数据的重要一步。以下的例子将阐明定义的应用。

掷飞镖 一个数据矩阵的每一个数被想象附着于一块飞镖板上,而个数是随机的若干根镖被掷到板上。每个被飞镖击中的数将丢失。在这个例子中,每个观测值丢失的概率不依赖于已观测到的值。这个模型的一个用途是用于键控穿孔误差,数据明显为 *MAR* 的。

遗漏预先测验 在一个教育成果的研究中,教师忘了举行几个预先测验中的一个测验,则遗失的预先测验的得分是随机遗漏的。

随机化实验 在比较处理与控制(未处理)的一个实验中,每个研究单元被随机地分到一组或另一组中。所有被分到处理组的单元遗漏了倘若它们被分到控制组而不是处理组将能得到的得分。它们的控制得分被随机遗漏。在这个意义上,所有随机实验的数据遗漏是随机的。因而在对数据分析时,忽略引起数据遗漏的原因,即随机性,是可以允许的。

火鸡生嫩度 在一个研究根据生嫩度对火鸡进行分级的方法的实验中,一个样本有 $n=17$ 只火鸡, X = 生嫩度的经验估计。 X 为一个从 1 (非常嫩) 到 5 (非常老) 的一个得分。然后火鸡被退毛、冰冻,并且在固定的一段时间后被解冻并烧好。然后得到对 Y = 实际生嫩度的一个实验室度量。但是在贮藏中,3 只等级标为 1 的火鸡被窃,故它们的 Y 值无法被记录。因为引起遗漏的原因可能是 X 的一个函数,但不是 Y 的一个函数,尽管 X 和 Y 被假设是

相关的，仍认为响应变量 Y 是 MAR 。

以下是非 MAR 的例子。

删截 假设研究中的一个变量为到达失效的时间，但某些单元在试验期间没有失效，则它们的失效时间没有记录。在这里随机遗漏明显是不合适的。

小的反应 如果一种化学物质的浓度在某些单元中的真实值小于所用仪器能够测量的最小值，则在这些单元中浓度可能是不可观测的。因此观测失效依赖于倘若能被观测将会得到的值，所以 MAR 不成立。

以下是一个 MAR 假设可疑的例子

受试验者退出试验 在一个试验开始时，受试验者被随机地分为若干组。在每一组内，应用一个不同的处理方法。在试验结束时，可能是数周之后，要得到处理后的分数，但某些受试验者不再有用。对这些受试验者响应变量遗漏。

如果受试验者退出试验的原因与遗漏变量无关，那么 MAR 是合理的。受试验者若是人，他的离开一般使数据为 MAR 。另一方面，受试验者退出是因为他们在试验中表现不佳，则会违背 MAR 假设。

MAR 和非 MAR 的区别的重要性在于：如果观测数据不是 MAR ，则任何忽略引起数据遗漏的原因的推断将有严重的误差。对非 MAR 数据，一个可靠的分析需要建立模型以说明没有被观测的数据。对非 MAR 问题的一个系统处理的主要例子为对删截和生存数据的研究。对此可以得到象书一样长的处理 (Kalbfleisch 和 Prentice, 1980, 及 Cox 和 Oakes, 1984)。当数据为 MAR 时，引起数据遗漏的过程可以被忽略。这里描述的技术和方法将是有益的。

对产生数据的过程进行仔细考虑是研究 MAR 假设的最好的诊断工具，这在前面的各个例子中已经说明。

10.2 通过填入和删除来处理不完全数据

对有几个观测值遗漏的数据的最简单的分析方法是删除案例、变量或两者，从而得到完全的数据集合。只要 MAR 的假设成立，由于通过删除得到的数据集合能代表全部，所以通过一些调整，通常的方法可被用于完全数据。当然，如果 MAR 假设是有问题的，则由通过删除得到的数据集合所作的推断也是值得怀疑的。

如果只有几个案例有未被观测的数据，因为需要最少的假设，故案例删除是最有吸引力的可用方法。遗漏响应变量或所有自变量的案例必须被删除。有若干未被观测自变量的案例也很有可能要被删除，因为它们经常包含很少的有用信息。

变量删除更为复杂。如果考虑的线性模型已知完全正确，则对部分地被观测的变量的删除可能难以证明其合理性，因为这可能导致不正确的模型。由于大部分线性模型的使用只是作为一个近似，则对部分地被观测的自变量的删除可能仍是有用的。如果一组完全观测的自变量的组合与一个部分观测的自变量高度相关，则这些完全观测的自变量可用于取代这个部分观测的自变量。这里，共线性对研究者是有利的。

填入的方法 填入遗漏数据可以得到与删除大致相同的结果：导致一个完整的数据集合，并且通过某些修改，可以使用通常的估计方法（检验和置信断言不是这么清楚）。问题是决定填入遗漏数据的值。对这一问题的一个谨慎的方法是使用所有可以得到的信息，对遗漏数据给出一个可信的值，填入它们的各种组合，并试图控制遗漏数据对参数估计以及模型的建立的影响。

如果可以得到数据外部的关于遗漏值的附加信息，它可以用于帮助选择填入的数据。例如，对按时间次序收集的数据，遗漏值可以从紧挨遗漏值的前一个和后一个变量的观测值可靠地估计得到。在“伯克来指导学习”（问题 2.1），一个对儿童生长的纵向

研究中,得到对每个儿童以 $\frac{1}{2}$ 年为间隔直到18岁的身高和体重的观测值。如果一个儿童在8岁时的体重数据丢失,可以通过将那个儿童在 $7\frac{1}{2}$ 岁和 $8\frac{1}{2}$ 岁时测得的体重求平均来获得估计。在其它问题里,试验中类似的考虑可能严重地将可信值限制到一个相对小的集合。

缺少直接对要估计的值有影响的附加信息时,完全观测的数据可用于获得关于部分观测的数据的一个估计方程。作为一个简单的例子,考虑两个自变量 X_1 和 X_2 。两者偶尔都有遗漏。对完全观测的数据,假设 X_1 关于 X_2 的散点图如图10.1给出。事实上,这一图形是很理想的,因为它通过使 (X_1, X_2) 配对为二元正态而生成的。点云集成的图形多少象是椭圆,且 X_2 关于 X_1 的回归明显是线性的。我们感兴趣的一个问题是,对 X_1 被观测, X_2 未被观测的案例得到填入 X_2 的值。我们可以由完全数据计算 X_2 关于 X_1 的回归,并用这一拟合方程对 X_1 被观测而 X_2 未被观测的案例来估计 X_2 。这相当于填入案例使它们沿图10.1所画的实线列出。该实线为 X_2 关于 X_1 的回归。类似地,为填入 X_1 的值,填入的值将沿虚线列出。虚线表示 X_1 关于 X_2 的回归。这一过程基于这样一个假设:填入的案例与哪些完全的案例是类似的,故对后者拟合的一个模型给出关于前者的信息。

由于这一方法试图使填入的案例尽可能接近数据集的中心,这使得这些案例对数据分析的结果影响较小。如果单个案例中的几个变量通过回归填入,这一案例的影响事实上可能很小。例如,在一个四个变量问题中,对完全被观测的案例, X_1, X_2, X_3 中的每一个都可以被 X_4 回归。从而对于 X_4 被观测,但 X_1, X_2 和 X_3 未被观测的案例,可以估计填入值。如果事实上这一案例 X_j 的真实值远离数据中心,且这一案例是有影响的,则通过回归填入将遗失这一信息。

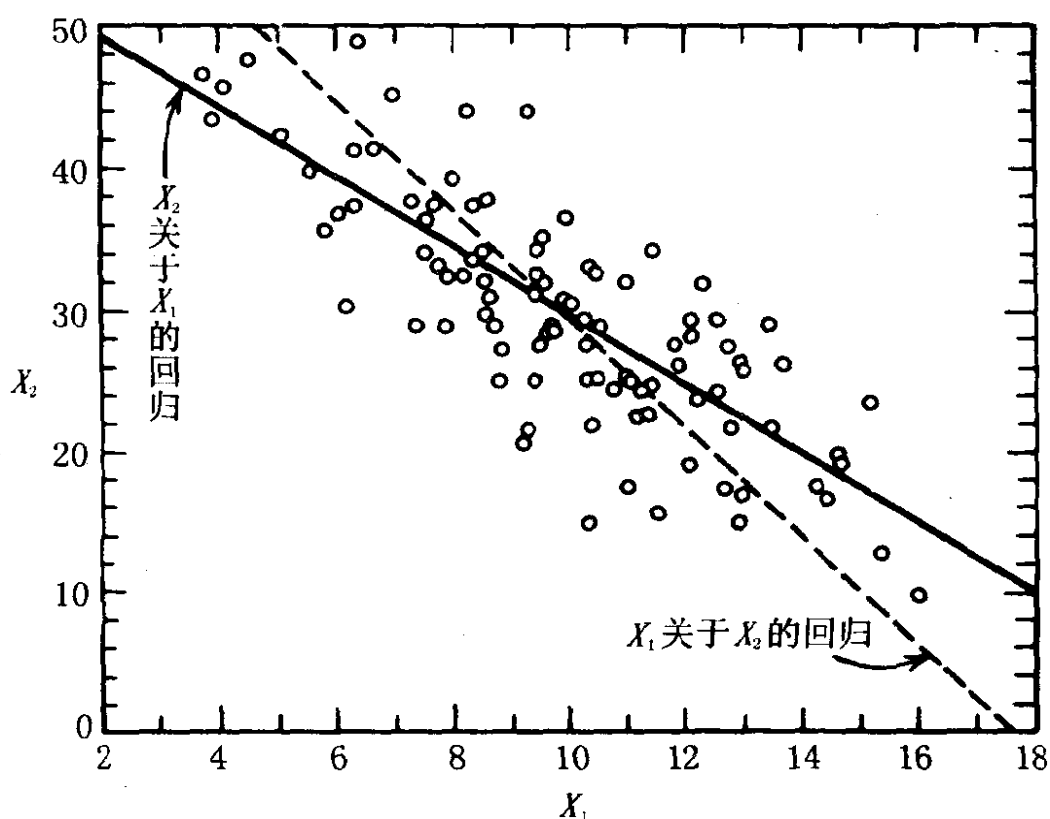


图 10.1 通过回归填入

通过回归来填入有许多变化。通过使用本书导出的方法可以建立一个自变量作为其它自变量的函数的模型。这可能包括子集选择，尺度变换等等。也可使用更简单的方法，如使用只有一个或两个变量的模型来填入。

调整含填入值的分析 如果一个数据集合用任何前面描述的方法填入遗漏值，则可以进行一个通常的分析，只是误差的自由度对每个填入值必须减 1。自然地，这将填入值的个数限制为 $n - p'$ ，即误差的自由度。这是相当合理的，因为 $n - p'$ 个值可能使 $RSS = 0$ 。依赖于填入数据的性质，检验和置信断言可能有严重误差。任何关于估计的最优性的断言，如无偏性或最小方差，一般不能作出。这部分是因为填入值“太好”，而没有反映出真实数据中明显的通常有的变异性。另外，我们对应用于填入数据的诊断过程的情况知道得很少。

10.3 正态性假设下的极大似然估计

如果我们能够假设所有数据，包括被观测和未被观测的为取自一个多元分布的样本，一般是正态的，则可以利用所有数据计算参数估计。尽管这些方法在计算上要比线性最小二乘法更复杂，只要多元正态的假设有意义，极大似然估计的方法仍是有吸引力的。两种不同的计算方法被提出。Hartley 和 Hocking (1971) 提出一种使用标准递归过程的算法，称为划线法。Dempster, Larid 和 Rubin (1977) 主要针对不完全数据的情况，给出一种方法，称为 EM 算法。他们的方法是继 Buck (1960), Orchard 和 Woodbury (1972) 及 Beale 和 Little (1975) 的早期工作之后的。Rubin (1974) 描述了可以用非递归计算的条件。

尽管对不完全数据求极大似然估计的方法已有若干年了，但由于多种原因，它们的使用受到限制。首先，多元正态性的假设不总是成立的，尽管 Little (1979) 给出，在完全被观测的自变量的分布不是正态分布时如何修改假设。其次，较常用的 EM 算法没有给出对系数的标准误差的估计（见 Louis, 1982，一种估计标准误差的方法）。第三，关于对小至中等样本的检验和推断，我们知道得很少。最后，对不完全数据问题的诊断过程未被仔细研究过。不过，暂且不论这些困难，使用不完全数据的极大似然估计的方法仍可能是有用的。

10.4 遗漏观测值相关

综合分析的所有的计算依赖于少数几个主要统计量，即样本均值、方差和协方差。处理遗漏数据的一个常用方法是使用所有的观测数据计算这每一个样本统计量，然后用最小二乘分析，仿佛它们是由完全数据计算得到的。这一方法常被称为遗漏观测值

相关的方法。

例如,假设一个数据集合由如下所示的四个变量的 $n=8$ 个案例组成。

案例编号	Y	X_1	X_2	X_3
1	X	X	X	
2	X	X	X	X
3	X	X		
4	X	X		X
5	X	X		
6	X	X		X
7	X		X	X
8	X		X	

注: X=被观测的, 空格=遗漏的。

用前面 6 个案例可以计算 Y 与 X_1 的相关系数, 基于案例 2, 4, 6 和 7, 可以计算 Y 与 X_3 的相关系数, 等等。对这些数据, 多至 6 个案例, 少至 2 个案例 (如 X_2 , X_3 之间的相关系数) 将被用于估计一个相关。

这一方法的成功依赖于假设: 数据为总体的一个样本, 故在遗漏观测值的相关矩阵中的每一个相关是总体相关的一个估计。只要所有的相关是小的, 并且样本容量不是太小, 这一方法的结果可以是合理的。然而, 如果出现大的相关, 可能引起严重问题, 包括计算得到负的平方和, 或 R^2 大于 1。其原因是, 当每个相关基于不同案例时, 计算得到的相关矩阵可能不是正定的。因此, 必须避免按常规使用遗漏观测值的相关矩阵。如果用它, 则对检验、估计和预测的解释几乎是不可能的, 并且不能给出有意义的案例分析。

10.5 一般推荐

由于并不总是能避免遗漏数据，因而需要有处理不完全数据集合的某些一般的准则。首先，研究遗漏数据的结构。这常可以用一个特别的程序，如 BMDP 系列中的 BMDPAM (Dixon, 1983) 来完成。通过决定哪些变量是部分被观测的，哪些案例有许多遗漏数据，以及遗漏数据的整体结构，可以得到许多有用信息。

如果随机遗漏的假设是合理的，则可以作出删除案例或变量的决定。对于删除后剩余的数据，可以应用填入的方法。或者，可能更好地，应用并比较多种方法。

我们总结得到一个警告：不要让计算机程序决定如何处理遗漏数据。许多大规模的软件包可以处理遗漏数据。它们常是使用这里描述的一种或多种方法，或是这些方法的变形。记住，程序的编写者并不知道你的问题的细节。程序中使用的缺省分析。对你的问题几乎是一定不合适的。

11

非最小二乘估计

本书几乎通篇使用的最小二乘估计不是能用于线性回归问题的唯一估计。之所以使用是因为它们易于计算，几何上漂亮，并且给出非常严格的假设，在若干重要意义上，它们是最优的。最小二乘估计的一个主要合格证明是它已被成功地使用了 150 多年。

近来，特别是可以廉价而高速地进行计算以后，提出了许多很有竞争力的估计。它们主要是针对最小二乘估计中发现的不足的。例如，我们在第 5 章中看到，数据集合的单个异常值对检验以及系数的最小二乘估计有相当大的影响。因此，我们可能希望有另一种估计，它对出现的异常值不要太敏感，或能更“稳健”。出于对偶尔出现的“坏的”数据的考虑，产生了稳健的回归方法，在 11.1 节中有简要的讨论。

我们发现的最小二乘法的另一个缺点是其在出现共线性时的表现。11.2 节讨论的岭回归及其同类方法，是用于对相关的自变量给出更好的估计。目前这些特别的估计似乎并不流行，因为它们并不比最小二乘法有太多的优越性。

11.1 稳健回归

好的统计过程，即使在基本假设略微有些误差时，仍应该工作良好。这关系到稳健性——一个首先被 Box 在这一场合下使用的术语（1953）——为近来统计研究中的一个重要组成部分。在第五和第六章讨论的诊断方法代表解决稳健性问题的一种途径。诊断过程被设计为，扰动假设模型、数据或假设的某些方面，并估计它对导致的结论的影响。例如，Cook 距离度量，通过删除一个案例来扰动数据产生的影响。对非常数方差的得分检验是一个诊断，它给出关于常数方差的假设的信息。

稳健估计代表在统计中稳健性思想的另一个应用。与其使用诊断来监视扰动的影响，我们不如求一种估计，即使有扰动也会工作良好。稳健估计的一个经典例子是基于一个随机样本 y_1, \dots, y_n 估计对称分布的中心的问题。通常的估计，样本均值 \bar{y} ，众所周知为非稳健的，因为一个非常大的 y_i 几乎能决定这个估计的取值。另一方面， y_i 的样本中位数对一些大的观测值是不敏感的，因而是稳健的。Andrews et al. (1972) 给出对称分布中心的各种估计的性质的一个大型研究，其中包括的不止均值与中位数。

稳健估计中重要的一类，由 P. Huber 提出，称为 M -估计，它是极大似然类型估计的一个缩写。它们是通过选择一个估计使残差的一个函数，非残差平方和，达到最小而得到的。对单个样本，令 μ 是对称分布的中心，并令 $\tilde{\mu}$ 为其估计。然后选择 $\tilde{\mu}$ ，它是 μ 的使函数

$$\sum_{i=1}^n \rho(y_i - \mu) / \sigma \quad (11.1)$$

达到最小的值，其中 ρ 是被指定的一个函数， σ 是一个标量因子。如果 $\rho(z) = z^2$ ，则 (11.1) 是通常的最小二乘准则。另一个普遍的选择的 $\rho(z) = |z|^f$ ，其中 f 常取为小于 2 的一个正常数。选

择 $f=1$, 给出最小绝对偏差估计。它给出单个样本的中位数。较小的 f 值试图对大的残差给出更小的权值。另一个重要的选择为

$$\rho(z) = \begin{cases} z^2/2 & , \quad \text{如果 } |z| \leq c \\ c|z| - c^2/2 & , \quad \text{如果 } |z| > c \end{cases} \quad (11.2)$$

其中 c 是一个固定常数。对 ρ 的这一选择导致有大的残差的案例权值降低。在单个样本案例中, 这等同于对极端的 y_i 给以较少的注意力。Huber (1981) 对这些估计提供了一个完整的讨论, 包括 c 的选择, 估计 σ^2 的方法, 计算方面的情况以及性质。

对线性回归问题, M -估计用 (11.1) 计算, 唯一不同的是用 $x_i^T \beta$ 代替 μ 。此外, 所有的方法都是相同的。

M -估计被设计使得对一个数据集中的异常值、非通常响应变量是稳健的。它们并非被设计为对在回归分析中所作的任何其它假设都是不敏感的。例如, 如果自变量的尺度 (如对数或平方根) 的正确性是值得怀疑的, 则 M -估计并不比最小二乘法好。另外, M -估计对高位势值, 或大的 h_{ii} 的案例的敏感性, 与最小二乘法相同。与回归分析中的许多其它问题相比较, 对响应变量中异常值不敏感的需要似乎并不重要。因此, 回归中 M -估计的作用似乎是有限的。

最为普遍接受的稳健回归的一个用处是作为最小二乘估计的一个检验。许多工作者提出同时使用最小二乘以及稳健分析。如果它们相一致, 则我们多少能相信最小二乘分析。如果它们不一致, 则试图找出原因。这里, 稳健估计被提倡作为诊断。

在最近这几年中关于稳健方法的文献出现很多。至此尚未摘录的一些重要参考文献, 包括 Andrews (1974), Mosteller 和 Tukey (1977), Devlin, Gnanadesikan 和 Kettenring (1975) 以及 Krasker 和 Welsch (1983)

11.2 有偏回归

寻求代替最小二乘结论的第二个途径放宽寻求无偏估计的必要性。考虑以样本均值偏差形式给出的线性模型，

$$Y = \mathbf{1}\beta_0 + \mathcal{X}\beta + e \quad (11.3)$$

其中 \mathcal{X} 为 $n \times p$, β 为 $p \times 1$, 不包括截距项, 且 $\text{var}(e) = \sigma^2 \cdot \mathbf{I}$ 。最小二乘估计为 $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T Y$ 。如第二章中所提到的, $\hat{\beta}$ 是最小方差无偏估计, $\text{var}(\hat{\beta}) = \sigma^2 (\mathcal{X}^T \mathcal{X})^{-1}$ 。

假如我们扩大所考虑的估计类, 包括有偏估计, 并且考虑将 β_j 的估计的均方误差和作为一个准则函数, 记为 $SMSE$, 其中

$$\begin{aligned} SMSE &= \sum_{j=1}^p E(\hat{\beta}_j - \beta_j)^2 \\ &= \sum_{j=1}^p \{\text{var}(\hat{\beta}_j) + [\text{bias}(\hat{\beta}_j)]^2\} \\ &= E(\hat{\beta} - \beta)^T (\hat{\beta} - \beta) \end{aligned} \quad (11.4)$$

这一准则函数与导致最小二乘估计的准则是不同的, 所以按这一准则的其它估计可能更好也就不值得惊讶了。在描述这些估计量之前, 仔细地研究一下 (11.4) 是有用的。其重要的特征为: (1) 所有 β_j 均等地加入, 这意味着对它们的每一个, 我们的兴趣是相同的; (2) $SMSE$ 不是对尺度不变的, 所以 X 的尺度选择是关键性的; (3) 兴趣集中于对参数的估计, 而非回归分析的其它方面; (4) 估计间的协方差可以忽略。

为处理缺乏不变性, \mathcal{X} 的每一列通常被标准化, 使 $\mathcal{X}^T \mathcal{X}$ 为样本相关矩阵 (Marquardt 和 Snee, 1975; Obenchain, 1975)。作为进一步的标记, 令 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 为排好序的 $\mathcal{X}^T \mathcal{X}$ 的特征值。然后, Hoerl 和 Kennard (1970a) 证明了, 对 $\hat{\beta}$ = 最小二乘估计,

$$SMSE = E(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)$$

$$= \sigma^2 \text{trace}(\mathcal{X}^T \mathcal{X})^{-1} = \sigma^2 \sum_{j=1}^p \lambda_j^{-1} \quad (11.5)$$

但是 $E(\hat{\beta} - \beta)^T(\hat{\beta} - \beta) = E(\hat{\beta}^T \hat{\beta}) - \beta^T \beta$ 。将此代入(11.5),

$$E(\hat{\beta}^T \hat{\beta}) = \beta^T \beta + \sigma^2 \sum_{j=1}^p \lambda_j^{-1} \geq \beta^T \beta + \sigma^2 \lambda_p^{-1} \quad (11.6)$$

这样,即使 $\hat{\beta}$ 对 β 是无偏的, $\hat{\beta}^T \hat{\beta}$ 对 $\beta^T \beta$ 不是无偏的,并且如果最小特征值 λ_p 接近 0,则平均而言, $\hat{\beta}^T \hat{\beta}$ 将会太大。我们回想到 λ_p 接近 0 是共线性的征兆。当 λ_p 较小,并且我们对(11.4)感兴趣,则可能得到相对于最小二乘的实质性收获。

这里考虑的大多数可选估计有一个共同特征:它们将给出关于 β 的一个估计 $\tilde{\beta}$, 它比最小二乘估计更短 ($\tilde{\beta}^T \tilde{\beta} < \hat{\beta}^T \hat{\beta}$)。故这些技术将压缩最小二乘估计。一般是朝原点 $\mathbf{0}$ 的压缩。我们已经遇到过一次这样的压缩:选择子集,其中足够多的系数置于 $\mathbf{0}$,使得用于(11.6)的最小特征值相对较大。

在转向其它具体的估计前,用标准形式表述初始问题(11.3)是有用的,其中 \mathcal{X} 的列被 p 个正交变量(主成分)所代替。由 7.6 节,有一个 $p \times p$ 正交矩阵,其列为 $\mathcal{X}^T \mathcal{X}$ 的特征向量,记为 U ($UU^T = U^T U = I$) 以及一个 $p \times p$ 对角矩阵 D , 其对角线元素 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, 使得

$$\mathcal{X}^T \mathcal{X} = U D U^T \quad (11.7)$$

令 $Z = \mathcal{X}U$ (故 z 的列为 \mathcal{X} 的主成分), 和 $\alpha = U^T \beta$, 则

$$\begin{aligned} Y &= \mathbf{1}\beta_0 + \mathcal{X}\beta + e \\ &= \mathbf{1}\beta_0 + \mathcal{X}(UU^T)\beta + e \\ &= \mathbf{1}\beta_0 + Z\alpha + e \end{aligned} \quad (11.8)$$

且模型(11.8)和(11.3)等价。 β 和 α 的估计通过方程

$$(\alpha \text{ 的估计}) = U^T (\beta \text{ 的估计}) \quad (11.9)$$

相联系。故或者 $\hat{\beta}$, 或者 $\hat{\alpha}$ 能被计算。然而,由于 $Z^T Z = D$, $\text{var}(\hat{\alpha}) = \sigma^2 (Z^T Z)^{-1} = \sigma^2 D^{-1}$, 故 $\hat{\alpha}_j$ 互相独立。另外, $\text{var}(\hat{\alpha}_p) = \sigma^2 \lambda_p^{-1}$, 这样 $\hat{\alpha}_p$ 比其它任何可能的估计有更大的方差——在数据中关于 Z 的第 p 列的信息比其它任何为原始 X 的线性组合的变量的信

息要少。

岭回归 岭回归估计 $\tilde{\beta}(RR)$ 由 Hoerl 和 Kennard (1970a; 1970b) 定义, 对某个 $k \geq 0$, 有

$$\tilde{\beta}(RR) = (\mathcal{X}^T \mathcal{X} + k \cdot I)^{-1} \mathcal{X}^T Y \quad (11.10)$$

如果 $k = 0$, $\tilde{\beta}(RR) = \hat{\beta}$, 即最小二乘估计, 而较大的 k 将把 $\tilde{\beta}(RR)$ 从最小二乘估计拉开, 并且增加估计的偏差。岭参数 k 是无穷多个可能的岭估计的索引。

图 11.1 表示增加 k 对岭估计的影响。用于产生这一图形的数据也用于图 4.3。在那里, 给出了基于最小二乘的 (β_1, β_2) 的 95% 的置信椭圆。岭估计由 (11.10) 计算得到。曲线给出当 k 从 0 增加时的 $\tilde{\beta}(RR)$ 的图。在 $k = 0$, $\tilde{\beta}(RR) = \hat{\beta}$ 。当 $k \cong 0.6$, $\tilde{\beta}(RR)$ 落至 95% 置信椭圆的边缘。对任何 $k < 0.6$, $\tilde{\beta}(RR)$ 在椭圆内, 而大的值将 $\tilde{\beta}(RR)$ 置于椭圆之外。

一般地, k 是一个未知的调整常数, 可以由分析者设定。大的 k 值对应于增加的偏差但方差更小, 故 k 值的选择必须能平衡偏差与方差。岭回归的基本好处由 Hoerl 和 Kennard 以下面的结论给出了总结 (1970a): 对模型 (11.3) 中每个固定的 \mathcal{X} 和 β , 存在一个 k_0 , 使得对所有 $0 < k < k_0$, $\tilde{\beta}(RR)$ 的 SMSE 小于 $\hat{\beta}$ 的 SMSE。然而, Thisted (1978b) 证明了, 对任何固定的 $k > 0$ 及任何 \mathcal{X} , 在一个回归问题 (即, β 的一个真实值), 使得 $\hat{\beta}$ 的 SMSE 小于 $\tilde{\beta}(RR)$ 的 SMSE。这样如果使用岭回归的目的试图使 SMSE 最小, 必须从数据中估计 k 。估计 k 的方法由许多文献给出, 包括 Hoerl 和 Kennard (1970a)。Draper 和 Van Nostrand (1978) 给出对提出的方法的一个概括的研究。

标准形式 在标准形式中, α_j 的岭估计 $\tilde{\alpha}_j(RR)$ 可以证明是等于

$$\tilde{\alpha}_j(RR) = \frac{\lambda_j}{\lambda_j + k} \cdot \hat{\alpha}_j \quad (j=1, 2) \quad (11.11)$$

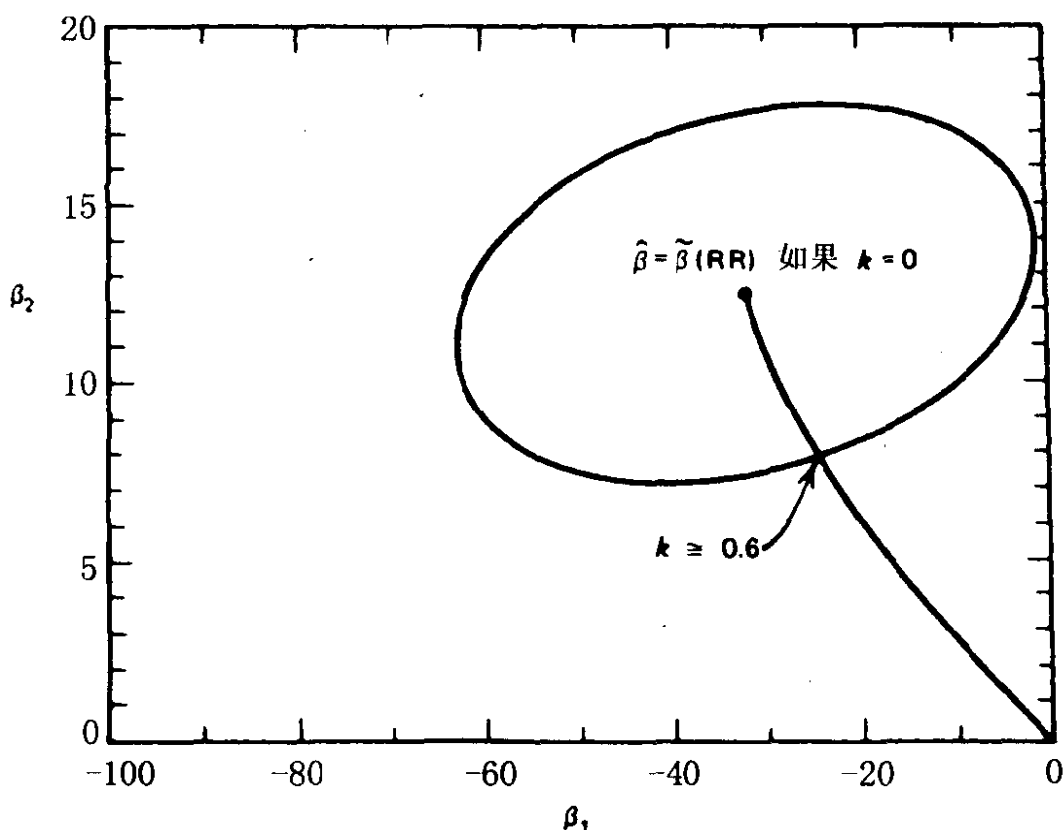


图 11.1 岭估计的路径 (基于图 4.3)

其中 $\hat{\alpha}_j$ 是最小二乘估计。如果 λ_j 比 k 大得多, 则岭回归的作用是使最小二乘估计几乎不变。但如果 λ_j 比 k 小, 则对应的 $\tilde{\alpha}_j$ (RR) 将比 $\hat{\alpha}_j$ 小得多。

与贝叶斯法则的关系 通过使用贝叶斯方法, 遵照一定格式把关于参数的先验信息吸收入问题中。如果我们假设 (11.8) 中的向量 α 取自一个正态分布

$$\alpha \sim N(0, k^{-1}I) \quad (11.12)$$

并且我们又假设 $e \sim N(0, \sigma^2 I)$, 则 (11.3) 中的 β 的贝叶斯估计由 (11.10) 给出。值 k^{-1} 是每个 α_j 的方差, 代表 α 的先验变异性。如果所作的假设合理, 估计 (11.10) 是很有吸引力的。然而, 先验均值为 0 的假设很少是合理的 (当然, 对其它先验均值可以修改方法)。倘若它成立的话, 数据的收集与分析大概永远不会完成。

使用岭回归的例子由 Marquardt 和 Snee (1975) 给出, 另外,

可参见 Smith 和 Campbell (1980)

广义岭回归 岭回归法则的一个推广是用一个参数向量 (k_1, k_2, \dots, k_p) 代替岭参数 k 。在标准公式中, 广义岭估计 $\tilde{\alpha}_j(GR)$ 为

$$\tilde{\alpha}_j(GR) = \frac{\lambda_j}{\lambda_j + k_j} \cdot \hat{\alpha}_j \quad (j=1, 2, \dots, p) \quad (11.13)$$

所以广义岭回归法则考虑列收缩估计的不同方式, 对每个 λ_j 使用不同的岭参数。利用原始坐标系, 定义 $p \times p$ 矩阵 G , $G = UKU^T$, 其中 K 是对角矩阵, 对角线上元素为 k_1, k_2, \dots, k_p , 在 (11.8) 邻近处有 U 的定义。则广义岭估计为 (Hoerl 和 Kennard, 1970a; Bingham 和 Larntz, 1977)

$$\tilde{\alpha}(GR) = (\mathcal{X}^T \mathcal{X} + G)^{-1} \mathcal{X}^T Y \quad (11.14)$$

特别地, 如果 $k_1 = k_2 = \dots = k_p = k$, 那么 $K = kI$ 并且 $G = U(kI)U^T = kI$, 那么岭回归是广义岭回归的一个特例。

估计或者确定 k_1, \dots, k_p 的各种方法已被提出。当我们对类似 (11.4) 的一个准则感兴趣, 则 Strawderman (1978), Berger (1975) 及 Thisted (1978a) 提出的方法是适当的。还有导致对 k_j 的另外选择的其它准则, 其中的一个在下面考虑。

主成分回归 广义岭回归的一个特例将导致对某些主成分向量进行回归。如果每个 k_j 或者被置为 0, 或者被允许趋于 $+\infty$, 则对主成分回归估计 $\tilde{\alpha}_j(PC)$, $j=1, 2, \dots, p$, 为

$$\tilde{\alpha}_j(PC) = \begin{cases} \hat{\alpha}_j & , \text{ 如果 } k_j = 0 \\ 0 & , \text{ 如果 } k_j \rightarrow +\infty \end{cases} \quad (11.15)$$

对应的 $\tilde{\beta}(PC)$ 通过将 $\tilde{\alpha}(PC)$ 代入 (11.9) 得到 (Marquardt, 1970; Mansfield et al. 1977)。如果对应于小的 λ_j 的 k_j 被允许趋于 $+\infty$, 则主成分回归可比最小二乘有小得多的 SMSE (Dempster, Schatzoff 和 Wermuth, 1977)。主成分回归的计算程序包含在 BMDP 系列中 (BMDP4R, Dixon, 1983)。

詹姆斯—斯坦 (James-Stein) 估计量 这些估计量是基于 Stein (1956) 以及 James 和 Stein (1961) 的结果, 在估计三元或更多元正态分布均值的问题中, 样本均值向量是不可容许的——即在某种意义下, 存在比样本均值总是更好的估计量。这些改进的估计量使得样本均值向量朝零或其它某点收缩。细节参见 Efron 和 Morris (1973; 1975)。

常用的詹姆斯—斯坦估计量可以在一个贝叶斯框架中得到。假设 α_j 是相互独立的正态分布, 其均值为零, 方差与 σ^2/λ_j 成比例。为得到岭估计量, 我们假设方差与 σ^2 成比例 (Dempster, 1973; Goldstein 和 Smith, 1974; Sclove, 1968; Rloph, 1976)。然后有詹姆斯—斯坦估计量为

$$\tilde{\beta}(JS) = (1 - \hat{B}) \hat{\beta} \quad (11.16)$$

$\hat{\beta}$ 的一个常用选择为

$$\hat{B} = \text{minimum} \left\{ 1, \frac{(p-2)(n-p')\hat{\sigma}^2}{\hat{\beta}^T \hat{\beta} (n-p'+2)} \right\}$$

詹姆斯—斯坦法则的这一形式 (还存在有其它形式) 具有按比例地缩小每个统计量这一不合需要的性质。这一法则也可被写为

$$\tilde{\beta}(JS) = [\mathcal{X}^T \mathcal{X} + \hat{\beta} (1 - \hat{B})^{-1} \mathcal{X}^T \mathcal{X}]^{-1} \mathcal{X}^T \mathbf{Y} \quad (11.17)$$

以用来与本章中的其它估计量作比较。这一类按比例缩小的詹姆斯—斯坦法则是通过改变 β 的定义产生的。Copas (1983) 给出这类估计的一个非贝叶斯的有趣的解释。

有偏估计量概述 所有的有偏估计量可导致在某些场合下改进 $SMSE$ 。Draper 和 Van Nostrand (1978) 指出, 只要参数 β 被较好地估计——即, 如果共线性不成问题, 并且 β 不太接近于 0, 则对最小二乘的改进将会很小。另一方面, 如果 β 被较差地估计, 或者因为共线性, 或者是 β 接近于 0, 有偏估计可能对最小二乘给出一个较大的改进。但这一改进的重要性还不清楚。如果 β 接近

于 0, 自变量与响应变量只是轻度相关, 并且当模型的形式值得怀疑时, 更精确的估计可能没有什么价值。如果数据是共线性的, 则对参数的某些组合, 数据包含很少可用于估计的可靠信息。而最小二乘对这些组合产生较差的估计, 有偏的方法给出对较差地决定的量的在某种程度上更精确的估计。

12

线性回归的推广

线性回归的范例给出模型的一个充分大而且复杂的范围，以满足许多分析者的需要。然而线性回归不能对所有问题都合适。有时，响应变量与自变量通过一个已知的，如 12.1 节所描述的非线性函数相联系。它与线性回归的不同只是它的响应变量作为参数的非线性函数而变化。12.2 节有对逻辑斯谛回归的简要介绍。当响应变量或者是一次成功，或者是一次失败，或是在固定次数的试验中成功的次数时，逻辑斯谛回归通常是合适的。逻辑斯谛回归是广义线性模型的一个例子。广义线性模型是由 Nelder 和 Wedderburn 在 1972 年一篇具有里程碑意义的论文中首先提出的一类模型。这些模型在 12.3 节中给出简要描述。

12.1 非线性回归

考虑问题，将 $y = \text{重量增量}$ 作为 $x = \text{给小动物的每日喂食量}$ 的一个函数建立模型。我们如何对 y 与 x 的关系建立模型？

由于某些原因，一个简单的直线回归模型对这一问题似乎是不合适的。响应变量可能被限制下界，而下界是不喂食动物的生长增量。响应变量也可能由喂食引起的某些生物的最大生长增量

而被限制上界。直线不能模拟这一表现，因为它们对自变量的单位增量有一个常数增量，并且最终将跨越任何上界或下界。然而，一个三参数的 S 形曲线可能可以充分地描述这一表现。三个参数可以对应于最小值或截距，对任何喂食量的最大值，称为渐近值，第三个参数控制从最小到最大的增长的速率。一个有这些性质的模型为

$$y_i = \theta_1 + \theta_2 [1 - \exp(-\theta_3 x_i)] + e_i \quad (12.1)$$

只要 $\theta_3 < 0$ ，当 x_i 增大时，响应变量接近作为渐近值的 $\theta_1 + \theta_2$ 。如果 $x_i = 0$ ， y_i 降为 θ_1 。第三个参数 θ_3 是速度参数。方程 (12.1) 只是渐近回归的许多模型之一，而渐近回归只是许多非线性模型中的一种。不过，它确实展示了非线性模型范例的重要特征：(1) 联系响应变量与自变量的函数是参数的非线性函数。模型 (12.1) 对 θ_3 是非线性的。(2) 与线性模型不同，在自变量和参数之间不需要有直接的对应。在 (12.1) 中，我们只有一个自变量，但有三个参数。(3) 参数化不是唯一的，即许多非线性回归模型是等价的。例如，模型

$$y_i = \eta_1 + \eta_2 (\eta_3)^{x_i} + e_i \quad (12.2)$$

与 (12.1) 等价，因为通过令 $\eta_1 = \theta_1 + \theta_2$ ， $\eta_2 = -\theta_2$ 及 $\eta_3 = \exp(\theta_3)$ ，即可从 (12.1) 得到 (12.2)。非线性模型参数化的非唯一性使得拟合和解释这些模型变得复杂得多。(4) 类似于线性回归模型，误差 e_i 被假设为相互独立，它们通过对响应变量增加一个量而进入模型。通常我们会作常数方差的假设，但它可以被减弱，如在线性回归 (4.1 节) 中，用加权最小二乘。对 e_i 的正态性假设与它在线性模型中有着相同的作用：它用于作推断陈述。

估计 在非线形回归中估计参数的标准方法是最小二乘法。对一般的非线性模型，

$$y_i = f(x_i; \theta) + e_i \quad (12.3)$$

其中 x_i 是自变量的 $p \times 1$ 向量， θ 是参数的 $q \times 1$ 向量， f 是 θ 的非线性函数，且 $\text{var}(e_i) = \sigma^2/w_i$ ，其中 $w_i > 0$ 已知。估计量 $\hat{\theta}$ 的选

择是使下列加权残差平方和函数取最小值的 θ 的值,

$$RSS(\theta) = \sum_{i=1}^n w_i [y_i - f(x_i; \theta)]^2 \quad (12.4)$$

如果 e_i 是相互独立的 $N(0, \sigma^2/w_i)$, 则 $\hat{\theta}$ 是 θ 的极大似然估计。 σ^2 的极大似然估计是 $\hat{\sigma}^2 = RSS(\hat{\theta})/n$, 尽管我们常用除数 $n-p$ 代替 n 。

计算 求最小二乘估计一般需要将一个迭代函数最小化。众所周知, 没有一个算法能应用于所有可能的非线性回归问题, 极小化 (12.4) 的若干程序是需要的。一个算法对最小二乘估计的收敛性可能关于初始值的选择以及模型的参数化是敏感的。许多算法要求计算 (12.4) 中函数 $f(x_i; \theta)$ 关于每个参数的一阶或可能二阶导数。不要求导数公式的程序通常会用数值来近似。

推断陈述 对非线性回归的推断陈述强烈地依赖于正态性, 并且仅对非常大的样本是精确的。在较小的样本中, 大样本结论的精确性会随问题而变化, 并且会依赖于参数化的选择。使用大多数计算机软件包的通常的大样本计算产生的标准误对某些问题可能有严重误差, 并可能低估或者高估一个估计的精确度。不过, 作为一阶近似, 标准误可以如它们在线性回归中那样被使用。例如, 一个估计与其标准误的比能为一个参数等于零的假设给出一个检验统计量, 近似的 P -值是从标准正态分布, 而不是从一个 t 分布得到的。为比较各种有竞争力的模型的检验也是有的, 如例 12.1 中所示。

例 12.1 补充饮食的影响

这里使用的数据来自于一个试验, 用刚出生的火鸡来比较两种来源的蛋氨酸对生长的影响。火鸡被分栏, 每栏 15 只。在每栏中, 来自于两种来源之一的蛋氨酸以从饮食总量的 0.04% 到 0.44% 不同的五种剂量中的一种加入到标准饮食中。响应变量是栏中动物在出生四星期时的平均体重(克)。剂量/来源的每一种组合在五个栏中被重复。五个重复试验的平均响应值列于表 12.1 中, 并由图 12.1 给出。这些数据取自于一个大型研究 (Noll et al. 1984), 其中处理相同的栏间的均方为 343.3, 自由度为 72。假设对所有处理

组的残差方差是相同的，则它给出了 σ^2 的一个纯误差估计。

在分析这个实验时，我们的目标是将生长量作为剂量的一个函数建立模型，并且对两种来源比较响应曲线。我们使用模型 (12.1) 的一个略微一般的形式。令 x_{i1} = 给第 i 组各栏的来源为 A 的蛋氨酸剂量，并令 x_{i2} = 给第 i 组各栏的来源为 B 的蛋氨酸剂量。 x_{i1} 和 x_{i2} 的值在表 12.1 中的第 2 列和第 3 列给出。然后我们考虑模型

$$y_i = \theta_1 + \theta_2 [-\exp(\theta_3 x_{i1} + \theta_4 x_{i2})] + e_i \quad (12.5)$$

这一模型使得来源 A 和来源 B 的响应线具有共同的截距 θ_1 及共同的渐近值 $\theta_1 + \theta_2$ ，但它们可以有不同的速率系数 θ_3 和 θ_4 。我们也可以考虑两种来源有不同渐近值的模型，但那种方法这里不作讨论。

程序 BMDPAR (Dixon, 1983) 是使用数值差分的求函数最小值的一般程序。通过它的帮助，我们拟合了模型 (12.5)。为使用这一程序，我们必须用 Fortran 一样的语言说明模型的形式。另外，我们必须给出对 θ_j 的初始值猜测。当剂量为其最小值 0.04% 时，截距 θ_1 大概小于响应变量 y_i 的值。故

表 12.1 补充来自于两种来源之一的蛋氨酸后
出生四星期的雄火鸡的平均体重

平均体重*	剂 量 (%)	
	来源 A	来源 B
	x_{i1}	x_{i2}
672	0.04	0
709	0.10	0
729	0.16	0
778	0.28	0
797	0.44	0
680	0	0.04
721	0	0.10
750	0	0.16
790	0	0.28
799	0	0.44

来源: Noll et al (1984)。

* 对剂量/来源的每一种组合取平均，平均指对每一栏中的 15 只火鸡和 5 个栏取平均。

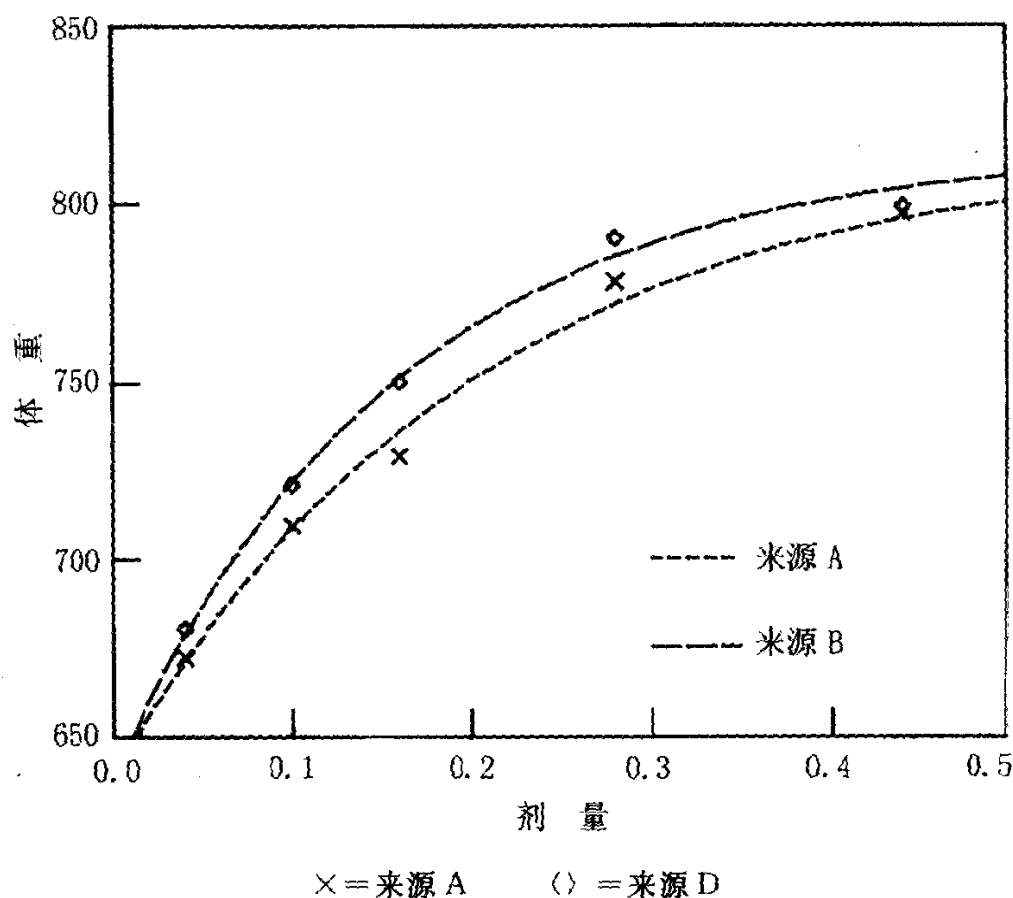


图 12.1 给定来自两种来源的各种剂量的蛋氨酸，
5 栏小火鸡的平均重量

将 600 作为 θ_1 的初始值猜测是合理的。类似地，渐近值 $\theta_1 + \theta_2$ 应该约为 800。则对 θ_2 的初始值猜测为 $800 - 600 = 200$ 。如果没有更多的经验，对速率参数的初始值猜测更为困难。不论如何，它们应该为负的。我们对 θ_3 和 θ_4 分别用值 -8 和 -6。尽管我们假设所有组有相同的残差方差，但我们将使用加权最小二乘，其中所有的 $w_i = 5$ ，即每个平均的栏数。这将变换残差平方和的尺度，用栏作为分析单位，而不是五个栏的平均。它和方差的纯误差估计有相同的单位。

使用这些值以及程序中给出的所有缺省值，需要 17 次迭代使程序收敛到最小二乘估计。这在表 12.2(a) 中给出概述。对应的响应曲线作于图 12.1。得到的标准误使用回归中的 6 个自由度的残差平方和，而不是为 σ^2 的一个估计的纯误差的 72 个自由度。曲线似乎与点集匹配非常密切。和 4.3 节一样，可以得到拟合失真的 F -检验。它是比值

$$F = \frac{RSS(\hat{\theta})/d.f.}{\hat{\sigma}^2(p.e.)} = \frac{732.805/6}{343.3} = 0.35$$

若是这样的话, F 可以近似类比于 $F(6, 72)$ 分布。由于观测值很小, 故对这一模型没有拟合失真的证据。

表 12.2 非线性回归分析简要

(a) 模型 (12.5)			(b) 模型 (12.6)		(c) 模型 (12.7)	
	值	标准误	值	标准误	值	标准误
θ_1	638.839	6.588	640.240	11.462	638.837	6.587
θ_2	175.904	6.212	175.307	11.153	175.905	6.210
θ_3	-5.053	0.620				
θ_4	-6.387	0.802			-6.387	0.802
θ_5			-5.554	1.204		
θ_6					0.791	0.049
RSS	732.805		2 509.595		732.805	
d. f.	6		7		6	

我们现在感兴趣的是比较两种来源。这可以用若干种方法完成。首先我们可以拟合模型

$$y_i = \theta_1 + \theta_2 \{1 - \exp[\theta_5(x_{i1} + x_{i2})]\} + e_i \quad (12.6)$$

它与 (12.5) 的不同只是在于共同速率系数 θ_5 对两种来源都合适。我们再次使用 BMDPAR 拟合这模型, 其结果在表 12.2 (b) 中概述。模型 (12.5) 对模型 (12.6) 的一个似然比检验为

$$LRT = -n \left[\ln \left(\frac{RSS(\hat{\theta} | \text{模型 (12.5)})}{RSS(\hat{\theta} | \text{模型 (12.6)})} \right) \right] = 12.31$$

似然比统计量可以与自由度为大模型中参数个数减去小模型中参数个数的 χ^2 -分布相比较, 这里的自由度等于 1。对这一问题 p -值很小, 模型 (12.5) 给出对数据的一个好得多的描述。我们总结得到, 两种来源的速率参数是不同的。

另外, 我们也可以考虑使用一个再参数化模型。代替 (12.5), 我们可以拟合

$$y_i = \theta_1 + \theta_2 \{1 - \exp[\theta_4(\theta_6 x_{i1} + x_{i2})]\} + e_i \quad (12.7)$$

在这个模型中, θ_6 是来源 A 关于来源 B 的相对位势。 $\theta_6=1$ 对应于相等的位势。我们可以拟合 (12.7), 并通过对 θ_6 的置信陈述来比较两种来源。这一模型的结论概述于表 12.2 (c) 中。模型 (12.5) 和 (12.7) 的拟合是等价的, 只是在计算过程中得到的答案有少量区别。例如, 由 $\hat{\theta}_6 = \hat{\theta}_3 / \hat{\theta}_4$, $\hat{\theta}_6$ 与 (12.5) 中的这些 $\hat{\theta}$ 相关联。再参数化的一个优点是它直接给出一个能解释的相对位势参数及其标准误。基于正态理论, 来源 A 相对来源 B 的位势的一个 95% 的置信区间为

$$(0.79 - 1.96(0.0488), 0.79 + 1.96(0.0488)) = \\ (0.69, 0.89)$$

附加评注 Ratkowsky 给出有一本书长度的对非线性回归的处理, 并给出它在除渐进回归以外的模型中的使用的许多特殊例子。更重要的关于非线性最小二乘的计算算法在 Kennedy 和 Gentle (1980, 第 10 章) 中有描述。许多软件包有非线性最小二乘的子程序, 但它们的质量不一样。

统计学家正在研究非线性最小二乘问题的几何性质。这一研究基于以下事实: 大部分计算算法及推断过程可被看作用线性回归问题近似于非线性回归问题。我们希望当 θ 的值接近真值时, 线性回归问题接近于非线性问题。在某些问题中, 这一近似效果很好, 但在另一些中, 效果可能很差。Beale (1960) 和 Box (1971) 首先描述了由于对非线性模型的糟糕的线性近似引起的推断里的这些问题, 而近期 Bates 和 Watts (1980) 的论文激起了在这一领域的广泛兴趣。这一正在进行的工作可能对拟合非线性模型的实践具有深远的影响。

12.2 逻辑斯谛回归

在某些回归问题中, 响应变量是分类的, 经常是或者成功, 或者失败。对这些问题, 正态线性模型显然是不合适的, 因为正态误差不对应于一个 0-1 响应。在这种情况下, 可用的一种重要方法称为逻辑斯谛回归。

表 12.3 概括了 R. Norell 进行的一项试验。我们感兴趣的是小的电流对农场动物的影响，其最终目标是了解高压电线对牲畜的影响。实验中有 7 头牛，6 种电击强度，0，1，2，3，4，5 毫安（15 毫安级的电击对许多人是痛苦的；Dalziel et al., 1941）。每头牛被电击 30 下，每种强度 5 下，按随机的次序进行。然后重复整个实验，故每头牛总共被电击 60 下。对每次电击，响应变量—嘴巴运动，或者出现，或者未出现。表 12.3 中的数据给出每种电击强度的 70 次试验中响应的总次数。这里我们暂时忽略牛之间的区别以及区组（试验）之间的区别。

表 12.3 7 头牛对 6 种不同强度的非常小的电击的响应

电流（毫安）	试验次数	响应次数	响应的比例
0	70	0	0.000
1	70	9	0.129
2	70	21	0.300
3	70	47	0.671
4	70	60	0.857
5	70	63	0.900

来源：Rick Norell。

令 y_i 为 n_i 次试验中测得的成功的次数。 n_i 和 y_i 在表 12.3 的第 2、3 列给出。假设 y_i 是一个在 n_i 次试验中的二项随机变量，其中任何一次试验的成功概率为 θ_i ， $0 \leq \theta_i \leq 1$ ，未知。在逻辑斯谛回归中，我们将 θ_i 作为自变量建立模型，这里 θ_i 是强度 x_i 的一个函数。如果没有其它理由，则因为 θ_i 在 0 到 1 之间，而 $\beta_0 + \beta_1 x_i$ 没有界限，简单线性回归模型 $\theta_i = \beta_0 + \beta_1 x_i$ 对这一目的将不是合适的。况且，图 12.2 中测得的成功比例 y_i/n_i 关于 x_i 的图更显现出 S-形，而不是直线形。这类 S 型表现形式在将一个二项响应变量作为自变量的函数建立模型时是很常见的。我们需要使用一个不同的函数形式来联系 θ_i 和 x_i 。

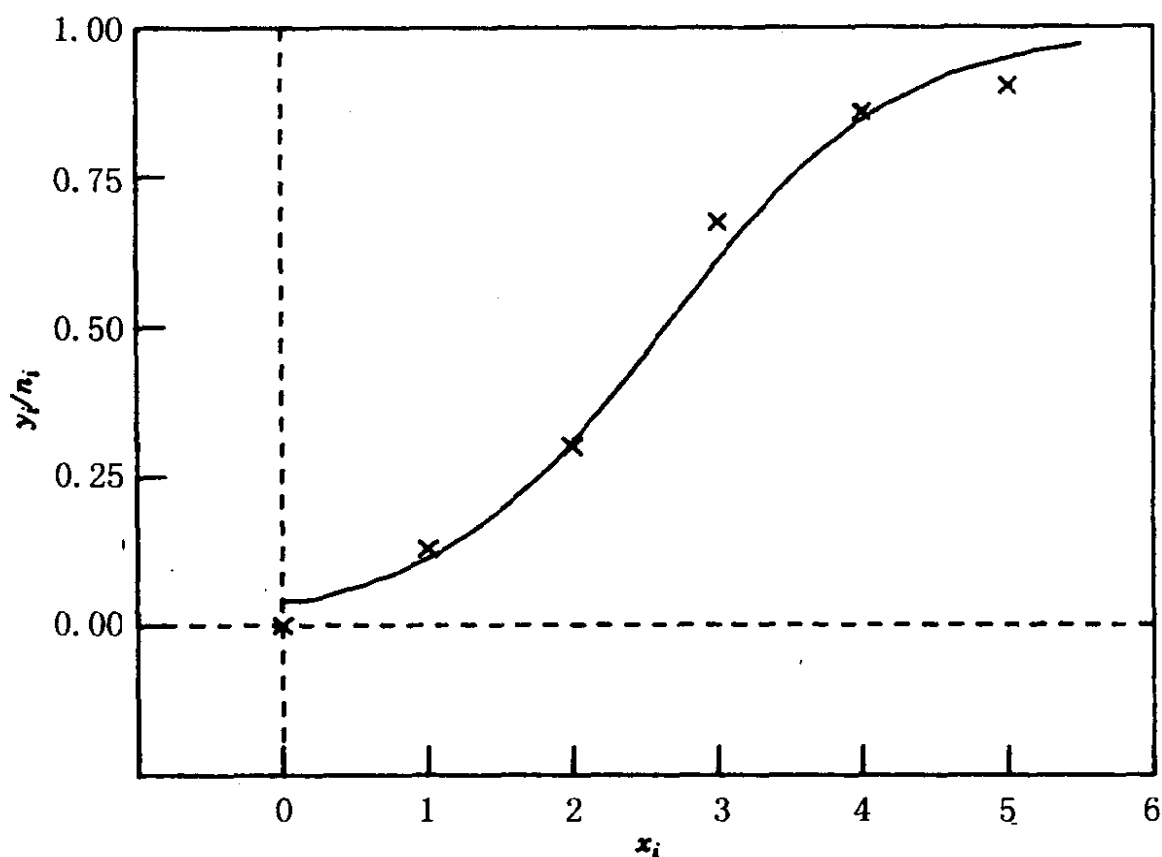


图 12.2 成功比例 $= y_i/n_i$ 关于强度 x_i 。

拟合曲线是逻辑斯谛回归曲线

通过使用对 θ_i 的 logit 变换，其定义为

$$\text{logit}(\theta_i) = \ln \left[\frac{\theta_i}{(1-\theta_i)} \right] \quad (12.8)$$

我们可以实现上节描述的目标。logit 是成功的优势 (odds)，即成功的概率与失败的概率的比值的对数。它有很多好的性质。首先，当 θ_i 增加时， $\text{logit}(\theta_i)$ 也增加。其次，尽管 θ_i 被限制在 0 到 1 之间， $\text{logit}(\theta_i)$ 在整个实数范围内变化。如果 $\theta_i < 0.5$ ， $\text{logit}(\theta_i)$ 是负的。如果 $\theta_i > 0.5$ ，则 $\text{logit}(\theta_i)$ 是正的。

这样逻辑斯谛回归可以被写成两个等价形式。首先，我们可以用 logit 尺度拟合一个线性模型，

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 x_i \quad (12.9)$$

这几乎与将成功比例 y_i/n_i 的 logit 值作为自变量的一个线性函数建立模型是一样的。由 (12.9) 解 θ_i 。使用 (12.8)，我们得到

$$E\left(\frac{y_i}{n_i}\right) = \theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (12.10)$$

方程 (12.10) 表示了原始概率尺度下，形如 S-型曲线的模型。方程 (12.9) 和 (12.10) 是等价的。

逻辑斯谛回归模型的结构包括三部分。首先，响应变量是由在已知次数的试验中的成功次数构成的相互独立的二项计数组成的。其次，成功的概率只是以线性的形式依赖于自变量。第三，存在一个函数，这里是 logit 变换，它将自变量的线性形式与二项计数的期望值相联系。这三部分结构是 12.3 节中讨论的一类广义线性模型的基础。

估计 得到的数据为 (y_i, x_i, n_i) , $i=1, 2, \dots, N$ ；我们这里用 N ，以区别二项的个数与每个二项中的试验次数 n_i 。由于 $\text{var}(y_i/n_i) = \theta_i(1-\theta_i)/n_i$ ，故每个二项的方差是不同的。权值为 $w_i = n_i/[\theta_i(1-\theta_i)]$ 的自变量的 logit (y_i/n_i) 的加权最小二乘回归看来似乎是合适的。不幸的是， θ_i ，从而 w_i 是未知的。不过可以使用一个迭代的过程。首先估计这些 θ ，然后计算给定这些 θ 后的 w_i 。对这一思想做略为细致的工作可以得到极大似然估计 (McCullagh 和 Nelder, 1983, 2.5 节及第 4 章)。算法如下：

1. 求得 β 的初始估计，并且由 (12.10)，求得 θ_i 的初始估计。这一算法中取 β 的初始估计值为 0 常是合适的。这如同取所有 θ_i 的初始估计值为 0.5。称当前估计为 $\tilde{\beta}_i$ 和 $\tilde{\theta}_i$ 。

2. 给定当前估计，计算调整响应变量 z_i ，其中

$$z_i = \text{logit}(\tilde{\theta}_i) + \frac{(y_i - n_i \tilde{\theta}_i)}{n_i \tilde{\theta}_i (1 - \tilde{\theta}_i)}$$

在迭代中使用调整响应变量可以得到极大似然估计。

3. 令 $w_i = n_i/[\tilde{\theta}_i(1-\tilde{\theta}_i)]$ 。用 w_i 作为权值，计算 z_i 关于自

变量的线性回归。用得到的 β_j 的估计来更新对 θ_i 的估计。

4. 重复执行步骤 2 和 3, 直到满足一个停止准则, 可能是直到 β_j 的最大变化量充分小。称最后的估计为 $\hat{\beta}_j$ 和 $\hat{\theta}_i$ 。通常的加权标准误即为 $\hat{\beta}_j$ 的恰当的大样本标准误。

偏差 与线性回归中的残差平方和相类似的是偏差。逻辑斯谛回归的偏差被定义为

$$\text{偏差} = 2 \sum_{i=1}^N \left[y_i \ln \left(\frac{y_i}{n_i \hat{\theta}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i - n_i \hat{\theta}_i} \right) \right]$$

如同通常的残差平方和, 偏差有 $N - p'$ 的自由度, 其中 p' 是在线性形式中 β 的个数。偏差在某些拟合优度的检验中是有用的。不同模型间偏差的变化被用于显著性检验。

对表 12.3 中的数据, 用程序 GLIM (Baker 和 Nelder, 1979) 估计系数。结果见表 12.4。得到的逻辑斯谛函数作在图 12.2 中。响应变量对自变量的依赖性的一个粗略的检验为 β_1 的估计与其标准误之比, $1.246/0.1119=11.13$ 。将它与标准正态分布相比较, 得到一个很小的 p -值。证实响应率随电流的增加而增加。一个更可靠的检验可通过拟合模型。

$$E(y_i) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \quad (12.11)$$

而得到, 如表 12.4 的最后二列所概括的。为比较象这样的两个相互套入的模型, 我们可以计算偏差的变化, $250.5 - 9.4 = 241.1$, 自由度为 $5 - 4 = 1$ 。这可以与自由度为 1 的 χ^2 -检验相比较, 以得到 p -值。我们再次清楚地证实了, 响应率随强度而增加。关于线性模型, 为检验两个有竞争力的线性模型, 引出的 t 和 F 检验是等价的。而关于其它模型, 如逻辑斯谛回归, 它们是不等价的, 并可能给出矛盾的结论。偏差变化的方法更好一些。

附加评注 在分类数据问题的对数线性方法中, 自然地产生出了逻辑斯谛回归模型; 见 Fienberg (1981)。这一模型在 McCullagh 和 Nelder (1983) 及 Cox (1970) 中有比这里更详细的讨论。

它自然地扩展到多于一个自变量的，甚至到因子设计的。这就如线性模型扩展了简单回归。偏差的变化可用于比较各种子模型。程序 GLIM 能够达到这一目的。不过，总的来说我们并不完全了解偏差统计量的使用。

在我们的例子中，所有的 n_i 都相等且大于 1。没有理由来说明为什么某些，甚至所有的 n_i 不能等于 1。例如，我们可能希望对一项手术成功的差异建立模型，它依赖于病人的年龄、性别、疾病的严重性、其它疾病等等。在这种类型的研究中，可能每一主项有一个唯一的自变量集合。

在这个简要的介绍中，省略了逻辑斯谛回归的诊断。这并不是它们不重要，而是因为至今没有一个被接受的方法体系。Pregibon (1980, 1981), Landwehr, Pregibon 和 Shoemaker (1984) 给出了某些近似的诊断。它们大致与第五和第六章中的许多方法等价。见 Cook 和 Weisberg (1982, 5.4 节)。

最后，本节讨论的对逻辑斯谛回归的计算方法并不是求极大似然估计的唯一算法。对非线性回归，有使用二阶导数的算法，对某些问题它们可能更好。Kennedy 和 Gentle (1980, 第 10 章) 是计算方法的另一份有用的参考资料。

12.3 广义线性模型

逻辑斯谛回归是广义线性模型的一个特例。广义线性模型 (GLMs) 首先由 Nelder 和 Wedderburn (1972) 提出。这些模型要求响应变量只能通过线性形式依赖于自变量，从而保持了线性自变量的思想。它们对线性模型进行了两个方面的推广：通过设定一个连接函数，将响应变量的期望与线性自变量相联系，以及对误差的分布给出一个误差函数。这些推广允许许多用于线性模型的方法能被用于更一般的问题。在线性回归中，我们的目标是将响应变量 y_i 作为 p 个自变量 $x_{1i}, \dots, x_{pi}, i=1, \dots, n$ 的函数建立模型。

线性自变量 第 i 个响应变量的期望 $E(y_i)$ 只是通过线性自变量 $\beta'x_i$ 而依赖于 x_i , 其中如通常一样, β 是未知参数的 $p' \times 1$ 向量, 可能包含截距。

连接函数 连接函数说明线性自变量和 $E(y_i)$ 的关系, 给出了线性模型的第一个方面的推广。到目前为止我们遇到了两个连接函数。第一个是恒等连接, 表述为

$$E(y_i) = x_i^T \beta$$

这是用于通常线性模型的连接函数。逻辑斯谛回归的连接函数由 (12.9) 给出,

$$x_i^T \beta = \ln \left[\frac{E(y_i/n_i)}{1 - E(y_i/n_i)} \right] \quad (12.12)$$

由 (12.12) 解出 $E(y_i/n_i)$, 我们得到回归模型

$$E\left(\frac{y_i}{n_i}\right) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad (12.13)$$

若干其它的常见连接函数列于表 12.5。例如, 对数连接表示

$$x_i^T \beta = \ln [E(y_i)]$$

它对应于回归模型

$$E(y_i) = \exp(x_i^T \beta)$$

如果 y_i 严格为正, 这可能是个合理的模型。表 12.5 的第二列给出

表 12.4 两个模型的逻辑斯谛回归分析简要

	模型 (12.9)		模型 (12.11)	
	估计	标准误	估计	标准误
β_0	-3.301	0.3238	-0.0953	0.0977
β_1	1.246	0.1119		
偏差	9.35 (4 d.f.)		250.5 (5 d.f.)	

对应于连接函数的模型。

误差函数 广义线性模型的最后一部分为随机成份。我们保留案例为相互独立的假设, 但去掉可加和正态误差的假设。我们

可以从指数型分布族中任意选取一个作为误差函数。最常见的选取列于表 12.5 中的最后一列。这样, 例如为得到正态线性模型, 我们假设 y_i 是正态分布的, 均值为 $x_i^T \beta$, 未知方差 σ^2 。如果我们假设 y_i 是泊松随机变量, 均值为 $\exp(x_i^T \beta)$, 我们得到一个泊松回归模型。

表 12.5 常见连接和误差函数*

	连接函数	逆连接函数(回归模型)	典则误差函数
恒等	$x^T \beta = E(y)$	$E(y) = x^T \beta$	正态
对数	$x^T \beta = \ln E(y)$	$E(y) = \exp(x^T \beta)$	泊松
Logit	$x^T \beta = \text{logit} E(y)$	$E(y) = \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)}$	二项式
逆	$x^T \beta = \frac{1}{E(y)}$	$E(y) = \frac{1}{x^T \beta}$	伽玛

* 为简单起见, 表示案例编号的下标被删去。

在逻辑斯谛回归模型中有 *GLIM* 的所有三个成份。假设响应概率只是通过一个线性函数依赖于自变量。连接函数为 logit, 误差函数是二项式的。拟合 *GLIM* 常通过极大似然完成。上节中描述的算法推广到拟合所有的 *GLMs*, 并且是程序 *GLIM* (Bakert 和 Nelder, 1979) 的基础。类似于残差平方和, 偏差统计量是拟合 *GLM* 的例程部分。

线性自变量, 连接函数及误差函数完整地表述了广义线性模型。如在逻辑斯谛回归情况中所指出的, 推广的代价是增加了计算以及对结论解释的困难。收益是大的。*GLMs* 将适用于广泛得多的一类问题。例如, *GLMs* 提供了一种简单的方法, 将离散数据的分析纳入到与连续数据的分析的同样结构。由线性回归模型发展而成的许多直觉知识, 例如方差分析问题的因子结构, 协方差分析, 以及在不同组中比较回归的方法, 能被应用于一类更为广泛的问题。总之, 广义线性模型看来为回归模型提供了一个重要的统一研

究法。

附加评注 McCullagh 和 Nelder(1983)对广义线性模型的理论 and 实践给出了一个综合性的介绍。表 12.5 含有在同一行的连接和误差函数必须相互适合的意思。尽管这些一对对的使用已被证实有意义,但如 McCullagh 和 Nelder 所指出的,可能会使用其它的结合。

问 题

12.1 表 12.6 中的数据来自于与 12.1 节中描述的相关的一个试验,只是在这个试验中,对每种水平的蛋氨酸使用不同数目的栏数,其分配是根据使方差最小的最优化设计(见 Noll et al., 1984)。另外,还包括一个对照组($x_1 = x_2 = 0$)。对这些数据重复文中所作的分析。基于这些数据,来源 A 是否与来源 B 一样有效?

表 12.6 补充来自于两种来源之一的蛋氨酸后,
出生四星期的雄火鸡的平均体重(Noll et al., 1984)

平均体重*	栏数 $m_i = w_i$	剂量(%)	
		来源 A	来源 B
		x_{i1}	x_{i2}
674	10	0	0
764	5	0.12	0
795	2	0.22	0
796	2	0.32	0
826	5	0.44	0
782	5	0	0.12
834	2	0	0.22
836	2	0	0.32
830	5	0	0.44

* 对剂量/来源的第 i 种组合取平均是指对每一栏中的 15 只火鸡和 m_i 个栏取平均。方差的纯误差估计为 449.2, 自由度为 66。

12.2 表 12.7 给出 12.2 节中描述的电击强度试验的详细结果。数据是对七头牛中的每一头牛在电流的六种强度中的每一种的五次试验中的正响应数。在第二区组中的所有观测是在第一区组中的所有观测进行完后进行的。由于动物的疲劳或体验,不同区组中试验的结果可能不相同。考虑区组和牛,以及电流的效应,拟合逻辑斯谛回归模型,分析这些数据。总结你的结论。

表 12.7 完整的电击数据*

牛 编号	区组 编号	电流强度(毫安)					
		0	1	2	3	4	5
1	1	0	1	1	5	5	5
1	2	0	0	4	5	5	5
2	1	0	0	2	5	5	5
2	2	0	0	0	4	5	5
3	1	0	0	1	3	5	5
3	2	0	0	0	3	5	5
4	1	0	0	3	3	4	4
4	2	0	2	2	3	5	5
5	1	0	1	1	3	4	5
5	2	0	1	0	1	4	3
6	1	0	2	2	2	5	5
6	2	0	0	0	2	0	1
7	1	0	2	3	5	5	5
7	2	0	0	2	3	3	5

* 表中数据为在五次试验中的正响应数。

附 录

1A.1 简单回归模型的形式上的展开

假设我们有两个量,一个自变量 X 和一个响应变量 Y ,并且 X 与 Y 之间的真实关系由某个未知函数 f 决定,使

$$Y=f(X) \quad (1A.1)$$

通过收集数据,我们希望研究 f ,从而研究 X 与 Y 之间的关系。为此,对 n 个单元中的每一个单元或每个案例,我们观察 X 的值 x_i 和 Y 的值 $y_i, i=1,2,\dots,n$,其中

$$y_i=f(x_i)+\epsilon_i \quad (1A.2)$$

并且 ϵ_i 是一个随机误差,它表示在观测过程中由于测量误差、被忽略的因素等等而引起的变化。

现在假设未知的 f 的形状可以由一条直线来近似。为了使之合理,可能需要改变 X 和,或 Y 的尺度或者限制 X 的值域。在任何一种情况下,把直线看作对 f 的第一次近似,如果后面的分析表明这个模型是不合适的,则必须用其它的分析来取代。这样, $f(x)$ 由 $\beta_0+\beta_1x$ 近似,并且

$$f(x_i)=\beta_0+\beta_1x_i+\delta_i \quad (1A.3)$$

其中 δ_i 为确定的或由直线在匹配 f 中的不合适而造成的拟合失真误差, $\delta_i=f(x_i)-\beta_0-\beta_1x_i$ 。为了使简单回归模型真正有用,与 ϵ_i 比较, δ_i 必须充分小(可忽略)。结合(1A.2)和(1A.3)并定义 $e_i=\epsilon_i+\delta_i$,我们得到简单回归模型。

$$y_i=\beta_0+\beta_1x_i+e_i(i=1,2,\dots,n) \quad (1A.4)$$

它由一个确定的分量和一个随机分量组成,其中的 e_i 和正文中的一样。

在这一展开的过程中,我们认为 x_i 的测量是没有误差的。诸 x_i 中包含的误差将使某些分析复杂化,并且只要可能,假设 X 中的误差相当小是有用的。检验这一假设的方法是第3章所讨论的问题。

1A.2 随机变量的均值和方差

假设我们令 u_1, u_2, \dots, u_n 为随机变量, 并令 a_0, a_1, \dots, a_n 为 $n+1$ 个已知常数。

记号 E 符号 $E(u_i)$ 读作随机变量 u_i 的期望值。术语“期望值”与术语“均值”是相同的, 或者不严格地, 为一个容量很大的样本的算术平均, 语句 $E(u_i) = 0$ 的意义是如果我们从 u_i 的分布中反复抽样, 将得到 u_i 的平均值为 0; 不过, 我们观测到的 u_i 的任何一个特殊的实现很可能不是 0。

随机变量的和的期望可以由下面两个等式记号化地表示:

$$E = (a_0 + a_1 u_1) = a_0 + a_1 E(u_1) \quad (1A.5)$$

$$E = (a_0 + \sum a_i u_i) = a_0 + \sum a_i E(u_i) \quad (1A.6)$$

例如, 假设 u_1, u_2, \dots, u_n 组成一个随机样本, 并且对所有的 $i = 1, 2, \dots, n$, $E(u_i) = \mu$ 为一常数。于是, 诸 u_i 的样本平均的期望值, $\bar{u} = \sum u_i / n = (1/n)u_1 + (1/n)u_2 + \dots + (1/n)u_n$, 可以在等式 (1A.6) 中令 $a_i = 1/n$, $i = 1, 2, \dots, n$ 及 $a_0 = 0$ 得到。由此,

$$E(\bar{u}) = \sum \left(\frac{1}{n} \right) E(u_i) = \left(\frac{1}{n} \right) \sum \mu = \left(\frac{1}{n} \right) n\mu = \mu \quad (1A.7)$$

所以样本均值是总体均值 μ 的一个无偏估计。

记号 var 符号 $\text{var}(u_i)$ 读作 u_i 的方差。方差由下面的等式定义: $\text{var}(u_i) = E[u_i - E(u_i)]^2 = u_i$ 的一个观测值与它的平均值的差的平方的期望值。 $\text{var}(u_i)$ 越大, 表示 u_i 的观测值的波动越大。符号 σ^2 常被用于一个方差, 在讨论多个方差时, 对 u 的方差用 σ_u^2 表示。

对随机变量和的方差的一般规则 (如果变量是不相关的) 为

$$\text{var}(a_0 + \sum a_i u_i) = \sum a_i^2 \text{var}(u_i) \quad (1A.8)$$

a_0 项消失: 常数的方差为零。现在我们可以用这个等式求样本均值的方差, 假设诸 u_i 是不相关的, 具有共同的方差 $\text{var}(u_i) = \sigma_u^2$:

$$\text{var}\left(\sum \left(\frac{1}{n}\right) u_i\right) = \sum \left(\frac{1}{n}\right)^2 \text{var}(u_i) = n \left(\frac{1}{n}\right)^2 \sigma_u^2 = \frac{\sigma_u^2}{n}$$

记号 cov 符号 $\text{cov}(u_i, u_j)$ 读作随机变量 u_i 与 u_j 的协方差, 定义为 $\text{cov}(u_i, u_j) = E[(u_i - E(u_i))(u_j - E(u_j))]$ 。协方差描述了两个随机变量联合变化的方式。如果两个变量是独立的, 则它们是不相关的, 但反之不一定成立。如果在定义中令 $i = j$, 则由上一符号的定义, 我们可以看出 $\text{cov}(u_i, u_i) = \text{var}(u_i)$ 。对协方差的规则为

$$\text{cov}(a_0 + a_1 u_1, a_3 + a_2 u_2) = a_1 a_2 \text{cov}(u_1, u_2) \quad (1A.9)$$

我们常常取而代之使用一种与尺度无关的协方差, 称为相关系数, 缩写成 $\text{corr}(u_i, u_j)$, 它由下式定义

$$\text{corr}(u_i, u_j) = \frac{\text{cov}(u_i, u_j)}{\sqrt{\text{var}(u_i) \text{var}(u_j)}} \quad (1A.10)$$

相关系数不依赖于随机变量的测量单位, 并且在 +1 与 -1 之间取值。如果相关系数为 0, 则变量 u_i 与 u_j 是不相关的; 这种情况只有在 $\text{cov}(u_i, u_j) = 0$ 时发生。

随机变量的线性组合的方差的一般形式依赖于变量的方差及它们的协方差, 遵从下列规则:

$$\text{var}(a_0 + \sum a_i u_i) = \sum_{i=1}^n a_i^2 \text{var}(u_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \text{cov}(u_i, u_j) \quad (1A.11)$$

1A.3 最小二乘

在简单回归中 β_0 和 β_1 的最小二乘估计是使下面的残差平方和函数取到最小值的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$,

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (1A.12)$$

极小化 (1A.12) 的一个方法是分别对 β_0 和 β_1 求导, 令导数等于零, 并解得到的方程。由此, 有

$$\begin{aligned} \frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned} \quad (1A.13)$$

通过移项, (1A.13) 变成

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i \end{aligned} \quad (1A.14)$$

方程 (1A.14) 称为模型 (1.2) 的正规方程。数据只是通过综合或充分统计量 $\sum x_i$, $\sum y_i$, $\sum x_i^2$ 和 $\sum x_i y_i$ 或等价地, 通过数值上更稳定的 \bar{x} , \bar{y} , SXX 和 SYY 而被使用; 如果使用样本均值偏差形式的模型, \bar{x} , \bar{y} , SXX 和 SYY 为得到的综合统计量。任何对这些量有相同值的两组数据集将有相同的估计

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 。解两个线性方程 (1A.14), 得

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{SXY}{SXX}\end{aligned}\quad (1A.15)$$

1A.4 最小二乘估计的均值和方差

最小二乘估计是 y_i 的线性函数, $i=1, \dots, n$, 并且由于 y_i 是 e_i 的线性函数, 我们可以对 1A.3 中求得的估计应用 1A.2 中的结论, 以求得估计的均值, 方差和协方差。特别地, 假设简单回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (i=1, 2, \dots, n)$$

是正确的。由结论 (1A.6) 和 (1A.8), $E(y_i) = \beta_0 + \beta_1 x_i$, 且 $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$ 。现在, 考虑由 (1A.15) 给出的估计 $\hat{\beta}_1$ 。假定我们定义常数 c_1, c_2, \dots, c_n (对每个 i)

$$c_i = \frac{(x_i - \bar{x})}{SXX} \quad (i=1, 2, \dots, n)$$

因为 x_i 被认为是固定的数, 所以 c_i 也是固定的数。估计量 $\hat{\beta}_1$ 等于 $\sum c_i y_i$, 为 y_i 的一个线性组合。于是求得 $\hat{\beta}_1$ 的均值为

$$\begin{aligned}E(\hat{\beta}_1) &= E(\sum c_i y_i) = \sum c_i E(y_i) \\ &= \sum c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i\end{aligned}$$

但我们由直接求和 $\sum c_i = 0$, $\sum c_i x_i = 1$, 得

$$E(\hat{\beta}_1) = \beta_1 \quad (1A.16)$$

这表明, 只要 $E(y_i) = \beta_0 + \beta_1 x_i$, $\hat{\beta}_1$ 是 β_1 的一个无偏估计。另外, 我们可以容易地证明 $E(\hat{\beta}_0) = \beta_0$ 。

$\hat{\beta}_1$ 的方差:

$$\text{var}(\hat{\beta}_1) = \text{var} \sum (c_i y_i) = \sum c_i^2 \text{var}(y_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n c_i c_j \text{cov}(y_i, y_j)$$

但由假设 $\text{cov}(y_i, y_j) = \text{cov}(\beta_0 + \beta_1 x_j + e_j, \beta_0 + \beta_1 x_i + e_i) = \text{cov}(e_j, e_i) = 0$ 。另外, 由假设 $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$ 。因此,

$$\text{var}(\hat{\beta}_1) = \sigma^2 \sum c_i^2$$

但 $\sum c_i^2 = 1/SXX$, 所以

$$\text{var}(\hat{\beta}_1) = \sigma^2 \frac{1}{SXX} \quad (1A.17)$$

为了求 $\hat{\beta}_0$ 的方差, 有

$$\text{var}(\hat{\beta}_0) = \text{var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \text{var}(\bar{y}) + \bar{x}^2 \text{var}(\hat{\beta}_1) - 2\bar{x} \text{cov}(\bar{y}, \hat{\beta}_1)$$

现在, $\text{var}(\bar{y}) = \sigma^2/n$, $\text{var}(\hat{\beta}_1)$ 由 (1A.17) 给出, 并且 $\text{cov}(\bar{y}, \hat{\beta}_1) = 0$ 。最后一个结论可以应用上节的规则加以证明, 但在直观上这是很清楚的, 因为平均值 \bar{y} 不应以任何方式依赖于拟合的斜率 $\hat{\beta}_1$ 。这样, 我们得到

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) \quad (1A.18)$$

最后,

$$\begin{aligned} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\ &= \text{cov}(\bar{y}, \hat{\beta}_1) - \bar{x} \text{cov}(\hat{\beta}_1, \hat{\beta}_1) \\ &= -\frac{\sigma^2 \bar{x}}{SXX} \end{aligned} \quad (1A.19)$$

进一步应用这些结论可以给出拟合值 $\hat{\beta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 的方差:

$$\begin{aligned} \text{var}(\hat{\beta}_i) &= \text{var}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \text{var}(\hat{\beta}_0) + x_i^2 \text{var}(\hat{\beta}_1) + 2x_i \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) + \sigma^2 x_i^2 \frac{1}{SXX} - 2\sigma^2 x_i \frac{\bar{x}}{SXX} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \right] \end{aligned} \quad (1A.20)$$

对在 x_* 的未来值 y_* 的预测 \tilde{y}_* 的方差要略为复杂一些。 \tilde{y}_* 和 y_* 有相同的均值, 所以我们有

$$\begin{aligned} E(\tilde{y}_* - y_*)^2 &= E \{ [\tilde{y}_* - E(y_*)] - [y_* - E(y_*)] \}^2 \\ &= E [\tilde{y}_* - E(y_*)]^2 + E [y_* - E(y_*)]^2 \\ &\quad - 2E [\tilde{y}_* - E(y_*)] [y_* - E(y_*)] \end{aligned}$$

因为 \tilde{y}_* 由过去的的数据计算而得, y_* 是未来的观测值, 预测和未来值是不相关的, 上一表达式中的协方差项为零。方括号中的第一项与 (1A.20) 中的相同, 只是用 x_* 代替 x_i 。方括号中的第二项恰好为未来值的方差, 等于 σ^2 。于是上一表达式为

$$\text{var}(\tilde{y}_*) = \sigma^2 \left[\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX} \right] + \sigma^2 \quad (1A.21)$$

通过移项, 可以得到 (1.36)。

1A.5 舍入, 舍入误差及回归计算的精确性

大部分回归计算是在计算机上完成的, 经常使用标准统计软件包。由这些程序得到的结论看来可以精确到很多位; 得到八位甚至十位的精度也并非

罕见。但是，我们并不鼓励对这些看来精确的数字不加辨别地使用。首先，统计问题的解答很少能获得比测量本身的精度更高的精度。例如，如果温度被测量到三至四位准确数字，则估计和预测应给出基本相同的精度。其次，某些程序可能使用不精确的计算步骤，有时会使所有得到的数字不准确。最后，有一个实际问题：我们需要多少位数能使得出的结果为他人所理解？

我们首先通过计算一个平方和来考虑在回归中的计算误差问题。假设 $x_1=12\ 541$, $x_2=12\ 537$, $x_3=12\ 548$ 。使用 SXX 的未校正的公式，但在做加法前将乘法的结果舍入至七位数字，就如某些计算机所做的那样，我们有

$$\begin{aligned}\sum x_i^2 &= (12\ 541)^2 + (12\ 537)^2 + (12\ 548)^2 \\ &= 157\ 276\ 700 + 157\ 176\ 400 + 157\ 452\ 300 \\ &= 471\ 905\ 400\end{aligned}$$

现在

$$\bar{x} = (12\ 541 + 12\ 537 + 12\ 548)/3 = 12\ 542$$

所以

$$n\bar{x}^2 = 3 \cdot (12\ 542)^2 = 3 \cdot (157\ 301\ 800) = 471\ 905\ 400$$

得到

$$SXX = \sum x_i^2 - n\bar{x}^2 = 0$$

正确的计算给出

$$SXX = (12\ 541 - 12\ 542)^2 + (12\ 537 - 12\ 542)^2 + (12\ 548 - 12\ 542)^2 = 62$$

精度为七位数字的未校正的公式在计算结果中给出不准确的数字。

Chan, Golub 和 LeVeque (1983) 讨论了计算平方和的若干方法，它们比刚才使用的“通常”的方法可能更精确，但比直接计算 SXX 的两步方法：首先计算 \bar{x} ，第二步计算每个 $x_i - \bar{x}$ ，平方并求和要快。另一种可选的方法称为更新法，在读入 x_1, \dots, x_m 后，我们计算

$$\bar{x}_m = \frac{1}{m} \sum_{i=1}^m x_i \quad \text{和} \quad SXX_m = \sum_{i=1}^m (x_i - \bar{x}_m)^2$$

当得到下一个观测值 x_{m+1} 时，我们利用方程

$$SXX_{m+1} = SXX_m + \frac{m}{m+1} (x_{m+1} - \bar{x}_m)^2$$

$$\bar{x}_{m+1} = \bar{x}_m + \frac{1}{m+1} (x_{m+1} - \bar{x}_m)$$

更新这些估计。如果我们定义 $\bar{x}_0 = SXX_0 = 0$ ，则对算法的描述就是完整的。

在计算机上对更新法编程并不比通常的方法或两步方法编程更困难,但一般它比前者更精确,比后者更快。给定新观测值 (x_{m+1}, y_{m+1}) 和当前的估计 \bar{x}_m, \bar{y}_m 和 SXY_m , 更新叉积 SXY 的公式为:

$$SXY_{m+1} = SXY_m + \frac{m}{m+1}(x_{m+1} - \bar{x}_m)(y_{m+1} - \bar{y}_m)$$

甚至当计算是精确的,也常常需要把最后的结果舍入到少数几位有效数字,但不是舍入中间结果。大部分人不能区分相关系数 0.752 和 0.773,两者都可以方便地舍入到 0.8 而不丢失任何重要信息。类似地,Forbes 数据中回归的标准误可以被舍入成 0.38,因为响应变量在计算前被舍入至小数点后二位。

Ehrenberg (1981) 讨论了许多人在吸取数值信息时的共同问题。他把这一问题归于分析者不能灵敏地给出数据,而不是人类品性的缺陷。他提出给出数值信息的若干好的规则,包括舍入至两位数,以及有效地利用数字的安排指导我们的眼睛容易地进行比较。尽管在本书中常常违反他的规则,因为为了有兴趣的读者进行重复练习,计算给出了足够的精度,这些规则对于回归用户是敏感的。

2A.1 对矩阵和向量的简要介绍

本书并不试图给出对矩阵和向量的完整介绍。关于线性代数在统计中的应用的两本有用的参考书为 Graybill (1969) 和 Seavle (1982), 尽管本书必须的材料应该包含在任何一本好的线性代数书中。

一个矩阵是数字的一个矩形排列。我们说 X 是一个 $r \times c$ 矩阵,如果它是 r 行 c 列数字的一个排列。一个特定的 4×3 矩阵 X 为

$$X = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 5 \\ 1 & 3 & 4 \\ 1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} = (x_{ij})$$

矩阵 X 的一个元素由 x_{ij} 给出,表示 X 的第 i 行第 j 列的数。例如,在前面的矩阵中, $x_{23} = 5$ 。本书通常的惯例是用黑体字母命名矩阵,用小写带下标的字母表示矩阵的元素。

向量是具有一列的矩阵。一个特定的 4×1 矩阵 Y (长度为 4 的向量) 为

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -2 \\ 0 \end{bmatrix}$$

向量也用黑体字母表示，向量中的元素有单个下标。这样 $y_3 = -2$ 是 Y 的第三个元素。我们对特殊向量的情况把符号略微复杂化一下，即参数向量 β 定义为

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

其中 β 通常由 β_0 开始，而不是 β_1 ，并且如果模型中包含截距， β 是 $(p+1) \times 1$ 的。

行向量是具有一行的矩阵。在本书中，所有的向量为列向量。如果一个向量需要用一行表示，将使用一个列向量的转置（见下面）。

一个方阵的行数 r 等于列数 c 。一个方阵 X 是对称的，如果对所有的 i 和 j ， $x_{ij} = x_{ji}$ 。一个方阵是对角阵，如果所有非主对角线的元素为零，即除了 $i = j$ ， $x_{ij} = 0$ 。下面的矩阵 C 和 D 分别是对称的和对角阵，

$$C = \begin{bmatrix} 7 & 3 & 2 & 1 \\ 3 & 4 & 1 & -1 \\ 2 & 1 & 6 & 3 \\ 1 & -1 & 3 & 8 \end{bmatrix}, D = \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

对角线上的所有元素等于 1 的对角阵称为单位阵，用符号 I 表示。有时单位阵可以写成 I_n ，表示单位阵为 $n \times n$ ：

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

一个标量是一个 1×1 矩阵，一个普通的数。标量通常是不带下标的。

矩阵运算：加法和减法 只有当两个矩阵具有相同的行数和列数时，它们可以相加减。如果 A 和 B 都是 $n \times p$ 矩阵，则它们的和 $C = A + B$ 也是 $n \times p$ 的。加法是逐个元素相加：

$$\begin{aligned}
 C = A + B &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \\
 &= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \\ a_{31} + b_{31} & a_{32} + b_{32} \end{pmatrix}
 \end{aligned}$$

减法用同样的方式操作，只是（+）号换成（-）号。通常对数的加法规则适用于矩阵的加法，即交换律

$$A + B = B + A$$

及结合律

$$(A + B) + C = A + (B + C) = (A + C) + B$$

与标量相乘 假设 k 是一个实数或标量。如果 A 是一个元素为 (a_{ij}) 的 $r \times c$ 矩阵，则 kA 是元素等于 ka_{ij} 的 $r \times c$ 矩阵。这一表示法用于指定一个随机向量的方差—协方差矩阵。表达式 $\text{var}(e) = \sigma^2 I_n$ 表示 e 的方差—协方差矩阵由单位矩阵乘以 σ^2 得到，所以 e 的每个元素具有方差 $\sigma^2(1) = \sigma^2$ ， e 的两个元素之间的协方差为 $\sigma^2(0) = 0$ 。更一般地，一个 $p \times 1$ 向量 z 的方差—协方差矩阵经常记作 $\sigma_z^2 \Sigma$ ，并且如果 s_{ij} 是 Σ 的第 (i, j) 个元素，则 z 的第 i 个与第 j 个元素之间的协方差为 $\sigma_z^2 s_{ij}$ 。

两个矩阵相乘 矩阵相乘的规则比加法和减法的规则复杂。两个矩阵按次序 AB 相乘， A 的列数（第二个维数）必须等于 B 的行数（第一个维数）。例如，如果 A 是 $n \times p$ ， B 是 $p \times q$ ，则我们得到乘积 $C = AB$ ，它是一个 $n \times q$ 矩阵。如果 A 的元素为 a_{ij} ， B 的元素为 b_{ij} ，则 C 的元素 c_{ij} 的公式为

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}$$

用语言表达，公式的含义是 c_{ij} 是通过取 A 的第 i 行和 B 的第 j 列，将 A 的指定行中的第一个元素与 B 的指定列中的第一个元素相乘，再第二个元素相乘，等等以此类推，然后将乘积相加得到的。

两个矩阵 A 和 B 相乘的最简单的情况可能是当 A 为 $1 \times p$ 而 B 为 $p \times 1$ 。得到的矩阵将是 1×1 ，为一个标量或一个普通的数。例如，如果 A 和 B 为

$$A = (1 \quad 3 \quad 2 \quad -1) \quad , \quad B = \begin{pmatrix} 2 \\ 1 \\ -2 \\ 4 \end{pmatrix}$$

则乘积 AB 为

$$AB = 1(2) + 3(1) + 2(-2) + (-1)(4) = -3$$

AB 与 BA 并不相同。事实上，对于前面的矩阵，乘积 BA 是一个 4×4 矩阵：

$$BA = \begin{pmatrix} 2 & 6 & 4 & -2 \\ 1 & 3 & 2 & -1 \\ -2 & -6 & -4 & 2 \\ 4 & 12 & 8 & -4 \end{pmatrix}$$

考虑一个小例子。用符号表示，一个 3×2 矩阵 A 乘以一个 2×2 矩阵 B 为

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{pmatrix}$$

用数字，两个矩阵相乘的例子为

$$\begin{pmatrix} 3 & 1 \\ -1 & 0 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 5 & 1 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 15+0 & 3+4 \\ -5+0 & -1+0 \\ 10+0 & 2+8 \end{pmatrix} = \begin{pmatrix} 15 & 7 \\ -5 & -1 \\ 10 & 10 \end{pmatrix}$$

在这个例子中，不仅 $AB \neq BA$ ，而且对给出的矩阵， BA 是没有定义的。不过，结合律仍然成立： $A(BC) = (AB)C$ 。

矩阵的转置 一个 $r \times c$ 矩阵 X 的转置是一个 $c \times r$ 矩阵，记为 X^T ，使得如果 x_{ij} 是 X 的元素，并且 x'_{ij} 是 X^T 的元素，则 $x_{ij} = x'_{ji}$ 。对前面给出的 X ，

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 4 \\ 1 & 5 & 4 & 6 \end{pmatrix}$$

一个列向量的转置是一个行向量。乘积的转置 $(AB)^T$ 是转置的乘积，但按相反的次序， $(AB)^T = B^T A^T$ 。

假设 A 是一个 $n \times 1$ 向量，其元素为 a_1, a_2, \dots, a_n ，则乘积 $A^T A$ 是有定义的，它是 1×1 的矩阵，由下式给出：

$$A^T A = a_1^2 + a_2^2 + \dots + a_n^2 = \sum a_i^2 \quad (2A.1)$$

即它是向量 A 的元素的平方和。这个量的平方根称为向量 A 的模或长度。例

如, 设 Y 是观测得到的 $n \times 1$ 数据向量, \hat{Y} 是拟合值的 $n \times 1$ 向量。则残差向量由 $\hat{e} = Y - \hat{Y}$ 给出, 并且残差平方和简单地为 $\hat{e}^T \hat{e} = (Y - \hat{Y})^T (Y - \hat{Y})$ 。

在本书中, X 是 $n \times p'$ 矩阵, 它给出自变量的值。 X 的第 i 行用 x_i^T 表示。在包含截距的模型中, x_i 是 $(p+1) \times 1$ 的, 其第一个元素为 1。由 X 得到的一个重要的矩阵是 $X^T X$, 它是未校正的平方和及叉积的一个 $p' \times p'$ 的对称矩阵。

分块矩阵 有时, 对矩阵的一部分的标记是有用的。 $n \times p'$ 矩阵 X 可以按列划分为矩阵 X_1 和 X_2 ,

$$X = (X_1 \quad X_2) \quad (2A.2)$$

X_1 是 X 的前 q 列, X_2 是 X 的后 $p' - q$ 列。矩阵乘积 $X^T X$ 为

$$\begin{aligned} X^T X &= (X_1 \quad X_2)^T (X_1 \quad X_2) \\ &= \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} \end{aligned}$$

在某些应用中, 将 X 和 Y 认为是数的单一的排列是方便的, 即 $Z = (X \ Y)$ 。则 $Z^T Z$ 是一个 $(p' + 1) \times (p' + 1)$ 矩阵

$$Z^T Z = \begin{pmatrix} X^T X & X^T Y \\ Y^T X & Y^T Y \end{pmatrix}$$

它给出对 X 和 Y 的未校正的平方和及叉积。

矩阵的逆 假设我们有一个 $k \times k$ 矩阵 C 。如果我们能得到另一个 $k \times k$ 矩阵, 比如说 D , 使得 $CD = I_k$, 则我们说 C 有一个逆矩阵, 通常记作 C^{-1} , 或 $C^{-1} = D$ 。如果逆矩阵存在, 它是唯一的。

只有在特殊情况下逆是容易计算的。最简单的是单位矩阵 I , 它是它自身的逆。如果 C 是一个对角矩, 如

$$C = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

则 C 的逆是对角阵

$$C^{-1} = \begin{pmatrix} \frac{1}{3} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

容易验证 $C^{-1}C = CC^{-1} = I_n$ 。只要对角元素不为零, 这对任何对角阵求逆都行得通。如果对角线上有等于 0 的元素, 则逆不存在。

逆矩阵易于求得的最重要的一类矩阵为正交矩阵。一个 $n \times n$ 矩阵 C 是正交的, 如果 $Q^T Q = Q Q^T = I_n$ 。因此, $Q^{-1} = Q^T$ 。例如, 矩阵

$$Q = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

是正交的,

$$Q^T = Q^{-1} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

在大多数需要求逆阵的回归问题中, 最好的途径是将原来的问题变换为一个等价的问题, 使得或者求逆变得容易, 或者不再需要求逆。这条途径的一个例子在附录 2A.3 中描述。

矩阵的秩 并非所有的方阵都有逆。对实数只有零没有逆; 任何非零实数 k 的逆是 $1/k$ 。如果一个方阵有逆, 我们说它是满秩的或是非奇异的。

为了说明矩阵运算, 我们将 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 代入 $RSS = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$ 并化简。首先, 进行指出的乘法, RSS 为

$$\begin{aligned} RSS &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= Y^T Y - \hat{\beta}^T X^T Y - Y^T X \hat{\beta} + \hat{\beta}^T X^T X \hat{\beta} \end{aligned} \quad (2A.3)$$

(2A.3) 右边所有的项都是 1×1 的, 所以 $(Y^T X \hat{\beta}) = (Y^T X \hat{\beta})^T = \hat{\beta}^T X^T Y$ 且

$$RSS = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \quad (2A.4)$$

用 $(X^T X)^{-1} X^T Y$ 替换 (2A.4) 中最后一项,

$$\begin{aligned} RSS &= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T (X^T X) (X^T X)^{-1} X^T Y \\ &= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T I X^T Y \\ &= Y^T Y - \hat{\beta}^T X^T Y \end{aligned} \quad (2A.5)$$

因为 $\hat{\beta}^T IX^T Y = \hat{\beta}^T X^T Y$ 。我们可以继续替换 (2A.5) 中的 $\hat{\beta}^T$ ，得到

$$\begin{aligned} RSS &= Y^T Y - [(X^T X)^{-1} X^T Y]^T X^T Y \\ &= Y^T Y - Y^T X (X^T X)^{-1} X^T Y \end{aligned}$$

最后一个结论需要承认转置的逆 $(X^T X)^{-T}$ 等于 $(X^T X)^{-1}$ ，因为 $X^T X$ 是对称的。类似的计算将给出 (2.22) 的其它形式。

2A.2 随机向量

一个元素为随机变量的向量称为随机向量。在回归中，误差的 $n \times 1$ 向量 e 是一个随机向量。回归中其它重要的随机向量有被估计的参数向量 β ，现测值向量 Y 和拟合值向量 \hat{Y} ，以及残差向量 \hat{e} 。

一个随机向量的均值或期望值为那个向量中的随机变量的均值的向量。这样，例如 e 的均值是一个全为零的向量。我们记为 $E(e) = 0$ 。

正如在附录 1A.2 中所述的，均值是线性的，即如果 z 是一个 $n \times 1$ 的随机向量， C 是任意一个 $q \times n$ 矩阵， d 是任意一个 $q \times 1$ 的确定的向量，则随机变量 $Cz + d$ 的均值为 $E(Cz + d) = CE(z) + d$ 。我们可以用这一规则求 β 的均值，

$$E(\hat{\beta}) = E[(X^T X)^{-1} X^T Y] = E[(X^T X)^{-1} X^T (X\beta + e)]$$

因为根据模型 $Y = X\beta + e$ 。于是

$$E(\hat{\beta}) = (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T E(e) \quad (2A.6)$$

但由于 $E(e) = 0$ ，

$$E(\hat{\beta}) = (X^T X)^{-1} X^T X\beta = \beta$$

并且 $\hat{\beta}$ 是 β 的一个无偏估计。附带地，它准确地证明了如果模型不正确并且如果 $E(e) \neq 0$ ，模型中的偏倚将是怎样的。例如，假设 $E(e) = Z\gamma$ ， z 是一些变量构成的某个 $n \times q$ 矩阵，尽管 Z 可能是 X 的列的变换或组合，这些变量还是与 X 中的不同，并且 γ 是一个 $q \times 1$ 的未知参数向量。则

$$E(\hat{\beta}) = \beta + (X^T X)^{-1} X^T Z\gamma$$

且 $\hat{\beta}$ 作为 β 的一个估计，其偏倚为

$$\text{偏倚} = \beta - E(\hat{\beta}) = - (X^T X)^{-1} X^T Z\gamma \quad (2A.7)$$

矩阵 $(X^T X)^{-1} X^T Z$ 称为混淆矩阵，因为 $\hat{\beta}$ 中的每个元素由混淆矩阵所决定的方式与 γ 中的元素相混淆。如果 (1) 乘积 $X^T Z = 0$ ，或 (2) $\gamma = 0$ ，偏倚将为

零。如果我们拟合模型 $Y = X\beta + e$ ，但真实的模型实际上为 $Y = X\beta + Z\gamma + e$ ，则会产生这种情况。作为练习试证明：如果较小的模型是真实的，而拟合的是较大的模型，则 β 的估计量是无偏的。

方差—协方差 一个随机向量有一个与之相联的方差—协方差矩阵。这个矩阵的对角元素的值为随机向量的元素的方差，而非对角元素的值为元素间的协方差；方差—协方差矩阵的第 (i, j) 个元素为随机向量的第 i 个元素与第 j 个元素间的协方差。我们用符号 $\text{var}(z)$ 表示随机向量 z 的方差—协方差矩阵。

假设误差向量 e 的元素具有共同的方差和为零的协方差。这被概括为 $\text{var}(e) = \sigma^2 I_n$ 。一个有不同的方差但互不相关的元素的随机向量由对角阵给出，

$$\begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & & \ddots \\ 0 & & & \sigma_n^2 \end{bmatrix}$$

$\text{var}(Cz + d)$ 的公式为

$$\text{var}(Cz + d) = C[\text{var}(z)]C^T$$

把它应用于

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y = (X^T X)^{-1} X(X\beta + e) \\ &= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T e \end{aligned}$$

我们看到第一项不包括 e ，因此对应于公式中的 d 。将 C 与 $(X^T X)^{-1} X^T$ 相联系，

$$\begin{aligned} \text{var}(\hat{\beta}) &= (X^T X)^{-1} X^T [\text{var}(e)] [(X^T X)^{-1} X^T]^T \\ &= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

如 (2.20) 所给出的。

这个结论的另一重要的应用为求与 $p' \times 1$ 向量 x 相对应的拟合值的方差。拟合值由 $\hat{y} = x^T \hat{\beta}$ 给出，因此

$$\text{var}(\hat{y}|x) = x^T [\text{var}(\hat{\beta})] x = \sigma^2 x^T (X^T X)^{-1} x \quad (2A.8)$$

对在 x_0 的目前尚未观测的未来值 y_0 的预测，预测值为 $\hat{y}_0 = x_0^T \hat{\beta}$ ，其方差为

$$\text{var}(\bar{y}_* | x_*) = \sigma^2 [1 + x_*^T (X^T X)^{-1} x_*] \quad (2A.9)$$

2A.3 最小二乘

我们这里并非重复第一章中最小二乘估计的导出,而是用一种方式求最小二乘估计,它给出了一种重要的计算方法。估计 β 是使函数

$$RSS(\beta) = (Y - X\beta)^T(Y - X\beta) \quad (2A.10)$$

取最小值的 β 的值。我们的目标是用一个能简单地求解的问题代替这一最小化问题。假设我们可以求得一个 $n \times p'$ 矩阵 Q , 使 $Q^T Q = I_{p'}$, 以及一个 $p' \times p'$ 上三角矩阵 R (所有主对角线以下的元素为零) 使

$$X = QR \quad (2A.11)$$

我们推迟讨论这些矩阵的存在性及如何求得它们。将 (2A.10) 乘开, 并用 (2A.11) 代替 X , 得

$$\begin{aligned} RSS(\beta) &= Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta \\ &= Y^T Y - 2Y^T QR\beta + \beta^T R^T Q^T QR\beta \end{aligned}$$

在这个方程的右边加上和减去 $Y^T QQ^T Y$, 使我们可以将 $RSS(\beta)$ 写成两项之和, 其中只有一项包含 β :

$$\begin{aligned} RSS(\beta) &= Y^T Y - Y^T QQ^T Y + (Y^T QQ^T Y - 2Y^T QR\beta + \beta^T R^T R\beta) \\ &= Y^T (I - QQ^T) Y + (Q^T Y - R\beta)^T (Q^T Y - R\beta) \end{aligned}$$

使第二项为零将使 $RSS(\beta)$ 取得最小值。这可以通过置

$$Q^T Y - R\beta = 0 \quad (2A.12)$$

或

$$R\beta = Q^T Y$$

达到, 并且只要 R 有逆,

$$\hat{\beta} = R^{-1} Q^T Y \quad (2A.13)$$

(2A.13) 等价于 (2.15), 这一命题留作练习。

最小二乘估计的这一导出用到 QR 因子分解, 这是由 A. S. Householder (1958) 和 G. Golab (1965) 引入的。关于将这一过程作为一个计算方法的基础的精彩出处是 Stewart. (1974, 第7章)。对任意 X , Q 和 R 的存在性及不计符号的唯一性可以通过找一个计算它们的算法而得到证明。基于这一因子分解的对最小二乘计算的高质量的 Fortran 子程序可以在 Linpack 软件包 (Dongarra et al, 1979) 中得到。在 QR 因子分解中的矩阵 R 称为 $X^T X$ 的乔勒斯基因子, 其它基于 R 避免计算 Q 的计算方法见 Stewart (1974)。

5A.1 相联回归方程

本书中的诊断统计量是实用的,因为当删除案例时可以导出简单的公式来得到各种统计量。假设 X 为 $n \times p'$, Y 为 $n \times 1$, 并已计算了矩阵 $(X^T X)^{-1}$ 。为了从 $(X^T X)^{-1}$ 计算 $(X_{(i)}^T X_{(i)})^{-1}$, 我们利用下面的基本恒等式:

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} X_i X_i^T (X^T X)^{-1}}{1 - h_{ii}} \quad (5A.1)$$

这个不寻常的公式被 Gauss (1821) 使用; 它的历史和许多变化由 Henderson 和 Searle (1981) 给出。它可以被用于给出我们想要的包含和不包含第 i 个案例的相联回归的结果。例如, 从 x_i 到余下的 $n-1$ 个案例的中心的距离定义为 $x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$ 。利用 (5A.1)

$$\begin{aligned} x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i &= x^T (X^T X)^{-1} x_i + \frac{x_i^T (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} x_i}{1 - h_{ii}} \\ &= h_{ii} + \frac{h_{ii}^2}{1 - h_{ii}} = \frac{h_{ii}}{1 - h_{ii}} \end{aligned} \quad (5A.2)$$

我们还可以进行类似的计算以求得从数据中删除第 i 个案例后的回归的任何统计量。 β 的估计

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^T X)^{-1} x_i \hat{e}_i}{1 - h_{ii}} \quad (5A.3)$$

σ^2 的估计:

$$\hat{\sigma}_{(i)}^2 = \frac{1}{n - p' - 1} \hat{\sigma}^2 (n - p' - r_i^2) \quad (5A.4)$$

8A.1 C_p 的由来

一个确定的子集模型由划分 $X = (X_1 \ X_2)$ 指定, 使得子集模型为

$$Y = X_1 \beta_1 + e$$

定义进一步的符号。设子集模型的第 i 个拟合值由 \hat{y}_i 给出, 并令 u_{ii} 为 $U = X_1 (X_1^T X_1)^{-1} X_1^T$ 的对应的对角线元素, $\text{var}(\hat{y}_i) = \sigma^2 u_{ii}$ 。类似地, 对完全模型, 如通常令 $H = X (X^T X)^{-1} X^T$, 对角线上元素为 h_{ii} , 并令 \hat{Y}_i 为第 i 个拟合值; 这是仅在本节使用的新的符号, 以区别两组拟合值。如文中所定义的

$$J_p = \frac{1}{\sigma^2} \sum_{i=1}^n \text{mse}(\hat{y}_i) \quad (8A.1)$$

其中

$$\text{mse}(\hat{y}_i) = \text{var}(\hat{y}_i) + (\text{bias})^2 = \sigma^2 u_{ii} + [E(\hat{y}_i) - E(y_i)]^2$$

由于完全模型被假设为无偏的, $E(\hat{y}_i) = E(\hat{Y}_i)$, 并且因此有

$$[E(\hat{y}_i) - E(y_i)]^2 = [E(\hat{y}_i) - E(\hat{Y}_i)]^2$$

可以证明 (Weisberg, 1981)

$$[E(\hat{y}_i) - E(\hat{Y}_i)]^2 = E(\hat{y}_i - \hat{Y}_i)^2 - \sigma^2(h_{ii} - u_{ii})$$

这个结论不是明显的, 可以通过变换成 X_1 和 X_2 为正交的问题后最方便地得到证明。 $E(\hat{y}_i - \hat{Y}_i)^2$ 用它的观测值, σ^2 用完全模型中它的估计 $\hat{\sigma}^2$ 代入 mse 的公式,

$$\widehat{\text{mse}}(\hat{y}_i) = (\hat{y}_i - \hat{Y}_i)^2 + \hat{\sigma}^2[u_{ii} - (h_{ii} - u_{ii})]$$

从而

$$\begin{aligned} C_P &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \widehat{\text{mse}}(\hat{y}_i) \\ &= \sum_{i=1}^n \left[\frac{(\hat{y}_i - \hat{Y}_i)^2}{\hat{\sigma}^2} + u_{ii} - (h_{ii} - u_{ii}) \right] \quad (8A.2) \end{aligned}$$

最后的公式有几个有趣的特征。由于 $(\hat{y}_i - \hat{Y}_i)^2 = [(\hat{y}_i - y_i) - (y_i - \hat{Y}_i)]^2 =$ 完全模型和子集模型中第 i 个残差的变化平方, 我们看到和式中的每一项有三个部分: 起因于残差的部分, 补偿 u_{ii} , 即 X_1 的第 i 行的势以及势的变化 $h_{ii} - u_{ii}$ 。(8A.2) 中方括号内的项称为 C_{pi} , 因为 $\sum C_{pi} = C_P$ 。 C_{pi} 对 u_{ii} 的图给出关于 C_P 统计量的一个诊断检验。对好的模型, 几乎所有的点应落在直线 $C_{pi} = u_{ii}$ 附近。

为了得到 C_P 的通常形式, 我们需要三个结论 (留给有兴趣的读者证明):

1. $\sum (\hat{y}_i - \hat{Y}_i)^2 =$ 继 X_1 后关于 X_2 的回归的附加平方和。
2. $\sum u_{ii} = p$ 。
3. $\sum h_{ii} = k'$ 。

将它们代入 (8A.2) 中得到 (8.21)。

表

表 A 学生 t -分布

表中的值为双尾值 $t(\alpha; \nu)$, 它使

$$\text{prob} \{ |\text{变量 } t_\nu| \geq t(\alpha; \nu) \} = \alpha$$

ν	α				
	0.200	0.100	0.050	0.010	0.001
1	3.08	6.31	12.71	63.66	636.62
2	1.89	2.92	4.30	9.92	31.60
3	1.64	2.35	3.18	5.84	12.92
4	1.53	2.13	2.78	4.60	8.61
5	1.48	2.02	2.57	4.03	6.87
6	1.44	1.94	2.45	3.71	5.96
7	1.41	1.89	2.36	3.50	5.41
8	1.40	1.86	2.31	3.36	5.04
9	1.38	1.83	2.26	3.25	4.78
10	1.37	1.81	2.23	3.17	4.59
11	1.36	1.80	2.20	3.11	4.44
12	1.36	1.78	2.18	3.05	4.32
13	1.35	1.77	2.16	3.01	4.22
14	1.35	1.76	2.14	2.98	4.14
15	1.34	1.75	2.13	2.95	4.07
16	1.34	1.75	2.12	2.92	4.01
17	1.33	1.74	2.11	2.90	3.97
18	1.33	1.73	2.10	2.88	3.92

表中的值是在明尼苏达大学的 CDC. Cyber. 172 型计算机上, 使用 IMSL 子程序 MDSTI 计算得到的。

表 A (续)

ν	α				
	0.200	0.100	0.050	0.010	0.001
19	1.33	1.73	2.09	2.86	3.88
20	1.33	1.72	2.09	2.85	3.85
21	1.32	1.72	2.08	2.83	3.82
22	1.32	1.72	2.07	2.82	3.79
23	1.32	1.71	2.07	2.81	3.77
24	1.32	1.71	2.06	2.80	3.75
25	1.32	1.71	2.06	2.79	3.73
26	1.31	1.71	2.06	2.78	3.71
27	1.31	1.70	2.05	2.77	3.69
28	1.31	1.70	2.05	2.76	3.67
29	1.31	1.70	2.05	2.76	3.66
30	1.31	1.70	2.04	2.75	3.65
31	1.31	1.70	2.04	2.74	3.63
32	1.31	1.69	2.04	2.74	3.62
33	1.31	1.69	2.03	2.73	3.61
34	1.31	1.69	2.03	2.73	3.60
35	1.31	1.69	2.03	2.72	3.59
36	1.31	1.69	2.03	2.72	3.58
37	1.30	1.69	2.03	2.72	3.57
38	1.30	1.69	2.02	2.71	3.57
39	1.30	1.68	2.02	2.71	3.56
40	1.30	1.68	2.02	2.70	3.55
41	1.30	1.68	2.02	2.70	3.54
42	1.30	1.68	2.02	2.70	3.54
43	1.30	1.68	2.02	2.70	3.53
44	1.30	1.68	2.02	2.69	3.53
45	1.30	1.68	2.01	2.69	3.52
46	1.30	1.68	2.01	2.69	3.51
47	1.30	1.68	2.01	2.68	3.51
48	1.30	1.68	2.01	2.68	3.51
49	1.30	1.68	2.01	2.68	3.50
50	1.30	1.68	2.01	2.68	3.50
60	1.30	1.67	2.00	2.66	3.46
70	1.29	1.67	1.99	2.65	3.44
80	1.29	1.66	1.99	2.64	3.42
90	1.29	1.66	1.99	2.63	3.40
100	1.29	1.66	1.98	2.63	3.39
120	1.29	1.66	1.98	2.62	3.37
∞	1.28	1.64	1.96	2.58	3.29

表 B F -分布

表中的值为 $F(\alpha, \nu_1, \nu_2)$, 它使

$$\text{prob} \{ \text{变量 } F(\nu_1, \nu_2) \geq F(\alpha, \nu_1, \nu_2) \} = \alpha$$

$$(\alpha=0.05)$$

$\nu_1 \backslash \nu_2$		分子的自由度									
		1	2	3	4	5	6	7	8	9	10
分 母 的 自 由 度	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	2	18.51	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

取自 Draper 和 Smith(1966)并在取得 E. S. Pearson 和 H. O. Hartley(1966)同意后重新生成, *Biometrika. Tables. for. Statisticians*, 第一卷第三版, 伦敦, 剑桥大学。

表 B (续 1)

 $(\alpha=0.05)$

ν_2	ν_1	分子的自由度							
		12	15	20	24	30	40	60	120 ∞
1	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.43	19.45	19.45	19.46	19.47	19.48	19.48	19.49	19.50
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.85	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
分	15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11
母	16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06
的	17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01
自	18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97
由	19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93
度	20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90
	21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87
	22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84
	23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81
	24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79
	25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77
	26	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75
	27	2.13	2.06	1.97	1.95	1.88	1.84	1.79	1.73
	28	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71
	29	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70
	30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68
	40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58
	60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47
	120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35
	∞	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22

表 B (续 2)

 $(\alpha=0.01)$

$\nu_2 \backslash \nu_1$	分子的自由度									
	1	2	3	4	5	6	7	8	9	10
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
分 15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
母 16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
的 17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
自 18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
由 19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
度 20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

表 B (续 3)

 $(\alpha=0.01)$

$\nu_2 \backslash \nu_1$									
	12	15	20	24	30	40	60	120	∞
分子的自由度									
1	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.84	2.80	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

表 C χ^2 -分布的分位点表中的值为 $\chi^2(\alpha; n)$, 它使 $\text{prob}\{\text{变量 } \chi^2(n) \geq \chi^2(\alpha; n)\} = \alpha$

d.f.	α				
	0.20	0.10	0.05	0.01	0.001
1	1.64	2.71	3.84	6.64	10.81
2	3.22	4.60	5.99	9.22	13.69
3	4.64	6.25	7.82	11.32	16.29
4	5.99	7.78	9.49	13.28	18.43
5	7.29	9.24	11.07	15.09	20.75
6	8.56	10.65	12.60	16.81	22.68
7	9.80	12.02	14.07	18.47	24.53
8	11.03	13.36	15.51	20.08	26.32
9	12.24	14.69	16.93	21.65	28.06
10	13.44	15.99	18.31	23.19	29.76
11	14.63	17.28	19.68	24.75	31.43
12	15.81	18.55	21.03	26.25	33.07
13	16.99	19.81	22.37	27.72	34.68
14	18.15	21.07	23.69	29.17	36.27
15	19.31	22.31	25.00	30.61	37.84
16	20.47	23.55	26.30	32.03	39.39
17	21.62	24.77	27.59	33.44	40.93
18	22.76	25.99	28.88	34.83	42.44
19	23.90	27.21	30.15	36.22	43.95
20	25.04	28.42	31.42	37.59	45.44
21	26.17	29.62	32.68	38.96	46.92
22	27.30	30.82	33.93	40.31	48.39
23	28.43	32.01	35.18	41.66	49.85
24	29.56	33.20	36.42	43.00	51.29
25	30.68	34.38	37.66	44.34	52.73
26	31.80	35.57	38.89	45.66	54.16
27	32.91	36.74	40.12	46.99	55.58
28	34.03	37.92	41.34	48.30	57.00
29	35.14	39.09	42.56	49.61	58.41
30	36.25	40.26	43.78	50.91	59.81
40	47.26	51.80	55.75	63.71	73.49
50	58.16	63.16	67.50	76.17	86.74

表 C (续)

d. f.	α				
	0.20	0.10	0.05	0.01	0.001
60	68.97	74.39	79.08	88.40	99.68
70	79.71	85.52	90.53	100.44	112.38
80	90.40	96.57	101.88	112.34	124.90
90	101.05	107.56	113.14	124.13	137.27
100	111.66	118.49	124.34	135.82	149.50

注：表中的值是在明尼苏达大学，使用IMSL (1977) 库中的子程序 MDCHI 计算得到的。

表 D $n \leq 20^*$ 时的 Rankits

i	n									
	1	2	3	4	5	6	7	8	9	10
1	0	-0.56	-0.85	-1.03	-1.16	-1.27	-1.35	-1.42	-1.49	-1.54
2		0.56	0.00	-0.30	-0.50	-0.64	-0.76	-0.85	-0.93	-1.00
3			0.85	0.30	0.00	-0.20	-0.35	-0.47	-0.57	-0.66
4				1.03	0.50	0.20	0.00	-0.15	-0.27	-0.38
5					1.16	0.64	0.35	0.15	0.00	-0.12
6						1.27	0.76	0.47	0.27	0.12

i	n									
	11	12	13	14	15	16	17	18	19	20
1	-1.59	-1.63	-1.67	-1.70	-1.74	-1.77	-1.79	-1.82	-1.84	-1.87
2	-1.06	-1.12	-1.16	-1.21	-1.25	-1.28	-1.32	-1.35	-1.38	-1.41
3	-0.73	-0.79	-0.85	-0.90	-0.95	-0.99	-1.03	-1.07	-1.10	-1.13
4	-0.46	-0.54	-0.60	-0.66	-0.71	-0.76	-0.81	-0.85	-0.89	-0.92
5	-0.22	-0.31	-0.39	-0.46	-0.52	-0.57	-0.62	-0.66	-0.71	-0.75
6	0.00	-0.10	-0.19	-0.27	-0.34	-0.40	-0.45	-0.50	-0.55	-0.59
7	0.22	0.10	0.00	-0.09	-0.17	-0.23	-0.30	-0.35	-0.40	-0.45
8	0.46	0.31	0.19	0.09	0.00	-0.08	-0.15	-0.21	-0.26	-0.31
9	0.73	0.54	0.39	0.27	0.17	0.08	0.00	-0.07	-0.13	-0.19
10	1.06	0.79	0.60	0.46	0.34	0.23	0.15	0.07	0.00	-0.06

* 经允许从 E. S. Pearson 和 H. O. Hartley (1966) 的表 28 中节录。

Biometrika Tables for Statisticians, 第一卷第三版, 伦敦, 剑桥大学。

注：本表未给出的 rankit 的值（或正态次序统计量的期望值）可以由对称得到。例如，对一个容量 $n=17$ 的样本，第 15 位大的 rankit 等于对 $n=17$ 的第 3 位 rankit 的负数，即 1.03。

表 E 对异常值检验的临界值

 $(\alpha=0.05)$

$n \backslash p'$	1	2	3	4	5	6	7	8	9
6	4.85	6.23	10.89	76.39					
7	4.38	5.07	6.58	11.77	89.12				
8	4.12	4.53	5.26	6.90	12.59	101.9			
9	3.95	4.22	4.66	5.44	7.18	13.36	114.6		
10	3.83	4.03	4.32	4.77	5.60	7.45	14.09	127.3	
11	3.75	3.90	4.10	4.40	4.88	5.75	7.70	14.78	140.1
12	3.69	3.81	3.96	4.17	4.49	4.98	5.89	7.94	15.44
13	3.65	3.74	3.86	4.02	4.24	4.56	5.08	6.02	8.16
14	3.61	3.69	3.79	3.91	4.07	4.30	4.63	5.16	6.14
15	3.58	3.65	3.73	3.83	3.95	4.12	4.36	4.70	5.25
16	3.56	3.62	3.68	3.77	3.87	4.00	4.17	4.41	4.76
17	3.54	3.59	3.65	3.72	3.80	3.90	4.04	4.21	4.46
18	3.53	3.57	3.62	3.68	3.75	3.83	3.94	4.08	4.26
19	3.52	3.56	3.60	3.65	3.71	3.78	3.86	3.97	4.11
20	3.51	3.54	3.58	3.62	3.67	3.73	3.81	3.89	4.00
21	3.50	3.53	3.57	3.60	3.65	3.70	3.76	3.83	3.92
22	3.50	3.52	3.55	3.59	3.63	3.67	3.72	3.78	3.86
23	3.49	3.52	3.54	3.57	3.61	3.65	3.69	3.75	3.81
24	3.49	3.51	3.53	3.56	3.59	3.63	3.67	3.71	3.77
25	3.48	3.50	3.53	3.55	3.58	3.61	3.65	3.69	3.73
26	3.48	3.50	3.52	3.54	3.57	3.60	3.63	3.66	3.70

表 E (续 1)

 $(\alpha=0.05)$

$n \backslash p'$	10	11	12	13	14	15	20	25	30
6									
7									
8									
9									
10									
11									
12	152.8								
13	16.08	165.5							
14	8.37	16.69	178.2						
15	6.25	8.58	17.28	191.0					
16	5.33	6.36	8.77	17.85	203.7				
17	4.82	5.40	6.47	8.95	18.40	216.4			
18	4.51	4.88	5.47	6.57	9.13	18.93			
19	4.30	4.55	4.93	5.54	6.67	9.30			
20	4.15	4.33	4.59	4.98	5.60	6.76			
21	4.03	4.18	4.37	4.64	5.03	5.67			
22	3.95	4.06	4.21	4.40	4.68	5.08	280.1		
23	3.88	3.98	4.09	4.24	4.44	4.71	21.41		
24	3.83	3.91	4.00	4.12	4.27	4.47	10.07		
25	3.79	3.85	3.93	4.02	4.14	4.30	7.17		
26	3.75	3.81	3.87	3.95	4.05	4.17	5.95		

表 E (续 2)

 $(\alpha=0.05)$

$n \backslash p'$	1	2	3	4	5	6	7	8	9
27	3.48	3.50	3.52	3.54	3.56	3.58	3.61	3.65	3.68
28	3.48	3.50	3.51	3.53	3.55	3.58	3.60	3.63	3.66
29	3.48	3.49	3.51	3.53	3.55	3.57	3.59	3.62	3.64
30	3.48	3.49	3.51	3.52	3.54	3.56	3.58	3.60	3.63
31	3.48	3.49	3.50	3.52	3.54	3.55	3.57	3.59	3.62
32	3.48	3.49	3.50	3.52	3.53	3.55	3.57	3.59	3.61
33	3.48	3.49	3.50	3.52	3.53	3.54	3.56	3.58	3.60
34	3.48	3.49	3.50	3.51	3.53	3.54	3.56	3.57	3.59
35	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.57	3.58
36	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.56	3.58
37	3.48	3.49	3.50	3.51	3.52	3.53	3.55	3.56	3.57
38	3.48	3.49	3.50	3.51	3.52	3.53	3.54	3.56	3.57
39	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.57
40	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.56
50	3.51	3.51	3.51	3.52	3.53	3.53	3.54	3.54	3.55
60	3.53	3.53	3.53	3.54	3.54	3.54	3.55	3.55	3.56
70	3.55	3.55	3.55	3.55	3.56	3.56	3.56	3.56	3.57
80	3.57	3.57	3.57	3.57	3.57	3.58	3.58	3.58	3.58
90	3.58	3.59	3.59	3.59	3.59	3.59	3.59	3.60	3.60
100	3.60	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.61
200	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73
300	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81
400	3.87	3.87	3.87	3.87	3.87	3.87	3.87	3.88	3.88
500	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92

表 E (续 3)

 $(\alpha=0.05)$

$n \backslash p'$	10	11	12	13	14	15	20	25	30
27	3.72	3.77	3.83	3.89	3.97	4.07	5.29	343.8	
28	3.70	3.74	3.79	3.84	3.91	3.99	4.88	23.63	
29	3.68	3.71	3.76	3.81	3.86	3.93	4.61	10.74	
30	3.66	3.69	3.73	3.77	3.82	3.88	4.42	7.53	
31	3.64	3.67	3.71	3.74	3.79	3.84	4.28	6.18	
32	3.63	3.66	3.69	3.72	3.76	3.80	4.17	5.47	407.4
33	3.62	3.64	3.67	3.70	3.74	3.77	4.08	5.03	25.66
34	3.61	3.63	3.66	3.68	3.71	3.75	4.01	4.74	11.34
35	3.60	3.62	3.64	3.67	3.70	3.73	3.96	4.53	7.84
36	3.60	3.61	3.63	3.66	3.68	3.71	3.91	4.37	6.39
37	3.59	3.61	3.62	3.65	3.67	3.69	3.87	4.26	5.62
38	3.58	3.60	3.62	3.64	3.66	3.68	3.84	4.16	5.16
39	3.58	3.59	3.61	3.63	3.65	3.67	3.81	4.09	4.84
40	3.58	3.59	3.60	3.62	3.64	3.66	3.79	4.03	4.62
50	3.56	3.57	3.57	3.58	3.59	3.60	3.66	3.75	3.88
60	3.56	3.57	3.57	3.57	3.58	3.59	3.62	3.67	3.73
70	3.57	3.58	3.58	3.58	3.59	3.59	3.61	3.64	3.67
80	3.58	3.59	3.59	3.59	3.60	3.60	3.61	3.63	3.66
90	3.60	3.60	3.60	3.60	3.61	3.61	3.62	3.63	3.65
100	3.61	3.61	3.62	3.62	3.62	3.62	3.63	3.64	3.65
200	3.73	3.73	3.73	3.73	3.73	3.74	3.74	3.74	3.74
300	3.81	3.82	3.82	3.82	3.82	3.82	3.82	3.82	3.82
400	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88
500	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92

表 E (续 4)

 $(\alpha=0.01)$

$n \backslash p'$	1	2	3	4	5	6	7	8	9
6	7.53	10.87	24.46	382.0					
7	6.35	7.84	11.45	26.43	445.6				
8	5.71	6.54	8.12	11.98	28.26	509.3			
9	5.31	5.84	6.71	8.38	12.47	29.97	573.0		
10	5.04	5.41	5.96	6.97	8.61	12.92	31.60	636.6	
11	4.85	5.12	5.50	6.07	7.01	8.83	13.35	33.14	700.3
12	4.71	4.91	5.19	5.58	6.17	7.15	9.03	13.75	34.62
13	4.60	4.76	4.97	5.25	5.66	6.26	7.27	9.22	14.12
14	4.51	4.64	4.81	5.02	5.32	5.73	6.35	7.39	9.40
15	4.44	4.55	4.68	4.85	5.08	5.37	5.80	6.43	7.50
16	4.38	4.48	4.59	4.72	4.90	5.12	5.43	5.86	6.51
17	4.34	4.41	4.51	4.62	4.76	4.94	5.17	5.48	5.92
18	4.30	4.36	4.44	4.54	4.66	4.80	4.98	5.21	5.53
19	4.26	4.32	4.39	4.47	4.57	4.69	4.83	5.01	5.25
20	4.23	4.29	4.35	4.42	4.50	4.60	4.72	4.86	5.05
21	4.21	4.26	4.31	4.37	4.44	4.52	4.62	4.74	4.89
22	4.19	4.23	4.28	4.33	4.39	4.46	4.55	4.65	4.77
23	4.17	4.21	4.25	4.30	4.35	4.41	4.49	4.57	4.67
24	4.15	4.19	4.22	4.27	4.32	4.37	4.43	4.51	4.59
25	4.14	4.17	4.20	4.24	4.28	4.33	4.39	4.45	4.53
26	4.12	4.15	4.18	4.22	4.26	4.30	4.35	4.41	4.47

表 E (续 5)

 $(\alpha=0.01)$

$n \backslash p'$	10	11	12	13	14	15	20	25	30
6									
7									
8									
9									
10									
11									
12	763.9								
13	36.03	827.6							
14	14.48	37.40	891.3						
15	9.57	14.82	38.71	954.9					
16	7.60	9.73	15.15	39.98					
17	6.59	7.70	9.88	15.46	41.21				
18	5.98	6.66	7.80	10.03	15.76	42.41			
19	5.57	6.03	6.72	7.89	10.17	16.05			
20	5.29	5.62	6.08	6.79	7.98	10.31			
21	5.08	5.33	5.66	6.13	6.85	8.06			
22	4.92	5.11	5.36	5.70	6.18	6.91			
23	4.80	4.95	5.14	5.40	5.74	6.22	47.94		
24	4.70	4.82	4.98	5.17	5.43	5.78	17.36		
25	4.62	4.72	4.85	5.00	5.20	5.46	10.92		
26	4.55	4.64	4.74	4.87	5.03	5.23	8.43		

表 E (续 6)

 $(\alpha=0.01)$

$n \backslash p'$	1	2	3	4	5	6	7	8	9
27	4.11	4.14	4.17	4.20	4.24	4.27	4.32	4.37	4.43
28	4.10	4.13	4.15	4.18	4.21	4.25	4.29	4.33	4.38
29	4.09	4.12	4.14	4.17	4.20	4.23	4.26	4.30	4.35
30	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.28	4.32
31	4.08	4.10	4.12	4.34	4.14	4.19	4.22	4.26	4.29
32	4.07	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.27
33	4.07	4.08	4.10	4.12	4.14	4.17	4.19	4.22	4.25
34	4.06	4.08	4.09	4.11	4.13	4.15	4.18	4.20	4.23
35	4.06	4.07	4.09	4.11	4.12	4.14	4.16	4.19	4.21
36	4.05	4.07	4.08	4.10	4.12	4.13	4.15	4.18	4.20
37	4.05	4.06	4.08	4.09	4.11	4.13	4.14	4.16	4.19
38	4.05	4.06	4.07	4.09	4.10	4.12	4.13	4.15	4.17
39	4.04	4.06	4.07	4.08	4.10	4.11	4.13	4.14	4.16
40	4.04	4.05	4.06	4.08	4.09	4.10	4.12	4.14	4.15
50	4.03	4.03	4.04	4.05	4.06	4.07	4.07	4.08	4.09
60	4.03	4.03	4.04	4.04	4.05	4.05	4.06	4.06	4.07
70	4.03	4.03	4.04	4.04	4.05	4.05	4.05	4.06	4.06
80	4.04	4.04	4.04	4.05	4.05	4.05	4.06	4.06	4.06
90	4.05	4.05	4.05	4.05	4.06	4.06	4.06	4.06	4.07
100	4.06	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.07
200	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15
300	4.21	4.21	4.21	4.21	4.21	4.21	4.22	4.22	4.22
400	4.26	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27
500	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31

表 E (续 7)

($\alpha=0.01$)

$n \backslash p'$	10	11	12	13	14	15	20	25	30
27	4.49	4.57	4.66	4.76	4.89	5.05	7.17		
28	4.44	4.51	4.59	4.68	4.78	4.91	6.43	52.90	
29	4.40	4.46	4.53	4.60	4.69	4.80	5.94	18.50	
30	4.36	4.42	4.47	4.54	4.62	4.71	5.60	11.44	
31	4.33	4.38	4.43	4.49	4.56	4.64	5.35	8.75	
32	4.31	4.35	4.39	4.45	4.50	4.57	5.16	7.40	
33	4.28	4.32	4.36	4.41	4.46	4.52	5.01	6.60	57.43
34	4.26	4.29	4.33	4.37	4.42	4.47	4.89	6.09	19.51
35	4.24	4.27	4.31	4.34	4.39	4.43	4.79	5.72	11.90
36	4.22	4.25	4.28	4.32	4.36	4.40	4.71	5.46	9.03
37	4.21	4.24	4.26	4.29	4.33	4.37	4.64	5.26	7.60
38	4.20	4.22	4.25	4.27	4.31	4.34	4.59	5.10	6.76
39	4.18	4.21	4.23	4.26	4.28	4.32	4.54	4.97	6.21
40	4.17	4.19	4.22	4.24	4.27	4.29	4.49	4.87	5.83
50	4.10	4.12	4.13	4.14	4.15	4.17	4.25	4.38	4.59
60	4.08	4.08	4.09	4.10	4.11	4.12	4.17	4.23	4.32
70	4.07	4.07	4.08	4.08	4.09	4.09	4.13	4.17	4.22
80	4.07	4.07	4.07	4.08	4.08	4.09	4.11	4.13	4.17
90	4.07	4.07	4.07	4.08	4.08	4.08	4.10	4.12	4.14
100	4.07	4.08	4.08	4.08	4.08	4.09	4.10	4.11	4.13
200	4.15	4.15	4.15	4.15	4.15	4.15	4.16	4.16	4.16
300	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22
400	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27
500	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31

注: 表中第 n 行第 p' 列的值为 $t(\alpha/n; n-p'-1)$, $\alpha=0.01$ 及 0.05 。表的排列方式是 Christopher Bingham 建议的。表是在明尼苏达大学的 CDC6400 型计算机上用 IMSI 子程序 MDSTI 计算得到的。

参考文献

- Afifi. A. A. and R. M. Elashoff (1966). "Missing values in multivariate statistics I. Review of the literature." *J. Am. Statist. Assoc.*, **61**, 595—604.
- Aitchison. J. and I. R. Dunsmore (1975). *Statistical Prediction Analysis*. New York; Cambridge University.
- Aitken. A. C. (1934) "Note on selection from a multivariate normal population." *Proc. Edinburgh Math. Soc.*, **4**, 106—110.
- Allen. D. M. (1974). "The relationship between variable selection and prediction." *Technometrics*, **16**, 125—127.
- Allison. T. and D. V. Cicchetti (1976). "Sleep in mammals: Ecological and constitutional correlates," *Science*, **194**, 732—734.
- Anderson, T. W. (1976). "Estimation of linear functional relationships: Approximate distributions and connections with simultaneous equations in econometrics (with discussion)" *J. Roy. Statist. Soc. Ser. B.* **38**, 1—36.
- Andrews, D. F. (1974) "A robust method for multiple linear regression." *Technometrics*, **16**, 523—531.
- Andrews. D. E., P. Bickel, F. Hampel, P. Huber, W. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location*. Princeton; Princeton University Press.
- Anscombe. F. J. (1973). "Graphs in statistical analysis." *Am. Statist.* **27**, 17—21.
- Atkinson, A. C. (1973). "Testing transformations to normality." *J. Roy. Statist. Soc. Ser. B.* **35**, 473—479.
- Atkinson. A. C. (1981). "Robustness, transformations and two graphical

- displays for outlying and influential observations in regression." *Biometrika*, **68**, 13—20.
- Atkinson, A. C. (1982). "Regression diagnostics, transformations and constructed variables (With discussion)." *J. Roy. Statist. Soc. Ser. B*, **44**, 1—35.
- Atkinson, A. C. (1983). "Diagnostics, regression analysis and shifted power transformations." *Technometrics*, **25**, 23—34.
- Baes, C. F. and H. H. Kellogg (1953). "Effect of dissolved sulphur on the surface tension of liquid copper." *J. Metals*, **5**, 643—648.
- Baker, R. J. and J. A. Nelder (1978). "The GLIM System, Release 3, Generalized Linear Interactive Modeling." Oxford: Numerical Algorithms Group.
- Barnett, V. and T. Lewis (1978). *Outliers in Statistical Data*. Chichester: Wiley.
- Bates, D. and D. Watts (1980). "Relative curvature measures of nonlinearity (with dicussion)." *J. Ror, Statist, Soc. Ser, 22*, 41—88.
- Beale, E. M. L (1960). "Confidence regions in nonlinear estimation (with discussion)." *J. Roy, Statist. Soc. Ser. B*, **22**, 41—88.
- Bcale, E. M. L., M. G. Kendall, and D. W. Mann (1967). "The discarding of variables in multivariate analysis." *Bimetrika*, **54**, 537—566.
- Beale, E. M. L. and R. J. Little (1975). "Missing values in multivarite analysis." *J. Roy. Statist. Soc. Ser. B*, **37**, 129—145.
- Beaton, A. E. (1964). "The use of special matrix operators in statistical calculations." Unpublished Ph. D. dissertation, Harvard University.
- Beaton, A. E., D. B. Rubin, and J. L. Barone (1976). "The acceptability of regression solutions: Another look at computational stability." *J. Am. Statist. Assoc.*, **71**, 158—168.
- Beckman, R. and R. D. Cood (1983). "Outlier... s." *Technometrics*, **25**, 119—149.
- Belsley, D. A. (1984). "Demeaning condition diagnostics through centering." *Am. Statist.*, **38**, 73—77.
- Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics*. New York: Wiley.

- Berger, J. (1975). "Minimax estimation of location vectors for a wide variety of densities." *Ann. Statist.*, **3**, 1318—1328.
- Berd, K. (1977). "Tolerance and condition in regression computations." *J. Am. Statist. Assoc.*, **72**, 863—866.
- Berkson, J. (1950). "Are there two regressions?" *J. Am. Statist. Assoc.*, **45**, 164—180.
- Bickel, P. and K. Doksum (1981). "An analysis of transformations revisited." *J. Am. Statist. Assoc.*, **76**, 296—311.
- Bingham, C. and K. Larntz (1977). "Comment on 'A simulation study of alternatives to ordinary least squares'." *J. Am. Statist. Assoc.*, **72**, 97—102.
- Bland, J. (1978). "A comparison of certain aspects of ontogeny in the long and short shoots of McIntosh apple during one annual growth cycle." Unpublished Ph. D. dissertation, University of Minnesota, St. Paul, MN.
- Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*. New York: Wiley.
- Box, G. E. P. (1953). "Non-normality and tests on variances." *Biometrika*, **40**, 318—335.
- Box, G. E. P. (1980). "Sampling and Bayes' inference in scientific modelling and robustness (with discussion)." *J. Roy. Statist. Soc. Ser. A*, **143**, 383—430.
- Box, G. E. P. and D. R. Cox (1964). "An analysis of transformations (with discussion)." *J. Roy. Statist. Soc. Ser. B*, **26**, 211—246.
- Box, G. E. P., J. S. Hunter, and W. G. Hunter (1978). *Statistics for Experimenters*. New York: Wiley.
- Box, G. E. P. and P. W. Tidwell (1962). "Transformations of the independent variables." *Technometrics*, **4**, 531—550.
- Box, G. E. P. and K. B. Wilson (1951). "On the experimental attainment of optimal conditions (with discussion)" *J. Roy. Statist. Soc. Ser. B*, **13**, 1—45.
- Box, M. J. (1971) "Bias in non-linear estimation (with discussion)" *J. Roy. Statist. Soc. Ser. B*, **33**, 171—201.
- Buck, S. F. (1960). "A method of estimating missing values in multivariate data suitable for use with an electronic computer." *J. Roy. Statist. Soc.*

- Ser. B.* **22**, 302—306.
- Burt, C. (1966). "The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart." *Br. J. Psych.*, **57**, 137—53.
- Butler, R. W. (1984). "The significance attained by the best fitting regressor variable." *J. Am. Statist. Assoc.*, **79**, 341—48.
- Carroll, R. (1980). "A robust method for testing transformations to achieve approximate normality." *J. Roy. Statist. Soc. Ser. B.* **42**, 71—78.
- Carroll, R. J. (1982). "Adapting for heteroscedasticity in linear models." *Ann. Statist.*, **4**, 1224—1233.
- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. Tukey (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.
- Chan, T., G. Golub, and R. LeVeque (1983). "Algorithms for computing the sample variance: Analysis and recommendations." *Am. Statist.*, **37**, 242—47.
- Chen, C. F. (1983). "Score tests for regression models." *J. Am. Statist. Assoc.*, **78**, 158—161.
- Clapham, A. W. (1934). *English Romanesque Architecture After the Conquest*. Oxford: Clarendon Press.
- Cook, R. D. (1977). "Detection of influential observations in linear regression" *Technometrics*. **19**, 15—18.
- Cook, R. D. (1979). "Influential observations in linear regression." *J. Am. Statist. Assoc.*, **74**, 169—174.
- Cook, R. D. (1984). "Comment on Belsley (1984)" *Am. Statist.*, **38**, 78—79.
- Cook, R. D. and J. O. Jacobsen (1978). "Analysis of 1977 West Hudson Bay snow goose surveys." Unpublished report, Canadian Wildlife Service.
- Cook, R. D. and P. Prescott (1981). "Approximate significance levels for detecting outliers in linear regression." *Technometrics*. **23**, 59—64.
- Cook, R. D. and Wang, P. C. (1983). "Transformations and influential cases in regression." *Technometrics*, **25**, 337—344.
- Cook, R. D. and S. Weisberg (1980). "Characterizations of an empirical influence function for detecting influential cases in regression." *Technometrics*.

22, 495—508.

Cook, R. D. and S. Weisberg (1982a). *Residuals and Influence in Regression*. London: Chapman Hall.

Cook, R. D. and S. Weisberg (1982b). "Criticism in regression." in Leinhardt, S. (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass. Chapter 8.

Cook, R. D. and S. Weisberg (1983). "Diagnostics for heteroscedasticity in regression." *Biometrika*, **70**, 1—10.

Copas, J. B. (1983). "Regression, prediction and shrinkage (with discussion)." *J. Roy. Statist. Soc. Ser. B.* **45**, 311—354.

Cox, D. R. (1958). *The Planning of Experiments*. New York: Wiley.

Cox, D. R. (1970). *The Analysis of Binary Data*, London: Chapman Hall.

Cox, D. R. (1977). "Nonlinear models, residuals and transformations." *Math. Operationsforsch. Statist. Ser. Statist.*, **8**, 3—22.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. London: Chapman Hall.

Daniel, C. (1976). *Applications of Statistics to Industrial Experiments*. New York: Wiley.

Daniel, C. and F. Wood (1980). *Fitting Equations to Data*, 2nd ed. New York: Wiley.

Dalziel, C. F., J. B. Lagen, and J. L. Thurston (1941) "Electric Shocks." *Trans. IEEE*, **60**, 1073—1079.

Dawies, R. B. and B. Hutton (1975). "The effects of errors in the independent variables in linear regression." *Biometrika*, **62**, 383—391.

Dempster, A. P. (1973). "Alternatives to least squares in multiple regression," in Kale, D. G. and R. P. Gupta (eds.) *Multivariate Statistical Inference*. Amsterdam: North-Holland.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *J. Roy. Statist. Soc. Ser. B*, **39**, 1—38.

Demster, A. P., M. Schatzoff, and N. Wermuth (1977). "A simulation study of alternatives to ordinary least squares (with discussion)." *J. Am. Statist. Assoc.*, **72**, 77—104.

- Devlin, S. J. R. Gnanadesikan, and J. Kettenring (1975). "Robust estimation and outlier detection with correlation coefficients." *Biometrika*, **62**, 531—546.
- Dixon, W. J. (ed.) (1983). *BMDP Biomedical Computer Programs*. Berkeley: University of California.
- Dongarra, J. , J. P. Bunch, C. B. Moler, and G. W. Stewart (1979). *The LINPACK Users Guide*. Philadelphia: Society for Industrial and Applied Mathematics.
- Draper, N. and W. G. Hunter (1969). "Transformations: Some examples revisited." *Technometrics*, **11**, 23—40.
- Draper, N. and H. Smith (1966). *Applied Regression Analysis*. New York: Wiley.
- Draper, N. R. and R. C. Van Nostrand (1978). "Ridge regression: Is it worthwhile?" Technical Report No. 501, Dept. of Statistics, University of Wisconsin.
- Draper, N. R. and R. C. Van Nostrand (1979). "Ridge regression and James-Stein estimation: Review and comments." *Technometrics*, **21**, 451—465.
- Durbin, J. and G. S. Watson (1950). "Testing for serial correlation in least squares regression I." *Biometrika*, **58**, 1—19.
- Durbin, J. and G. S. Watson (1951). "Testing for serial correlation in least squares regression II." *Biometrika*, **38**, 159—178.
- Durbin, J. and G. S. Watson (1971). "Testing for serial correlation in least squares regression III." *Biometrika*, **58**, 1—19.
- Efron, B. and C. Morris (1973). "Stein's estimation rule and its competitors—An empirical Bayes approach." *J. Am. Statist. Assoc.*, **68**, 117—130.
- Efron, B. and C. Morris (1975). "Data analysis using Stein's estimator and its generalizations." *J. Am. Statist. Assoc.*, **70**, 311—319.
- Ehrenberg, A. S. C. (1981). "The problem of numeracy." *Am. Statist.*, **35**, 67—71.
- Ehrenberg, A. S. C. (1982). "How good is best?" *J. Roy. Statist. Soc. Ser. A*, **145**, 364—366.
- Ezediel, M. and F. A. Fox (1959). *Methods of Correlation and Regression*

Analysis, New York: Wiley.

- Fienberg. S. E. (1977). *The Analysis of Cross Classified Categorical Data*, 2nd ed. Cambridge: MIT Press.
- Finkelstein. M. O. (1980). "The judicial reception of multiple regression studies in race and sex discrimination cases" *Columbia Law Rev.*, **80**, 734—757.
- Forbes. J. D. (1857). "Further experiments and remarks on the measurement of heights by the boiling point of water." *Trans. R. Soc. Edinburgh*, **21**, 135—143.
- Freedman, D. (1983). "A note on screening regression equations." *Am. Statist.*, **37**, 152—155.
- Freedman. D. and S. Peters (1984). "Bootstrapping a regression equation: Some empirical results." *J. Am. Statist. Assoc.*, **79**, 97—106.
- Freeman. M. F. and J. W. Tukey (1950). "Transformations related to the angular and the square root." *Ann. Math. Statist.*, **21**, 607—611.
- Furnival. G. and R. Wilson (1974). "Regression by leaps and bounds." *Technometrics*, **16**, 499—511.
- Garside, M. J. (1971). "Some computational procedures for the best subset problem." *Appl. Statist.*, **20**, 8—15.
- Gauss, C. F. (1821, collected works 1873). "Theoria combinationis observationum erroribus minimis obnoxiae," in Werde 4, Section 35, Gottengen.
- Geisser. S. (1980). "A predictivistic primer." in Zellner, A. (ed.) *Bayesian Analysis in Econometrics and Statistics*. Amsterdam: North-Holland.
- Geisser, S. and W. F. Eddy (1979). "A predictive approach to model selection." *J. Am. Statist. Assoc.*, **74**, 153—160.
- Gentry, A. H. and J. Lopez-Parodi (1980). "Deforestation and increased flooding in the upper Amazon." *Science*. **210**, 1354—1356.
- Gnanadesikan, R. (1977), *Methods for Statistical Analysis of Multivariate Data*. New York: Wiley.
- Goldstein, M. and A. F. M. Smith (1974). "Ridge type estimators for regression analysis." *J. Roy. Statist. Soc. Ser. B*, **36**, 284—301.
- Golub. G. H. (1965). "Numerical methods for solving linear least squares problems." *Numerical Mathematics*, **7**, 206—216.

- Goodnight, J. H. (1979). "A tutorial on the SWEEP operator." *Am. Statist.* , **33**, 149—158.
- Gould, S. J. (1966). "Allometry and size in ontogeny and phylogeny." *Biol. Rev.* , **41**, 587—640.
- Gould, S. J. (1973). "The shape of things to come." *Syst. Zool.* , **22**, 401—404.
- Gray, J. and R. Ling (1984). "K-clustering as a detection tool for influential subsets in regression." *Technometrics* , **26**, 304—318.
- Graybill, F. A. (1969). *Introduction to Matrices with Statistical Applications*. Belmont, CA: Wadsworth.
- Hald, A. (1960). *Statistical Theory with Engineering Applications*. New York: Wiley.
- Hartley, H. O. and R. R. Hocking (1971). "The analysis of incomplete data." *Biometrics* , **27**, 783—808.
- Hawkins, D. M. (1980). *Identification of Outliers*. London: Chapman Hall.
- Hawkins, D. M. , D. Bradu, and G. Kass (1984). "Location of several outliers in multiple regression using elemental sets." *Technometrics* , **26**, 197—208.
- Henderson, H. V. and S. R. Searle (1981). "On deriving the inverse of a sum of matrices." *SIAM Rev.* , **23**, 53—60.
- Hernandez, F. and R. A. Johanson (1980). "The large sample behavior of transformations to normality." *J. Am. Statist. Assoc.* , **75**, 855—861.
- Hinkley, D. V. and G. Runger (1984). "The analysis of transformed data (with discussion)." *J. Am. Statist. Assoc.* , **79**, 302—319.
- Hoaglin, D. C. and R. Welsch (1970). "The hat matrix in regression and ANOVA." *Am. Statist.* , **32**, 17—22.
- Hocking, R. R. and R. N. Leslie (1967) . "Selection of the best subset in regression analysis." *Technometrics* , **2**, 531—540.
- Hocking, R. R. and O. J. Pendleton (1982). "The regression dilemma." *Commun. Statist. Ser. A.* **12**, 497—527.
- Hodges, S. D. and P. G. Moore (1972). "Data uncertainties and least squares regression." *Appl. Statist.* , **21**, 185—195.
- Hoerl, A. E. and R. W. Kennard (1970a). "Ridge regression: biased estimation

- for nonorthogonal problems." *Technometrics*, **12**, 55—67.
- Hoerl, A. E. and R. W. Kennard (1970b). "Ridge regression: Applications to nonorthogonal problems." *Technometrics*, **12**, 69—82.
- Holt, D. and A. J. Scott (1981). "Regression analysis using survey data." *The Statistician*, **30**, 169—178.
- Householder, A. S. (1958). "The approximate solution of matrix problems." *J. Assoc. Comput. Mach.*, **5**, 204—243.
- Huber, P. J (1981). *Robust Statistics*. New York: Wiley.
- IMSL (1979). *The IMSL Library*. Houston: International Mathematics and Statistics Library.
- James, W. and C. Stein (1961). "Estimation with quadratic loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California.
- Jenson, R. (1977). "Evinrude's computerized quality control productivity." *Qual. Prog.*, **10**, 12—16.
- Jevons, W. S. (1868). "On the condition of the gold coinage of the United Kingdom. with reference to the question of international currency." *J. [Roy.] Statist. Soc.*, **31**, 426—464.
- John. P. W. M (1971). *Statistical Design and Analysis of Experiments*. New York: Macmillan.
- Johnson, M. P. and P. H. Raven (1973). "Species number and endemism: The Galapagos Archipelago revisited" *Science*. **179**, 893—895.
- Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kennard. R. W. (1971). "A note on the C_p statistic." *Technometrics*, **13**, 899—900.
- Kennedy. W. and T. Bancroft (1971). "Model building for prediction in regression based on repeated significance tests." *Ann. Math. Statist.*, **42**, 1273—1284.
- Kennedy. W. and J. Gentle (1980). *Statistical Computing*. New York: Marcel Dekker.
- Kerr, R. A. (1982). "Test fails to confirm cloud seeding effect." *Science*, **217**, 234—236.

- Krasker, W. S. and R. E. Welsch (1982). "Efficient bounded influence regression estimation." *J. Am. Statist. Assoc.* **77**, 595—604.
- LaMotte, L. R. and R. R. Hocking (1970). "Computational efficiency in the selection of variables." *Technometrics*, **12**, 83—93.
- Land, C. E (1974). "Confidence interval estimation for means after data transformations to normality." *J. Am. Statist. Assoc.*, **69**, 795—802 (correction. *ibid.* 71. 255).
- Landwehr, J. , D. Pregibon, and A. Shoemaker (1984). "Graphical methods for assessing logistic regression models (with discussion)." *J. Am. Statist. Assoc.*, **79**, 61—83.
- Larson, W. A. and S. A. McCleary (1972). "The use of partial residual plots in regression analysis," *Technometrics*, **14**, 781—790.
- Lawley, D. N. (1943). "A note on Karl Pearson's selection formulae." *Proc. R. Soc. Edinburgh.* **62**, 28—30.
- Lindgren, B. L (1976). *Statistical Theory*, 3rd ed. New York: Macmillan.
- Little, R. J. A. (1979), "Maximum likelihood inference for multiple regression with missing values: A simulation study." *J. Roy. Statist. Soc Ser. B*, **41**, 76—87.
- Longley, J. W. (1967). "An appraisal of least squares programs for the electronic computer from the point of view of the user." *J. Am. Statist. Assoc.*, **62**, 819—841.
- Louis, T. A. (1982). "Finding the observed information matrix when using the EM algorithm." *J. Roy. Statist. Soc. Ser. B.* **44**, 226—233.
- Madansky, A. (1959). "The fitting of straight lines when both variables are subject to error." *J. Am. Statist. Assoc.* **54**, 173—206.
- Mallows, C. L. (1973) "Some comments on C_p " *Technometrics*, **15**, 661—676.
- Mansfield, E. R. , J. T. Webster, and R. F. Gunst (1977). "An analytic variable selection technique for principal component regression." *Appl. Statist.*, **26**, 34—40.
- Mantel, N. (1970). "Why stepdown procedures in variable selection?" *Technometrics*, **12**, 621—625.
- Marler, G. D. (1969). *The Story of Old Faithful*. W. Yellowstone, Wyoming:

Yellowstone Library and Museum Association.

Marquardt, D. W. (1970). "Generalized inverses, ridge regression and biased linear estimation." *Technometrics*, **12**, 591—612.

Marquardt, D. W. and R. Snee (1975). "Ridge regression in practice." *Am. Statist.*, **12**, 3—19.

McCullagh, P. and J. Nelder (1983). *Generalized Linear Models*. London: Chapman and Hall.

Miller, R. (1981). *Simultaneous Inference*, 2nd ed. New York: Springer.

Moore, J. A. (1975). "Total Biochemical Oxygen Demand of Animal Manures." Unpublished Ph. D. dissertation, University of Minnesota.

Morgan, J. A. and J. F. Tartar (1972). "Calculation of residual sum of squares for all possible regressions." *Technometrics*, **14**, 317—325.

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Linear Regression*. Reading, MA: Addison-Wesley.

Myers, R. H. (1971). *Response Surface Methodology*. Boston: Allyn and Bacon.

Narula, S. C. and J. W. Wellington (1977). "Prediction, Linear regression and minimum sum of relative errors." *Technometrics*, **19**, 185—190.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models." *J. Roy. Statist. Soc. Ser. A*, **135**, 370—384.

Noll, S. L., P. E. Waibel, R. D. Cook, and J. A. Witmer (1984). "Biopotency of methionine sources for young turkeys." *Poult. Sci.* **63**, 2458—2470.

Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." *Technometrics*, **17**, 431—445.

Orchard, T. and M. A. Woodbury (1972). "A missing information principle: Theory and applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California.

Pearson, K. (1930). *Life and Letters and labours of Francis Galton*, Vol. IIIa. Cambridge, UK: Cambridge University Press.

Pereira, B. (1977). "Discriminating among separate models: A bibliography." *Int. Statist. Rev.*, **45**, 163—172.

- Picard, R. R. and R. D. Cook (1984). "Cross-validation of regression models." *J. Am. Statist. Assoc.*, **79**, 575—583.
- Pregibon, D. (1980). "Goodness of link tests for generalized linear models." *Appl. Statist.*, **29**, 15—24.
- Pregibon, D. (1981). "Logistic regression diagnostics." *Ann. Statist.*, **9**, 705—724.
- Ralston, A. (1960). *Mathematical Methods for Digital Computers*. New York: Wiley.
- Ratkowsky, D. A. (1983). *Nonlinear regression modeling*. New York: Marcel Dekker.
- Rencher, A. C. and F. C. Pun (1980). "Inflation of R^2 in best subset regression." *Technometrics*, **22**, 49—53.
- Renshaw, E. (1958). "Scientific appraisal." *Natl. Tax J.*, **11**, 314—322.
- Rolph, J. E. (1976). "Choosing shrinkage estimators for regression problems." *Commun Statist. Ser. A*, **5**, 789—802.
- Royston, J. P. (1982a). "An extension of Shapiro and Wilk's W test for normality to large samples." *Appl. Statist.*, **31**, 115—124.
- Royston, J. P. (1982b). "Expected normal order statistics (exact and approximate), Algorithm AS177." *Appl. Statist.*, **31**, 161—168.
- Royston, J. P. (1982c). "The W test for normality. Algorithm AS181." *Appl. Statist.*, **31**, 176—180.
- Rubin, D. B. (1974). "Characterizing the estimation of parameters in incomplete data problems." *J. Am. Statist. Assoc.*, **69**, 467—474.
- Rubin, D. B. (1976). "Inference and missing data." *Biometrika*, **63**, 581—592.
- Rubin, D. B. (1980). "Using empirical Bayes' techniques in the Law School validity studies." *J. Am. Statist. Assoc.*, **67**, 801—827.
- Ryan, T., B. Joiner and B. Ryan (1985). *Minitab Student Handbook*, 2nd ed. Belmont, CA: Duxbury.
- Sandberg, J. S., M. J. Basso, and B. A. Okin (1978). "Winter rain and summer ozone: A predictive relationship." *Science*, **200**, 1051—1054.
- Saw, J. G. (1966). "A conservative test for concurrence of several regression lines and related problems." *Biometrika*, **53**, 272—275.

- Schatzoff, M., S. Fienberg and R. Tsao (1968). "Efficient calculations of all possible regressions." *Technometrics*, **10**, 769—779.
- Scheffe, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Sclove, S. (1968). "Improved estimators for coefficients in linear regression." *J. Am. Statist. Assoc.*, **63**, 596—606.
- Sclove, S. (1972). "(Y vs X) or (log Y vs X)?" *Technometrics*, **14**, 391—403.
- Searle, S. R. (1971). *Linear Models*. New York: Wiley.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.
- Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.
- Shapiro, S. S. and M. B. Wilk (1965). "An analysis of variance test for normality (complete samples)." *Biometrika*, **52**, 591—611.
- Smith, B. T., J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler (1976). *Matrix Eigensystem Routines—EISPACK Guide*, 2nd ed. Lecture Notes in Computer Science, No. 6. New York: Springer.
- Smith, G. and F. Campbell (1980). "A critique of some ridge regression methods (with discussion)." *J. Am. Statist. Assoc.*, **75**, 74—103.
- Snee, R. D. (1977). "Validation of regression models. Methods and examples." *Technometrics*, **19**, 415—428.
- Sprent, P. (1972). "The mathematics of size and shape." *Biometrics*, **28**, 23—37.
- Stein, C. (1956) "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." in *Third Berkeley Symposium on Probability and Statistics*, Vol. 1. Berkeley: University of California.
- Stewart, G. W. (1974). *Introduction to Matrix Computations*. New York: Academic.
- Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions (with discussion)." *J. Roy. Statist. Soc. Ser. B*, **36**, 111—147.
- Strawderman, W. (1978). "Minimax adaptive generalized ridge regression estimators." *J. Am. Statist. Assoc.*, **73**, 623—627.
- Thisted, R. (1978a). "On generalized ridge regression." Technical Report No. 57, Dept. of Statistics, University of Chicago.

- Thisted, R. (1978b). "Multicollinearity, information and ridge regression." Technical Report No. 58, Dept. of Statistics, University of Chicago.
- Titterton, D. M. (1978). "Estimation of correlation coefficients by elliptical trimming." *Appl. Statist.*, **27**, 227—234.
- Tuddenham, R. D. and M. M. Snyder (1954). "Physical growth of California boys and girls from birth to age 18." *Calif. Publ. Child Develop.*, **1**, 183—364.
- Weisberg, H., E. Beier, H. Brody, R. Patton, K. Raychaudhari, H. Takeda, R. Thern, and R. Van Berg (1978). "S-dependence of proton fragmentation by hadrons. II. Incident laboratory momenta 30—250 GeV/c." *Phys. Rev. D*, **17**, 2875—2887.
- Weisberg, S. (1981). "A statistic for allocating C_p to individual cases." *Technometrics*, **23**, 27—31.
- Weisberg, S. (1983). "Principles for regression diagnostics and influence analysis, discussion of a paper by R. R. Hocking." *Technometrics*, **25**, 240—244.
- Weisberg, S. and C. Bingham (1975). "An analysis of variance test for normality suitable for machine calculation." *Technometrics*, **17**, 133.
- West, D. H. D. (1979). "Updating mean and variance estimates: An improved method." *Commun. ACM*, **22**, 532—535.
- Wilk, M. B. and R. Gnanadesikan (1968). "Probability plotting methods for the analysis of data." *Biometrika*, **55**, 1—17.
- Williams, E. (1959). *Regression Analysis*. New York: Wiley.
- Wilm, H. G. (1950). "Statistical control in hydrologic forecasting." *Res. Notes*, **61**, Pacific Northwest Forest Range Experiment Station, Oregon.
- Wood, F. S. (1973). "The use of individual effects and residuals in fitting equations to data." *Technometrics*, **15**, 677—695.
- Woodley, W. L., J. Simpson, R. Biondini, and J. Berkelcy (1977). "Rainfall results 1970 75: Florida arca cumulus experiment." *Science*, **195**, 735—742.

标题索引

A

- Added variable plot 附加变量图, 39—43, 55—58, 134, 149
- Additive effects 可加效应, 178
- Adjusted R^2 校正的 R^2 , 229
- Amazon River 亚马逊河, 32
- Analysis of covariance 协方差分析, 193
- Analysis of variance 方差分析, 15—19, 50—55
- Asymptote 渐近值, 278
- Atkinson's score test Athinson 得分检验, 157—159
- Backward elimination 后向消去, 221—226
- Bayes method: 贝叶斯方法:
 James - Stein estimates 詹姆斯—斯坦估计量, 275
 Prediction 预测, 253

B

- Bayes rules 贝叶斯法则, 273
- Biased regression 有偏回归, 270—276
- Binomial distribution 二项式分布, 284—290
- BMDP programs BMDP 程序, 193, 230, 254, 266, 274, 280, 282
- Bonferroni inequality 邦弗伦尼不等式, 121, 223
 t values t 值, 149
- Box and Cox procedure Box 和 Cox 方法, 153—159, 166, 168, 169, 174

Box and Tidwell procedure

Box 和 Tidwell 方法, 159—164

C

Case statistics, see Diagnostic statistics

案例统计量, 见诊断统计量

Causal relationships

因果关系, 59, 69, 231—233, 244—246

Censored data

删截数据, 260

Centering

中心化, 173, 292—294

Central composite design

中心复合设计, 174

Cholesky decomposition

乔勒斯基分解, 50, 307

Coefficient of determination

测定系数, 19, 51

Collinearity

共线性, 205—212, 222, 267

Comparison of regression lines

比较回归直线, 187—194

Computation,

计算, 33, 50, 64—67, 109, 297—299, 307—308

all possible regressions

所有可能回归, 230

cross validation and PRESS

相互证实和 PRESS, 253, 254

leaps and bounds

高速, 230

polynomial regression

多项式回归, 172, 173

Condition number

条件数, 209

Confidence;

置信:

interval

区间, 20, 21

regions

区域, 101, 102

Cook's distance

Cook 距离, 122—130

Correlated errors

相关误差, 167—168

Correlation, Partial

相关, 偏, 42—43

Correlation coefficient

相关系数, 9, 295

Criticism

评判, 133—134

Cross product terms

叉积项, 173

Cross validation

相互证实, 229, 244, 247, 249

Computations

计算, 253—254

D

Dependent variable, See Response

Deviance

Diagnostic methods

statistics

Dummy variables

Durbin - Watson statistic

应变变量, 见响应

偏差, 287

诊断方法, 133—166

统计量, 106

虚拟变量, 177—194

Durbin—Watson 统计量, 168

E

Eigenvalues and eigenvectors

Eispack,

EM algorithm

Envelope, for probability plots

Error function

Errors

Comparison to residuals

distribution

fixed and random

multiplicative

Experiments

Exponential family

Extrapolation

特征值和特征向量, 196

Eispack, 197

EM 算法, 264

信封, 对于概率图, 166—168

误差函数, 289

误差, 4

与残差相比较, 7

分布, 6

固定和随机, 4—6, 293—294

乘法的, 147

试验, 74

指数族, 290

外推, 242

F

Fitted values

Forward selection

F-tests

relation to t

Gamma distribution

Gaussian elimination, See Sweep
algorithm

Gauss - Markov theorem

拟合值, 8, 23

前向选择, 222, 224—226

F-检验, 18, 51—52, 100—101, 106

和 t 有关, 22, 53—54

伽玛分布, 290

高斯消去法, 见扫描算法

高斯—马尔科夫定理, 14

- Generalized least squares
Generalized linear models

Generalized ridge regression
Geometric mean
GLIM
- H**
- Hat matrix
- I**
- Identity matrix
IMSL
Incomplete data
Independent variable
Index plots
Indicator variable, see Dummy variables
Influence
Intercept
Interpolation
Inverse regression
Iteratively reweighted least squares
- J**
- James - Stein estimators
- L**
- Lack of fit tests
 approximate
 Variance
 known
 unknown
- 广义最小二乘, 85—87
广义线性模型, 148, 277, 286, 288—291
广义岭回归, 274
几何平均, 154—155
GLIM, 193, 287, 288, 290
- 帽子矩阵, 113, 114—116
- 单位矩阵, 300
IMSL, 197, 230
不完全数据, 258—266
自变量, 2
索引图, 135
指示变量, 见虚拟变量
- 影响, 110, 122—130
截距, 5
内插, 242, 250—252
逆回归, 109
迭代重加权最小二乘, 90
- 詹姆斯—斯坦估计量, 275
- 拟合失真检验:
 近似, 99
 方差:
 已知, 92—94
 未知, 94—100

Lagged variables
Leaps and bounds
Least Squares estimates

generalized
iterative reweighted
means and variances
nonlinear models
weighted

Leverage, See Potential

Likelihood ratio test

Linear dependence

Linear independence

Linear transformation

Link function

Linpack

Logistic function

Logistic regression

Logit transformation

Log - likelihood

M

Mahalanobis distance

Mallows' Cp,

in prediction

Matrix algebra

Matrix plot

Maximum likelihood:

estimates

incomplete data

Mean square

Mean square error

of prediction

• 344 •

滞后变量, 249

高速, 230

最小二乘估计, 7—15, 44—46,

243, 295, 307

广义, 84—86

迭代重加权, 90

均值和方差, 296—297

非线性模型, 278

加权, 84—91, 98—99

杠杆, 见位势

似然比检验, 282

线性相关, 72

线性无关, 72

线性变换, 70—71

连接函数, 289

Linpack, 50, 307

逻辑斯帝函数, 147

逻辑斯帝回归, 277, 283—288

Logit 变换, 285

对数似然, 154—155

马哈拉诺比斯距离, 116

Mallows' Cp, 228—230, 308—309

在预测, 243

矩阵代数, 293—299

矩阵图, 134—135

极大似然:

估计, 14, 133, 286—287

不完全数据, 264

均方, 12

均方误差, 243—244

预测的, 227—230, 308—309

- Measurement error
 - M-estimates
 - Minimum covering ellipsoid
 - Minitab,
 - Missing data
 - Missing observation correlation
 - Missing at random
 - Mixture experiments
 - Model expansion
 - Model selection
 - prediction
 - Model varidation
 - Multicollinearity, see Collinearity
 - Multiple correlation coefficient
 - Multiplicative errors
 - Nonconstant variance
 - Nonlinearity
 - Nonlinear least squares
 - computations
 - models
 - regression
 - Normal distribution
 - Collinearity
 - Normal equation
- O**
- Observables
 - Odds ratio
 - Order statistics
 - Orthogonal matrices
 - Orthogonal projection
 - Orthogonal variables
 - 测量误差, 79—82
 - M-估计, 268—269
 - 最小覆盖椭圆, 252
 - Minitab, 197, 254
 - 遗漏数据, 258—264
 - 遗漏观测值相关, 264—265
 - 随机遗漏, 258—260
 - 混料试验, 58
 - 模型扩展, 134, 153, 159—160
 - 模型选择, 212—235
 - 预测, 253—254
 - 模型的确认, 243—244
 - 复共线性, 见共线性
 - 复相关系数, 19—20, 51—52
 - 乘法型误差, 147
 - 非常数方差, 136, 139—146
 - 非线性, 136, 146—153
 - 非线性最小二乘:
 - 计算, 279
 - 模型, 147
 - 回归, 277—283
 - 正态分布, 6, 13, 75—79, 164—168
 - 共线性, 208
 - 正规方程, 295
 - 可观测的, 243
 - 优比, 285
 - 次序统计量, 165
 - 正交矩阵, 131
 - 正交投影, 114
 - 正交变量, 43

- Outliers
- Overparameterized models
- P**
 - Parallel regressions
 - Parameters
 - interpretation
 - Partial correlation
 - Partial residual plots
 - Picket fence
 - Plots:
 - added variable
 - index
 - matrix
 - Partial residual
 - Probability
 - envelope
 - residual plus component
 - residuals in
 - scatter plots
 - usefulness
 - Poisson distribution
 - Polynomial regression
 - Potential
 - simple repression
 - Power transformations
 - Prediction
 - Predictive distribution
 - predictive intervals
 - Predictor, see Independent variable
 - PRESS
 - computation
- 异常值, 118—122, 136, 267
- 参数过多的模型, 71
- 平行回归, 181, 187, 188
- 参数, 4, 7
 - 解释, 68—74
- 偏相关, 42—43
- 偏残差图, 56, 149
- 尖桩篱笆, 207—208
- 图:
 - 附加变量, 39—43, 55—58, 134, 149
 - 索引, 135
 - 矩阵, 134—135
 - 偏残差, 56, 149
 - 概率, 164—168
 - 信封, 166—168
 - 残差加分量, 57
 - 残差, 136—139
 - 散点图, 2, 134—139
 - 有用性, 110—112
- 泊松分布, 290
- 多项式回归, 93, 159, 172—176
- 位势, 110, 115, 133
 - 简单回归, 116
- 幂变换, 153—155
- 预测, 22—24, 240—257, 297
 - 预测分布, 252—253
 - 预测区间, 23
- 预报因子, 见自变量
- PRESS, 229, 244, 247, 249
 - 计算, 253

Principal components

regression

Probability plots

Pure error

P-values

Q

QR factorization

R

Random variables, means and variances

Range of fitting

Range of validity

Rank deficiency

Rankits, see Probability plots

Rank of model or matrix

Regression, origin of the term

Regression through the origin

Residual mean square

Residual plus component plots

Residuals

correlations

plots

studentized

supernormality

Residual sum of squares

Response

Response surfaces

Reversion

Ridge regression

canonical form

generalized

主成分, 195, 196—199

回归, 274

概率图, 164—168

纯误差, 95

P-值, 19

QR 因子分解, 63, 131, 195, 307

随机变量, 均值和方差, 294

拟合范围, 241

有效范围, 242

秩不足, 71

Rankits, 见概率图

模型或矩阵的秩, 71, 304

回归, 术语的起源, 103—106

通过原点的回归, 31, 58

残差均方, 12

残差加分量图, 57

残差, 24, 110, 113—122, 133

相关, 114

图, 136

学生化, 117, 120, 121

超正态性, 164

残差平方和, 10

响应, 2

响应曲面, 174

恢复, 104

岭回归, 267, 272—276

典则形式, 272

广义, 274

Robust estimates
Rounding error
Royal Statistical Society

稳健估计, 267, 268—269
舍入误差, 297—299
(英国) 皇家统计协会, 193

S

Sample reuse, see PRESS
Sampling models
Scalar
Scale invariance
Scatter plot
Score test
 nonconstant variance
 transformations of predictors
 transformations of response
Significant test
Simulation
Singular models
Slope
 Comparison between groups
Square root of a matrix
Standard error of regression
Standardized coefficients
Stepwise regression
 criticism of
Studentization of residuals
 external
 internal
Subset selection
Supernormality of residuals
Sweep algorithm

重复使用样本, 见 PRESS
抽样模型, 74—79
标量, 301
尺度不变性, 32, 194
散点图, 4, 134—139
得分检验:
 非常数方差, 141—146
 变换自变量, 159—162
 变换响应变量, 157—159
显著性检验, 19
模拟, 82
退化模型, 72
斜率, 4
 组间比较, 187—194
矩阵的平方根, 85
回归的标准误, 13
标准化系数, 195
逐步回归, 221—226
 评判, 226—227
学生化残差, 133
 外部的, 119—120
 内部的, 117
子集选择, 271
残差的超正态性, 164
扫描算法, 50, 65—67, 109

T

Tolerance

容差, 66, 222, 223

Transformations

Atkinson's score Atkinson
Box and Cox method
Box and Tidwell method

Predictors

response
variance stabilizing
zeroes or negative values
t value

relation to F

U

Updating method

regression equations

V

Variable constants

Variable selection

Variance inflation factor

Variance stabilizing transformations

Vector

random

W

Weighted least squares

Symbols, specific definitions

AH, 18—19
corr (), 295
cov (), 294
 C_p , 228

变换:

得分, 157—159
Box 和 Cox 方法, 153—157
Box 和 Tidwell 方法, 159—163
预报变量, 自变量, 159—163
响应, 153—157
方差稳定化, 139—140
零或负值, 141
t 值, 21—22, 53—54, 120, 149, 193
和 F 有关, 22, 54

更新法, 297—299

回归方程, 308

变量常数, 241

变量选择, 212—236

方差扩大因子, 207

方差稳定化变换, 139—140

向量, 299

随机, 305—306

加权最小二乘, 85—88

符号, 特殊定义的:

D_i , 124
DFFITS, 130
 E (), 305
 e , 43

\hat{e} , 46	SXX , 9
e_i , 6	SXY , 9
\hat{e}_i , 7	S_{XY} , 9
F_p , 228	$SY Y$, 9
$F(\alpha, v_1, v_2)$, 18	T , 45
G , 158	t_D , 158
GLM , 288	t_i , 120
$GM(y)$, 154	$t(\alpha, n)$, 20
H , 113	$\text{var}(\quad)$, 15
h_{ii} , 115	v_{ii} , 见 h_{ii}
h_{ij} , 115	VIF_j , 207
J_p , 228	W , 86—87
$L(\lambda)$, 154	w_i , 86—87
MCE , 252	X_{\cdot} , 22
NH , 18—19	Y^{λ} , 154
$N(\mu, \sigma^2)$, 44	\hat{y}_i , 8
NID , 6	y_{\cdot} , 22
P' , 44	\sim
$PRESS$, 229	y_{\cdot} , 22
R^2 , 19	\hat{y} , 46
\bar{R}^2 , 229	Z^{λ} , 155
r_i , 117	β_0 , 4
r_{XY} , 9	β_1 , 4
$RSS(\beta_0, \beta_1)$, 10	β , 43
$RSS(\beta)$, 45	$\hat{\beta}$, 45
RSS , 10	β^* , 46
S , 142	k , 209
SD , 9	λ , 141
$se(\quad)$, 15	λ , 153
$sefit$, 23, 50	σ^2 , 6
$sepred$, 23, 50	$\chi^2(v)$, 13
SS_{reg} , 16	