

目 录

符号说明	1
第一章 线性回归的基本方法	1
§1.1 概说	1
§1.2 参数的点估计	8
§1.3 线性假设的检验	24
§1.4 参数的区间(域)估计	42
§1.5 预测和校准问题	54
§1.6 自变量为随机变量的情况	63
第二章 回归诊断	91
§2.1 引言	91
§2.2 残差	95
§2.3 残差图	106
§2.4 方差稳定化变换	121
§2.5 正态化变换	126
§2.6 影响函数	134
§2.7 异常点	146
第三章 自变量选择	151
§3.1 引言	151
§3.2 变量选择对估计、预测的影响	153
§3.3 基于残差平方和的准则	163
§3.4 C_p -准则	173
§3.5 预测平方和准则	189

§3.6	AIC准则	193
§3.7	计算方法: 扫描运算和Gauss消去法	196
§3.8	所有可能子集回归	202
§3.9	逐步型回归	212
第四章	回归系数的有偏估计	217
§4.1	复共线性	218
§4.2	岭估计	226
§4.3	广义岭估计	242
§4.4	Stein估计	256
§4.5	主成分估计	259
§4.6	特征根估计	268
§4.7	小结	277
第五章	其它方法	279
§5.1	权函数估计法	280
§5.2	广义线性模型	299
§5.3	截尾回归	328
§5.4	M估计法	341
参考文献		356
名词索引		363

第一章 线性回归的基本方法

§1.1 概 说

设想一个问题牵涉变量 X 和 Y ，为说明的方便，可把问题设想成通过 X 之值去预测 Y 之值，相应地我们称 X 为自变量而称 Y 为因变量。不过，这些名称只是反映 X 、 Y 在问题中的地位的不同，而并不意味着二者之间存在着因果关系。本书总假定因变量是一维的。

在许多情况下，变量 X 与 Y 有一定关系，但又没有密切到可以通过 X 唯一决定 Y 的程度。人的身高(X)体重(Y)即为一例。要用数学的方法通过 X 去预测 Y ，首先就需要找到一种数学方法以描述它们的关系，这就必然含有假定的成份。我们的要求是：在已知 X 值的条件下，变量 Y 取值的不确定性可以通过一定的概率分布去描述。这就是说，我们要求因变量 Y 在概率论的意义下是一个随机变量，在 X 的值已知为 x 时，有一定的条件分布 $F(\cdot | X = x)$ ，或简记为 $F(\cdot | x)$ 。这个要求使我们能够用统计方法去研究 X 和 Y 的关系，这种研究也就构成本书的主题——回归分析。

从概率论的观点看，概率分布虽是随机变量的统计性质的最完整的描述，但为着某种特定的应用目的，更有关系且更方便的往往是这概率分布的某些特征数字。例如，分布的均值就是一个

最重要的特征数字，以 $f(x)$ 记分布 $F(\cdot|x)$ 的均值，即

$$f(x) = \int_{-\infty}^{\infty} yF(dy|x) = E(Y|X=x)$$

我们则把这个函数 f 称为 Y 对 X 的**回归函数**，更确切地应称为**均值回归函数**。因为，我们也可以考察其他的数字特征，例如中位数，相应地就可以引进“中位回归函数”。在本书中，我们仅限于考虑均值回归函数。回归函数的一个重要应用，就是本节开头提到的那个由 X 预测 Y 的问题：若知道了回归函数 f ，则在已知 X 时，可以用 $f(X)$ 去预测 Y ，这个预测在均方误差最小的意义下是最优的。当然，回归函数的作用，以及整个回归分析的目的，都不只是预测问题。

统计学还有一类很重要的应用课题——相关分析问题。它的任务也是研究变量之间的关系的，因而与回归分析有密切的关系。二者的差别主要有两点：1.在回归分析中，有一个变量，即因变量 Y 处在特殊的地位，而在相关分析中，各变量的地位都是平等的。这意味着二者在研究的着重方面，所引出的统计推断问题有很大的不同。2.在回归分析中，因变量 Y 是随机的，但自变量 X 可以是随机，也可以是非随机。而在相关分析中，所涉及的变量全是随机的。无论从内容的深度、广度以至在应用上的重要性等方面看，回归分析都远超出相关分析。目前，回归分析已确立为统计学中最重要的分支学科之一，出版了大量的著作，而相关分析则未能形成统计学的一个独立分支。

回归分析的方法以至“回归”这个名称的起源，统计史上一般归功于英国生物学家兼统计学家F. Galton(1822~1911)。基本意思是这样的：考察人的某项指标，以 X 记父代此指标之值， Y 记子代此指标之值，某些指标，如身高是有一定遗传性的，对这类指标而言， X 与 Y 有关，而 X 又不能决定 Y ，因此正好构成上

述那种关系。在不太长的时间内，人体的一项指标平均值大体上是稳定的(子代平均与父代平均相去不远)。但另一方面，该指标又有遗传性，因而当 X 增加(减少)时， Y 的平均值要增加(减少)，把这两点结合起来，会得出这样的结论：随着时间的推移，该项指标会发生两极分化的作用，但在现实中我们并未观察到这种两极分化。因此，除了随机性的影响外，必然存在一种力量把个体指标值“拉向中心”；或者说，子代指标有“向中心回归”的作用，这一点得到了观察结果的证实。Galton的学生、现代统计学的奠基者之一K. Pearson(1856—1936)观察了1078对夫妇，以每对夫妇的平均身高作为 X ，而取他们的一个成年儿子的身高作为 Y ，标出这1078对夫妇身高平均为 $X=68$ 吋，而子代平均身高为 $Y=69$ 吋，即子代比父代平均高了1吋。可是，对那些 X 值在72吋附近的夫妇而言(这些夫妇属于高个子一类)，其子代身高平均 $E(Y|72)=71$ ，他们虽仍属于高个子一类(71比子代全体平均69要大)，但不仅未能比其父母平均身高多出1吋，反而比父母平均身高下降了。就是说，父代向中心的偏离在子代被拉回来了；反之，对那些 X 值在64吋附近的夫妇而言(他们属于矮个子一类)，则 $E(Y|64)=67$ 吋，其子代平均虽仍低于总平均69吋(这反映遗传的影响)。但比父代提高了3吋，大于子代总平均提高的值1吋。就是说，子代平均向中心回归了。正是由于这个现象，Galton引进了“回归”这个名词来描述 X 与 Y 的关系。自然，这种“向中心回归”的现象，只是在这个特定领域里观察到的，并不具有普遍性。从这一点看，用“回归”这个名词描述变量之间的关系未必妥善，不过这名词已在统计学中沿用成习，没有可能也不必要再去设法改变它了。

再回到本题。由于回归函数的特殊重要性，回归分析所研究的主要问题，就是如何利用 X 、 Y 的观察值(样本)，对回归函数

进行统计推断，包括对它进行估计，及检验与它有关的假设等等。为在理论上进行讨论并导出种种具体的方法，需要引进一定的假定。这些假定产生种种回归模型，有的一般些，有的则性质较为特殊。容易理解：较弱的假定产生比较一般的模型，其适用面广，但针对这种模型而构造的统计推断方法，性能可能较差。反之，较强的假定产生性质特殊的模型，针对它而构造的统计推断方法性能较好，但适用面较小。下面来介绍几种常见的回归模型。

首先且主要的分类标准是回归函数 $f(x) = E(Y|x)$ 的形式。如果只假定这函数存在，而对其形式不作任何假定，则称为**非参数回归模型**。这种模型很一般，在理论上的发展为时也不长，大致是近二十年左右，才成为统计学家注意的对象，应用上也还不多。我们将在本书第五章中简略地考察一下这方面的问题。

与此相对的就是对回归函数 f 的一般数学形式假定为已知，但其中包含若干未知的参数，这其中尤其重要的是所谓“**线性回归模型**”，它是指 f 相对于未知参数而言是线性的。例如若 X 为 1 维，而

$$f(x) = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \quad (1.1)$$

$$f(x) = \alpha + \beta e^x, \quad c \text{ 已知} \quad (1.2)$$

都属于这种情况。此处 $\alpha, \beta_1, \cdots, \beta_p, \beta$ 等是模型中的未知参数。注意“线性”一词是针对参数而言，而不是针对自变量而言。在上述两例中， f 都不是自变量 X 的线性函数。在 (1.2) 中，若 c 不为已知而也是未知参数，则模型不是线性的。对线性模型而言，形式上引进一些新的“自变量”，可以把回归函数写成对“自变量”也是线性的形式。如在 (1.1) 中，引进新“自变量”

$$X_1 = X, X_2 = X^2, \cdots, X_p = X^p$$

可以把 (1.1) 写为

$$f(x_1, \cdots, x_p) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p \quad (1.3)$$

的形式。(1·2)也可类似处理，只须令 $X_1 = e^{cx}$ 即可。注意：引进的新自变量只能依赖于原自变量，而不能与未知参数有关。故如在(1·2)中 c 未知，则不可以令 $X_1 = e^{cx}$ 。(1·3)称为线性回归的标准形式。

线性回归在理论上和应用上都极为重要。到目前为止还可以说：线性回归的理论和方法，占据了回归分析这个分支的主要部分。在理论上，只有在回归函数为线性的假定下，才有比较深入和一般的结果。在应用上，许多实际问题，当自变量局限于一定范围时，可以满意地取这种模型作为真实模型的近似，其误差从实用的观点看无关紧要。当然，是否选用线性模型或其他（非线性）模型，要根据问题的具体情况而定。有时，现有的理论给出的就是一个非线性的模型，或者根据所收集的数据看，宜选用某一种非线性的模型，这也是不得已的，削足适履的作法也不足取。

有关因变量 Y 的矩和分布的假定，也是回归模型分类的一个重要标准。有一些问题只涉及到因变量 Y 的一、二阶矩，这时只须对 Y 的观察值的二阶矩结构有所假定即可（一阶矩已在回归函数中有所规定）。本书中有不少内容属于这个方面。但一涉及有关的检验和区间估计之类的问题，则必须进一步对 Y 的观察值的分布有所假定，例如常见的假定有 Y 的各次观察值独立，并服从正态分布。回归函数为线性而 Y 的观察值又是正态的模型，称为正态线性模型，这是在整个回归分析中，从理论到方法都研究得最完备的一种模型。在自变量代表时间的情况下，通常不假定因变量 Y 的各次观察值独立，而具有某种非独立的结构，例如构成一个平稳序列。这种回归模型的研究被划入统计学的另一个重要分支——时间序列统计分析的范围，本书不讨论这个范围内的问题。另外，也可以假定因变量 Y 具有某种截尾结构等。因变量 Y 本身也

可以是多维的，这种情况称为**多重回归分析**^{*)}，它通常划入多元统计分析的范围(广义地说，整个回归分析也可看作是多元统计分析的一部分)。本书也不讨论这方面的问题。

最后，自变量 X 有随机与否之分(因变量 Y 总是随机的)。设有 (X, Y) 的观察样本 $(X_1, Y_1), \dots, (X_n, Y_n)$ 。把 X 视为非随机的，就是认为 X_1, \dots, X_n 是已知的常数，不论它在何处出现，都以处理常数的规矩去对待它。若 X 视为随机的，则尽管从实用的观点看， X_1, \dots, X_n 都已知，但在进行理论上的探讨时，则必须记住它们本身也是随机变量，有一定的分布。举一个例子：设 X, Y 都是1维，回归函数是简单的线性形式 $E(Y|X) = \alpha + \beta X$ ，又设在给定 X 时 Y 的条件分布为 $N(\alpha + \beta X, \sigma^2)$ ， σ^2 与 X 无关。若进行了 n 次独立观察得 (X_i, Y_i) ， $i=1, \dots, n$ ，则通常统计教本中给出的回归系数 β 的估计是

$$\beta = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \left(\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \right)$$

若 X_1, \dots, X_n 被视为非随机，则易知 β 有正态分布，因为这里 X_i 都是普通常数，而 β 不过是独立正态变量 Y_1, \dots, Y_n 的线性组合。反之，若 X 也随机，则这个性质不复成立，而 β 的分布还与样本 X_1, \dots, X_n 的分布有关。

自变量 X 是否应视为随机的，这个问题在原则上不难由实际决定。在有些问题中， X 是一些对指标 Y 有影响的因素，其取值在一定限度内可由人控制，且在进行试验时， X 的取值也确是经过试验者一定的考虑而选择的，则这时 X 可以而且应当视为非随机的。在工农业试验中， X 往往是某些对产品质量或数量有影响

^{*)}注意不要把这个名词与统计学上常用的“多元回归分析”一词相混，习惯上，多元回归分析指的是自变量 X 的维数大于1的情况。

的因素——工艺、配方、施肥量、播种量之类，属于可由人控制的情况。在另一些问题中， X 和 Y 同是从随机地抽得的个体上测出的值， X 取值并不能事先预定。社会调查性质的问题，有不少是这个情况，这时把 X 看成非随机就不大符合实际。比如人的身高 X 和体重 Y ，一般是同时从一个随机抽得的人身上测得，不大可能在固定身高的人群中去随机抽样以观察其体重。

不过，由于一般讲当 X 非随机时，回归分析的理论要简化些，所以在实际应用中，有时往往不顾 X 的本质为随机的现实，而强行按 X 非随机的方法去处理。对这种做法可提出两点辩解。一是可以把所作的分析推断作为“条件化”的结论，即是在“ X 于观察中取值 X_1, \dots, X_n 的条件下”而作出的结论——自然，这一来结论的普遍程度就差一些：当另一名研究者想重复你的试验以判定你的研究是否可靠时，他未见得能碰巧观察到你所观察得的 X_1, \dots, X_n 。这样，在条件化的意义上你的试验原则上不可重复。二是若假定 (X, Y) 的联合分布为正态，则从一切实用的方面看，是否把 X 看成随机的都无关紧要，以后我们将解释这一点。

在 X 取值可以由人控制的情况下，就存在一个怎样去选定 X 在试验中所取的值的问题。这当然就需要确定一种目标，例如，使在这种选择下，据所得样本而进行的统计推断有某种良好的性质，以至在一种优良性准则之下有最优性等。这部分问题构成“回归设计”这个主题。本书将不涉及这方面。国内已有若干著作讨论这种问题，例如〔1〕，〔2〕，可以参考。

由以上的介绍可以看出，回归分析这个分支的内容是很丰富的，但本书主要限于讨论 X 为非随机时的线性回归问题，其他情况只是简单地涉及。

§1.2 参数的点估计

在本节及以下两节中, 我们集中讨论回归函数为线性, 且自变量 X 为非随机的情况。为了强调后面这一点, 我们将把 X 的样本值用小写字母 x_1, x_2, \dots 去记。因变量 Y 的样本值则用大写记为 Y_1, Y_2, \dots , 以强调它为随机。在线性回归分析中要广泛地使用向量和矩阵的工具。本书假定读者具有这方面的初步知识。所用记法也按一般通行的习惯。诸如用不加 “, ” 的字母记列向量, 加 “, ” 则表示转置, 两同维向量 a, b 的内积记为 $a'b$ 。若 $a'b=0$, 称 a, b 正交。 $a'a$ 有时记为 $\|a\|^2$, $\|a\|$ 称为向量 a 之长。当 $\|a\|=1$ 时称 a 为单位向量。 m 行 n 列的矩阵 A 有时说成是 $m \times n$ 矩阵 A , 或写成 $(a_{ij})_{m \times n}$ 的形式, a_{ij} 是 A 的 (i, j) 元, 即第 i 行第 j 列处的元。 A 的秩记为 $R(A)$, 而方阵 A 的行列式和迹(主对角线元之和)则分别记为 $\det(A)$ 和 $\text{tr}(A)$ 。 n 阶单位阵记为 I_n 或 I 。若无特别声明, 本书中提到的向量、矩阵等, 都是实的。 k 维欧氏空间记为 R^k 。若按向量理解, R^k 就是一切 k 维实列向量之集。设 \mathcal{S} 是 R^k 之一子集(\mathcal{S} 是一些 k 维向量之集), 则一切形如 $c_1 a_1 + \dots + c_n a_n$ 的向量之集, 其中 a_i 是 \mathcal{S} 中任一向量, c_i 是任意实常数, $i=1, \dots, n$, n 为任意自然数, 构成一个线性子空间, 称为由 \mathcal{S} 生成或张成的线性子空间, 记为 $\mathcal{M}(\mathcal{S})$ 。特别, 当 \mathcal{S} 是某矩阵 A 的一切列向量所构成时, 则称为由 A 所生成的线性子空间, 并记为 $\mathcal{M}(A)$ 。在这个场合, $\mathcal{M}(A)$ 的维数就是 A 的秩 $R(A)$ 。若方阵 A 为正定的, 则记为 $A > 0$; 若 A 为半正定, 则记为 $A \geq 0$ 。 $A > B$ ($A \geq B$) 表示 $A - B > 0$ (≥ 0)。正定及半正定的方阵总设为对称的。主对角元依次为 $\lambda_1, \dots, \lambda_n$ 的对角方阵记为 $\text{diag}(\lambda_1, \dots, \lambda_n)$ 。

(一) 最小二乘估计

设自变量 X 为 p 维, 因变量 Y 对 X 的回归函数有形状 $E(Y|X=x) = a + \beta'x$ 。此式中 a 称为常数项, $\beta = (\beta_1, \dots, \beta_p)'$ 称为回归系数(向量)。设在第 i 次观察中 X 取值为 $x_i = (x_{1i}, \dots, x_{pi})'$, Y 取值为 Y_i , $i=1, \dots, n$ 。形式地可以把 Y_i 表为

$$Y_i = a + \beta'x_i + e_i = a + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + e_i \quad (2.1)$$

这里 e_i 是随机变量, 有时称之为第 i 次观察的随机误差。按回归函数的定义, 应有 $Ee_i = 0$ 。

引进以下的向量和矩阵

$$\begin{aligned} \tilde{x}_i &= (1, x_i)', \quad i=1, \dots, n. \\ \tilde{X}' &= (\tilde{x}_1 : \tilde{x}_2 : \dots : \tilde{x}_n) \\ Y &= (Y_1, \dots, Y_n)', \quad e = (e_1, \dots, e_n)' \\ \gamma &= (a, \beta')' \end{aligned} \quad (2.2)$$

可以把(2.1)写成简洁的矩阵形式

$$Y = \tilde{X}\gamma + e \quad (2.3)$$

且附有条件

$$Ee = 0 \quad (2.4)$$

注意(2.3)式中的 Y 的意义与§1.1不同。这不致引起混淆。在以下, 记号 Y 多用在(2.3)式的意义上。在统计理论中, 当提到线性回归模型时, 一般常是指(2.3)式而言。因为(2.3)式是在有了数据以后模型的具体化, 而统计推断必须有数据才行。

现在的问题是要利用已有的数据 (x_i, Y_i) , $i=1, \dots, n$, 对 a 和 β 作出点估计。我们使用最小二乘法(简称为LS法)。其法归结为找 a, β , 使偏差平方和

$$H(a, \beta) = \sum_{i=1}^n (Y_i - a - x_{1i}\beta_1 - \dots - x_{pi}\beta_p)^2$$

$$= \sum_{i=1}^n (Y_i - \tilde{x}_i' \gamma)^2 = \|Y - \tilde{X} \gamma\|^2 \quad (2.5)$$

达到最小。在数学史上，一般把这个方法的发明归功于伟大数学家 Gauss 在 1799—1809 年之间的工作，但也还有些争议，参看 [3]。这方法直观上看很自然，仔细分析起来它也包含了某些对数据结构的假定。首先是各次观察在表达式 $H(\alpha, \beta)$ 中自占一项，不与其他项牵扯。这只有在各次观察独立或至少不相关时，才显得合理。其次，各次观察在表达式 $H(\alpha, \beta)$ 中具有等权，这需要各次观察具有大致相同的方差。事实上，确实只有当这些条件满足时，LS 估计才有优良的性质，我们在后面将论证这一点。

回到 H 的极值问题。把 H 对 $\alpha, \beta_1, \dots, \beta_p$ 求偏导数并命之为 0，或直接用求向量导数的公式，求 $\|Y - \tilde{X} \gamma\|^2$ 对 γ 的导数并命之为 0，得出方程组

$$\tilde{S} \gamma = \tilde{X}' Y, \quad \tilde{S} = \tilde{X}' \tilde{X} \quad (2.6)$$

此方程组有唯一解的充要条件为 $R(\tilde{S}) = p+1$ ，这等价于条件

$$R(\tilde{X}') = p+1 \quad (2.7)$$

以后我们总假定这条件满足。这意味着当我们选择 X 在试验中所取的值 x_1, \dots, x_n 时，不能违反这个条件。在这个条件成立时，得 (2.6) 的解为

$$\hat{\gamma} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \tilde{S}^{-1} \tilde{X}' Y \quad (2.8)$$

$\hat{\gamma}$ 是 Y_1, \dots, Y_n 的线性函数。就是说，LS 估计 $\hat{\gamma}$ 是线性估计。从理论上讲，还须验证一下，(2.8) 所确定的 $\hat{\gamma}$ 确实是 H 的最小值点。这很容易，事实上，对任何 γ 有

$$\begin{aligned} \|Y - \tilde{X} \gamma\|^2 &= \|Y - \tilde{X} \hat{\gamma} + \tilde{X} (\hat{\gamma} - \gamma)\|^2 \\ &= \|Y - \tilde{X} \hat{\gamma}\|^2 + (\hat{\gamma} - \gamma)' \tilde{S} (\hat{\gamma} - \gamma) \end{aligned}$$

$$+ 2(\hat{\gamma} - \gamma)' \tilde{X}' (Y - \tilde{X} \hat{\gamma}) \quad (2.9)$$

由(2.8)易见 $\tilde{X}' (Y - \tilde{X} \hat{\gamma}) = 0$, 又因 $\tilde{S} > 0$, 有 $\|Y - \tilde{X} \hat{\gamma}\|^2 \geq \|Y - \tilde{X} \gamma\|^2$ 对任何 γ , 等号当且仅当 $\gamma = \hat{\gamma}$ 时成立。这证明了 $\hat{\gamma}$ 是 H 的唯一的极小值点。

例如, 设 $p=1$ 这时

$$\tilde{X}' = \begin{pmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 \\ x_1 & x_2 & \cdot & \cdot & \cdot & x_n \end{pmatrix}$$

$R(\tilde{X}') = p+1=2$ 的充要条件是 $n \geq 2$, 且 x_1, \dots, x_n 不全相同。当这个条件满足时, 得到 α, β (即 β_1) 的 LS 估计分别为

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}, \quad \hat{\beta} = \sum_{i=1}^n (x_i - \bar{x}) Y_i / \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.10)$$

其中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$

(二) 中心化和标准化

从上一段的讨论中看出方阵 \tilde{S} 的作用, 以后还将看到它更多的作用。因此, 如能对模型作一些适当的变换以简化这个方阵, 将是有益的。下文讨论的中心化和标准化, 就是为了这个目的。

中心化就是把自变量空间(即 R^p)的原点移至它在 n 次试验中所取值的中心点处。换言之, 记

$$\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)' = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.11)$$

而把方程(2.1)改写为

$$Y_i = \alpha_0 + \sum_{j=1}^p \beta_j (x_{ji} - \bar{x}_j) + \epsilon_i, \quad i=1, \dots, n \quad (2.12)$$

其中 α_0 与(2.1)式中的 α 的关系为

$$\alpha_0 = \alpha + \beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p \quad (2.13)$$

因此, 从参数的角度看, 中心化不过是把原参数 $\alpha, \beta_1, \dots, \beta_p$ 作了一个线性变换而已, 且在这变换中 β_1, \dots, β_p 维持不动。若记

$$X_0 = \begin{bmatrix} 1 & (x_1 - \bar{x})' \\ 1 & (x_2 - \bar{x})' \\ \vdots & \dots \dots \dots \\ 1 & (x_n - \bar{x})' \end{bmatrix}, \quad S_0 = X_0' X_0 \quad (2.14)$$

以及

$$X^* = \begin{bmatrix} (x_1 - \bar{x})' \\ (x_2 - \bar{x})' \\ \dots \dots \dots \\ (x_n - \bar{x})' \end{bmatrix}, \quad S^* = X^{*'} X^* \quad (2.15)$$

则易见

$$S_0 = \begin{pmatrix} n & 0 \\ 0 & S^* \end{pmatrix}, \quad S_0^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & S^{*-1} \end{pmatrix} \quad (2.16)$$

而 α_0, β 的 LS 估计为

$$\begin{pmatrix} \hat{\alpha}_0 \\ \hat{\beta} \end{pmatrix} = S_0^{-1} X_0' Y: \quad \begin{cases} \hat{\alpha}_0 = \bar{Y}, \\ \hat{\beta} = S^{*-1} X^{*'} Y \end{cases} \quad (2.17)$$

这样, 在经过中心化后, 为得出 β 的 LS 估计 $\hat{\beta}$, 只须解一个 p 阶线性方程组^{*}), 而在原模型下要解一个 $p+1$ 阶方程组, 这在计算上有所简化。不过, 中心化最大的好处尚不在此。主要在于, 在参数 $\alpha, \beta_1, \dots, \beta_p$ 中, α 称为常数项, 而 β_1, \dots, β_p 称为回归系数, 二者意义有别。一般讲, 回归系数更重要一些。中心化提供了把

^{*}) 在解 (2.17) 中要求 S^* 可逆, 即 $R(S^*) = p$. 这条件自然与条件 (2.7) 等价。这一点可用纯矩阵方法证明, 留给读者作为一个练习。

二者分开处理的可能性，这在理论研究中很有用。以后我们将有机会看到这一点。

现在来讨论标准化。标准化是在先作中心化的基础上进行的。记

$$\begin{aligned}s_{jj}^2 &= \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2, \quad j=1, \dots, p \\ r_{jk} &= \sum_{i=1}^n (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) / (s_{jj}s_{kk}) \\ \beta_j^0 &= s_{jj}\beta_j, \quad j=1, \dots, p\end{aligned}\quad (2.18)$$

则可将经过中心化后的模型(2.12)再改变为

$$Y_i = \alpha_0 + \sum_{j=1}^p \beta_j^0 [(x_{ji} - \bar{x}_j) / s_{jj}] + e_i, \quad i=1, \dots, n \quad (2.19)$$

就这个模型，用一般公式(2.8)求 α_0 , β_1^0 , \dots , β_p^0 的LS估计，可得 α_0 的LS估计仍为 $\hat{\alpha}_0 = \bar{Y}$ ，而 $\beta^0 = (\beta_1^0, \dots, \beta_p^0)'$ 的LS估计则有

$$\hat{\beta}^0 = P^{-1}TY \quad (2.19)$$

的形式，其中 P 为 p 阶方阵，其 (i, j) 元为 r_{ij} ，而 $r_{ii}=1$ 。 T 为 $p \times n$ 矩阵，其 (j, i) 元为 $(x_{ji} - \bar{x}_j) / s_{jj}$ 。从解方程的角度看，公式(2.19)与(2.17)比，并未带来多少简化，重要的是方阵 P 。从(2.18)看出，如果自变量 X 是随机的，则(2.18)式所定义的 r_{jk} ，正是 X 的 j, k 两个分量的样本相关系数，所以 P 就是自变量 X 的样本相关矩阵。这是在统计相关分析中最基本的统计量，它的分析对回归分析有重要意义。除此之外，由于在某些问题中， X 的各分量所用单位的不同以及其考虑范围大小之不同，可能使 X 的某些分量在试验中的取值显著地大于另一些分量的取值，这不利于放在同一标准上进行比较，标准化可以克服这个缺点。

(三) Gauss-Markov定理

定理2.1 在记号(2.2)之下，假定(2.3)及(2.4)成立，则 α ,

β 的LS估计 $\hat{\alpha}, \hat{\beta}$ 是无偏估计。

证 只须证明 $E\hat{\gamma} = \gamma$ 。由(2.8), 并利用由(2.3)和(2.4)而得出的 $EY = \tilde{X}\gamma$, 得

$$E\hat{\gamma} = \tilde{S}^{-1} \tilde{X}' \tilde{X} \gamma = \tilde{S}^{-1} \tilde{S} \gamma = \gamma$$

于是得证。

在应用上, 我们不仅需要估计 α, β 本身, 还需要估计其线性函数。例如, 若要估计回归函数在某点 $X=d$ 处的值, 就是要估计 α, β 的线性函数 $\alpha + d'\beta$, 可写为 $(1, d')\gamma$ 的形状。可一般地提出估计 $c'\gamma$ 的问题, c 为一已知的 $p+1$ 维向量。按定理2.1, 有

系2.1 若 $\hat{\gamma}$ 为 γ 的LS估计, 则对任何 $p+1$ 维向量 c , $c'\hat{\gamma}$ 是 $c'\gamma$ 的无偏估计。习惯上也把 $c'\hat{\gamma}$ 称为 $c'\gamma$ 的LS估计。

要考察LS估计的精度如何, 就需计算 $\hat{\gamma}$ 的协方差阵 $\text{COV}(\hat{\gamma})$ 。为此就需对 Y 的协方差阵有所假定。在LS法的研究中常被采用的一种假定是

$$\text{COV}(Y) = \sigma^2 I_n \quad (2.20)$$

这里 $\sigma^2 > 0$ 是未知参数, I_n 为 n 阶单位阵。这假定称为 Gauss-Markov (GM) 假定。它的意义是: 各次观察结果 Y_1, \dots, Y_n 互不相关且有等方差。在(一)中引进LS法时我们曾提到过, 这方法的背后实际上隐含了这种要求。因此可以推想, 在GM假定下, LS估计应有良好表现。这一点下面即将给予证明。现在先直接由(2.20)得到

定理2.2 在条件(2.20)之下有

$$\text{COV}(\hat{\gamma}) = \sigma^2 \tilde{S}^{-1} \quad (2.21)$$

证 只须利用公式: 若 A 为常数矩阵, 则

$$\text{COV}(AY) = A \cdot \text{COV}(Y) \cdot A' \quad (2.22)$$

由于 $\hat{\gamma} = \tilde{S}^{-1} \tilde{X}' Y$, 以及 $\tilde{S} = \tilde{X}' \tilde{X}$ 对称, 有 $(\tilde{S}^{-1})' = \tilde{S}^{-1}$ 。

故

$$\begin{aligned}\text{COV}(\hat{\gamma}) &= \tilde{S}^{-1} \tilde{X}' \cdot \sigma^2 I_n \tilde{X} \tilde{S}^{-1} = \sigma^2 \tilde{S}^{-1} \tilde{X}' \tilde{X} \tilde{S}^{-1} \\ &= \sigma^2 \tilde{S}^{-1} \tilde{S} \tilde{S}^{-1} = \sigma^2 \tilde{S}^{-1}\end{aligned}$$

明所欲证。由此，再利用(2.22)，得

系2.2 $c'\gamma$ 的LS估计 $\hat{c'\gamma}$ 的方差是

$$\text{Var}(c'\hat{\gamma}) = \sigma^2 c' \tilde{S}^{-1} c' \quad (2.23)$$

由公式(2.21)看出中心化的作用：当我们把模型写成中心化的形式(2.12)时，其常数项 α_0 和回归系数 β 的LS估计由(2.17)决定。按公式(2.21)及(2.16)，应有

$$\text{COV} \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\beta} \end{pmatrix} = \sigma^2 S^{-1} = \sigma^2 \begin{pmatrix} n^{-1} & 0 \\ 0 & S^{*-1} \end{pmatrix}$$

于是得到

$$\text{Var}(\hat{\alpha}_0) = \frac{1}{n} \sigma^2, \text{COV}(\hat{\beta}) = \sigma^2 S^{*-1}, \text{COV}(\hat{\alpha}_0, \hat{\beta}) = 0 \quad (2.24)$$

(2.24)最后一式表明， $\hat{\alpha}_0$ 和 $\hat{\beta}$ 不相关。在假定 Y_1, \dots, Y_n 有正态分布时，由此就可推出 $\hat{\alpha}_0$ 与 $\hat{\beta}$ 独立。这更落实了下述想法：**中心化**使我们可以把模型中的常数项与回归系数分开处理。

系2.1表明： $c'\hat{\gamma}$ 是 $c'\gamma$ 的无偏估计。当然，除此以外，一般还可能存在其他的线性且无偏的估计。那么，在 $c'\gamma$ 的全部线性无偏估计类中，其LS估计 $c'\hat{\gamma}$ 是否有任何特出之处呢？下面的重要定理回答了这个问题。

定理2.3(GM定理) 设(2.20)成立，则在 $c'\gamma$ 的全部线性无偏估计类中，其LS估计 $c'\hat{\gamma}$ 是唯一的一个方差一致最小的估计(“一致最小”是指不论参数 α, β, σ^2 取什么值，方差都是最小)。

证 设 $a'Y$ 是 $c'\gamma$ 的一个线性无偏估计。由无偏性得出

$$c' \gamma = E(a' Y) = a' \tilde{X} \gamma, \text{ 对一切 } \gamma \in R^{p+1}$$

于是应有 $a' \tilde{X} = c'$, 现有

$$\begin{aligned} \|a\|^2 &= \|(a - \tilde{X} \tilde{S}^{-1} c) + \tilde{X} \tilde{S}^{-1} c\|^2 \\ &= \|a - \tilde{X} \tilde{S}^{-1} c\|^2 + \|\tilde{X} \tilde{S}^{-1} c\|^2 \\ &\quad + 2c' \tilde{S}^{-1} \tilde{X}' (a - \tilde{X} \tilde{S}^{-1} c) \end{aligned}$$

上式右边第三项为 0, 此因 $a' \tilde{X} = c'$, 故

$$\begin{aligned} \tilde{X}(a - \tilde{X} \tilde{S}^{-1} c) &= \tilde{X}' a - \tilde{X}' \tilde{X} \tilde{S}^{-1} c \\ &= c - \tilde{S} \tilde{S}^{-1} c = c - c = 0. \end{aligned}$$

因而

$$\|a\|^2 = \|a - \tilde{X} \tilde{S}^{-1} c\|^2 + \|\tilde{X} \tilde{S}^{-1} c\|^2 \quad (2.25)$$

但由条件(2.20), 有 $\text{Var}(a' Y) = \sigma^2 \|a\|^2$, $\text{Var}(c' \hat{\gamma}) = \sigma^2 \|\tilde{X} \tilde{S}^{-1} c\|^2$. 由此及(2.25)即知, 只要 $a' Y$ 为 $c' \gamma$ 的无偏估计, 则总有 $\text{Var}(a' Y) \geq \text{Var}(c' \hat{\gamma})$. 这证明了 $c' \hat{\gamma}$ 确为 $c' \gamma$ 的无偏估计中, 方差一致最小者。而且由(2.25)知, 要 $\text{Var}(a' Y) = \text{Var}(c' \hat{\gamma})$, 必须 $\|a - \tilde{X} \tilde{S}^{-1} c\|^2 = 0$ 即 $a = \tilde{X} \tilde{S}^{-1} c$, 这时有 $a' Y = c' \tilde{S}^{-1} \tilde{X}' Y = c' \hat{\gamma}$. 因此证实了: $c' \hat{\gamma}$ 为唯一的一个具有这个性质的估计。定理证毕。

通常把这个定理作为 LS 估计优越性的一个重要论据。在承认这一点的同时也应当注意到: 1. 此定理是在 GM 假定(2.20)的前提下才成立, 而在一个具体场合下, 此假定是否合理是可以讨论的, 并非自然而然的。2. 此定理中“方差最小”的结论, 是在“无偏”的限制下得到的。如果不局限于无偏估计, 可否找到另一些估计, 在某种意义上优于 LS 估计? 这也是可以讨论的问题。事实上, 近几十年来有许多学者在沿着这个方向做工作。本书的第四章中将研究这方面的问题。

(四) Y的协方差阵为一般正定方阵的情况

维持假定(2.3)和(2.4), 但将GM假定(2.20)修改为

$$\text{COV}(Y) = \sigma^2 G \quad (2.26)$$

其中 $\sigma^2 > 0$ 为未知参数, 而 G 为已知的正定方阵。当 $G = I_n$ 时回到(2.20)。因此有时也把(2.26)称为广义GM假定。

在广义GM假定下, $c'\hat{\gamma}$ 仍为 $c'\gamma$ 的线性无偏估计, 但方差最小的性质一般就不成立了。事实上, 利用定理2.3, 我们不难求出在假设(2.26)之下, $c'\gamma$ 的方差一致最小线性无偏估计, 如下定理所示。

定理2.4 在假定(2.26)之下, 线性函数 $c'\gamma$ 的一切线性无偏估计中, 唯一的方差一致最小者是 $c'(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1}Y$ 。

证 因 $G > 0$, 令 $Z = G^{-1/2}Y$, $\eta = G^{-1/2}\varepsilon$ ($G^{-1/2}$ 为一正定方阵, 满足条件 $G^{-1/2}G^{-1/2} = G^{-1}$ 。在矩阵论中证明了, 满足这些条件的 $G^{-1/2}$ 存在唯一), 有

$$Z = G^{-1/2}\tilde{X}\gamma + \eta \quad (2.27)$$

因 $E\varepsilon = 0$, 有 $E\eta = 0$ 。又

$$\text{COV}(Z) = G^{-1/2}\text{COV}(Y)G^{-1/2} = \sigma^2 G^{-1/2}GG^{-1/2} = \sigma^2 I_n$$

这说明, 线性模型(2.27)满足GM条件。据定理2.3, 即得 $c'\gamma$ 的唯一的方差一致最小线性无偏估计是

$$\begin{aligned} & c'((G^{-1/2}\tilde{X})'(G^{-1/2}\tilde{X}))^{-1}(G^{-1/2}\tilde{X})'Z \\ &= c'(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1/2}Z \\ &= c'(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1}Y \end{aligned} \quad (2.28)$$

这证明了定理的结论。显然, 本定理给出的估计(2.28), 与 $c'\gamma$ 的LS估计 $c'\hat{\gamma} = c'\tilde{S}^{-1}\tilde{X}'Y$ 一般不同。

从本定理看出，在一般假定(2·26)之下，原来在GM假定下LS估计 $\hat{\gamma}$ 的地位，由 $(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1}Y$ 所取代。这一点可从另一角度进行解释：设我们引进二次型 $(Y - \tilde{X}\gamma)'G^{-1}(Y - \tilde{X}\gamma)$ ，而去找 γ 使上式达到最小。则很容易证明：唯一的最小值点，正是 $(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1}Y$ 。这意思是说：由于 Y 的协方差阵有了变化，需要把最小二乘原则作相应的修改——即把目标函数由 $(Y - \tilde{X}\gamma)'(Y - \tilde{X}\gamma)$ 修改为 $(Y - \tilde{X}\gamma)'G^{-1}(Y - \tilde{X}\gamma)$ 。这证实了我们在引进LS法时所作的一种解释：这个方法(LS法)隐含了关于各次观察值不相关且等方差的要求。若这一点不对，则LS法将丧失其优越性。

一个有兴趣的情况是 G 为对角形时。设 $G = \text{diag}(g_1, \dots, g_n)$ ，即 G 的主对角元为 g_1, \dots, g_n 而其他元为0。这相当于 Y_1, \dots, Y_n 互不相关且 $\text{Var}(Y_i) = \sigma^2 g_i$ 。这时，目标函数为

$$(Y - \tilde{X}\gamma)'G^{-1}(Y - \tilde{X}\gamma) = \sum_{i=1}^n (Y_i - \tilde{x}_i'\gamma)^2/g_i$$

记号意义见(2·2)。于是我们得到以 $1/g_1, \dots, 1/g_n$ 为权的**加权最小二乘法**，权数与方差成反比。就是说，若某次观察的误差大(g_i 大)，则它在目标函数中占的地位要下降一些。方差不等的情况，在实用中是常见的(例如，方差随均值的增加而增加的情况)，但是，由于一般无法确定各次观察的方差的比值，故无法确定权数。因此加权最小二乘法不易付诸实施。

(五) 方差 σ^2 的点估计

仍回到GM假定(2·20)。在此假定下，方差 σ^2 仍是一未知参数。此参数反映观察误差的大小，因而在实用上很重要。

以 $\hat{\gamma}_n$ 记 γ 的LS估计，则

$$\delta_i = Y_i - \tilde{x}_i' \hat{\gamma} \quad (2.29)$$

称为第 i 次观察的残差，有时也简称为 Y_i 的残差。这个名称的来由，是因为按模型(2.1)， $Y_i - \tilde{x}_i' \gamma$ 就是第 i 次观察的误差。因此 δ_i 可视为这误差的估计。另一种看法是：若 X 与 Y 严格地有线性关系，则残差应为0。因此，残差反映了在因变量中消去自变量的线性影响后“残留”下来的东西。正因为残差具有这样的意义，它可用于“诊断”模型之用。比方说，不相关性与等方差性是否合理，是否需要作什么修正，这部分问题在实用上有重要意义，将在下一章讨论，这里先讨论残差的一种简单应用——用于估计方差 σ^2 。

δ_i 既然作为第 i 次观察误差的估计，故一般说，当 σ^2 大(小)时， $|\delta_i|$ 倾向于大(小)。由此可知， δ_i 的平方和，即

$$RSS = \sum_{i=1}^n \delta_i^2 \quad (2.30)$$

是衡量 σ^2 大小的一个合理指标。RSS称为**残差平方和**“(Residual Sum of Squares)，在回归分析中有多方面的应用。

先给RSS求出一个方便的表达式。按定义，有

$$\begin{aligned} RSS &= \|Y - \tilde{X} \hat{\gamma}\|^2 = Y'Y - 2\hat{\gamma}' \tilde{X}' Y + \hat{\gamma}' \tilde{S} \hat{\gamma} \\ &= Y'Y - \hat{\gamma}' \tilde{X}' Y \quad (\text{利用 } \tilde{S} \hat{\gamma} = \tilde{X}' Y) \end{aligned} \quad (2.31)$$

用(2.31)计算RSS很方便：因为在计算RSS时，通常已算出了LS估计 $\hat{\gamma}$ ，而 $\tilde{X}' Y$ 作为求解 $\hat{\gamma}$ 的方程组(2.6)的右边，也早已算出了。再以 $\hat{\gamma} = \tilde{S}^{-1} \tilde{X}' Y$ 代入，也可以把RSS表成 Y 的二次型：

$$RSS = Y' (I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') Y \quad (2.32)$$

易见 $I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}'$ 是 n 阶对称幂等方阵，即其平方等于其自身。

现在来计算RSS的均值。为此要利用下述结果：

引理2.1 设 ξ 为 n 维随机向量, $E\xi = a$, $\text{COV}(\xi) = V$, 又 A 为 n 阶常数方阵, 则

$$E(\xi' A \xi) = a' A a + \text{tr}(AV) \quad (2.33)$$

特别, 当 $\text{COV}(\xi) = \sigma^2 I_n$ 时, 有

$$E(\xi' A \xi) = a' A a + \sigma^2 \text{tr}(A) \quad (2.34)$$

证 要利用公式 $\text{tr}(CD) = \text{tr}(DC)$ 。将 ξ 表为 $\xi = a + \xi_0$, 则 $E(\xi_0 \xi_0') = \text{COV}(\xi) = V$. 故

$$\begin{aligned} E(\xi' A \xi) &= E\{(\xi_0 + a)' A (\xi_0 + a)\} \\ &= a' A a + 2a' A E\xi_0 + E(\xi_0' A \xi_0) \\ &= a' A a + E\{\text{tr}(\xi_0' A \xi_0)\} \\ &= a' A a + E\{\text{tr}(A \xi_0 \xi_0')\} \\ &= a' A a + \text{tr}(A \cdot E(\xi_0 \xi_0')) \\ &= a' A a + \text{tr}(AV) \end{aligned}$$

将此引理用于RSS的表达式(2.32), 并注意 $EY = \tilde{X}\gamma$, $\text{COV}(Y) = \sigma^2 I_n$, 而

$$\begin{aligned} &(\tilde{X}\gamma)'(I_n - \tilde{X}\tilde{S}^{-1}\tilde{X}')\tilde{X}\gamma \\ &= \gamma'(\tilde{X}'\tilde{X} - \tilde{X}'\tilde{X}\tilde{S}^{-1}\tilde{X}'\tilde{X})\gamma \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{tr}(I_n - \tilde{X}\tilde{S}^{-1}\tilde{X}') &= \text{tr}(I_n) - \text{tr}(\tilde{X}\tilde{S}^{-1}\tilde{X}') \\ &= n - \text{tr}(\tilde{S}^{-1}\tilde{X}'\tilde{X}) \\ &= n - \text{tr}(I_{p+1}) \\ &= n - p - 1 \end{aligned}$$

即得 $E(\text{RSS}) = (n - p - 1)\sigma^2$ 。于是得到,

定理2.5 令

$$\hat{\sigma}^2 = \text{RSS}/(n - p - 1) \quad (2.35)$$

则 $E\hat{\sigma}^2 = \sigma^2$, 即 $\hat{\sigma}^2$ 为 σ^2 的无偏估计。为使这估计有意义, 自然要有 $n > p + 1$. 有时也把 $\hat{\sigma}^2$ 称为“基于残差平方和的无偏估计”。 $n - p - 1$ 则称为 RSS 的“自由度”。

也可用类似方法求出在一般假定 (2.26) 之下, 方差 σ^2 的无偏估计。为此只须利用变换后的模型 (2.27)——它满足 GM 假定。按定理 2.5, 在 (2.27) 之下, σ^2 的基于残差平方和的无偏估计为

$$\begin{aligned} & Z'(I_n - G^{-1/2}\tilde{X}(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1/2})Z \\ &= Y'(G^{-1} - G^{-1}\tilde{X}(\tilde{X}'G^{-1}\tilde{X})^{-1}\tilde{X}'G^{-1})Y \end{aligned} \quad (2.36)$$

(六) 正态情况

GM 定理肯定了(在 GM 假定之下) LS 估计在一切线性无偏估计类中方差一致最小性。如果把考虑的估计类放宽到一切无偏估计的类(不论线性与否), 则这个方差最小性是否仍成立? 不难举例证明情况不必如此; 这一点取决于随机误差 e_i 的分布如何。在一个特别重要的场合, 即 e_i 为正态的情况, 这一点是成立的。不仅如此, 这时还可以证明: (2.35) 式确定的 σ^2 的无偏估计 $\hat{\sigma}^2$, 也是在 σ^2 的一切无偏估计类中, 方差一致最小的估计。这些结果的证明, 依赖于无偏估计理论中的 Rao-Blackwell-Lehmann-Scheffe 定理, 及统计量的充分性与完全性的理论(都参看 [4])。不熟悉这些理论的读者, 可略去下一定理的证明。本书在今后不会利用这定理的结果。

定理 2.6 设 GM 假定成立且 e_1, \dots, e_n 的联合分布为正态(这等价于要求 e_1, \dots, e_n 独立同分布, 且每一个有分布 $N(0, \sigma^2)$), 则 $c'\hat{\gamma}$ 是 $c'\gamma$ 的 UMVUE, $\hat{\sigma}^2$ 是 σ^2 的 UMVUE (UMVUE 是“一致最小方差无偏估计”的缩写)

证 把 n 维随机向量 Y 的密度 $f(y; \gamma, \sigma)$ 写出。注意到 $\|Y -$

$\tilde{X}'Y\|^2 = \|Y\|^2 - 2\gamma' \tilde{X}'Y + \gamma' \tilde{S} \gamma$, 得

$$f(y; \gamma, \sigma) = C(\tilde{X}, \gamma, \sigma) \exp \left\{ -\frac{1}{2\sigma^2} (\|Y\|^2 - 2\gamma' \tilde{X}'Y) \right\}$$

由此表达式, 利用因子判别定理, 知 $(\|Y\|^2, \tilde{X}'Y)$ 为 (γ, σ) 的充分统计量。又因当 σ 在 $(0, \infty)$ 变化而 γ 在 R^{p+1} 变化时, 点 $(1/\sigma^2, \gamma'/\sigma^2)$ 跑遍 R^{p+2} 的一半空间, 当然有内点, 故 $(\|Y\|^2, \tilde{X}'Y)$ 也是完全统计量。由于 $\hat{c}'\gamma$ 和 $\hat{\sigma}^2$ 都只与统计量 $(\|Y\|^2, \tilde{X}'Y)$ 有关且分别是 $c'\gamma$ 和 σ^2 的无偏估计, 因此也就分别是 $c'\gamma$ 和 σ^2 的UMVUE。

(七) 大样本性质

一个统计方法(如一个估计、一个检验等)的大样本性质, 是指当样本大小无限增加时, 该方法所具有的极限性质。与此相对的是小样本性质, 它是指统计方法在样本大小固定时的性质(一般要求这性质对任意的样本大小都成立)。例如, 本节前面几段所讨论的有关 $\hat{\gamma}$ 和 $\hat{\sigma}^2$ 的性质, 全是在样本大小 n 固定时进行的, 且对一般的 n 成立(有些必要的限制, 如 $n > p$, $n > p+1$ 之类), 故都是小样本性质。较狭义的说法, 当谈到小样本性质时, 往往与该统计方法中所使用的统计量的分布相关联, 而由这一点又引出所谓“大样本方法”与“小样本方法”之分: 一个统计方法若依据的是有关统计量的精确分布, 则称为小样本方法。反之, 若依据的是有关统计量当样本大小趋于 ∞ 时的极限分布, 则称为大样本方法。总之, 按照这些说法, 大小样本之分并不在于实际样本的大小(事实上, 多少样本算大算小, 并无客观标准)。不过, 在使用大样本方法时, 要求样本要“充分的大”, 而这个界限也难于明确规定。

有些久已确立的大样本方法，人们依据长期使用的经验，提出了一些参考性的界限。

首要的大样本性质是**相合性**。对参数估计而言，这是指当样本大小 $n \rightarrow \infty$ 时，估计量在一定的意义上收敛于被估计的参数。经常提到的有弱相合(依概率收敛)、强相合(以概率1收敛)、矩相合(矩收敛于0。在二阶矩的情况也称均方相合)。自六十年代以来，线性回归模型中LS估计的相合性问题，在文献中有很多研究。就本书的性质而言，不宜在此详细介绍这方面的工作，而只打算不加证明地提到几点最基本的事实，有兴趣的读者可参看[4]，或者[5]。

在讨论大样本性质时，要标明一些量与样本大小 n 的关系。为此，我们把由(2.2)式定义的矩阵 \tilde{X}' 记为 \tilde{X}_n ，把 $\tilde{X}_n' \tilde{X}_n$ 记为 \tilde{S}_n ，而把 γ 的LS估计记为 $\hat{\gamma}_n$ 。我们只须讨论 $\hat{\gamma}_n$ 本身的相合性即可。因若 $\hat{\gamma}_n$ 有某种意义下的相合性，则 $c' \hat{\gamma}_n$ 作为 $c' \gamma$ 的估计，有同一意义下的相合性。以下介绍几个关于相合性的结果。

1. 在GM条件满足时， $\hat{\gamma}_n$ 的均方相合性与其弱相合性等价，且二者的充要条件都是

$$\lim_{n \rightarrow \infty} \tilde{S}_n^{-1} = 0。$$

当 $\tilde{S}_n^{-1} \rightarrow 0$ (即 \tilde{S}_n^{-1} 的每一元都趋于0)时， $\hat{\gamma}_n$ 为 γ 的均方相合估计一事，立即从公式(2.21)看出。因为均方收敛导致弱收敛，知这时 $\hat{\gamma}_n$ 也为弱相合。其逆则是一个较深刻的结果，到七十年代中期以后才解决。

2. 当GM条件满足时， $\hat{\gamma}_n$ 的强相合性的问题，由本书作者之一解决。其要求是 $\tilde{S}_n^{-1} \rightarrow 0$ 要有一定的速度，例如 $\tilde{S}_n^{-1} = O((\log n)^{-2} (\log \log n)^{-(1+\varepsilon)})$ 就可以(对某个 $\varepsilon > 0$ 。这里 $1 + \varepsilon$ 不能改为1。如果进一步假定 e_1, e_2, \dots 相互独立，则黎子良等在只要求 $\tilde{S}_n^{-1} \rightarrow 0$

的条件下, 证明了 $\hat{\gamma}_n$ 的强相合性。

3. 关于 $\hat{\sigma}_n^2$ (即(2.35)式定义的 $\hat{\sigma}^2$)的相合性问题, 只假定GM条件已不够。1965-66年时, Gleser在 e_1, e_2, \dots 独立同分布(当然, 有 $Ee_1=0, Ee_1^2=\sigma^2$)的假定下, 证明了 $\hat{\sigma}_n^2$ 为 σ^2 的强相合估计。79-80年时, 陈希孺和赵林城解决了 e_1, e_2, \dots 相互独立但不必同分布的情况。

举一个例子: 设 $p=1$, 则

$$\tilde{X}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}$$

而

$$\tilde{S}_n^{-1} = \begin{pmatrix} n^{-1} & -\bar{x}_n/s_n \\ -\bar{x}_n/s_n & 1/s_n \end{pmatrix}, \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_n = \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

由此不难推出: $S_n^{-1} \rightarrow 0$ 与 $s_n \rightarrow \infty$ 等价。这归结为证明, 当 $s_n \rightarrow \infty$ 时有 $\bar{x}_n/s_n \rightarrow 0$ 。我们把这个简单证明留给读者。

§1.3 线性假设的检验

这一节我们要讨论涉及 α 和 β 的假设检验问题。仍记 $\gamma = (\alpha, \beta')'$, 我们要考虑的原假设有线性的形式, 即 $c'\gamma = 0$ 的形式, c 为已知的 $p+1$ 维向量。往往需要同时考虑几个这样的假设, 即原假设有形式 $H: \{c_i'\gamma = 0, i=1, \dots, k\}$ 。若引进一个 $k \times (p+1)$ 矩阵 H , 其 k 个行向量分别为 c_1', \dots, c_k' , 则上述原假设可表为

$$H: H\gamma = 0 \quad (3.1)$$

举几个例子 1. 如要检验回归方程 $y = \alpha + \beta'x$ 是否通过原点, 则原假设为 $\alpha = 0$, 即 $(1, 0, \dots, 0)\gamma = 0$ 。2. 若要检验某几个回归系数, 例如 β_1, \dots, β_k 是否为0, 则原假设 $\beta_1 = \dots = \beta_k = 0$ 可写为 $H\gamma = 0$, 其中 $k \times (p+1)$ 矩阵 H 有形状

$$H = (0 \quad I_k \quad 0)$$

这问题的实际含义是： $\beta_1 = 0$ 意味着 X 的第1分量，即变量或因素 X_1 ，不出现在回归方程中。也就是说，当 $\beta_1 = 0$ 时，因素 X_1 其实对指标 Y 无影响（或者，因素 X_1 的作用可以由其他因素 X_2, \dots, X_p 担当起来，故无必要再引入它）。当回归系统中包含众多自变量时，往往希望能舍去其中一部分不大重要的（对 Y 影响不大的）量，为了确定某些指定的自变量能否舍去，就需要进行这个检验。我们这里已提及“选择自变量”的问题，这个重要问题将在后面专章讨论。3. 上述问题的一个特例是当 $k = p$ 时，即 H 为 $\beta_1 = \dots = \beta_p = 0$ 。如果这成立，就表示所有的自变量都与指标无关，即这个回归完全无用。反之，若这一点不成立，则说明至少有若干个因素与指标有关，而这个回归不是空的。由于其这一意义，这个检验问题有时被称为“回归显著性检验”。

我们仍是从模型的一般形式(2.3)出发。就此可说明一个问题。形式上看我们也可以考虑非齐次的线性假设 $H\gamma = h$, h 为已知向量。这种非齐次假设在应用上也是可能出现的。例如，别人的研究表明， X_1 的回归系数 β_1 可取为2。现在你想要通过自己收集的数据来检验一下这项结论是否站得住，这就需要检验非齐次假设 $\beta_1 = 2$ 。不难证明，很容易将一般的非齐次假设 $H\gamma = h$ 转化为齐次假设(3.1)。事实上，任取方程 $H\gamma = h$ 之一解 γ_0 （若这种 γ_0 不存在，则任何参数值 γ 都不适合这个假设：原假设是空的，没有意义）。引进新参数 $\gamma^* = \gamma - \gamma_0$ ，则原假设 $H\gamma = h$ 转化为齐次假设 $H\gamma^* = 0$ 。在模型(2.3)中，只须引进 $Y^* = Y - X\gamma_0$ ，则可将它写成以 γ^* 为参数的形式 $Y^* = X\gamma^* + e$ 。因此，局限于齐次线性假设(3.1)并不影响讨论的普遍性。

(一) 一般线性假设的检验

现在来讨论假设(3.1)的检验问题。先提出检验方法的思想。在原模型(2.3)之下（不带假设(3.1)），据LS法，得到残差平方和RSS如(2.31)。根据我们在引进RSS时所作的解释，它反映了实际数据与假定的模型(2.3)的偏离程度。用统计学上的习惯说法，反映了数据与模型的“拟合程度”如何（RSS愈小，拟合愈好）。

一旦把原假设(3.1)考虑进来,那就等于说我们有了一个新模型,它是原模型的“缩小”——因新模型增加了(3.1)的限制,当然缩小了模型的范围。现有数据对这个缩小了的模型的拟合程度(以新模型下的残差平方和 RSS_H 来刻画),自然不能优于原模型,因为可供选择的范围小了。可是,若假设(3.1)的确成立,则新旧模型事实上同一,故数据对它们的拟合程度也应一样好。总之,这表明:当假设(3.1)成立时, $RSS_H - RSS$ 应倾向于小*,而当(3.1)不成立时则倾向于大。因此就得出这样的想法:当 $RSS_H - RSS$ 较大时,否定原假设(3.1),不然就接受它。但这个考虑还有一点毛病: $RSS_H - RSS$ 大小的意义要与 RSS 相比较去考虑,在此 RSS 起一种“背景噪声”的作用。好比在一把秤上称出甲比乙重3公斤,这结果的意义如何,只有联系到秤的误差大小(相当于此处的 RSS)去考虑才合理。

这样,我们最终达到以下的结论:应当根据统计量 $(RSS_H - RSS)/RSS$ 的大小来决定是否否定还是接受(3.1)。这统计量乘以适当的常数因子(见下文),就成为统计学中著名的F统计量,而相应的检验则称为F检验。

于是问题就转到 RSS_H 的计算。这无非是要解一个有约束的极值问题:

$$\text{在 } H\gamma = 0 \text{ 的约束下, 使 } \|Y - \tilde{X}\gamma\|^2 \text{ 最小} \quad (3.2)$$

这个问题在数学上易处理,且不难用显式表出 $\|Y - \tilde{X}\gamma\|^2$ 的最小值即 RSS_H (在约束 $H\gamma = 0$ 之下)。但我们不去做这个工作,因为一则理论上并非必要(至少就本书的目的而言),一则在实用上并不方便。在实用上解决极值问题(3.2)时,一般是采取将约束“融

*)应注意,总有 $RSS_H \geq RSS$ 。即使(3.1)成立,也不见得就会有 $RSS_H = RSS$ 。这是因为,用残差平方和去描述数据与模型的拟合程度,只是一种具体的方法,并非由真实模型与数据唯一决定,而与真实模型处在的背景有关。

化”于原模型中，得出一个无约束的线性模型，然后用无约束时计算RSS的公式去算。下面我们就一、二个具体例子说明这是怎样做的。下一段还将就若干重要特例给出表达式。

1. 若原假设(3·1)是 $\alpha=0$ ，则将它“融化”入原模型(2·1)后，我们得到线性模型 $Y_i = x_i'\beta + e_i, i=1, \dots, n$ 。这仍可写成(2·3)的形式，只是其中的 γ 现改为 β ，而新的 \tilde{X} 是由旧的 \tilde{X} 去掉第1列(全由1构成)而得到。在作完这步工作后即可按(2·31)式计算 RSS_H (为此需要在新模型下算出 β 的LS估计，这与旧模型下算得的自然不必相同)。

2. 若原假设(3·1)是 $\beta_1 = \beta_2$ ，则将此式代入(2·1)后，得 $Y_i = \alpha + (x_{1i} + x_{2i})\beta_1 + x_{3i}\beta_3 + \dots + x_{pi}\beta_p + e_i, i=1, \dots, n$ 。这仍是(2·3)的形式，只是 γ 应改为 $(\alpha, \beta_1, \beta_3, \dots, \beta_p)'$ 而新的 \tilde{X} 是由旧的 \tilde{X} 删去第3列，第2列由旧 \tilde{X} 的第2、3列之和构成，其他各列不动。这样做了以后即可按(2·31)式计算 RSS_H 。

用这样的算法，一般比直接用计算 RSS_H 的普遍公式要方便。现在为定出检验的界限，有必要求出统计量 $(RSS_H - RSS)/RSS$ 在原假设(3·1)成立时的分布。要实现这一点，需要假定误差 e_1, \dots, e_n 服从正态分布。

定理3·1 设在线性模型(2·3)中， e 服从 n 维正态分布 $N_n(0, \sigma^2 I)$ ，即 e 的各分量 e_1, \dots, e_n 独立，且都有分布 $N(0, \sigma^2)$ 。又以 r 记矩阵 H 的秩，则在原假设(3·1)成立时有

$$1^\circ \quad RSS/\sigma^2 \sim \chi_{n-p-1}^2, (RSS_H - RSS)/\sigma^2 \sim \chi_r^2$$

$$2^\circ \quad RSS \text{ 与 } RSS_H - RSS \text{ 独立}$$

由 1° 、 2° 即得到：若定义统计量

$$F_H = \frac{1}{r} (RSS_H - RSS) / \frac{1}{n-p-1} RSS \quad (3.3)$$

则有

$$3^{\circ} \quad F_H \sim F_{r, n-p-1}$$

此处 x_r^2 记自由度为 r 的 χ^2 分布, 而 $F_{m,n}$ 记自由度为 (m, n) 的 F 分布, 也常称 r 为 $RSS_H - RSS$ 的自由度。

这个定理是线性回归的线性假设检验理论中的基本定理, 其证明依赖下面两点预备事实。

引理3.1 设 ξ 为 n 维随机向量, $\xi \sim N_n(\mu, I_n)$ 。又 A 为 n 阶对称幂等方阵, $\mu' A \mu = 0$, 则有 $\xi' A \xi \sim x_r^2$, 其中 $r = R(A)$ 。

证 由对 A 之假定, 知存在正交方阵 P , 使

$$A = P' J P, \text{ 其中 } J = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \quad (3.4)$$

记 $\eta = P\xi$, 则因 P 正交, $\xi \sim N_n(\mu, I_n)$, 知 $\eta \sim N_n(\nu, I_n)$, 其中 $\nu = P\mu$ 。由 $\mu' A \mu = 0$ 知

$$0 = \mu' A \mu = \mu' P' J P \mu = \mu' P' J' J P \mu = \|J P \mu\|^2$$

故 $0 = J P \mu = J \nu$ 。即 ν 的前 r 个分量 ν_1, \dots, ν_r 皆为 0。故由 $\eta \sim N_n(\nu, I_n)$ 知, η 的前 r 个分量 η_1, \dots, η_r 独立, 且各有分布 $N(0, 1)$, 故

$$\sum_{i=1}^r \eta_i^2 \sim x_r^2.$$

但

$$\xi' A \xi = \xi' P' J P \xi = \eta' J \eta = \sum_{i=1}^r \eta_i^2$$

于是得证

引理3.2 设 A_1, A_2 都是 n 阶对称幂等方阵, $R(A_i) = r_i, i = 1, 2$ 且 $A_1 \geq A_2$, 则存在正交方阵 P 使

$$P' \begin{pmatrix} I_{r_i} & 0 \\ 0 & 0 \end{pmatrix} P = A_i, \quad i = 1, 2 \quad (3.5)$$

由此立即得出 $A_1 - A_2$ 为幂等方阵, $R(A_1 - A_2) = r_1 - r_2$ 。

证 先找正交阵 P_1 , 使

$$P_1' \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0 \end{pmatrix} P_1 = A_1$$

由于 $A_1 \geq A_2$, 从 $P_1 A_1 P_1' = \text{DIAG}(I_{r_1}, 0)$, 知存在 r_1 阶非负定方阵 C , 使

$$P_1' \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} P_1 = A_2$$

找 r_1 阶正交阵 Q , 使 $Q' C Q = \text{DIAG}(I_{r_2}, 0)$. 这种 Q 存在, 因为由 A_2 幂等及 $R(A_2) = r_2$, 知 C 幂等且 $R(C) = r_2$, 故 C 的特征根有 r_2 个为 1, 其余为 0, 作正交阵 $P_2 = \text{DIAG}(Q', I_{n-r_1})$, $P = P_2 P_1$, 则有

$$\begin{aligned} P' \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0 \end{pmatrix} P &= P_1' P_2' \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0 \end{pmatrix} P_2 P_1 \\ &= P_1' \begin{pmatrix} I_{r_1} & 0 \\ 0 & 0 \end{pmatrix} P_1 \\ &= A_1 \\ P' \begin{pmatrix} I_{r_2} & 0 \\ 0 & 0 \end{pmatrix} P &= P_1' P_2' \begin{pmatrix} I_{r_2} & 0 \\ 0 & 0 \end{pmatrix} P_2 P_1 \\ &= P_1' \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} P_1 = A_2 \end{aligned}$$

于是得证

定理 3.1 的证明。先设 $\sigma = 1$. 由 RSS 的表达式 (2.32), 注意到 $I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}'$ 为幂等且有秩 $n - p - 1$ (幂等易直接验证, 由秩 $R(I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') = \text{tr}(I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') = n - p - 1$ 可知。后一式已在证明 (2.35) 时验证过), 而且

$$\begin{aligned} (EY)' (I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') (EY) &= Y' \tilde{X}' (I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') \tilde{X} Y \\ &= 0 \end{aligned}$$

按引理3.1即得 $RSS \sim x_{n-p-1}^2$. 对一般的 σ , 用 Y/σ 代 Y , 得 $RSS/\sigma^2 \sim x_{n-p-1}^2$. 注意, 这一部分证明不依赖原假设是否成立.

为考虑 RSS_H , 先求出 $H\gamma=0$ 的一般解. 因 $R(H)=r$, 知存在秩为 $p+1-r$ 的 $(p+1) \times (p+1-r)$ 矩阵 B , 使通解为 $\gamma=B\delta$, $\delta \in R^{p+1-r}$ 任意. 这样, 有约束 $H\gamma=0$ 的模型化为无约束模型

$$Y = X^* \delta + e, \quad X^* = \tilde{X}B \quad (3.6)$$

且 $n \times (p+1-r)$ 矩阵 X^* 的秩与其列数 $p+1-r$ 一样. 此因一方面有 $n+1-r = R(B) \geq R(\tilde{X}B) = R(X^*)$, 另一方面因 $\tilde{X}'\tilde{X}$ 满秩, 有 $R(B) = R(\tilde{X}'\tilde{X}B) \leq R(\tilde{X}B) = R(X^*)$. 这样, 我们面临与 RSS 一样的情况, 即 $RSS_H = Y'A_1Y$, 其中 $A_1 = I_n - X^*(X^{*'}X^*)^{-1}X^{*'}$ 为幂等, 其秩等于 n 减去 X^* 之秩, 即等于 $n-p-1+r$. 若记 $A_2 = I_n - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$, 则显然有 $A_1 \geq A_2$, 因为 $RSS_H \geq RSS$. 这样, 引理3.2的条件全满足. 据该引理, 存在正交阵 P , 使(3.5)成立, 其中

$$r_1 = n - p - 1 + r, \quad r_2 = n - p - 1$$

记 $PY = Z = (Z_1, \dots, Z_n)'$, $EZ = \mu$. 则由 $Y \sim N_n(\tilde{X}\gamma, \sigma^2 I_n)$ 及 P 为正交, 知 $Z \sim N_n(\mu, \sigma^2 I_n)$, 故, Z_1, \dots, Z_n 相互独立. 由于(3.5), 有

$$\begin{aligned} RSS &= Y'A_2Y = Z'PA_2PZ \\ &= Z' \begin{pmatrix} I_{r_2} & 0 \\ 0 & 0 \end{pmatrix} Z = \sum_{i=1}^{n-p-1} Z_i^2 \end{aligned}$$

$$\begin{aligned} RSS_H - RSS &= Y'(A_1 - A_2)Y = Z'(PA_1P' - PA_2P')Z \\ &= Z' \text{DIAG}(0, I_{r_1-r_2}, 0)Z \\ &= \sum_{i=n-p}^{n-p+r-1} Z_i^2 \end{aligned}$$

这证明了 RSS 与 $RSS_H - RSS$ 独立. 注意这一条也与原假设成立与否无关. 最后, 为证明 $(RSS_H - RSS)/\sigma^2 \sim x_r^2$, 据引理3.1, 只须

证

$(EY)'(A_1 - A_2)(EY) = 0$, 当原假设(3.1)成立时
为此只须证 $(EY)'A_1(EY) = 0$ (当原假设(3.1)成立), 因前面已证
 $(EY)'A_2(EY) = 0$ (不论原假设成立与否)。但当(3.1)成立时有
(3.6), 故有 $EY = X^*\delta$, 而

$$(EY)'A_1(EY) = \delta'X^{*'}(I_n - X^*(X^{*'}X^*)^{-1}X^{*'})X^*\delta = 0 \quad (3.7)$$

这就证明了定理的全部论断, 注意这最后一步依赖(3.1)成立之事实。不然, 在(3.7)式中 EY 应以 $\tilde{X}\gamma$ 代入, 而(3.7)式就不成立了。

由这个定理, 立即就得到假设(3.1)的检验法。因其重要性我们将它写为一个定理。

定理3.2 设有线性回归模型(2.3), 其中随机误差向量 e 服从 n 维正态分布 $N_n(0, \sigma^2 I)$, 要检验线性假设(3.1). 令

$$RSS_H = \min \{ \|Y - \tilde{X}\gamma\|^2 : \gamma \in R^{p+1}, H\gamma = 0 \} \quad (3.8)$$

F_H 定义如(3.3), 其中 RSS 定义如(2.32), $r = R(H)$. 则在指定 $\alpha \in (0, 1)$ 后, 下述检验

$$\text{“当 } F_H > F_{r, n-p-1}(\alpha) \text{ 时否定(3.1), 不然就接受(3.1)”} \quad (3.9)$$

是水平 α 的检验。此检验称为(3.1)的水平 α 的 F 检验。 $F_{m,n}(\alpha)$ 是自由度为 (m, n) 的 F 分布的上侧 α 分位点。

以上我们是由 LS 法意义下的“拟合优度”的考虑, 而导出 F 统计量(3.3)。也可以从似然比的角度导出这个统计量。因此, F 检验(3.9)是假设(3.1)的似然比检验。

在理论上证明了: F 检验有多方面的优良性。最早研究这个问题的是我国著名统计学家许宝騄教授(1910—1970)。许的结果后来又由 Wald 改进。有关这方面的讨论可参看[4]§5.3, 及[6]第七

(二) 若干重要例子

从计算的角度看,要使用 F 检验(3·9)去检验线性假设(3·1),主要的工作是计算 RSS_H 。在这一段中,我们就在回归分析应用中很重要的几个特例来进行这个计算,并指出个别可以使计算简便的地方。

例3·1 检验一切回归系数为0,即

$$\text{原假设: } \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad (3\cdot10)$$

这个假设的意义,已在本节的引言部分交代过了。为算 RSS_H ,把(3·10)代入模型(2·1),得到无约束的模型 $Y_i = \alpha + \varepsilon_i$, $i = 1, \cdots, n$ 。于是

$$RSS_H = \min_{\alpha} \sum_{i=1}^n (Y_i - \alpha)^2 = \sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2 \quad (3\cdot11)$$

表达式 $\sum_{i=1}^n (Y_i - \bar{Y})^2$ 在回归分析(及方差分析)中常称为“总变差平方和”,或简称为“总平方和”(Total Sum of Squares,简化为TSS)。它衡量 Y 的各次取值 Y_1, \cdots, Y_n 的变异程度,或者说不均匀程度,散布程度。显然,对假设(3·10),相应于(3·1)中的矩阵 H 之秩为 p ,故得相应的 F 检验为

$$\text{“当 } \frac{1}{p} \left\{ \sum_{i=1}^n (Y_i - \bar{Y})^2 - RSS \right\} / \frac{1}{n-p-1} RSS > F_{p, n-p-1}(\alpha) \text{ 时}$$

$$\text{否定(3·10), 不然就接受(3·10)”} \quad (3\cdot12)$$

假设(3·10)常被称为“回归显著性假设”,相应的检验(3·12)则被称为“回归显著性检验”。不过,关于这个检验结果的解释,需要慎重。如果假设(3·10)被接受了,则应解释为(在所给的水平 α 之下,下同):相对于误差的影响而言,自变量对 Y 的影响是不重要的,因而通过这些自变量去预测 Y (及其他种种推断目的)也就没有多大意义。但这里所谓“误差的影响”有两种可能性:或者是因为有其他重要自变量遗漏了,或模型非线性因而增大了模型误差;或者纯系由于试验作得粗糙,而使随机误差增大了。若是前者,则检验结果指示我们,要进一步去搜索那些被遗漏了的重要因素,并考虑模型为线性的假定是否合理。若是后者,则指示我们要提高试验质量。当然一

般可能是二者兼而有之。在实用上，有时难于判定那一个是主要的。在试验有重复的情况下，可以通过数据去检验模型是否为线性，因而可以部分地回答这个问题。不过，从问题的专业角度去考察仍是主要的。

反之，如果假设(3·10)被否定了，那也只是说明：所选定的自变量对 Y 有一定的影响，因而这个回归模型在一定程度上说明了自变量 X 与 Y 的关系。但这并不意味着它已完善了。因为，虽则检验的结果告诉我们：所选自变量中至少有一部分是重要的，但不能排斥尚遗漏了其他重要因素的可能性。本书作者认为：此检验只宜用于辅助性的，事后验证性质的目的。说清楚一点：研究者在事前根据专业知识及经验，认为已把较重要的自变量选入了，且在一定误差限度内认为模型为线性是合理的。经过试验得出数据后，他可以用这个检验验证一下，原先的考虑是否有毛病。这时，若(3·10)被否定，他可以把这解释为，至少数据并不与他原来的设想矛盾；反之，若(3·10)被接受了，则提醒他，原来的想法也许在什么地方欠周到，最好再通盘考虑一下。这个说法的精神是强调事前的研究工作，而不把一切寄托在这个检验上。

可以从方差分析的角度对这个检验加以解释。总变差平方和 $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$ 分解为两项：

$$TSS = RSS + (RSS_H - RSS) = SS_e + SS_{\text{回}} \quad (3 \cdot 13)$$

TSS 之值是由于 Y_1, \dots, Y_n 不尽相同而来。为什么 Y_1, \dots, Y_n 会不尽相同呢？有两个原因：一是误差的影响，其量由 RSS 刻划，因此常把这一项称为

“误差平方和”并记为 SS_e ；一是由于自变量 X 取值对 Y 有影响，而 X 在各次试验中取值不同。这后一原因对形成 TSS 的贡献，就由 $TSS - RSS$ 或 $RSS_H - RSS$ 表达。因此常把这一项称为“回归平方和”并记为 $SS_{\text{回}}$ 。因 X 的重要性如何，就要看它在形成 TSS 中的贡献上与 SS_e 去对比，这正是 F 统计量。不仅如此，由这个分解式(3·13)，我们还可以对“ X 与 Y 关系的程度”作一个数量上的刻划，即利用 $SS_{\text{回}}$ 在 TSS 中所占份额：

$$R = SS_{\text{回}} / TSS \quad (3 \cdot 14)$$

\sqrt{R} (取正根) 称为 X 对 Y 的“样本复相关系数”。它衡量作为一个整体的 X ，与 Y 的线性关系的大小。当 X 的维数 $p = 1$ 时，复相关系数转化为通常相关系

数的绝对值。按这个观点，检验(3.12)无非是说，如果 X 对 Y 的样本复相关系数超过一定的界限，则认为 X 对 Y 的影响是显著的。但这个解释用于自变量 X 本身也为随机变量时，更为贴切一些。

例3.2 检验一部分回归系数为0，例如

$$\text{原假设: } \beta_k = \cdots = \beta_p = 0 \quad (1 < k \leq p) \quad (3.15)$$

我们往证：此假设的 $RSS_H - RSS$ 可用以下的步骤算得：先求出 β_k, \dots, β_p 的LS估计 $\hat{\beta}_k, \dots, \hat{\beta}_p$ 。把方阵 $S = \tilde{X}' \tilde{X}$ 写成 $\tilde{S} = \begin{pmatrix} B & C \\ C' & D \end{pmatrix}$ 的分块形式，其中 D 为 $p - k + 1$ 阶方阵，则

$$RSS_H - RSS = (\hat{\beta}_k, \dots, \hat{\beta}_p) (D - C' B^{-1} C) \begin{pmatrix} \hat{\beta}_k \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \quad (3.16)$$

证 把矩阵 \tilde{X}' 写为分块形式 $\tilde{X}' = \begin{pmatrix} \tilde{X}'_1 \\ \tilde{X}'_2 \end{pmatrix}$ ，其中 \tilde{X}_2 为 $p - k + 1$ 行。记 $\xi_i =$

$\tilde{X}'_i Y, i = 1, 2$ 。把 \tilde{S}^{-1} 记为 $\begin{pmatrix} B_1 & C_1 \\ C_1 & D_1 \end{pmatrix}$ ，则按(2.32)式，分别用于原模型及受到约束(3.15)的模型，有

$$RSS = \|Y\|^2 - (\xi'_1, \xi'_2) \begin{pmatrix} B_1 & C_1 \\ C_1 & D_1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

$$RSS_H = \|Y\|^2 - \xi'_1 B^{-1} \xi_1$$

故

$$RSS_H - RSS = \xi'_1 (B_1 - B^{-1}) \xi_1 + 2\xi'_1 C_1 \xi_2 + \xi'_2 D_1 \xi_2 \quad (3.17)$$

另有 $\hat{\gamma} = \tilde{S}^{-1} \tilde{X}' Y$ ，记 $a = (\hat{\beta}_k, \dots, \hat{\beta}_p)'$ ，得 $a = C_1' \xi_1 + D_1 \xi_2$ 。于是有

$$\begin{aligned} a' (D - C' B^{-1} C) a &= \xi'_1 C_1 (D - C' B^{-1} C) C_1' \xi_1 + 2\xi'_1 C_1 \\ &\quad (D - C' B^{-1} C) D_1 \xi_2 \\ &\quad + \xi'_2 D_1 (D - C' B^{-1} C) D_1 \xi_2 \end{aligned} \quad (3.18)$$

比较(3.17)与(3.18)，知(3.16)式从下述引理推出：

引理3.3 在上述记号之下有

$$B_1 = (B - C D^{-1} C')^{-1}, \quad D_1 = (D - C' B^{-1} C)^{-1}, \quad C_1 = -B_1 C D^{-1}$$

事实上,由

$$\begin{pmatrix} B & C \\ C' & D \end{pmatrix} \begin{pmatrix} B_1 & C_1 \\ C_1' & D_1 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

得

$$\begin{aligned} BB_1 + CC_1' &= I, \quad BC_1 + CD_1 = 0, \quad C'B_1 + DC_1' = 0 \\ C'C_1 + DD_1 &= I \end{aligned} \quad (3.20)$$

由第2、3两式解得 $C_1 = -B^{-1}CD_1$, $C_1' = -D^{-1}C'B_1$. 以后者代入 (3.20) 第1式, 即得 (3.19) 第1式。第2式类似得到。

利用此引理, (3.18) 右边第2、3项分别为 $2\xi_1' C_1 \xi_2$ 和 $\xi_2' D_1 \xi_2$ 。第1项可论证如下: 在 (3.19) 中调换 $\begin{pmatrix} B & C \\ C' & D \end{pmatrix}$ 与 $\begin{pmatrix} B_1 & C_1 \\ C_1' & D_1 \end{pmatrix}$ 的位置, 得 $B^{-1} = B_1 - C_1 D_1^{-1} C_1' = B_1 - C_1 (D - C' B^{-1} C) C'$ 。故 (3.18) 右边第1项为 $\xi_1' (B_1 - B^{-1}) \xi_1$ 。

公式 (3.16) 方便之处在于, 其中所有的量都是在原模型 (2.3) 之下算出的。例如 $\hat{\beta}_1, \dots, \hat{\beta}_p$ 是原模型下的 LS 估计, 计算麻烦之处在于涉及逆矩阵 B^{-1} 。可是在第三章中我们将看到: 整个计算可以通过一种叫做 “S 运算” 的简单程序实现。若不使用计算机, 则 $B^{-1}C(\hat{\beta}_1, \dots, \hat{\beta}_p)'$ 可如下计算: 算出 $b = C(\hat{\beta}_1, \dots, \hat{\beta}_p)'$ 。解方程 $Bx = b$, 其解 x 即为 $B^{-1}C(\hat{\beta}_1, \dots, \hat{\beta}_p)'$ 。

还有一点需要注意: 由于中心化不改变回归系数之值而使 \tilde{S} 有 (2.16) 式的 S_0 的形状。易见在公式 (3.16) 中, B, C, D 可取成按把 S^* 分块而得的矩阵。

例3.3 例3.2有一个重要特例, 即 $k = p$ 的情况:

原假设: $\beta_p = 0$ (3.21)

按公式 (3.16) 和 (3.19), $RSS_H - RSS$ 有 $\hat{\beta}_p^2/c$ 的形式, 其中 c 是 \tilde{S}^{-1} 主对角线上最末一元。但按公式 (2.21), 将有 $\text{Var}(\hat{\beta}_p) = \sigma^2 c$ 。于是就得到 假设 (3.21) 的 $RSS_H - RSS$ 的一个算法如下: 先找出 β_p 的 LS 估计 $\hat{\beta}_p$ 。算出 $\hat{\beta}_p$ 的方差, 有 $c\sigma^2$ 的形式, 则 $RSS_H - RSS$ 即为 $\hat{\beta}_p^2/c$ 。因为当 $\hat{\beta}_p = a_1 Y_1 + \dots + a_n Y_n$ 时有 $\text{Var}(\hat{\beta}_p) = \sigma^2 \sum_{i=1}^n a_i^2$, 故这个算法比直接从 (3.16) 式去计算远为简单。

另外, 在本例中, 相当于(3.1)中的矩阵 II 之秩为1, 故当原假设(3.21)成立时, 有

$$(\hat{\beta}_0^2/c) / \frac{1}{n-p-1} \text{RSS} \sim F_{1, n-p-1} \quad (3.22)$$

因此, $\sqrt{\frac{n-p-1}{c}} \hat{\beta}_0 / \sqrt{\text{RSS}}$ 服从自由度为 $n-p-1$ 的 t 分布, 而我们可用 t 分布表去检验假设(3.21)。

本例所得结果可略加推广。一般, 设要检验假设 $h'\gamma=0$, h 为已知的 $p+1$ 维非零向量。我们可以作一个可逆线性变换, 把 γ 变为 γ^* , 但使 $h'\gamma$ 恰好是 γ^* 的 $p+1$ 元。这样在新模型中, 我们所要检验的假设, 正好是(3.21)这一类型。因为检验统计量(3.22)显然不依赖于参数的变换, 故我们可这样去检验假设 $h'\gamma=0$: 算出 γ 的LS估计 $\hat{\gamma}$ 。计算 $\text{Var}(h'\hat{\gamma})=c^{-2}$, 则 F 统计量就是 $((h'\hat{\gamma})^2/c) / \frac{1}{n-p-1} \text{RSS} \sim F_{1, n-p-1}$ 。让我们来考察几个例子。

例3.4 要检验的是

$$\text{原假设: } \alpha = 0 \quad (3.23)$$

即回归平面通过原点。

解方程(2.6)算出 α 。我们已知有

$$\hat{\alpha} = \overline{Y} - \overline{x'}\hat{\beta}, \quad \overline{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad (3.24)$$

这不意味着要用(3.24)算 $\hat{\alpha}$ 。当然, $\hat{\beta}$ 可通过解(2.17)算出, 这是低一阶的方程, 因而用(3.24)也许方便些(计算程序往往也是如此安排的)。现用(3.24)来算 $\text{Var}(\hat{\alpha})$ 。有

$$\text{Var}(\hat{\alpha}) = \text{Var}(\overline{Y}) + \text{Var}(\overline{x'}\hat{\beta}) - 2E(\overline{x'}(\hat{\beta} - E\hat{\beta})(\overline{Y} - E\overline{Y}))$$

但由(2.24)式(注意(2.17)), 知上式右边最后一项为0再利用 $\text{COV}(\hat{\beta}) = \sigma^2 S^{*-1}$, 以及 $\text{Var}(\overline{Y}) = \sigma^2/n$, 得

$$\text{Var}(\hat{\alpha}) = \left(\frac{1}{n} + \overline{x'} S^{*-1} \overline{x} \right) \sigma^2$$

于是得到(3.23)的 F 统计量为

$$(\hat{a}^* (\frac{1}{n} + \bar{x}' S^{*-1} \bar{x}))' / \frac{1}{n-p-1} \text{RSS} = \frac{n(n-p-1)}{1 + \bar{x}' S^{*-1} \bar{x}} \frac{\hat{a}}{\text{RSS}}$$

自由度为 $(1, n-p-1)$. 也可用 t 检验, 即以

$$|\hat{a}| > t_{n-p-1} \left(\frac{\alpha}{2} \right) \sqrt{\text{RSS} \cdot (1 + \bar{x}' S^{*-1} \bar{x})} / \sqrt{n(n-p-1)}$$

为否定域。 S^* 见 (2.15), $t_m(\delta)$ 为上侧 δ 分位点。

例3.5 检验回归方程通过某点 $c = (c_1, \dots, c_p)'$. 不妨写成中心化的形式, 即要检验

$$\text{原假设: } \alpha_0 + (c - \bar{x})' \beta = 0 \quad (3.25)$$

$\alpha_0 + (c - \bar{x})' \beta$ 的 LS 估计为 $\bar{Y} + (c - \bar{x})' \hat{\beta}$, 其方差为 $\sigma^2 \left(\frac{1}{n} + (c - \bar{x})' S^{*-1} (c - \bar{x}) \right)$, 其余与例3.4无别。

例3.6 检验两个一元回归方程的斜率相同。即设

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, m \quad (3.26)$$

$$Y_i^* = \alpha^* + \beta^* x_i^* + e_i^*, \quad i = 1, \dots, n \quad (3.27)$$

此处假设 $e_1, \dots, e_m, e_1^*, \dots, e_n^*$ 全体独立, 且各服从正态分布 $N(0, \sigma^2)$. 要检验的是

$$\text{原假设: } \beta = \beta^*$$

根据上面交代的一般方法, 首先要作出 $\beta - \beta^*$ 的 LS 估计。显然, 这 LS 估计就是 $\hat{\beta} - \hat{\beta}^*$, 其中 $\hat{\beta}$ 和 $\hat{\beta}^*$ 分别就是从 (3.26) 和 (3.27) 各自作 LS 估计而得, 即

$$\hat{\beta} - \hat{\beta}^* = \frac{\sum_{i=1}^m (x_i - \bar{x}) Y_i}{\sum_{i=1}^m (x_i - \bar{x})^2} - \frac{\sum_{i=1}^n (x_i^* - \bar{x}^*) Y_i^*}{\sum_{i=1}^n (x_i^* - \bar{x}^*)^2}$$

又其方差, 即 $\hat{\beta}$ 的方差与 $\hat{\beta}^*$ 的方差之和:

$$\text{Var}(\hat{\beta} - \hat{\beta}^*) = \sigma^2 \left\{ \left(\sum_{i=1}^m (x_i - \bar{x})^2 \right)^{-1} + \left(\sum_{i=1}^n (x_i^* - \bar{x}^*)^2 \right)^{-1} \right\}$$

这样就可写出检验的形式。若用 t 检验，则否定域为

$$|\hat{\beta} - \hat{\beta}^*| > t_{\alpha/2, m+n-4} \left(\frac{\alpha}{2} \right) \sqrt{\text{RSS} \sqrt{\left(\sum_{i=1}^m (x_i - \bar{x})^2 \right)^{-1} + \left(\sum_{i=1}^n (x_i^* - \bar{x}^*)^2 \right)^{-1}}} \\ / \sqrt{m+n-4}$$

此处RSS为(3.26)的RSS及(3.27)的RSS之和，即

$$\text{RSS} = \sum_{i=1}^m (Y_i - \bar{Y})^2 - \left(\sum_{i=1}^m (x_i - \bar{x}) Y_i \right)^2 / \sum_{i=1}^m (x_i - \bar{x})^2 \\ + \sum_{i=1}^n (Y_i^* - \bar{Y}^*)^2 - \left(\sum_{i=1}^n (x_i^* - \bar{x}^*) Y_i^* \right)^2 / \sum_{i=1}^n (x_i^* - \bar{x}^*)^2$$

共有4个参数 $\alpha, \beta, \alpha^*, \beta^*$ ，故RSS的自由度为 $m+n-4$ 。

(三) 模型为线性的检验

到目前为止，所有的讨论都是在承认回归模型为线性的前提下进行的。我们曾说过，模型为线性这个基本前提，要由研究者根据有关的专业理论知识和经验，根据所造成的误差是否在所容许的限度内去决定之。这里所谓“经验”，自然包括以往所积累的经验，及当前所得的数据。在实际应用中，我们往往是把数据标在直角坐标系中形成“散点图”，根据其趋势去衡量线性假定是否可用。这在自变量为1维时很方便。当自变量为多维时要复杂些，但原则上也是可行的。从理论的角度看，可以提出“模型为线性”这个假设的检验问题。这里有两种情况。

一是随机误差 e_i 的方差 σ^2 大体上已知(知道它不超过某个限度)。这时，我们可以拿 $\hat{\sigma}^2$ (见(2.35)式)与 σ^2 比较。道理是这样的：当模型确为线性时，我们已知 $\hat{\sigma}^2$ 为 σ^2 的无偏估计，因此比值 $\hat{\sigma}^2/\sigma^2$ 不应过大。反之，若模型并非线性，则RSS的构成中，不仅有随机误差的成份，还有由模型与线性的偏离而带来的影响。因此RSS将倾向于增大，就是说，以 $\hat{\sigma}^2$ 作为 σ^2 的估计将偏高。这样，

根据比值 $\hat{\sigma}^2/\sigma^2$ (或 RSS/σ^2)的大小,可以给出一个判定准则,以判定模型是否显著偏离线性。

二是 e_i 的方差 σ^2 完全未知。这时,我们手头只有 RSS ,单凭它是无从判定模型是否偏离线性的。因为即使 RSS 很大,那也可能全是由于随机误差而来,不见得模型一定偏离线性。总之,失掉了比较的依据。

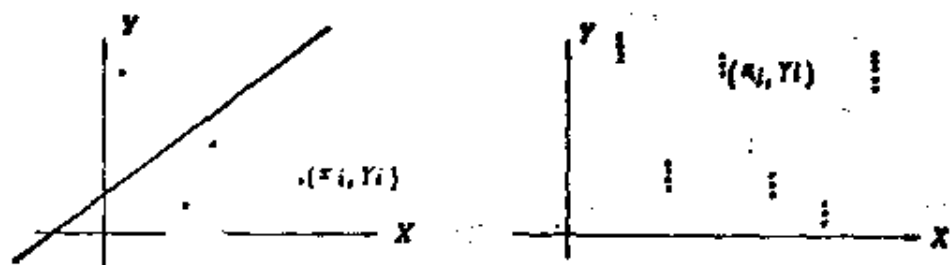


图1.3.1

但是,如果在自变量 X 的若干个值处进行了重复试验,则情形不同。因为重复试验的结果可用来估计随机误差的方差,而提供了比较的标准。直观上不难从上面两个图来说明。左图是无重复试验的情况。在此,散点图与直线相离甚远。但也有可能,理论上的回归直线就是图中画的那条,只不过因为随机误差太大,把试验点“赶”到了离此线甚远之处,因而散点图上已难于找到线性的踪迹。反之,右图是一个有重复的情况,在每个固定 x 处, Y 的试验值相聚很近,说明随机误差并不大。因此,散点图与直线趋势的严重不符,就不能委过于随机误差,而更可能是因为,模型本非线性的。

所以,让我们假定,在自变量 X 取值 x_i 时重复作 m_i 次试验观察 Y ,结果记为 Y_{ij} , $j=1, \dots, m_i$, 而 $i=1, \dots, n$ 。要求 $m_i \geq 1$, 且至少有一个 m_i 大于1。模型是

$$Y_{ij} = f(x_i) + e_{ij}, \quad j=1, \dots, m_i, \quad i=1, \dots, n \quad (3.28)$$

这里 f 是一个完全未知的函数(任何可能的模型都在考虑之列),而

$\{e_{ij}, j=1, \dots, m_i, i=1, \dots, n\}$ 全体独立, 且各有分布 $N(0, \sigma^2)$ 。在这个背景之下, 要检验以下的假设:

$$\text{“存在 } \alpha, \beta, \text{ 使 } f(x) = \alpha + \beta'x \text{”} \quad (3.29)$$

检验的步骤是这样的: 先就线性模型

$$Y_{ij} = \alpha + \beta'x_i + e_{ij}, \quad j=1, \dots, m_i, \quad i=1, \dots, n \quad (3.30)$$

算出其残差平方和, 记为 RSS 。其次, 算出“组内平方和” A :

$$A = \sum_{i=1}^n \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2, \quad \bar{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij} \quad (3.31)$$

把 RSS 分解为 $RSS = A + B$ 作统计量:

$$\mathcal{F} = \frac{B}{n-p-1} \bigg/ \frac{A}{N-n}, \quad N = \sum_{i=1}^n m_i \quad (3.32)$$

有如下的定理:

定理3.3 设自变量为 p 维, $n > p+1$, 而 m_1, \dots, m_n 中至少有一个大于 1, $\{e_{ij}, j=1, \dots, m_i, i=1, \dots, n\}$ 全体独立且各有分布 $N(0, \sigma^2)$ 。则当 (3.29) 成立时, 由 (3.22) 确定的统计量 \mathcal{F} 服从自由度为 $(n-p-1, N-n)$ 的 F 分布。

因此, 以

$$\mathcal{F} > F_{n-p-1, N-n}(\alpha) \quad (3.33)$$

为否定域的检验, 是假设 (3.29) 的水平 α 检验。注意的是, $\max(m_1, \dots, m_n) > 1$ 保证了 A (以概率 1) 大于 0, 且 $N-n > 0$ 。

这个检验的思想已在前面解释了: 把反映数据与线性模型的偏离之量 RSS 分解为 A 、 B 两部分, A 反映随机误差的影响, 而 B 则反映回归函数与线性函数的差异的影响。以 A 为基准去衡量 B , 而作出“ B 这么大的值是否达到了显著”的结论。

现往证定理3.3 记 $Y = (Y_{11}, \dots, Y_{1m_1}, \dots, Y_{n1}, \dots, Y_{nm_n})$ 。又以 J_k 记一 k 阶方阵, 其各元都是 $1/k$ 作方阵

$$C_1 = \text{DIAG} (I_{m_1} - J_{m_1}, \dots, I_{m_s} - J_{m_s})$$

则易见 $A = Y' C_1 Y$. 又易见 C_1 为幂等阵, 其秩为 $N - n$. 据 (2.32), 此处的 RSS 可表为 $Y' C Y$ 的形式, 其中 C 为幂等阵, C 的秩为 $N - (p + 1)$. 现往证 $B = \text{RSS} - A \geq 0$. 事实上 (我们全在假设 (3.29) 正确的条件下去讨论, 因为本定理是要在这个条件下去证明的), 以 $\hat{\alpha}, \hat{\beta}$ 记 α, β 的 LS 估计, 并记 $\hat{Y}_i = \hat{\alpha} + \hat{\beta}' x_i, i = 1, \dots, n$. 则由 $Y_{ij} - \hat{Y}_i = Y_{ij} - \bar{Y}_i + \bar{Y}_i - \hat{Y}_i$, 有

$$\sum_{j=1}^{m_i} (Y_{ij} - \hat{Y}_i)^2 = \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2 + m_i (\bar{Y}_i - \hat{Y}_i)^2$$

因而, 按残差平方和 RSS 的原始定义 (2.29), (2.30), 有

$$\text{RSS} = \sum_{i=1}^n \sum_{j=1}^{m_i} (Y_{ij} - \hat{Y}_i)^2 = A + \sum_{i=1}^n m_i (\bar{Y}_i - \hat{Y}_i)^2$$

故

$$B = \text{RSS} - A = \sum_{i=1}^n m_i (\bar{Y}_i - \hat{Y}_i)^2 \geq 0 \quad (3.34)$$

这证明了所要结果。又若记 $\mu = EY$, 则在 (3.29) 成立时, 前已证得 $\mu' C \mu = 0$, 故由 $C \geq C_1$ (此由 $\text{RSS} \geq A$ 推出) 知 $0 \leq \mu' C_1 \mu \leq \mu' C \mu = 0$, 即 $\mu' C_1 \mu = 0$. 这样一来, 由引理 3.1 和 3.2 导出定理 3.1 的论证, 在此完全适用, 从而推出本定理的结论。定理证毕。

反映回归函数 $f(x) = E(Y|X=x)$ 与线性函数的偏离程度的那一项 B 的表达式 (3.34), 使我们对这个检验的意义获得进一步的理解: 为估计 $f(x_i) = E(Y|X=x_i)$, 有两种方法: 一是用 \hat{Y}_i , 这一估计法不依赖模型的形式; 一是用 $\hat{\alpha} + \hat{\beta}' x_i = \hat{Y}_i$, 这一估计则有赖于模型为线性之假设, 故若线性假设正确, 则 \bar{Y}_i 与 \hat{Y}_i 作为同一量 $f(x_i)$ 之估计, 应相去不远, 而 B 之值将偏小。否则 \bar{Y}_i 和 \hat{Y}_i 将有差距, 而使 B 增大, 故以 B 之大小来衡量回归函数与线性函数偏离的程度是合理的。以 A 作分母, 只是意味着, B 的大小应以

随机误差大小为基准去衡量之。

§1.4 参数的区间(域)估计

(一) 线性函数的置信区域

仍将线性回归模型写成(2.3), 记号的意义如(2.2). 假定 $e \sim N_n(0, \sigma^2 I_n)$. 又本节与上两节一样, 自变量 X 视为非随机的。

给定 $k \times (p+1)$ 矩阵 A , 记 $\eta = A\gamma$, 我们要考虑 η 的置信区域。我们总假定 A 的秩为 k (因而必有 $k \leq p+1$). 先证明两个预备性的结果。

引理4.1 以 RSS 和 $\hat{\gamma}$ 分别记线性回归模型(2.3)的残差平方和及参数 γ 的LS估计。若 $e \sim N_n(0, \sigma^2 I_n)$, 则 RSS 与 $\hat{\gamma}$ 独立。

证 由(2.32)式, 注意到 $I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}'$ 为幂等阵, 有 $RSS = \|(I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}')Y\|^2$. 另一方面, 又有 $\hat{\gamma} = \tilde{S}^{-1} \tilde{X}'Y$ 因为

$$\begin{aligned}(I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}')(\tilde{S}^{-1} \tilde{X}')' &= (I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}') \tilde{X} \tilde{S}^{-1} \\ &= (\tilde{X} - \tilde{X} \tilde{S}^{-1} \tilde{S}) \tilde{S}^{-1} \\ &= 0\end{aligned}$$

及 Y 服从正态分布, 知 $(I_n - \tilde{X} \tilde{S}^{-1} \tilde{X}')Y$ 与 $\tilde{S}^{-1} \tilde{X}'Y$ 独立, 因而 RSS 与 $\hat{\gamma}$ 也独立. 证毕。

引理4.2 设随机向量 $\xi \sim N_n(0, G)$, $G > 0$. 则 $\xi' G^{-1} \xi \sim \chi_n^2$.

证 令 $\eta = G^{-1/2} \xi$. 则 η 为正态, $E\eta = 0$, $COV(\eta) = G^{-1/2} COV(\xi) G^{-1/2} = G^{-1/2} G G^{-1/2} = I_n$. 于是 $\eta \sim N_n(0, I_n)$, 因而 $\eta' \eta \sim \chi_n^2$. 但

$$\eta' \eta = \xi' G^{-1/2} G^{-1/2} \xi = \xi' G^{-1} \xi$$

于是得证.

现有

$$\begin{aligned} A(\hat{\gamma} - \gamma) &= A\tilde{S}^{-1}\tilde{X}'Y - A\gamma \\ &= A\tilde{S}^{-1}\tilde{X}'(\tilde{X}\gamma + e) - A\gamma \\ &= A\tilde{S}^{-1}\tilde{X}'e \end{aligned}$$

易见 $R(A\tilde{S}^{-1}\tilde{X}') = R(A) = k$, 此因 $k = R(A) \geq R(A\tilde{S}^{-1}\tilde{X}') \geq R(A\tilde{S}^{-1}\tilde{X}'\tilde{X}) = R(A) = k$. 因此, $G = (A\tilde{S}^{-1}\tilde{X}')(A\tilde{S}^{-1}\tilde{X}')' = A\tilde{S}^{-1}A' > 0$. 由引理4.2, 得

$$(\hat{\gamma} - \gamma)' A' (A\tilde{S}^{-1}A')^{-1} A(\hat{\gamma} - \gamma) \sim \chi_k^2$$

再利用引理3.4即得

$$\frac{1}{k}(\hat{\gamma} - \gamma)' A' (A\tilde{S}^{-1}A')^{-1} A(\hat{\gamma} - \gamma) / \hat{\sigma}^2 \sim F_{k, n-p-1} \quad (4.1)$$

此处 $\hat{\sigma}^2 = \text{RSS} / (n - p - 1)$. 把 $A\hat{\gamma}$ 和 $A\gamma$ 写为 $\hat{\eta}$ 和 η , 由(4.1)得

$$\begin{aligned} P \{ (\hat{\eta} - \eta)' (A\tilde{S}^{-1}A')^{-1} (\hat{\eta} - \eta) / k \hat{\sigma}^2 \leq F_{k, n-p-1}(\alpha) \} \\ = 1 - \alpha \end{aligned} \quad (4.2)$$

在有了样本之后, $\hat{\eta}$ 和 $\hat{\sigma}^2$ 也定了, 而适合关系式 $(\hat{\eta} - \eta)' (A\tilde{S}^{-1}A')^{-1} (\hat{\eta} - \eta) / k \hat{\sigma}^2 \leq F_{k, n-p-1}(\alpha)$ 的一切 η 构成一个以 $\hat{\eta}$ 为中心的椭球体. 根据(4.2), 这个椭球体就是 η 的一个置信区域, 有置信系数 $1 - \alpha$.

有两个特例值得注意

1. $A = I_{p+1}$, 即 $\eta = \gamma$. 这时(4.2)成为

$$\begin{aligned} P \{ (\hat{\gamma} - \gamma)' \tilde{S}(\hat{\gamma} - \gamma) / (p+1) \hat{\sigma}^2 \leq F_{p+1, n-p-1}(\alpha) \} \\ = 1 - \alpha \end{aligned} \quad (4.3)$$

2. $k=1$. 即 $\eta = a'\gamma$, a 为 $p+1$ 维非零向量. 这时(4.2)成为

$$\begin{aligned} P \{ (a'\hat{\gamma} - a'\gamma)^2 / (a'\tilde{S}^{-1}a \hat{\sigma}^2) \leq F_{1, n-p-1}(\alpha) \} \\ = 1 - \alpha \end{aligned} \quad (4.4)$$

注意 $\text{Var}(a'\hat{\gamma}) = \sigma^2 \cdot a'\tilde{S}^{-1}a$. 故在得到 $a'\hat{\gamma}$ 的表达式 $c'Y$ 后, 即有 $a'\tilde{S}^{-1}a = \|c\|^2$. 这样计算比用 \tilde{S}^{-1} 方便. 又(4.4)式也可通过 t 分

布表达:

$$P\left\{a' \hat{\gamma} - \sqrt{a' \hat{S}^{-1} a} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2}\right) \leq a' \gamma \leq a' \hat{\gamma} + \sqrt{a' \hat{S}^{-1} a} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha \quad (4.5)$$

取向量 a 为种种特殊形状, 可得常数项及各回归系数的区间估计, 以及两回归系数之差的区间估计等等。

例4.1 回归函数的区间估计。我们把回归函数写为中心化的形式 $\alpha_0 + \beta'(x - \bar{x})$ 较方便些。 α_0 的LS估计为 $\hat{\alpha}_0 = \bar{Y}$, β 的LS估计仍记为 $\hat{\beta}$, 回归函数的估计为 $\bar{Y} + \hat{\beta}'(x - \bar{x})$ 。我们已知 \bar{Y} 与 $\hat{\beta}$ 的协方差为0 (见(2.24)式), 有

$$\text{Var}(\bar{Y} + \hat{\beta}'(x - \bar{x})) = \sigma^2 \left(\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x}) \right)$$

S^* 见(2.15)式。要注意 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 是 X 的 n 次取值的平均, 而 x 是 R^p 中的一个流动点。用(4.5)式, 得

$$\begin{aligned} & P \left\{ \bar{Y} + \hat{\beta}'(x - \bar{x}) - \sqrt{\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2}\right) \right. \\ & \leq \alpha_0 + \beta'(x - \bar{x}) \leq \bar{Y} + \hat{\beta}'(x - \bar{x}) \\ & \left. + \sqrt{\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2}\right) \right\} \\ & = 1 - \alpha \end{aligned} \quad (4.6)$$

当 $p=1$ 时, 此式成为

$$\begin{aligned} & P \left\{ \bar{Y} + \hat{\beta}(x - \bar{x}) - \sqrt{\frac{1}{n} + (x - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2} \hat{\sigma} t_{n-2} \left(\frac{\alpha}{2}\right) \right. \\ & \leq \alpha_0 + \beta(x - \bar{x}) \leq \bar{Y} + \hat{\beta}(x - \bar{x}) \\ & \left. + \sqrt{\frac{1}{n} + (x - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2} \hat{\sigma} t_{n-2} \left(\frac{\alpha}{2}\right) \right\} \\ & = 1 - \alpha \end{aligned}$$

(二) 误差方差 σ^2 的区间估计

仍设在线性模型(3.2)中, $e \sim N_n(0, \sigma^2 I_n)$. 并如我们在其他地方所做的那样, 假定(2.7)成立, 要作 σ^2 的区间估计。

以RSS记残差平方和, σ^2 的区间估计是基于定理3.1, 1°中所证明的性质

$$RSS/\sigma^2 \sim \chi^2_{n-p-1}$$

在定理3.1的证明中我们曾交代过, 这个性质之成立与假设(3.1)是否成立无关。若以 $\chi^2_{n-p-1}(\alpha)$ 记自由度为 $n-p-1$ 的 χ^2 分布的上侧 α 分位点, 则由这个性质有

$$P\left(\chi^2_{n-p-1}\left(1 - \frac{\alpha}{2}\right) \leqslant RSS/\sigma^2 \leqslant \chi^2_{n-p-1}\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

此式可改写为以下的形状:

$$\begin{aligned} P\left(RSS/\chi^2_{n-p-1}\left(\frac{\alpha}{2}\right) \leqslant \sigma^2 \leqslant RSS/\chi^2_{n-p-1}\left(1 - \frac{\alpha}{2}\right)\right) \\ = 1 - \alpha \end{aligned} \quad (4.7)$$

于是, (4.7)式左边的不等式规定了 σ^2 的一个区间估计, 其置信系数等于指定的 $1 - \alpha$. 对 $\alpha = 0.05, 0.01$ 等值, (4.7)式中涉及的量 $\chi^2_{n-p-1}\left(\frac{\alpha}{2}\right), \chi^2_{n-p-1}\left(1 - \frac{\alpha}{2}\right)$ 可由通常的 χ^2 分布表查得。当 $n-p-1$

不大时, (4.7)式确定的区间往往很长, 因而使区间估计的意义不大。总的说, 在实际应用中人们并不常去作 σ^2 的区间估计。

(三) 同时区间估计

以 $g_1(x)$ 和 $g_2(x)$ 分别记(4.6)式中的区间的上下端点, 则当 x 在 R^p 内流动时, R^{p+1} 中的点 $(x', g_1(x))$ 和 $(x', g_2(x))$ 分别画出

曲面 l_1, l_2 , 把经验回归平面^{*} $y = \bar{Y} + \hat{\beta}'(x - \bar{x})$ 夹在当中。在 $p=1$ 时可画图, 如图1·4·1所示

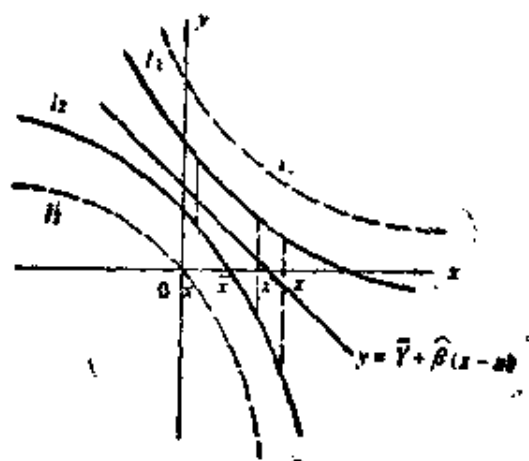


图1·4·1

要注意的是: (4·6)式只是当 x 固定时成立。就是说, 若我们关心的只是回归函数在一个特定的点 x 处的区间估计, 则(4·6)完全解决了问题, 然而, 在应用上我们往往不是只关心回归函数在一点处之值, 而是关心它在许多点的值, 一般是在一定范围内的一切点处之值。换句话说, 我们希望, 对许多 x 值, 甚至是一切 x 值, (4·6)中的不等式同时成立。但这个概率便不能保持为 $1-\alpha$, 而会低于 $1-\alpha$ 。因为, 许多个事件(每个不同的 x 产生一个事件, 即“(4·6)中的不等式成立”这个事件)同时发生的概率, 不超过(且往往是小于)其中任一个事件的概率。因此, 在图4·1中, 理论回归直线 $y = a_0 + \hat{\beta}'(x - \bar{x})$ 落在两条曲线 l_1, l_2 之间的区域内的概率, 要低于 $1-\alpha$ 。如果仍想保持这个概率, 就必须把区域扩大一些(把 l_1 上移, l_2 下移)。这一段我们就来讨论此问题。

同时区间估计的问题不止是在估计回归函数时发生。例如, 要同时对 β_1, \dots, β_p 作区间估计。我们固然可以用(4·5)对每个 β_i 作出置信系数为 $1-\alpha$ 的区间估计, 但把这 p 个区间估计联合起来, 其总的置信系数就会低于 $1-\alpha$ 。若要保持这个置信系数, 就须修改每个 β_i 的区间估计。在某种意义上说, 这个问题虽然只涉及到

*)此名称的来由, 是由于 \bar{Y} 和 $\hat{\beta}$ 是通过数据得到的。真实的回归平面 $y = a_0 + \beta'(x - \bar{x})$ 有时称为“理论回归平面”

有限个量(即 β_1, \dots, β_p), 但比前一问题(它涉及无限个量 $\alpha_0 + \beta^x(x - \bar{x}) : x \in R^p$)更为麻烦, 以后将解释这一点。

同时区间估计是统计学中的一个重要研究题目, 有一些学者针对种种具体情况, 提出了种种方法, 其详可参看专著[7]或[8]。对此处的情况而言, 只有Scheffe在1953年提出的方法(文献中常称为S法)比较合用, 现在我们就来介绍这个方法。其基本定理是

定理4.1 设有线性回归模型(2.3), 其中 $e \sim N_n(0, \sigma^2 I)$, $\hat{\gamma}, \hat{\sigma}^2$ 的意义同前。设 \mathcal{L} 是 R^{p+1} 的一个线性子空间, 其维数为 d 。则有

$$\begin{aligned} P\{l' \hat{\gamma} - \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} (l' \tilde{S}^{-1} l)^{1/2} \leq l' \gamma \\ \leq l' \hat{\gamma} + \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} (l' \tilde{S}^{-1} l)^{1/2}, \text{ 对一切 } l \in \mathcal{L}\} \\ = 1 - \alpha \end{aligned} \quad (4.8)$$

这个定理给出了无限多个线性函数($l' \gamma : l \in \mathcal{L}$)的同时区间估计, 其置信系数(对全体而言)等于指定值 $1 - \alpha$ 。把(4.8)与(4.5)比较, 我们看出: 对一个固定的 $l' \gamma$ 而言, 二者的差别在于(4.5)中的因子 $t_{n-p-1}\left(\frac{\alpha}{2}\right)$, 即 $\sqrt{F_{1, n-p-1}(\alpha)}$, 在(4.8)中改为 $\sqrt{d F_{d, n-p-1}(\alpha)}$ 。

这个修改不依赖 l 。不难看出, 当 $d > 1$ 时必有 $d F_{d, n-p-1}(\alpha) > F_{1, n-p-1}(\alpha)$ (我们把这一很易证明的结果留给读者作为练习), 因而对一个固定的 $l' \gamma$ 而言, (4.8)所给的区间估计确比(4.5)扩大了, 以这个扩大为代价, 保持了置信系数(对一切 $l \in \mathcal{L}$ 同时)不变。这个定理的局限之处在于: \mathcal{L} 必须是一个线性子空间, 而不能是任意集合。

现转到定理的证明。以 Q' 记 $(p+1) \times d$ 矩阵, 使 $\mathcal{L}(Q')(Q'$ 的各列张成的子空间)即为 \mathcal{L} 。令

$$\delta = Q' \gamma, \quad \hat{\delta} = Q' \hat{\gamma}$$

则易见当 $l \in \mathcal{L}$ 时, 存在唯一的 $\lambda \in R^d$, 使 $l'\gamma = \lambda'\delta$ 。反之, 若 $\lambda \in R^d$, 则必存在 l (即为 $Q'\lambda$), 使 $\lambda'\delta = l'\gamma$ 。因为 $\hat{\gamma} \sim N_{p+1}(\gamma, \sigma^2 \tilde{S}^{-1})$, 有 $\hat{\delta} \sim N_d(\delta, \sigma^2 Q \tilde{S}^{-1} Q')$ 。故

$$\begin{aligned} l' \hat{\gamma} &= \lambda' \hat{\delta} \sim N(\lambda' \delta, \sigma^2 \lambda' Q \tilde{S}^{-1} Q' \lambda) \\ &= N(l' \gamma, \sigma^2 l' \tilde{S}^{-1} l) \end{aligned}$$

因 Q 为 $d \times (p+1)$ 矩阵有秩 d , 知 $Q \tilde{S}^{-1} Q' > 0$ 。故存在 d 阶非异方阵 P , 使 $PQ \tilde{S}^{-1} Q' P' = I_d$ 。令

$$\eta = P\delta, \quad \hat{\eta} = P\hat{\delta}, \quad \text{则 } \hat{\eta} \sim N_d(\eta, \sigma^2 I)$$

因而, 注意到 $\hat{\eta}$ 只依赖于 $\hat{\gamma}$, 故与 $\hat{\sigma}^2$ 独立 (引理 4.1), 因此 $\|\hat{\eta} - \eta\|^2 \frac{1}{\hat{\sigma}^2} \sim F_{d, n-p-1}$, 有

$$P\{\|\hat{\eta} - \eta\|^2 \leq d F_{d, n-p-1}(\alpha) \hat{\sigma}^2\} = 1 - \alpha$$

易见: 若 t 为一固定的 d 维向量, $c > 0$ 为常数, 则

$$\|t\|^2 \leq c^2 \iff |a't| \leq c \|a\| \text{ 对一切 } a \in R^d \quad (4.9)$$

我们把这个简单事实的证明留给读者。应用这个事实, 得

$$\begin{aligned} P\{|a'(\hat{\eta} - \eta)| \leq \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} \|a\|, \text{ 对一切 } a \in R^d\} \\ = 1 - \alpha \end{aligned}$$

即

$$\begin{aligned} P\{|a' P(\hat{\delta} - \delta)| \leq \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} \|a\|, \text{ 对一切 } a \in R^d\} \\ = 1 - \alpha \end{aligned}$$

由于 P 非异, 当 a 跑遍 R^d 时, $P'a = \lambda$ 跑遍 R^d 。故上式可写为

$$\begin{aligned} P\{|\lambda'(\hat{\delta} - \delta)| \leq \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} \|P'^{-1}\lambda\|, \text{ 对一切 } \lambda \in R^d\} \\ = 1 - \alpha \end{aligned}$$

回忆前面讲到的 $l'\gamma$ 与 $\lambda'\delta$ 的对应关系, 知上式可写为

$$\begin{aligned} P\{|l'(\hat{\gamma} - \gamma)| \leq \sqrt{d F_{d, n-p-1}(\alpha)} \hat{\sigma} \|P'^{-1}Q'^{-1}l\|, \text{ 对一切 } l \in R^{p+1}\} \\ = 1 - \alpha \end{aligned} \quad (4.10)$$

因为 $P'^{-1}Q'^{-1}l = (Q'P')^{-1}l$, 有

$$\|P'^{-1}Q'^{-1}l\|^2 = l'(PQ)^{-1}(Q'P')^{-1}l$$

再注意到 $PQ\tilde{S}^{-1}Q'P' = I_d$, 得 $\|P'^{-1}Q'^{-1}l\| = \sqrt{l'\tilde{S}^{-1}l}$. 于是(4.10)转化为(4.8). 定理证毕.

这个定理直接用于回归函数还有点小问题. 因为回归函数 $\alpha + \beta'x$ 写成 $l'\gamma$ 的形状时, 有 $l' = (1, x')$, 当 x 在 R^p 内变化时, 形如 $l' = (1, x')$ 的一切向量, 并不构成线性子空间. 但这只须对定理 4.1 稍加一点修改即可:

系 4.1 设 \mathcal{L} 为 R^{p+1} 的一个 $d-1$ 维线性子空间, h 为不属于 \mathcal{L} 的 $p+1$ 维向量. 以 \mathcal{L} 记向量集合 $\{l: l = h + \tilde{l}, \text{ 其中 } \tilde{l} \in \mathcal{L}\}$. 则对这个 \mathcal{L} , (4.8) 式仍成立.

证 记 $\mathcal{L}_1 = \{ch + \tilde{l} : -\infty < c < \infty, \tilde{l} \in \mathcal{L}\}$, 则 \mathcal{L}_1 为包含 \mathcal{L} 的线性子空间, 其维数为 d . 往证

$$\begin{aligned} & \{|l'\hat{\gamma} - l'\gamma| \leq \sqrt{dF_{d,n-p-1}(\alpha)} \hat{\sigma}(l'\tilde{S}^{-1}l)^{1/2}, \\ & \text{对一切 } l \in \mathcal{L}\} \\ &= \{|l'\hat{\gamma} - l'\gamma| \leq \sqrt{dF_{d,n-p-1}(\alpha)} \hat{\sigma}(l'\tilde{S}^{-1}l)^{1/2}, \\ & \text{对一切 } l \in \mathcal{L}_1\} \end{aligned} \quad (4.11)$$

因为 $\mathcal{L} \subset \mathcal{L}_1$, 故上式左边 \supset 右边 (即右边事件成立时, 左边必成立). 反之, 设左边事件成立. 往证右边事件也成立. 证明了这一点也就证明了 (4.11). 再由 (4.8) 知 (4.11) 右边的概率为 $1-\alpha$, 故左边的概率也是 $1-\alpha$, 而系 4.1 得证.

易见, 若对某个 $l \in R^{p+1}$ 有

$$|l'\hat{\gamma} - l'\gamma| \leq \sqrt{dF_{d,n-p-1}(\alpha)} \hat{\sigma}(l'\tilde{S}^{-1}l)^{1/2} \quad (4.12)$$

则当 l 改为 cl (c 为常数) 时, 此式仍成立. 现取 \mathcal{L}_1 中形如 $ch + \tilde{l}$ 的 l , 其中 $c \neq 0$, $\tilde{l} \in \mathcal{L}$, 则 $h + \tilde{l}/c \in \mathcal{L}$. 因为 (4.11) 左边的事件发生, 故以 $h + \tilde{l}/c$ 代 (4.12) 中的 l 时, (4.12) 成立. 据上述, (4.12) 式对 $c(h + \tilde{l}/c) = ch + \tilde{l}$ 也成立. 以 $ch + \tilde{l}$ 代入 (4.12) 式并令 c

$\rightarrow 0$, 知(4.12)式对形如 $0 \cdot h + \tilde{l}$ 的 l 也成立, 即(4.12)式对一切 $l \in \mathcal{L}_1$ 成立. 这证明了(4.11)式.

在应用中, 常把回归方程写成中心化的形式: $y = a_0 + \beta'(x - \bar{x})$. 若以 γ 记 $(a_0, \beta')'$, 而 $\hat{\gamma}$ 则相应地记 $(\hat{a}_0, \hat{\beta}')'$, 则(4.8)式中的 $l'S^{-1}l$ 用下式代替:

$$l'S^{-1}l = l_1^2/n + l^{*'}S^{*-1}l^*, \quad l = (l_1, l^{*'})' \quad (4.13)$$

S^* 见(2.15)式.

例4.2 回归函数的同时区间估计. 写成中心化的形式: $\gamma = (a_0, \beta')'$, 此问题是要同时作一切 $l'\gamma$ 的区间估计, 其中 l 有 $(1, (x - \bar{x})')$ 的形状(注意 \bar{x} 固定, x 在 R^p 内流动), 因为 $\tilde{\mathcal{L}} = \{(0, (x - \bar{x})')': x \in R^p\}$ 构成一个维数为 p 的线性子空间, 而 $h = (1, 0, \dots, 0)' \in \tilde{\mathcal{L}}$, 故 $\mathcal{L} = \{(1, (x - \bar{x})')': x \in R^p\} = \{h + \tilde{l} : \tilde{l} \in \tilde{\mathcal{L}}\}$ 正好有系4.1的结构, $d-1 = p$ 即 $d = p+1$. 由(4.8), (4.13), 得

$$\begin{aligned} & P\{ |(\bar{Y} + \hat{\beta}'(x - \bar{x})) - (a_0 + \beta'(x - \bar{x}))| \\ & \leq \sqrt{(p+1)F_{p+1, n-p-1}(\alpha)} \hat{\sigma} \left(\frac{1}{n} + (x - \bar{x})'S^{*-1}(x - \bar{x}) \right)^{\frac{1}{2}}, \end{aligned}$$

$$\text{对一切 } x \in R^p \} = 1 - \alpha \quad (4.14)$$

此式与(4.6)比较, 不同之处只在于用 $\sqrt{(p+1)F_{p+1, n-p-1}(\alpha)}$ 代替了 $t_{n-p-1}(\frac{\alpha}{2})$ 即 $\sqrt{F_{1, n-p-1}(\alpha)}$. (4.6) 只对一个 x 成立, 而(4.14)则对一切 $x \in R^p$ 成立.

例4.3 有时, 我们要同时对回归函数在若干个点处作区间估计, 这些点的一部分坐标固定不动, 而另外的坐标则任意变化, 即这样一些点的集

$$\begin{aligned} H = \{ & a_0 + (x_1 - \bar{x}_1)\beta_1 + \dots + (x_k^0 - \bar{x}_k)\beta_k + (x_{k+1} - \bar{x}_{k+1})\beta_{k+1} \\ & + \dots + (x_p - \bar{x}_p)\beta_p : x_1^0, \dots, x_k^0 \text{ 固定, } x_{k+1}, \dots, x_p \text{ 任意} \} \end{aligned}$$

换句话说, 某些自变量取的值固定, 而另一些可自由变化. 对此只须令 $h = (1, x_1^0 - \bar{x}_1, \dots, x_k^0 - \bar{x}_k, 0, \dots, 0)'$, $\tilde{\mathcal{L}}$ 为一切形如 $(0, 0, \dots, 0, x_{k+1} - \bar{x}_{k+1}, \dots, x_p - \bar{x}_p)'$ 的向量构成的子空间, 即可纳入系4.1的范围. 此处 $d = p - k$. 因此, 只须把(4.14)中的 $(p+1)F_{p+1, n-p-1}(\alpha)$ 改为 $(p-k+1)$

$F_{p-k+1, n-p-1}(\alpha)$, 并把“对一切 $x \in R$ ”一语改为“对一切 $x \in H$ ”即可。

现在设我们只需要同时对少数几个线性函数 $l'_1\gamma, \dots, l'_k\gamma$ 作区间估计, 以 \mathcal{L} 记 l_1, \dots, l_k 生成的线性子空间, 其秩为 d , 则有 (4.8), 因而将有

$$P\left\{l'_i\hat{\gamma} - \sqrt{dF_{d, n-p-1}(\alpha)}\hat{\sigma}(l'_i\hat{S}^{-1}l_i)^{\frac{1}{2}} \leq l'_i\gamma \leq l'_i\hat{\gamma} + \sqrt{dF_{d, n-p-1}(\alpha)}\hat{\sigma}(l'_i\hat{S}^{-1}l_i)^{\frac{1}{2}}, i=1, \dots, k\right\} \geq 1-\alpha \quad (4.15)$$

就是说, 我们找到了 $l'_1\gamma, \dots, l'_k\gamma$ 的同时区间估计, 即 (4.15) 左边的不等式, 其总的置信系数不低于 $1-\alpha$ 。但到底高出多少则无法估计, 因此, 这样找出的区间估计, 可能由于不必要地增大了置信度, 而变得太粗糙了 (太长)。问题在于, 当我们给 $l'_1\gamma, \dots, l'_k\gamma$ 作区间估计时, 我们附带地给一切 $l'\gamma (l \in \mathcal{L} \text{ 如上})$ 也作了。故这方法针对性不强。难于找到简便的, 专门针对 $l'_1\gamma, \dots, l'_k\gamma$ 的方法。故前面说过: 有限个量的同时区间估计问题更麻烦。

如果 k 不很大, 则可以用另一个简单方法。按 (4.5), 有

$$P\left\{l'_i\hat{\gamma} - \sqrt{l'_i\hat{S}^{-1}l_i}\hat{\sigma}t_{n-p-1}\left(-\frac{\alpha}{2k}\right) \leq l'_i\gamma \leq l'_i\hat{\gamma} + \sqrt{l'_i\hat{S}^{-1}l_i}\hat{\sigma}t_{n-p-1}\left(\frac{\alpha}{2k}\right)\right\} = 1 - \frac{\alpha}{k} \quad i=1, \dots, k \quad (4.16)$$

因为对任意 k 个事件 A_1, \dots, A_k 有

$$P(A_1 \cap A_2 \cap \dots \cap A_k) \geq 1 - \sum_{i=1}^k (1 - P(A_i))$$

故由 (4.16) 有

$$P\left\{l'_i\hat{\gamma} - \sqrt{l'_i\hat{S}^{-1}l_i}\hat{\sigma}t_{n-p-1}\left(\frac{\alpha}{2k}\right) \leq l'_i\gamma \leq l'_i\hat{\gamma} + \sqrt{l'_i\hat{S}^{-1}l_i}\hat{\sigma}t_{n-p-1}\left(\frac{\alpha}{2k}\right), i=1, \dots, k\right\} \geq 1-\alpha \quad (4.17)$$

这样, 我们找到了 $l'_i\gamma, i=1, \dots, k$ 的同时区间估计, 即 (4.17) 左边的不等式, 其总的置信系数不低于 $1-\alpha$ 。在文献上一般把这叫 **Bonferroni 方法**。

这方法与 Scheffe 方法 (4.15) 何者为优? 就取决于 $\left(t_{n-p-1}\left(\frac{\alpha}{2k}\right)\right)^2$ 即

$F_{1,n-p-1}\left(\frac{\alpha}{2k}\right)$ 与 $dF_{d,n-p-1}(\alpha)$ 的比较, Dunn在1958年[9]曾就 $d=1, 2, \dots, 8, k=1, 2, \dots, 10, 15, 20, 50$ 及 $n-p-1=5, 10, 15, 20, 24, 30, 40, 60, 120, \infty$ 及 $\alpha=0.05$ 对二者作了比较。结果显示: 当 d 与 k 大小差不多时, 有 $F_{1,n-p-1}\left(\frac{\alpha}{2k}\right) < dF_{d,n-p-1}(\alpha)$, 只有当 k 显著地大于 d 时才反过来。因此, Scheffe方法适用于所考察的线性函数个数(即 k)较大, 但其张成的子空间维数较小的情况, 这在直观上也不难理解。这个比较也显示: 在许多常用的情况下, Bonferroni方法优于Scheffe方法。这是很有意思的: 一个经过精细论证作出的方法, 未见得比基于粗略考虑作出的方法为优。

(四) 回归直线的等距界限

再回到回归函数的同时区间估计问题。Scheffe方法确定的界限(4.14), 在几何上在空间 R^{p+1} 中定出了两个曲面。

$$L_1: \left\{ (x': \bar{Y} + \hat{\beta}'(x - \bar{x}) + \sqrt{(p+1)\overline{F}_{p+1, n-p-1}(\alpha)} \hat{\sigma} \left(\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x}) \right)^{1/2} : x \in R^p \right\}$$

$$L_2: \left\{ (x': \bar{Y} + \hat{\beta}'(x - \bar{x}) - \sqrt{(p+1)\overline{F}_{p+1, n-p-1}(\alpha)} \hat{\sigma} \left(\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x}) \right)^{1/2} : x \in R^p \right\}$$

当 $p=1$ 时, L_1 和 L_2 如图1.4.1所示。(4.14)式在几何上可表述为: 真正的回归平面落在 L_1 和 L_2 所夹的区域内的概率, 等于指定的 $1-\alpha$ 。但在应用上, 我们通常只对一个有界范围内的 x 感兴趣。因此提出问题: 可否用两个与经验回归平面平行的平面 L_1 和 L_2 去代替 L_1 和 L_2 , 使在所指定的有界范围 A 内的 x 保持上述性质? 即使

$$P\{\text{点}(x': \alpha_0 + \beta'(x - \bar{x})) \text{落在 } L_1, L_2 \text{ 之间, 对一切 } x \in A\} = 1 - \alpha$$

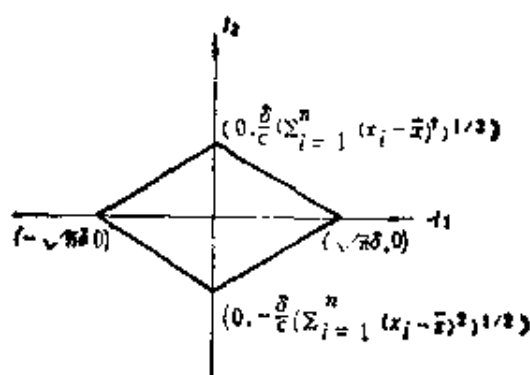


图1.4.2

Gafarian在 $p=1$ 的情况下讨论了这个问题。他取 A 为一个以 \bar{x} 为中心，长为 $2c$ 的区间。结果可表述如下：任给 $\delta > 0$ ，有

$$P\{ |(\bar{Y} + \hat{\beta}(x - \bar{x}) - a_0 - \beta'(x - \bar{x}))| \leq \hat{\delta}\sigma,$$

$$\text{对一切 } x \in [\bar{x} - c, \bar{x} + c] \} = \frac{1}{2\pi} \iint_B \left(1 + \frac{t_1^2 + t_2^2}{n-2}\right)^{-\frac{n}{2}} dt_1 dt_2$$

(4.18)

其中 B 为图1.4.2中所示的菱形。Gafarian 对 $n=4(2)20, 30, 50, \infty$, $\sqrt{n}\delta=1(0.05)2.5(0.1)4(0.2)5(0.5)7(1)10(5)20(10)50$

$$\frac{1}{c\sqrt{n}} \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{1/2}$$

$$= 1(0.1)2(0.2)3(0.4)5, 6.8, 10, 20, \infty$$

作了(4.18)式中的积分的表。在应用时，用已知的 n, c 和 $\sum_{i=1}^n (x_i - \bar{x})^2$ 之值及指定的 $1-\alpha$ 值，从表上决定 δ 。

(4.18)式证明不难，但我们不在此给出了，因为这方法在应用上意义不大。本方法在原则上也可以推广到 $p>1$ 的情况，但形式复杂得多，实际上无法造表，因而就没有多大意义了。

§1.5 预测和校准问题

(一) 预测因变量的值

在本章开始时我们就提出了这个问题，实际上，这个问题是回归分析的主要应用，虽然不是唯一的应用。问题的提法是：指定了自变量 X 之值 x ，设想我们在这个条件下做试验以观察因变量的值 y 。问预期 y 等于多少？在应用上有两种可能：一是根本没有进行所说的试验，我们只是想弄清楚：“如果”给试验以什么条件，预测它可能有什么结果。一是试验是做了，但结果要等到以后某个时间(如几个月以后)才能出来，但我们想现在就给将要出现的结果以一个预测。

为解决这个问题，我们先设想一个单纯的情况：设已知随机变量 ξ 的均值为 a 。对 ξ 进行一次观察，要预测 ξ 之值。这时，直观上告诉我们，应取 a 为预测值。因为 a 是 ξ 的最有代表性的数值，舍 a 而去用其他值，显得没有道理。这一点也有其理论上的根据：若以均方误差作为优良性准则，那么若以 b 为预测值时，预测均方误差为

$$E(\xi - b)^2 = E(\xi - a)^2 + (b - a)^2$$

此式当且仅当 $b = a$ 时达到最小，即均值 a 是预测均方误差最小者。

如果 a 未知，而我们从某种途径对 a 作出了一个估计 \hat{a} ，则根据上述分析，取 \hat{a} 作为 ξ 的预测值是合理的。

这个思想可立即用到回归预测上来。给充 $X = x$ 后，因变量 Y 的条件分布是 $Y|X = x$ 。在线性回归的假定下，此分布有均值 $a + \beta'x$ 。故当 a, β 已知时，根据上面的分析，应使用 $a + \beta'x$ 作为 y 的

预测值。一般 α, β 是未知的。若我们通过以往积累的样本 $(x_i, Y_i), i=1, \dots, n$ 对 α, β 作了估计 α^*, β^* , 则我们将用 $\alpha^* + \beta^* x$ 去预测 y . 在GM假定下, 通常使用LS估计 $\hat{\alpha}$ 和 $\hat{\beta}$, 因而我们用 $\hat{\alpha} + \hat{\beta}'x$ 作为 y 的预测值。

所以, 从形式上看, 在 $X=x$ 时预测 y 之值的解, 与估计回归函数在 x 点处之值 $\alpha + \beta'x$ 的解完全一样。但应注意二者有重要差别, 可归纳为以下两点。

1. 在估计 $\alpha + \beta'x$ 的问题中, 我们要估计的量是一个虽然未知, 但本身是确定的, 无随机性的量。而在预测 y 之值时, 预测的对象本身就是随机的。通常在统计推断理论中, 我们对“统计推断”一词的理解是: 利用样本去作出有关总体参数的结论(如总体为非参数型的, 说法略作些改变)。按这一理解, 预测问题不能纳入统计推断的范围之内, 不过习惯上人们不去坚持这个细节上的差别。

2. 从精度上说, 预测 y 的精度要比估计回归函数的精度差, 原因是被预测的量 y 可表为

$$y = \alpha + \beta'x + e \quad (5.1)$$

的形式。为预测 y , 我们分别去预测 $\alpha + \beta'x$ 和 e . 前者就用 $\hat{\alpha} + \hat{\beta}'x$, 而后者, 因为 $Ee=0$, 就用0作为预测值。故预测误差由两部分构成: 一部分是预测 $\alpha + \beta'x$ 时产生的误差, 另一部分即本身。特别是: 纵然前者在数据量充分多时可以达到任意小, 而后者则与已有的数据无关: 不论你有了多少数据, 不能把预测误差(实际上是指其方差)压到一定限度以下。因此, 若原来的回归模型 $y = \alpha + \beta'x + e$ 中误差 e 的影响显著的话, 增加观察次数对改进预测效果的作用是有限的。直接计算不难看出这一点: 以

$$\hat{y} = \hat{\alpha} + \hat{\beta}'x \quad (5.2)$$

作为 y 的预测值, 则依(5.1), 预测均方误差 (Mean Square Er-

ror of Prediction, 简记为MSEP)为

$$\begin{aligned} \text{MSEP}(\hat{y}) &= E(\hat{y} - y)^2 = E\{(\hat{\alpha} + \hat{\beta}'x - \alpha - \beta'x) - e\}^2 \\ &= \text{Var}(\hat{\alpha} + \hat{\beta}'x) + \text{Var}(e) \\ &= \sigma^2 \left\{ 1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x}) \right\} \quad (5.3) \end{aligned}$$

在推导出(5.3)式时要注意, e 与 $\hat{\alpha} + \hat{\beta}'x$ 独立, 或至少不相关。这是因为, $\hat{\alpha}$, $\hat{\beta}$ 只与在 x_1, \dots, x_n 处作的试验结果有关, 而 y 或者说 e , 只与在 x 点作的试验有关。不论假设这 $n+1$ 个试验结果 Y_1, \dots, Y_n, y 独立或不相关, 都推出(5.3)。在(5.3)式右边三项中, 后两项可随 $n \rightarrow \infty$ 而任意小(当然对 x_1, \dots, x_n 也有些条件), 但第1项那个1是不动的。

我们还可以用一个形象化的例子来说明预测 y 与估计回归函数的差别。设以 X 和 Y 分别记一个人的身高、体重。指定 $X=1.70$ 米, 把所有身高为1.70米的人找来, 构成一个子总体 A 。预测问题的意义是: 在 A 中随机地抽出一人, 要预测其体重。估计回归函数(在 $X=1.70$ 处之值)的意义是: 估计总体 A 中的人的平均体重。即使在直观上也不难设想, 后一问题要好办得多。

预测量(5.2)有以下性质:

1. \hat{y} 是 y 的无偏预测, 即 $E(y - \hat{y}) = 0$

这是因为, 我们已证明过, $\hat{\alpha}$, $\hat{\beta}$ 分别是 α , β 的无偏估计, 故 $E\hat{y} - Ey = (\alpha + \beta'x) - (\alpha + \beta'x) = 0$ 。

2. 若GM假定满足, 则在 y 的一切线性无偏预测中, y 有最小预测方差。

事实上, 若 $c'Y$ 为 y 的任一线性无偏预测(此处 $Y = (Y_1, \dots, Y_n)'$), 则有 $E(c'Y) = E(y) = \alpha + \beta'x$, 即 $c'Y$ 为 $\alpha + \beta'x$ 的线性无偏估计。但根据GM定理, $\hat{\alpha} + \hat{\beta}'x$ 在 $\alpha + \beta'x$ 的一切线性无偏估计中, 有最小方差, 即 $\text{Var}(c'Y) \geq \text{Var}(\hat{\alpha} + \hat{\beta}'x)$ 。故

$$\begin{aligned} E(c'Y - y)^2 &= \sigma^2 + \text{Var}(c'Y) \geq \sigma^2 + \text{Var}(\hat{a} + \hat{\beta}'x) \\ &= E(\hat{y} - y)^2 \end{aligned}$$

如所欲证。

3. 若误差 e_1, e_2, \dots 独立且各有正态分布 $N(0, \sigma^2)$, 则在 y 的一切无偏预测(不必为线性的)中, \hat{y} 有最小预测方差。

证明显然, 是利用在正态假定下, $\hat{a} + \hat{\beta}'x$ 在 $a + \beta'x$ 的一切无偏估计中, 有最小方差。

从预测方差的表达式(5.3)看出: 当预测点 x 落在 n 个试验点 x_1, \dots, x_n 的重心 \bar{x} 附近时, 预测方差小, 因而预测准一些。 x 与 \bar{x} 离得愈远, 预测精度愈差。一般, 人们把在试验点所在范围内使用(经验)回归方程称为“内插”, 而把在这个范围之外使用回归方程称作“外推”。上述事实告诉我们: 回归方程的内插使用较可靠, 而外推则不可靠。故一般都告诫, 不宜对回归方程作过远的外推。实际上, 理由还不止此一端: 在具体问题中, 回归函数总不会是严格线性的。在一个较小范围内, 用线性函数去逼近它还比较可靠。当有了样本以后, LS法的精神, 就是在样本所在范围内, 以一个适当的线性函数去取代原来的回归函数, LS法的平均性质照顾了试验区域内的各处(当然愈近中心 \bar{x} 处愈好), 故在试验区域内情况一般不致太差。当外推出去时, 回归函数的线性已不成立, 而LS法也没有照顾到这个范围, 效果自然就不会好。即使原回归函数确是严格线性的。但两条直线即使在一个有界区域内极接近, 在远处的分歧仍能任意大、所谓“差之毫厘, 失之千里”。

(二) 区间预测与同时区间预测

就是要找一个区间, 使 y 之值落在这区间内的概率达到指定的 $1 - \alpha$ 。为此目的要假定误差为正态的, 即假定 n 次试验的误差

e_1, \dots, e_n , 以及预定在 x 点 (预测点) 所作试验的误差 e , 全体独立且各有分布 $N(0, \sigma^2)$ 。由此假定, 及引理 4.1, 即知 $y - \hat{y}$ 与 $\hat{\sigma}^2$ 独立。此处 \hat{y} 定义如 (5.2), 而 $\hat{\sigma}^2$ 见 (2.35) 式。又因 $y - \hat{y}$ 服从正态分布, 其均值为 0, 方差为 $\sigma^2 \left(1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})' \right)$,

而 $(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-p-1}^2$, 知

$$\left((y - \hat{y}) / \sqrt{1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})'} \right) / \hat{\sigma} \sim t_{n-p-1}$$

因此有

$$\begin{aligned} P \left\{ \hat{y} - \sqrt{1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})'} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2} \right) \leq y \right. \\ \left. \leq \hat{y} + \sqrt{1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})'} \hat{\sigma} t_{n-p-1} \left(\frac{\alpha}{2} \right) \right\} \\ = 1 - \alpha \end{aligned} \quad (5.4)$$

(5.4) 式左边的不等式给出了 y 的一个预测区间, 此区间包含 y 的概率恰为 $1 - \alpha$ 。与 (4.6) 比较, 看出与回归函数 $a_0 + \beta'(x - \bar{x})$ 的置信区间相比, 长度大了 $\sqrt{\left(1 + \frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})' \right) / \left(\frac{1}{n} + (x - \bar{x})' S^{*-1} (x - \bar{x})' \right)}$ 倍, 从另一个侧面反映出, 预测 y 的精

度低于估计回归函数的精度。

如果需要同时在若干个 $x_1^0, \dots, x_k^0 (k > 1)$ 处预测试验值 y_1^0, \dots, y_k^0 , 则可将 (5.4) 用于每一个 y_i^0 。可是这样一来, 其联合置信系数就不能保持 $1 - \alpha$, 而要低于它。若要保持这个 $1 - \alpha$ 值, 就需要把预测区间加长。Lieberman 在 1961 年 (见 [10]) 提供了以下类似于 Scheffe 方法的处理办法: 记

$$\hat{y}_i^0 = \hat{a}_0 + \hat{\beta}'(x_i^0 - \bar{x}), \quad i = 1, \dots, k$$

它们分别是 y_1^0, \dots, y_k^0 的点预测, 这里我们是写成中心化的形式。
Lieberman证明了:

$$P \left\{ |y_i^0 - \hat{y}_i^0| \leq (kF_{k, n-p-1}(\alpha))^{1/2} \hat{\sigma} \left(1 + \frac{1}{n} + (x_i^0 - \bar{x})' S^{*-1} (x_i^0 - \bar{x}) \right)^{1/2}, \text{ 对 } i=1, \dots, k \text{ 同时成立} \right\} \geq 1 - \alpha \quad (5.5)$$

为证明此式, 先利用由引理4.2推出之下述结果: 若 $\xi \sim N^k(0, \sigma^2 \Lambda)$, $\Lambda > 0$, $(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-p-1}^2$, 则

$$P(\xi' \Lambda^{-1} \xi / \hat{\sigma}^2 \leq kF_{k, n-p-1}(\alpha)) = 1 - \alpha \quad (5.6)$$

记 $\Lambda^{-1} = P'P$, $W = P\xi$, 此处 P 为 k 阶满秩方阵, 上式可写为

$$P(W'W \leq \hat{\sigma}^2 kF_{k, n-p-1}(\alpha)) = 1 - \alpha \quad (5.7)$$

用(4.9), 可由此式得出

$$P\{|a'W| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} \|a\|, \text{ 对一切 } a \in R^k\} = 1 - \alpha \quad (5.8)$$

记 $l = P'a$, 并注意到当 a 跑遍 R^k 时, l 跑遍 R^k , 由(5.8)得

$$P\{|l' \xi| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} \|P'^{-1}l\|, \text{ 对一切 } l \in R^k\} = 1 - \alpha \quad (5.9)$$

但 $\|P'^{-1}l\|^2 = l'P^{-1}P'^{-1}l = l'(P'P)^{-1}l = l'\Lambda l$, 故(5.9)可写为

$$P\{|l' \xi| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} (l'\Lambda l)^{1/2}, \text{ 对一切 } l \in R^k\} = 1 - \alpha \quad (5.10)$$

以 λ_{ii} 记 Λ 的 (i, i) 元。令 $l_1 = (1, 0, \dots, 0)'$, $l_2 = (0, 1, 0, \dots, 0)'$, \dots , $l_k = (0, \dots, 0, 1)$, 则 $l_i' \Lambda l_i = \lambda_{ii}$, $l_i' \xi = \xi_i$ —— ξ 的第 i 个分量。故由(5.10), 得

$$\begin{aligned} & P\{|\xi_i| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} \lambda_{ii}^{1/2}, i=1, \dots, k\} \\ &= P\{|l_i' \xi| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} (l_i' \Lambda l_i)^{1/2}, i=1, \dots, k\} \\ &\geq P\{|l' \xi| \leq \hat{\sigma} \sqrt{kF_{k, n-p-1}(\alpha)} (l' \Lambda l)^{1/2}, \text{ 对一切 } l \in R^k\} \end{aligned}$$

$$= 1 - \alpha \quad (5.11)$$

将(5.11)式用于 $\xi = (y_1^0 - \hat{y}_1^0, \dots, y_k^0 - \hat{y}_k^0)'$, 在此处的假定下, ξ 有 k 维正态分布, 均值向量为 0, 而协方差阵的 (i, i) 元为 $\text{Var}(y_i^0 - \hat{y}_i^0) = \sigma^2 \left(1 + \frac{1}{n} + (x_i^0 - \bar{x})' S^{*-1} (x_i^0 - \bar{x}) \right)$, 于是由 (5.11) 立即得出 (5.5)。

(5.5) 给出了 y_1^0, \dots, y_k^0 的同时区间预测, 其总的置信系数不小于指定的 $1 - \alpha$, 而有可能超过它, 是以这个同时预测的可靠度比要求的过高了一些, 而以牺牲预测精度为代价 (即预测区间过长), 这情况与 (4.17) 式相似。还可以指出: 在预测问题中, 并没有与 (4.8) 式相当的结果。事实上, 在此处的正态假定下, 不难证明: 若要对一切 $x \in R^p$ 同时预测所有的 y_x (y_x 记在 x 点处观察 y 的结果), 则不论你把预测区间取得多长 (但为有限长), 总的置信系数只能为 0。

另一个可用的方法是 Bonferroni 的方法, 与 (4.16) 式相当:

$$P \{ |y_i^0 - \hat{y}_i^0| \leq \hat{\sigma} \left(1 + \frac{1}{n} + (x_i^0 - \bar{x})' S^{*-1} (x_i^0 - \bar{x}) \right)^{1/2} \\ t_{n-p-1} \left(\frac{\alpha}{2k} \right), i=1, \dots, k \} \geq 1 - \alpha \quad (5.12)$$

(5.12) 与 (5.11) 的比较, 正如 (4.16) 与 (4.15) 的比较。在此即取决于 $t_{n-p-1}^2 \left(\frac{\alpha}{2k} \right)$ 即 $F_{1, n-p-1} \left(\frac{\alpha}{2k} \right)$ 与 $k F_{k, n-p-1}(\alpha)$ 何者为大。

在上节的末尾处提到的 Dunn 的计算, 可用于解决这个问题。

(三) 预测的反问题——校准问题

此问题只用于自变量为 1 维的情况。问题是: 在 $X=x$ 处观察了因变量 Y 之值 y , 要通过 y 去估计 x 。可以这样设想: 事先并不知

道 x ，或过去有过记录但记录遗失了。或认为过去对 x 的量度不准，现想通过 y 之值来校准一下。因最后这一层意思，这问题在统计学上有时称为“校准问题”。这个问题与预测问题正好相反：后者是已知 x 预测 y ，而本问题是已知 y 去“预测”（校准） x 。

如果 X, Y 都随机，则这两个问题的地位是对称的。我们可反过来把 Y 作为自变量而 X 作为因变量。这时当然应取 X 对 Y 的回归直线，它与 Y 对 X 的回归直线并不重合。一般，这问题多用在自变量非随机的场合，因此不能象这样反转过去做。事实上，在这个场合下，校准问题的性质与预测问题很不一样：在前者， x 虽然未知，但是是一个非随机的、确定的数；在后者，预测对象 y 就是随机的。

问题的处理其实很简单：作为 x 的点估计，就用 $\hat{x} = (y - \hat{a})/\hat{\beta}$ 。为作其区间估计，从下面的等式出发：

$$P\left\{|y - \bar{Y} - \hat{\beta}(x - \bar{x})| \leq \hat{\sigma} \left(1 + \frac{1}{n} + (x - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2} t_{n-2} \left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha \quad (5.13)$$

这里，所作的假定与导出(5.4)式时作的假定相同：即假定回归函数有线性形式 $y = a_0 + \beta'(x - \bar{x})$ ，且各次试验误差独立并各有分布 $N(0, \sigma^2)$ 。已有了 y 以后，解 x 的二次方程

$$(y - \bar{Y} - \hat{\beta}(x - \bar{x}))^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n} + (x - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2\right) t_{n-2}^2 \left(\frac{\alpha}{2}\right) \quad (5.14)$$

容易看出：(5.13)左边的不等式所确定的 x 的集合（注意此不等式中，除 x 外，其余都已知），必不是空集。事实上，若 $\hat{\beta} = 0$ ，则当

$|x|$ 充分大时, x 显然属于此集; 若 $\hat{\beta} \neq 0$, 则点 $\hat{x} = \bar{x} + (y - \bar{Y})/\hat{\beta}$ 及其近旁之点, 必属于此集(这进一步肯定了此集包含不止一个点)。肯定了这个事实, 就不难确定(5.13)式左边所决定的 x 集合 T 如下:

1. 若方程(5.14)有两个实根 $c < d$, 则

a. 当 $\hat{\beta}^2 > \hat{\sigma}^2 t_{n-2}^2 \left(\frac{a}{2} \right) / \sum_{i=1}^n (x_i - \bar{x})^2$ 时, $T = [c, d]$;

b. 当 $\hat{\beta}^2 < \hat{\sigma}^2 t_{n-2}^2 \left(\frac{a}{2} \right) / \sum_{i=1}^n (x_i - \bar{x})^2$ 时, $T = (-\infty, c] \cup$

$[d, \infty)$

2. 若方程(5.14)只有一实根或无实根, 则 $T = R^1$. 问题的解最后写成形式

$$P\{x \in T\} = 1 - \alpha \quad (5.15)$$

当然, 常见的情况, 且是问题的解有意义的唯一的一种情况, 是

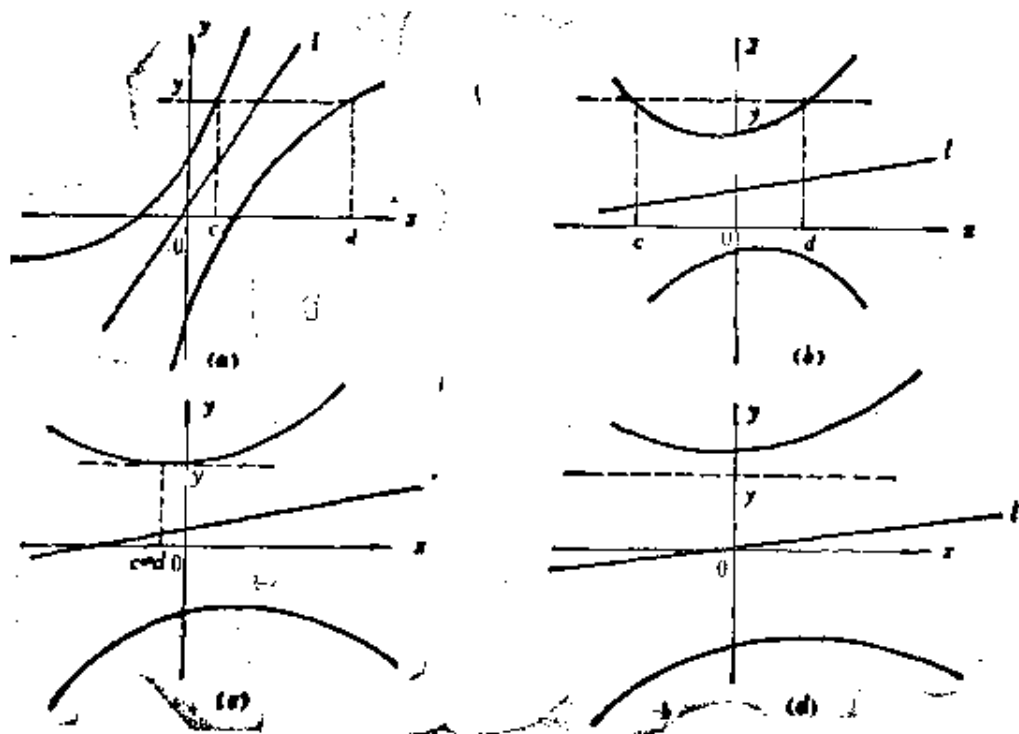


图1.4.3

1a. 其中条件的意义就是 $|\hat{\beta}|$ 不能太小(与有关的量比较而言)。这条件在直观上的意义是很清楚的: 当 $|\hat{\beta}|$ 很小时, x 变化很大还引不起 $y = \bar{Y} + \hat{\beta}(x - \bar{x})$ 多少变化, 由 y 去定 x 也就很难了。四种情况通过作图更能看清楚: 图1·4·3的(a)、(b)分别相当于上述1a、1b的情况, 而(c)、(d)则分别相当于重根及无实根的情况。图中那条直线 l 就是经验回归直线, 而两条曲线相当于图1·4·1的 l_1 和 l_2 。

与预测问题一样, 这里也有“同时校准”的问题。在 k 个点 x_1^0, \dots, x_k^0 处独立地观察了因变量之值 y_1^0, \dots, y_k^0 。但不知道 x_1^0, \dots, x_k^0 , 要利用 y_1^0, \dots, y_k^0 去估计它们。一个方法是根据(5·5)式。按照上面的方法, 每个不等式(取(5·5)中的 $p=1$, $(x_i^0 - \bar{x})'$

$$S^{*-1}(x_i^0 - \bar{x}) = (x_i^0 - \bar{x})^2 / \sum_{i=1}^k (x_i - \bar{x})^2$$

$$(y_i^0 - \bar{Y} - \beta(x_i^0 - \bar{x}))^2 \leq k F_{k, n-2}(\alpha) \hat{\sigma}^2 \left(1 + \frac{1}{n} + (x_i^0 - \bar{x})^2 / \sum_{i=1}^k (x_i - \bar{x})^2 \right)$$

决定 x_i^0 的一个集合 T_i , 而将(5·5)写为

$$P\{x_i^0 \in T_i, i=1, \dots, k\} \geq 1 - \alpha \quad (5.16)$$

这样得出的解, 其总的置信系数超过指定之值 $1 - \alpha$, 而使估计的精度下降了。与(5·5)一样, 在此要得出严格的等式是困难的*)。

也可以利用Bonferroni方法, 这只需从(5·12)式出发即可。

§1.6 自变量为随机变量的情况

在以上几节中, 我们一直假定自变量 X 是非随机的。就是说,

*) 不难指明: 不论是在预测问题还是校准问题中, 若我们愿意接受椭圆形的预测区域, 则可以得到严格的等式。但把在一个点处的预测范围与在其他点处的预测值联系起来, 显得不合理。

X 在 n 次试验或观察中取的值 x_1, \dots, x_n , 就当作已知的 p 维常向量看。我们也曾指出, 尽管在不少问题中, 自变量 X 在试验中的取值可由人控制, 因而将 X 视作非随机是合理的。但也存在不少问题, 其中自变量 X 的本性是随机的, 在这种情况下, X 的取值 x_1, \dots, x_n 本身就是随机变量。当一个统计量涉及这些 x_i 时, 该统计量的分布应结合 x_1, \dots, x_n 的分布去考虑, 而不能把它们当常数看。因此总的说, X 也是随机时, 理论复杂化了。不过, 在某些假定下, X 为非随机时所发展的方法, 也可些基本上不加改动地移用于此处。这要看模型中的具体假定如何。这一点我们在下面就会具体交代清楚。

在实际问题中, X 为随机的情况, 可分为以下两大类:

1. (X, Y) 是从一个总体中随机地抽得的个体上的某些指标值。由于抽样的随机性而带来 (X, Y) 的随机性。因此, 不论是 Y 还是 X , 其取值都无法事先规定。身高(X)和体重(Y)是一个典型的例子。对这种情况而言, X, Y 的地位可以说是平等的, 只是由于指标的具体含义及问题的着重点等等的不同, 才把其中一个变量(即 Y)突出来作为因变量。

(X, Y) 的样本, 就是从该总体中随机抽出的 n 个个体的指标 $(X_1, Y_1), \dots, (X_n, Y_n)$ (在本节中, 为强调 X 也随机, 我们把 X 的取值用大写字母记)。一般可以假定它们是独立同分布。有时还须提出另外的要求。

2. X 就其本性来说, 在试验中之值可随人控制, 不过控制不准, 或者说测量不准。例如, 这亩地播种量经秤量是15公斤, 而实际确值是14.8公斤; 这一工艺过程的反应温度本打算控制在 120°C , 而实际上是 118°C 。这种测量误差的存在, 使自变量 X 成为随机的。

由此可见, 在这种情况下, 虽则 X, Y 同属随机, 其本性却有

所不同。 Y 的随机性的来由,是因为自变量 X 不包括影响 Y 的全部因素,因而 X 之值不能唯一决定 Y 之值。比如即使确切知道一人的身高也不足以决定其体重。 Y 的这种随机性有其更本质的根由,而不仅是由量测不准而来*。至于 X ,其随机性纯来自量测。相对比起来,量测误差在一般应用中不甚重要,故为一般目的,这时把 X 视为非随机也无伤大局。只在以下两种情况之一才有其重要性:*a.* X 的量测误差大到可以与 Y 的随机性相比的程度(比如,以方差作比较标准,量测误差的方差达到 $\text{Var}(Y|X)$ 的一个可观的百分比)。*b.* Y 的随机性主要也是来自量测误差。换言之,在应用所要求的限度内, X, Y 其实是有严格函数关系(例如,严格线性关系)的,只是由于 X, Y 都存在量测误差,故在观察数据的表现上,二者无严格关系。可以举一个具体例子。为求某种金属的膨胀系数,准备该金属的一条均匀细杆。加热到温度 X ,记下这时杆长 Y 。在一般应用的限度内,可认为 X, Y 有严格线性关系 $Y = \alpha + \beta X$ 。可是,由于温度 X 与长度 Y 在量测时都有误差,故如我们做试验得到 n 组数据 $(X_i, Y_i), i = 1, \dots, n$,它们不会严格地落在一条直线上。由于这个解释,这种模型——即在变量有误差时,估计函数关系的模型,常称为**结构关系模型**。

上面这第2种情况是近几十年来数理统计学研究的一个热门题目。这里我们只介绍其一点点最基本的、有实用意义的结果,且不能都给以仔细的论证。对这方面有兴趣的读者,可参看[11], [12], [13]及其所引文献。总的说,这方面的结果在论证上都很难,而达到的结果的实用意义还不能和以上几节比拟。这是为什

*) 诚然,因变量 Y 也有测量误差的问题。但如我们假定这种量测误差(的分布)与 X 及 Y 之值无关,则等于以 $Y + \eta$ 作为 Y 的确值, η 为均值为0的随机变量,其分布一定。这样所起的效果,不过相当于增大 Y 的方差,其他一切无变化,故不构成本质上新的模型。

么这种模型还较少见诸实用的原因。

(一) X 、 Y 的联合分布为正态情况

本段设 X 为 p 维, Y 为1维, $(X', Y)'$ 为 $(p+1)$ 维非退化正态分布 $N_{p+1}(\mu, \Lambda)$ 。记 $v = (\mu_1, \dots, \mu_p)'$, 而

$$\Lambda = (\sigma_{ij}) = \begin{pmatrix} \Lambda_1 & c \\ c' & \sigma_{p+1,p+1} \end{pmatrix}$$

则在概率论中证明了: 在给定 X 之值时, Y 的条件分布为正态, 其均值为

$$E(Y|X) = \mu_{p+1} + c' \Lambda_1^{-1} (X - v) = \alpha + \beta' X \quad (6.1)$$

其中 $\alpha = \mu_{p+1} - c' \Lambda_1^{-1} v$, $\beta = \Lambda_1^{-1} c$, 而方差为

$$\text{Var}(Y|X) = \sigma_{p+1,p+1} - c' \Lambda^{-1} c = \sigma^2 \quad (6.2)$$

这些结果的证明, 可参看[8]或[14]

从(6.1)看出, Y 对 X 的回归函数是线性的, 因此, 在这个情况下谈论线性回归有意义, 而(6.2)又表明, Y 的条件方差 σ^2 与 X 无关, 这就说明, 若我们有了 (X, Y) 的独立同分布样本 $(X_1, Y_1), \dots, (X_n, Y_n)$, 而在给定 X_1, \dots, X_n 之值的条件下来考察 Y_1, \dots, Y_n , 则在这个条件化之下, $(Y_1, \dots, Y_n)'$ 构成一个如(2.3)式一样的线性回归模型, 且其中的 e 有分布 $N_n(0, \sigma^2 I_n)$ 。于是在这个条件化(即给定 X_1, \dots, X_n)之下, 一切都可以按以往的方法去做。例如为估计 $\gamma = (\alpha, \beta)'$, 可用由(2.8)式确定的LS估计 $\hat{\gamma}$, 它仍是 γ 的无偏估计*, 因为在给定 X_1, \dots, X_n 的值时, $\hat{\gamma}$ 的条件均值为 γ , 而 $\hat{\gamma}$ 的(无条件)均值, 等于其条件均值再求均值, 故仍为 γ 。又如要检验线性假设(3.1), 我们先在给定 X_1, \dots, X_n 的

* 且不难证明: 在为条件 X_1, \dots, X_n 下 $\hat{\gamma}$ 仍为 γ 的UMVUE, 证明与定理2.6的证法完全一样。

条件下进行,得到由(3.3)定义的统计量 F_H . 这统计量在给定 X_1, \dots, X_n 时的条件分布, 是一个与 X_1, \dots, X_n 之值无关的分布 $F_{n, n-p-1}$, 故后者就是 F_H 的(无条件)分布。因此, 检验(3.9)仍是(3.1)的水平 α 检验。区间估计也一样: 为要作 $\eta = A\gamma$ 的置信区域, 先在给定 X_1, \dots, X_n 的条件下证明了事件

$$\{(\hat{\eta} - \eta)' (A \tilde{S}^{-1} A')^{-1} (\hat{\eta} - \eta) / k \sigma^2 \leq F_{n, n-p-1}(\alpha)\} \quad (6.3)$$

的条件概率为 $1 - \alpha$, 故其(无条件)概率也是 $1 - \alpha$, 即(4.2)式在这里仍成立, 而(6.3)式确定一个置信系数 $1 - \alpha$ 的置信椭圆, 一如在 X 为非随机时。

当然, 也不是每一件事情都是严格平行的。例如, 要找 γ 的 LS 估计 $\hat{\gamma}$ 的分布。在 X 非随机而误差 $e \sim N_n(0, \sigma^2 I_n)$ 时, 有 $\hat{\gamma} \sim N_{p+1}(\gamma, \sigma^2 \tilde{S}^{-1})$. 在本段的假定下(X 也随机, X, Y 联合分布为 $p+1$ 维非退化正态), $\hat{\gamma}$ 的分布就复杂得多。以 $p=1$ 为例, 记 $\sigma_1^2 = \text{Var}(X), \sigma_2^2 = \text{Var}(Y), \rho = X, Y$ 的相关系数

$$(6.4)$$

则读者不难自己证明:

$$\frac{\sigma_1 \sqrt{n-1}}{\sigma_2 \sqrt{1-\rho^2}} (\hat{\beta} - \beta) \sim t_{n-1} \quad (6.5)$$

已不再是正态分布。从实用的观点看, 象(6.5)这类结果可能用得并不多, 因为通常总是按上述条件化的观点去处理问题。但(6.5)这样的“非条件化”结果也有用处, 在于对 β 的精度作一个综合性的评价。“综合”一词在此的意义是“综合”所有可能的 X 样本。详细解释如下: 设甲、乙两个人都抽 n 个样本, 都用条件化的方法去作 β 的区间估计。在给定置信系数 $1 - \alpha$ 时, 按一般公式(4.5), 知置信区间的长将与 $\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^{1/2}$ 成反比。设某甲抽

样正好碰到较大的 $\sum_{i=1}^n (X_i - \bar{X})^2$, 则他的区间短些 (精度大些), Z 抽样碰到的 $\sum_{i=1}^n (Y_i - \bar{X})^2$ 小些, 则其区间长些 (精度小些)。由于二者是在各自的不同条件 (即不同的 X 样本) 去处理的, 对“ $\hat{\beta}$ 的精度究竟好坏如何”, 就难作出统一的解释。若从 (6.5) 出发就可作出一个公认的解释。比如, 由 (6.5) 知 $\hat{\beta}$ 的方差 $\text{Var}(\hat{\beta}) = \frac{1}{n-3} \frac{\sigma_2^2(1-\rho^2)}{\sigma_1^2}$. 这个表达式制约了 $\hat{\beta}$ 的“综合”精度。

在结束这一段之前我们再补充两点。

1. 在自变量为 1 维 ($p=1$) 时, 采用记号 (6.4), 易得出 $E(Y|X) = \alpha + \beta X$ 中的回归系数 β 有表达式

$$\beta = \rho\sigma_2/\sigma_1 \quad (6.6)$$

若 $\sigma_1 = \sigma_2$, 则 $|\beta| = |\rho| < 1$. 这一点解释了我们在本书开始处在说明“回归”一词的来由时, 所谈到的由 Galton 指出现象。因为在 Galton 的例中, σ_1^2 是父代身高方差, σ_2^2 是子代身高方差。在一代之间, 身高变化不大, 故可以很准确地假定有 $\sigma_1 = \sigma_2$, 因而据上述有 $\beta = \rho$. 在此例中 $0 < \rho < 1$, 故 $0 < \beta < 1$. 暂以 a 记 $E(X)$, b 记 $E(Y)$, 从回归方程

$$E(Y|X) - b = \beta(X - a)$$

及 $0 < \beta < 1$ 看出: 当 X 大于其平均 a 时, $E(Y|X)$ 确要大于 Y 的平均 b . 但 X 增长 1 单位时, $E(Y|X)$ 并不相应增 1 单位, 而是少一些。这就从理论上给“回归到中心的趋势”这个现象以圆满的解释。当然, 这事还不象上面说起来那么轻易: 回归 (regression) 的名词 Galton 在 1886—1888 年间才使用, 而相关系数 (coefficient of correlation) 一词则是 Edgeworth 在 1892 年才开始使用的。

不言而喻, 在许多问题中可以有 $\sigma_2 > \sigma_1$, 这时 $|\beta|$ 可以大于 1,

“回归于中心”也就不成立了。

2. 本段结果可稍作推广：只要求在给定 X 之值时， Y 的条件分布有 $N(a + \beta'X, \sigma^2)$ 的形式，其中 a, β, σ^2 与 X 无关，则在有了样本 $(X_1, Y_1), \dots, (X_n, Y_n)$ 后，从固定 X_1, \dots, X_n 的条件下去考察 Y_1, \dots, Y_n ，一切与前无异。这里推广的地方，就在于并不要求 X 的分布是正态。这时，象(6.5)这类的结果自然就不再成立。

(二) 一般情况

当 X, Y 的联合分布非正态时，一般说来回归函数 $f(X) = E(Y|X)$ 并非线性(指不能表为一些参数的线性函数，见§1.1的说明。其实，在一般情况下， (X, Y) 的分布族是非参数性的，根本谈不上通常的实参数的问题)，因此也就不存在估计线性回归系数之类的问题。在这种情况下，当然没有理由非坚持用线性回归不可——我们可以就拿 $f(X)$ 本身作为估计对象，这将在第五章中讨论。目下我们想做的是，可否用另一种方式来引进一种可解释为“线性回归”的东西。

我们先证明回归函数 $f(X)$ 的一个性质，即它是在均方误差最小的意义下，用 X 的函数去逼近 Y 的最佳逼近。即对 X 的任何函数 $g(X)$ ，有

$$E(Y - g(X))^2 \geq E(Y - f(X))^2, \quad f(X) = E(Y|X) \quad (6.7)$$

证明如下：有

$$\begin{aligned} E[Y - g(x)]^2 &= E[(Y - f(x)) + (f(x) - g(x))]^2 \\ &= E(Y - f(x))^2 + E(f(x) - g(x))^2 \\ &\quad + 2E\{(Y - f(X))(f(X) - g(X))\} \end{aligned}$$

由于

$$\begin{aligned}
& E[(Y - f(X))(f(X) - g(X))] \\
&= E\{E[(Y - f(X))(f(X) - g(X)) | X]\} \\
&= E\{(f(X) - g(X))E[E(Y|X) - f(X)]\} \\
&= 0
\end{aligned}$$

得

$$E(Y - g(X))^2 = E(Y - f(X))^2 + E(f(X) - g(X))^2 \quad (6.8)$$

于是证明了(6.7).

此结果给我们一个启发:回归函数是在一切(X 的)函数类中,对 Y 的最佳逼近。这“一切函数类”太大,我们把所允许的函数局限在一定范围内,也可得出对 Y 的最佳逼近。虽在逼近效果上可能差一些,但应用上较方便,较易求出来。线性函数类就是一个自然的选择。假定 Y 的方差 $\text{Var}(Y) = \sigma^2$ 存在, X 的协方差阵 $\Lambda = (\sigma_{ij})$ 存在且为正定,又记 $\mu = EX$, $v = EY$, $\lambda = E[(X - \mu)(Y - v)]$. 令

$$H(\alpha, \beta) = E(Y - \alpha - \beta'X)^2$$

去找 α, β , 使上式达到最小值。用求偏导数并命之为0的方法, 不难得到使 H 达到最小的 α, β 为

$$\beta = \Lambda^{-1}\lambda, \quad \alpha = v - \beta'\mu \quad (6.9)$$

于是, 由(6.9)式确定的线性函数 $M(X) = \alpha + \beta'X$, 是 Y 的(在均方误差最小的意义下的)最佳线性逼近。 $Y - M(X)$ 称为剩余。由(6.9)易见 $E(Y - M(X)) = 0$, 故

$$\sigma_r^2 = \text{Var}(Y - M(X)) = E(Y - M(X))^2 \quad (6.10)$$

($= H(\alpha, \beta)$, α, β 由(6.9)确定)

它称为**最佳线性逼近的剩余方差**。如果 σ_r^2 相对于实际应用中要求的精度来说是比较小, 则用 $M(X)$ 代替 Y 是可以的, 因而 $M(X)$ 往往形式地被说成是 Y 对 X 的“线性回归”。不过, 由于事实上

$M(X)$ 并不等于 $E(Y|X)$ ，它的意义只能从上述逼近的观点去解释。如果剩余方差 σ^2 并不很小， $M(X)$ 是没有多大意义的。不比回归函数 $f(X) = E(Y|X)$ 。尽管在不少情况下 $E(Y - f(X))^2$ 并不很小，但 $f(X)$ 的基本作用——即它是 Y 在给定 X 时的均值——并不受影响。这是二者不同之处。

(6.9) 式定义的 α 、 β 依赖于 μ 、 ν 、 Λ 和 λ ，它们可通过样本去估计，可以用矩法：设有了 (X, Y) 的独立同分布样本 (X_i, Y_i) ， $i=1, \dots, n$ ，则我们容易证明：以下统计量

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \\ \hat{\Lambda} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})', \\ \hat{\lambda} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})\end{aligned}\tag{6.11}$$

分别是 μ 、 ν 、 Λ 和 λ 的无偏估计，故可用

$$\hat{\beta} = \hat{\Lambda}^{-1} \hat{\lambda}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}' \bar{X}\tag{6.12}$$

去估计 β 和 α ，它们已不再有无偏性，但由大数定律知，(6.11) 是 μ 等的强相合估计，因而 (6.12) 是 β 和 α 的强相合估计。除此以外，由于我们对 (X, Y) 的分布没有作多少假定，进一步的统计推断问题——检验和区间估计问题，就无法讨论了。而且，在目前情况下，关于 α 、 β 的线性假设也丧失了其实际背景，没有多大意义。基本上我们能做的事，就是用 (6.12) 去估计出 α 、 β ，并对剩余方差 σ^2 作一估计（见下）。如估计值较小，则可以把 $\hat{\alpha} + \hat{\beta}' \bar{X}$ 付诸应用，如此而已。

为估计 σ^2 ，有两个方法：一是 σ^2 的表达式 (6.11) 直接提示我们用

$$\hat{\sigma}_r^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{M}(X_i))^2, \quad \hat{M}(X) = \hat{\alpha} + \hat{\beta}'X \quad (6.13)$$

第二种方法是：利用(6.9)和(6.10)算出

$$\sigma_r^2 = \sigma^2 - \lambda' A^{-1} \lambda \quad (6.14)$$

用 $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ 估计 $\sigma^2 = \text{Var}(Y)$ ，其余的 λ ， A 分别用

(6.11)估计，得 σ_r^2 的估计

$$\tilde{\sigma}_r^2 = \hat{\sigma}^2 - \hat{\lambda}' \hat{A}^{-1} \hat{\lambda} \quad (6.15)$$

我们留给读者去证明下述简单事实：

$$\tilde{\sigma}_r^2 = \frac{n}{n-1} \hat{\sigma}_r^2 \quad (6.16)$$

因此二者只相差一个常数因子 $\frac{n}{n-1}$ 。这二者都不是 σ_r^2 的无偏估计，也难说那一个好些。当 n 较大时二者实质上无差别，当 n 不大时，按习惯上的考虑可用(6.15)。最后我们把上述未证明的几点的证明途径提一下，建议读者循此把仔细证明写出来：

1. (6.14)的证明：以(6.9)代入 $E(Y - \alpha - \beta'X)$ ，得 $\sigma_r^2 = E[(Y - \nu) - \beta'(X - \mu)]^2$ ，把平方展开逐项求均值。

2. 注意到(6.12)式就是在§1.2意义下， α ， β 的LS估计，于是 $\sum_{i=1}^n (Y_i - \hat{M}(X_i))^2$ 就是在§1.2(五)所述意义下的残差平方和(2.30)。

3. 利用残差平方和的表达式(2.32)，但要在中心化以后去使用它(中心化不改变残差平方和)：以 $(Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})'$ 代替 $Y = (Y_1, \dots, Y_n)'$ ，以(2.15)式的 X^* 和 S^* 代替(2.32)式中的 \tilde{X} 和 \tilde{S} 。这样就不难看出：(6.15)定义的 $\tilde{\sigma}_r^2$ ，等于残差平方和除以

$(n-1)$.

相关比

在统计教本中，常提到的用来刻划两变量的相关性的指标，是相关系数。相关比也是一种刻划变量相关性的指标，有着某些与相关系数不同的特点。在这里顺便讨论一下并无困难。以下设 X 的维数为 1。

仍以 $f(X)$ 记回归函数 $E(Y|X)$ ，此处并不假定 $f(X)$ 为线性。在 (6.8) 式中，以 $E(Y) = v$ 代替 $g(X)$ ，得

$$\text{Var}(Y) = E(Y - f(X))^2 + E(f(X) - v)^2 \quad (6.17)$$

(6.17) 式把 $\text{Var}(Y)$ 分解成两部分。其第一部分 $E(Y - f(X))^2$ ，是 Y 的最佳均方逼近 $f(X)$ 的剩余方差（易见 $E(Y - f(X)) = 0$ ，故 $E(Y - f(X))^2 = \text{Var}(Y - f(X))$ ）。这一部分愈小，就表示剩余 $Y - f(X)$ 很小，换句话说，在 Y 中消除由 X 的影响所导致的部分后，剩下已不多，这表明 X 与 Y 的关系密切。类似地，当剩余方差较大时，表明 X 对 Y 的影响不大，即 X 与 Y 的关系不密切。依这个考虑，把

$$\begin{aligned} \eta_{Y \cdot X}^2 &= 1 - E(Y - f(X))^2 / \text{Var}(Y) \\ &= E(f(X) - v)^2 / \text{Var}(Y) \end{aligned}$$

定义为 Y 对 X 的**相关比**。由于 $E(f(X)) = E(Y) = v$ ，有 $E(f(X) - v)^2 = \text{Var}(f(X))$ ，故 $\eta_{Y \cdot X}^2$ 可写为两方差之比：

$$\eta_{Y \cdot X}^2 = \text{Var}(f(X)) / \text{Var}(Y)$$

类似地可定义 X 对 Y 的相关比 $\eta_{X \cdot Y}^2$ 。值得注意的是， $\eta_{Y \cdot X}^2$ 与 $\eta_{X \cdot Y}^2$ 一般并不相等（例见下）。

相关比 $\eta_{Y \cdot X}^2$ 有以下的性质：

1. $0 \leq \eta_{Y \cdot X}^2 \leq 1$ ，这显然。

2. $\eta_{Y \cdot X}^2 = 1$ 当且仅当 Y 完全由 X 确定，即以概率 1 有 $Y = f(X)$ 。

$\eta^2_{Y \cdot X} = 0$ 当且仅当(以概率1) $f(X) = v$, 即 $E(Y|X)$ 不依赖于 X 。后者的直观意义是: 当 X 变化时, Y 的(条件)均值毫无反应, 这从一个侧面反映出 X 与 Y 的关系不大。

若以 ρ 记 X, Y 的(线性)相关系数, 则 $\rho^2 = 1$, 当且仅当 Y 与 X 有严格线性关系, 正是基于这一点, 把 ρ 理解为 X, Y 之间线性关系程度的刻划。相关比则要广一些: 不论 Y 与 X 有什么函数关系, 都导致 $\eta^2_{Y \cdot X} = 1$ 。故可以说, 相关比是 Y 与 X 的“一般”关系程度的刻划。

3. 若 X, Y 独立, 则 $\eta^2_{Y \cdot X} = 0$ 。但当 $\eta^2_{Y \cdot X} = 0$ 时并不能推出 X, Y 独立, 当 (X, Y) 为二维正态分布时这二者等价。

这一条的证明很容易, 留给读者, 为了说明 $\eta^2_{Y \cdot X} = 0$ 不一定推出 X, Y 独立, 可举单位圆内的均匀分布作例子。这一条说明: 虽则从刻划关系的广度这一点上看, 相关比似比相关系数有所改进, 但仍不能免除这个基本缺点: 即使 $\eta^2_{Y \cdot X} = 0$, Y 与 X 仍可存在“某种”关系——某种不能在 $\eta^2_{Y \cdot X}$ 这个量上面反映出来的关系。这从根本上说也许并不能算是相关比的缺点, 因为两随机变量的“关系”的性质多种多样, 并非一个简单定义的常数所能全面刻划。

4. $\eta^2_{Y \cdot X} \geq \rho^2$, 等号当且仅当 $f(X)$ 有 $\alpha + \beta X$ 的形式时才成立, 即除非 Y 对 X 的回归函数确为线性的, 相关比总大于(相关系数)²。

证明如下: 以 $\alpha + \beta X$ 记 Y 的线性逼近中均方误差最小者。按(6.8), 有

$$E(Y - f(X))^2 \leq E(Y - \alpha - \beta X)^2 \quad (6.18)$$

但由(6.10), (6.14), 并使用记号(6.4), 得

$$E(Y - \alpha - \beta X)^2 = \sigma^2_Y(1 - \rho^2) \quad (6.19)$$

因 $\sigma^2_Y = \text{Var}(Y)$, 由(6.18), (6.19), 得

$$\begin{aligned} 1 - \eta^2_{Y \cdot X} &= E(Y - f(X))^2 / \text{Var}(Y) \leq \sigma^2_Y(1 - \rho^2) / \sigma^2_Y \\ &= 1 - \rho^2 \end{aligned}$$

因而有 $\eta^2_{Y \cdot X} \geq \rho^2$ 。若 $f(X)$ 就等于 $a + \beta X$ ，自然有 $\eta^2_{Y \cdot X} = \rho^2$ 。

直观上说这结果是不难理解的。因 $\eta^2_{Y \cdot X}$ 是刻划 Y 与 X 的“一切”关系，而 ρ 只刻划线性关系。有可能， Y 与 X 毫无线性关系，但有密切的函数关系，甚至有可能： $\eta^2_{Y \cdot X} = 1$ 而 $\rho^2 = 0$ 。

例6.1 设 (X, Y) 的联合分布是

$$P(X=0, Y=0) = P(X=1, Y=1) = P(X=-1, Y=1) = \frac{1}{3}$$

则易见 $\rho = 0$ 。又因 X 唯一决定 Y ，有 $\eta^2_{Y \cdot X} = 1$ 。

当 $\eta^2_{Y \cdot X} > \rho^2$ 时，表示 Y 与 X 之间有一种非线性的关系。当然，无法指出这种关系的性质。这还要依靠对回归函数 $E(Y|X)$ 直接进行估计，见第五章。

从以上几条性质看，给人的印象是，用相关比刻划变量的关系，似优于用相关系数。但为什么在应用上，相关比远不如相关系数流行呢？原因有以下一些：

1. 在最重要的正态情况，二者（指 $\eta^2_{Y \cdot X}$ 和 ρ^2 ）并无区别，且这时 ρ 还有一个优点：其符号指出了相关的方向（正、负相关）。在非正态情况下，一般说来，相关性指标的意义大大减少。

2. 相关比缺乏对称性：一般 $\eta^2_{Y \cdot X}$ 不等于 $\eta^2_{X \cdot Y}$ 甚至有可能，一者为1而他者为0。读者不难验证：例6.1就是这个情况*）。

3. 相关系数只涉及到 X, Y 的二阶矩，而相关比就复杂得多，这在理论和应用上都造成许多不便。例如，不难算出相关系数的均值、方差的渐近表达式，而对相关比则不行。尤其是，相关系数可以通过样本矩去估计，对样本 $(X_i, Y_i), i=1, \dots, n$ 的形式无要求。相关比则不然，它必须有大量数据以构成分组形式，然后才能用于估计。

*) [11]的P.318断言，由 $\eta^2_{Y \cdot X}$ 和 $\eta^2_{X \cdot Y}$ 中的一个为1，必推出另一个为1。本例明这个论断是错误的。

4. 相关系数的作用并非相关比所能完全代替的。因前者特别着眼于线性关系，而线性关系是应用上最常见，最重要的一种关系。

由于这些原因，相关比在统计学中的地位不如相关系数，也就可以理解了。

(三) 结构关系模型

现在转向于讨论在本节引言部分中指出的第二种情况，即就其本性而言， X 之取值本是可由人控制的，但在测量上有误差。为说法上的方便，我们把因变量 Y 的误差也说成是量测误差。

我们先讨论自变量为1维的情况，把模型中的假定详细陈述如下。

假定 X 与 Y 有严格的线性关系

$$Y = \alpha + \beta X \quad (6.20)$$

α 、 β 自然是未知的。注意(6.20)式丝毫不带误差。我们一共进行 n 次观测，在第 i 次观测时， X 所取的真正值是 X_i ，而从仪表上读出的值是 ξ_i 。实际情况是这样的：做第 i 次试验时，原来我们计划让 X 取某个值 c_i ，但因控制不准，在该次试验中 X 取的值实际上是 X_i ，而又因量测不准，仪表上读出的值不是 X_i 而是 ξ_i 。这样， X_i 未知而 ξ_i 则是已知。二者的关系是

$$\xi_i = X_i + e_i, \quad i = 1, \dots, n \quad (6.21)$$

既然 X 在第 i 次试验中取的值是 X_i ，根据(6.20)，在这次试验中 Y 应取值 $Y_i = \alpha + \beta X_i$ ，可由于量测有误差，实际读出的值是 η_i 。二者的关系是

$$\eta_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, \dots, n \quad (6.22)$$

我们手头的样本是 (ξ_i, η_i) ， $i = 1, \dots, n$ ，而 α 、 β 及 X_1, \dots, X_n 都是未知参数。当然，我们主要关心的参数是 α, β ，至于诸 X_i ，

多少有点“不速之客”的性质：只要你作一次观察，就有这样一位不速之客来临。在某些情况下， X_i 之值也是试验者关心的。但不论怎样，这些未知的 X_i 的出现，使模型大大复杂化了。

e_i, ε_i 等测量误差都是随机变量，我们假定它们全体(共 $2n$ 个)独立， $e_i \sim N(0, \sigma_1^2)$ ， $i=1, \dots, n$ ； $\varepsilon_i \sim N(0, \sigma_2^2)$ ， $i=1, \dots, n$ 。等一会我们会指明： σ_1^2 和 σ_2^2 不能完全未知，需要知道其比 σ_2^2/σ_1^2 。通常就假定

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \quad (6.23)$$

这个假定在实际应用中未必总合理，因为 X, Y 的量测误差未必可以认为相当(X, Y 两个量的性质可以很不一样)。不过，如知道 σ_2^2/σ_1^2 ，不难转化为(6.23)的情况。

以下我们就一直在所有上述假定之下来讨论问题，不再随处指明。在文献中有时把这种模型叫**结构关系模型** (Structural relationship model)，它受到许多研究者的注意，积累下来的文献很多。

首先来考虑诸参数的极大似然估计。由以上假定，知 (ξ_i, η_i) ， $i=1, \dots, n$ 的概率密度函数是

$$\begin{aligned} & f(\xi_i, \eta_i, X_i, i=1, \dots, n, \alpha, \beta, \sigma) \\ &= \left(\frac{1}{\sqrt{2\pi}} \right)^{2n} \sigma^{-2n} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (\xi_i - X_i)^2 \right. \right. \\ & \quad \left. \left. + \sum_{i=1}^n (\eta_i - \alpha - \beta X_i)^2 \right) \right] \end{aligned} \quad (6.24)$$

于是 $\alpha, \beta, X_i, i=1, \dots, n$ 的极大似然估计，要使表达式

$$\begin{aligned} & F(\xi_i, \eta_i, X_i, i=1, \dots, n, \alpha, \beta) \\ &= \sum_{i=1}^n (\xi_i - X_i)^2 + \sum_{i=1}^n (\eta_i - \alpha - \beta X_i)^2 \end{aligned} \quad (6.25)$$

达到最小。在求解这个极值问题前，我们先注意下列事实：若在

上述模型中我们不假定 $\sigma_1 = \sigma_2$ ，则(6·24)应当用

$$f(\xi_i, \eta_i, X_i, i=1, \dots, n, \alpha, \beta, \sigma_1, \sigma_2) \\ = \left(\frac{1}{\sqrt{2\pi}} \right)^{2n} \sigma_1^{-n} \sigma_2^{-n} \exp \left[-\frac{1}{2\sigma_1^2} \sum_{i=1}^n (\xi_i - X_i)^2 \right] \exp \\ \left[-\frac{1}{2\sigma_2^2} \sum_{i=1}^n (\eta_i - \alpha - \beta X_i)^2 \right]$$

来代替。为求 $X_i, i=1, \dots, n$ 及 $\alpha, \beta, \sigma_1, \sigma_2$ 的极大似然估计，应使这表达式达到最大。但这个极值问题无解。因为，令 $X_i = \xi_i$ 而 $\sigma_1 \rightarrow 0$ ，可使这表达式任意大。由此可知， σ_1 和 σ_2 不能完全未知，一如我们在前面所提到的。

再回到使(6·25)达到最小的极值问题，它实际上是我们模型下的最小二乘原则的使用。因由(6·21)及(6·22)，有 $\sum_{i=1}^n (e_i^2 + \varepsilon_i^2) = (6·25)$ 右边。按最小二乘原则，这正是应让之达到最小的表达式。在此还要澄清一个问题。把(6·21)解出的 $X_i = \xi_i - e_i$ 代入(6·22)，得

$$\eta_i = \alpha + \beta \xi_i + \delta_i, \quad \delta_i = e_i - \beta e_i, \quad i=1, \dots, n \quad (6·26)$$

这里，按我们的假定， $\delta_i, \dots, \delta_n$ 为独立同分布，且每一个有均值为0的正态分布。因此，初一看可能会认为：经过这样一个手续，就把 (ξ_i, η_i) 化成为一个简单的正态线性模型。然而不然，其一， ξ_i 是随机变量。这还不是最要紧的，因为在前面我们已指出过用“条件化”的方法去处理问题，在正态场合下并无区别。其二，这是关键之一点，就是 ξ_i 与 δ_i 并不独立，尤其是 $E(\delta_i | \xi_i)$ 并不为0，因而 $E(\eta_i | \xi_i)$ 并不等于 $\alpha + \beta \xi_i$ —— $\alpha + \beta \xi_i$ 根本不是 η 对 ξ 的回归函数，故用以前的方法去处理完全失掉了根据。 $E(\delta_i | \xi_i)$ 的计算如下：

$$E(\delta_i | \xi_i) = E(e_i | \xi_i) - \beta E(e_i | \xi_i)$$

因 e_i 与 ξ_i 独立因而和 ξ_i 也独立, 故 $E(e_i|\xi_i) = Ee_i = 0$. 又因 $e_i = \xi_i - X_i$ 而 X_i 为常数, 有 $E(e_i|\xi_i) = \xi_i - X_i$, 最后得 $E(\delta_i|\xi_i) = -\beta(\xi_i - X_i)$, 不为0(只有当 X_i 的量测无误差时, 才有 $E(\delta_i|\xi_i) = 0$. 不言而喻, 这回到了习见的情况)。

提醒这些是为了使读者看到这模型的复杂性, 不要以为它是通常正态线性模型的轻易推广。

再回到使(6.25)达到最小的问题。求 $\partial F/\partial X_i$, $\partial F/\partial \alpha$ 和 $\partial F/\partial \beta$ 并命之为0, 得方程组

$$\xi_i - X_i + \beta(\eta_i - \alpha - \beta X_i) = 0, \quad i=1, \dots, n \quad (6.27)$$

$$\sum_{i=1}^n (\eta_i - \alpha - \beta X_i) = 0 \quad (6.28)$$

$$\sum_{i=1}^n X_i (\eta_i - \alpha - \beta X_i) = 0 \quad (6.29)$$

以下我们以 $\hat{\alpha}$, $\hat{\beta}$, \hat{X}_i , $i=1, \dots, n$ 记这方程组的解, 并以 \bar{X} 记 $\sum_{i=1}^n \hat{X}_i/n$, $\bar{\xi}$, $\bar{\eta}$ 的意义类推。由(6.28)得

$$\bar{\eta} = \hat{\alpha} + \hat{\beta} \bar{X} \quad (6.30)$$

把(6.27)的 n 个式子相加并用(6.28), 得

$$\bar{X} = \bar{\xi}$$

从(6.30)解出 $\hat{\alpha} = \bar{\eta} - \hat{\beta} \bar{X}$ 并代入(6.27), 再利用上式把 $\xi_i - \hat{X}_i$ 写成 $\xi_i - \bar{\xi} - (\hat{X}_i - \bar{X})$, (6.27)成为

$$\begin{aligned} & \xi_i - \bar{\xi} - (\hat{X}_i - \bar{X}) + \hat{\beta} (\eta_i - \bar{\eta} - \hat{\beta} (\hat{X}_i - \bar{X})) \\ & = 0 \end{aligned} \quad (6.31)$$

解出

$$\hat{X}_i - \bar{X} = [\xi_i - \bar{\xi} + \hat{\beta} (\eta_i - \bar{\eta})] / (1 + \hat{\beta}^2) \quad (6.32)$$

利用(6.28), 可将(6.29)改写为

$$\sum_{i=1}^n (X_i - \bar{X}) (\eta_i - \hat{\alpha} - \hat{\beta} \hat{X}_i) = 0 \quad (6.33)$$

以 $\hat{\alpha} = \bar{\eta} - \hat{\beta} \bar{X}$ 代入, 然后用(6.32)式, 易将(6.33)式整理为

$$\sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) \hat{\beta}^2 + \left(\sum_{i=1}^n (\xi_i - \bar{\xi})^2 - \sum_{i=1}^n (\eta_i - \bar{\eta})^2 \right) \hat{\beta} - \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) = 0 \quad (6.34)$$

记 $S_1^2 = \sum_{i=1}^n (\xi_i - \bar{\xi})^2$, $S_2^2 = \sum_{i=1}^n (\eta_i - \bar{\eta})^2$, $S_{12} = \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$, 可将(6.34)的解写为

$$\hat{\beta} = \{S_1^2 - S_2^2\} \pm [(S_1^2 - S_2^2)^2 + 4S_{12}^2]^{1/2} / (2S_{12}) \quad (6.35)$$

(6.25)的最小值必在有限的 X_i , α , β 处达到, 且必为方程组(6.27)–(6.29)之解。因而, 使(6.25)达到最小的 β , 只能是(6.35)的两个解之一。我们来证明: 应在(6.35)的分子中取正号。

为证实这一点, 把 $\xi_i - X_i$ 写为 $\xi_i - \bar{\xi} - (\hat{X}_i - \bar{X})$ (因为 $\bar{\xi} = \bar{X}$), 由(6.31)式得

$$[\xi_i - \bar{\xi} - (\hat{X}_i - \bar{X})]^2 = \hat{\beta}^2 [\eta_i - \bar{\eta} - \hat{\beta}(\hat{X}_i - \bar{X})]^2$$

又 $\eta_i - \hat{\alpha} - \hat{\beta} \hat{X}_i = \eta_i - \bar{\eta} - \hat{\beta}(\hat{X}_i - \bar{X})$, 因而(6.25)式成为

$$H = (1 + \hat{\beta}^2) \sum_{i=1}^n [\eta_i - \bar{\eta} - \hat{\beta}(\hat{X}_i - \bar{X})]^2$$

再以(6.32)式代入上式并稍加整理, 得

$$\begin{aligned} H &= (1 + \hat{\beta}^2)^{-1} \sum_{i=1}^n (\eta_i - \bar{\eta} - \hat{\beta}(\xi_i - \bar{\xi}))^2 \\ &= (1 + \hat{\beta}^2)^{-1} (S_2^2 + S_1^2 \hat{\beta}^2 - 2S_{12} \hat{\beta}) \end{aligned} \quad (6.36)$$

由此式可知, 欲 H 达到最小, $\hat{\beta}$ 必须与 \hat{S}_{12} 同号。再看(6.35)式知, 要 $\hat{\beta}$ 与 \hat{S}_{12} 同号, 必须右边的分子非负, 而这只有在分子中的“ \pm ”号处取“ $+$ ”号才行。这样最后证明了 β 的极大似然估计, 也就是 LS 估计, 为

$$\hat{\beta} = \{(S_2^2 - S_1^2) + [(S_2^2 - S_1^2)^2 + 4S_{12}^2]^{1/2}\} / (2S_{12}) \quad (6.37)$$

既得 $\hat{\beta}$, $\hat{\alpha}$ 和 \hat{X}_i 等次第得出:

$$\hat{\alpha} = \bar{\eta} - \hat{\beta} \bar{\xi} \quad (6.38)$$

$$\hat{X}_i = \bar{\xi} + (\xi_i - \bar{\xi} + \hat{\beta}(\eta_i - \bar{\eta})) / (1 + \hat{\beta}^2) \quad (6.39)$$

(6.39)式可用于修正最初从仪表上读出的 ξ_i 。若采用(6.26)的记法, 则不难得出

$$\hat{X}_i = \xi_i + \frac{\hat{\beta}}{1 + \hat{\beta}^2} (d_i - \bar{d}) \quad (6.40)$$

此式更明显看出 \hat{X}_i 作为 ξ_i 的“修正”的作用。

σ^2 的极大似然估计不难求得。只须把(6.24)式改为 $(\sqrt{2\pi})^{-2n} \sigma^{-2n} \exp\left(-\frac{H}{2\sigma^2}\right)$, 其中 H 由(6.36)式确定, 就求得这极大似然估计 $\hat{\sigma}^2$ 为

$$\hat{\sigma}^2 = \frac{1}{2n} H, \quad (H \text{ 由 (6.36)、} \hat{\beta} \text{ 由 (6.37) 确定}) \quad (6.41)$$

然而, 这个估计必须加以修正。因为可以证明(见以下(五)), 这估计甚至不是相合的。将其修正为

$$\tilde{\sigma}^2 = \frac{1}{n-2} H = \frac{1}{n-2} (S_2^2 + S_1^2 \hat{\beta}^2 - 2S_{12} \hat{\beta}) / (1 + \hat{\beta}^2) \quad (6.42)$$

则在比较宽广的条件下可证明其强相合性, 即以概率1有 $\tilde{\sigma}^2 \rightarrow \sigma^2$ 当 $n \rightarrow \infty$ 。证明也见(五)。这个修正其实不难理解: 模型(6.21)、(6.22)统共有 $2n$ 个观察值 $\xi_i, \eta_i, i=1, \dots, n$, 而参数(不算 σ^2)有 $\alpha, \beta, X_i, i=1, \dots, n$ 等 $n+2$ 个。故留给 σ^2 的估计的自由度, 只有 $2n - (n+2) = n-2$ 。

在与证明 $\tilde{\sigma}^2$ 的相合性的同样条件下,也可以证明 $\hat{\alpha}$ 、 $\hat{\beta}$ 的强相合性,证明也见(五)。更进一步还可以证明这些估计的渐近正态性,其结果将在(四)中在更一般的模型中陈述而不给证明。利用这些大样本性质,可以在 n 很大时对 α 、 β 进行检验和构造置信区间。然而,这些估计的小样本性质则很困难,可以说实质上所知极少。这不难理解:这个模型因包含了众多的附带参数 X_1, \dots, X_n (其个数随样本大小 n 而增加!)而变得极为复杂,不是好对付的。

例6.2 (本例取自[11], P.412), 从确定的关系式

$$Y = 2X + 4 \quad (6.43)$$

出发,通过模拟得9组数据 $(\xi_1, \eta_1), \dots, (\xi_9, \eta_9)$ 。方法如下:指定9个数 X_1, \dots, X_9 ,从随机正态数表(其表中的数按 $N(0, 1)$ 而分布,故 $\sigma^2 = 1$)中抽取18个数 $e_1, \dots, e_9, e_1, \dots, e_9$,而后令 $\xi_i = X_i + e_i, \eta_i = 2X_i + 4 + e_i, i = 1, \dots, 9$ 。得到9组数据为:

(1.8, 6.9), (4.1, 12.5), (5.8, 20.0), (7.5, 15.7)

(9.3, 24.9), (10.6, 23.4), (13.4, 30.2), (14.7, 35.6),

(18.9, 39.1)

算出 $\bar{\xi} = 9.57, \bar{\eta} = 23.14, S_{\xi}^2 = 238, S_{\eta}^2 = 906, S_{\xi\eta} = 451$ 。于是由(6.37)、(6.38)分别算出 α, β 的估计为

$$\begin{aligned} \hat{\beta} &= \{ (906 - 238) + [(906 - 238)^2 + 4(451)^2]^{1/2} \} / (2 \times 451) \\ &= 1.99 \end{aligned}$$

$$\hat{\alpha} = 23.14 - 1.99 \times 9.57 = 4.01$$

此与(6.43)很接近。若按通常LS法,将得到 β 和 α 的估计分别为1.89和5.05,差距就较大。又利用(6.42)式得到 σ^2 的估计值

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{9-2} \{ 906 + 238(1.99)^2 - 2 \times 451 \times 1.99 \} / (1 + 1.99^2) \\ &= 1.53 \end{aligned}$$

此值超过1较多,这是否表示估计(6.42)有偏高的倾向?因本例数据少还不好说,也许更大规模的模拟有助于了解这个问题。

以上我们全是在 $\sigma_1^2 = \sigma_2^2$ 的条件下讨论的。一般, 若 $\sigma_2 = \lambda\sigma_1$ 而 λ 已知, 则可令

$$\eta'_i = \eta_i/\lambda, \quad \alpha' = \alpha/\lambda, \quad \beta' = \beta/\lambda, \quad \varepsilon'_i = \varepsilon_i/\lambda$$

而将(6.22)写为

$$\eta'_i = \alpha' + \beta' X_i + \varepsilon'_i, \quad i=1, \dots, n \quad (6.44)$$

ε'_i 的方差为 $\sigma_1^2/\lambda^2 = \sigma_1^2$, 故(6.21)和(6.44)联合成的模型, 适合等方差条件。利用前述方法, 通过 (ξ_i, η'_i) , $i=1, \dots, n$, 可对 α' , β' , X_i 和 σ_1^2 进行估计, 从而得到 α , β , σ_1^2 的估计。 λ 这个值在应用上可通过以往积累的经验, 通过对测量仪器精度的了解去确定之。它不能作为一个未知参数出现在模型中, 这一点在前面已交代过了。

(四) 自变量 X 为多维时(续(三))的讨论

当自变量 X 的维数 $p > 1$ 时, 模型(6.21)、(6.22)显然地推广为以下的形式:

$$\xi_i = X_i + e_i, \quad \eta_i = \alpha + \beta' X_i + \varepsilon_i, \quad i=1, \dots, n \quad (6.45)$$

(6.45)中各个量的性质与前无异, 只是 ξ_i , X_i , e_i 和 β 等如今为 p 维。仍假定 e_i , ε_i , $i=1, \dots, n$ 全体独立, $e_i \sim N_p(0, \sigma^2 I_p)$, $\varepsilon_i \sim N(0, \sigma^2)$, 这是等方差情况。方差不等但知道其比时, 不难通过简单变换化归(6.45)的情况。

下面我们不加证明列举关于此模型的几个基本结果。证明可在文献[12]中找到。

1. α 、 β 的极大似然估计

按对模型(6.45)的假定, α 、 β 的极大似然估计, 是要使表达式

$$H = \sum_{i=1}^n \| \xi_i - X_i \|^2 + \sum_{i=1}^n (\eta_i - \alpha - \beta' X_i)^2 \quad (6.46)$$

达到最小。其解的构造过程如下：先作 $p+1$ 阶方阵

$$W = \begin{pmatrix} \sum_{i=1}^n (\xi_i - \bar{\xi})(\xi_i - \bar{\xi})' & \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) \\ \sum_{i=1}^n (\xi_i - \bar{\xi})'(\eta_i - \bar{\eta}) & \sum_{i=1}^n (\eta_i - \bar{\eta})^2 \end{pmatrix} \quad (6.47)$$

以 $\lambda_1 \geq \dots \geq \lambda_{p+1} \geq 0$ 记 W 的特征根, Λ 为对角阵 $\text{diag}(\lambda_1, \dots, \lambda_{p+1})$. 找 $p+1$ 阶正交方阵 $P = (p_{ij})$, 使

$$W = P \Lambda P' \quad (6.48)$$

则 α 、 β 的极大似然估计 $\hat{\alpha}$ 、 $\hat{\beta}$ 分别为

$$\hat{\beta} = - (p_{1,p+1}, \dots, p_{p,p+1})' / p_{p+1,p+1}, \quad \hat{\alpha} = \bar{\eta} - \hat{\beta}' \bar{\xi} \quad (6.49)$$

2. σ^2 的估计

表达式 (6.46) 的最小值 $\min H$ 等于矩阵 W 的最小特征根 λ_{p+1} . 又一共有 $n(p+1)$ 个数据, 而未知参数 X_i , α , β (不算 σ^2 本身) 共有 $np + p + 1$ 个。剩下自由度为 $n - p - 1$. 故用

$$\tilde{\sigma}^2 = \frac{1}{n - p - 1} \lambda_{p+1} \quad (6.50)$$

作为 σ^2 的估计

X_1, \dots, X_n 的估计也可写出, 此处不赘.

3. 相合性 在本模型及下面两条假定之下, $\hat{\alpha}$ 、 $\hat{\beta}$ 、 $\tilde{\sigma}^2$ 分别是 α 、 β 、 σ^2 的强相合估计:

$$a. \lim_{n \rightarrow \infty} \bar{X} = c \text{ 存在有限, } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$b. \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' = \Delta \text{ 存在有限且为正定.}$$

此定理中正态性条件还可放宽, 只须要求 $(e'_i, \varepsilon_i)'$, $i=1, 2, \dots$ 为独立同分布, 且对每个 i , $(e'_i, \varepsilon_i)'$ 的 $p+1$ 个分量独立, 有

均值向量 0 和协方差阵 $\sigma^2 I_{p+1}$.

4. 渐近正态性 仍假定上述条件 a, b , 则当 $n \rightarrow \infty$ 时, 有

1°. $\sqrt{n}(\hat{\alpha} - \alpha, (\hat{\beta} - \beta)')'$ 依分布收敛于 $N_{p+1}(0, G)$, 其中 $G = (g_{ij})$ 而

$$g_{11} = \sigma^2 \{1 + c' [\sigma^2 \Delta^{-1} (I_p + \beta \beta')^{-1} \Delta^{-1} + \Delta^{-1}] c\} (1 + \beta' \beta) \quad (6.51)$$

$g_{1i} = g_{i1} = \sigma^2 (1 + \beta' \beta) \times$ 向量 $(\sigma^2 \Delta^{-1} (I_p + \beta \beta')^{-1} \Delta^{-1} + \Delta^{-1})$ c 的 $(i-1)$ 元, $i=2, \dots, p+1$.

$$\begin{pmatrix} g_{22} & \cdots & g_{2,p+1} \\ \cdots & \cdots & \cdots \\ g_{p+1,2} & \cdots & g_{p+1,p+1} \end{pmatrix} = \sigma^2 \Delta^{-1} (I_p + \beta \beta')^{-1} \Delta^{-1} \quad (6.52)$$

其中 c 和 Δ 由条件 a, b 规定。特别, $\sqrt{n}(\hat{\alpha} - \alpha)$ 和 $\sqrt{n}(\hat{\beta} - \beta)$ 分别依分布收敛于 $N(0, g_{11})$ 和 $N_p(0, G_1)$, 其中 G_1 就是 (6.52) 式决定的方阵。

2°. $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ 依分布收敛于 $N(0, 2\sigma^4)$.

当用这些结果对 α, β, σ^2 进行大样本检验和区间(域)估计时, 极限分布中涉及的未知量可以其估计值代替之: β, σ^2 用其估计 $\hat{\beta}, \hat{\sigma}^2$ 代替, c 用 $\bar{\xi}$ 代替, 而 Δ 用

$$\begin{aligned} \tilde{\Delta} = & \frac{1}{n} (I_p + \hat{\beta} \hat{\beta}')^{-1} \{ (I_p : \hat{\beta}) W (I_p : \hat{\beta})' - \lambda_{p+1} \\ & (I_p + \hat{\beta} \hat{\beta}') \} (I_p + \hat{\beta} \hat{\beta}')^{-1} \end{aligned}$$

来代替。可以证明: 在前述 a, b 两条假定之下, $\bar{\xi}$ 和 $\tilde{\Delta}$ 分别是 c 和 Δ 的强相合估计。

例6.3 对 $p=1$ 的情况, 记

$$S^2_1 = \sum_{i=1}^n (\xi_i - \bar{\xi})^2, \quad S^2_2 = \sum_{i=1}^n (\eta_i - \bar{\eta})^2, \quad S_{12} = \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) \quad (6.53)$$

则 $W = \begin{pmatrix} S_{11}^2 & S_{12}^2 \\ S_{12}^2 & S_{22}^2 \end{pmatrix}$, 其最小特征根为

$$\lambda_1 = \{ S_{11}^2 + S_{22}^2 - [(S_{11}^2 + S_{22}^2)^2 - 4(S_{11}^2 S_{22}^2 - S_{12}^4)]^{1/2} \} / 2$$

据(6.49)式下面一段开头所述, 此式应等于(6.36)右边, 其中 $\hat{\beta}$ 由(5.37)确定。我们把这个事实的验证留给读者去完成。这虽只用到初等代数, 但并不很容易, 需要某些技巧。

其次, 用(6.49)式解出的 $\hat{\beta}$, 就与(6.37)一致。我们把这个(容易)验证的事实留给读者。

(五)附录

这里我们给出第(三)段中的 α 、 β 、 σ^2 的估计 $\hat{\alpha}$ 、 $\hat{\beta}$ 、 $\tilde{\sigma}^2$ 的强相合性的证明。先给出几条预备事实。

1. 设 U_1, U_2, \dots 为一串独立同分布随机变量, $EU_1 = a$ 存在有限, 则当 $n \rightarrow \infty$ 时以概率 1 有

$$\bar{U} = \frac{1}{n} \sum_{i=1}^n U_i \rightarrow a \quad (6.54)$$

这就是著名的Kolmogorov强大数律。类似地, 若 EU_1^2 也有有限, 则当 $n \rightarrow \infty$ 时以概率 1 有

$$\frac{1}{n} \sum_{i=1}^n (U_i - \bar{U})^2 \rightarrow \text{Var}(U_1) \quad (6.55)$$

2. 设 $(U_1, V_1), (U_2, V_2), \dots$ 为一串独立同分布的二维随机向量, $EU_1^2 < \infty, EV_1^2 < \infty$. 则当 $n \rightarrow \infty$ 时以概率 1 有

$$\frac{1}{n} \sum_{i=1}^n (U_i - \bar{U})(V_i - \bar{V}) \rightarrow \text{Cov}(U_1, V_1) \quad (6.56)$$

此因 $\sum_{i=1}^n (U_i - \bar{U})(V_i - \bar{V}) = \sum_{i=1}^n U_i V_i - n \bar{U} \bar{V}$. 由 Kolmogorov 强大数律, 有 $\bar{U} \rightarrow EU_1, \bar{V} \rightarrow EV_1, \sum_{i=1}^n U_i V_i / n \rightarrow EU_1 V_1$ (都是以概

率 1), 由此立得(6.56)。

3. 设 U_1, U_2, \dots 为一串独立同分布随机变量, $U_1 \sim N(0, \sigma^2)$. 又对任何自然数 n , 给定 n 个常数 a_{n1}, \dots, a_{nn} , 满足条件

$$A_n^2 = \sum_{i=1}^n a_{ni}^2 \leq C/n, \quad C \text{ 与 } n \text{ 无关} \quad (6.57)$$

则当 $n \rightarrow \infty$ 时, 以概率 1 有

$$T_n = \sum_{i=1}^n a_{ni} U_i \rightarrow 0 \quad (6.58)$$

证 记 $A_n^2 \sigma^2 = B_n^2$, 则 $T_n \sim N(0, B_n^2)$, 由(6.57)知 $B_n^2 \leq C'/n$, $C' = C\sigma^2$. 有

$$\begin{aligned} P(|T_n| \geq \varepsilon) &= 2 \frac{1}{\sqrt{2\pi}} \int_{\varepsilon/B_n}^{\infty} e^{-\frac{t^2}{2}} dt \\ &\leq \frac{2}{\sqrt{2\pi}} \cdot \frac{B_n}{\varepsilon} \int_{\varepsilon/B_n}^{\infty} t e^{-\frac{t^2}{2}} dt \\ &< B_n \exp\left(-\frac{\varepsilon^2}{2B_n^2}\right) / \varepsilon < \exp(-dn) \end{aligned}$$

对某个 $d > 0$ (d 与 n 无关), 当 n 充分大, 因此, $\sum_{i=1}^n P(|T_n| \geq \varepsilon) < \infty$

对任何 $\varepsilon > 0$, 从而推得(6.58)(见[16], P.228)

注 这个定理的条件可放宽: 只须假定 U_1, U_2, \dots 独立同分布, $EU_1 = 0$, $EU_1^2 < \infty$ 即可. 这个推广是 Y.S.Chow 在1966年证明的。

有了这些准备, 现在我们可以证明下面的结果: 设 X_1, X_2, \dots 满足条件 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 有界, 以及

$$0 < c_1 \leq \sum_{i=1}^n (X_i - \bar{X})^2 / n \leq C_2 < \infty, \quad \text{当 } n \text{ 充分大} \quad (6.59)$$

其中 C_1, C_2 与 n 无关, 则由 (6.37)、(6.38) 和 (6.42) 定义的 $\hat{\beta}, \hat{\alpha}$ 和 $\hat{\sigma}^2$ 分别是模型 (6.21) — (6.22) 中的参数 β, α 和 σ^2 的强相合估计。

证 按 (6.53) 定义 S_1^2, S_2^2 和 S_{12} . 以下我们将统一地用 $o(1)$ 这个记号去记一个以概率 1 收敛于 0 的量 (当 $n \rightarrow \infty$ 时), 由 (6.21) 有

$$S_1^2/n = \sum_{i=1}^n (X_i - \bar{X})^2/n + \sum_{i=1}^n (e_i - \bar{e})^2/n + \sum_{i=1}^n \frac{2(X_i - \bar{X})(e_i - \bar{e})}{n} \quad (6.60)$$

据 (6.55), 有 $\sum_{i=1}^n (e_i - \bar{e})^2/n \rightarrow \sigma^2$ (以下, 记号 “ \rightarrow ” 都是指当 $n \rightarrow \infty$ 时以概率 1 成立)。又由条件 (6.59) 知 $\sum_{i=1}^n \left(\frac{2(X_i - \bar{X})(e_i - \bar{e})}{n} \right)^2 \leq$

$4C_2/n$. 故由预备事实 3, 知 (6.60) 右边第三项为 $o(1)$. 故有

$$S_1^2/n = \sum_{i=1}^n (X_i - \bar{X})^2/n + \sigma^2 + o(1) \quad (6.61)$$

类似地有

$$S_2^2/n = \beta^2 \sum_{i=1}^n (X_i - \bar{X})^2/n + \sigma^2 + o(1) \quad (6.62)$$

又

$$\begin{aligned} S_{12}/n &= \beta \sum_{i=1}^n (X_i - \bar{X})^2/n + \beta \sum_{i=1}^n (X_i - \bar{X})(e_i - \bar{e})/n \\ &\quad + \sum_{i=1}^n (X_i - \bar{X})(e_i - \bar{e})/n + \sum_{i=1}^n (e_i - \bar{e})(e_i - \bar{e})/n \end{aligned}$$

据预备事实 3, 知上式右边中间两项为 $o(1)$ 。又由 (6.56), 知 $\sum_{i=1}^n (e_i - \bar{e})(e_i - \bar{e})/n \rightarrow \text{Cov}(e_1, e_1) = 0$. 于是得

$$S_{12}/n = \beta \sum_{i=1}^n (X_i - \bar{X})^2/n + o(1) \quad (6.63)$$

把 $\hat{\beta}$ 的表达式的分子分母同除以 n , 利用 (6.61) — (6.63), 并记

$D_n = \sum_{i=1}^n (X_i - \bar{X})^2/n$, 即知分子、分母分别有形式 $2\beta^2 D_n + o(1)$

和 $2\beta D_n$. 因为据 (6.59), 知 $D_n \geq c_1 > 0$ 当 n 充分大, 故当 $\beta \neq 0$ 时有

$$(2\beta^2 D_n + o(1)) / 2\beta D_n \rightarrow \beta$$

因此在 $\beta \neq 0$ 时证明了相合性.

若 $\beta = 0$, 则仍由 $D_n \geq c_1 > 0$ 当 n 充分大, 知以概率 1 有 $S_1^2/n > S_2^2/n$ 当 n 充分大, 因此以概率 1

$$\begin{aligned} |\hat{\beta} \text{ 的分子}| &= [(S_1^2 - S_2^2)^2 + 4S_{12}^2]^{1/2} - (S_1^2 - S_2^2) \\ &\leq 2S_{12}^2 / (S_1^2 - S_2^2) \end{aligned}$$

当 n 充分大, 故

$$|\hat{\beta}| \leq |S_{12}| / (S_1^2 - S_2^2)$$

以概率 1 当 n 充分大成立. 把上式分子分母除以 n , 用 (6.61) — (6.63) 以及 $D_n \geq c_1 > 0$, 即知 $\hat{\beta} \rightarrow 0$ 以概率 1 成立. 故相合性对 $\beta = 0$ 也有效.

为证 $\hat{\alpha}$ 的相合性, 利用 (6.38), 及已证的 $\hat{\beta} \rightarrow \beta$. 有

$$\begin{aligned} \hat{\alpha} &= \alpha + \beta \bar{X} + \bar{e} - \hat{\beta} (\bar{X} + \bar{e}) \\ &= \alpha + (\beta - \hat{\beta}) \bar{X} + \bar{e} - \bar{e} \end{aligned}$$

因为当 $n \rightarrow \infty$ 时 \bar{X} 保持有界, 而由预备事实 1 有 $\bar{e} \rightarrow 0$, $\bar{e} \rightarrow 0$. 又已证得 $\hat{\beta} \rightarrow \beta$, 故立得 $\hat{\alpha} \rightarrow \alpha$.

最后证明 $\tilde{\sigma}^2$ 的强相合性. 利用 (6.42), 以及 (6.61) — (6.63), 有

$$\begin{aligned} \frac{1}{n-2} (S_2^2 + S_1^2 \hat{\beta}^2 - 2S_{12} \hat{\beta}) &= (\beta^2 D_n + \sigma^2) + \beta^2 (D_n \\ &\quad + \sigma^2) - 2\beta^2 D_n + o(1) \\ 1 + \hat{\beta}^2 &= 1 + \beta^2 + o(1) \end{aligned}$$

故

$$\tilde{\sigma}^2 = \{(1 + \beta^2)\sigma^2 + o(1)\} / (1 + \beta^2 + o(1)) \rightarrow \sigma^2$$

从以上证明中我们注意两点：一是“ X 有界”这条件只用于证明 $\hat{\alpha}$ 的相合性*¹⁾（我们留给读者设法证明：由 $\{\sum_{i=1}^n (X_i - X)^2/n, n=1, 2, \dots\}$ 的有界性并不能推出 $\{\sum_{i=1}^n X_i/n, n=1, 2, \dots\}$ 的有界性）。二是若利用预备事实3的注，则可以把正态性条件去掉，只要求 $e_1, e_2, \dots, e_1, e_2, \dots$ 全体独立同分布， $Ee_1=0, Ee_1^2=\sigma^2$ 有限即可。

*¹⁾建议读者在直观上想一想，为何 \bar{X} 不有界会影响 $\hat{\alpha}$ 的相合性，而不影响 $\hat{\beta}$ 的相合性。

第二章 回归诊断

§2.1 引言

在第一章,我们讨论了线性回归模型

$$Y = \tilde{X}\gamma + e \quad (1.1)$$

的参数估计与检验问题。这里 Y 为 $n \times 1$ 观测向量, $\tilde{X} = (1_n : X)$ 为 $n \times (p+1)$ 设计阵, $1_n = (1, \dots, 1)$, 即 n 个元素皆为1的 n 维向量, 未知参数向量为 $\gamma' = (\alpha, \beta')$. 对于随机误差 e , 一个特别重要的情形是 $E(e) = 0$, 且满足GM假定:

$$\text{COV}(e_i, e_j) = \begin{cases} \sigma^2, & i=j \\ 0, & i \neq j \end{cases}$$

即 e 满足条件

$$E(e) = 0, \text{COV}(e) = \sigma^2 I_n \quad (1.2)$$

在进一步研究假设检验和区间估计时,还假设了 e_i 服从正态分布, 即 $e_i \sim N(0, \sigma^2)$, 在目前的线性回归的应用中, 绝大多数都采用了这些假设。这里有一个重要问题, 就是在一个具体场合, 当手头有了数据之后, 如何考察这些假设是否合理。如果实际数据与这些假设偏离比较大, 那么上一章我们所得到的的一些结果如GM定理、假设检验以及区间估计就不再成立。如果经过分析, 已经确认对所研究的具体问题, 上面的假设不成立, 那么我们又希望探讨对数据作怎样的修正后, 能够使它们满足或近似满足这些假设。这些就是我们在回归诊断中所要解决的第一个问题。这个问

题的重要性不言自明。

回归诊断所要研究的另一个重要问题,是探查对统计推断(如估计或预测等)有较大影响的试验点 $(x_{1i}, \dots, x_{pi}, Y_i)$, 这样的点称为强影响点(Influence Case).在回归分析中,因变量 Y 的取值 Y_i 具有随机性,自变量的值 $x'_i = (x_{1i}, \dots, x_{pi})$, $i=1, \dots, n$ 虽然是给定的,但毕竟是很多可能取到的值中的 n 组。我们希望每组数据 (x'_i, Y_i) 对未知参数的估计或其它推断有一定影响,但影响不要过大,这样我们所得到的估计关于数据就具有一定的稳定性。不然的话,如果个别一、两组数据是强影响点,在去掉它们之后,我们所得到的估计或经验回归模型与原来相比变化很大,那么我们对原来所建立的经验回归模型就会产生“不信任感”,怀疑它是否真正刻画了因变量与自变量的相互关系。当然,这时我们还需要对强影响点做进一步的仔细分析,还不能一概认为含强影响点的回归分析结果是不可取的。如果对获得数据的全过程作了检查之后,认为强影响点产生于试验或记录中的失误,那么这种数据应该剔除掉。不然的话,应考虑收集更多的数据,或采用稳健估计以缩小强影响点对估计或其它推断的影响,以求得到更稳定的估计和经验回归模型。关于已确认的强影响点的处理问题,难以给出统一的法则,这需要根据不同的具体情况斟酌对待。后面我们主要是讨论强影响点的探查方法。

现举一例说明回归诊断的必要性。

例1.1 (Anscombe, 1973)。表1.1给出了两个变量 Y 与 X 的四组数据。每组数据有11对观测值 (Y_i, x_i) , 其中第2例是前3组数据自变量 X 的值,第3、4、5列分别是第1、2、3组数据因变量 Y 的值。最后两列为第4组数据。对这四组数据分别作一元线性回归分析,常数项 α 和回归系数 β 的LS估计都分别为 $\hat{\alpha} = 3.0$, $\hat{\beta} = 0.5$ 。即它们有相同的经验回归方程 $Y = \hat{\alpha} + \hat{\beta}X = 3.0 + 0.5X$ 。误差方差 σ^2 的估计都为 $\hat{\sigma}^2 = 13.75$,复相关系数的平方 $R =$

0.667。因为这些回归统计量都有相同的值，人们很自然得出结论：一元线性回归模型 $Y = \alpha + \beta X + e$ 对这四组数据适合程度是一样的。但是，我们说这个结论是不对的。事实上，只要稍微考察一下这些数据在直角坐标系中的散点图(图2·1·1)就可以醒悟到这一点。

图2·1·1(a)表明，对第一组数据来讲，一元线性回归确实是适合的。而从图2·1·1(b)容易看出，同样的模型对第二组数据是不妥当的。很可能要考虑某种光滑曲线，或许是二次三项式模型。图2·1·1(c)显示了，一元线性回归对多数数据是适合的，唯独第三个观测点(13.00, 12.74)远离回归直线。以后我们把这种点称为异常点(Outlier)。如果把异常点暂时剔除掉，对剩下的数据重新配回归直线，这时经验回归方程为 $Y = 4 + 0.346X$ ，与原来的经验回归方程 $Y = 3.0 + 0.5X$ 有相当的不同，所以这个异常点很可能是强影响点。对最后一组数据，情况与前三组都不相同，从图2·1·1(d)可以看出，我们不能认为 Y 与 X 之间存在某种线性依赖关系。这里，第八组数据对应的

表11 Anscombe数据

数据号	数 据 组 号					
	1—3	1	2	3	4	4
	X	Y	Y	Y	X	Y
1	0.0	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	6.44	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

点(19.0, 12.50)远离其它点。如果把它剔除掉, R 就减少很多, 以致于从 R 就可断言 Y 与 X 没有什么相关关系。于是第八组数据可称为强影响点。对这种情况, 我们说数据不够好, 应该对自变量 X 在 $[8, 19]$ 这个区间上再收集一些数据。

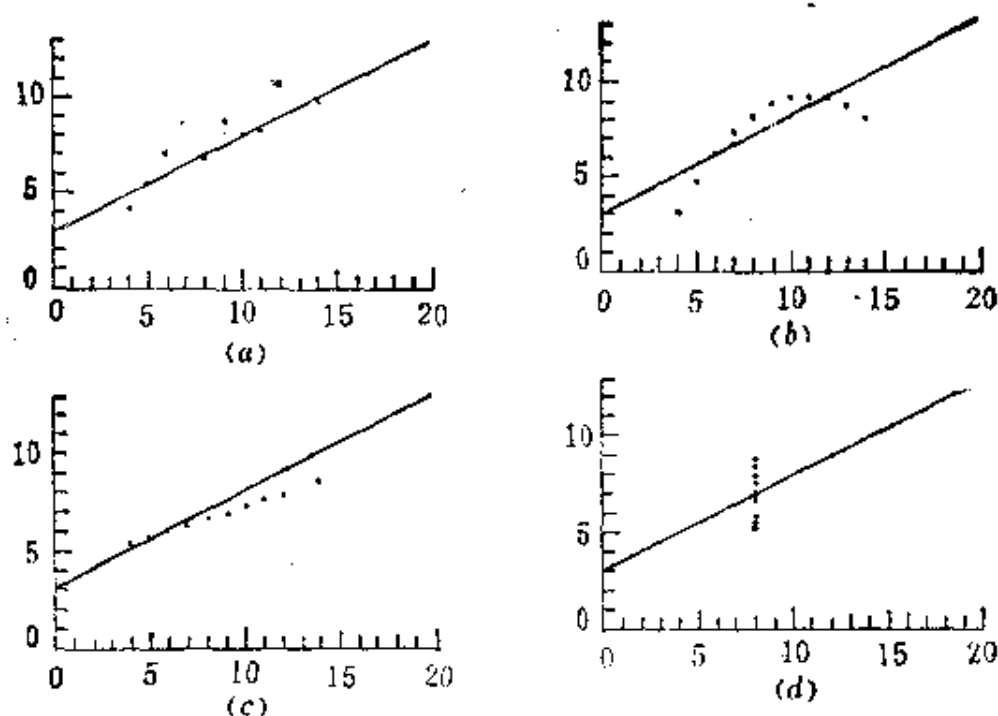


图2.1.1 Anscombe数据的回归

本章我们主要讨论上面提出的两个问题。以下诸节先引进一些回归诊断所用的统计量, 并研究它们在基本假设成立时的一些性质, 然后考察某些假设条件不满足会对这些统计量产生怎样的影响, 用这些信息来判断在一具体场合基本假设是否成立。另外, 我们还引进一些判断一组数据影响大小的统计量。

回归诊断是一个比较复杂的问题。它有点象医生给病人治病。有时一个症状往往是多种不同疾病的征兆, 必须仔细地从多方面作检查分析, 才能断言毛病出现在什么地方。在这个方向上研究的时间还不长, 许多统计量和实施方法只是从直观角度提出的。系统理论工作见到的还不多。这一章里我们要介绍的大多是已见

诸实用的一些内容。

现在约定一些记号。我们用 $\tilde{X}(i)$, $Y(i)$, $e(i)$ 表示从矩阵 \tilde{X} , Y 和 e 中剔除第 i 行所得到的矩阵。 $\hat{\gamma}$ 表示从完全数据算出的 γ 的LS估计, 而 $\hat{\gamma}(i)$ 表示剔除第 i 组数据后, 从剩下的 $n-1$ 组数据算出的 γ 的LS估计, 即 $\hat{\gamma}(i)$ 为从线性回归模型

$$Y(i) = \tilde{X}(i)\gamma + e(i) \quad (1.3)$$

求到的 γ 的LS估计。

§2.2 残差

在§1.2我们已经定义

$$\delta_i = Y_i - \tilde{x}_i' \hat{\gamma}, \quad i=1, \dots, n \quad (2.1)$$

为第 i 次观测的残差, 或称 Y_i 的残差。它是因变量 Y 的实际观测值 Y_i 与由经验回归方程得到的拟合值 $\hat{Y}_i = \tilde{x}_i' \hat{\gamma}$ 之差。如果模型(1.1)正确的话, 我们可以把 δ_i 看作误差 e_i 的观测值, 它应该具有 e_i 的一些性状。因此, 我们可以通过这些 δ_i 以及基于它们的一些统计量来考察模型假设的合理性。

(一) 普通残差

为了与后面定义的其它残差相区别, 以后我们把(2.1)所定义的残差称为**普通残差**。若记 $\hat{Y}' = (\hat{Y}_1, \dots, \hat{Y}_n)$, 则

$$\hat{Y} = \tilde{X}' \hat{\gamma} = \tilde{X} (\tilde{X}' \tilde{X})^{-1} \tilde{X}' Y \triangleq HY, \quad (2.2)$$

这里矩阵

$$H = \tilde{X} (\tilde{X}' \tilde{X})^{-1} \tilde{X}'$$

作用在向量 Y 的结果是给 Y 戴上了帽子“ $\hat{}$ ”, 故通常称 H 为**帽子矩阵**, 容易验证

$$H' = H, \quad H^2 = H. \quad (2.3)$$

即帽子矩阵是一个对称幂等阵。利用 H , 残差向量 $\delta' = (\delta_1, \dots, \delta_n)$ 可表为

$$\delta = Y - \hat{Y} = (I - H)Y = (I - H)e. \quad (2.4)$$

此处应用了 $(I - H)\tilde{X} = 0$.

若随机变量 Z_1, \dots, Z_n 的联合分布为正态, 且 $Z' = (Z_1, \dots, Z_n)$ 的均值 $E(Z) = \mu$, $\text{COV}(Z) = V$, 则称 Z 为均值为 μ , 协方差阵为 V 的 n 维正态向量, 记为 $Z \sim N_n(\mu, V)$, 有时也记为 $Z \sim N(\mu, V)$. 对于模型(1.1)的随机误差向量 e , 若GM假设成立, 且 e_1, \dots, e_n 联合分布为正态, 则 $e \sim N(0, \sigma^2 I)$.

关于普通残差向量 δ , 我们有如下定理.

定理2.1 在假设(1.2)下

- (1) $E(\delta) = 0$, $\text{COV}(\delta) = \sigma^2(I - H)$,
- (2) $\text{COV}(\hat{Y}, \delta) = 0$,
- (3) 若 $e \sim N(0, \sigma^2 I)$, 则 $\delta \sim N(0, \sigma^2(I - H))$.

证 (1) $E(\delta) = 0$ 是显然的. 第二条的证明只需利用公式(1.2.22), 即

$$\begin{aligned} \text{COV}(\delta) &= \text{COV}((I - H)e) \\ &= (I - H)\text{COV}(e)(I - H). \end{aligned}$$

从(1.2)及 $(I - H)^2 = I - H$, 立得

$$\text{COV}(\delta) = \sigma^2(I - H)^2 = \sigma^2(I - H).$$

- (2) 利用公式: 若 A, B 为两个常数矩阵, 则

$$\text{COV}(AY, BY) = A\text{COV}(Y)B'; \quad (2.5)$$

由于 $\hat{Y} = HY$, $\delta = (I - H)Y$, 所以

$$\begin{aligned} \text{COV}(\hat{Y}, \delta) &= H\text{COV}(Y)(I - H)' \\ &= \sigma^2 H(I - H)' = 0 \end{aligned}$$

这里利用了(2.3).

- (3) 利用多元正态分布的性质: 若 $U \sim N(\mu, V)$, 则

$$W = AU \sim N(A\mu, AVA') \quad (2.6)$$

(这个事实的证明见[8]定理2.1.6)由 $\delta = (I - H)e$ 及(2.3)立得所要结论。定理证毕。

我们已经看到, 普通残差和帽子矩阵 H 有密切关系, 后者以后我们常常要用到它。为此我们要进一步讨论 $H = (h_{ij})$ 的元素的一些性质。

定理2.2 (1) $0 \leq h_{ii} \leq 1$, 且 $h_{ii} = 1$ 时, $h_{ij} = 0, j \neq i$.

$$(2) \sum_{i=1}^n h_{ii} = p + 1,$$

$$(3) h_{ii} = \frac{1}{n} + (x_i - \bar{x})'(X^*X^*)^{-1}(x_i - \bar{x}).$$

这里

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}, \quad X^* = \begin{bmatrix} (x_1 - \bar{x})' \\ (x_2 - \bar{x})' \\ \vdots \\ (x_n - \bar{x})' \end{bmatrix}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

证 (1) 因为 $H^2 = H, H' = H$, 所以

$$h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2,$$

故有 $h_{ii} \geq h_{ii}^2$. 由此可推得 $0 \leq h_{ii} \leq 1$. 再由上式立得剩下的结论。

(2) $\sum_{i=1}^n h_{ii} = \text{tr}(H) = \text{tr}(\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')$, 利用迹的性质:

$\text{tr}(AB) = \text{tr}(BA)$, 得

$$\sum_{i=1}^n h_{ii} = \text{tr}((\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{X}) = \text{tr}(I_{p+1}) = p + 1$$

(3) 依定义

$$h_{ii} = (1, x_i')(\tilde{X}'\tilde{X})^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= (1 \ x_i') \begin{pmatrix} n & 1'X \\ X'1 & X'X \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

应用分块矩阵的逆矩阵公式(证明见〔8〕的定理1·2·4)

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{pmatrix}, \quad (2.7)$$

其中 $B = A_{22} - A_{21}A_{11}^{-1}A_{12}$, 得到

$$h_{ii} = (1 \ x_i') \left\{ \begin{array}{cc} \frac{1}{n} + \frac{1}{n} 1'X(X^{*'}X^*)^{-1} \frac{1}{n} X'1 & -\frac{1}{n} 1'X(X^{*'}X^*)^{-1} \\ - (X^{*'}X^*)^{-1} \frac{1}{n} X'1 & (X^{*'}X^*)^{-1} \end{array} \right\} \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= \frac{1}{n} + (x_i - \bar{x})'(X^{*'}X^*)^{-1} (x_i - \bar{x}).$$

定理证毕.

从最后的结果我们可以看到 h_{ii} 的几何意义。 h_{ii} 表达式的第二项表示在自变量空间中, 第 i 个试验点 x_i 到试验中心 \bar{x} 的距离。不过这个距离不是通常的欧氏距离, 而是 Mahalanobis 距离, 简称为 **马氏距离**。因此, h_{ii} 刻画了第 i 个试验点 x_i 距离试验中心 \bar{x} 的远近。

例2·1 对一元线性回归

$$Y_i = a + \beta x_i + e_i, \quad i = 1, \dots, n$$

有

$$\tilde{X}' = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} = \begin{pmatrix} 1' \\ X' \end{pmatrix},$$

$$X^{*'} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}),$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

容易验证

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \quad (2.8)$$

h_{ii} 的最小值 $1/n$ 当 $x_i = \bar{x}$ 时达到, 且随着 x_i 远离中心点 \bar{x} , h_{ii} 增大. 当 x_i 离 \bar{x} 充分远时, h_{ii} 能够充分接近于 1.

因为 $\text{Var}(\delta_i) = \sigma^2(1 - h_{ii})$, 所以 h_{ii} 愈大, $\text{Var}(\delta_i)$ 就愈小. 特别当 $h_{ii} = 1$ 时, 从 $\text{Var}(\delta_i) = 0$ 和 $E(\delta_i) = 0$ 可知 $\delta_i = 0$ (此事实以概率为 1 成立, 如果将概率为 0 的集合不予考虑, 就认为 $\delta_i = 0$), 即 $Y_i = \hat{Y}_i$, 这说明, 当 h_{ii} 很大, 即当试验点 x_i 距离试验中心很远时, 残差 $\delta_i \approx 0$, 即不论观测值 Y_i 等于什么值, 总有拟合值 $\hat{Y}_i \approx Y_i$. 从几何上看, 这个结论就是, 在自变量空间 R^p 中 x_i 远离试验中心 \bar{x} , 则在空间 R^{p+1} 中, 点 (x_i, Y_i) 就把回归直线拉向它自己. 这种点对回归系数的 LS 估计的影响可能很大, 在回归诊断中应予特别注意. 通常称这种点为 **高杠杆点** (High Leverage Case) 因为 $0 \leq h_{ii} \leq 1$, 所以, 所谓 h_{ii} 很大, 只不过是它很靠近 1. 那么究竟 h_{ii} 大到多少, 所对应的点才是高杠杆点呢? 一般说来, 很难给出一个处处适用的标准. 有一种做法是, 将 x_1, x_2, \dots, x_n 看作服从正态分布的随机向量 $X' = (X_1, \dots, X_p)$ 的简单随机样本, 则可以证明 [18]

$$\mathcal{J} = \frac{n-p-1}{p} \frac{h_{ii} - \frac{1}{n}}{1 - h_{ii}} \sim F_{p, n-p-1}. \quad (2.9)$$

注意到 \mathcal{J} 是 h_{ii} 的单调增函数, 所以 h_{ii} 很大等价于 \mathcal{J} 很大. 如果从某个 h_{ii} 算出的 $\mathcal{J} > F_{p, n-p-1}(0.05)$ 或 $> F_{p, n-p-1}(0.10)$, 就认为 h_{ii} 很大, 对应的点 (x_i, Y_i) 就判为高杠杆点.

(二) 学生化残差

残差的重要应用之一是根据它的绝对值大小判定异常点. 但是, 对于普通残差 δ_i , $\text{Var}(\delta_i) = \sigma^2(1 - h_{ii})$, 这个方差与因变量 Y 的度量单位以及 h_{ii} 有关. 因此在判定异常点的情形, 直接比较

普通残差 δ_i 是不适宜的。为此将它们标准化, 得到

$$\frac{\delta_i}{\sigma\sqrt{1-h_{ii}}}, i=1, \dots, n \quad (2.10)$$

但其中 σ 未知, 用其估计 $\hat{\sigma} = [(\|Y - \tilde{X}\hat{y}\|^2/(n-p-1))]^{1/2}$ 代替, 获得

$$r_i = \frac{\delta_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, i=1, \dots, n \quad (2.11)$$

称为学生化残差。这里要注意的是, 在误差 e 的正态假设 $e \sim N(0, \sigma^2 I)$ 条件下, 虽然 $\delta_i \sim N(0, \sigma^2(1-h_{ii}))$, $\hat{\sigma}^2 \sim X_{n-p-1}^2$, 但二者并不独立, 所以 r_i 并不服从通常的 t_{n-p-1} 分布。诸 r_i 彼此也不独立。关于 r_i 的分布我们不加证明的给出下面的结果, 证明可以在(20)中找到。

记 $r' = (r_1, \dots, r_n)$, $I = \{i_1, \dots, i_m\}$

$r'_I = (r_{i_1}, \dots, r_{i_m})$.

H_I 为 H 的第 i_1, \dots, i_m 行和第 i_1, \dots, i_m 列交叉处元素构成的方阵,

$D = \text{diag}(1-h_{i_1 i_1}, \dots, 1-h_{i_m i_m})$.

$C_I = D^{-1/2}(I - H_I)D^{-1/2}$

定理2.3 若 $e \sim N(0, \sigma^2 I)$, 则 r_I 的分布密度为

$$f(r_I) = \begin{cases} \frac{\left(\frac{n-p-1}{2}\right) |I - H_I|^{-1/2}}{\Gamma\left(\frac{n-m-p-1}{2}\right) ((n-p-1)\pi)^{m/2}} \\ \prod_{i=1}^m (1-h_{i_i i_i})^{1/2} \left[-\frac{r'_I C_I^{-1} r_I}{n-p-1} \right]^{\frac{n-m-p-1}{2}}, \\ \quad \text{当 } r'_I C_I^{-1} r_I \leq n-p-1 \text{ 时,} \\ 0, \quad \text{其它} \end{cases}$$

特别 $m=1$ 时, 得到任一个 r_i 的分布密度为

$$f(r_i) = \begin{cases} \frac{\Gamma\left(\frac{n-p-1}{2}\right)}{\Gamma\left(\frac{n-p}{2}-1\right) \Gamma\left(\frac{1}{2}\right) (n-1)^{1/2}} \left(1 - \frac{r_i^2}{n-p-1}\right)^{\frac{n-p}{2}-1}, & \text{当 } |r_i| \leq (n-p-1)^{1/2} \text{ 时} \\ 0, & \text{其它} \end{cases}$$

根据这个定理, 能够得到如下推论

推论2.1 (1) $\frac{r_i^2}{(n-p-1)}$ 服从参数为 $\frac{1}{2}$, $\frac{n-p-2}{2}$ 的Beta

分布, 记为

$$\frac{r_i^2}{n-p-1} \sim B\left(\frac{1}{2}, \frac{n-p-2}{2}\right)$$

(2) $E(r_i) = 0$, $\text{Var}(r_i) = 1$,

$$\text{COV}(r_i, r_j) = -\frac{h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}, \quad i \neq j$$

因为 $\text{COV}(r_i, r_j)$ 一般都很小, 所以在应用上常常近似地认为 r_i 与 r_j 不相关。并进一步用正态分布作为 r_i 的近似分布。于是 r_i 近似服从 $N(0, 1)$ 且相互独立。下一节要讲的残差图主要依据这个事实进行模型假设合理性诊断。

另外一种学生化残差定义为

$$r_i^* = \frac{\hat{o}_i}{\hat{\sigma}(i) \sqrt{1-h_{ii}}}, \quad i=1, \dots, n \quad (2.12)$$

这里 $\hat{\sigma}(i)$ 是从模型(1.3)得到的 σ 的估计, 即

$$\hat{\sigma}^2(i) = Y(i)[I - \tilde{X}(i)(\tilde{X}(i)' \tilde{X}(i))^{-1} \tilde{X}(i)']$$

$$\hat{Y}(i)/(n-p-2) \quad (2.13)$$

(2.12)是在(2.11)中用 $\hat{\sigma}(i)$ 代替 $\hat{\sigma}$ 得到的,其用意是,在对普通残差 δ_i 作“标准化”时,排除 δ_i 自身在误差方差估计中的份额,一些学者研究了 r_i^* 的分布,指出在许多应用场合, r_i^* 近似服从 t_{n-p-2} .

在计算 r_i^* 时,我们并不必对每个 i 应用(2.13)计算 $\sigma(i)^2$.下式给出了计算 $\hat{\sigma}(i)$ 的一个简便方法.

$$\hat{\sigma}(i)^2 = \frac{n-p-1-r_i^2}{n-p-2} \hat{\sigma}^2 \quad (2.14)$$

现在我们证明(2.14),对 $A = \tilde{X}'\tilde{X}$, $u=v=\tilde{x}_i$ 应用公式

$$(A-uv')^{-1} = A^{-1} + \frac{A^{-1}u \cdot v' A^{-1}}{1-v' A^{-1}u}$$

得到

$$\begin{aligned} (\tilde{X}(i)' \tilde{X}(i))^{-1} &= (\tilde{X}' \tilde{X})^{-1} \\ &\quad + \frac{(\tilde{X}' \tilde{X})^{-1} \tilde{x}_i \tilde{x}_i' (\tilde{X}' \tilde{X})^{-1}}{1-h_{ii}} \end{aligned} \quad (2.15)$$

将两边右乘 $\tilde{X}'Y$,利用

$$\begin{aligned} \tilde{X}'Y &= \tilde{X}'(i)Y(i) + Y_i \tilde{x}_i, \\ \hat{\gamma}(i) &= (\tilde{X}'(i)\tilde{X}(i))^{-1} \tilde{X}'(i)Y(i), \end{aligned}$$

我们有

$$\begin{aligned} \hat{\gamma} - \hat{\gamma}(i) &= (\tilde{X}'(i)\tilde{X}(i))^{-1} \tilde{x}_i Y_i - (\tilde{X}'\tilde{X})^{-1} \tilde{Y}_i \tilde{x}_i \\ &\quad / (1-h_{ii}) \end{aligned} \quad (2.16)$$

再用 \tilde{x}_i 右乘(2.15),得到

$$(\tilde{X}'(i)\tilde{X}(i))^{-1} \tilde{x}_i = (\tilde{X}'\tilde{X})^{-1} \tilde{x}_i / (1-h_{ii})$$

代入(2.16)有

$$\hat{\gamma} - \hat{\gamma}(i) = (\tilde{X}'\tilde{X})^{-1} \delta_i \tilde{x}_i / (1-h_{ii}) \quad (2.17)$$

利用(2·17)及 $H\hat{\sigma} = 0$, 我们有

$$\begin{aligned}
 (n-p-2)\hat{\sigma}^2(i) &= \sum_{j \neq i} (Y_j - \tilde{x}'_j \hat{\gamma}(i))^2 \\
 &= \sum_{j=1}^n [(Y_j - \tilde{x}'_j \hat{\gamma}) + (\tilde{x}'_j \hat{\gamma} - \tilde{x}'_j \hat{\gamma}(i))]^2 \\
 &\quad - (Y_i - \tilde{x}'_i \hat{\gamma}(i))^2 \\
 &= \sum_{j=1}^n [\delta_j + h_{ji}\delta_i / (1 - h_{ii})]^2 - \delta_i^2 / (1 - h_{ii})^2 \\
 &= (n-p-1)\hat{\sigma}^2 + \frac{2\delta_i}{1 - h_{ii}} \sum_{j=1}^n h_{ji}\delta_j \\
 &\quad + \frac{\delta_i^2}{(1 - h_{ii})^2} \sum_{j=1}^n h_{ji}^2 - \frac{\delta_i^2}{(1 - h_{ii})^2} \\
 &= (n-p-1)\hat{\sigma}^2 - \delta_i^2 / (1 - h_{ii}) \\
 &= (n-p-1-r_i^2)\hat{\sigma}^2
 \end{aligned}$$

这就证明了(2·14)。

(三) 预测残差

前面讨论的普通残差和学生化残差都是从数据与模型的拟合角度提出的。因为预测是回归模型的一种重要应用, 于是我们有必要从预测角度定义残差。

从模型(1·1)剔除第 i 组数据 (x'_i, Y_i) 后得到模型

$$Y(i) = \tilde{X}(i)\gamma + e(i)$$

从此模型算出的 γ 的LS估计为 $\hat{\gamma}(i) = (\tilde{X}'(i)\tilde{X}(i)^{-1}\tilde{X}'(i)Y(i))$ 。

利用这个估计, 可以对第 i 个试验点 x_i 处 Y 的值作预测, 预测值为 $\tilde{x}'_i \hat{\gamma}(i)$ 。我们称

$$e_i = Y_i - \tilde{x}'_i \hat{\gamma}(i), \quad i = 1, \dots, n \quad (2·18)$$

为 x_i 处的预测残差。

利用(2.17), 容易得到预测残差 ε_i 与普通残差 δ_i 的关系.

$$\begin{aligned}\varepsilon_i &= Y_i - \tilde{x}_i' \hat{\gamma} \\ &= Y_i - \tilde{x}_i' (\hat{\gamma} - (\tilde{X}' \tilde{X})^{-1} \delta_i \tilde{x}_i' / (1 - h_{ii})) \\ &= \delta_i / (1 - h_{ii})\end{aligned}\quad (2.19)$$

因为 $0 \leq h_{ii} \leq 1$, 所以对 h_{ii} 较大的数据, 对应的 ε_i 也就很大, 可见用预测残差作回归诊断更看重具有较大 h_{ii} 的试验数据, 也就是远离试验中心的那些数据。

记 $\varepsilon' = (\varepsilon_1, \dots, \varepsilon_n)$, $D = \text{diag}(1 - h_{11}, \dots, 1 - h_{p+1, p+1})$, 则预测残差向量 $\varepsilon = D^{-1}\delta$. 于是若 $\varepsilon \sim N(0, \sigma^2 I)$, 则从 $\delta \sim N(0, \sigma^2(I - H))$ 及多元正态分布的性质(2.6), 可以推得

$$\varepsilon \sim N(0, \sigma^2 D^{-1}(I - H)D^{-1}).$$

特别, $\varepsilon_i \sim N(0, \sigma^2/(1 - h_{ii}))$, $i = 1, \dots, n$. 由此可以看出, 将 ε_i 施“学生化”程序, 就得到了前面讨论过的学生化残差 r_i 和 r_i^* .

预测残差 ε_i 除了应用在回归诊断之外, 另一个重要应用是回归自变量选择。 ε_i 的平方和 $\sum_{i=1}^n \varepsilon_i^2$ 可以作为回归自变量选择的一种准则, 称为PRESS(Prediction Error Sum of Squares)准则。的回归模型应该有较小的PRESS. 关于这个准则的讨论将在下一个好章进行。

(四)不相关残差

前面讨论的几种残差的各分量彼此都是相关的, 现在我们引进普通残差的一种线性变换, 使得变换后的残差分量互不相关。特别若模型的误差向量 e 服从正态分布, 这种残差的分量也就相互独立。

从定理2.1我们知道, 对普通残差 δ , $\text{COV}(\delta) = \sigma^2(I - H)$, 但是, 这个协方差阵是不可逆的。事实上, 前面已经指出过, H

域内，且不呈现任何的趋势，如图2·3·1(a).这时数据与假设 $e \sim N(0, \sigma^2 I)$ 没有不一致的征兆，可以认为这个假设基本上是合理的。而图2·3·(b)-(d)显示了误差等方差即 $\text{Var}(e_i) = \sigma^2, i=1, \dots, n$ 的假设(或称方差齐性假设)并不满足。其中图(b)表示了误差方差随 \hat{Y}_i 增大而增大。而图(c)正好相反，它表示了误差方差随 \hat{Y}_i 增大而减小。图(d)则表示对较大和较小的 \hat{Y}_i ，误差方差比较小，而对中等大小的 \hat{Y}_i ，误差方差比较大。接下来的图(e)和(f)表明回归函数可能是非线性的，或误差 e_i 之间具有一定相关性或漏掉了一个甚至多个重要自变量。究竟属于何种情况，还需要作进一步的诊断。这种“一种症状可能产生于多种不同的疾病”正是回归诊断的困难所在。在处理回归诊断时，很象医生治病，方法和经验都是不可缺少的。

诊断出来“疾病”即知道某些假设条件不成立之后，进一步的问题就是“对症下药”。如果残差图显示了误差非齐性，则可采用两种“治疗”方案。一种是对自变量或因变量作变换，使得对变换过的数据，误差方差近似相等。这将在后面仔细讨论。另一种治疗方案是改用加权最小二乘法。关于加权最小二乘法。上一章(见§1·2)已经讨论过了。这种方法的困难往往在于权是未知的，需要设法给出估计。如果是图(e)和(f)的情形，应该仔细分析实际问题，试探各种治疗方案，如增加新自变量，或用本节后面讨论的Durbin-Watson检验，考察误差是否有正相关或负相关，等等。

例3·1 人工降雨试验(取自[20])

70年代美国曾多次进行人工降雨试验。其中有一次试验是在Florida州进行的，目的是考察碘化银对增加降雨量的作用。目标区域是Coral Gables北部和东部大约3千平方哩的范围。所考虑的因变量和自变量分别是

Y ——在每个试验天的6小时中，目标区域的总降雨量。

x_1 ——示性变量, $x_1 = 1$ 表示实施人工降雨, $x_1 = 0$, 表示不实施人工降雨。

x_2 ——表示从试验第一天(算作0)算起的试验天的日历年数。例如, 对第一天, $x_{21} = 0$, 而后隔了两天才进行试验, $x_{22} = 3$ 。这个变量的引入是为了反映自然降雨本身可能存在的趋势以及随着试验的进行, 试验方法上所作的修正。

x_3 ——确定试验天的一种气象指标, 当 $x_3 \geq 1.5$ 时, 认为是适于试验的天,

x_4 ——用雷达测定的目标区域云层复盖率,

x_5 ——在实施人工降雨措施前 1 小时, 目标区域总降雨量。

x_6 ——示性变量, 取 1 表示移动雷达的响应值, 取 2 表示非移动雷达的响应值。

原始数据如表 3·1, 因为从气象专业理论还不能对模型提出明确的形式, 所以我们考虑回归模型

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{15} x_1 x_5 + \beta_{16} x_1 x_6 + e \quad (3.1)$$

这里引进了交叉项 $x_1 x_3$, $x_1 x_4$, $x_1 x_5$, $x_1 x_6$, 以反映人工降雨措施与其它因素之间的交互作用。如在第一章曾指出的, 只要引进一些新“自变量”, 上述模型就化为一般线性回归模型。实施人工降雨和未实施人工降雨的降雨量之差

$$\begin{aligned} \Delta Y &= Y(x_1 = 1) - Y(x_1 = 0) \\ &= \beta_1 + \beta_{13} x_3 + \beta_{14} x_4 + \beta_{15} x_5 + \beta_{16} x_6, \end{aligned} \quad (3.2)$$

它主要与交互作用项的系数 β_{13} 、 β_{14} 、 β_{15} 、 β_{16} 有关。

假设 $\text{COV}(e) = \sigma^2 I$ 成立, 应用最小二乘法求得未知参数的 LS 估计及有关统计量, 见表 3·2。于是经验回归方程为

$$\begin{aligned} Y &= -3.4991 + 16.2452x_1 - 0.0450x_2 + 0.4198x_3 + 0.3879x_4 \\ &\quad + 4.1083x_5 + 3.1528x_6 - 3.1972x_1x_3 - 0.4863x_1x_4 \\ &\quad - 2.5571x_1x_5 - 0.5622x_1x_6. \end{aligned}$$

图 2·3·2 是以拟合值 \hat{Y} 为横坐标、学生化残差 r_i 为纵坐标的残差图。从

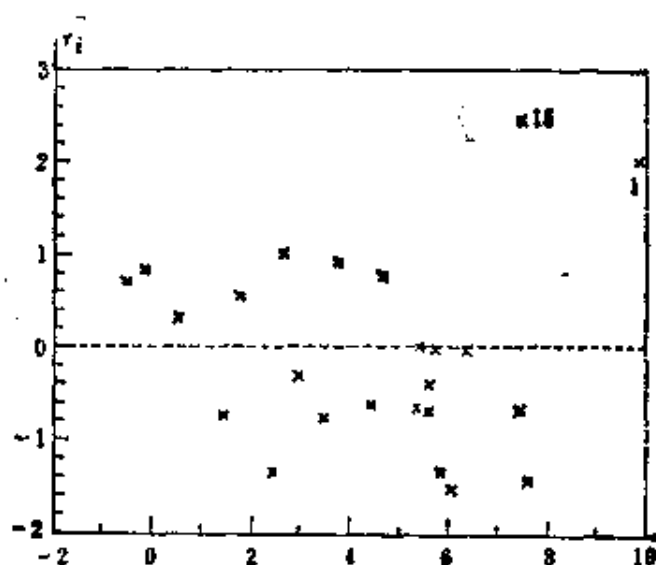
表3.1 人工降雨试验数据

数据号	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	0	0	1.75	13.40	0.274	2	12.85
2	1	1	2.70	37.90	1.267	1	5.52
3	1	3	4.10	3.90	0.198	2	6.29
4	0	4	2.35	5.30	0.526	1	6.11
5	1	6	4.25	7.10	0.250	1	2.45
6	0	9	1.60	6.90	0.018	2	3.61
7	0	18	1.30	4.60	0.307	1	0.47
8	0	25	3.35	4.90	0.194	1	4.56
9	0	27	2.85	12.10	0.751	1	6.35
10	1	28	2.20	5.20	0.084	1	5.06
11	1	29	4.40	4.10	0.236	1	2.76
12	1	32	3.10	2.80	0.214	1	4.05
13	0	33	3.95	6.80	0.796	1	5.74
14	1	35	2.90	3.00	0.124	1	4.84
15	1	38	2.05	7.00	0.144	1	11.86
16	0	39	4.00	11.30	0.398	1	4.45
17	0	53	3.35	4.20	0.237	2	3.66
18	1	55	3.70	3.30	0.960	1	4.22
19	0	56	3.80	2.20	0.230	1	1.16
20	1	59	3.40	6.50	0.142	2	5.45
21	1	65	3.15	3.10	0.073	1	2.02
22	0	68	3.15	2.60	0.136	1	0.82
23	1	82	4.01	8.30	0.123	1	1.09
24	0	83	4.65	7.40	0.168	1	0.28

表2·2 人工降雨试验的LS估计

参 数	LS 估计	参 数	LS 估计
α	-3.4991	β_9	3.1528
β_1	16.2452	β_{13}	-3.1972
β_2	-0.0450	β_{14}	-0.4863
β_3	0.4198	β_{15}	-2.5571
β_4	0.3879	β_{16}	-0.5622
β_5	4.1083		

这个图可以看出一些问题来。在残差图中，第1组和第15组数据对应的点 (\hat{Y}_1, r_1) 和 (\hat{Y}_{15}, r_{15}) 远离其余点，经计算，对这两组数据，降雨量的预测值是负的。另外，随着 \hat{Y}_i 的增加，学生化残差 r_i 有减少的趋势。对于这个情况，我们可能采取的“治疗”方案是：(1)对因变量作变换，消去回归函数可能有的非线性形式，或许能够纠正了在第1组和第15组数据的降雨量的负预测值；(2)对回归自变量作变换。

图2·3·2 人工降雨试验的学生化残差(横轴为 Y_i)

欲进一步考察有关假设的合理性，一种方法是作以某个自变量为横坐标的残差图。图2·3·3是以 x_2 为横坐标的残差图。因为大部分比较大的残差

对应的 x_3 的值比较小,因而误差方差很可能是 x_3 的减函数,即不是等方差的。综合图2·3·2和图2·3·3,说明对人工降雨试验数据,通常的GM假设是不合理的,因变量或自变量的变换,或引进另外一些新自变量是需要考虑的。

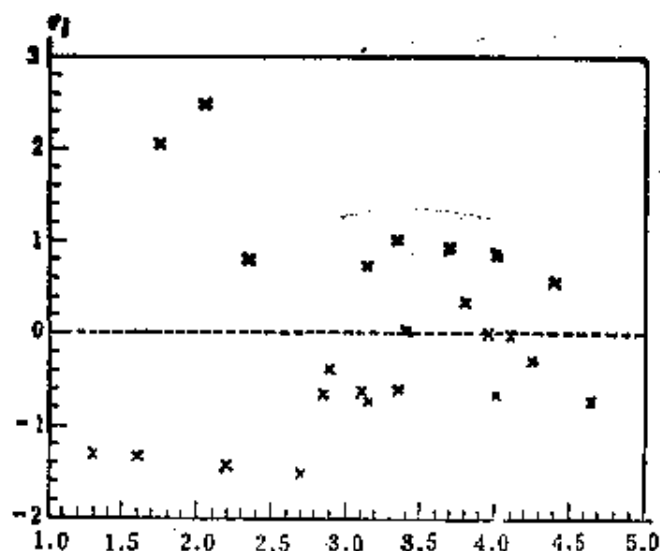


图2·3·3 人工降雨试验的学生化残差(横轴为 x_3)

(二)时间为横坐标的残差图与Durbin-Watson检验

如果因变量 Y 的观测值 Y_1, Y_2, \dots, Y_n 是依一定时间顺序观测得到的,则它们构成一个时间序列,在研究经济、商业以及一些工程问题时,常常会碰到这样的数据。对于时间序列数据,误差 e_i 往往是自相关的,例如可能有关系

$$e_i = \rho e_{i-1} + u_i \quad (3.3)$$

这里 $u_i \sim N(0, \sigma^2)$,且相互独立, $\rho (|\rho| < 1)$ 是自相关参数。(3.3)称为一阶自回归模型。如果 $\rho > 0$,称为正相关。不然, $\rho < 0$ 称为负相关。当误差 e_i 是正相关时,残差符号具有“集团”性,即有一段全是正号,另一段全是负号,然后又有一段全是正号。这时残差符号的改变很不频繁,如图2·3·4(a)。相反,如果误差 e_i 是负

相关，则残差符号大致有正负相间的趋势。符号的改变过于频繁，如图2·3·4(b)。

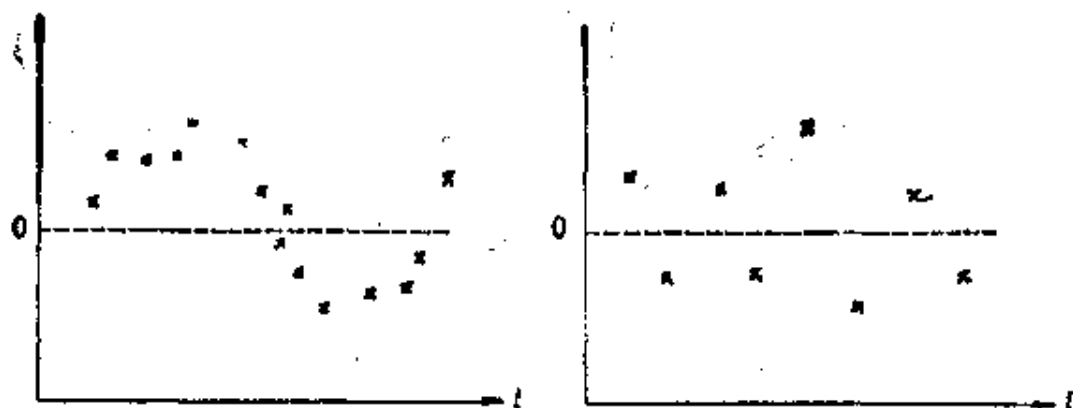


图2·3·4 误差相关情形的残差

对残差图上观察到的相关性还需要作进一步的检验。最常用的一种检验是Durbin-Watson的d-检验。这种检验是针对误差一阶自回归模型(3·3)提出的。将

$$e_{i-j} = \rho e_{i-j-1} + u_{i-j}, \quad j=1, 2, \dots$$

重复代换到(3·3)，得到

$$e_i = \sum_{k=0}^{\infty} \rho^k u_{i-k},$$

据此可以证明

$$E(e_i) = 0,$$

$$\text{Var}(e_i) = \sigma^2 \left(\frac{1}{1 - \rho^2} \right) \quad (3.4)$$

$$\text{COV}(e_i, e_{i+k}) = \rho^{|k|} \sigma^2 \left(\frac{1}{1 - \rho^2} \right)$$

可见，如果误差有关系(3·3)，则它们的均值为零，且方差相等，但当 $\rho \neq 0$ 时，彼此是相关的。

为了检验 e_i 的正相关，即检验假设

$$H_{01}: \rho = 0 \longleftrightarrow H_{11}: \rho > 0,$$

定义

$$D = \frac{\sum_{i=2}^n (\delta_i - \delta_{i-1})^2}{\sum_{i=1}^n \delta_i^2} \quad (3.5)$$

当 $\{e_i\}$ 是正相关时，从(3.3)知，相邻的项比较接近且符号大体一致，于是差 $e_i - e_{i-1}$ 应当倾向于比较小，因而 $\delta_i - \delta_{i-1}$ 也应当比较小。基于这种直观考虑，Durbin 和 Watson^[21]提出了检验：对给定的 α ，当 D 的观测值 $d < d_\alpha$ 时拒绝 H_0 ，并证明了这种检验具有一定的最优性质。但遗憾的是，当 H_0 成立时， D 的分布与设计阵 \tilde{X} 有关，于是临界点 d_α 必须根据设计阵 \tilde{X} 来计算，这在应用上多有不便。Durbin 和 Watson 提出了一种近似方法，证明了存在两个分布与设计阵 \tilde{X} 无关的随机变量 D_L 和 D_U ，使得： $D_L \leq D \leq D_U$ ，并对不同的 n 和 p ，计算出了 D_L 和 D_U 的1%，2.5%和5%分位点 d_L 和 d_U 的表，如表3.3.检验法则为

若 $d < d_L$ ，拒绝 H_0 ，

表3.3(a) d_L 和 d_U 的分位点($\alpha = 1\%$)

n	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69

(续上页)

23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.26	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

表3·3(b) d_L 和 d_U 的分位点($\alpha = 5\%$)

n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78

(续上页)

50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

表3.3(c) d_L 和 d_U 的分位点($\alpha=2.5\%$)

n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
16	0.98	1.24	0.86	1.40	0.75	1.59	0.64	1.80	0.53	2.03
17	1.01	1.25	0.90	1.40	0.79	1.58	0.68	1.77	0.57	1.98
18	1.03	1.26	0.98	1.40	0.82	1.56	0.72	1.74	0.62	1.93
19	1.06	1.28	0.96	1.41	0.86	1.55	0.76	1.72	0.66	1.90
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
21	1.10	1.30	1.01	1.41	0.92	1.54	0.83	1.69	0.73	1.84
22	1.12	1.31	1.04	1.42	0.95	1.54	0.86	1.68	0.77	1.82
23	1.14	1.32	1.06	1.42	0.97	1.54	0.89	1.67	0.80	1.80
24	1.16	1.33	1.08	1.43	1.00	1.54	0.91	1.66	0.83	1.79
25	1.18	1.34	1.10	1.43	1.02	1.54	0.94	1.65	0.86	1.77
26	1.19	1.35	1.12	1.44	1.04	1.54	0.96	1.65	0.88	1.76
27	1.21	1.36	1.13	1.44	1.06	1.54	0.99	1.64	0.91	1.75
29	1.22	1.37	1.15	1.45	1.08	1.54	1.01	1.64	0.93	1.74

(续上页)

n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
29	1.24	1.38	1.17	1.45	1.10	1.54	1.03	1.63	0.96	1.73
30	1.25	1.38	1.18	1.46	1.12	1.54		1.63	0.98	1.73
31	1.26	1.39	1.20	1.47	1.13	1.55	1.07	1.63	1.00	1.72
32	1.27	1.40	1.21	1.47	1.15	1.55	1.08	1.63	1.02	1.71
33	1.28	1.41	1.22	1.48	1.16	1.55	1.10	1.63	1.04	1.71
34	1.29	1.41	1.24	1.48	1.17	1.55	1.12	1.63	1.06	1.70
35	1.30	1.42	1.25	1.48	1.19	1.55	1.13	1.63	1.07	1.70
36	1.31	1.43	1.26	1.49	1.20	1.56	1.15	1.63	1.09	1.70
37	1.32	1.43	1.27	1.49	1.21	1.56	1.16	1.62	1.10	1.70
38	1.33	1.44	1.28	1.50	1.23	1.56	1.17	1.62	1.12	1.70
39	1.34	1.44	1.29	1.50	1.24	1.56	1.19	1.63	1.13	1.69
40	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
45	1.39	1.48	1.34	1.53	1.30	1.58	1.25	1.63	1.21	1.69
50	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
55	1.45	1.52	1.41	1.56	1.37	1.60	1.33	1.64	1.30	1.69
60	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
65	1.49	1.55	1.46	1.59	1.43	1.62	1.40	1.66	1.36	1.69
70	1.51	1.57	1.48	1.60	1.45	1.63	1.42	1.66	1.39	1.70
75	1.53	1.58	1.50	1.61	1.47	1.64	1.45	1.67	1.42	1.70
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
85	1.56	1.60	1.53	1.63	1.51	1.65	1.49	1.68	1.46	1.71
90	1.57	1.61	1.55	1.64	1.53	1.66	1.50	1.69	1.48	1.71
95	1.58	1.62	1.56	1.65	1.54	1.67	1.52	1.69	1.50	1.71
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72

若 $d > d_U$, 接受 H_0 ,

(3.6)

若 $d_L \leq d \leq d_U$, 不能做决定。

对于负相关的检验，即检验假设

$$H_0: \rho = 0 \longleftrightarrow H_1: \rho < 0$$

改用统计量 $4-D$ ，法则与(3.6)同。对于双边检验，即检验假设

$$H_0: \rho = 0 \longleftrightarrow H_1: \rho \neq 0$$

只要把前面两个单边检验联立起来，每个检验的水平取作 $\alpha/2$ ，就可得到水平为 α 的双边检验。

表3.4 饮料销售数据

年 份	年度销 售销量	年 度 广告费	普通残差 (δ_i)	δ_i^2	$(\delta_i - \delta_{i-1})^2$	该地区 人 口
1960 1	3083	75	-32.330	1045.2289		825000
1961 2	3149	78	-26.603	707.7196	32.7985	830445
1962 3	3218	80	2.215	4.9062	830.4771	838750
1963 4	3239	82	-16.967	287.8791	367.9491	842940
1964 5	3295	84	-1.148	1.3179	250.2408	846315
1965 6	3374	88	-2.512	6.3101	1.8605	852240
1966 7	3475	93	-1.967	3.8691	0.2970	860760
1967 8	3569	97	11.669	136.1656	185.9405	865925
1968 9	3597	99	-0.513	0.6232	148.4011	871640
1969 10	3725	104	27.032	730.7290	758.7270	877745
1970 11	3794	109	-4.422	19.5541	989.3541	886520
1971 12	3959	115	40.032	1602.5610	1976.1581	894500
1972 13	4043	120	23.577	555.8749	270.7670	900400
1973 14	4194	127	33.940	1151.9236	107.3918	904005
1974 15	4318	135	-2.787	7.7674	1348.8725	908525
1975 16	4493	144	-8.606	74.0632	33.8608	912160
1976 17	4683	153	0.575	0.3306	84.2908	917630
1977 18	4850	161	6.848	46.8951	39.3505	922220
1978 19	5005	170	-18.971	359.8988	666.6208	925910
1979 20	5236	182	-29.063	844.6580	101.8485	929610

例3.2 饮料销售问题(取自[22])

某公司欲预测一种饮料在一某个地区的年销售量，把自变量取作本公司在该区域内年度广告费用。从1960年—1979年的数据如表3.4。应用最小二乘法配一元回归，有关统计量如表3.5。因变量 Y 的观测值 Y_1, \dots, Y_n 是一个时间序列，对普通残差与时间作残差图，如图2.3.5。从这个图我们可以看到，残差有一个先向上后向下的趋势，因此误差很可能是正相关的。考虑假设 $H_0: \rho = 0 \longleftrightarrow H_1: \rho > 0$ 的检验。Durbin-Watson 统计量 $d = 1.08$ 。取显著性水平 $\alpha = 0.05$ ，从表3.3(b) $n = 20, p = 1$ 查到 $d_L = 1.20, d_U = 1.41$ 。现在 $d = 1.08 < d_L = 1.20$ ，于是我们拒绝原假设。即从现有的数据看，误差是正相关的。

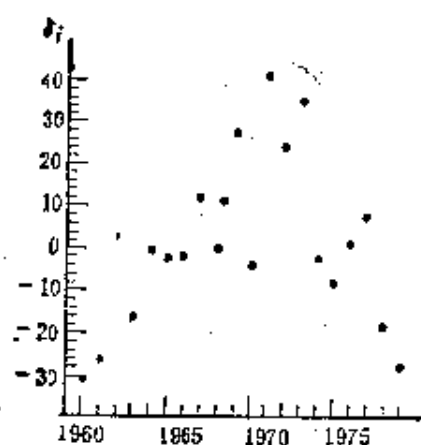


图2.3.5 残差 e_i 的残差图

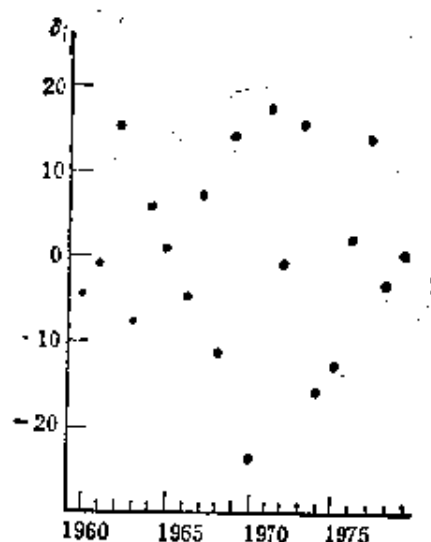


图2.3.6 残差 e_i 的残差图

表3.5 饮料销售问题的回归(一元)

参 数	最小二乘估计	t -统计量
α	1608.508	94.49
β_1	20.091	140.71
$n = 20,$	$R = 0.9991$	$\hat{\sigma}^2 = 421.5484$

表3.6 饮料销售问题的回归(二元)

参 数	最小二乘估计	t-统计量
α	320.340	1.47
β_1	18.434	63.23
β_2	0.002	5.83
$n = 20$	$R = 0.9997$	$\delta^2 = 145.3408$

残差呈现正相关,有可能是本质性的,即实际数据本身就有这种相依关系。也可能是由于漏掉了重要的回归自变量。分析现在的例子,很可能所考虑的地区的人口总数是影响销量的一个因素,应该把它考虑进来。表3.4最后一列为该地区历年人口总数。作二元线性回归,有关统计量如表3.6。相应的Durbin-Watson统计量 $d = 3.06$ 。对水平 $\alpha = 0.05$,查表3.3(b), $n = 20$, $p = 2$,得 $d_L = 1.10$, $d_U = 1.54$,推得 $d > d_U$ 。于是,从现在的数据看,还没有证据能够说明误差有正相关。普通残差对时间的残差图如图2.3.6。从图上可明显看出,添加了总人口这个自变量之后,残差图不再有正相关的趋势。

如果用添加自变量的方法不能消去残差的相关趋势,这时我们就认为误差彼此相关。对于这种情况,原则上我们总可以按协方差阵 $\text{COV}(e) = \sigma^2 G$ 的最一般情形用“两步估计法”来处理。即先假设 G 已知,用本章的方法求出 α, β 的LS“估计”,然后用 G 的估计 \hat{G} 代替LS估计中的 G ,得到所谓两步估计。假设误差 $\{e_i\}$ 满足(3.3),这时问题较简单, G 只包含一个未知参数 ρ ,它是 e_i 和 e_{i-1} 的相关系数,用残差 δ_i 作为 e_i 的“估计”,可以求出样本相关系数 $\hat{\rho}$,于是得到 G 的估计 \hat{G} 。由此求得 α, β 的两步估计。

§2.4 方差稳定化变换

从第一章我们知道,在GM假定

$$\text{COV}(e_i, e_j) = \begin{cases} \sigma^2, & i=j \\ 0, & i \neq j \end{cases}$$

下，常数项 α 和回归系数 β 的LS估计都是无偏估计，且具有方差最小性。但是，如果其中的等方差条件不满足，即方差具有非齐性，这时虽然LS估计仍然是无偏的，但一般不再具有方差最小性，这意味着回归系数估计的精度降低了。因此，如果残差图显示了误差方差的非齐性，人们往往考虑对因变量作变换，使得对变换过后的数据，误差方差能够近似相等，即方差比较稳定，所以通常称这种变换为**方差稳定化变换**。本节我们先对一般情况导出方差稳定化变换，然后给出常用的几种特殊变换。最后举一个例子。

设随机变量 Y 的均值为 μ ，方差为 σ^2 ，方差与均值有关系 $\sigma = \phi(\mu)$ ，这里 μ 是未知的， ϕ 是已知的。例如，若 Y 服从二项分布 $B(n, p)$ ，则 $E(Y) = np$ ， $\sigma^2 = np(1-p)$ 。于是 $\sigma = \phi(\mu) = [\mu(1-\mu)]^{1/2}$ ，其中 $\mu = np$ 。回到一般情况，我们欲寻找变换 $U = f(Y)$ ，使得 U 的方差等于或近似等于事先给定的常数 σ_u^2 。暂时假定 f 已经求到，对 U 在 $Y = \mu$ 附近作Taylor展开，取到一次项，得到如下近似关系

$$U = f(\mu) + f'(\mu)(Y - \mu). \quad (4.1)$$

求方差，有

$$\sigma_u^2 = \text{Var}(U) = [f'(\mu)]^2 \sigma^2$$

于是 $\sigma_u = f'(\mu)\sigma$ 。从而

$$f'(\mu) = \frac{\sigma_u}{\sigma}$$

积分得到

$$f(\mu) = \sigma_u \int \frac{d\mu}{\sigma} = \sigma_u \int \frac{d\mu}{\phi(\mu)}$$

于是所求的变换为

$$U=f(Y)=\sigma\int\frac{dY}{\phi(Y)} \quad (4.2)$$

根据(4.2), 容易推得下列方差稳定化变换:

(1) 若 $\sigma^2 \propto E(Y)$, 则作变换 $U=Y^{1/2}$, 例如 Y 服从 Poisson 分布就属于这种情况。

(2) 若 $\sigma^2 \propto E(Y)(1-E(Y))$, 则作变换 $U=\text{Sin}^{-1}\sqrt{Y}$

(3) 若 $\sigma^2 \propto [E(Y)]^2$, 则作变换 $U=\ln(Y)$.

(4) 若 $\sigma^2 \propto [E(Y)]^3$, 则作变换 $U=Y^{-1/2}$.

(5) 若 $\sigma^2 \propto [E(Y)]^4$, 则作变换 $U=Y^{-1}$.

在应用上, 首先从残差图粗略地考察一下 σ^2 与 $E(Y)$ 可能存在的几种关系, 然后从(4.2)求出对应的变换。对几种变换过的数据分别作 LS 处理, 作新的残差图, 看哪一种变换的残差图无方差非齐性的征兆。从中提出最好的方差稳定化变换。

在应用方差稳定化变换时, 我们需要注意的是, 当对因变量 Y 施以变换 $U=f(Y)$, 固然使方差变得近似相等, 但同时也使 Y 的均值发生变化。从(4.1)知, 新变量 U 的均值 $E(U) \approx f(\mu)$. 对均值的这种变化我们可能受益也可能受害。因为 μ 是自变量 x_1, \dots, x_p 的函数, 因此我们把它写成更明确的形式, 即 $\mu(x_1, \dots, x_p)$, 而 $E(U) \approx f(\mu) = f(\mu(x_1, \dots, x_p))$. 因为, 无论在变换前还是在变换后, 我们都是用线性回归模型 $\alpha + \beta_1 x_1 + \dots + \beta_p x_p$ 去近似 $\mu(x_1, \dots, x_p)$ 和 $f(\mu) = f(\mu(x_1, \dots, x_p))$, 若后者与线性回归模型更接近, 则从均值部分讲我们受益于变换, 不然的话, 变换后的均值距离线性形式更远了。举两个变量的简单例子来说明这种可能性。设 Y 的均值为

$$\mu(x_1, x_2) = \exp\{\beta_1 x_1 + \beta_2 x_2\},$$

用线性回归实际上就是用 $\mu(x_1, x_2)$ 的 Taylor 展开的一次近似。

如果做变换 $U = \ln(Y)$, 则 $E(U) \approx f(\mu) = \ln[\mu(x_1, x_2)] = \beta_1 x_1 + \beta_2 x_2$, 它是线性函数。相反的例子就更多了。

例4.1 用电量问题

研究用电高峰每小时的用电量 Y 与每月总用电量 x 之间的关系。53户居民1979年八月用电记录如表4.1。应用最小二乘法, 经验回归方程为 $Y = 0.8313 + 0.00368x$ 。复相关系数的平方 $R = 0.7046$ 。以拟合值 \hat{Y}_i 为横坐标, 普通残差 δ_i 为纵坐标的残差图如图2.4.1。从图上可以看出, 随着 \hat{Y}_i 的增大, 残差向外散开, 很象一个漏斗, 这是误差方差非齐性的一个证据。考虑对 Y 作方差稳定化变换。先试 $U = \sqrt{Y}$ 。用最小二乘法, 得到经验回归方程

$$U = 0.5822 + 0.0009529x.$$

图2.4.2是 U 的普通残差对拟合值 \hat{U}_i 的残差图。看上去 U 的方差已经稳定化, 所以, 所选的变换是适宜。

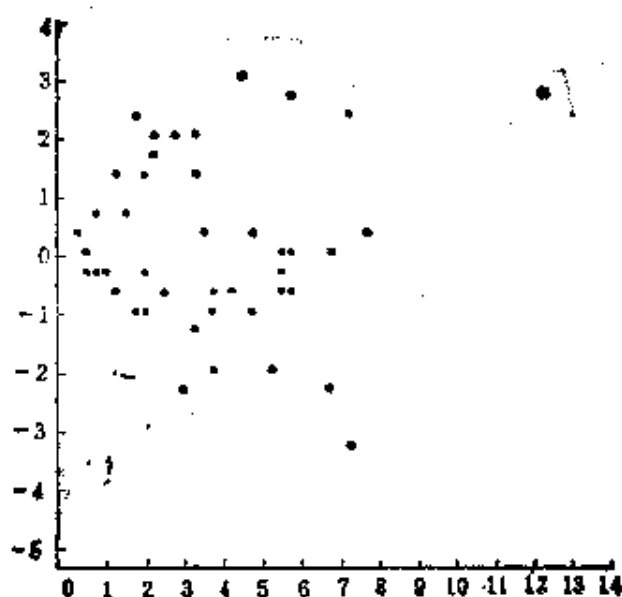


图2.4.1 普通残差 δ_i 的残差图

表4.1 用电量数据

用户	X	Y	用户	X	Y
1	679	.79	14	2030	4.43
2	292	.44	15	1643	3.16
3	1012	.56	16	414	.50
4	493	.79	17	354	.17
5	582	2.70	18	1276	1.88
6	1156	3.64	19	745	.77
7	997	4.73	20	435	1.39
8	2189	9.50	21	540	.56
9	1097	5.34	22	874	1.56
10	2078	6.85	23	1543	5.28
11	1818	5.84	24	1029	.64
12	1700	5.21	25	710	4.00
13	747	3.25	26	1434	.31
27	837	4.20	41	783	3.29
28	1748	4.88	42	406	.44
29	1381	3.48	43	1242	3.24
30	1428	7.58	44	658	2.14
31	1255	2.63	45	1746	5.71
32	1777	4.99	46	468	.64
33	370	.59	47	1114	1.90
34	2316	8.19	48	413	.51
35	1130	4.79	49	1787	8.33
36	463	.51	50	3560	14.94
37	770	1.74	51	1495	5.11
38	724	4.10	52	2221	3.85
39	808	3.94	53	1526	3.93
40	790	.96			

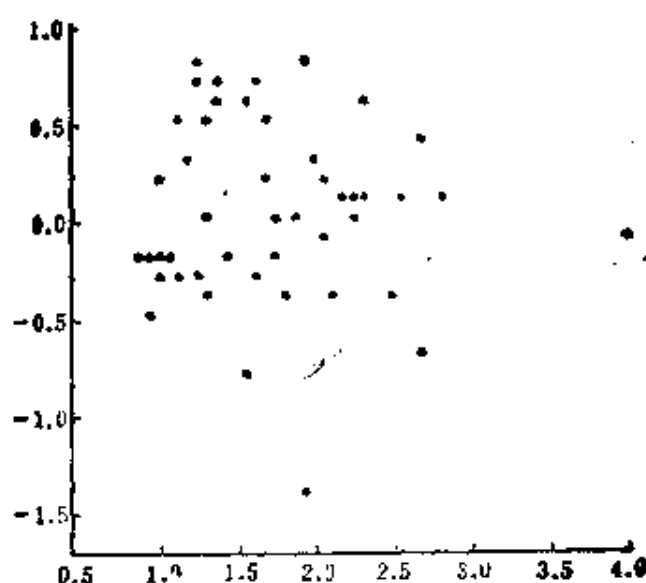


图2.4.2 变换后的残差图

§2.5 正态化变换

在回归分析中，当我们对回归系数的LS估计的分布性质作深入讨论以及对它们或回归方程作显著性检验或构造回归系数的置信区间时都需要假设模型误差服从正态分布，因此误差的正态性检验在回归诊断也是很重要的。检验一个分布是否为正态，可用的方法很多，例如 χ^2 -检验、Kolmogorov-Смирнов检验等等。本节我们介绍在应用上十分简便的一种方法——正态概率纸法。如果经检验，认为误差的分布是非正态的，那么需要对因变量作适当变换，使其分布更接近于正态分布。所以本节要讨论的第二个问题，就是如何选择适当的正态化变换。

(一) 正态概率纸

下面我们暂记 $\Phi(x, \mu, \sigma)$ 为正态分布 $N(\mu, \sigma^2)$ 的分布函数，

并简记 $\Phi(x)$ 为标准正态分布 $N(0, 1)$ 的分布函数, 即 $\Phi(x) = \Phi(x, 0, 1)$.

正态概率纸是一种特殊刻度的直角坐标纸。它的横坐标是普通的等分刻度, 而纵坐标采用一种不均匀刻度, 使得标准正态分布的点子 $(x, \Phi(x))$ 标在该坐标纸上, 落在第一象限的平分线 $y=x$ 上。而对一般的正态分布 $Z \sim N(\mu, \sigma^2)$, 我们知道 $X = (Z - \mu)/\sigma \sim N(0, 1)$. 于是 $Z = \mu + \sigma X$, $X \sim N(0, 1)$. 因为

$$\Phi(z, \mu, \sigma) = \Phi\left(\frac{z - \mu}{\sigma}, 0, 1\right) = \Phi(x)$$

所以

$$(z, \Phi(z, \mu, \sigma)) = (\mu + \sigma x, \Phi(x)). \quad (5.1)$$

由 $(x, \Phi(x))$ 是直线 $y=x$, 知 (5.1) 就是直线 $y = \mu + \sigma x$. 于是, 对任意的正态分布 $N(\mu, \sigma^2)$, 点集 $(z, \Phi(z, \mu, \sigma))$ 在正态概率纸上是一条直线, 均值 μ 和标准差 σ 分别为该直线的截距与斜率。图 2.5.1 是一张正态概率纸。这里纵轴刻度标的是 $100\Phi(x)$, 且最大值为 99.99, 而不是 1, 这是因为 $\Phi(\infty) = 1$ 的缘故。

假设 x_1, \dots, x_n 为来自总体分布为 $F(x)$ 的简单随机样本, 现在欲用正态概率纸来判定 $F(x)$ 是否正态分布。先将 x_1, \dots, x_n 依大小次序重排, 得到次序样本 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, 用经验分布函数

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{i}{n}, & x_{(i)} \leq x < x_{(i+1)} \\ 1, & x_{(n)} \leq x \end{cases} \quad (5.2)$$

作为 $F(x)$ 的近似。将点 $(x_{(i)}, F_n(x)) = (x_{(i)}, \frac{i}{n})$ 标在正态概率纸上, 如果 $F(x)$ 是正态分布, 则 $(x_{(i)}, \frac{i}{n})$, $i = 1, \dots, n$ 应呈直

线状。注意到正态概率纸的纵轴坐标最大值为0.9999，所以最后一个点 $(x_{(n)}, 1)$ 无法标上。我们把 $(x_{(i)}, \frac{i}{n})$ 修正为 $(x_{(i)}, \frac{i}{n+1})$ 或 $(x_{(i)}, \frac{i-1/2}{n})$ 。如果这些点在正态概率纸上呈直线状，则可以认

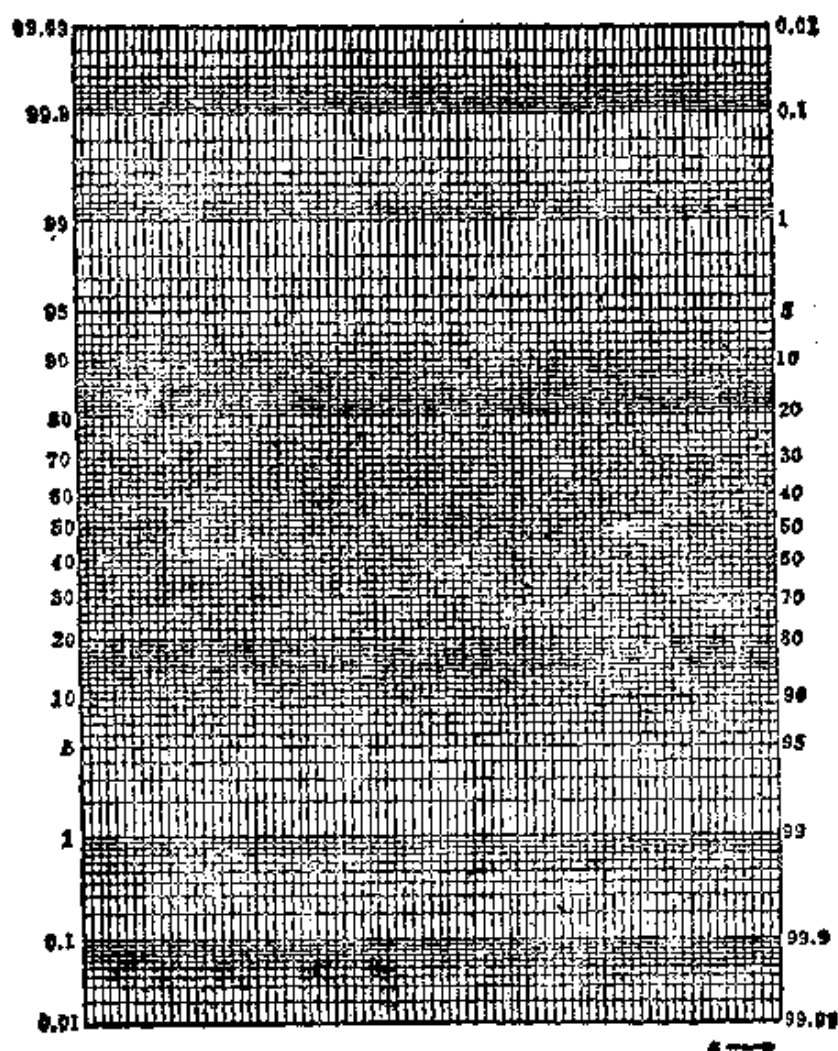


图2.5.1 正态概率纸

为分布 $F(x)$ 是正态分布。

用上面的方法可以检验线性回归模型误差正态性假设是否合理。从 §2.2 的讨论我们已知知道, 若 $e \sim N(0, \sigma^2 I)$, 学生化残差 r_i 近似服从 $N(0, 1)$, 且近似地相互独立, 所以可以将 r_1, \dots, r_n 视为来自 $N(0, 1)$ 的简单随机样本。将 r_1, \dots, r_n 依大小重新排序: $r_{(1)} \leq r_{(2)} \leq \dots \leq r_{(n)}$, 并把 $\left(r_{(i)}, \frac{i}{n+1}\right)$ 或 $\left(r_{(i)}, \frac{(i-1/2)}{n}\right)$,

$i=1, \dots, n$ 标在正态概率纸上, 若这些点呈直线状, 则可以认为误差正态性假设是合理的。不然的话, 认为误差与正态分布偏离较大, 需要考虑正态化变换。又若 $e \sim N(0, \sigma^2 I)$, 不相关残差 $\hat{e} \sim N(0, \sigma^2 I_{n-p-1})$ (见 (2.20)), 所以 $\hat{e}_1, \dots, \hat{e}_{n-p-1}$ 为来自 $N(0, \sigma^2)$ 的简单随机样本。因而也可以利用 \hat{e}_i 作正态概率纸检验。

(二) 正态化变换

如果经过检验认为误差不服从正态分布, 即因变量 Y 不服从正态分布, 则可以考虑对 Y 作适当变换, 使变换后新变量更接近正态分布。假设因变量 Y 取正值, Tukey 于 1957 年提出了如下幂变换族

$$Y^{(\lambda)} = \begin{cases} Y^\lambda, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases} \quad (5.3)$$

为了避免在 $\lambda=0$ 的间断性, Box 和 Cox 于 1964 年^[23] 提出了如下修正变换族

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases} \quad (5.4)$$

对不同的 λ , 所作的变换就不同, 这个变换族包括了对数变换 ($\lambda =$

0)"平方根变换($\lambda=1/2$)和倒数变换($\lambda=-1$)等上节所讨论过的变换。如果记 $Y^{(\lambda)}=(Y_1^{(\lambda)}, \dots, Y_n^{(\lambda)})'$ (这个记号与(5.3)、(5.4)的 $Y^{(\lambda)}$ 一样,但意义不同,不过这不会引起混淆),并假定对固定的 λ ,变换过的数据满足正态线性回归模型

$$Y_{n \times 1}^{(\lambda)} = \tilde{X} \gamma + e, \quad e \sim N(0, \sigma^2 I) \quad (5.5)$$

其中变换参数 λ 和模型参数 γ, σ^2 一样,将通过数据来确定,以下我们以 Box-Cox 变换(5.4)为例,讨论如何确定 λ ,有了 λ ,从(5.5)按通常LS法就可得到 γ, σ^2 的LS估计。

因为 $Y^{(\lambda)} \sim N(\tilde{X}\gamma, \sigma^2 I_n)$,所以基于变换前数据 $Y'=(Y_1, \dots, Y_n)$, σ^2 的似然函数为

$$L(\gamma, \sigma^2; Y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} (Y^{(\lambda)} - \tilde{X}\gamma)' (Y^{(\lambda)} - \tilde{X}\gamma)\right\} J, \quad (5.6)$$

这里, J 为Jacobi行列式的绝对值

$$J = \prod_{i=1}^n \left| \frac{dY_i^{(\lambda)}}{dY_i} \right| = \prod_{i=1}^n Y_i^{\lambda-1}$$

对固定的 λ ,除了常数因子 J 以外, $L(\gamma, \sigma^2; Y)$ 就是通常正态线性回归模型的似然函数。于是, γ 的极大似然估计(等于LS估计),残差平方和以及似然函数的最大值分别为

$$\hat{\gamma}^{(\lambda)} = (\tilde{X}' \tilde{X})^{-1} \tilde{X}' Y^{(\lambda)} \quad (5.7)$$

$$\begin{aligned} \text{RSS}(\lambda, Y^{(\lambda)}) &= n \hat{\sigma}^{(\lambda)2} \\ &= Y^{(\lambda)'} (I_n - \tilde{X}(\tilde{X}' \tilde{X})^{-1} \tilde{X}') Y^{(\lambda)} \end{aligned} \quad (5.8)$$

$$L_{\max}(\lambda) = e^{-\frac{n}{2}} \cdot J (2\pi \hat{\sigma}^{(\lambda)2})^{-\frac{n}{2}} \quad (5.9)$$

若略去常数因子,则(5.9)的对数为

$$\ln L_{\max}(\lambda) = -\frac{n}{2} \ln [\text{RSS}(\lambda, Y^{(\lambda)})] + \ln J$$

$$= -\frac{n}{2} \ln[\text{RSS}(\lambda, Z^{(1)})] \quad (5.10)$$

这里

$$\begin{aligned} \text{RSS}(\lambda, Z^{(1)}) &= Z^{(1)'} (I_n - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}') Z^{(1)} \\ Z^{(1)} &= (Z_1^{(1)}, \dots, Z_n^{(1)})', \\ Z_i^{(1)} &= \frac{Y_i^{(1)}}{J^{1/2}} = \begin{cases} \frac{Y_i^{(1)}}{[g(Y)]^{\lambda-1}}, & \lambda \neq 0, \\ g(Y) \ln(Y_i), & \lambda = 0. \end{cases} \end{aligned} \quad (5.11)$$

$g(Y) = g(Y_1, \dots, Y_n)$ 表 Y_1, \dots, Y_n 的几何平均, 即

$$g(Y) = \sqrt[n]{Y_1 Y_2 \dots Y_n} \quad (5.12)$$

按照似然原理, 我们选择使 $\ln L_{\max}(\lambda)$ 达到最大的 λ 作为变换参数值。这等价于解方程

$$\frac{d \ln L_{\max}(\lambda)}{d\lambda} = 0 \quad (5.13)$$

并把 $\hat{\lambda}$ 称为极大似然估计。因为 λ 是一个数, 于是在应用上可以作 $\ln L_{\max}(\lambda)$ 的图, 从图上找出最大值点 λ 。

Box-Cox 变换 (5.4) 只能应用于 Y 的观测值为正的情形。如果 Y 可能取到负值, 则我们需要对 (5.4) 作一些修正。例如采用推广的幂变换族

$$Y^{(1)} = \begin{cases} \frac{(Y + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, & \lambda_1 \neq 0 \\ \ln(Y + \lambda_2), & \lambda_1 = 0 \end{cases} \quad (5.14)$$

这里 $Y + \lambda_2 > 0$ 。这实际上可以看作将 Y 作了平移对平移后数据再作 Box-Cox 变换。对变换 (5.14), 相应地 (5.11) 变为

$$Z_i^{(1)} = \begin{cases} \frac{(Y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1 [g(Y + \lambda_2)]^{\lambda_1 - 1}}, & \lambda_1 \neq 0 \\ g(Y + \lambda_2) \ln(Y_i + \lambda_2), & \lambda_1 = 0 \end{cases} \quad (5.15)$$

其中 $\mathbf{1} = (1, \dots, 1)'$, 即 n 个元素皆为 1 的 $n \times 1$ 向量。在这里左端 $\lambda = (\lambda_1, \lambda_2)'$ 。因此, 对这种变换, (5.13) 是以 λ_1, λ_2 为未知量的方程组, 求解比较困难。即便对 $\ln L_{\max}(\lambda_1, \lambda_2)$ 作图也不可能, 因为这是一个二元函数。在应用上一种简便方法是, 根据数据直观确定 λ_2 的值, 然后应用 Box-Cox 变换。

虽然 Box-Cox 变换常常用于误差正态化的目的, 但是从 (5.5) 及选择 λ 的极大似然原则, 我们知道, Box-Cox 变换要求选择 λ , 使得 $Y^{(1)}$ 从均值的线性性、误差等方差和不相关、以及误差正态性等多个方面都与正态线性回归模型很接近, 因此, Box-Cox 变换不仅可用于误差正态化, 也可用于其它多种目的。

例5.1 用电量问题(续例4.1)

在例4.1中, 我们应用平方根变换 $Z = \sqrt{Y}$ 使方差稳定化。现在 we 再用 Box-Cox 变换研究这个问题。从 (5.10) 知, 求 $\ln L_{\max}(\lambda)$ 的最大值等价于求残差平方和 $RSS(\lambda, Z^{(1)})$ 的最小值。对不同的 λ 计算 $RSS(\lambda, Z^{(1)})$, 得到下表

表5.1 变换后的残差平方和

λ	-2	-1	-0.5	0	0.125	0.25
RSS	34,101.04	986.04	291.59	134.10	119.20	107.21
λ	0.375	0.5	0.625	0.75	1	2
RSS	100.26	96.95	97.29	101.69	127.87	1,275.56

残差平方和 $RSS(\lambda, Z)$ 在 $\lambda = 0.5$ 达到最小。这就证实了上节所作的平方根变换是合适的。

例5.2 树干体积问题

表5.2是树干的体积 Y 与离地面一定高度的树干直径 x_1 和树干高度 x_2 的

一组数据。目的是要研究 Y 与 x_1, x_2 之间的关系,以便能够用简单的方法从 x_1 和 x_2 估计一棵树的体积,进而估计一片森林的木材储量。因为树木可以看成是一个圆柱或圆锥,所以它的体积 Y 并不是 x_1, x_2 的线性函数。若对模型 $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$,应用最小二乘法,求得LS估计: $\hat{\alpha} = -57.987$, $\hat{\beta}_1 = 4.708$, $\hat{\beta}_2 = 0.339$,以 X_1 为横轴学生化残差为纵轴的残差图如图2.5.2.这个图显示了非线性趋势。对 $\ln L_{\text{mix}}(\lambda)$ 作图,图2.5.3.最大值约在 $\hat{\lambda} = 0.307$ 达到,近似地取 $\hat{\lambda} = 1/3$.这个上结果建议了对因变量作变换 $Z = Y^{(1/3)}$.对变换后的数据作LS估计后再一次作残差图,可以看到,原来的非线性趋势有了明显改进。

表5.2 树干体积数据

X_1 (直径)	X_2 (高)	Y (体积)	X_1 (直径)	X_2 (高)	Y (体积)
3.3	70	10.3	12.9	85	33.8
8.6	65	10.3	13.3	86	27.4
8.8	63	10.2	13.7	71	25.7
10.5	72	26.4	13.8	64	24.9
10.7	81	18.8	14.0	78	34.5
10.8	83	19.7	14.2	80	31.7
11.0	66	15.6	14.5	74	36.3
11.0	75	18.2	16.0	72	38.3
11.1	80	22.6	16.3	77	42.6
11.2	75	19.9	17.3	81	55.4
11.3	79	24.2	17.5	82	55.7
11.4	76	21.0	17.9	80	58.3
11.4	76	21.4	18.0	80	51.5
11.7	69	21.3	18.0	80	51.0
12.0	75	19.1	20.6	87	77.0
12.9	74	22.2			

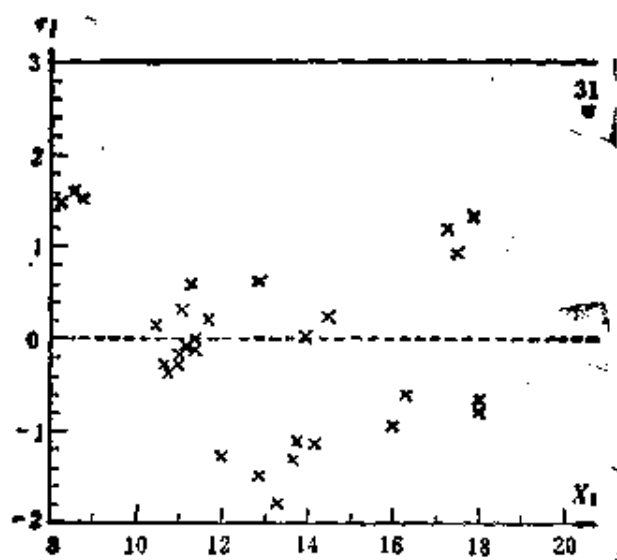


图2.5.2 残差图

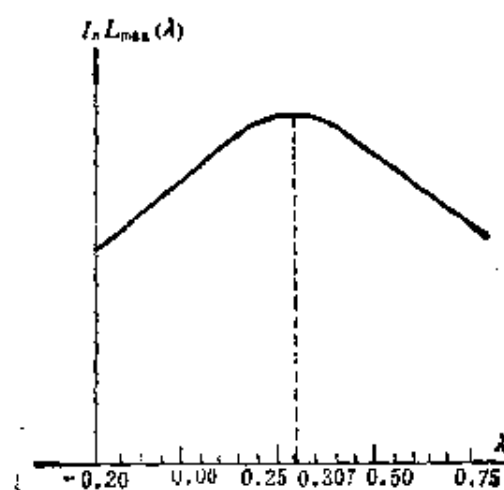


图2.5.3

§2.6 影响函数

前面几节我们集中讨论了回归诊断的第一个问题，即考察模型假设的合理性。其中残差是最重要的诊断统计量。本节我们转到回归诊断的第二个问题，即探查对回归推断(如估计、预测)具

有很大影响的数据，即强影响点。

我们仍然讨论线性回归模型(1.1)。度量其中的一组数据 $(x_{1i}, \dots, x_{pi}, Y_i)$ 对回归估计影响大小的一种重要统计量是**经验影响函数**，定义为 $IF_i = \hat{\gamma}(i) - \hat{\gamma}$ 。和前面一样，这里 $\hat{\gamma}(i)$ 表剔除第 i 组数据之后，从剩余的 $n-1$ 组数据算出的 γ 的LS估计。很显然， IF_i 刻画了由于第 i 组数据的剔除，回归系数的LS估计变化的大小，因此它是第 i 组数据对估计影响大小的度量。因为 IF_i 是一个向量，应用上很不方便，于是需要考虑它的某种数量函数。

Cook等^[24]引进的 IF_i 的一种函数是距离

$$D_i(M, c) = \frac{(\hat{\gamma}(i) - \hat{\gamma})' M (\hat{\gamma}(i) - \hat{\gamma})}{c} \quad (6.1)$$

这里方阵 M 是给定的正定阵， c 为给定的正数。 $D_i(M, c)$ 愈大表示第 i 组数据剔除之后， $\hat{\gamma}$ 移动的距离愈大。因此 $D_i(M, c)$ 度量了第 i 组数据 $(x_{1i}, \dots, x_{pi}, Y_i)$ 对估计 $\hat{\gamma}$ 影响的大小。将(2.17)代入(6.1)，得

$$\begin{aligned} D_i(M, c) &= \frac{\delta_i^2}{\hat{\sigma}^2(1-h_{ii})} \cdot \frac{\hat{\sigma}^2}{c} \\ &\quad \cdot \frac{\tilde{x}_i' (\tilde{X}' \tilde{X})^{-1} M (\tilde{X}' \tilde{X})^{-1} \tilde{x}_i}{(1-h_{ii})} \\ &= r_i^2 \frac{\hat{\sigma}^2}{c} P_i(M) \end{aligned} \quad (6.2)$$

这里 r_i 为学生化残差，而

$$P_i(M) = \frac{\tilde{x}_i' (\tilde{X}' \tilde{X})^{-1} M (\tilde{X}' \tilde{X})^{-1} \tilde{x}_i}{1-h_{ii}}$$

(6.2)式给出了从完全数据回归分析的结果计算 $D_i(M, c)$ 的简便

方法。更进一步, (6.2) 把 $D_i(M, c)$ 分解成三个因子。其中 $\hat{\sigma}^2/c$ 与 i 无关, 不去讨论它。第一个因子 r_i^2 度量了模型在点 (x_i', Y_i) 处拟合的好坏。第三个因子 $P_i(M)$ 本质上刻画了点 x_i 在自变量空间 R^p 中的位置。可见, 一组数据对 LS 估计影响的大小既与模型在该点拟合好坏有关, 又与该点的位置有关。若 $D_i(M, c)$ 很大, 则称对应的点 (x_i', Y_i) 为强影响点。但是究竟多大的 $D_i(M, c)$ 才算很大呢? 这不能一概而论。它与 M, c 的选择以及具体问题都有关。 M 和 c 的一种常用选择是取 $M = \tilde{X}' \tilde{X}$, $c = (p+1)\hat{\sigma}^2$, 此时 (6.1) 变为

$$\begin{aligned} D_i &\triangleq D_i(\tilde{X}' \tilde{X}, (p+1)\hat{\sigma}^2) \\ &= \frac{(\hat{\gamma}(i) - \hat{\gamma})' \tilde{X}' \tilde{X} (\hat{\gamma}(i) - \hat{\gamma})}{(p+1)\hat{\sigma}^2} \end{aligned} \quad (6.3)$$

应用置信椭球可以对 M, c 的这种选法给予一定的理论支持。在误差正态假设下, $\hat{\gamma} \sim N(\gamma, \sigma^2 (\tilde{X}' \tilde{X})^{-1})$, 所以

$$\begin{aligned} \frac{(\gamma - \hat{\gamma})' \tilde{X}' \tilde{X} (\gamma - \hat{\gamma})}{\sigma^2} &\sim \chi^2_{p+1} \\ \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-p-1} \end{aligned}$$

且二者相互独立, 因而

$$\frac{(\gamma - \hat{\gamma})' \tilde{X}' \tilde{X} (\gamma - \hat{\gamma})}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-p-1} \quad (6.4)$$

集合

$$S = \left\{ \gamma_i: \frac{(\gamma - \hat{\gamma})' \tilde{X}' \tilde{X} (\gamma - \hat{\gamma})}{(p+1)\hat{\sigma}^2} \leq F_{p+1, n-p-1}(\alpha) \right\}$$

称为未知参数 γ 的置信系数为 $1-\alpha$ 的置信椭球。虽然(6.3)与(6.4)很相似,但前者并不服从 F 分布。然而借助于后者我们可以对 D_i 值大小给出概率解释。例如,若 $D_i = F_{p+1, n-p-1}(0.50)$,则表明第 i 组数据 (x_i', Y_i) 剔除后, γ 的估计 $\hat{\gamma}(i)$ 移动到了 γ 的置信系数为0.50的置信椭球边界上,若 $D_i = F_{p+1, n-p-1}(0.80)$,则显然第 j 组数据 (x_j', Y_j) 对估计的影响比第 i 组数 (x_i', Y_i) 来得大。对于 M, c 的这种选择,一种观点认为只要 $D_i > 1$,就可以认为 (x_i', Y_i) 为强影响点。将 $M = \tilde{X}' \tilde{X}, c = (p+1)\hat{\sigma}^2$ 代入(6.2),得到

$$D_i = \frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}} \quad (6.5)$$

此时, $P_i(M) = h_{ii}/(1-h_{ii})$ 是 h_{ii} 的单调增函数。可见, h_{ii} 愈大, $P_i(M)$ 愈大。因为 h_{ii} 在几何上是数据 x_i 在自变量空间 R^p 中距离试验区域中心远近的度量,所以 $P_i(M)$ 也就度量了这种距离的大小。在前面的讨论中,我们曾经把 h_{ii} 很大的点称为高杠杆点。另一方面, r_i 为学生化残差,一组数据 (x_i', Y_i) 如果对应的残差 $\hat{\delta}_i$ 或 r_i 很大,我们称 (x_i', Y_i) 或 x_i 为异常点(outlier)。(6.5)表明,异常点和高杠杆点都可能是强影响点,但又不一定都是强影响点。因此(6.5)从数学上描述了这三种点的关系。

方阵 M 和数 c 的另一种选择是 $M = \tilde{X}' \tilde{X}, c = \hat{\sigma}^2(i)^2$, 这里 $\hat{\sigma}^2(i)$ 由(2.13)定义。此时

$$D_i(\tilde{X}' \tilde{X}, \hat{\sigma}^2(i)^2) = \frac{(\hat{\gamma}(i) - \hat{\gamma})' \tilde{X}' \tilde{X} (\hat{\gamma}(i) - \hat{\gamma})}{\hat{\sigma}^2(i)^2} \quad (6.6)$$

它是从预测角度提出的。记 $\hat{Y}_i = \tilde{x}_i' \hat{\gamma}$, 它是从完全数据回归得到的 x_i 处的拟合值,但也可以看作对 x_i 处 Y 值的预测。 $\hat{Y}_i(i) = \tilde{x}_i' \hat{\gamma}(i)$ 为剔除第 i 组数据 (x_i', Y_i) 之后,从剩余的 $n-1$ 组数据回归算得的 x_i 处 Y 值的预测。于是 $\hat{Y}_i - \hat{Y}_i(i) = \tilde{x}_i' (\hat{\gamma} - \hat{\gamma}(i))$ 刻画了第 i

组数据对 x_i 处预测影响的大小。将(2·17)代入, 得到

$$\hat{Y}_i - \hat{Y}_{i(i)} = \frac{h_{ii}}{1 - h_{ii}} \delta_i, \quad (6.7)$$

因为 $\hat{Y}_i = \tilde{x}_i' \hat{\gamma}$ 的标准差为 $\sigma h_{ii}^{1/2}$, 用 $\hat{\sigma}(i)$ 代替 σ , 去除(6·7), 得到

$$W_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{[\text{Var}(\hat{Y}_i)]^{1/2}} = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} r_i^*, \quad (6.8)$$

这里 r_i^* 由(2·12)所定义。容易验证, $W_i^2 = D_i(\tilde{X}'\tilde{X}, \hat{\sigma}(i)^2)$.

从 W_i^2 (或 W_i)的定义我们知道, W_i^2 度量了第 i 组数据对 x_i 处的预测影响的大小。一个问题是, W_i^2 与第 i 组数据对其它点的预测影响的关系如何呢? 为了回答这个问题, 我们先证明一个引理。

引理6·1 设 $A_{n \times n} \geq 0$, $B_{n \times n} > 0$, 记 $\lambda_1 \geq \dots \geq \lambda_n$ 为 $|A - \lambda B| = 0$ 的根, 则

$$\sup_x \frac{x'Ax}{x'Bx} = \lambda_1$$

证明 先考虑 $B=I$ 的特殊情况。因为 $A \geq 0$, 故存在正交方阵 P , 致

$$A = P \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix} P',$$

这里 $\mu_1 \geq \dots \geq \mu_n \geq 0$ 为 A 的特征根。记 $y = Px$, 则

$$\sup_x \frac{x'Ax}{x'Bx} = \sup_y \frac{y' \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix} y}{y'y} = \sup_y \frac{\sum_i \mu_i y_i^2}{\sum_i y_i^2} = \mu_1$$

注意, 此时 $\mu_1 = \lambda_1$. 这就完成 $B=I$ 时的证明。

对一般的正定阵 B , 记 $B^{1/2}$ 为 B 的平方根方阵, 即 $B=(B^{1/2})^2$, 又记 $y=B^{1/2}x$, 于是利用已证事实有

$$\sup_x \frac{x'Ax}{x'Bx} = \sup_y \frac{y' B^{-\frac{1}{2}} A B^{-\frac{1}{2}} y}{y'y} = \delta_1$$

这里 δ_1 为 $B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ 的最大特征根, 即 $|B^{-\frac{1}{2}} A B^{-\frac{1}{2}} - \delta I| = 0$ 的最大根。也就是 $|A - \delta B| = 0$ 的最大根。所以, $\delta_1 = \lambda_1$. 引理证毕.

根据这个引理

$$\begin{aligned} & \sup_{\tilde{x}} \frac{|\tilde{x}'(\hat{\gamma} - \hat{\gamma}(i))|}{(\tilde{x}'(\tilde{X}'\tilde{X})^{-1}\tilde{x})^{1/2}} \\ &= \left[\sup_{\tilde{x}} \frac{\tilde{x}'(\hat{\gamma} - \hat{\gamma}(i))(\hat{\gamma} - \hat{\gamma}(i))'\tilde{x}}{\tilde{x}'(\tilde{X}'\tilde{X})^{-1}\tilde{x}} \right]^{1/2} \\ &= [(\hat{\gamma} - \hat{\gamma}(i))'\tilde{X}'\tilde{X}(\hat{\gamma} - \hat{\gamma}(i))]^{1/2} \end{aligned}$$

于是, 我们证明了, 对一切 \tilde{x} 有

$$\frac{[\tilde{x}'\hat{\gamma} - \tilde{x}'\hat{\gamma}(i)]^2}{\hat{\sigma}(i)^2 \tilde{x}'(\tilde{X}'\tilde{X})^{-1}\tilde{x}} \leq W_i^2 = D_i(\tilde{X}'\tilde{X}, \hat{\sigma}(i)^2)$$

(6.9)

(6.9)左边的分子为完全数据和剔除第 i 组数据的两种回归在 x 处预测值之差的平方。分母为预测 $\tilde{x}'\hat{\gamma}$ 的方差估计。所以(6.9)的左边度量了第 i 组数据对任一点 x 处预测影响的大小。这个不等式表明 W_i^2 为这种预测影响的上界。

除了上面讨论的 M 、 c 的两种选择之外, 近年来, 许多作者还提出了另外一些选择。见表6.1.一般它们都是从某种直观意义提出的。目前尚未见到对这些距离的统计性质的理论或模拟比较。表面上看来, 似乎它们所能提供的信息大体相当。

表6.1 $D_i(M, c)$ 的几种选择

M	c	化简形式
$\tilde{X}'\tilde{X}$	$(p+1)\hat{\sigma}^2$	$\frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}}$
$\tilde{X}'\tilde{X}$	$(p+1)\hat{\sigma}(i)^2$	$\frac{1}{p+1} r_i^{*2} \frac{h_{ii}}{1-h_{ii}}$
$\tilde{X}'_{(i)}\tilde{X}_{(i)}$	$(p+1)\hat{\sigma}^2$	$\frac{1}{p+1} r_i^2 h_{ii}$
$\tilde{X}'_{(i)}\tilde{X}_{(i)}$	$(p+1)\hat{\sigma}(i)^2$	$\frac{1}{p+1} r_i^{*2} h_{ii}$
I	$(p+1)\hat{\sigma}^2$	$\frac{1}{p+1} r_i^2 \frac{\tilde{x}'_i(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i}{1-h_{ii}}$
$\tilde{X}'\tilde{X}$	$\hat{\sigma}(i)^2$	$r_i^{*2} \frac{h_{ii}}{1-h_{ii}}$

凡与其它点比较起来, 具有很大的 $D_i(M, c)$ 的点 (x'_i, Y_i) 称为强影响点。强影响点探查出来之后, 应该如何处理, 是剔除、保留还是缩小它在LS估计中的影响(如采用稳健估计, 见第五章), 没有一个简单化的答案。必须根据问题的专业知识、数据收集的实际情况, 分析强影响点产生的原因, 然后斟酌对待。一些人认为, 如果强影响点来自于数据收集的失误, 则应剔除。如果这样的数据是所考察的系统本身可能产生的, 则应予保留, 并仔细加以分析, 有时这样的数据能够提供普通数据所不能提供的有用信息。

例6.1 老鼠试验

试验者研究老鼠肝脏中某种药物的含量。他们随机地抽取19只老鼠, 称过体重, 然后放在喷过微量乙醚麻醉剂的容器中, 并喂给某种口服药。一般认为, 肝脏愈大, 药物吸收量愈大, 而肝重又与体重密切相关。于是, 他们给每只老鼠实际服药量近似地按每千克体重服40毫克药的比例确定。在服药一定时间后, 所有老鼠都死去。解剖后称其肝重, 并计算肝里所含药量与肝

重的百分比。

记 Y 为肝中含药量与肝重的百分比， x_1 为鼠体重， x_2 为肝重， x_3 为服药量。试验数据和样本相关系数分别列在表6·2和表6·3。从表6·3最后一行看到， Y 与三个自变量 x_1 ， x_2 ， x_3 的样本相关系数都很小。如果将 Y 对 x_1 或 x_2 或 x_3 作一元回归，计算一元回归系数的 t 值，即(1·3·22)的统计量 F 值的平方根。现在 $n=19$ ， $p=1$ ，故此 t 分布的自由度为17。取显著性水平 $\alpha=0.05$ ， $t_{17}(0.05)=2.11$ ，仅当 t 值大于2.11时， t 检验才是显著的。从表6·4看到，所有一元回归系数都是不显著的。我们再看 Y 对 x_1 ， x_2 和 x_3 的三元

表6·2 老鼠试验数据

x_1 (克)	x_2 (克)	x_3	Y
176	6.5	0.88	0.42
176	9.5	0.88	0.25
190	9.0	1.00	0.56
176	8.9	0.88	0.23
200	7.2	1.00	0.23
167	8.9	0.83	0.32
168	8.0	0.94	0.37
195	10.0	0.98	0.41
176	8.0	0.88	0.33
165	7.9	0.84	0.38
158	6.9	0.80	0.27
148	7.3	0.74	0.36
149	5.2	0.75	0.21
163	8.4	0.81	0.28
170	7.2	0.85	0.34
186	6.8	0.94	0.28
146	7.3	0.73	0.30
181	9.0	0.90	0.37
149	6.4	0.75	0.46

表6.3 样本相关系数

	x_1	x_2	x_3	Y
x_1	1.000			
x_2	0.500	1.000		
x_3	0.990	0.490	1.000	
Y	0.151	0.203	0.228	1.000

表6.4 回归系数

	模型所含的自变量			
参数	x_1	x_2	x_3	(x_1, x_2, x_3)
a	0.196 (0.89)	0.220 (1.64)	0.133 (0.63)	0.266 (1.37)
β_1	0.0008 (0.63)			-0.0212 (-2.67)
β_2		0.0147 (0.86)		0.0143 (0.83)
β_3			0.235 (0.96)	4.178 (2.74)

注：括号中的数为对应的 t 值

线性回归，此时情况就不同了。自变量 x_1 与 x_3 的回归系数的 t 值都大于 $t_{1,0.05} = 2.13$ （注意，此时 $p = 3$ ）。这表明， Y 与 x_1 和 x_3 有线性回归关系。如果把 x_2 剔除，我们仍然得到同样的结论。这就与一元情况所得到的结论相矛盾。

现在我们就来诊断问题所在。表6.5列出了 Y 对 x_1, x_2, x_3 的三元回归的主要诊断统计量。从普通残差 δ_i 和学生化残差 r_i 看不出多少异常可以解释上述矛盾。但是从 h_{ii} ， D_i 两列发现， $h_{33} = 0.8509$ ，而其余 $0 \leq h_{ii} \leq 0.4$ ，同样， $D_3 = 0.9296$ ，其余 $D_i \leq 0.273$ 。这表明第三组数据对回归系数估计的影响特别大，相应的自变量与其余试验点也有很大不同，它远离试验区域的

表6.5 Y对 x_1, x_2, x_3 的三元线性回归的诊断统计量(全部数据)

序号	Y_i	δ_i	r_i	h_{ii}	D_i
1	0.4200	0.1238	1.7660	0.1780	0.1688
2	0.2500	-0.8914E-01	-1.2730	0.1793	0.0885
3	0.5600	0.2409E-01	0.8072	0.8509	0.9296
4	0.2300	-0.1006	-1.3772	0.1076	0.0572
5	0.2300	-0.6771E-01	-1.1231	0.3915	0.2029
6	0.3200	0.7131E-02	0.1007	0.1612	0.0005
7	0.3700	0.5658E-01	0.7880	0.1369	0.0246
8	0.4100	0.4958E-01	0.7426	0.2537	0.0469
9	0.3300	0.1231E-01	0.1649	0.670	0.0005
10	0.3800	-0.2845E-02	-0.0392	0.1197	0.0001
11	0.2700	-0.8015E-01	-1.1051	0.1195	0.0414
12	0.3600	0.4236E-01	0.6024	0.1724	0.0189
13	0.2100	-0.9815E-01	-1.5357	0.3162	0.2726
14	0.2800	-0.2714E-01	-0.3768	0.1314	0.0054
15	0.3400	0.3161E-01	0.4256	0.0762	0.0037
16	0.2800	-0.5875E-01	-0.8589	0.2166	0.0510
17	0.3000	-0.1835E-01	-0.2647	0.1952	0.0042
18	0.3700	0.6068E-01	0.8509	0.1487	0.0316
19	0.4600	0.1347	1.9220	0.1780	0.1999

注：表中E-01表示 10^{-1} ，余类推

中心。于是第三组数据为强影响点，它也是高杠杆点。如果我们剔除第三组数据，重新作最小二乘回归，结果见表6.6，此时所有的回归系数的 t 值都不显著了。这说明了 Y 与 x_1, x_2, x_3 无线性回归关系。通过上述分析，我们可以得到这样的结论：原来的矛盾是由第三组数据引起的。再仔细检查表6.2，第三只老鼠体重190克，而服药量为1个单位，这个数字比它应该服的药量要大。例如第八只老鼠体重195克，才服了0.95个单位的药。这时应该再仔细核对一下记录以及实验过程，看是否有什么失误。

表6.6 删除第三组数据后的回归

参 数	LS估计	t 值	
α	0.311	1.52	$\hat{\sigma}^2 = 0.612E-02$
β_1	$-0.738E-02$	-0.42	$R = 0.0211$
β_2	$0.899E-02$	0.48	
β_3	1.485	0.40	

这个例子说明，仅仅一组不寻常数据，就能够导致完全相反的结论。因此，在应用回归分析处理实际问题时，我们必须对数据获得过程有较详细的了解。那种把数据拿来就套用现成公式的机械搬用，难以取得好的应用效果，是不可取的。

前面我们讨论了单组数据对回归影响大小的度量。从应用角度看，每次考察一组数据就基本上可以得到数据对回归影响的大部分信息。但是也存在这样的情况，一组数据对回归影响不大，但它们的联合影响却很大。图2.6.1可以直观地说明这一点。例如，

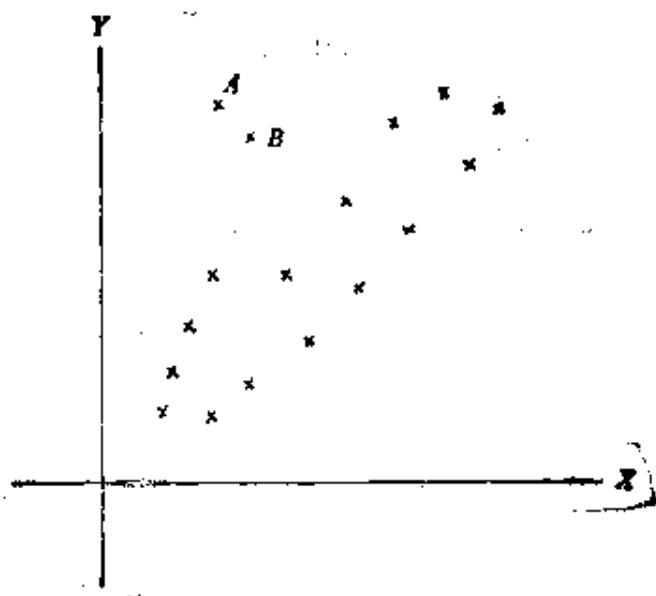


图2.6.1

点A或B被剔除掉, 经验回归直线几乎没有什么变化或变化很小, 但如果把这两点同时剔除掉, 则经验回归直线就将有较大的移动。下面我们简略地讨论一下多组数据对回归影响大小的度量问题。

记 $I = \{i_1, i_2, \dots, i_k\}$ 为剔除的数据序号。用 $Y(I)$ 、 $\tilde{X}(I)$ 、 $e(I)$ 分别表示从观测向量 Y 、设计阵 \tilde{X} 以及随机误差向量 e 剔除第 i_1, \dots, i_k 行之后剩下的矩阵或向量。于是得到线性回归模型

$$Y(I) = \tilde{X}(I)\gamma + e(I) \quad (6.10)$$

从(6.10)得到 γ 的LS估计为

$$\begin{aligned} \hat{\gamma}(I) &= (\tilde{X}(I)' \tilde{X}(I))^{-1} \tilde{X}(I)' Y(I) \\ &= (\tilde{X}' \tilde{X} - \tilde{X}'_I \tilde{X}_I)^{-1} (\tilde{X}' Y - \tilde{X}'_I Y_I) \end{aligned} \quad (6.11)$$

这里 \tilde{X}_I 、 Y_I 表 \tilde{X} 、 Y 的第 i_1, \dots, i_k 行组成的矩阵或向量。应用公式: 设 A 、 D 可逆, 则

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B + D^{-1})^{-1}B'A^{-1} \quad (6.12)$$

(6.11)变为

$$\begin{aligned} \hat{\gamma}(I) &= [(\tilde{X}' \tilde{X})^{-1} + (\tilde{X}' \tilde{X})^{-1} \tilde{X}'_I (I - H_I)^{-1} \tilde{X}_I \\ &\quad (\tilde{X}' \tilde{X})^{-1}] [\tilde{X}' Y - \tilde{X}'_I Y_I] \\ &= \hat{\gamma} - (\tilde{X}' \tilde{X})^{-1} \tilde{X}'_I [- (I - H_I)^{-1} \tilde{X}_I \hat{\gamma} \\ &\quad + (I + (I - H_I)^{-1} H_I) Y_I] \end{aligned}$$

这里 $H_I = \tilde{X}_I (\tilde{X}' \tilde{X})^{-1} \tilde{X}'_I$. 利用 $(I - H_I)^{-1} = I + (I - H_I)^{-1} H_I$, 上式变为

$$\hat{\gamma}(I) = \hat{\gamma} - (\tilde{X}' \tilde{X})^{-1} \tilde{X}'_I (I - H_I)^{-1} \delta_I \quad (6.13)$$

这里 $\delta_I = (\delta_{i_1}, \dots, \delta_{i_k})$. 易见, (6.13) 是 (2.17) 的推广。

类似于 (6.3), 定义

$$D_I = D_I(\tilde{X}' \tilde{X}, (p+1)\hat{\sigma}^2)$$

$$= \frac{(\hat{\gamma}(I) - \hat{\gamma})'(\tilde{X}'\tilde{X})(\hat{\gamma}(I) - \hat{\gamma})}{(p+1)\hat{\sigma}^2} \quad (6.14)$$

它度量了第 i_1, \dots, i_k 组数据对LS估计影响的大小。将(6.13)代入(6.14)得到

$$D_I = \frac{\delta_I'(I - H_I)^{-1}H_I(I - H_I)^{-1}\delta_I}{(p+1)\hat{\sigma}^2} \quad (6.15)$$

除了应用距离(6.1)度量一组或多组数据对回归的影响大小之外,近年来许多作者还提出了另外一些度量影响的标准。例如, Cook和Weisberg^[24]提出的置信椭圆, Johnson和Geisser^[25]基于Bayes方法提出的Kullback-Leibler偏差, 以及 Andrews和Pregibon^[26]的另一种体积比等。目前, 这些方法还没有象距离(6.1)那样广泛地为人们所接受。想对这些内容作深入了解的读者可以去看上述文献。

§2.7 异常点

一组数据 (x_i, Y_i) , 如果它的残差 $(\delta_i$ 或 r_i 等)比其它组数据的残差大得多, 我们称 (x_i, Y_i) 为异常点, 有时也称 Y_i 为异常值。从公式(6.5), 我们已经看到, 异常点很可能是强影响点, 因此它有可能对回归的估计以及其它推断产生非同小可的影响。在回归分析的实际应用中, 如何探查异常点, 判断出哪些点是异常点之后应该如何处理, 是近年来倍受重视的问题。本节我们讨论探查异常点的一种检验。至于对异常点的处理问题, 原则上与强影响点一样。需要根据对数据本身以及数据的获得过程进行仔细的研究, 分别情况酌情处理。如果异常值产生于试验过程的失误, 则当然应该剔除。对于其它情况, 也可以剔除也可以用稳健估计

以缩小异常点的影响。无论哪一种情况，对异常点所对应的试验作进一步分析是必要的，往往是很有益处的。下面我们讨论异常点的检验问题。

我们把线性回归模型(1.1)改写为

$$Y_i = \tilde{x}_i' \gamma + e_i, \quad i=1, \dots, n \quad (7.1)$$

如果 (x_i, Y_i) 是一个异常点，那末它的残差之所以很大，是因为 $E(Y_i) \neq \tilde{x}_i' \gamma$ ，而是 $E(Y_i) = \tilde{x}_i' \gamma + \eta_i$ ，这里 η_i 是一个非随机量。

因此，若 Y_{i_1}, \dots, Y_{i_k} 对应于 k 个异常点，则线性回归模型为

$$Y_i = \begin{cases} \tilde{x}_i' \gamma + \eta_i + e_i, & i \in I = \{i_1, \dots, i_k\} \\ \tilde{x}_i' \gamma + e_i, & i \in \bar{I} \end{cases} \quad (7.2)$$

记 $\tilde{d}_i = (0, \dots, 0, 1, 0, \dots, 0)$ ，即第 i 个元素为1，其余 $n-1$ 个元素均为0的向量，又记 $D = (\tilde{d}_{i_1}, \dots, \tilde{d}_{i_k})$ ， $\eta' = (\eta_{i_1}, \dots, \eta_{i_k})$ ，则(7.2)为

$$Y = \tilde{X} \gamma + D \eta + e \quad (7.3)$$

这个模型可以看作将模型(1.1)对均值作平移后得到的，所以称为**均值平移线性回归模型**。要判定 (x_i, Y_i) ， $i \in I$ 都不是异常点，等价于检验线性假设 $H: \eta = 0$ 。

为了导出检验统计量，我们需要如下引理。

引理7.1 对模型(7.3)， γ 和 η 的LS估计分别为

$$\gamma^* = \hat{\gamma}(I) \quad (7.4)$$

和

$$\eta^* = (I - H_I)^{-1} \delta_I. \quad (7.5)$$

这里 $\hat{\gamma}(I)$ 由(6.13)定义， $H_I = \tilde{X}_I' (\tilde{X}_I' \tilde{X}_I)^{-1} \tilde{X}_I'$ ， δ_I 为残差向量 δ 的第 i_1, \dots, i_k 分量组成的子向量。 \tilde{X}_I 的定义类似于 δ_I 。

证明 因为 $\tilde{X}' D = \tilde{X}_I'$ ， $D' Y = Y_I$ ，所以

$$\begin{aligned} \begin{pmatrix} \gamma^* \\ \eta^* \end{pmatrix} &= \begin{pmatrix} \tilde{X}'\tilde{X} & \tilde{X}'D \\ D'\tilde{X} & D'D \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}'Y \\ D'Y \end{pmatrix} \\ &= \begin{pmatrix} \tilde{X}'\tilde{X} & \tilde{X}'_I \\ \tilde{X}'_I & I_k \end{pmatrix}^{-1} \begin{pmatrix} \tilde{X}'Y \\ Y_I \end{pmatrix} \end{aligned}$$

利用分块逆矩阵公式(2.7), 有

$$\begin{aligned} \begin{pmatrix} \gamma^* \\ \eta^* \end{pmatrix} &= \begin{pmatrix} (\tilde{X}'\tilde{X})^{-1} + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'_I(I-H_I)^{-1}\tilde{X}'_I(\tilde{X}'\tilde{X})^{-1} \\ - (I-H_I)^{-1}\tilde{X}'_I(\tilde{X}'\tilde{X})^{-1} \\ - (\tilde{X}'\tilde{X})^{-1}\tilde{X}'_I(I-H_I)^{-1} \\ (I-H_I)^{-1} \end{pmatrix} \begin{pmatrix} \tilde{X}'Y \\ Y_I \end{pmatrix} \\ &= \begin{pmatrix} \hat{\gamma} + (\tilde{X}'\tilde{X})^{-1}\tilde{X}'_I(I-H_I)^{-1}(\tilde{X}'_I\hat{\gamma} - Y_I) \\ (I-H_I)^{-1}(Y_I - \tilde{X}'_I\hat{\gamma}) \end{pmatrix} \\ &= \begin{pmatrix} \hat{\gamma} - (\tilde{X}'\tilde{X})^{-1}\tilde{X}'_I(I-H_I)^{-1}\delta_I \\ (I-H_I)^{-1}\delta_I \end{pmatrix} \\ &= \begin{pmatrix} \hat{\gamma} (I) \\ (I-H_I)^{-1}\delta_I \end{pmatrix} \end{aligned}$$

引理证毕.

利用这个引理, 对模型(7.3), 最小二乘分析之后残差平方和

$$RSS = Y'Y - (\gamma^{*'}, \eta^{*'}) \begin{pmatrix} \tilde{X}'Y \\ D'Y \end{pmatrix}$$

$$\begin{aligned} &= Y'Y - \hat{\gamma}'_{(I)} \tilde{X}'Y - \eta^{*'} D'Y \\ &= Y'Y - \hat{\gamma}' \tilde{X}'Y + \delta'_I (I-H_I)^{-1} \tilde{X}'_I \hat{\gamma} - \eta^{*'} Y_I \end{aligned}$$

而在假设 $H: \eta=0$ 下, 残差平方和 RSS_H 就等于模型(1.1)作 LS 估计后的残差平方和, 于是

$$RSS_H = Y'Y - \hat{\gamma}'\tilde{X}'Y$$

容易验证

$$RSS_H - RSS = \delta_I'(I - H_I)^{-1}\delta_I.$$

于是, 根据(1.3.3), 检验 $H: \eta=0$ 的 F -统计量为

$$\begin{aligned} \mathcal{F} &= \frac{\frac{1}{k}(RSS_H - RSS)}{\frac{1}{n-p-k-1}RSS} \\ &= \frac{n-p-k-1}{k} \cdot \frac{\delta_I'(I - H_I)^{-1}\delta_I}{(n-p-1)\hat{\sigma}^2 - \delta_I'(I - H_I)^{-1}\delta_I} \end{aligned}$$

当 H 为真时, 它服从 $F_{k, n-p-k-1}$. 于是我们证明了如下定理.

定理7.1 对模型(7.3), $\eta=0$ 的 F -检验统计量为

$$\mathcal{F} = \frac{(n-p-k-1)Q(I)}{k[(n-p-1)\hat{\sigma}^2 - Q(I)]} \quad (7.6)$$

当 $\eta=0$ 时, $\mathcal{F} \sim F_{k, n-p-k-1}$. 其中 $Q(I) = \delta_I'(I - H_I)^{-1}\delta_I$.

如果经检验, 假设 $\eta=0$ 被拒绝, 则推断在 k 组数据 (x_i', Y_i) , $j=i_1, \dots, i_k$ 中, 至少有一组是异常点. 特别对 $k=1$, 我们有下面的推论.

推论7.1.

$$\mathcal{F}_1 = \frac{(n-p-2)r_1^2}{n-p-1-r_1^2} \sim F_{1, n-p-2} \quad (7.7)$$

等价地

$$t_i = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{1/2} \sim t_{n-p-2} \quad (7.8)$$

这里 r_i 为学生化残差.

对给定的点 (x_i', Y_i) , 如果 $|t_i| > t_{n-p-2}\left(\frac{\alpha}{2}\right)$, 则以水平 α ,

我们拒绝原假设 $H: \eta_i = 0$, 即认为 (x_i', Y_i) 是一个异常点.

例7.1 人工降雨试验(续例3.1)

在这个试验中, 试验者预先对变量 x_4 有一个要求, 仅对 $x_4 \leq 13\%$ 的天进行试验. 但从表3.1我们看到, 在第二组数据中, $x_4 = 37.90\% > 13\%$. 这一天是一个“干扰天”. 据此先验信息, 我们来检验这组数据是否异常. 对应的 $t_2 = 1.60 < t_{13}(0.025) = 2.178$, 所以在水平 $\alpha = 0.05$, 我们接受原假设, 即第二组数据不是异常点.

异常点的检验是一个尚未很好解决的问题. 特别是数据中含多个异常点时更是如此. 因为此时, 会出现真正的异常点“伪装”成非异常点、而非异常点却又呈现出异常的复杂情况. 因此在应用上, 对从残差图上所看到的残差很大的点一定要谨慎处理. 关于异常点的研究, 目前仍比较活跃, 想对这一方向作深入了解的读者可参看[27]

第三章 自变量的选择

§3.1 引言

当我们应用回归分析去处理实际问题时，碰到的头一个重要问题就是选择回归自变量。一般说来，根据问题本身的专业理论及有关经验，人们罗列出来可能与因变量有关的自变量往往太多，其中有一些变量对因变量可能根本没有影响或影响很小。如果回归模型把这样一些变量都包含进来，不但计算量大，而且估计和预测的精度也会下降。在一些情况，某些自变量的观测数据的获得代价较大，如果这些自变量本身与因变量的关系很小或根本就没有关系，但错误地选进模型，会使模型应用的费用不必要的升高。正是由于这样一些原因，在应用回归分析时，对进入模型的自变量作精心选择是十分必要的。本章的目的就是对自变量选择作一些理论分析，提出一些变量选择准则，并介绍有关的计算方法。

回归自变量选择所涉及的计算量都很大。所以在本世纪六十年代以前，人们多局限于从理论上讨论剔除或添加一个自变量所引起的后果。随着高速电子计算机日益广泛的应用，这个方向得到了迅速发展，提出了许多变量选择准则、可供实用的计算方法和程序。就选择准则或标准而言，不少是从某种直观想法出发的。某些基于残差平方和的方法可以归入这一类；也有的是从某种目标出发，如：要求回归系数估计准确些，要求预测偏差的方

差小些等等。不同的标准导致了不同的选择方法，因而所选到的“最优”变量组(或称“最优”变量子集)也不必相同。关于计算方法的重要性是不言自明的。因为自变量选择的问题所涉及的计算量都很大，一个好的选择准则即便理论上有着相当的吸引力，但如果缺乏有效的计算方法也不能付之实用。在本世纪六十年代和七十年代，人们提出了很多有效的计算方法，不仅计算时间省，而且存储量也控制在一定的范围内，使许多变量选择准则得以见诸应用。

还需要指出，变量选择问题不能孤立于其它问题来考虑。例如，在上一章我们讨论了异常点、高杠杆点和强影响点。这些点对变量选择影响也很大。在下一章我们要讨论的自变量之间的复共线性也会对变量选择产生非同小可的影响。这其中的关系也很复杂，它们互相影响。另一个有关的问题是估计方法。在第四章读者将会看到，在回归系数的另外一些估计方法中，有一些特殊的变量选择方法。

关于变量选择，理论分析方面的文章比较少。例如，对每种选择准则进行统计理论上的分析，判别其用于何种情况，不同选择准则的比较等，目前知道得不多。某些企图作一些理论分析的文章，也往往加了一些从应用观点看来不太现实的条件。从实际方面看，这个问题也有其困难之处。变量选择本质上是一个模型选择问题。当讨论变量选择时，我们必须在一定模型下来考虑。例如，某问题中一切可能有关的自变量有50个，假定它们与因变量一起，适合一个线性回归模型。在这个条件下来考虑从50个变量中选择一部分的问题。在这里，重要的前提是“这50个自变量和因变量适合线性回归模型”这个假定。如果这个假定不成立，那末根据这个假定，依某种变量选择准则所选取的自变量子集，也就不具有任何优越性了。由于模型在一个具体问题中往往是一

件没有十足把握的假定，于是建立在其上的一套变量选择方法的效力也就没有坚实的基础。当然，变量选择，乃至更一般的模型选择基本上不能算是一个数学问题。尽管数学方法对模型的正确选择可能有一些帮助，但在处理一个具体问题时，模型的正确选择在根本上要依赖于所研究的问题本身的专业知识和实践经验。这一点很重要。当应用某种准则和方法选出了一个“最优”变量子集，明显地与实际问题本身的专业理论不一致时，需要首先重新考虑我们的统计结论，仔细从数据中是否含有异常点、复共线性、计算错误等方面找一找原因。那种把变量选择方法看成僵死的“教条”机械搬用，是不可取的。只有把它看作一种辅助工具，与实际问题本身的专业知识和实践经验相结合，才能取得好的实际效果。

§3.2 变量选择对估计、预测的影响

假定根据实际问题的专业知识和经验，初步认为一切可能与因变量有关系的变量共有 p 个，它们与因变量一起适合线性回归模型。在有了实际观测数据之后，我们有模型

$$Y = X\gamma + e, E(e) = 0, \text{COV}(e) = \sigma^2 I \quad (2.1)$$

这里 Y 为 $n \times 1$ 观测向量， X 为 $n \times (p+1)$ 的设计矩阵，这和前面两章记号略有不同。当模型含有常数项时，我们约定 X 的第1列的元素皆为1。为确定计，我们把常数项也视为自变量，这时，不论模型是否含常数项，在(2.1)中，自变量的个数总是 X 的列数。

将设计阵 X 写成分块形式 $X = (X_q : X_t)$ ，相应地将 γ 分块为 $\gamma' = (\gamma'_q : \gamma'_t)$ ，于是，(2.1)可改写为

$$Y = X_q \gamma_q + X_t \gamma_t + e, \quad (2.2)$$

这里，我们约定 X_q 中包含了常数项，且 X_q 和 X_t 分别有 q ， t 列，

$q+t=p+1$. 并且 X_p 和 X_t 都是列满秩矩阵, 即 $R(X_q)=q$, $R(X_t)=t$. 对 (2.2) 这种写法, 我们可以作双重解释:

1. 真实模型为 $Y=X\gamma+e$, 而我们误认为 $Y=X_q\gamma_q+e$, 这时错误地丢掉了一些自变量;

2. 真实模型为 $Y=X_q\gamma_q+e$, 而我们误认为 $Y=X\gamma+e$. 这时错误地把一些不必要的自变量引进了模型。

下面我们将分析这两种情况的后果, 以便对回归自变量选择的意义有更进一步的理解。

为方便计, 我们称 (2.1) 为**全模型**, 而称

$$Y=X_q\gamma_q+e \quad (2.3)$$

为**选模型**. 如在第一章所讨论的, 在全模型下, γ 和 σ^2 的 LS 估计分别为

$$\hat{\gamma}=(X'X)^{-1}X'Y, \quad (2.4)$$

$$\hat{\sigma}^2=Y'(I-X(X'X)^{-1}X')Y/(n-p-1) \quad (2.5)$$

而在选模型下, γ_q 和 σ^2 的 LS 估计为

$$\tilde{\gamma}_q=(X_q'X_q)^{-1}X_q'Y \quad (2.6)$$

$$\tilde{\sigma}_q^2=Y'(I-X_q(X_q'X_q)^{-1}X_q')Y/(n-q) \quad (2.7)$$

对 $\hat{\gamma}$ 作相应的分块: $\hat{\gamma}'=(\hat{\gamma}_q':\hat{\gamma}_t')$

定理 2.1 假设全模型 (2.1) 正确, 则

$$(1) E(\tilde{\gamma}_q)=\gamma_q+A\gamma_t, \text{ 这里 } A=(X_q'X_q)^{-1}X_q'X_t$$

$$(2) \text{COV}(\tilde{\gamma}_q) \geq \text{COV}(\hat{\gamma}_q)$$

其中, 对两个同阶方阵 A 与 B , 记号 $A \geq B$ 定义为 $A-B \geq 0$.

证明 (1) 从 (2.6) 易见

$$\begin{aligned} E(\tilde{\gamma}_q) &= (X_q'X_q)^{-1}X_q'E(Y) \\ &= (X_q'X_q)^{-1}X_q'(X_q:X_t)\begin{pmatrix} \gamma_q \\ \gamma_t \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
&= (I : A) \begin{pmatrix} \gamma_q \\ \gamma_t \end{pmatrix} \\
&= \gamma_q + A\gamma_t
\end{aligned}$$

于是(1)得证。

(2) 根据分块矩阵的逆矩阵公式(2.2.7), 有

$$\begin{aligned}
(X'X)^{-1} &= \begin{pmatrix} X_q'X_q & X_q'X_t \\ X_t'X_q & X_t'X_t \end{pmatrix}^{-1} \\
&= \begin{pmatrix} (X_q'X_q)^{-1} + (X_q'X_q)^{-1}X_q'X_tDX_t'X_q(X_q'X_q)^{-1} & \\ -DX_t'X_q(X_q'X_q)^{-1} & \\ - (X_q'X_q)^{-1}X_q'X_tD & \\ D & \end{pmatrix} \\
&= \begin{pmatrix} (X_q'X_q)^{-1} + ADA' & B \\ B' & D \end{pmatrix} \quad (2.8)
\end{aligned}$$

这里 $B = - (X_q'X_q)^{-1}X_q'X_tD$.

$$D^{-1} = X_t'X_t - X_t'X_q(X_q'X_q)^{-1}X_q'X_t. \quad (2.9)$$

由 $(X'X)^{-1} > 0$ 知 $D^{-1} > 0$. 又由

$$\text{COV}(\hat{\gamma}) = \text{COV} \begin{pmatrix} \hat{\gamma}_q \\ \hat{\gamma}_t \end{pmatrix} = \sigma^2 (X'X)^{-1}$$

推得 $\text{COV}(\hat{\gamma}_q) = \sigma^2 [(X_q'X_q)^{-1} + ADA']$, 但 $\text{COV}(\tilde{\gamma}_q) = \sigma^2 (X_q'X_q)^{-1}$, 所以

$$\text{COV}(\hat{\gamma}_q) - \text{COV}(\tilde{\gamma}_q) = \sigma^2 ADA'$$

由 $D^{-1} > 0$, 知 $D > 0$. 故 $ADA' \geq 0$. 这就证明了(2). 定理证毕.

从定理2.1的结论(1)我们看到, 要使 $\tilde{\gamma}_q$ 为 γ_q 的无偏估计, 必须 $A\gamma_t = 0$. 即 $\gamma_t = 0$ 或 $A = 0$ 至少一个成立. 前者表示后 t 个自变量与因变量没有什么关系, 即选模型本来就是正确的. 后者仅当 $X_q'X_t = 0$ 时成立. 这表示设计矩阵 $X = (X_q : X_t)$ 的两部分正交. 这时, 为了估计 γ_q , 后 t 个自变量不起任何作用。

定理2.1的(2)表明, 即使全模型(2.1)正确, 丢掉一部分自变量之后, 总使剩下的那部分自变量的回归系数LS估计的方差减小, 但此时的估计是有偏的, 总的效果如何就不一定了。

对于未知参数向量 θ 的有偏估计 $\tilde{\theta}$, 其协方差阵不能作为衡量估计精度之用, 更恰当的是平均平方误差矩阵 (Mean Square Error Matrix) 简称为均方误差矩阵, 记为MSEM, 定义为

$$\text{MSEM}(\tilde{\theta}) = E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)' \quad (2.10)$$

容易证明, 它和协方差阵有如下关系

$$\text{MSEM}(\tilde{\theta}) = \text{COV}(\tilde{\theta}) + (E\tilde{\theta} - \theta)(E\tilde{\theta} - \theta)' \quad (2.11)$$

事实上

$$\begin{aligned} \text{MSEM}(\tilde{\theta}) &= E(\tilde{\theta} - E\tilde{\theta} + E\tilde{\theta} - \theta)(\tilde{\theta} - E\tilde{\theta} \\ &\quad + E\tilde{\theta} - \theta)' \\ &= E(\tilde{\theta} - E\tilde{\theta})(\tilde{\theta} - E\tilde{\theta})' + (E\tilde{\theta} - \theta) \\ &\quad (E\tilde{\theta} - \theta)' + E(\tilde{\theta} - E\tilde{\theta})(E\tilde{\theta} - \theta)' \\ &\quad + E(E\tilde{\theta} - \theta)(\tilde{\theta} - E\tilde{\theta})' \\ &= E(\tilde{\theta} - E\tilde{\theta})(\tilde{\theta} - E\tilde{\theta})' + (E\tilde{\theta} - \theta) \\ &\quad (E\tilde{\theta} - \theta)' \end{aligned}$$

此即(2.11). 对于 $\tilde{\gamma}_q$, 应用(2.11)有

$$\text{MSEM}(\tilde{\gamma}_q) = \sigma^2(X_q'X_q)^{-1} + A\gamma_t\gamma_t'A' \quad (2.12)$$

注意到, $\hat{\gamma}_q$ 是 γ_q 的无偏估计, 所以

$$\begin{aligned} \text{COV}(\hat{\gamma}_q) &= \text{MSEM}(\hat{\gamma}_q) \\ &= \sigma^2(X_q'X_q)^{-1} + \sigma^2ADA' \end{aligned} \quad (2.13)$$

又因

$$\text{COV}(\hat{\gamma}_t) = \sigma^2D \quad (2.14)$$

比较(2.12)和(2.13), 并利用(2.14)得到下述定理.

定理2.2 假设全模型(2.1)正确, 则当 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t\gamma_t'$ 时

$$\text{COV}(\hat{\gamma}_q) \geq \text{MSEM}(\tilde{\gamma}_q) \quad (2.15)$$

这个定理表明：即使被丢掉的自变量的影响确实存在（即 $\gamma_i \neq 0$ ），但若 γ_i 难于准确估计（用 $\text{COV}(\hat{\gamma}_i) \geq \gamma_i \gamma_i'$ 来刻画）时，丢掉这样的变量还是有好处的。这可以使余下的那些自变量的回归系数的LS估计的精度提高。因此，对那些与因变量关系不是很大或难于掌握的自变量从模型中剔除是有利的。

定理2.3 假设全模型(2.1)正确，则对选模型导出的 $\hat{\sigma}^2$ 的估计 $\hat{\sigma}_q^2$ ，有

$$E(\tilde{\sigma}_q^2) = \sigma^2 + \gamma_i' D^{-1} \gamma_i / (n - q), \quad (2.16)$$

这里 D 由(2.9)所定义。

证明 对(2.7)应用公式(1.2.33)

$$\begin{aligned} E(\tilde{\sigma}_q^2) &= (X\gamma)'(I_n - X_q(X_q'X_q)^{-1}X_q')X\gamma / (n - q) \\ &\quad + \sigma^2 \text{tr}(I_n - X_q(X_q'X_q)^{-1}X_q') / (n - q) \end{aligned}$$

因为 $I_n - X_q(X_q'X_q)^{-1}X_q'$ 为幂等方阵，且

$$(I_n - X_q(X_q'X_q)^{-1}X_q')X = (0 : (I_n - X_q(X_q'X_q)^{-1}X_q')X_i)$$

所以

$$\begin{aligned} E(\tilde{\sigma}_q^2) &= \gamma_i' X_i' (I_n - X_q(X_q'X_q)^{-1}X_q') X_i \gamma_i / (n - q) \\ &\quad + \sigma^2 [n - \text{tr}(X_q(X_q'X_q)^{-1}X_q')] / (n - q) \\ &= \sigma^2 + \gamma_i' D^{-1} \gamma_i / (n - q) \end{aligned}$$

定得得证。

这个定理说明，丢掉了一些与因变量有关的变量之后，误差方差 σ^2 的估计 $\tilde{\sigma}_q^2$ 会偏高。直观上这也是显然的。因为丢掉的那些变量的影响进入了误差项，当然使得误差方差的估计偏大。

上面我们是从参数估计的角度考察了变量选择的后果。因为预测是回归分析的一个重要应用，所以我们有必要讨论变量选择对预测的影响。

假设我们要在点 $x' = (x'_q, x'_t)$ 处预测因变量的值 $y = x'\gamma + e$. 由第一章知, 在全模型下, 我们用 $\hat{y} = x'\hat{\gamma}$ 作为 y 的预测值, 此时预测偏差 $U = y - x'\hat{\gamma}$. 而在选模型(2.3)下, 用 $\tilde{y} = x'_q\tilde{\gamma}_q$ 预测 y , 预测偏差 $U_q = y - x'_q\tilde{\gamma}_q$. 若全模型(2.1)正确, 预测 \hat{y} 是无偏的, 即 $E(U) = 0$. 在 e 与 $\hat{\gamma}$ 不相关的条件下, U 的方差

$$\begin{aligned}\text{Var}(U) &= \sigma^2 + \text{Var}(x'\hat{\gamma}) \\ &= \sigma^2[1 + x'(X'X)^{-1}x]\end{aligned}\quad (2.17)$$

但对于 U_q , 利用定理2.1之(1), 有

$$E(U_q) = x'_q\gamma_t - x'_qA\gamma_t \quad (2.18)$$

只要 $\gamma_t \neq 0$, U_q 就是有偏预测. 和估计情形一样, 这时 U_q 的方差不能度量预测的优劣, 需要考虑**预测均方误差**(Mean Square Error of Prediction, 简记为MSEP)

$$\text{MSEP}(\tilde{y}) \triangleq E(U_q^2) = \text{Var}(U_q) + [E U_q]^2 \quad (2.19)$$

因为在 ϵ 与 e 不相关的假设下

$$\text{Var}(U_q) = \sigma^2[1 + x'_q(X'_qX_q)^{-1}x_q]. \quad (2.20)$$

再利用(2.18), 得到

$$\text{MSEP}(\tilde{y}) = \sigma^2[1 + x'_q(X'_qX_q)^{-1}x_q] + (x'_qA\gamma_t - x'_q\gamma_t)^2 \quad (2.21)$$

利用这些事实能够证明下面的结果.

定理2.4 若全模型(2.1)正确, 则

$$(1) \quad E(U_q) = x'_q\gamma_t - x'_qA\gamma_t,$$

$$(2) \quad \text{Var}(U) \geq \text{Var}(U_q)$$

其中 $A = (X'_qX_q)^{-1}X_q'X_t$.

证明 (1)已证. 现证(2). 利用公式(2.8)

$$\begin{aligned}& [\text{Var}(U) - \text{Var}(U_q)]/\sigma^2 \\ &= x' \begin{pmatrix} (X'_tX_t)^{-1} + ADA' & B \\ B' & D \end{pmatrix} x - x'_q(X'_qX_q)^{-1}x_q\end{aligned}$$

$$\begin{aligned}
&= x_q' ADA' x_q + 2x_q' Bx_t + x_t' Dx_t \\
&= (A' x_q - x_t)' D (A' x_q - x_t) \geq 0
\end{aligned} \tag{2.22}$$

证毕.

这个定理表明, 当全模型(2.1)正确时, 用选模型所作的预测 $\tilde{y} = x_q' \tilde{\gamma}_q$ 一般是有偏的, 但这时预测偏差的方差下降。总的预测效果如何, 需要比较MSEP.

定理2.5 若全模型(2.1)正确, 则当 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t \gamma_t'$ 时, 有 $\text{Var}(U) \geq \text{MSEP}(\tilde{y}) = E(U_q^2)$

证明: 根据假设条件及(2.18)

$$\begin{aligned}
(EU_q)^2 &= (x_q' A \gamma_t - x_t' \gamma_t)^2 \\
&= (x_q' A - x_t') \gamma_t \gamma_t' (A' x_q - x_t) \\
&\leq (A' x_q - x_t)' \text{COV}(\hat{\gamma}_t) (A' x_q - x_t)
\end{aligned}$$

因为

$$\text{COV}(\hat{\gamma}_t) = \sigma^2 D$$

并利用(2.22), 得

$$(EU_q)^2 \leq \text{Var}(U) - \text{Var}(U_q).$$

从而有

$$\begin{aligned}
\text{Var}(U) &\geq \text{Var}(U_q) + (EU_q)^2 \\
&= E(U_q^2) \\
&= \text{MSEP}(\tilde{y})
\end{aligned}$$

定理得证.

这个定理的意义与定理2.2相似: 即使全模型正确, 但如果有一些自变量的影响很小或难以估计(用 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t \gamma_t'$ 来刻画), 则丢掉这些变量之后可以使预测精度提高。显然, 如果模型包含了不必要的自变量($\gamma_t = 0$), 它总使预测精度降低。此因 $\gamma_t = 0$ 时, 条件 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t \gamma_t' = 0$ 成立。

仔细考虑条件 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t \gamma_t'$, 还可以看出, 因 $\text{COV}(\hat{\gamma}_t)$ 与

σ^2 成正比, 所以 σ^2 愈大, 即试验误差愈大, 模型中可丢弃的自变量也就愈多, 因为这时上述条件更容易满足些。这一点从直观上也不难理解。因为误差愈大, 愈不宜保留过多的自变量。

把上面的讨论归纳起来, 我们得到如下一般性结论: 在选择回归自变量时, 基本原则应当是少而精。丢掉一些变量总会使剩余变量回归系数的LS'估计和预测的方差减少。但这时的估计和预测一般是有偏的。如果丢掉的那些变量确实影响比较小, 则剩余变量的回归系数LS'估计和预测的均方误差减少。本着这样的理由, 那些可有可无的自变量应当尽量剔除掉。

在假设检验中, 对假设 $H: \gamma_i = 0$, 利用 F -统计量

$$F = \frac{\hat{\gamma}_i' D^{-1} \hat{\gamma}_i}{t \hat{\sigma}^2} \quad (2.23)$$

作检验, 这本身也是一种选择变量的方法。这里 $\text{COV}(\hat{\gamma}_i) = \sigma^2 D$ 。现在我们把它和定理2.2和定理2.5作一比较。为此我们需要如下代数事实。

引理2.1 设 $A_{n \times n} > 0$, x 为 $n \times 1$ 向量, $a > 0$ 为数, 则

$$aA \geq xx' \iff x' A^{-1} x \leq a.$$

证明 必要性的证明比较容易。在 $aA \leq xx'$ 的两边都左乘 $x' A^{-1}$ 、右乘 $A^{-1}x$, 即得 $x' A^{-1} x \leq a$ 。至于充分性, 依Cauchy-Schwarz不等式

$$(x'u)^2 \leq x' A^{-1} x \cdot u' A u, \quad \text{对一切 } u$$

及假设 $x' A^{-1} x \leq a$, 得

$$(x'u)^2 \leq a u' A u, \quad \text{对一切 } u$$

此即

$$u'(aA)u \geq (x'u)^2 = u'(xx')u, \quad \text{对一切 } u$$

由 u 的任意性, 这就证明了 $aA \geq xx'$. 证毕。

在定理2·2和定理2·5中, 后 t 个自变量丢掉可以提高估计和预测精度的条件为 $\text{COV}(\hat{\gamma}_t) \geq \gamma_t \gamma_t'$, 即

$$\sigma^2 D \geq \gamma_t \gamma_t'.$$

利用引理2·1, 上式等价于

$$\gamma_t' D^{-1} \gamma_t \leq \sigma^2. \quad (2.34)$$

将 γ 和 σ^2 用LS估计代替, 得

$$F_1 = \frac{\hat{\gamma}_t' D^{-1} \hat{\gamma}_t}{t \hat{\sigma}^2} \leq \frac{1}{t} \quad (2.35)$$

此式说明, 定理2·2和定理2·5所给出的丢掉后 t 个变量的条件, 相当于按法则(2·35)决定变量的取舍。但是, 在通常的5%, 10%等显著水平, 所得到的临界值比(2·35)给出的 $1/t$ 大, 所以, 与从定理2·2和定理2·5导出的法则(2·35)相比, 通常的 F 检验倾向于剔除较多的自变量。

一般说来, 从不同的目的出发, 所建立的剔除变量的法则也就不同。例如, 如果我们的兴趣只是在于对试验值所在区域进行预测, 那末下面的条件看来比较合理:

$$\sum_{i=1}^n [\text{Var}(U_i) - E(U_{qi}^2)] \geq 0 \quad (2.36)$$

这里 $U_i = y_i - x_i' \hat{\gamma}$ 和 $U_{qi} = y_i - x'_{qi} \tilde{\gamma}_q$ 分别为对第 i 个试验点全模型和选模型(2·3)的预测偏差, 这里

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x'_{q1} & x'_{t1} \\ \vdots & \vdots \\ x'_{qn} & x'_{tn} \end{pmatrix} = (X_q : X_t)$$

下面从(2·36)出发, 导出另外一种变量取舍的条件, 我们先证明

$$\sum_{i=1}^n \text{Var}(U_i) = \sigma^2(n + p + 1) \quad (2.37)$$

$$\sum_{i=1}^n E(U_{qi}^2) = \sigma^2(n+q) + \gamma_i' D^{-1} \gamma_i \quad (2.38)$$

这里 q 为 X_q 的列数, D 由 (2.9) 定义, 即

$$D^{-1} = X_i' X_i - X_i' X_q (X_q' X_q)^{-1} X_q' X_i. \quad (2.39)$$

利用 (2.17), 有

$$\begin{aligned} \sum_{i=1}^n \text{Var}(U_i) &= \sigma^2 \left(n + \sum_{i=1}^n x_i' (X' X)^{-1} x_i \right) \\ &= \sigma^2 \left[n + \sum_{i=1}^n \text{tr}(x_i' (X' X)^{-1} x_i) \right] \\ &= \sigma^2 \left[n + \sum_{i=1}^n \text{tr}((X' X)^{-1} x_i x_i') \right] \\ &= \sigma^2 \left[n + \text{tr}((X' X)^{-1} \sum_{i=1}^n x_i x_i') \right] \\ &= \sigma^2(n + \text{tr}(I_{p+1})) \\ &= \sigma^2(n + p + 1) \end{aligned}$$

于是 (2.37) 得证.

再证 (2.38). 因为

$$E(U_{qi}^2) = \text{Var}(U_{qi}) + (EU_{qi})^2. \quad (2.40)$$

利用 (2.20), 采用与 (2.37) 同样证法, 可证明

$$\sum_{i=1}^n \text{Var}(U_{qi}) = \sigma^2(n+q) \quad (2.41)$$

再由定理 2.4 之 (1), 有

$$\begin{aligned} \sum_{i=1}^n (EU_{qi})^2 &= \sum_{i=1}^n (x_{ii}' \gamma_i - x_{qi}' A \gamma_i)^2 \\ &= \sum_{i=1}^n \gamma_i' (A' x_{qi} - x_{ii}) (x_{qi}' A - x_{ii}') \gamma_i \end{aligned}$$

将 $A = (X_q' X_q)^{-1} X_q' X_i$ 代入, 化简得

$$\sum_{i=1}^n (EU_{qi})^2 = \gamma_i' D^{-1} \gamma_i \quad (2.42)$$

由 (2.40)、(2.41) 和 (2.42), 即得 (2.38).

将(2·37)和(2·38)代入(2·36), 注意到 $t = p + 1 - q$, 得

$$\sigma^2 t - \gamma_i' D^{-1} \gamma_i \geq 0,$$

即

$$\frac{\gamma_i' D^{-1} \gamma_i}{t \sigma^2} \leq 1$$

将 γ_i 、 σ^2 用LS估计代替, 得到

$$F_1 = \frac{\hat{\gamma}_i' D^{-1} \hat{\gamma}_i}{t \hat{\sigma}^2} \leq 1 \quad (2.43)$$

这就是说, 按照条件(2·36)取舍变量, 相当于依法则(2·43)取舍变量。比较(2·43)和(2·35), 前者丢弃的变量要多一些。这个结果说明, 如果所建立的回归模型将用于原试验区域内的预测, 则丢弃的变量比用于一般目的的估计或预测情形要多一些。

§3.3 基于残差平方和的准则

上节我们从丢掉一组自变量对剩下自变量的回归系数LS估计和因变量预测的影响, 对自变量选择提出了若干一般性考虑。把这些考虑概括起来, 就是所谓少而精原则。但是法则(2·35)或(2·43), 都难以付诸实用。这主要是因为, 我们已选出一个特定的自变量子集时, 用(2·35)或(2·43)固然可以判断该子集是否该丢掉, 但在不该丢掉的情况下, 并不等于说这个子集中每个自变量都该保留下来。因此在这种情况下, 结论包含了某种程度的不确定性。并且, 在实际应用上往往更重要的问题, 不在于判定一个特别选出的自变量子集是否该去掉, 而在于对各种不同子集的比较, 进而从全部可能的变量子集中挑出一组“最优”的, 这时就需要使用对这种问题能作出确定回答的准则。

在实用上，从数据与模型拟合优劣的直观考虑出发，基于残差平方和RSS的变量选择准则使用得最多。因此，本节首先介绍三个这种类型的准则。

在选模型(2.3)下，残差平方和RSS为

$$RSS_q = \|Y - X_q \tilde{\gamma}_q\|^2 = Y'(I - H_q)Y \quad (3.1)$$

这里 $H_q = X_q(X_q'X_q)^{-1}X_q'$ 。如果在选模型(2.3)中再增加一个变量，设对应的设计阵为 $X_{q+1} = (X_q : b)$ ，则残差平方和为

$$RSS_{q+1} = Y'(I - H_{q+1})Y, \quad (3.2)$$

其中 $H_{q+1} = X_{q+1}(X_{q+1}'X_{q+1})^{-1}X_{q+1}'$ 。利用分块矩阵的逆矩阵公式(2.2.7)，不难证明 $H_{q+1} \geq H_q$ 。由此可得

$$RSS_{q+1} \leq RSS_q. \quad (3.3)$$

即当自变量子集在扩大时，残差平方和随之减少。因此，如果按“ RSS_q 愈小愈好”的原则来选择自变量子集，则毫无疑问应该选全部自变量。所以，“ RSS_q 愈小愈好”不能作为一个选择自变量的法则。

问题是，我们应该设法防止选取过多的自变量。为此，有的作者提出事先给定变量个数的最大值，尔后选择使 RSS_q 达到最小的子集。这个方法的困难之处在于变量个数的限定值往往难于事先给出。另外一种常见的作法是，在残差平方和RSS上添加对增加变量的惩罚因子。

(一)平均残差平方和(Residual Mean Squares, RMS_q)

对选模型(2.3)，平均残差平方和为

$$RMS_q = \frac{RSS_q}{n - q} \quad (3.4)$$

这里 q 为选模型(2.3)设计阵 X_q 的列数。根据我们在上节的约定，如果有常数项的话，它一定含在 X_q 中，并也把它视为一个自变量。

所以无论模型是否含有常数项， q 总被认为是自变量的个数。根据(2·7)知， RMS_q 实际上就是从选模型(2·3)给出的误差方差 σ^2 的估计 $\tilde{\sigma}_q^2$ 。在(3·4)中，因子 $(n-q)^{-1}$ 随着自变量个数 q 的增加而增加，它体现了对变量个数的增加所施加的惩罚。 RMS_q 的图形大致如图3·3·1。由(3·3)知， RSS_q 随着 q 的增加而减少，所以，一开始当 q 增加时， RMS_q 先是减少，而后稳定下来，最后又增加。之所以会如此，是因为刚开始时，随着自变量个数的增加，虽然因子 $(n-q)^{-1}$ 增大了，但此时 RSS_q 减少很多，故总起来 RMS_q 还是减少。当自变量增加到一定程度，重要自变量基本上都已选上了，此时再增加自变量， RSS_q 减少不多，以致于抵消不了 $(n-q)^{-1}$ 的增加，最终还是导致了 RMS_q 的增加。依 RMS_q 准则，按“ RMS_q 愈小愈好”选择自变量子集。

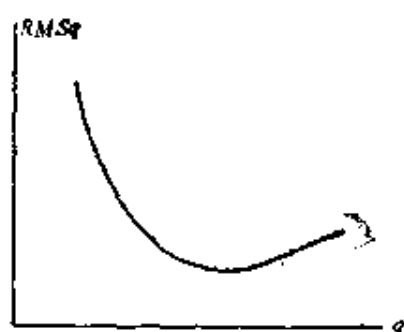


图3·3·1

也有一些作者提出修正相复关系数准则：按照“修正复相关系数之平方

$$\bar{R}_q = 1 - \frac{n-1}{n-q}(1 - R_q) \quad (3.5)$$

愈大愈好”的原则选择子集，这里

$$R_q = 1 - \frac{RSS_q}{TSS}$$

$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. $\sqrt{\bar{R}_q}$ 为 q 个自变量对 Y 的复相关系数。(参见第34页)因为

$$\begin{aligned}\bar{R}_q &= 1 - \frac{n-1}{n-q} \cdot \frac{RSS_q}{TSS} \\ &= 1 - \frac{n-1}{TSS} RMS_q\end{aligned}$$

所以, 这个准则等价于 RMS_q 准则.

(二) 预测偏差的方差 $(n+q)RMS_q$

这个准则是从预测角度提出的. 记 $X'_q = (x_{q1}, \dots, x_{qn})$. 考虑选模型(2.3)下, 在 n 个试验点 x_{qi} , $i=1, \dots, n$ 预测的偏差 $U_{qi} = y_i - x'_{qi} \tilde{\gamma}_q$ 的方差之和. 由(2.41)知

$$\sum_{i=1}^n \text{Var}(U_{qi}) = \sigma^2(n+q)$$

因为 RMS_q 是从选模型得到的 σ^2 的估计, 用它代替上式的 σ^2 , 得到 $(n+q)RMS_q$. 这个准则是按照 “ $(n+q)RMS_q$ 愈小愈好” 选择自变量子集. 因为

$$(n+q)RMS_q = \frac{n+q}{n-q} RSS_q$$

可见, 它的惩罚因子为 $(n+q)/(n-q)$. 易见, 这个准则对变量个数增加的惩罚要比 RMS_q 准则更严厉些.

需要指出的是, 这个准则只考虑了预测方差, 而未能兼顾预测偏差. 对于选模型, 往往预测是有偏的, 所以, 预测均方误差才是预测精度的全面度量.

(三) 平均预测均方误差 S_q

虽然这个准则也是从预测角度提出的, 但考虑问题的方法与

上段略有不同。 S_q 定义为

$$S_q = \frac{1}{n-q-1} \text{RMS}_q. \quad (3.6)$$

这个准则是按“ S_q 愈小愈好”的原则选择自变量子集。下面我们给出 S_q 的详细推导。因为其中应用了一些多元分析的知识，兴趣在应用方面的读者，这一段可以略去不读。

将选模型(2.3)写成分量形式，注意到 $\gamma_i = (\alpha, \beta_1, \dots, \beta_{q-1})$ ，得

$$Y_i = \alpha + \beta_1 x_{1i} + \dots + \beta_{q-1} x_{q-1i} + e_i, \quad i=1, \dots, n, \quad (3.7)$$

记 $\bar{x} = (\bar{x}_1, \dots, \bar{x}_{q-1})'$, $\bar{x}_i = \sum_j x_{ji}/n$. 将(3.7)中心化，得

$$Y_i = \alpha_0 + \beta_1 (x_{1i} - \bar{x}_1) + \dots + \beta_{q-1} (x_{q-1i} - \bar{x}_{q-1}) + e_i, \quad i=1, \dots, n, \quad (3.8)$$

其中 $\alpha_0 = \alpha + \beta_1 \bar{x}_1 + \dots + \beta_{q-1} \bar{x}_{q-1}$. 暂记

$$X = \begin{bmatrix} x_{11} & \dots & x_{q-11} \\ \vdots & & \vdots \\ x_{1n} & & x_{q-1n} \end{bmatrix}, \quad \bar{X} = \begin{bmatrix} \bar{x}_1 & \dots & \bar{x}_{q-1} \\ \vdots & & \vdots \\ \bar{x}_1 & \dots & \bar{x}_{q-1} \end{bmatrix} \quad (3.9)$$

由第一章知，对中心化模型(3.8)， α_0 和 $\beta' = (\beta_1, \dots, \beta_{q-1})$ 的LS估计为

$$\hat{\alpha}_0 = \bar{Y}, \quad \hat{\beta} = S^{-1}(X - \bar{X})'Y \quad (3.10)$$

其中 $S = (X - \bar{X})'(X - \bar{X})$. 在任一点 $x' = (x_1, \dots, x_{q-1})$ 处，因变量的预测值为

$$\hat{y} = \hat{\alpha}_0 + (x - \bar{x})' \hat{\beta} \quad (3.11)$$

预测偏差 $U = y - \hat{y}$ 的方差为

$$\text{Var}(U) = \sigma^2 \left[1 + \frac{1}{n} + (x - \bar{x})' S^{-1} (x - \bar{x}) \right] \quad (3.12)$$

现在假设 $q-1$ 个自变量 X_1, \dots, X_{q-1} 与因变量 Y 服从 q 维正态分布 $N(\mu, V)$ ，把 $(x_{1i}, \dots, x_{q-1i}, Y_i)$, $i=1, \dots, n$ 看作从此

总体抽出的大小为 n 的随机样本. 又 $x' = (x_1, \dots, x_{q-1})$ 为从 X_1, \dots, X_{q-1} 的边缘分布中抽取的独立样本, 现欲预测对应的 y . 在给定 X, x 的条件下, 预测偏差 U 的方差为(3.12). 我们可以证明

$$E(\text{Var}(U|X, x)) = \frac{n+1}{n} \cdot \frac{n-2}{n-q-1} \sigma^2 \quad (3.13)$$

这里 σ^2 由 V 决定. 若记

$$V = \begin{pmatrix} V_1 & v \\ v' & v_{qq} \end{pmatrix}$$

由(1.5.2), 知 $\sigma^2 = v_{qq} - v'V_1^{-1}v$.

现在证明(3.13). 由多元分析知(见[14]. P.138)

$$S \sim W_{q-1}(n-1, V),$$

这里 $W(\cdot, \cdot)$ 表Wishart分布. 又从正态分布的性质, 知

$$x - \bar{x} \sim N\left(0, \left(1 + \frac{1}{n}\right)V\right)$$

故统计量

$$T^2 = \frac{n(n-1)}{n+1} (x - \bar{x})' S^{-1} (x - \bar{x})$$

为Hotelling统计量. 因此(见[14]P.143)

$$\frac{n-q+1}{q-1} \cdot \frac{T^2}{n-1} = \frac{n-q+1}{q-1} \cdot \frac{n}{n+1} (x - \bar{x})' S^{-1} (x - \bar{x})$$

$$(x - \bar{x}) \sim F_{q-1, n-q+1}$$

利用事实: $u \sim F_{m, n}$, $E(u) = n/(n-2)$, 从(3.12)得

$$E(\text{Var}(U|X, x)) = \sigma^2 \frac{(n+1)(n-2)}{n(n-q-1)} \quad (3.14)$$

(3.13)得证.

在(3.13)中, 略去与 q 无关的因子, 并用 RMS_q 代替 σ^2 , 即得(3.6).

下面我们讨论两个实例、这些例子在回归分析的文献中常被用来讨论有关自变量选择、计算方法以及有偏估计等问题。

例3.1 汽车油耗问题

为了研究汽车耗油量与汽车各种结构参数之间的关系，R. J. Freund 考虑了可能影响耗油量的如下十个自变量：

x_1 = 发动机的形状，分两种， x 分别取为 1 和 0 .

x_2 = 汽缸个数.

x_3 = 传动类型，分手动($x_3 = 1$)、自动($x_3 = 0$)两种.

x_4 = 传动速度的种数，

x_5 = 发动机的体积，

x_6 = 发动机的功率，

x_7 = 汽化器的个数，

x_8 = 末级传动比，

x_9 = 汽车重量，

x_{10} = 行驶1/4哩所用时间，

因变量 Y 为每哩耗油量。共收集了 $n = 32$ 组数据。有关结果最早发表在美国杂志“Motor trend”上。

在进一步讨论之前，我们先对回归自变量子集引进秩的概念。一个回归问题，如果共有 p 个自变量，对给定的 q ， $1 \leq q \leq p+1$ ，包含 $q-1$ 个自变量的子集共有 $k = C_{p+1}^{q-1}$ 个。设有了某种变量选择准则，它归结为对每个自变量子集去计算一定的统计量 T 的值。按 T 值大小将这 k 个自变量子集排序。如果准则是 T 值愈小愈好，则我们把这 k 个自变量子集中具有最小 T 值那个子集称为秩为 1，具有次最小的称为秩为 2，余类推。秩反映了在该准则下，这个自变量子集的优劣。

图3.3.2是汽车油耗问题的 RMS_q 和 S_q 图。对 $q = 1, 2, \dots, 11$ ，画出了秩为 1 的自变量子集的 RMS_q 和 S_q 值。对于 RMS_q 准则， $q = 6$ ，即含 5 个自变量的子集具有最小的 RMS_q 。计算结果表明，它们是 $x_1, x_5, x_6, x_9, x_{10}$ （见表3.1）。而对 S_q 准则， $q = 4$ 具有最小的 S_q 。它们是自变量子集 x_3, x_9, x_{10} 。对这两组自变量子集，对应的选模型回归系数LS估计列在表3.1

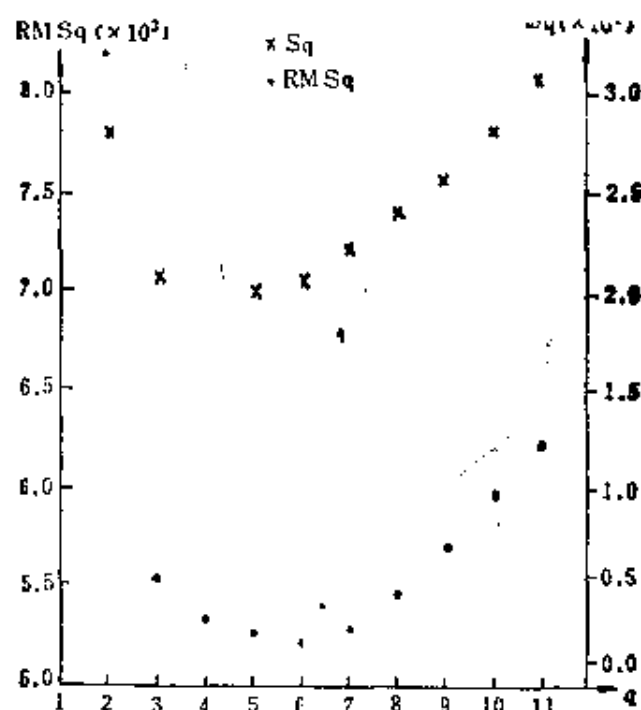


图3.3.2 汽车油耗问题 RMS_q 和 S_q 图

表3.1 汽车油耗问题两个最优子集LS估计($\times 10$)

变 量 q	3	5	6	9	10
4	2.43	—	—	-6.36	3.63
6	2.87	0.231	-2.41	-6.63	2.99

例3.2 Hald水泥问题

这组数据取自Hald, A.的书“Statistical Theory with Engineering Applications(1952), P.647.”它也是回归分析文献中经常引用的经典例子。问题是考察含如下四种化学成份

$x_1 = 3\text{CaO} \cdot \text{Al}_2\text{O}_3$ 的含量(%),

$x_2 = 3\text{CaO} \cdot \text{SiO}_2$ 的含量(%),

$x_3 = 4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_3$ 的含量(%),

$x_4 = 2\text{CaO} \cdot \text{SiO}_2$ 的含量(%),

表3.2 Hald数据

序 号	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

的某种水泥，每一克所释放出的热量（卡路里）Y 与这四种成份含量之间的关系。数据共 $n=13$ 组，列在表3.2.

这个问题有四个自变量. 含1个自变量的子集有 C_4^1 个, 含两个的有 C_4^2 个, 含三个的有 C_4^3 个, 含所有自变量的有 $C_4^4=1$ 个, 把仅含常数项的也算上一个, 共有 $C_4^1 + C_4^2 + C_4^3 + 1 = 2^4 = 16$ 个. 这16个变量子集的LS估计和 RMS_q 列在表3.3. 对 $q=1, 2, 3, 4, 5$, 图3.3.3画出了秩为1的自变量子集的 RMS_q . 从表或图可以看出, RMS_q 在 (x_1, x_2, x_4) 达到最小值5.3303. 另外, 子集 (x_1, x_2) 的 $RMS_q = 5.7904$ 也比较小. 如要选两个自变量的回归, 应该选 (x_1, x_2) . 于是, 对于Hald水泥问题, 含两个、三个自变量的在 RMS_q 准则下“最优”子集回归分别为

$$Y \approx 52.577 + 1.468x_1 + 0.662x_2$$

和

$$\hat{Y} = 71.648 + 1.452x_1 + 0.416x_2 - 0.237x_4.$$

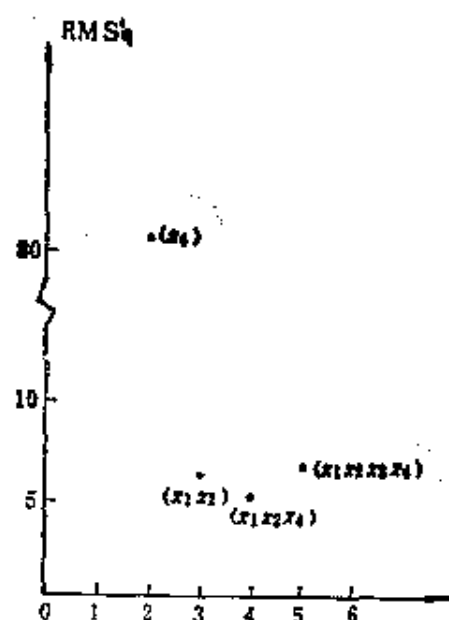


图3-3-3 Hald水泥问题RMS_i图

表3-3 Hald水泥问题的LS估计

模型中的 自变量	α	β_1	β_2	β_3	β_4	RMS _i
x_1	81.479	1.869				115.0624
x_2	57.424		0.789			82.3942
x_3	110.203			-1.256		176.3029
x_4	117.568				-0.738	80.8515
x_1x_2	52.577	1.468	0.662			5.7904
x_1x_3	72.349	2.312		0.494		122.7073
x_1x_4	103.097	1.440			-0.614	7.4762
x_2x_3	72.075		0.731	-1.008		41.5443
x_2x_4	94.160		0.311		-0.457	86.8880
x_3x_4	131.282			-1.200	-0.724	17.5738
$x_1x_2x_3$	48.194	1.696	0.657	0.250		5.3456

(续上表)

$x_1 x_2 x_4$	71.648	1.452	0.416		-0.237	5.3303
$x_2 x_3 x_4$	203.642		-0.923	-1.448	-1.557	5.6485
$x_1 x_3 x_4$	111.684	1.052		-0.410	-0.643	8.2017
$x_1 x_2 x_3 x_4$	62.405	1.551	0.510	0.102	-0.144	5.9329

§3.4 C_p -准则

近年来, 愈来愈得到广泛重视的一种变量选择准则是基于 C.L. Mallows^[28, 29]的 C_p 统计量. 和上节讨论的 $(n+p)$ RMS_q 和 S_q 准则一样, C_p 统计量也是从预测观点出发, 基于残差平方和的一个准则. 鉴于它的重要性及内容上的独特性, 我们把它单独列为一节, 以便进行较为详细的讨论.

(一) C_p 统计量的定义

对于选模型(2·3), C_p 统计量定义为

$$C_p = \frac{\text{RSS}_q}{\hat{\sigma}^2} - (n - 2q) \quad (4.1)$$

和前面一样, RSS_q 为在模型(2·3)下的残差平方和, $\hat{\sigma}^2$ 为从全模型(2·1)导出的 σ^2 的估计, q 是选模型(2·3)的设计阵 X_q 的列数, 即包括常数项在内的自变量个数. 按照 C_p 的本来意义, 下标 p 表示选模型的自变量个数(包括常数项在内). 因为我们前面一直把 p 记作全模型的自变量个数, 所以在(4·1)中, C_p 的下标 p 无任何意义. 好在随着 C_p 统计量的日益广泛应用, 人们已经把 C_p 作为一个整体, 用作一个专门术语了. 鉴于这一点, 虽然(4·1)右边用的是 q , 但已没有必要把左边的 C_p 改为 C_q 了.

如果采用选模型(2·3), 那末在任一点 $x' = (x'_q, x'_i)$ 处, 我们用 $\tilde{y} = x'_q \tilde{\gamma}_q$ 预测 $y = x' \gamma + \varepsilon$ 。此时量

$$d = E(\tilde{y} - E(y))^2 \quad (4.2)$$

度量了这种预测的优劣。对 d 作分解

$$\begin{aligned} d &= E(x'_q \tilde{\gamma}_q - E(x'_q \tilde{\gamma}_q) + E(x'_q \tilde{\gamma}_q) - E(y))^2 \\ &= E((x'_q \tilde{\gamma}_q - E(x'_q \tilde{\gamma}_q))^2 + (E(x'_q \tilde{\gamma}_q) - y)^2) \\ &= \text{Var}(x'_q \tilde{\gamma}_q) + (EU_q)^2 \end{aligned} \quad (4.3)$$

其中第一项为预测方差, 第二项中 $U_q = x'_q \tilde{\gamma}_q - y$, $EU_q = x'_i \gamma_i - x'_q A \gamma_i$, 这里 $A = (X'_q X_q)^{-1} X'_q X_i$ 。因此第二项度量了预测偏差。对 n 个试验点 $x' = (x'_{qi}, x'_{ii})$, $i = 1, \dots, n$, 计算出

$$d_i = \text{Var}(x'_{qi} \tilde{\gamma}_q) + (x'_{ii} \gamma_i - x'_{qi} A \gamma_i)^2, \quad i = 1, \dots, n$$

并对 i 求和, 除以 σ^2 , 得到

$$\Gamma_q = \sum_{i=1}^n d_i / \sigma^2 = \sum_{i=1}^n x'_{qi} (X'_q X_q)^{-1} x_{qi} + \sum_{i=1}^n (x'_{ii} \gamma_i - x'_{qi} A \gamma_i)^2 / \sigma^2$$

与(2·37)的证明相类似, 可证得上式第一项为 q 。再利用(2·42), 第二项则为 $\gamma'_i D^{-1} \gamma_i / \sigma^2$ 。 D 由(2·9)所定义, 所以

$$\Gamma_q = q + \frac{\gamma'_i D^{-1} \gamma_i}{\sigma^2} \quad (4.4)$$

Γ_q 是采用选模型(2·3)时, 在 n 个试验点预测优劣的一个总度量, 它反映了选模型(2·3)的好坏。但是(4·4)中包含有未知参数, 因此还不能直接用作自变量选择的标准。从(2·7)知, 在选模型(2·3)下, 残差平方和

$$\text{RSS}_q = (n - q) \tilde{\sigma}_q^2$$

再利用(2·16), 得到

$$\frac{\gamma'_i D^{-1} \gamma_i}{\sigma^2} = \frac{E(\text{RSS}_q)}{\sigma^2} + q - n \quad (4.5)$$

所以, Γ_q 可以改写为

$$\Gamma_q = \frac{E(\text{RSS}_q)}{\sigma^2} - (n - 2q) \quad (4.6)$$

在上式中, 用 RSS_q 代替 $E(\text{RSS}_q)$, 以 σ^2 在全模型的估计 $\hat{\sigma}^2$ 代替 σ^2 , 便得到 (4.1). 可见, C_p 统计量是作为 Γ_q 的一种估计产生的。

(二) C_p 统计量的性质

本段证明 C_p 统计量的一些性质。它们对于应用 C_p 统计量作自变量选择, 提供了理论基础。

定理 4.1 若全模型 (2.1) 误差 $e \sim N(0, \sigma^2 I)$, 则对选模型 (2.3) 的 C_p , 有

$$E(C_p) = q - t + \frac{n - p - 1}{n - p - 3} \left(t + \frac{\gamma'_t D^{-1} \gamma_t}{\sigma^2} \right) \quad (4.7)$$

这里 D 由 (2.9) 定义。

证明 问题归结为计算 $E(\text{RSS}_q / \hat{\sigma}^2)$. 对全模型 (2.1), 残差平方和 $\text{RSS} = (n - p - 1) \hat{\sigma}^2$, 根据定理 1.3.1

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p-1}$$

而选模型 (2.3) 下的残差平方和 RSS_q 就是在假设 $H: \gamma_t = 0$ 下, 模型的残差平方和。所以, 仍依定理 1.3.1, 知 $B = \text{RSS}_q - \text{RSS}$ 与 RSS 相互独立。故

$$\begin{aligned} E\left(\frac{\text{RSS}_q}{\hat{\sigma}^2}\right) &= (n - p - 1) E\left(\frac{\text{RSS}_q}{\text{RSS}}\right) \\ &= (n - p - 1) \left[1 + E(B) \cdot E\left(\frac{1}{\text{RSS}}\right) \right] \end{aligned}$$

记 $k = n - p - 1$, 从 $\text{RSS} \sim \sigma^2 \chi^2_k$, 得

$$E\left(\frac{1}{\text{RSS}}\right) = 2^{-\frac{k}{2}} \left[\Gamma\left(\frac{k}{2}\right) \right]^{-1} \int_0^\infty x^{-1} \cdot e^{-\frac{x}{2}} x^{\frac{k}{2}-1} dx$$

$$\begin{aligned}
&= 2^{-\frac{k}{2}} \left[\Gamma\left(\frac{k}{2}\right) \right]^{-1} 2^{\frac{k}{2}-1} \Gamma\left(\frac{k}{2}-1\right) \\
&= \frac{1}{k-2} = \frac{1}{n-p-3}
\end{aligned} \tag{4.9}$$

其中 $\Gamma(x)$ 为Gamma函数. 注意到现在假设 $H: \gamma_i=0$ 不一定成立, 所以 $B/\sigma^2 \sim \chi^2_{(q)}(\delta)$, 这里 $\chi^2_{(q)}(\delta)$ 表示非中心参数为 δ 的自由度为 q 的 χ^2 分布. 因为对非中心 χ^2 分布, 其均值等于自由度与非中心参数之和, 所以为求 $E(B)$, 我们需计算 δ . 注意到, RSS 是中心 χ^2 分布, 所以, B 的非中心参数即为 RSS_q 的非中心参数. 依非中心参数的定义

$$\delta = \frac{B}{\sigma^2} \Big|_{\text{将 } Y \text{ 改为 } EY} = \frac{RSS_q}{\sigma^2} \Big|_{\text{将 } Y \text{ 改为 } EY}$$

利用(2.7), 上式

$$\delta = \frac{\gamma_i' D^{-1} \gamma_i}{\sigma^2}$$

于是

$$E(B) = E(\chi^2_{(q)}(\delta)) = q + \frac{\gamma_i' D^{-1} \gamma_i}{\sigma^2} \tag{4.10}$$

将(4.9)、(4.10)代入(4.8), 整理便推得(4.7). 定理证毕.

这个定理说明, C_p 统计量不是 Γ_q 的无偏估计. 但如果 $n-p$ 比较大, 使得

$$\frac{n-p-1}{n-p-3} \approx 1 \tag{4.11}$$

则从(4.4)和(4.7)可以看到, 此时 $E(C_p) \approx \Gamma_q$. 根据 Γ_q 的实际意义, Γ_q 愈小愈好. 所以我们应该选择具有最小 C_p 值的变量子集.

推论4.1 在定理4.1条件下, 若 $\gamma_i=0$, 则

$$C_p = (q-t) + tu,$$

等价地

$$C_p - q = t(u - 1).$$

这里 $u \sim F_{t, n-p-1}$.

证明 沿用定理4.1的记号, 并记

$$u = \frac{\text{RSS}_q - \text{RSS}}{t \hat{\sigma}^2}$$

易见, u 为检验假设 $H: \gamma_t = 0$ 的 F 统计量. 所以, 若 $\gamma_t = 0$, 依定理 1.3.1, $u \sim F_{t, n-p-1}$. 借助于 u , C_p 可表为

$$\begin{aligned} C_p &= \left(\frac{\text{RSS}_q - \text{RSS}}{t \hat{\sigma}^2} + \frac{\text{RSS}}{t \hat{\sigma}^2} \right) t - (n - 2q) \\ &= tu + \frac{(n - p - 1) \hat{\sigma}^2}{\hat{\sigma}^2} - (n - 2q) \\ &= (q - t) + tu \end{aligned}$$

此处应用了 $p + 1 = q + t$. 推论得证

现在我们来解释上述性质如何应用于变量选择. 若 $\gamma_t = 0$, 即选模型 (2.3) 正确, 从 (4.7) 得到

$$E(C_p) = q - t + \frac{n - p - 1}{n - p - 3} t$$

若 $n - p$ 比较大, 使得 (4.11) 成立, 那末

$$E(C_p) \approx q.$$

这说明, 对于正确的选模型, 在平面直角坐标系中, 点 (q, C_p) 落在第一象限角平分线附近. 如果选模型不正确, $\gamma_t \neq 0$, 则由 (4.7) 看到, 在条件 (4.11) 下, 有

$$E(C_p) = q + \frac{\gamma_t' D^{-1} \gamma_t}{\sigma^2} > q$$

此时, 点 (q, C_p) 将会向第一象限角平分线上方移动. 再结合上

C_p 为 Γ_0 的估计,我们就得到如下变量选择法则:选择对应的点(q , C_p)最接近第一象限角平分线,且 C_p 最小的选模型。

推论4.1使我们可以通过 F 分布来表示 $\gamma_i=0$,即选模型(2.3)正确的条件下, C_p-q 的分布。这个分布既然与 $F_{t,n-p-1}$ 相联系,故也有两个“自由度”,表4.1对 $v=n-p-1$ 的30和 ∞ (即 $n-p$ 比较大的情形)两个值及 $t=1, 2, \dots, 15$,给出了 $\alpha=0.10$ 和0.05的分位点。例如,若 $n=40$,自变量个数(不包括常数项) $p=9$,选模型含 $q=5$ 个自变量(含常数项),此时 $t=5$ 。从表4.1, $v=30$, $t=5$, $\alpha=0.10$, 查出 $Z_{0.10}=5.25$ 。即

$$P(C_p - q \geq 5.25 | \gamma = 30, t = 5) = 0.10.$$

这表明,当 $\gamma_i=0$ 为真时,虽然 $E(C_p) \approx q$,但对 $v=30$, $t=5$, $C_p - q$ 仍有0.10的概率超过5.25,即使 $n \rightarrow \infty$,查 $v=\infty$, $t=5$, 到出 $Z_{0.10}=4.24$,下降仍不多。仔细观察表4.1可以发现,对固定的 v 和 α ,当 t 增大时, Z_α 随之增大。注意到 t 是选模型丢弃的自变量个数,所以, C_p 方法适用于丢掉的自变量不太多的情况。

表4.1 $C_p - q$ 的分位点、表中值为 Z_α : $P(C_p - q \geq Z_\alpha | v, t) = \alpha$

	$t \backslash \alpha$	1	2	3	4	5	6	7	8	9	10	15
		1	2	3	4	5	6	7	8	9	10	15
$v=30$												
	0.10	1.88	2.98	3.83	4.57	5.25	5.88	6.49	7.07	7.64	8.20	10.83
	0.05	3.17	4.63	5.77	6.76	7.67	8.52	9.34	10.13	10.90	11.65	15.22
$v=\infty$	$t \backslash \alpha$	1	2	3	4	5	6	7	8	9	10	15
		1	2	3	4	5	6	7	8	9	10	15
	0.10	1.71	2.61	3.25	3.78	4.24	4.65	5.02	5.36	5.68	5.99	7.31
	0.05	2.84	3.99	4.82	5.49	6.07	6.59	7.07	7.51	7.92	8.31	10.00

(三) C_p 图

对于每个选模型，或者说对每个变量子集计算出对应的 C_p 统计量，将点 (q, C_p) 标在平面直角坐标系中，这些点构成的散点图称为 (q, C_p) 图，简称 C_p 图。 C_p 图是应用 C_p 准则选择自变量的有效工具，前面已经指出，依 C_p 准则应选点 (q, C_p) 最接近第一象限角平分线且 C_p 最小的选模型。当然，在很多情况下，这两个条件

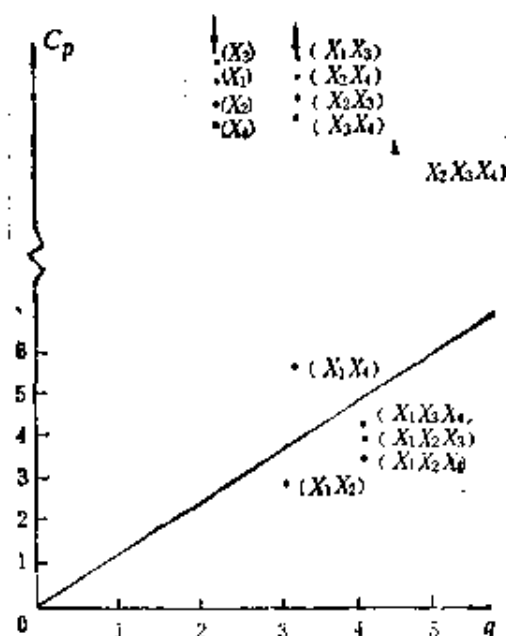
表4.2 Hald 水泥数据的 C 统计量

所选自变量个数	q	所选自变量	RMS_q	C_p
1	2	x_1	115.0624	202.55
1	2	x_2	82.3942	142.49
1	2	x_3	176.3092	315.16
1	2	x_4	80.3515	138.73
2	3	x_1x_2	5.7904	2.68
2	3	x_1x_3	122.7073	198.10
2	3	x_1x_4	7.4762	5.50
2	3	x_2x_3	41.5443	62.44
2	3	x_2x_4	86.8880	138.23
2	3	x_3x_4	17.5738	22.37
3	4	$x_1x_2x_3$	5.3456	3.04
3	4	$x_1x_2x_4$	5.3303	3.02
3	4	$x_1x_3x_4$	5.6485	3.50
3	4	$x_2x_3x_4$	8.2017	7.34
4	5	$x_1x_2x_3x_4$	5.9829	5.00

并不在同一个点上达到,此时要根据具体情况以及一些附加信息。来决定究竟选那一个模型。下面我们通过一些实例来说明 C_p 图的具体用法。

例4.1 Hald水泥问题(续例3.2)

对Hald水泥问题, 表4.2给出了所有选模型的 C_p 值. C_p 的最小值对应的变量子集为 (x_1, x_2) , $C_p = 2.68$. 另外一些较小的 C_p 统计量分别对应于 (x_1, x_2, x_3) , (x_1, x_2, x_4) 和 (x_1, x_3, x_4) . 对这三个变量子集, (x_1, x_3, x_4) 对应的 $(4, 3.50)$ 最接近第一象限的角平分线, 而 (x_1, x_2, x_4) 的 C_p 值最小. 如果没有别的附加考虑, 在 C_p 准则下 (x_1, x_3) 是“最优”子集

图3·4·1 Hald水泥的C₃图

与上节表3.3RMS_q准则相比,结果大体上相吻合。上面所列举的四个变量子集的RMS_q也比较小。但也有不尽一致的地方。RMS_q的最小值在(x_1, x_3, x_4)达到,而 C_p 的最小值在(x_1, x_2)达到。注意到(x_1, x_2)的RMS_q也比较小,所以,综合起来看, (x_1, x_2)是最适于采用的子集。

例4·2 汽车油耗问题(续例3·1)

图3·4·2是汽车油耗问题的 C_p 图。对每个 $q=1, 2, \dots, 11$, 标出了秩为1的自变量子集的 C_p 值。我们看到, C_p 的最小值在 $q=4$ 达到, 其值为0.103, 它所对应的子集为 (x_8, x_9, x_{10}) , 这与例3·1用 S_q 准则得到的结论是一致的。从图上也可以看出, $q=3, q=5$ 的秩为1的自变量子集的 C_p 值也比较小。

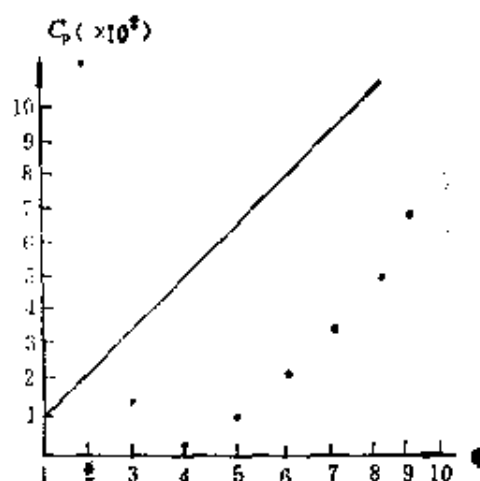


图3·4·2 汽车油耗问题 C_p 图

例4·3 空气污染问题

McDonald和Schwing^[80]曾经研究了死亡率与空气污染、气候以及社会经济状况等因素的关系。一共考虑了十五个因素:

x_1 = 年平均降雨量,

x_2 = 1月份平均气温,

x_3 = 3月份平均气温,

x_4 = 年令在65岁以上的人口占总人口的百分数,

x_5 = 每家的入口数, x_6 = 中学毕业年龄,

x_7 = 住符合标准的家庭比例数, x_8 = 每平方哩居民数,

x_9 = 非白种人占总人口的比例, x_{10} = “白领阶层”中受雇百分数,

x_{11} = 收入在300美元以上的家庭百分数,

x_{12} = 碳氢化合物的相对污染势,

x_{13} = 氮氧化物的相对污染势,

x_{14} = 二氧化硫的相对污染势,

x_{15} = 相对湿度。

共收集了 $n = 60$ 组数据。详细结果见原文。图3·4·3为 C_p 图。从这个图可以看到, $q = 7$ 时 C_p 达到最小, 它对应的自变量子集为 $x_1, x_2, x_3, x_6, x_8, x_{14}$ 。其次是 $q = 8$, 它对应的变量子集为在上面的子集中再增加 x_5 , 再其次是 $q = 9$, 对应的子集是再增加 x_4 。这个结果与用 RMS_q 准则所得的结果基本一致。这个问题我们将在后面多次讨论。

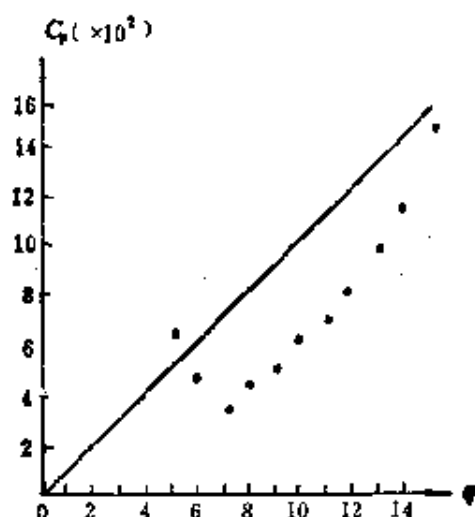


图3·4·3 空气污染问题 C_p 图

(四) 一般线性估计的 C_p

前面我们对LS估计讨论了 C_p 统计量及其在变量选择中的应用。不难看到, C_p 统计量的这种思想也可以用到其它一般的线性估计。

我们把常数项和回归系数分开来, 即在全模型(2·1)中, 将 $\gamma' = (\alpha, \beta')$ 代入, 得到

$$Y = \alpha \mathbf{1} + X\beta + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I. \quad (4 \cdot 12)$$

这里我们把 β 对应的设计阵仍记作 X , 并假定 X 已经中心化, $\mathbf{1}' = (1, 1, \dots, 1)$. 对常数项 α , 我们用 $\hat{\alpha} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ 来估计。考

考虑 β 的某个线性估计 $\tilde{\beta}_L = LY$, 这里 L 是给定的, 依赖于 X . 对于这样一个线性估计, 也存在一个变量选择问题. 我们需要先按一定准则选择一部分自变量, 然后在对应的选模型, 使用线性估计 $\tilde{\beta}_L = LY$.

因为 X 已经中心化, 根据线性估计可容许性的结果 (见 [8] 定理 4.7.7) 我们只需考虑满足条件 $L\mathbf{1} = 0$ 的矩阵 L . 此因在平方损失下, 为了估计 β , 我们只需讨论形如 $\hat{C}\beta = C(X'X)^{-1}X'Y$ 的估计, 即考虑形如 $L = C(X'X)^{-1}X'$ 的 L . 但由 X 的中心化知 $X'\mathbf{1} = 0$, 所以 $L\mathbf{1} = 0$. 用 $\tilde{\beta}_L = LY$ 估计 β , 其残差平方和为

$$RSS_L = \|Y - \bar{Y}\mathbf{1} - X\tilde{\beta}_L\|^2$$

作为 $\tilde{\beta}_L$ 的优良性指标, 对应于 Γ , 我们有

$$\begin{aligned}\Gamma_L &= \sum_{i=1}^n d_i / \sigma^2 = \frac{1}{\sigma^2} \sum_{i=1}^n E(\tilde{y}_i - E(y_i))^2 \\ &= \frac{1}{\sigma^2} E(\|X\tilde{\beta}_L - X\beta\|^2 + n(\bar{Y} - \alpha)^2)\end{aligned}\quad (4.13)$$

其中 \tilde{y}_i 为在第 i 个试验点, 用估计 $\tilde{\beta}_L$ 作出的预测.

我们先证明一个引理.

引理 4.1.

$$\Gamma_L = U_L + b_L / \sigma^2, \quad (4.14)$$

$$E(RSS_L) = \sigma^2 U_L^* + b_L, \quad (4.15)$$

这里

$$U_L = 1 + \text{tr}(X'XLL'),$$

$$b_L = \beta' X' (XL - I)' (XL - I) X \beta,$$

$$U_L^* = (n - 1) + \text{tr}(X'XLL') - 2\text{tr}(XL).$$

证明 利用 $L\mathbf{1} = 0$, 有

$$\begin{aligned}\|X\tilde{\beta}_L - X\beta\|^2 &= \|XL(\alpha\mathbf{1} + X\beta + e) - X\beta\|^2 \\ &= \|(XL - I)X\beta + XLe\|^2\end{aligned}$$

$$\begin{aligned}
&= \beta' X' (XL - I)' (XL - I) X \beta + e' L' X' (XL - I) X \beta \\
&\quad + \beta' X' (XL - I)' X L e \\
&\quad + e' L' X' X L e
\end{aligned}$$

利用迹的性质: $\text{tr}(AB) = \text{tr}(BA)$ 以及引理1·2·1, 有

$$\begin{aligned}
E\|\tilde{X}\beta_L - X\beta\|^2 &= b_L + E(e' L' X' X L e) \\
&= b_L + \sigma^2 \text{tr}(L' X' X L) \\
&= b_L + \sigma^2 (U_L - 1)
\end{aligned}$$

于是

$$\begin{aligned}
E(\Gamma_L) &= \frac{1}{\sigma^2} [b_L + \sigma^2 (U_L - 1) + nE(\bar{Y} - \alpha)^2] \\
&= \frac{1}{\sigma^2} \left(b_L + \sigma^2 (U_L - 1) + n \frac{\sigma^2}{n} \right) \\
&= U_L + b_L / \sigma^2.
\end{aligned}$$

(4·14)得证.

为证(4·15), 将 RSS_L 改写为

$$\begin{aligned}
\text{RSS}_L &= \left\| \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) Y \right\|^2 \\
&= Y' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right)' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) Y
\end{aligned}$$

再次利用引理1·2·1, 有

$$\begin{aligned}
E(\text{RSS}_L) &= \sigma^2 \text{tr} \left[\left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right)' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \right] \\
&\quad + (EY)' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right)' \\
&\quad \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) (EY) \\
&= a_1 + a_2
\end{aligned}$$

利用假设 $L\mathbf{1}=0$ 及 $X'\mathbf{1}=0$ 及矩阵迹的性质, 容易证得

$$\begin{aligned} a_1 &= \sigma^2 \text{tr} \left[\left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right)' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \right] \\ &= \sigma^2 [(n-1) - 2\text{tr}(XL) + \text{tr}(X'XLL')] \\ &= \sigma^2 U_L^*, \\ a_2 &= \beta' X' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right)' \left(XL - I + \frac{1}{n} \mathbf{1}\mathbf{1}' \right) X\beta \\ &= \beta' X' (XL - I)' (XL - I) X\beta = b_L \end{aligned}$$

明所欲证.

利用这个引理, 我们很易推得

定理4.2.

$$\Gamma_L = \frac{E(\text{RSS}_L)}{\sigma^2} - [n - 2(1 + \text{tr}(XL))] \quad (4.16)$$

证明 从(4.14)和(4.15), 有

$$\begin{aligned} \Gamma_L &= 1 + \text{tr}(X'XLL') + b_L/\sigma^2 \\ &= 2\text{tr}(XL) + U_L^* - (n-2) + b_L/\sigma^2 \\ &= \frac{1}{\sigma^2} (\sigma^2 U_L^* + b_L) + 2\text{tr}(XL) - (n-2) \\ &= \frac{E(\text{RSS}_L)}{\sigma^2} - [n - 2(1 + \text{tr}(XL))] \end{aligned}$$

证毕.

比较(4.16)与(4.6), 我们看到 Γ_L 与 Γ_e 很类似。和以前一样, 在(4.16)中用 $\hat{\sigma}^2$ 代替 σ^2 , 去掉 $E(\text{RSS}_L)$ 的均值符号, 得到 Γ_L 的估计

$$C_L = \frac{\text{RSS}_L}{\hat{\sigma}^2} - [n - 2(1 + \text{tr}(XL))] \quad (4.17)$$

如果取线性估计 $\hat{\beta}_L$ 为 LS 估计, (4.17) 就化为 C_p 统计量. 事实上,

考虑选模型

$$Y = \alpha \mathbf{1} + X_{q-1} \beta_{q-1} + X_t \beta_t + e,$$

这里 $X = (X_{q-1} : X_t)$, X_{q-1} 有 $q-1$ 列, X_t 有 t 列, $q-1+t=p$. β_{q-1} 的 LS 估计 $\tilde{\beta}_{q-1} = LY$, 其中

$$L = \begin{pmatrix} (X_{q-1}' X_{q-1})^{-1} X_{q-1}' \\ 0 \end{pmatrix},$$

所以

$$\begin{aligned} \text{tr}(XL) &= \text{tr}(X_{q-1} (X_{q-1}' X_{q-1})^{-1} X_{q-1}') \\ &= \text{tr}(I_{q-1}) \\ &= q-1 \end{aligned}$$

代入(4.17), 即得 C_p .

C_L 有两种用途. 其一是变量选择, 这和 C_p 方法无本质差别. 其二是用作在某一估计类中选择最优估计. 例如, 设 LY 是一个线性估计类, 它包含有一个或多个可供选择的参数, 因此算出的 C_L 是这个或这些参数的函数. 在该估计类中选择最优估计就归结为选择这些参数使 C_L 达到最小.

下面我们以岭估计为例说明 C_L 的第二种用途. 关于岭估计的详细讨论见第四章.

对于线性回归模型(4.12), 在设计阵 X 中心化的条件下, 常数项 α 总是用 $\hat{\alpha} = \bar{Y} = \frac{1}{n} \sum_i Y_i$ 来估计. 所以, 以下只讨论 β 的估计.

β 的岭估计定义为

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'Y \triangleq L_k Y, \quad (4.18)$$

其中 $k \geq 0$ 为待选参数. 对不同的 k , 构成了不同的估计, 因此岭估计 $\hat{\beta}(k)$ 是一个估计类.

因为 $X'X$ 为 p 阶对称方阵, 如众所周知, 存在正交方阵 P 致

$$P'X'XP = A = \text{diag}(\lambda_1, \dots, \lambda_p). \quad (4.19)$$

于是

$$\begin{aligned} \text{tr}(XL_k) &= \text{tr}[X(X'X + kI)^{-1}X'] \\ &= \text{tr}[(X'X + kI)^{-1}X'X] \\ &= \text{tr}[P'(X'X + kI)^{-1}PP'X'XP] \\ &= \sum_{i=1}^p \frac{\lambda_i}{\lambda_i + k} \end{aligned} \quad (4.20)$$

为了计算 C_L , 我们还需计算残差平方和 RSS_{Lk} . 将 L_k 代入 RSS_L 的表达式, 得到

$$\begin{aligned} \text{RSS}_{Lk} &= \|Y - \bar{Y}\mathbf{1} - X(X'X + kI)^{-1}X'Y\|^2 \\ &= \|Y - \bar{Y}\mathbf{1}\|^2 + Y'X(X'X + kI)^{-1}X'X(X'X \\ &\quad + kI)^{-1}X'Y - 2Y'X(X'X + kI)^{-1}X'Y \\ &\triangleq A_1 + A_2 - 2A \end{aligned} \quad (4.21)$$

由(4.19)有 $X'X = PAP'$, 作变换 $Z = PX'Y$, 得

$$\begin{aligned} A_2 &= Z'P(X'X + kI)^{-1}P'PX'XP' \cdot P(X'X + kI)^{-1}P'Z \\ &= \sum_{i=1}^p \frac{\lambda_i Z_i^2}{(k + \lambda_i)^2} \\ A_3 &= Z'P(X'X + kI)^{-1}Z \\ &= \sum_{i=1}^p \frac{Z_i^2}{k + \lambda_i} \end{aligned}$$

其中 $Z' = (Z_1, \dots, Z_p)$. 又

$$A_1 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

代入(4.21), 得到

$$\text{RSS}_{Lk} = \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2 \sum_{i=1}^p \frac{Z_i^2}{k + \lambda_i} + \sum_{i=1}^p \frac{\lambda_i Z_i^2}{(k + \lambda_i)^2}$$

于是, 对岭估计, C_L 统计量为

$$C_{Lk} = \hat{\sigma}^{-2} \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 - 2 \sum_{i=1}^p \frac{Z_i^2}{k + \lambda_i} + \sum_{i=1}^p \frac{\lambda_i Z_i^2}{(k + \lambda_i)^2} \right] + 2 \sum_{i=1}^p \frac{\lambda_i}{\lambda_i + k} - (n - 2) \quad (4.22)$$

我们的准则是：选择使 C_{Lk} 达到最小的 k 。

一个重要的问题是，由 C_{Lk} 最小准则所确定的岭估计，究竟性质如何？为此，我们假定设计是正交的，并假定设计阵已经中心标准化，此时 $X'X = I$ 。因此，一切 $\lambda_i = 1$ ， $P = I$ ， $Z = X'Y = \hat{\beta}$ 从而

$$C_{Lk} = \hat{\sigma}^{-2} \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{2}{1+k} \sum_{i=1}^p \hat{\beta}_i^2 + \frac{1}{(1+k)^2} \sum_{i=1}^p \hat{\beta}_i^2 \right] + \frac{2p}{1+k} - (n - 2)$$

将上式关于 k 求导数，并令其为零，解得使 C_{Lk} 达到最小值的 k 满足方程

$$\frac{1+k}{k} = -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^p \hat{\beta}_i^2$$

最后解得最优的 k 值为

$$k^* = \frac{\hat{\sigma}^2}{\|\hat{\beta}\|^2 - \hat{\sigma}^2}$$

对应的岭估计为

$$\beta^* = \left(1 - \frac{p\hat{\sigma}^2}{\|\hat{\beta}\|^2} \right) \hat{\beta}$$

这个估计也称为 James-Stein 估计，可以证明（见 [8]，定理 6.2.3）当

$$0 < c < \frac{2(n-p)(p-2)}{n-p+2}$$

时, 一切形如

$$\tilde{\beta} = \left(1 - \frac{c\sigma^2}{\|\hat{\beta}\|^2}\right) \hat{\beta}$$

的估计, 在平方损失下优于 LS 估计 $\hat{\beta}$.

§3.5 预测平方和准则

预测平方和(Prediction Sum Error of Square, PRESS)准则是由 D.M. Allen 于 1971 年提出的^[31, 32]. 虽然和前两节的多数准则一样, 也是基于预测精度, 但是在具体方法上有重要的区别。以前的几种基于预测的统计量都有一个共同的缺点: 即在计算某点的预测偏差时, 该点曾在建立经验回归方程中使用过。而 PRESS 准则克服了这一缺点, 因此可望有更好的效果。

考虑线性回归模型

$$Y = X\gamma + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I, \quad (5.1)$$

这里设计阵 $X' = (x_1, x_2, \dots, x_n)$. 如在第二章所作的一样, 剔除第 i 次试验, 得到模型

$$Y(i) = X(i)\gamma + e(i), \quad (5.2)$$

其中 $Y(i)$, $X(i)$, $e(i)$ 分别表示从 Y 、 X 、 e 中删去第 i 行得到的向量或矩阵。根据模型 (5.2), 我们可以导出 γ 的 LS 估计 $\hat{\gamma}(i)$. 然后利用这个估计对第 i 个试验点 x_i 处的因变量作预测, 得到预测值 $x_i' \hat{\gamma}(i)$, (请注意, 在作估计 $\hat{\gamma}(i)$ 时, 与 x_i 对应的数据都没有使用过), 预测偏差

$$d_i = Y_i - x_i' \hat{\gamma}(i) \quad (5.3)$$

对每一个点都这样作, 得到 d_1, d_2, \dots, d_n . 其平方和为

$$\text{PRESS} = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (Y_i - x_i' \hat{\gamma}(i))^2 \quad (5.4)$$

称为预测平方和. (5.4)是全模型的预测平方和, 它度量了全模型的优劣. 对于选模型

$$Y = X_q \gamma_q + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I, \quad (5.5)$$

用 X_q 代替(5.4)中 X , 得到该模型的预测平方和, 记为 PRESS_q . PRESS 准则是, 选择使 PRESS_q 达到最小的选模型.

因为(5.4)含有 $\hat{\gamma}(i)$, $i=1, 2, \dots, n$, 似乎为了计算 PRESS , 需要计算 n 个不同的回归. 实际不然, 在第二章我们已经知道 $\hat{\gamma}(i)$ 可以通过 $\hat{\gamma}$ 来表示, 因此利用完全数据的回归分析结果计算 PRESS 是可能的. 下面的公式就给出了计算 PRESS 的一个简单方法

$$\text{PRESS} = \sum_{i=1}^n \left(\frac{\delta_i}{1 - h_{ii}} \right)^2 \quad (5.6)$$

这里 $\delta_i = Y_i - x_i' \hat{\gamma}$ 为普通残差, h_{ii} 为帽子矩阵 $H = X(X'X)^{-1}X'$ 的第 i 个对角元.

现在我们证明(5.6). 根据公式(2.2.17)

$$\hat{\gamma}(i) = \hat{\gamma} - \frac{(X'X)^{-1}x_i\delta_i}{1 - h_{ii}}, \quad i=1, 2, \dots, n$$

于是

$$\begin{aligned} d_i &= Y_i - x_i' \hat{\gamma}(i) \\ &= Y_i - x_i' \hat{\gamma} + \frac{x_i' (X'X)^{-1} x_i \delta_i}{1 - h_{ii}} \\ &= \delta_i + \frac{h_{ii}}{1 - h_{ii}} \delta_i \\ &= \frac{\delta_i}{1 - h_{ii}}, \quad i=1, 2, \dots, n \end{aligned}$$

代入(5.4), 即得(5.6).

对于选模型的 PRESS_q , 只要将(5.6)的 δ_i 换成 $\delta_{qi} = Y_i - \tilde{x}'_{qi} \tilde{\gamma}_q$, h_{ii} 换成 $H_{ii} = X_q(X_q'X_q)^{-1}X_q'$ 的第 i 个对角之即可, 这里 $x'_i = (x'_{i1}, x'_{i2})$, $\gamma'_q = (x'_qX_q)^{-1}X_q'Y$

(5.6)把PRESS表成了普通残差 δ_i 的加权平方和, 根据 h_{ii} 的几何意义(见§2.2), 它度量了第 i 个试验点距离试验区域中心的远近。如果第 i 个试验点距离试验区域中心愈远, h_{ii} 就愈大, δ_i 的权 $w_i = \frac{1}{(1-h_{ii})^2}$ 也就愈大。因此 PRESS 准则比较强调高杠杆点的作用。

从这里多少可以看到, 变量选择与回归诊断有着密切的关系。事实上, 对于其它变量选择准则也是一样, 异常点、高杠杆点和强影响点以及下章要讲的复共线性的存在对变量选择的结果都有相当的影响。有一种观点认为, 对全模型先作最小二乘分析和诊断, 剔除异常点、高杠杆点和强影响点, 然后再进行变量选择, 但尚未见到有理论结果支持这种处理方法

PRESS的每次剔除一组数据的作法与统计学中目前很热门的刀切法(Jackknife法)有共同之处。这种处理方法在其它分支, 如多元分析、参数估计等都有一定的应用。

例5.1 Hald水泥问题(续例3.2)

在例3.2和例4.1, 我们已经用 RMS_q 准则和 C_p 准则分别讨论了Hald 水泥问题的变量选择。子集 (x_1, x_2) 具有最小的 C_p 值和较小的 RMS_q , 而子集 (x_1, x_2, x_4) 的 RMS_q 最小, C_p 值是次最小。对这两个子集, PRESS 计算列在表5.1。从PRESS值看, 两者相差不是很多。虽然 (x_1, x_2, x_4) 的PRESS的值略小些, 但它比 (x_1, x_2) 增加了一个自变量 x_4 , 从增加了变量但PRESS下降不多这个事实看, 很可能 x_4 与 x_1 或 x_2 有很高的相关性。从表5.2给出的相关系数阵可以发现, x_4 与 x_2 有高度的负相关, 这就解释了为什么添加 x_4 之后, PRESS下降不多的原因。总起来, 我们应该选择子集 (x_1, x_2) 。

表5.1 Hald水泥数据的两个选模型PRESS

数据号	$y = 52.58 + 1.468x_1 + 0.662x_2$			$y = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$		
	δ_i	h_{ii}	$(\delta_i/(1-h_{ii}))^2$	δ_i	h_{ii}	$(\delta_i/(1-h_{ii}))^2$
1	-1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	-1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	-1.5147	0.11890	2.9553	-1.8910	0.13315	4.7588
4	-1.6585	0.24225	4.7905	-1.8016	0.24431	5.6837
5	-1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	-1.3031	0.36180	4.1627	-1.4287	0.36341	5.0369
8	-2.0754	0.24119	7.4806	-3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5453
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	-2.8934	0.21420	13.5579	-2.2247	0.25923	9.0194
	PRESS $x_1, x_2 = 93.8827$			PRESS $x_1, x_2, x_4 = 85.3516$		

表5.2 Hald水泥数据的样本相关阵

	x_1	x_2	x_3	x_4	Y
x_1	1.000				
x_2	0.229	1.000			
x_3	-0.824	-0.139	1.000		
x_4	-0.245	-0.973	0.030	1.000	
Y	0.731	0.816	-0.535	-0.321	1.000

D. M. Allen于1971年在前面所引的同一篇文章中还提出了另外一种基于预测精度的变量选择准则,称为预测均方误差准则。对每个给定的 $x' = (x'_q, x'_r)$, 在该处从选模型作出的预测的预测偏差平方的均值 $E(x'_q \tilde{\gamma}_q - x' \beta)^2$, 在去掉一个只与全模型有关的量之后, 就只与所选的自变量有关, 预测均方误差准则在于选择这样的自变量子集, 使得在给定点 x 处的预测偏差平方的均值达到最小。这样对不同的 x , 所选的自变量子集也不一样。此方法的缺点正在这一点。因为在应用上, 所建立的回归模型在今后的应用中仅仅限于预测一个固定点的情况是很少见的, 故此种方法的可用范围受到很大限制。类似地, Aitkin^[33]还应用这个准则以及修正后的其它一些准则, 讨论了寻求“不显著地坏于”全模型的选模型的方法。

§3.6 AIC准则

众所周知, 极大似然原理是统计学中估计参数的一个重要方法。Akaike^[34]把这个方法加以修正, 提出了一种较为一般的模型选择准则, 文献中称为**Akaike信息量准则** (Akaike Information Criterion, 简记为AIC)。AIC准则应用比较广泛, 例如, 它可用于时间序列分析中自回归阶数的确定等。本节我们讨论如何把它应用于用回归自变量选择。

在选模型(5.5)中, 假设误差 $e \sim N(0, \sigma^2 I)$, 则 γ_q 和 σ^2 的似然函数为

$$L(\gamma_q, \sigma^2 | Y) = (2\pi\sigma^2)^{-T} \exp \left\{ -\frac{1}{2\sigma^2} \|Y - X_q \gamma_q\|^2 \right\} \quad (6.1)$$

容易求得 γ_q 和 σ^2 的极大似然估计

$$\begin{aligned}\tilde{\gamma}_q &= (X_q' X_q)^{-1} X_q' Y, \\ \sigma_q^{*2} &= \frac{RSS_q}{n} = \frac{Y'(I - X_q(X_q' X_q)^{-1} X_q') Y}{n}\end{aligned}$$

代入(6.1)并取对数, 得到对数似然函数的最大值

$$\ln L(\tilde{\gamma}_q, \sigma_q^{*2} | Y) = \left[\ln \left(\frac{n}{2\pi} \right)^{\frac{n}{2}} - \frac{n}{2} \right] - \frac{n}{2} \ln(RSS_q) \quad (6.2)$$

略去与 q 无关的项, 得到

$$\ln L(\tilde{\gamma}_q, \sigma_q^{*2} | Y) \propto -\frac{n}{2} \ln(RSS_q) \quad (6.3)$$

前面已经指出, 残差平方和 RSS_q 随着自变量个数 q 的增加而减少, 因此, 如按极大似然途径势必导致选择全模型。可见, 如把似然原理应用于变量选择, 适当的修正是必要的。

Akaike所提出的修正是这样的: 对一般的统计模型, 设 Y_1, \dots, Y_n 为一组样本, 如果它们服从某个含 k 个参数的模型, 对应的似然函数最大值记为 $L_k(Y_1, \dots, Y_n)$, Akaike建议选择使 $\ln L_k(Y_1, \dots, Y_n) - k$ 达到最大的模型, 这个量称为AIC统计量, 即

$$AIC = \ln L_k(Y_1, \dots, Y_n) - k. \quad (6.4)$$

可见, AIC统计量就是在对数似然函数最大值上添加了对变量个数的惩罚项。

对于选模型(5.5), 从(6.3)容易看到

$$AIC = -\frac{n}{2} \ln(RSS_q) - q$$

等价地, 可取

$$AIC = n \ln(RSS_q) + 2q. \quad (6.5)$$

于是, AIC准则归结为: 选择使(6.5)达到最小的自变量子集。

关于AIC准则, 近年来有一些不同形式的推广。考虑只有两

个备选模型的特殊情况。也就是说，我们有两组自变量子集，要决定从中选择哪一个。记它们的似然函数最大值分别为 $L_{k_i}(Y_1, \dots, Y_n)$, $i=1, 2$, 这里 k_1, k_2 表它们所含的自变量个数。从 (6.4) 知, AIC 准则等价于:

若

$$\ln \frac{L_{k_2}(Y_1, \dots, Y_n)}{L_{k_1}(Y_1, \dots, Y_n)} - (k_2 - k_1) > 0 \quad (6.6)$$

则选取模型 2, 否则选取模型 1. 如果记

$$\lambda = \frac{1}{2} \ln \frac{L_{k_2}(Y_1, \dots, Y_n)}{L_{k_1}(Y_1, \dots, Y_n)}$$

(6.6) 就变为

$$\lambda - 2(k_2 - k_1) \quad (6.7)$$

这是只有两个备选模型的 AIC. 近年来一些作者从 Bayes 观点提出的模型选择准则大致可统一表述为

$$\Delta(m) = \lambda - m(k_2 - k_1) \quad (6.8)$$

这里, 对不同的准则, m 取值不同。显然, 对 AIC 准则 (6.7), $m=2$. Schwarz^[85] 基于 Bayes 估计的大样本性质, 建议取 $m=\ln n$. 而 Smith 等^[86] 从另外的 Bayes 考虑提出取 $m=1$ 或 $m=3/2$ 等。这些准则的差别仅仅在于对变量个数的惩罚程度。除了 AIC 准则 (6.5) 之外, 其余尚未见诸实用, 故此处不再详细讨论。

关于自变量选择准则, 我们就讨论到这里。还有一些变量选择方法, 如岭回归法、主成分回归法, 我们将在下一章讨论。至于逐步回归方法, 与其看作一个变量选择准则, 不如当作一种子集回归的计算方法, 因此, 我们把它放在 §3.9 讨论。

§3.7 计算方法：扫描运算和Gauss消去法

前面我们总共讨论了六种回归自变量选择准则：平均残差平方和、预测偏差的方差、平均预测均方误差、 C_p 统计量、PRESS和AIC准则。其中除PRESS之外，对其余五种准则，作为选择标准的统计量都是选模型的残差平方和 RSS_q 的非常简单的函数。即便对PRESS准则，PRESS也易从其它统计量算出。因此，在下面我们讨论计算方法时都以计算 RSS_q 为例。

我们已经知道，无论哪一种变量选择准则都需要对不同的自变量子集进行比较，计算量都很大。例如，当自变量个数为 p 时，包含1个自变量的子集（即选模型）有 C_p^1 个，包含有2个的有 C_p^2 个，一般包含 k 个自变量的子集有 C_p^k 个，如果把仅含常数项的也算作一个，那么所有可能的自变量子集共有 2^p 个。例如当 $p=10$ 时，共有 $2^{10}=1024$ 。当 p 比较大时，数字 2^p 大得惊人。对这么多的自变量子集，要计算出对应的回归系数的估计及 RSS_q ，所需的计算量相当大，而且存储量、计算误差的积累也是值得注意的问题。因此，我们必须设计一个非常合理的计算次序，使得从一个自变量子集到另一个自变量子集所需计算量比较省，并且把误差积累、存储量都控制在一定的范围之内。这个问题不解决，前面所介绍的种种自变量选择方法就难以付诸实用。在60—70年代，回归自变量选择的计算问题获得了比较彻底的解决，提出了许多计算所有可能回归的有效算法，例如，Furnival[37]、Furnival和Wilson[38]、Garside[39]和Schatzoff等[40]。这些算法的基本方法是，在计算全部 2^p 个子集回归时，下一步要计算的子集回归和前一步的子集回归只相差一个自变量，而所用的计算都是扫描运算(Sweep operator, 以下简称S运算)或Gauss消去法。因为这两种运算

本身还有其它的用处,如[41],因而本节对它们作较详细的讨论。

(一) 扫描运算

设 $A = (a_{ij})_{n \times n}$. 若 $a_{ii} \neq 0$, 定义一个新的方阵 $B = (b_{ij})_{n \times n}$, 其中

$$\begin{aligned} b_{ii} &= 1/a_{ii} \\ b_{ij} &= a_{ij}/a_{ii}, \quad j \neq i \\ b_{ji} &= -a_{ji}/a_{ii}, \quad j \neq i \\ b_{kl} &= a_{kl} - a_{ki}a_{li}/a_{ii}, \quad k \neq i, \quad l \neq i. \end{aligned} \quad (7.1)$$

称由 A 到 B 的这种变换为以 a_{ii} 为枢轴的 S 运算, 记为 $B = S_i A$.

例如, 对 $i=1$, 即以 a_{11} 为枢轴的 S 运算, $B = S_1 A$ 为

$$B = \begin{bmatrix} a_{11}^{-1} & \vdots & \frac{a_{12}}{a_{11}} & \dots & \frac{a_{1n}}{a_{11}} \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{a_{21}}{a_{11}} & \vdots & & & \\ \vdots & \vdots & a_{ii} - \frac{a_{i1}a_{1i}}{a_{11}} & & \\ -\frac{a_{n1}}{a_{11}} & & & & \end{bmatrix} \quad (7.2)$$

根据定义, 容易验证 S 运算具有下列性质:

- (1) $S_i S_i A = A$;
- (2) $S_i S_j A = S_j S_i A$ (可交换性).

第一条性质说明, 以同一个对角元为枢轴的 S 运算连续作两次等于没有作 S 运算。把这一条与第二条可交换性结合起来, 可以推知, 在任意形如 $S_{i_1} S_{i_2} \dots S_{i_m} A$ 的表达式中, 若某一足标出现偶数次, 则与之相应的 S 可全部去掉。若某足标出现奇数次, 则与之相应的 S 只保留一个, 且对剩下的 S 可按任意次序排列。

S运算在回归分析中的应用，主要依据如下定理。

定理7·1 将 n 阶方阵 A 分块为

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

其中 A_{11} 为 r 阶可逆阵，则

$$S_1 S_2 \cdots S_r A = \begin{pmatrix} A_{11}^{-1} & A_{11}^{-1} A_{12} \\ -A_{21} A_{11}^{-1} & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{pmatrix} \quad (7.3)$$

证明 设 $x' = (x_1, \cdots, x_n)$ 和 $y' = (y_1, \cdots, y_n)$ 为两个 n 维向量，满足方程： $y + Ax = 0$ ，即

$$\begin{cases} y_1 + a_{11}x_1 + \cdots + a_{1n}x_n = 0, \\ y_2 + a_{21}x_1 + \cdots + a_{2n}x_n = 0, \\ \cdots \cdots \\ y_n + a_{n1}x_1 + \cdots + a_{nn}x_n = 0, \end{cases} \quad (7.4)$$

在这组关系中，把 x_1, \cdots, x_n 视为自变量，而 y_1, \cdots, y_n 为因变量。现在我们把 x_1 和 y_1 的位置调换一下，即把 y_1, x_2, \cdots, x_n 看作自变量，而 x_1, y_2, \cdots, y_n 作为因变量。设 $a_{11} \neq 0$ ，由(7.4)的第一式解出 x_1 ，得

$$x_1 = -\left(\frac{1}{a_{11}} y_1 + \frac{a_{12}}{a_{11}} x_2 + \cdots + \frac{a_{1n}}{a_{11}} x_n \right)$$

代入以后各式，整理后连同原来的第一式，即为

$$\begin{cases} x_1 + \frac{1}{a_{11}} y_1 + \frac{a_{12}}{a_{11}} x_2 + \cdots + \frac{a_{1n}}{a_{11}} x_n = 0, \\ y_2 - \frac{a_{21}}{a_{11}} y_1 + \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}} \right) x_2 + \cdots + \left(a_{2n} - \frac{a_{21}a_{1n}}{a_{11}} \right) x_n = 0 \\ \cdots \cdots \\ y_n - \frac{a_{n1}}{a_{11}} y_1 + \left(a_{n2} - \frac{a_{n1}a_{12}}{a_{11}} \right) x_2 + \cdots + \left(a_{nn} - \frac{a_{n1}a_{1n}}{a_{11}} \right) x_n = 0 \end{cases}$$

其中自变量 y_1, x_2, \dots, x_n 的系数矩阵为 $S_1 A$. 这说明, 把 x_1 和 y_1 的位置调换一下, 在自变量 x_1, \dots, x_n 的系数矩阵 A 上产生的后果, 是把 A 变为 $S_1 A$. 一般, $S_1 \cdots S_r A$ 相当于把 x_1, \dots, x_r 与 y_1, \dots, y_r 调换, 即自变量改为 $y_1, \dots, y_r, x_{r+1}, \dots, x_n$, 而因变量变为 $x_1, \dots, x_r, y_{r+1}, \dots, y_n$.

但上述调换在矩阵 A 上的后果也可以直接算出来. 记

$$y_{(1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_r \end{bmatrix}, \quad y_{(2)} = \begin{bmatrix} y_{r+1} \\ \vdots \\ y_n \end{bmatrix},$$

$$x_{(1)} = \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix}, \quad x_{(2)} = \begin{bmatrix} x_{r+1} \\ \vdots \\ x_n \end{bmatrix}$$

(7.4) 变形为

$$\begin{cases} y_{(1)} + A_{11}x_{(1)} + A_{12}x_{(2)} = 0 \\ y_{(2)} + A_{21}x_{(1)} + A_{22}x_{(2)} = 0 \end{cases}$$

将第一式两边左乘 A_{11}^{-1} , 得

$$A_{11}^{-1}y_{(1)} + x_{(1)} + A_{11}^{-1}A_{12}x_{(2)} = 0, \quad (7.5)$$

解出 $x_{(1)}$, 代入第二式, 整理后得

$$y_{(2)} - A_{21}A_{11}^{-1}y_{(1)} + (A_{22} - A_{21}A_{11}^{-1}A_{12})x_{(2)} = 0. \quad (7.6)$$

(7.5) 和 (7.6) 即为

$$\begin{cases} x_{(1)} + A_{11}^{-1}y_{(1)} + A_{11}^{-1}A_{12}x_{(2)} = 0 \\ y_{(2)} - A_{21}A_{11}^{-1}y_{(1)} + (A_{22} - A_{21}A_{11}^{-1}A_{12})x_{(2)} = 0 \end{cases} \quad (7.7)$$

在 (7.7) 中, 自变量 $y_{(1)}, x_{(2)}$ 的系数矩阵为

$$\begin{pmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}$$

根据前面的讨论, 这正是 $S_1 \cdots S_r A$. 定理证毕.

比较(7.2)和(7.3), 我们可以看出, 公式(7.3)可理解为: 若连续进行几个 S 运算, 则可以把枢轴的全体所在的行、列组成的子方阵看作一个数, 然后按照(7.2)的方式进行运算.

现在我们把上述结果应用于线性回归计算.

考虑线性回归模型

$$Y = a\mathbf{1} + X\beta + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I,$$

假设 $X_{n \times p}$ 已经中心化, 于是 $X'Y = X'(Y - \bar{Y}\mathbf{1})$. 将 X 分块, $X = (X_q; X_t)$, 假定 X_q 有 q 列, X_t 有 t 列, 此处 $q + t = p$. 记

$$S = X'X = \begin{pmatrix} X_q'X_q & X_q'X_t \\ X_t'X_q & X_t'X_t \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

$$A = \begin{pmatrix} S & X'Y \\ Y'X & \|Y - \bar{Y}\mathbf{1}\|^2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & X_q'Y \\ S_{21} & S_{22} & X_t'Y \\ Y'X_q & Y'X_t & \|Y - \bar{Y}\mathbf{1}\|^2 \end{pmatrix} \quad (7.8)$$

假设 $q \leq p$. 依定理 7.1, 有

$$S_1 \cdots S_r A = \begin{pmatrix} S_{11}^{-1} & * & S_{11}^{-1}X_q'Y \\ * & * & * \\ * & * & \|Y - \bar{Y}\mathbf{1}\|^2 - Y'X_t S_{11}^{-1}X_q'Y \end{pmatrix} \quad (7.9)$$

其中 “*” 表示没有必要写出的一些子块. 注意到 $S_{11}^{-1}X_q'Y$ 是从选模型

$$Y = a\mathbf{1} + X_q\beta_q + e \quad (7.10)$$

得到的 β_q 的 LS 估计 $\tilde{\beta}_q = (X_q'X_q)^{-1}X_q'Y$, 而

$$\|Y - \bar{Y}\mathbf{1}\|^2 - Y'X_t S_{11}^{-1}X_q'Y$$

是该选模型对应的残差平方和 RSS_q . 事实上, 利用 $X_q'\mathbf{1} = 0$, 有

$$\begin{aligned}
RSS_q &= \|Y - \mathbf{1}\bar{Y} - X_q \tilde{\beta}_q\|^2 \\
&= (Y - \mathbf{1}\bar{Y} - X_q \tilde{\beta}_q)'(Y - \mathbf{1}\bar{Y} - X_q \tilde{\beta}_q) \\
&= \|Y - \mathbf{1}\bar{Y}\|^2 - 2\tilde{\beta}_q' X_q'(Y - \mathbf{1}\bar{Y}) + \tilde{\beta}_q' X_q' X_q \tilde{\beta}_q \\
&= \|Y - \mathbf{1}\bar{Y}\|^2 - Y' X_q S_{11}^{-1} X_q' Y
\end{aligned}$$

又 S_{11}^{-1} 为估计 $\tilde{\beta}_q$ 的协方差阵(除去因子 σ^2 不计外)。可见,应用 S 运算得到的(7.9)给出了选模型(7.10)最小二乘回归的主要统计量。更一般,我们有如下定理。

定理7.2 设 $1 \leq i_1 < i_2 < \dots < i_q \leq p$, 对(7.8)定义的矩阵 A 施以 S 运算 $S_{i_1}, S_{i_2}, \dots, S_{i_q}$, 得到 $B = S_{i_1} \dots S_{i_q} A$, 则 B 的结构为

- (1) B 的 i_1, \dots, i_q 行和 i_1, \dots, i_q 列交叉处元素组成的子阵为 A 的相应子阵的逆;
- (2) $b_{i_1, p+1}, b_{i_2, p+2}, \dots, b_{i_q, p+1}$ 为选入回归的自变量 $x_{i_1}, x_{i_2}, \dots, x_{i_q}$ 的回归系数的LS估计;
- (3) $b_{p+1, p+1}$ 为对应于所选自变量子集的残差平方和;
- (4) 将(1)中所说的 B 的子方阵乘以 σ^2 , 等于(2)中所说的回归系数LS估计的协方差阵。

根据这个定理,我们就可以说明如何应用 S 运算来作回归自变量选择。假定自变量 x_{i_1}, \dots, x_{i_q} 已选入回归模型, 下一步无论是引入或删除一个自变量, 只需对 $S_{i_1} \dots S_{i_q} A$ (这个矩阵是存在计算机里的)作一次 S 运算。例如, 如要把 $x_l, l \neq i_1, \dots, i_q$ 引入回归模型, 只需施 S_l 于 $S_{i_1} \dots S_{i_q} A$ 就够了。反过来, 如要把某个自变量, 例如 x_{i_1} 从回归方程中剔除, 那末根据前面 S 运算的第一条性质, 只需施 S_{i_1} 运算到 $S_{i_1} \dots S_{i_q} A$, 得到 $S_{i_2} \dots S_{i_q} A$ 。根据定理7.2, $S_{i_2} \dots S_{i_q} A$ 就给出了自变量 x_{i_2}, \dots, x_{i_q} 在回归方程的LS回归的主要统计量。

(二) Gauss消去法

设有方阵 $A = (a_{ij})_{n \times n}$, 所谓以 a_{ii} 为枢轴的 Gauss 消去运算 (以下简称 G 运算), 是指这样一个运算

$$b_{kj} = \begin{cases} a_{ki} - a_{ki}a_{ij}/a_{ii}, & i+1 \leq k, j \leq n \\ a_{kj}, & \text{其它} \end{cases} \quad (7.11)$$

从定义容易看出, 若 $1 \leq i_1 < \dots < i_q \leq n-1$, 依次以 $a_{i_1 i_1}, \dots, a_{i_q i_q}$ 元为枢轴对 A 作 G 运算, 对 a_{nn} 的作用与 S 运算 S_{i_1}, \dots, S_{i_q} 对 a_{nn} 的作用相同. 再由定理 7.2 的结果 (3) 知, G 运算也可用来计算 RSS_q .

定理 7.3 设 $1 \leq i_1 < \dots < i_q \leq p$, 对 (7.8) 定义的方阵 A 依次作以 $a_{i_1 i_1}, \dots, a_{i_q i_q}$ 为枢轴的 G 运算之后, 所得方阵的 $(p+1, p+1)$ 元是以 x_{i_1}, \dots, x_{i_q} 为自变量的子集回归的残差平方和

从定理 7.3 我们看到, 和 S 运算不同, G 运算不能同时算出对应的回归系数的 LS 估计. 因此如采用 G 运算, 工作要分两步走. 第一步, 按某种变量选择准则求出“最优”自变量子集. 第二步, 就选出的“最优”自变量子集, 求出对应的 LS 估计.

另外, 由 (7.8) 定义的矩阵 A 的对称性, 在进行 S 或 G 运算时, 只需对主对角线及其上方的元素进行计算, 这可以节省相当大的计算量.

§3.8 所有可能子集回归

前已指出, 计算方法是回归自变量选择中一个十分重要的问题. 对 p 个自变量的线性回归问题, 所有可能的回归有 2^p 个. 从这 2^p 个回归中根据某种变量选择标准, 选出一个或几个最优的回归 (下称最优子集回归), 是本世纪 60—70 年代应用统计工作者十分

关注的问题，并且在这一时期内，比较彻底地解决了这个问题。在60年代提出的一些算法，基本上只能处理含10—12个自变量的回归问题。而Furnival和Wilson[38]提出的算法较完美地解决了节省计算量、存储量以及减少计算误差的问题，它可以计算含30多个自变量的所有可能的子集回归，而所需的计算时间与逐步回归大体相当。

Furnival-wilson程序所用的变量选择标准有三个： C_p 统计量、修正复相关系数的平方 \bar{R}_q （等价于 RMS_q ，见§3.3）和复相关系数的平方 R_q 。当然，象前面所指出的，将这些程序稍加修正，也可适用于其它变量选择准则。这里需要说明，如何把 R_q 用作变量选择准则。从 R_q 的定义

$$R_q = 1 - \frac{RSS_q}{TSS}$$

容易看出，“ R_q 最大”等价于“残差平方和最小”。根据前面的讨论知，若以“ R_q 愈大愈好”为原则，势必导致选择全部自变量。因此在应用 R_q 作标准时，一些人提出，在限定所选自变量个数 q 的前提下，选使 R_q 达到最大的自变量子集。另外一种考虑是，虽然 R_q 随着 q 增大而增大，但 q 增大到一定程度时， R_q 就增加比较缓慢。这表明此时新增加的自变量的贡献比较小。在选择变量时，就选择使 R_q 的曲线达到平缓的最小 q 。Furnival-wilson程序对用户事先确定的 k ，对 $q=1, 2, \dots, p$ 分别打印出最优的 k 个子集回归。当 $k \geq C_q$ 时，就打出全部的子集回归。

Furnival和wilson设计了三种计算程序：二进制、自然式和字典式。它们的共同点是利用 S 、 G 两种运算。不同点仅在于子集回归的计算次序上。下面是这三个计算所有可能子集回归的计算程序中的核心部分。

BINARY PROGRAM(二进制)

```

DO 1 L=1, P
  1 NK(L)=0
  NK(P+1)=1
  L=1
2  NK(L)=1
  DO 3 M=L, K
    IF(NK(M+1).EQ.1)GO TO 4
3  CONTINUE
4  CALL GAUSS(P-M+1, P-L+2, P-L+1, A, P+
  1)
  WRITE(6, )A(P-L+2, P+1, P+1), (NK(N), N
    =1, P)
  DO 5 L=1, P
    IF(NK(L).EQ.0)GO TO 2
5  NK(L)=0
  STOP

```

其中GAUSS(·, ·, ·, ·)表示Gauss运算的子程序.按照这个程序的计算方式,子集回归的计算次序是按二进制排列的.即它们所含的自变量的足标依次是1, 2, 12, 3, 13, 23, 123, 4, 14, 24, 124, 34, 134, 234, 1234等.数组NK是一个二进制计数器.从NK(I), $I=1, 2, \dots, p$ (p 为自变量总数), 可以看出计算的是哪一个子集回归.例如, $p=5$, 而

$$\{NK(I), I=1, \dots, 5\} = \{01001\}$$

则表明计算的是含自变量 x_2 、 x_5 的子集回归.数组NK在此程序中的另一个作用是控制计算次序, 以保证计算是按二进制次序进行.当NK中全为1时, 表示算出来的是全模型, 计算就停止.

在计算开始前, 先把按(7·8)算出的 A 存入 $A(1, 0, 0)$ 中.但

是, 需要特别说明的是, 在使用这个子程序时, 自变量排列次序完全是倒过来的, 即按 x_p, x_{p-1}, \dots 的排列次序去计算 A . 也就是说在 A 的 $(1, 1)$ 元放 $\sum_{i=1}^n (x_{pi} - \bar{x}_p)^2$, 第 $(1, 2)$ 元放 $\sum_{i=1}^n (x_{p-1,i} - \bar{x}_{p-1})(x_{pi} - \bar{x}_p)$ 等等。只有这样, 计算出来的子集回归的次序才是二进制的。因为我们用的是 Gauss 消去法, 所以此子程序只计算残差平方和。一旦选定了某个自变量子集, 再另外计算该子集的回归系数及其它统计量。

Furnival 和 wilson 设计的另一个子程序是, 把计算出的子集回归按字典列排列

LEXICOGRAPHIC PROGRAM(字典式)

IND(1)=0

M=1

1 M=M+1

IND(M)=IND(M-1)+1

2 CALL SEMI(IND(M-1)+1, IND(M)+1, IND(M),
A, P+1, IND, 2, M-1)

WRITE(6,) (IND(L), A(IND(M)+1, IND(L), P+1), L=2, M)

A(IND(M)+1, P+1, P+1)

IF(IND(M).LT.P)GO TO 1

M=M-1

IND(M)=IND(M)-1

IF(M.GT.1)GO TO 2

STOP

其中子程序 SEMI 为 S 运算程序, 因为 (7·8) 定义的方阵 A 为对称阵, 运算只需对 A 的主对角线上方完成, 所以此子程序取名为

“SEMI”，即“一半”的意思。

按这个子程序，计算出的子集回归次序是按字典式排列的。例如，若有4个自变量即 $P=4$ ，那末计算次序是1, 12, 123, 1234, 124, 13, 134, 14, 2, 23, 234, 24, 3, 34, 4。

Furnival和Wilson设计的第三个子程序是所谓自然式程序：

NATURAL PROGRAM(自然式)

DO 1 L=1, P

1 IND(L)=0

M=P

IB=0

IS=1

2 IB=MOD(IB, MAX)+1

DO 3 L=M, P

IF(IND(L).LT.L)GO TO 3

IND(L-1)=IND(L-1)+1

IND(L)=IND(L-1)

3 CONTINUE

4 IND(P)=IND(P)+1

IS=MOD(IS, MAX)+1

CALL SEMI(IB, IS, IND(P), A, P+1, IND, M,
P-1)

WRITE(6,)(IND(L), A(IS, IND(L), P+1), L=
M, P)

IF(IND(P).LT.P)GO TO 4

IS=IS-1

IF(IND(M).EQ.M) M=M-1

IF(M.GT.0)GO TO 2

STOP

按照这个程序，计算出的子集回归次序是按自变量下标的自然大小顺序排列的。例如，对 4 个自变量，其次序为 1, 2, 3, 4, 12, 13, 14, 23, 24, 34, 123, 124, 134, 234, 1234。

这些程序的计算次序的安排是很巧妙的。为了形象地说明这一点，我们引进回归树的概念。图 3·8·1 是 4 个自变量的回归树。图中每个顶点表示一个子集回归。圆点后面的数字表示已选入变量的足标，圆点前面的数字表示可能进入而又尚未进入回归的自变量足标。例如，树根 1234. 就表示只包含常数项的回归。顶点 234.1 则表示只含自变量 x_1 的子集回归。又如顶点 4.13 则表示含自变量 x_1 和 x_3 的回归，圆点前面的 4 表示沿这个顶点往下，进入回归的自变量只可能是 x_4 。

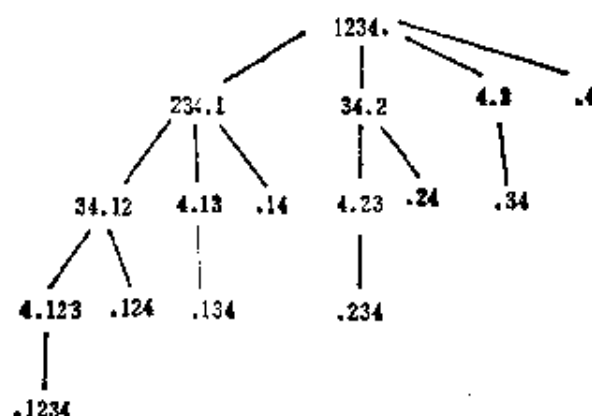


图 3·8·1 回归树

字典式程序是“垂直”地一支一支地计算回归树各顶点所表示的子集回归。从图 3·8·1 看到，最左边一支诸顶点的圆点后数字依次为 1, 12, 123, 1234，紧挨的第二支对应的数字为 124，再过来一支的诸顶点对应的数字为 13, 134, …。这个次序正是前面所说的字典式的计算次序。如果把回归树视为家族的家谱，则字典式是从树根即“祖父”开始，沿“父亲”到“长子”方向，一直访

问到一支的端点，然后转到“父亲”的“大弟弟”，依次下去。若“父亲”的“兄弟”都已访问完，则转到“祖父”的“大弟弟”。按照这样的次序，当所有的支都访问完时，计算结束。

自然式程序是按回归树的顶点水平位置自左至右自上至下的次序计算所有可能回归。例如，在图3·8·1中，除了树根以外，最上边第一行各顶点圆点后的数字自左至右依次是1，2，3，4。接下来第二行为12，13，14，23，24，34。第三行为123，124，134，234，最后一行为1234。这正是前面所说的自然式的计算次序。

前面我们所讨论的只是计算机内部的计算次序。而计算机的输出次序是按选入回归的自变量个数 $q=1, 2, \dots, p$ 排列。对给定的 q ，再按某变量选择准则，依次打出前 k 个(用户可预先确定)最优子集回归。

例8·1 Gorman-Toman车辙数据

Gorman和Toman于1966年提出的一组数据，是含五个因素(自变量)的一个回归问题。因变量为

$$Y = \ln(\text{车辙深度改变量/车轮碾过百万次}).$$

自变量为

$$x_1 = \ln(\text{沥青粘度}),$$

$$x_2 = \text{路面沥青的比例},$$

$$x_3 = \text{路底部沥青比例},$$

$$x_4 = \text{示性变量, 用以把数据分成两组},$$

$$x_5 = \text{路面好坏的一种指标}$$

$$x_6 = \text{路面空疏比例}$$

共有31组原始数据，列在表8·1。对这个问题，共有 $2^6 = 64$ 个子集回归。表7·2是按二进制次序计算出的所有可能子集回归的自变量及对应的 C_p 统计量。最优子集为 x_1, x_2, x_6, x_4 ，对应的最小 $C_p = 4.26$ 。

表8.1 Gorman-Toman车辙数据

观测值 序号	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	0.32930	0.44716	4.68	4.87	-1.00	8.4	4.916
2	1.11394	0.14613	5.19	4.50	-1.00	6.5	4.593
3	1.16879	0.14613	4.82	4.73	-1.00	7.9	5.321
4	1.10037	0.51851	4.85	4.76	-1.00	8.3	4.865
5	0.91645	0.23045	4.86	4.95	-1.00	8.4	3.776
6	1.02816	0.46240	5.16	4.45	-1.00	7.4	4.397
7	0.86213	0.56820	4.82	5.05	-1.00	6.8	4.867
8	1.10278	0.23045	4.86	4.70	-1.00	8.6	4.828
9	1.09968	-0.03621	4.78	4.84	-1.00	6.7	4.865
10	1.31387	-0.16749	5.16	4.76	-1.00	7.7	4.034
11	0.55388	0.77815	4.57	4.82	-1.00	7.4	5.450
12	0.84510	0.63347	4.61	4.65	-1.00	6.7	4.853
13	1.41830	-0.22185	5.07	5.10	-1.00	7.5	4.257
14	1.06707	0.25537	4.66	5.09	-1.00	8.2	5.144
15	0.88479	0.77815	5.42	4.41	-1.00	5.8	3.718
16	1.08814	0.64345	5.01	4.74	-1.00	7.1	4.715
17	-0.11919	1.94448	4.97	4.66	1.00	6.5	4.625
18	0.13033	1.79239	4.01	4.72	1.00	8.0	4.977
19	0.15836	1.69897	4.96	4.90	1.00	6.8	4.322
20	0.20412	1.76343	5.20	4.70	1.00	8.2	5.087
21	0.04139	1.95424	4.80	4.60	1.00	6.6	5.971
22	0.07058	1.81954	4.98	4.69	1.00	6.4	4.647
23	0.07918	2.14613	5.35	4.76	1.00	7.3	5.115
24	0.25181	2.38021	5.04	4.80	1.00	7.8	5.939
25	0.14267	2.62325	4.80	4.80	1.00	7.4	5.916
26	0.32790	2.69897	4.83	4.60	1.00	6.7	5.471
27	0.48149	2.25527	4.66	4.72	1.00	7.2	4.602

观测值 序号	Y	X_1	X_2	X_3	X_4	X_5	X_6
28	0.58503	2.43136	4.67	4.50	1.00	6.3	5.043
29	0.11919	2.23045	4.72	4.70	1.00	6.8	5.075
30	0.09691	1.99123	5.00	5.07	1.00	7.2	4.334
31	0.30102	1.54407	4.70	4.80	1.00	7.7	5.705

表8·2 Gorman-Toman车轱数据的所有可能子集回归(二进制次序)

变量子集	q	C_p	变量子集	q	C_p
常数项	1	835.6	x_5	2	748.8
x_1	2	20.43	x_1x_5	3	21.68
x_2	2	817.0	x_2x_5	3	716.1
x_1x_2	3	13.98	$x_1x_2x_5$	4	14.29
x_3	2	806.7	x_3x_5	3	748.3
x_1x_3	3	21.90	$x_1x_3x_5$	4	22.53
x_2x_3	3	770.0	$x_2x_3x_5$	4	709.0
$x_1x_2x_3$	4	15.95	$x_1x_2x_3x_5$	5	16.17
x_4	2	92.0	x_4x_5	3	84.2
x_1x_4	3	20.53	$x_1x_4x_5$	4	21.46
x_2x_4	3	70.23	$x_2x_4x_5$	4	57.56
$x_1x_2x_4$	4	12.22	$x_1x_2x_4x_5$	5	11.32
x_3x_4	3	88.6	$x_3x_4x_5$	4	85.1
$x_1x_3x_4$	4	22.42	$x_1x_3x_4x_5$	5	22.86
$x_2x_3x_4$	4	58.38	$x_2x_3x_4x_5$	5	53.92
$x_1x_2x_3x_4$	5	13.51	$x_1x_2x_3x_4x_5$	6	13.24

(续上表)

变量子集	q	C_p	变量子集	q	C_p
x_0	2	692	$x_5 x_6$	3	575.5
$x_1 x_6$	3	20.00	$x_1 x_5 x_6$	4	21.87
$x_2 x_0$	3	693.7	$x_2 x_5 x_6$	4	577.4
$x_1 x_2 x_6$	4	5.67	$x_1 x_2 x_5 x_6$	5	7.45
$x_3 x_6$	3	666.8	$x_3 x_5 x_6$	4	577.3
$x_1 x_3 x_6$	4	21.37	$x_1 x_3 x_5 x_6$	5	22.84
$x_2 x_3 x_6$	4	668.0	$x_2 x_3 x_5 x_6$	5	579.1
$x_1 x_2 x_3 x_6$	5	7.56	$x_1 x_2 x_3 x_5 x_6$	6	9.43
$x_4 x_6$	3	92.8	$x_4 x_5 x_6$	4	82.7
$x_1 x_4 x_6$	4	20.61	$x_1 x_4 x_5 x_6$	5	22.24
$x_2 x_4 x_6$	4	70.21	$x_2 x_4 x_5 x_6$	5	59.09
$x_1 x_2 x_4 x_6$	5	4.26	$x_1 x_2 x_4 x_5 x_6$	6	5.51
$x_3 x_4 x_6$	4	89.36	$x_3 x_4 x_5 x_6$	5	83.9
$x_1 x_3 x_4 x_6$	5	22.36	$x_1 x_3 x_4 x_5 x_6$	6	23.65
$x_2 x_3 x_4 x_6$	5	56.83	$x_2 x_3 x_4 x_5 x_6$	6	54.38
$x_1 x_2 x_3 x_4 x_6$	6	5.31	$x_1 x_2 x_3 x_4 x_5 x_6$	7	7.00

请注意，这个最优子集含示性变量 x_4 。从表8.1我们看到，它只取+1或-1。引进这个示性变量的目的仅仅出于某种原因把数据分成两类。如果把这个最优子集回归用于预测的话， x_4 无法取值，这是此最优子集回归的缺点。对这个问题一个处理方法是，考察原始数据的获得过程，搞清为什么要把数据分成两组，以决定如何对待 x_4 。另一种方法是从不含 x_4 的变量子集中选 C_p 最小者，这时最优子集为 x_1, x_2, x_6 ，对应的 $C_p=5.67$ 。

另外一种求最优子集回归的方法是基于所谓“分支定界法”(the branch-bound method)，即将自变量按某种原则(如回归

树的支)分成若干组(即所谓“支”),设 A 、 B 为其中的两组。若它们的残差平方和满足 $RSS_A \leq RSS_B$,则因 B 的一切子集的残差平方和不会再比 RSS_B 小,因此变量组 B 的一切可能子集回归就不需要再计算了。这样 RSS_A 就作为一个“界”,凡是残差平方和比它大的变量组,其子集回归不会是最优的,就不必计算。按照这个程序显然要节省不少计算量。关于这类方法的详细讨论可以在[19]中找到。

§3.9 逐步型回归

上节我们讨论了通过计算所有可能子集回归,寻求最优子集回归的方法。虽然Furnival-wilson等算法设计很巧妙,计算量相当省。但对自变量特别多(例如超过30个)的大型回归问题,计算量仍然是很大的。目前有另外一些不计算所有可能子集回归的变量选择算法,其中应用最普遍的是所谓逐步型回归法。这些方法分三种:向前法,向后法和逐步法。它们的共同点是每一步只引入或删除一个自变量。所以都归入逐步型方法。关于这些方法的计算细节容易在其它回归分析著作中找到,如[1]、[42],此处不予详细叙述。下面我们仅作一般性讨论。

(一) 向前法

这个方法是从回归方程仅含常数项开始,把自变量逐个引入回归方程。第一步,把与因变量 Y 有最大简单相关系数的变量,作回归系数的显著性检验,若它显著地异于零,则把该自变量选入方程。而后在余下的自变量中,考虑在消除了已选入变量的影响之后,对与 Y 有最大相关系数(即偏相关系数)的变量,作回归系数显著性检验,如用(1·3·22)定义的 F -统计量,以决定是否选

入。这样做下去，一直到在排除已选入变量对 Y 的影响之后，未选入变量对 Y 的回归系数的检验都不显著为止。

这个方法有一个明显的缺点，就是由于各自变量之间可能存在着相关关系，因此后续变量的选入可能会使前面已选入的自变量变得不重要。这样最后得到的“最优”回归方程可能包含一些对 Y 影响不大的自变量。

例9.1 Hald水泥数据(续例3.2)

对例3.2讨论过的Hald水泥数据，应用向前法的结果如表9.1。这里作 F 检验时，显著性水平取为 $\alpha = 0.10$ 。用向前法得到的回归方程为

$$Y = 71.6483 + 1.4519x_1 + 0.4161x_2 - 0.237x_4$$

从例3.2看到，若以平均残差平方和 RMS_q 为标准，该回归方程是最优的。而由例4.1，在 C_p 准则下，它也是接近最优的。

表9.1 Hald水泥问题的向前法

步号	所选变量	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_4$
1	x_4	117.5679			-0.7382
2	x_1x_4	103.0974	1.4400		-0.6140
3	$x_1x_2x_4$	71.6483	1.4519	0.4161	-0.2365

(二) 向后法

向后法与向前法正好相反。它是先将全部变量选入回归模型，即从全模型开始，然后逐个剔除对RSS贡献较小的变量。这里，贡献大小可以用(1.3.22)定义的 F -统计量来度量。若一开始，所有自变量的 F -统计量经检验后都显著，则“最优”回归方程就是全模型。不然，若有若干个 F 值不显著，则剔除具有最小 F 值的变量。然后对剩下的变量建立新的回归方程。重复这个过程，直到剩下的自变量都不能剔除为止。

例9.2 Hald水泥数据(续例3.2)

如果对例3.2的Hald问题应用向后法, 取检验的显著性水平 $\alpha=0.10$, 计算结果列在表9.2. 从例4.1看到, 这里用向后法选到的“最优”回归方程在 C_p 准则下是最优的, 而且它的 RMS_e 也比较小。对应的“最优”回归方程为

$$Y = 52.5733 + 1.4683x_1 + 0.6623x_2$$

表9.2 Hald水泥问题的向后法

步号	所选变量	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
1	$x_1x_2x_3x_4$	62.4054	1.5511	0.5102	0.1019	-0.1441
2	$x_1x_2x_4$	71.6483	1.4519	0.4161		-0.2365
3	x_1x_2	52.5773	1.4683	0.6623		

(三) 逐步法

逐步法即通常所称的逐步回归, 本质上是向前法, 但吸收了向后法的作法。刚才已经指出, 向前法有一个缺点, 就是后续变量的引入会使得一些已在回归方程的自变量变得不重要。因此在逐步法中, 在每一步增加了对已选入变量的显著性检验。即在每一步, 经 F -检验选择进入方程的变量, 而后又作 F -检验, 看是否要剔除某些自变量。这个过程一直进行到既没有变量需要引入, 也没有变量要剔除为止。

例9.3 Hald水泥数据(续例9.1和例9.2)

把逐步法应用于Hald数据, 计算的情况是: 先引入 x_4 , 然后引入 x_1 , 接下来引入 x_2 . 这和向前法一样。这几步都没有剔除任何变量。在含 x_1, x_2, x_4 的方程中, 经检验需剔除 x_4 . 剔除 x_4 后也没有什么变量需要再剔除。最后得到的“最优”回归方程与向后法相同。

在应用上, 逐步型方法面临着的一个较大的困难是 F -检验的

显著性水平 α 的选择。 α 选得太大了,最后所得到的方程含较多的自变量,相反,方程所含的自变量则偏少。另外,人们对逐步型方法的批评也多与所采用的 F -检验的合理性有关。事实上在每一步,我们是在一组相关的 F -变量中找出最大值或最小值作 F -检验。直观上,供选择的自变量愈多,所找出的最大值(或最小值)也就愈大(或小)。显然除了一些极端情况之外,这些量并不服从 F 分布。因而并不能保证所挑选出的回归方程在某种准则下是好的或较好的。但是,从长期应用实践看,一般情况下逐步型方法所给出的回归方程还比较好,加之计算量少,又有较成熟的计算程序,因而到目前为止它们仍是被广泛使用的变量选择计算方法。

对目前一般所使用的电子计算机,如果回归自变量有20-30个左右,那末上节介绍的Furnival-Wilson算法所需时间与逐步型法大致差不多。然而,如果回归自变量多于30,一般可采用如下两阶段法。第一阶段先用逐步型法筛选出一部分重要自变量。第二阶段对已选出的这些数量较少的自变量,用Furnival-Wilson算法,计算出所有可能子集回归,从中找出最优子集回归。

在结束这一节时,我们再介绍一种简单有效的变量选择法—— **t -直接法**,它似乎也可以算作一种逐步型方法。

从(1.3.22)知,对于全模型检验假设 $H_0: \beta_j=0$ 的 t 检验统计量为

$$t_{p,j} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{c_{jj}}}$$

这里 $t_{p,j}$ 的下标 p 表示全模型所含的自变量个数, c_{jj} 表 $\tilde{X}'\tilde{X}$ 的第 $j+1$ 个对角元,即 $\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj}$, $\hat{\sigma}^2$ 为从全模型得到的 σ^2 的估计。对全模型贡献大的自变量,对应的 $|t_{p,j}|$ 应该比较大。因此,在某种准则下,最优或较优的子集回归应该包含这样的自变量。

所谓 t -直接法,就是对全模型的每个自变量计算 $|t_{p,i}|$,将自变量按 $|t_{p,i}|$ 的大小排序,然后依 $|t_{p,i}|$ 由大到小的次序将对应的自变量引入方程,每次引入一个自变量。我们可以期望用 t -直接法产生的子集回归在某种准则下是最优或接近最优的。

例9.4 Hald水泥数据

下表是对Hald水泥数据用 t -直接法的计算结果.对 $q=2$,即含一个自变量的回归, t -直接法给出的是自变量 x_1 .从表4.2看到,在四个含一个自变量的回归中, x_1 的秩是3,并不算好。而对 $q=3$, t -直接法给出的变量子集为 $x_1, x_2, C_p=2.68$,它不仅是所有含2个变量的子集回归中 C_p 最小的,而且也是在 C_p 准则下最优子集。对 $q=4$,结果也很好。这些结果从比较表9.3和表4.2很容易看出。

表9.3 Hald水泥数据 t -直接法计算结果

自变量	$ t_{p,i} $	回归方程中的自变量	q	C_p
x_1	2.08	x_1	2	202.55
x_2	0.70	x_1, x_2	3	2.68
x_4	0.20	x_1, x_2, x_4	4	3.02
x_3	0.14	x_1, x_2, x_3, x_4	5	5.00

第四章 回归系数的有偏估计

在第一章我们证明了回归系数的LS估计具有许多优良性质,其中特别重要的是Gauss-Markov定理,它保证了LS估计在线性无偏估计类中的方差最小性。如果进一步假设误差服从正态分布,那末LS估计还具有更多更好的性质。基于这个原因,从上世纪初Gauss创立最小二乘法以来,在线性回归模型乃至更一般的线性模型参数估计中,LS估计一直是唯一被广泛应用的重要估计。然而,随着现代电子计算技术的飞速发展,人们愈来愈多地处理含较多自变量的大型回归问题。许多应用实践表明,在一些情况下LS估计并不很理想,在个别情况下可能很不好。从本世纪五十年代特别是六十年代以来,统计学家们作了种种努力,试图改进LS估计。前面两章所讨论的回归诊断和自变量选择都可以算作这种努力的一部分。它们都是从模型或数据角度考虑问题的。这种努力的另外一个重要方面就是寻求一些新的估计。Stein于1955年证明了,当维数 p 大于2时,正态均值向量LS估计的不可容许性。即能够找到另外一个估计在某种意义下一致优于LS估计^[8]。有些文献称此为Stein现象。以此为开端,在后来的二十年中,人们提出了许多新的估计。其中主要有岭估计、Stein估计、主成分估计以及特征根估计等。这些估计有一个共同点,就是有偏性,即它们的均值并不等于待估参数,于是人们把这些估计统称为有偏估计。从某种意义上讲,这些估计都改进了LS估计。本章的目的

是系统地讨论这些估计的性质和应用。

§4.1 复共线性

一个首要问题是,研究在什么情况下LS估计的性质才明显地变坏。为此我们先引进度量估计优劣的另一个标准——均方误差。

对于一般的参数估计问题,设 θ 为 $p \times 1$ 未知参数向量, $\tilde{\theta}$ 为它的某种估计。 $\tilde{\theta}$ 的均方误差(Mean Square Error, 简记为MSE)

$$\begin{aligned} \text{MSE}(\tilde{\theta}) &= E\|\tilde{\theta} - \theta\|^2 \\ &= E(\tilde{\theta} - \theta)'(\tilde{\theta} - \theta) \end{aligned} \quad (1.1)$$

度量了估计 $\tilde{\theta}$ 与未知参数 θ 偏离的大小。一个好的估计应该有较小的均方误差。为了进一步说明 $\text{MSE}(\tilde{\theta})$ 的意义,我们把它作如下分解

$$\begin{aligned} \text{MSE}(\tilde{\theta}) &= E(\tilde{\theta} - \theta)'(\tilde{\theta} - \theta) \\ &= E[(\tilde{\theta} - E\tilde{\theta}) + (E\tilde{\theta} - \theta)]'[(\tilde{\theta} - E\tilde{\theta}) + (E\tilde{\theta} - \theta)] \\ &= E(\tilde{\theta} - E\tilde{\theta})'(\tilde{\theta} - E\tilde{\theta}) + \|E\tilde{\theta} - \theta\|^2 \\ &= \text{trCOV}(\tilde{\theta}) + \|E\tilde{\theta} - \theta\|^2 \\ &\triangleq \Delta_1 + \Delta_2 \end{aligned} \quad (1.2)$$

若记 $\tilde{\theta}' = (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)$, 则(1.2)中的第一项

$$\Delta_1 = \text{trCOV}(\tilde{\theta}) = \sum_{i=1}^p \text{Var}(\tilde{\theta}_i) \quad (1.3)$$

它是估计 $\tilde{\theta}$ 的各分量方差和。第二项

$$\Delta_2 = \|E\tilde{\theta} - \theta\|^2 = \sum_{i=1}^p (E\tilde{\theta}_i - \theta_i)^2 \quad (1.4)$$

它是估计 $\tilde{\theta}$ 的各分量偏差的平方和。要均方误差小，必须 Δ_1 和 Δ_2 都比较小。

现在我们来计算LS估计的均方误差。考虑线性回归模型

$$Y = a\mathbf{1} + X\beta + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I, \quad (1.5)$$

这里 $n \times p$ 的设计阵 X 假定已经中心标准化，且 $R(X) = p$ 。于是 a 的LS估计 $\hat{a} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ ， β 的LS估计为 $\hat{\beta} = (X'X)^{-1}X'Y$ 。以下

我们只讨论回归系数 β 的LS估计改进问题。

因为 $\hat{\beta}$ 是 β 的无偏估计，于是在 $\text{MSE}(\hat{\beta})$ 中， $\Delta_2 = 0$ 。由 $\text{COV}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ ，从(1.2)有

$$\text{MSE}(\hat{\beta}) = \Delta_1 = \sigma^2 \text{tr}(X'X)^{-1} \quad (1.6)$$

记 $\lambda_1 \geq \dots \geq \lambda_p > 0$ 为 $X'X$ 的特征根。因为 $X'X$ 可逆，所以 $(X'X)^{-1}$ 的特征根为 λ_1^{-1} ， λ_2^{-1} ， \dots ， λ_p^{-1} 。故上式变为

$$\text{MSE}(\hat{\beta}) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \quad (1.7)$$

由此可以看出，如果 $X'X$ 至少有一个特征根非常小，即非常接近于零，那末 $\text{MSE}(\beta)$ 就很大。这时尽管 Gauss-Markov 定理保证了 $\Delta_1 = \sigma^2 \text{tr}(X'X)^{-1}$ 在线性无偏估计类中很小，但它本身的值却很大。在这种情况下， $\hat{\beta}$ 不再是 β 的一个良好估计。

下面我们进一步分析， $X'X$ 至少有一个特征根很小在设计阵 X 本身意味着什么？

如果 $X'X$ 至少有一个特征根很小，即很接近于零，我们则称设计阵 X 呈病态。由上面的讨论知，设计阵 X 呈病态时，LS估计的性质变得不好。现在我们证明，当 X 呈病态时， X 的列向量之间存在近似的线性关系。事实上，记 $X = (X_1, \dots, X_p)$ ，即 X_i ， $i=1, \dots, p$ 为 X 的 p 个列向量。若 c 为对应于 $X'X$ 的特征根 λ 的标准正交化特征向量，且 $\lambda \approx 0$ ，则

$$X'Xc = \lambda c \approx 0.$$

上式两边左乘 c' , 得到

$$c'X'Xc \approx 0,$$

从而有

$$Xc \approx 0. \quad (1.8)$$

若记 $c' = (c_1, \dots, c_p)$, 上式即为

$$c_1X_1 + c_2X_2 + \dots + c_pX_p \approx 0 \quad (1.9)$$

这说明, X 的列向量 X_1, \dots, X_p 之间有近似线性关系 (1.9). 通常称 (1.9) 为复共线关系。相应地, 称设计阵 X 或线性回归模型 (1.5) 存在复共线性 (Multi-Collinearity). 可见, 导致 LS 估计性质变坏的原因是复共线性。

关于复共线性的诊断以及复共线性严重程度的度量, 是近年来十分引人注目的研究课题, 已经提出了一些行之有效的方法。归纳起来主要有如下三种。

(1) 特征分析法

假设 X 呈病态, 则 $X'X$ 至少有一个特征根很接近于零。不妨设后 $p-r$ 个特征根 $\lambda_{r+1}, \dots, \lambda_p \approx 0$, 记 $\varphi_{r+1}, \dots, \varphi_p$ 为与它们对应的标准正交化特征向量。由 (1.8) 知对这些 φ_i , 有

$$X\varphi_i \approx 0, \quad i = r+1, \dots, p$$

若令 $\varphi_i' = (\varphi_{i1}, \dots, \varphi_{pi})$, 则有

$$\varphi_{i1}X_1 + \dots + \varphi_{pi}X_p \approx 0, \quad i = r+1, \dots, p \quad (1.10)$$

这是 $p-r$ 个复共线关系。可见, $X'X$ 有多少个特征根非常接近于零, 设计阵 X 就有多少个复共线关系, 并且这些复共线关系的系数向量就是接近于 0 的那些特征根对应的特征向量。

特征分析法计算简单, 且很容易地给出了全部的复共线关系。但是, $X'X$ 的特征根 “很接近于零” 是一个很模糊的说法, 在应

用上究竟什么样的数算是“很接近于零”较难于掌握。下面介绍的条件数在这方面会有一些帮助。

(2) 条件数

方阵 $X'X$ 的条件数(Condition Number)定义为

$$k = \frac{\lambda_1}{\lambda_p}$$

直观上,条件数度量了 $X'X$ 的特征根散布程度,可以用来判断复共线性是否存在以及复共线性严重程度。在应用经验中,若 $0 < k < 100$,则认为没有复共线性;若 $100 \leq k \leq 1000$,则认为存在中等程度或较强的复共线性;若 $k > 1000$,则认为存在严重的复共线性。

例1.1 Webster-Gunst-Mason数据

Webster、Gunst和Mason在[43]研究回归系数的特征根估计时(见

表1.1 Webster-Gunst-Mason非中心标准化数据

数据号	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	10.006	8.000	1.000	1.000	1.000	0.541	-0.099
2	9.737	8.000	1.000	1.000	0.000	0.130	0.070
3	15.087	8.000	1.000	1.000	0.000	2.116	0.115
4	8.422	0.000	0.000	9.000	1.000	-2.397	0.252
5	8.625	0.000	0.000	9.000	1.000	-0.046	0.017
6	16.289	0.000	0.000	9.000	1.000	0.365	1.504
7	5.958	2.000	7.000	0.000	1.000	1.996	-0.865
8	9.313	2.000	7.000	0.000	1.000	0.228	-0.055
9	12.960	2.000	7.000	0.000	1.000	1.380	0.502
10	5.541	0.000	0.000	0.000	10.000	-0.798	-0.399
11	8.756	0.000	0.000	0.000	10.000	0.257	0.161
12	10.937	0.000	0.000	0.000	10.000	0.440	0.432

§4.6), 构造了下面的例子。此例含六个自变量 x_1, x_2, \dots, x_6 和因变量 Y , 表1.1给出了原始数据, 共有12组数据, 除第一组外, 其余11组数据满足线性关系

$$x_1 + x_2 + x_3 + x_4 = 10 \quad (1.11)$$

将数据中心标准化, 并乃用 x_1, \dots, x_6 表示. 设 e_1, e_2, \dots, e_{12} 为从正态随机数表随机查出的12个数, 则 e_1, \dots, e_{12} 为来自 $N(0, 1)$ 的独立样本。根据关系

$$Y_i = 10 + 2.0x_{1i} + 1.0x_{2i} + 0.2x_{3i} - 2.0x_{4i} + 3.0x_{5i} + 10.0x_{6i} + e_i, \quad i = 1, 2, \dots, 12 \quad (1.12)$$

算出因变量 Y 的12个观测值. 这些值列在表1.1的第一列。这样我们得到了一个真正的正态线性回归模型。对于此模型, $X'X$ 为

$$\begin{bmatrix} 1.000 & 0.052 & -0.343 & -0.498 & 0.417 & -0.192 \\ & 1.000 & -0.432 & -0.371 & 0.485 & -0.317 \\ & & 1.000 & -0.355 & -0.505 & 0.494 \\ & & & 1.000 & -0.215 & -0.087 \\ & & & & 1.000 & -0.123 \\ & & & & & 1.000 \end{bmatrix}$$

它的六个特征根分别为

$$\lambda_1 = 2.24879, \lambda_2 = 1.54615, \lambda_3 = 0.92208,$$

$$\lambda_4 = 0.79399, \lambda_5 = 0.30789, \lambda_6 = 0.00111.$$

条件数为

$$k = \frac{\lambda_1}{\lambda_6} = \frac{2.24879}{0.00111} = 2025.94$$

因为 $k > 1000$, 所以我们认为模型(1.12)有严重的复共线性。表1.2列出了 $X'X$ 的标准正交化特征向量. 因最小特征根 $\lambda_6 = 0.00111 \approx 0$, 根据(1.10), 以 φ_6 为系数的关系

$$\begin{aligned} & -0.44768x_1 - 0.42114x_2 - 0.54169x_3 - 0.57337x_4 \\ & - 0.00605x_5 - 0.00217x_6 \approx 0 \end{aligned}$$

就是一个复共线关系。注意到 x_5 和 x_6 的系数相对于前面四个系数都十分小,

表1.2 Webster等数据 $X'X$ 的特征向量

φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
-.39072	-.33968	.67980	.07990	-.25104	-.44768
-.45560	-.05392	-.70013	.05769	-.34447	-.42114
.48264	-.45333	-.16078	.19103	.45364	-.54169
.18766	.73547	.13587	-.27645	.01521	-.57337
-.49773	-.09714	-.03185	-.56356	.65128	-.00605
.35195	-.35476	-.04864	-.74818	-.43375	-.00217

把它们略去, 得到

$$-0.44768x_1 - 0.42114x_2 - 0.54169x_3 - 0.57337x_4 \approx 0, \quad (1.13)$$

又因 x_1, x_2, x_3, x_4 的系数比较接近, 因此这个复共线性就体现了我们原来构造数据时所用的关系(1.11). 因为第一组数据并不满足(1.11), 所以(1.13)与(1.11)不尽相同。

(3) 方差扩大因子

复共线性严重程度的另一种度量是方差扩大因子 (Variance Inflation Factor, 简记为VIF). 记 $C = (c_{ij}) = (X'X)^{-1}$, $\sqrt{R_i}$ 为 x_i 对其余 $p-1$ 个自变量的复相关系数, 我们将证明

$$c_{ii} = (1 - R_i)^{-1}, \quad j=1, \dots, p \quad (1.14)$$

若记 $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$, 则 $\text{Var}(\hat{\beta}_i) = \sigma^2 c_{ii}$, 即 c_{ii} 与LS估计 $\hat{\beta}_i$ 的方差仅差一个因子, 或者说 c_{ii} 也是 $\text{Var}(\hat{\beta}_i)$ 的一个因子, 而且是很重要的因子. 文献中称 c_{ii} 为方差扩大因子. 我们知道, R_i 度量了自变量 x_i 与其余 $p-1$ 个自变量线性相依程度. 若这种相依程度愈高, 即自变量之间复共线性愈严重, R_i 就愈接近于1, c_{ii} 也就愈大. 反过来, 若 x_i 与其余 $p-1$ 个自变量线性相依程度愈低, 即复共线性愈弱, R_i 就愈接近于零, c_{ii} 也就愈接近于1 (注意,

$c_{ii} \geq 1$). 可见, c_{ii} 的大小也可以反映出自变量之间是否存在复共线性. 应用上的经验表明, 当 c_{ii} 大于 5 或 10 时, 就有严重的复共线性存在。

现在我们来证明 (1.14). 只证 $j=1$ 的情形. 将设计阵 X 分块, $X = (X_1; X_0)$, 这里 X_1 为 X 的第一列, X_0 为后 $p-1$ 列. 将 X_1 视为因变量 x_1 的 n 次观测, X_0 为其余 $p-1$ 个自变量的 n 次观测. 考虑线性回归模型

$$X_1 = X_0 \gamma_0 + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I,$$

这里 γ_0 为 $(p-1) \times 1$ 回归系数. 对此模型总平方和 $\text{TSS} = \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 = X_1' X_1 = 1$, 此因 X 已标准化. 而回归平方和 $\text{SS}_{\text{回}} = \hat{\gamma}_0' X_0' X_1$, $X_1 = X_1' X_0 (X_0' X_0)^{-1} X_0' X_1$. 依公式 (1.3.14), 有

$$R_1 = \frac{\text{SS}_{\text{回}}}{\text{TSS}} = X_1' X_0 (X_0' X_0)^{-1} X_0' X_1 \quad (1.15)$$

另一方面, 类似于分块矩阵的逆矩阵公式 (2.2.7), 有 (见 [8] 定理 1.2.4)

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} & * \\ * & * \end{pmatrix}$$

利用这个结果, 我们得到

$$\begin{aligned} C = \begin{pmatrix} c_{11} & * \\ * & * \end{pmatrix} &= (X'X)^{-1} = \begin{pmatrix} X_1'X_1 & X_1'X_0 \\ X_0'X_1 & X_0'X_0 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (1 - X_1'X_0(X_0'X_0)^{-1}X_0'X_1)^{-1} & * \\ * & * \end{pmatrix} \end{aligned}$$

这里应用了 $X_1'X_1 = 1$. 结合 (1.15), 得到

$$c_{11} = (1 - R_1)^{-1}.$$

这就证明了我们所要的结果。

例1.2 (续例1.1)

对Webster等的数据, 方差扩大因子如表1.3. 方差扩大因子的最大值 $297.72 \gg 10$, 可见模型存在严重的复共线性。从表1.3看出, $VIF > 10$ 的四个自变量对应于例1.1的复共线关系(1.13)所含的四个自变量。这个事实容易从关系式 $C_{jj} = (1 - R_j)^{-1}$ 及 R_j 的统计意义得到解释。推广到一般情况, 即凡方差扩大因子超过10的自变量, 都含在某个复共线关系中。

表1.3 Webster等数据的方差扩大因子

自变量	x_1	x_2	x_3	x_4	x_5	x_6
方差扩大因子 C_{jj}	182.05	161.36	266.26	297.72	1.92	1.46

复共线性产生的原因是多方面的。一种是由于数据“收集”的局限性所致。直观上, 如果设计阵 X 的 p 个列向量 X_1, \dots, X_p 近似地都落在维数低于 p 的 R^n 的超平面内, 那末自然 X 的列向量就有复共线性。虽然这样产生的复共线性是非本质性的, 原则上可以通过“收集”更多的数据来解决, 但具体实现起来会遇到许多困难。例如, 在一些问题中, 由于试验或生产过程已经完结或经费限制, 不可能再产生新的数据。另一方面, 对一些情况, 虽然客观上可以“收集”更多数据, 但对于多于三个自变量的情况, 往往难于确定“收集”什么样的数据, 才能“打破”复共线性。最后, 即便“收集”了一些新数据, 但为了“打破”复共线性, 这些新数据势必要远离原来数据, 这样它们很可能是高杠杆点或强影响点, 这又会产生新的问题(见第二章)。

另一种产生复共线性的重要情况是, 自变量之间客观上就有近似线性关系。例如, 在第二章例6.1老鼠试验中, 体重 x_1 和服药量 x_3 是成比例的, 如果把 x_1, x_3 都选作自变量, 那么在 x_1 和 x_3 之间

就有一个复共线关系。又如，在研究农村家庭用电问题中，如果把家庭收入 x_1 和住房面积 x_2 都算作自变量，那么因为家庭收入高的住房也相应宽敞一些，在变量 x_1 、 x_2 之间就有复共线性。一般说来，当处理含自变量较多的大型回归问题时，由于人们往往对自变量之间的关系缺乏认识，很可能把一些有复共线关系的自变量引入回归方程。这就是为什么在大型回归问题中，LS估计的性质往往不理想，甚至可能很坏的一个原因。

§4.2 岭估计

岭估计(Ridge estimate)是由Hoerl和Kennard于1970年提出的^[44,45]。自1970年以来，这种估计的研究和应用得到广泛的重视，成为目前最有影响的一种有偏估计。

(一) 定义及性质

对于线性回归模型(1.5)，回归系数 β 的岭估计定义为

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y \quad (2.1)$$

这里 $k > 0$ 称为岭参数或偏参数。如果 k 取与试验数据 Y 无关的常数，则 $\hat{\beta}(k)$ 为线性估计，不然的话， $\hat{\beta}(k)$ 就是非线性估计。取不同的 k ，得到不同的岭估计。所以(2.1)定义了一个很大的估计类。特别，取 $k=0$ ， $\hat{\beta}(0) = (X'X)^{-1}X'Y$ 就是 β 的LS估计。下面我们先研究 k 为常数的情况。

与LS估计 $\hat{\beta}$ 相比，岭估计是把 $X'X$ 换成 $X'X + kI$ 得到的。直观上这样作的理由也是明显的。因为当 X 呈病态时， $X'X$ 的特征根至少有一个非常接近于0，而 $X'X + kI$ 的特征根 $\lambda_1 + k, \dots, \lambda_p + k$ 接近于零的程度就会得到改善，从而“打破”原来设计阵的复共线性，使岭估计比LS估计有较小的均方误差，即 $MSE(\hat{\beta}(k))$

$< \text{MSE}(\hat{\beta})$. 后面我们将证明, 使这个不等式成立的 k 是存在的。

为了研究岭估计的性质以及后面讨论其它估计的需要, 我们引进线性回归模型(1.5)的典则形式。设 $\varphi_1, \dots, \varphi_p$ 为 $X'X$ 对应于特征根 $\lambda_1 \geq \dots \geq \lambda_p$ 的标准正交化特征向量。记 $\Phi = (\varphi_1, \dots, \varphi_p)$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$, 则

$$\begin{aligned} Y &= a_0 \mathbf{1} + X\beta + e \\ &= a_0 \mathbf{1} + Z\alpha + e, \end{aligned}$$

由 $\Phi' \Phi = I$
 $\Phi \Phi' = I$
 (2.2) 正交

这里把(1.5)中的常数项改记为 a_0 , $Z = X\Phi$, $\alpha = \Phi'\beta$. 我们称(2.2)为线性回归模型的典则形式, α 称为典则回归系数。因为 X 已中心化, 所以 Z 也是中心化的。于是对模型(2.2), 常数项 a_0 的LS估计仍为 $\hat{a}_0 = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. 因 $Z'Z = \Phi'X'X\Phi = \Lambda$, 所以从(2.2)导

出的 α 的LS估计为

$$\hat{\alpha} = \Lambda^{-1} Z'Y \quad (2.3)$$

原回归系数 β 的LS估计可表为

$$\hat{\beta} = \Phi \hat{\alpha} \quad (2.4)$$

相应的岭估计分别为

$$\hat{\alpha}(k) = (\Lambda + kI)^{-1} Z'Y \quad (2.5)$$

和

$$\hat{\beta}(k) = \Phi \hat{\alpha}(k) \quad (2.6)$$

因为均方误差在估计和参数的正交变换下保持不变, 所以典则回归系数和原回归系数的LS估计(或岭估计)有相同的均方误差。这个性质对后面其它估计也成立, 这对计算均方误差带来不少方便。

现在我们证明岭估计的一些性质。

(1) $\hat{\beta}(k) = A_k \hat{\beta}$, 这里 $A_k = (X'X + kI)^{-1} X'$ 这表明岭估

计是LS估计的一个线性变换。

(2) $E\hat{\beta}(k) = A_k\beta$, 只要 $A_k \neq I$, 岭估计就是 β 的有偏估计。而 $A_k = I \iff k=0$, 所以岭估计类中除了LS估计之外, 所有估计都是有偏估计。

(3) 对任意 $k>0$, $\|\hat{\beta}\| \neq 0$, 总有 $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$ 。

此因

$$\|\hat{\beta}(k)\| = \|\hat{\alpha}(k)\| = \|(\Lambda + kI)^{-1} \Lambda \hat{\alpha}\| < \|\hat{\alpha}\| = \|\hat{\beta}\|$$

这个性质表明, $\hat{\beta}(k)$ 是对 $\hat{\beta}$ 向原点的压缩。因为

$$\text{MSE}(\hat{\beta}) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = E\hat{\beta}'\hat{\beta} - \beta'\beta$$

利用(1.7), 有

$$E\|\hat{\beta}\|^2 = \|\beta\|^2 + \text{MSE}(\hat{\beta}) = \|\beta\|^2 + \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \quad (2.7)$$

当 X 病态时, 上式第二项很大。所以这时平均说来, LS估计 $\hat{\beta}$ 偏长, 对它作适当压缩是应该的。这个结果从一个侧面说明了岭估计的合理性。

(4) 存在 $k>0$, 使得

$$\text{MSE}(\hat{\beta}(k)) < \text{MSE}(\hat{\beta})$$

即存在 $k>0$, 使得在均方误差意义下, 岭估计优于LS估计。

证明 根据前面的讨论, 我们只需证明, 存在 $k>0$, 使

$$\text{MSE}(\hat{\alpha}(k)) < \text{MSE}(\hat{\alpha})$$

因为

$$\text{COV}(\hat{\alpha}(k)) = \sigma^2(\Lambda + kI)^{-1} \Lambda (\Lambda + kI)^{-1}$$

$$E(\hat{\alpha}(k)) = (\Lambda + kI)^{-1} Z'(\alpha_0 1 + Z\alpha)$$

$$= (\Lambda + kI)^{-1} Z' Z \alpha$$

$$= (\Lambda + kI)^{-1} \Lambda \alpha$$

应用(1.2), 我们得到

$$\text{MSE}(\hat{\alpha}(k)) = \text{tr COV}(\hat{\alpha}(k)) + \|E\hat{\alpha}(k) - \alpha\|^2$$

$$= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

$$\triangleq g_1(k) + g_2(k) \triangleq g(k) \quad (2.8)$$

对 k 求导数, 得

$$g'_1(k) = -2\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} \quad (2.9)$$

$$g'_2(k) = 2k \sum_{i=1}^p \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} \quad (2.10)$$

因为 $g'_1(0) < 0$, $g'_2(0) = 0$, 所以 $g'(0) < 0$. 而 $g'_1(k)$ 和 $g'_2(k)$ 在 $k \geq 0$ 都连续, 所以当 $k > 0$ 且充分小时, $g'(k) = g'_1(k) + g'_2(k) < 0$. 这就说明 $g(k) = \text{MSE}(\hat{\alpha}(k))$ 在 $k > 0$ 充分小时, 随着 k 的增大而减少. 故存在 $k > 0$, 致 $g(k) < g(0)$, 即 $\text{MSE}(\hat{\alpha}(k)) < \text{MSE}(\hat{\alpha})$. 结论得证.

岭估计还具有其它一些性质. 例如, 对任意 $k > 0$, $\hat{\beta}(k)$ 是 β 的可容许估计. 又岭估计是 Bayes 估计等. 这些事实的证明超出本书的范围. 对这些问题感兴趣的读者可参看 [8] 的第六章.

(二) 岭参数的选择

引进岭估计的目的是减少均方误差. 所以我们应该选择使 $\text{MSE}(\hat{\beta}(k))$ 达到最小的 k . 但从 $g'(k) = g'_1(k) + g'_2(k)$ 的表达式, 我们知道 k 的最优值不但依赖于模型未知参数 β, σ^2 , 而且这种依赖关系没有显示表示, 这使得 k 值的确定变得十分困难. 对于这个应用上非常重要的问题, 统计学家们作了很多工作, 提出了十余种选 k 的方法. 但是到目前为止还没有一个方法能够一致地优于其它方法. 下面我们介绍其中的几种方法.

(1) 岭迹法

岭估计 $\hat{\beta}(k) = (X'X + kI)^{-1} X'Y$ 的分量 $\hat{\beta}_i(k)$ 作为 k 的函

数, 当 k 在 $[0, +\infty)$ 变化时在平面直角坐标系所描出的图形称为岭迹(ridge trace). 选择 k 的岭迹法是: 将 p 个分量 $\hat{\beta}_i(k)$ 的岭迹画在同一个图上, 如图 4.2.1. 选择 k 使得各回归系数的岭估计大体稳定, 并且兼照回归系数没有不合理的符号, 残差平方和上升不太多等. 在图 4.2.1 中, 在 k^* 附近三条岭迹就大体稳定了, 可以考虑取 $k = k^*$.

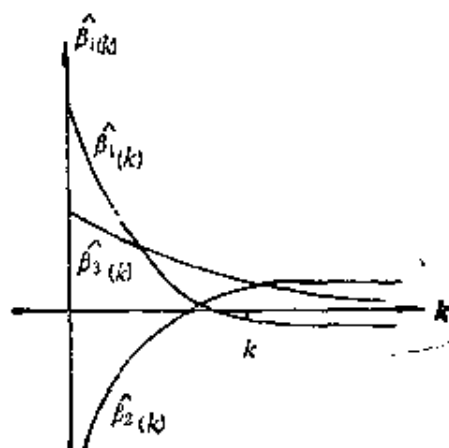


图 4.2.1 岭迹

这里还有一个岭迹计算问题。如果按照 $\hat{\beta}(k)$ 的定义去计算岭迹, 势必对每个 k 要计算一次逆阵 $(X'X + kI)^{-1}$, 这样作计算量太大。如果将 $\hat{\beta}(k)$ 变形为

$$\begin{aligned}\hat{\beta}(k) &= (\Phi \Lambda \Phi' + kI)^{-1} X'Y \\ &= \Phi (\Lambda + kI)^{-1} \Phi' X'Y \\ &= \sum_{i=1}^p \left(\frac{1}{\lambda_i + k} \right) \varphi_i \varphi_i' X'Y\end{aligned}\quad (2.11)$$

那么根据 $X'X$ 的特征根和特征向量 $\lambda_i, \varphi_i, i=1, \dots, n$, 从 (2.11) 就很容易计算出岭迹。要注意的是, 在 (2.11) 中 $X'Y$ 为正则方程 $X'X\beta = X'Y$ 的右端, 而 λ_i 和 φ_i 在作其它分析, 如复共线性等都很有用。因此用 (2.11) 计算岭迹并不需要很多附加的计算量。

岭迹还可用于研究因变量 Y 与诸自变量的关系以及回归自变

量选择, 这将在下一段讨论。

岭迹法的一个缺点是, 缺少严格的令人信服的理论依据, k 值的确定具有一定程度的主观随意性。

(2) 方差扩大因子法

上节已经指出, 方差扩大因子 c_{ji} 也度量了复共线性严重程度, 一般当 $c_{ji} > 10$ 时, 模型的复共线性就很严重。对于岭估计 $\hat{\beta}(k)$, 它的协方差阵为

$$\begin{aligned}\text{COV}(\hat{\beta}(k)) &= \sigma^2 (X'X + kI)^{-1} X'X (X'X + kI)^{-1} \\ &= \sigma^2 (c_{ij}(k))\end{aligned}$$

对角元 $c_{ji}(k)$ 就是岭估计的方差扩大因子。不难看出, $c_{ji}(k)$ 随着 k 的增大而减少。应用方差扩大因子选择 k 的经验作法是: 选择 k 使所有方差扩大因子 $c_{ji}(k) \leq 10$ 。

(3) C_p 准则

在 §3.4, 我们把 C_p 方法应用于一般线性估计时, 曾经计算出度量岭估计优劣的一个指标 C_{Lk} :

$$\begin{aligned}C_{Lk} &= \hat{\sigma}^{-2} \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 - 2 \sum_{i=1}^p \frac{Z_i^2}{\lambda_i + k} + \sum_{i=1}^p \frac{\lambda_i Z_i^2}{(\lambda_i + k)^2} \right] \\ &\quad + 2 \sum_{i=1}^p \frac{\lambda_i}{\lambda_i + k} - (n - 2)\end{aligned}$$

其中 $Z' = (Z_1, \dots, Z_p) = \Phi X'Y$, 作为 C_p 统计量的推广, C_{Lk} 愈小愈好。因此我们应该选择使 C_{Lk} 达到最小的 k 。应用上可以采用数值解法或图解法求解。

(4) Hoerl-Kennad 公式

$$\hat{k} = \frac{\hat{\sigma}^2}{\max_i c_i} \quad (2.12)$$

这个方法是基于如下的考虑。由 (2.9) 和 (2.10), 有

$$\begin{aligned} g'(k) &= g'_1(k) + g'_2(k) \\ &= 2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} (k\alpha_i^2 - \sigma^2) \end{aligned}$$

故当 $k\alpha_i^2 - \sigma^2 < 0$, $i=1, \dots, p$ 成立时, $g'(k) < 0$. 即当 $0 \leq k \leq \sigma^2 / \max \alpha_i^2$ 时, $g(k)$ 是 k 的单调下降函数. 故取

$$k^* = \frac{\sigma^2}{\max_i \alpha_i^2}$$

时, $g(k^*) < g(0)$, 即 $\text{MSE}(\hat{\beta}(k^*)) < \text{MSE}(\hat{\beta})$. 这是假定 σ^2 和 α_i^2 皆已知的情况. 事实上 σ^2 和 α_i 未知, 这时在上式中用 $\hat{\alpha}_i$ 和 $\hat{\sigma}^2$ 代替 α_i 和 σ^2 , 便得到公式 (2.12).

(5) McDorard-Garneau 法^[46]

我们知道, 当 X 呈病态时, LS 估计 $\hat{\beta}$ 偏长. McDonald 和 Garneau 把 $\hat{\beta}$ 的长度之平方 $\|\hat{\beta}\|^2$ 与 $\text{MSE}(\hat{\beta})$ 的估计 $\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1}$ 作比较. 如果

$$Q = \|\hat{\beta}\|^2 - \hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} > 0 \quad (2.13)$$

则认为 $\hat{\beta}$ 太长, 需要对它作压缩. 压缩量由 $\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1}$ 决定. 他们建议选择 k , 使得

$$\|\hat{\beta}\|^2 - \|\hat{\beta}(k)\|^2 \approx \hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1}$$

即选择 k , 使

$$\|\hat{\beta}(k)\|^2 \approx \|\hat{\beta}\|^2 - \hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} = Q \quad (2.14)$$

如果 $Q \leq 0$, 则认为 $\hat{\beta}$ 还不算太长, 此时对 $\hat{\beta}$ 不作压缩, 即选 $k=0$.

这里需要指出, 对前面几种方法所确定的 k , 我们并不能从理论上保证它们所给出的岭估计比 LS 估计有较小的均方误差. 但

是,大量的计算机模拟结果表明,当 X 呈病态时这些方法对降低均方误差都有一定作用。关于这些方法的比较,理论上也是困难的。从发表的计算机模拟结果来看,这些方法中没有一个绝对地优于所有其它方法,它们的优劣程度随着条件数不同而变化。但总的印象是McDonald-Galarneau法相对要好一些。

对下面要介绍的选 k 方法,我们可以从理论上证明,对一切 β 和 σ^2 它所给出的岭估计比LS估计有较小的均方误差。

(8) 双 h 公式

Vinod和Ullah^[47]把一些选 k 公式统一为

$$\hat{k}(h_1, h_2) = \frac{h_1 \hat{\sigma}^2}{\hat{\beta}' A \hat{\beta} + h_2 \hat{\sigma}^2} \quad (2.15)$$

其中 $A > 0$ 为已知方阵且 $\Phi' A \Phi$ 为对角阵,这里 Φ 的定义同(2.2)处。因为这个公式含两个待选参数 h_1, h_2 ,故得“双 h 公式”之名。如果岭参数由(2.15)确定,对应的岭估计常称为双 h 类岭估计(double h -class ridge estimate)。这个公式包含了其它一些常用的选 k 公式。例如,取 $A = X'X$, $h_1 = p$, $h_2 = 0$, (2.15)就变为Lawless-Wang公式^[48],

$$\hat{k} = \frac{p \hat{\sigma}^2}{\hat{\beta}' X' X \hat{\beta}} \quad (2.16)$$

又,若取 $A = I_p$, $h_1 = p$, $h_2 = 0$, (2.15)变为Hoerl-Kennard-Baldwin公式:^[49]

$$\hat{k} = \frac{p \hat{\sigma}^2}{\hat{\beta}' \hat{\beta}} \quad (2.17)$$

可以证明〔见〔8〕第六章〕,若 h_1, h_2 满足条件

$$0 < h_1 < \frac{2(n-p-1)\eta_p}{n-p+1} (\lambda_1^2 \sum \lambda_i^{-2} - 2)$$

$$h_2 \geq 0,$$

则对一切 β , σ^2 , 双 h 类岭估计比LS估计有较小的均方误差。这里 η_p 为 A 的最小特征根。

例2.1 汽车油耗问题(续例3.3.1)

对第三章例3.1讨论过的汽车油耗问题, $X'X$ 的10个特征根为5.760, 2.650, 0.597, 0.270, 0.222, 0.210, 0.133, 0.081, 0.054, 0.024. 这里后三个特征根较接近于零。又, 条件数 $\lambda_1/\lambda_{10} = 240$, 根据上节的讨论, 这个问题有较强的复共线性。

部分变量的岭迹如图4.2.2. 看来当 $0.15 \leq k \leq 0.25$ 时, 诸 $\beta_i(k)$ 比较稳定。另一方面, 当 $k=0$ 时, 方差扩大因子为21.6. 当 k 增大到0.20时, 减少到1.0. 依方差扩大因子法, 可取 $0.05 \leq k \leq 0.10$. 如果应用(2.17), 得到 $k=0.1$. 我们看到, 这几种方法所得的结论不一致。应用上需要结合具体情况及其它信息全面考虑, 最后确定一个较合适的 k 。

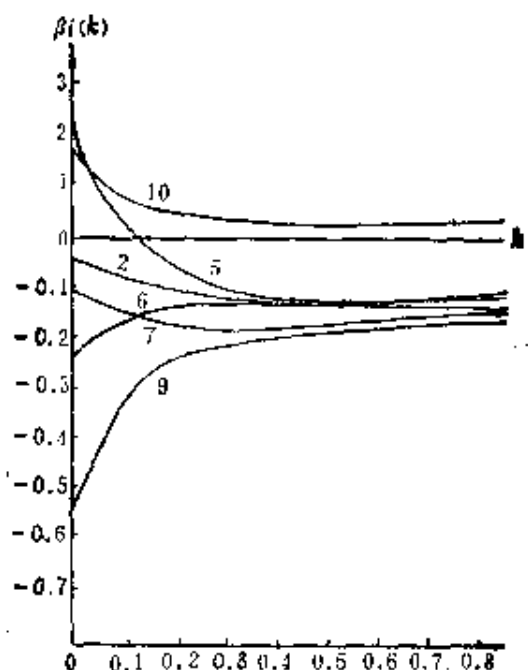


图4.2.2 汽车油耗问题的岭迹

(三) 岭迹分析与自变量选择

前面我们已经把岭迹应用于参数 k 的选择。本段要进一步说明，岭迹是分析自变量的作用、相互关系以及进行自变量选择的一种工具。下面我们举几个有代表性的情况来解释这一点。

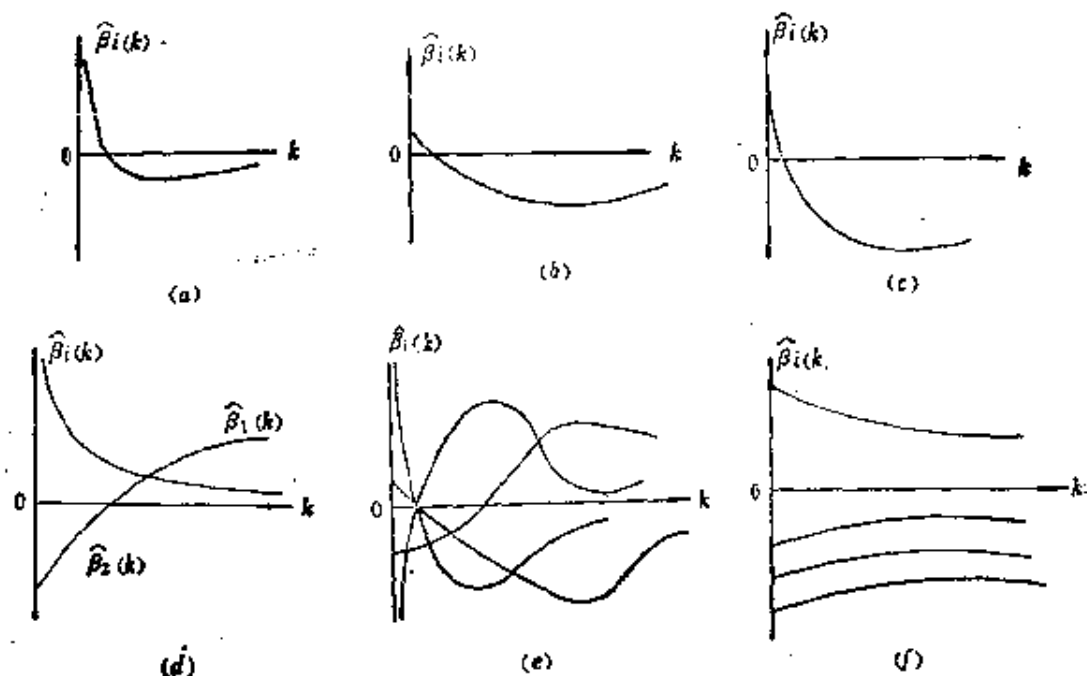


图4-2-3 岭迹

(1) 在图(a)中, $\hat{\beta}_i = \hat{\beta}_i(0) > 0$ 且比较大, 从最小二乘回归的观点看, 应将 x_i 视为对因变量 Y 有重要影响的因素。但是从岭迹看, $\hat{\beta}_i(k)$ 显示出相当的不稳定, 当 k 从零开始略增加时, $\hat{\beta}_i(k)$ 显著地下降, 而且迅速趋于零, 因而失去“预报能力”。因此, 从岭回归观点看, x_i 对因变量 Y 不起重要影响, 甚至可以去掉这个自变量。

(2) 与(a)相反的情况如图(b)所示。 $\hat{\beta}_i = \hat{\beta}_i(0) > 0$, 但很小。从最小二乘回归看, x_i 对因变量 Y 的作用不大。但随着 k 略增加, $\hat{\beta}_i(k)$ 骤然变为负的, 且绝对值还比较大。从岭回归的观点看,

x_i 对因变量有显著影响。

(3) 第三种情况如图(c)所示, $\hat{\beta}_1 = \hat{\beta}_1(0) > 0$ 比较大, 但当 k 增加时迅速下降且稳定为负值。在最小二乘回归看, x_1 是对因变量有“正”影响的重要因素。而在岭回归, x_1 要被看作对因变量有“负”影响的因素。

(4) 另一种有趣的情况如图(d)所示。这里 $\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 都很不稳定, 但其和却大体上变动不大。这种情况往往发生在自变量 x_1 和 x_2 相关性很大的场合, 即 x_1 和 x_2 之间存在复共线关系的场合。因此, 从变量选择的观点看, 两者只要保留其中一个就够了。这种情况往往有助于解释某些回归系数估计的符号不合理。比方说, 从实际观点看, β_1 和 β_2 不应有相反符号。岭回归分析的结果对这一点可以提供一种解释。

(5) 从全局看, 岭迹分析可以用来估价在某一具体场合LS估计是否合用。如图(e), 所有岭迹不稳定程度很大, 整个“系统”呈现比较“乱”的局面, 往往就使人怀疑LS估计是否很好地反映了真实情况。反过来, 若情况如图(f)那样, 则我们对LS估计可以有更大的信心。有时情况介于(e)、(f)之间, 此时我们必须适当的选择 k 值。

如果把岭迹应用于回归自变量选择, 其基本原则为

(1) 去掉岭回归系数比较稳定且绝对值比较小的自变量。这里岭回归系数可以直接比较大小的, 因为设计阵 X 是假定已经中心标准化了的。

(2) 去掉岭回归系数不稳定但随着 k 的增加迅速趋于零的自变量。

(3) 去掉一个或若干个具有不稳定岭回归系数的自变量。如果不稳定的岭回归系数很多, 究竟去掉几个, 去掉哪几个, 并无一般原则可遵循。这要结合已找出的复共线关系以及去掉后重新

进行岭回归分析的效果来决定。

下面我们举几个例子说明上面的方法。

例2.2 空气污染问题(续例3.4.3)

这个问题含15个自变量, $X'X$ 的15个特征根为: 4.5272, 2.7547, 2.0545, 1.3487, 1.2227, 0.9605, 0.6124, 0.4729, 0.3708, 0.2163, 0.1665, 0.1275, 0.1142, 0.0460, 0.0049. 我们看到后面两个特征根很接近于零。再看条件数 $\lambda_1/\lambda_{15} = 4.5272/0.0049 = 923.918$, 这个数比较接近1000。可见设计阵 X 含较严重的复共线性。根据 $\lambda_{15} = 0.0049$ 对应的特征向量, 得到复共线关系

$$-0.689x_{12} + 0.712x_{13} - 0.108x_{14} \approx 0.$$

这里 x_i 是标准化的变量。其余自变量的系数近似等于零。

岭迹如图4.2.4。从图上看, 当 $k = 0.20$ 时岭迹大体上达到稳定。依岭迹法应取 $k = 0.20$ 。如果采用方差扩大因子法, 因 $k = 0.18$ 时, 方差扩大因子接

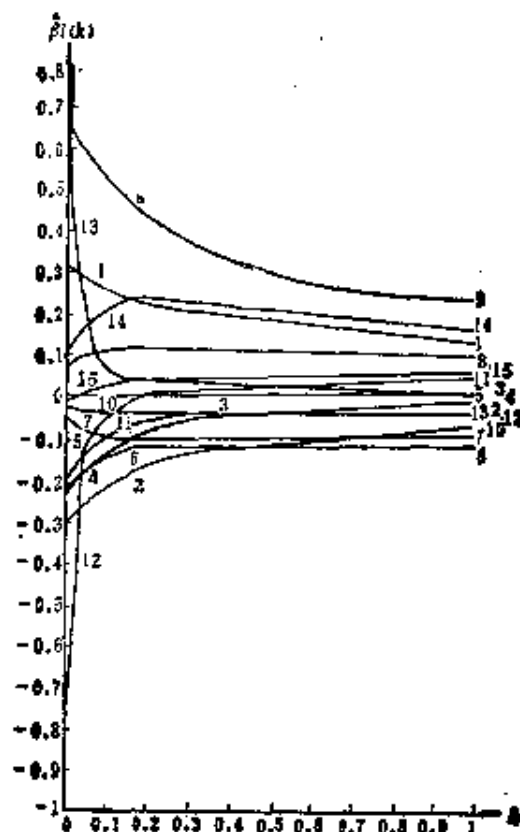


图4.2.4 空气污染问题的岭迹

近于1，当 k 在0.02—0.08时，方差扩大因子小于10，故应建议在此范围选取 k 。若用公式(2.17)，算得 $k=0.03$ ，三个方法所得结论不一致。这说明在应用上， k 的选取不是完全确定的。

至于变量选择，因为自变量 x_4, x_7, x_{10}, x_{11} 和 x_{15} 有较稳定且绝对值比较小的岭回归系数，根据变量选择的第一条原则，这些自变量可以去掉。又，自变量 x_{12} 和 x_{13} 的岭回归系数很不稳定，且随着 k 的增加很快趋于零，根据第二条原则这些自变量也应该去掉。再根据第三条原则去掉变量 x_8 和 x_5 。于是最后剩下的自变量是 $x_1, x_2, x_6, x_8, x_9, x_{14}$ 。

例2.3 Gorman-Torman例.

这是Gorman和Torman用岭回归仔细分析过的一个含10个自变量的例子。 $X'X$ 的特征根为：

3.692, 1.542, 1.293, 1.046, 0.972,

0.659, 0.357, 0.220, 0.152, 0.068.

最后一个特征根为0.068，较接近于零。条件数 $\lambda_1/\lambda_{10}=54.294 < 100$ 。如果从条件数角度看，似乎设计阵没有复共线性。但下面的研究表明，作岭回归还是必要的。关于条件数，这里附带说明它的一个缺陷，就是当 $X'X$ 所有特征根都比较小时，虽然条件数不很大，但复共线性却存在。本例就

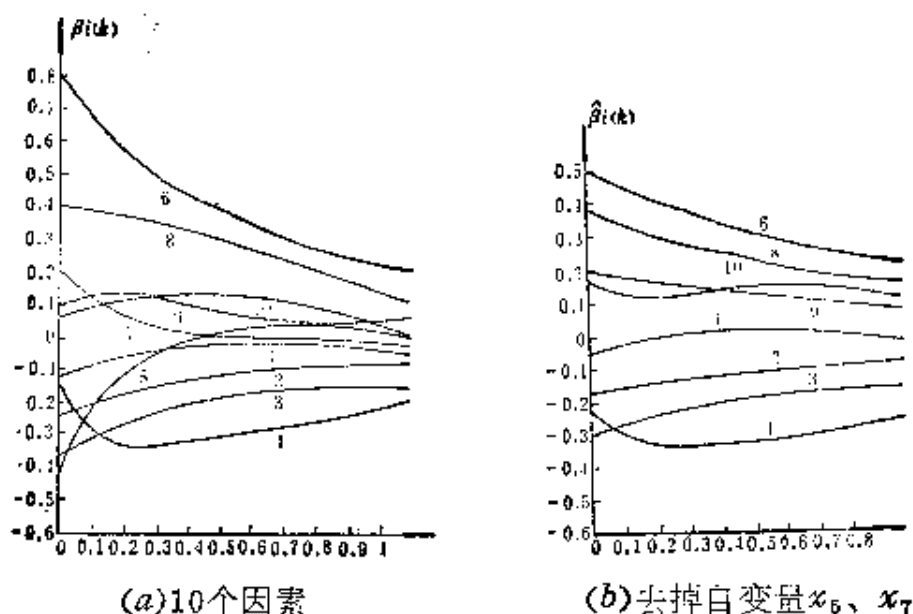


图4.2.5 Gorman Torman例的岭迹

是一个例证。

岭迹画在图4.2.5(a).从图上看, LS估计的稳定性较差.这反映在当 k 与零略有偏离时, $\hat{\beta}(k)$ 与 $\hat{\beta}$ 就有较大的差距,特别是 $|\hat{\beta}_5|$ 、 $|\hat{\beta}_6|$ 下降最多.计算结果表明,当 k 从0增加到0.1时, $\|\hat{\beta}(k)\|^2$ 下降到 $\|\hat{\beta}(0)\|^2 = \|\hat{\beta}\|^2$ 的59%.而在正交设计的情形只下降17%.这些现象在直观上使人怀疑, LS估计 $\hat{\beta}$ 是否反映了 β 的真实面目。

另外, 自变量 x_5 的回归系数LS估计 $\hat{\beta}_5$ 为负回归系数绝对值中最大的.当 k 增加时, $\hat{\beta}_5(k)$ 迅速上升且变为正的.与此相反, 对自变量 x_1 , $\hat{\beta}_1$ 为正的且绝对值最大.当 k 增加时, $\hat{\beta}_1(k)$ 迅速下降.如果根据原始数据计算 x_5 和 x_6 的相关系数, 得到0.84, 表明 x_5 和 x_6 之间有复共线关系.故这两个自变量可以去掉一个, 譬如去掉 x_5 .

我们再来看自变量 x_7 .它的回归系数LS估计 $\hat{\beta}_7$ 绝对值很大, 当 k 增加时它很快趋于零.根据前面讲的自变量选择的原则, 可以去掉 x_7 .至于 x_1 , 其回归系数LS估计绝对值看来有点偏低.当 k 增加时, $|\hat{\beta}_1(k)|$ 上升很快, 而且趋于稳定, 成为对因变量有“负”影响的重要因素.这意味着, 通常的LS估计可能对 x_1 的重要性估计过低。

从整体上看, 当 k 达到0.2—0.3范围时, 岭迹已大体稳定.因此在这个区间上取 k 值可望得到较好结果。

剔除自变量 x_5 和 x_7 , 重新作岭回归分析.岭迹画在图4.2.5(b).总体情况如图4.2.3(f).可见去掉这两个自变量是合理的。

(四) 岭估计的几何意义

前已证明, 岭估计 $\hat{\beta}(k)$ 是LS估计 $\hat{\beta}$ 的一种压缩.如果我们现在已经有了 $\hat{\beta}$, 希望将它的长度压缩到原来的 c 倍($0 < c < 1$), 并使残差平方和上升尽可能小, 我们将证明, 这样得到的估计就是岭估计。

设 b 为 β 的任一估计, 对应的残差平方和为

$$\begin{aligned} \text{RSS} &= \|Y - Y\bar{1} - Xb\|^2 \\ &= \|Y - Y\bar{1} - X\hat{\beta} + X(\hat{\beta} - b)\|^2 \end{aligned}$$

$$= \|Y - \bar{Y}\mathbf{1} - X\hat{\beta}\|^2 + (\hat{\beta} - b)'X'X(\hat{\beta} - b)$$

所以, 将 $\hat{\beta}$ 的长度压缩到原来的 c 倍且使RSS上升最少, 就是解极值问题

$$\begin{cases} (\hat{\beta} - b)'X'X(\hat{\beta} - b) \text{ 最小} \\ \|\hat{\beta}\|^2 = c^2 \|\hat{\beta}\|^2 \end{cases} \quad (2.18)$$

设 Φ 为正交阵, 致 $X'X = \Phi\Lambda\Phi'$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$. 又记 $d = \Phi'b$, $\hat{\alpha} = \Phi'\hat{\beta}$, 即 α 为典则回归系数, $\hat{\alpha}$ 为其LS估计. 显然(2.18)等价于

$$\begin{cases} (\hat{\alpha} - d)' \Lambda (\hat{\alpha} - d) \text{ 最小} \\ \|\hat{\alpha}\|^2 = c^2 \|\hat{\alpha}\|^2 \end{cases} \quad (2.19)$$

应用Lagrange乘子法, 作辅助函数

$$F(d, k) = (\hat{\alpha} - d)' \Lambda (\hat{\alpha} - d) + k(d'd - c^2 \|\hat{\alpha}\|^2)$$

其中 k 为乘子. 记 $\frac{\partial F}{\partial d} = \left(\frac{\partial F}{\partial d_1}, \dots, \frac{\partial F}{\partial d_p} \right)'$. 对上式关于 d 求导,

得

$$\frac{\partial F}{\partial d} = 2(\Lambda + kI)d - 2\Lambda\hat{\alpha}$$

令其等于零, 解出 d , 得

$$d = (\Lambda + kI)^{-1} \Lambda \hat{\alpha} \quad (2.20)$$

若能证明 $k \geq 0$, 上式就是典则回归系数 α 的岭估计. 对应地

$$\begin{aligned} b &= \Phi d = \Phi(\Lambda + kI)^{-1} \Lambda \Phi' \hat{\beta} \\ &= (X'X + kI)^{-1} X'Y \end{aligned} \quad (2.21)$$

就是原来回归系数的岭估计。

所以, 剩下的问题是证明(2.20)中 $k \geq 0$. 将(2.20)代入(2.19)的目标函数, 记之为 $Q(k)$. 于是

$$Q(k) = (\hat{\alpha} - d)' \Lambda (\hat{\alpha} - d)$$

$$\begin{aligned}
&= \hat{\alpha}' [(I - (A + kI)^{-1}A)A(I - (A + kI)^{-1}A)] \hat{\alpha} \\
&= \hat{\alpha}' \text{diag}\left(\frac{\lambda_1}{(\lambda_1 + k)^2}, \dots, \frac{\lambda_p}{(\lambda_p + k)^2}\right) \hat{\alpha} \cdot k^2
\end{aligned}$$

对 $k > 0$, 因 $(\lambda_i + k)^2 > (\lambda_i - k)^2$, $i = 1, \dots, p$, 所以 $Q(k) < Q(-k)$. 这说明 $Q(k)$ 的极小值不会在 $k < 0$ 达到. 这就证明了我们所要的结论.

从几何上来说, (2.18) 的约束条件 $\|b\|^2 = c^2 \|\hat{\beta}\|^2 \triangleq k^2$ 是一个中心在原点, 半径为 k 的球面. 对目标函数 $(b - \hat{\beta})' X' X (b - \hat{\beta})$ 作椭球

$$(b - \hat{\beta})' X' X (b - \hat{\beta}) = \delta^2 \quad (2.22)$$

因 $0 < c < 1$, 所以 $\hat{\beta}$ 在上述球之外. 故总可找到 $\delta > 0$, 使球 $\|b\|^2 = k^2$ 和椭球 (2.22) 相切于某点 $\hat{\beta}$. 显然这个 $\hat{\beta}$ 就是极值问题 (2.18) 的解, 也就是岭估计 $\hat{\beta}(k)$. 如图 4.2.6 所示.

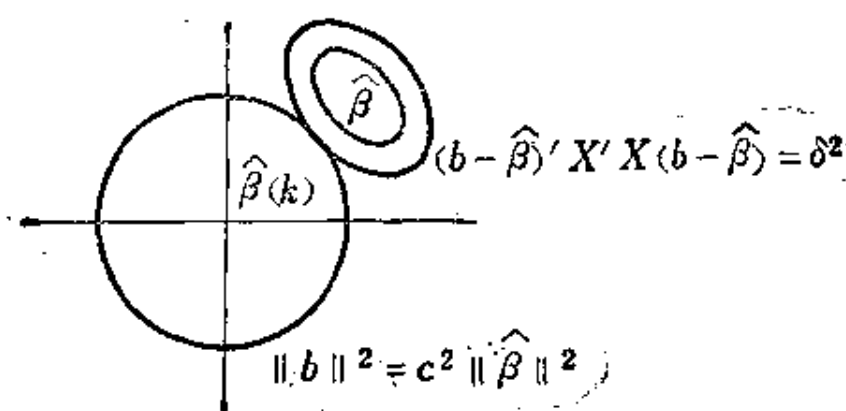


图4.2.6 岭估计的几何意义

岭估计是目前应用最广泛的一种有偏估计. 虽然 $\hat{\beta}(k) = (X'X + kI)^{-1}X'Y$ 呈线性估计的形式, 但象前面所看到的, 应用上 k 总是要通过数据来确定, 因而它依赖于 Y , 也是随机的, 故 $\hat{\beta}(k)$ 实为非线性估计. 除了对双 k 选 k 公式, 我们从理论上证明了对应的双 k 类岭估计能够一致地优于 LS 估计之外, 对其它选 k 方法目前

只有一些模拟研究结果支持了这些估计。因此在应用上除了双 h 类岭估计之外，我们并不能简单地一概用岭估计代替LS估计。事实上，若设计阵 X 不含复共线关系，LS估计仍不失为一个良好的估计。加之关于它的性质以及检验、预测等方面已有了相当丰富、完整的理论结果，因此，这时我们仍然应当考虑采用LS估计。仅当我们有充足理由认为 X 呈病态，从而复共线性较严重时，才使用岭估计以及后面要讨论的其它有偏估计。

§4.3 广义岭估计

Hoerl和Kennard在[44]中还提出了岭估计的一种推广形式，称为**广义岭估计** (Generalized ridge estimate). 为了说话方便，必要时我们把上节的岭估计称为**狭义岭估计**。

(一) 定义及性质

对线性回归模型的典则形式(2·2)，上节引进的典则回归系数 a 的狭义岭估计为

$$\hat{a}(k) = (\Lambda + kI)^{-1} Z'Y \quad (3.1)$$

若以对角元不必都相等的对角阵 $K = \text{diag}(k_1, \dots, k_p)$, ($k_i \geq 0$) 代替 kI ，可以期望均方误差能够进一步下降。基于这种考虑，定义估计

$$\hat{a}(K) = (\Lambda + K)^{-1} Z'Y \quad (3.2)$$

代回到原回归系数 β ，得

$$\begin{aligned} \hat{\beta}(K) &= \Phi \hat{a}(K) \\ &= \Phi(\Lambda + K)^{-1} \Phi' X'Y \\ &= (X'X + \Phi K \Phi')^{-1} X'Y \end{aligned} \quad (3.3)$$

称(3.2)和(3.3)分别为典则回归系数和原回归系数的**广义岭估计**

计。显然，当 $K = kI$ 时，(3.3) 就化为 (2.1)，即狭义岭估计是广义岭估计的特殊情况。

广义岭估计具有下列性质：

(1) $\hat{\beta}(K) = B_k \hat{\beta}$ ，其中 $B_k = (X'X + \Phi K \Phi')^{-1}(X'X)^{-1}$ ，即广义岭估计也是 LS 估计的一个线性变换。

(2) $E\hat{\beta}(K) = B_k \beta$ ，可见只要 $B_k \neq I$ ，等价地 $K \neq 0$ ，广义岭估计就是有偏估计。

(3) 对任意 $K = \text{diag}(k_1, \dots, k_p)$ ， $k_i > 0$ ， $\|\hat{\beta}\| > 0$ ，总有 $\|\hat{\beta}(K)\| < \|\hat{\beta}\|$ ，即广义岭估计也是 LS 估计向原点的一种压缩。

以上性质的证明都很简单，请读者自己完成。

(4) 存在 $K = \text{diag}(k_1, \dots, k_p) > 0$ ，使得

$$\text{MSE}(\hat{\beta}(K)) < \text{MSE}(\hat{\beta}) \quad (3.4)$$

证明 根据均方误差的性质，有

$$\text{MSE}(\hat{\beta}(K)) = \text{MSE}(\hat{a}(K))$$

$$\text{MSE}(\hat{\beta}) = \text{MSE}(\hat{a})$$

故要证 (3.4)，只需证明

$$\text{MSE}(\hat{a}(K)) < \text{MSE}(\hat{a}) \quad (3.5)$$

依公式 (1.2) 及 $Z'1 = 0$ ，有

$$\begin{aligned} \text{MSE}(\hat{a}(K)) &= \text{tr COV}(\hat{a}(K)) + \|E\hat{a}(K) - a\|^2 \\ &= \sigma^2 \text{tr}[(\Lambda + K)^{-1} \Lambda (\Lambda + K)^{-1}] \\ &\quad + \|(\Lambda + K)^{-1} Z'(a_0 1 + Z\alpha) - a\|^2 \\ &= \sigma^2 \text{tr}[(\Lambda + K)^{-2} \Lambda] \\ &\quad + \|((\Lambda + K)^{-1} \Lambda - I)\alpha\|^2 \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^p \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2} \end{aligned} \quad (3.6)$$

对 k_i ， $i = 1, \dots, p$ 求导数，并令之为零，解得 k_i 的最优值为

$$k_i^* = \frac{\sigma^2}{a_i^2}, \quad i=1, \dots, p \quad (3.7)$$

因为 $\lambda_i > 0$, 经简单计算知

$$\begin{aligned} \text{MSE}(\hat{\alpha}(K^*)) - \text{MSE}(\hat{\alpha}) &= -\sigma^4 \sum_{i=1}^p \frac{1}{\lambda_i a_i^2 \left(\lambda_i + \frac{\sigma^2}{a_i^2} \right)} \\ &< 0 \end{aligned}$$

其中 $K^* = \text{diag}(k_1^*, \dots, k_p^*)$. 这就证明了所要结论。

因为(3.7)中, k_i^* 不必都相等, 因此, 从理论上说, 广义岭估计能够比狭义岭估计达到更低的均方误差。

遗憾的是, 最优值 k_i^* 依赖于未知参数 σ^2 和 a_i . 因此和狭义岭估计一样, 必须通过数据来确定 K . 这样得到的广义岭估计就是非线性估计了。

另外还可以证明^[8], 对一切 $K > 0$, 广义岭估计是 β 的可容许估计, 且还是 Bayes 估计。

(二) 岭参数 K 的选择

和狭义岭估计一样, 从数据选择岭参数 K 是应用上十分重要的问题。目前已提出许多方法。这里介绍其中有代表性的几种。

(1) Hemmerle-Brantle 法

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{a}_i^2 - \hat{\sigma}^2/\lambda_i}, \quad i=1, \dots, p \quad (3.8)$$

当 $\hat{a}_i^2 - \hat{\sigma}^2/\lambda_i \leq 0$ 时, 取 $\hat{k}_i = \infty$

(3.8) 可以从两种不同的考虑导出。一种是在(3.7)中, 用 σ^2 和 a_i^2 的无偏估计 $\hat{\sigma}^2$ 和

$$\tilde{a}_i^2 = \hat{a}_i^2 - \frac{\hat{\sigma}^2}{\lambda_i}, \quad i=1, \dots, p \quad (3.9)$$

代替 σ^2 和 a_i^2 得到的。(3.9)是 a_i 的无偏估计, 容易从事实

$$\begin{aligned} E(\hat{a}_i^2) &= \text{Var}(\hat{a}_i) + (E\hat{a}_i)^2 \\ &= \frac{\sigma^2}{\lambda_i} + a_i^2 \end{aligned}$$

导出。

(3.8)的另一种导出方法是, Hemmerle和Brantle^[60]证明了(3.8)使 $\text{MSE}(\hat{a}(K))$ 的一个无偏估计达到最小。事实上, 利用 $\hat{a}(K) = (\Lambda + K)^{-1}\Lambda\hat{a} \triangleq D\hat{a}$, 有

$$\begin{aligned} E(\hat{a}(K) - \hat{a})'(\hat{a}(K) - \hat{a}) &= E(D\hat{a} - \hat{a})'(D\hat{a} - \hat{a}) \\ &= E\hat{a}'(I - D)^2\hat{a} \end{aligned}$$

因 $E\hat{a} = a$, $\text{COV}(\hat{a}) = \sigma^2\Lambda^{-1}$, 利用引理1.2.1, 得

$$\begin{aligned} E(\hat{a}(K) - \hat{a})'(\hat{a}(K) - \hat{a}) &= a'(I - D)^2a + \sigma^2(I - D)^2\Lambda^{-1} \\ &= \sum_{i=1}^p \frac{k_i^2 a_i^2}{(\lambda_i + k_i)^2} + \sigma^2 \sum_{i=1}^p \frac{k_i}{\lambda_i(\lambda_i + k_i)} \end{aligned}$$

结合(3.6), 得到

$$\begin{aligned} \text{MSE}(\hat{a}(K)) &= E(\hat{a}(K) - \hat{a})'(\hat{a}(K) - \hat{a}) \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i - k_i}{\lambda_i(\lambda_i + k_i)} \end{aligned}$$

在上式右端用 $\hat{\sigma}^2$ 代替 σ^2 , 不难得到 $\text{MSE}(\hat{a}(K))$ 的一个无偏估计

$$L = (\hat{a}(K) - \hat{a})'(\hat{a}(K) - \hat{a}) + \hat{\sigma}^2 \sum_{i=1}^p \frac{\lambda_i - k_i}{\lambda_i(\lambda_i + k_i)} \quad (3.10)$$

我们证明(3.8)使上式达到极小。记 $D = (\Lambda + K)^{-1}\Lambda = \text{diag}(d_1, \dots, d_p)$, 则 $d_i = \lambda_i/(\lambda_i + k_i)$, $i = 1, \dots, p$. 将 L 变形为

$$L = \sum_{i=1}^p \left[(d_i - 1)^2 \hat{a}_i^2 + \frac{\hat{\sigma}^2}{\lambda_i} (2d_i - 1) \right] \triangleq \sum_{i=1}^p f_i(d_i)$$

其中

$$\begin{aligned} f_i(d_i) &= (d_i - 1)^2 \hat{\sigma}_i^2 + \frac{\hat{\sigma}_i^2}{\lambda_i} (2d_i - 1) \\ &= \hat{\sigma}_i^2 \left[(d_i + \hat{\tau}_i^{-1} - 1)^2 - \frac{\hat{\tau}_i + 1}{\hat{\tau}_i^2} \right] \end{aligned}$$

其中

$$\hat{\tau}_i = \frac{\lambda_i \hat{\alpha}_i}{\hat{\alpha}^2}, \quad i=1, \dots, p \quad (3.11)$$

因每个 f_i 都是 d_i 的一元函数, 故欲使 L 达到极小等价于对 $i=1, \dots, p$ 求 d_i 使每个 f_i 达到极小。易见, d_i 的最优值为

$$d_i^* = \begin{cases} 1 - \hat{\tau}_i^{-1}, & \text{若 } \hat{\tau}_i > 1 \\ 0, & \text{若 } \hat{\tau}_i \leq 1 \end{cases}$$

等价地, k_i 的最优值为

$$k_i^* = \begin{cases} \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2 - \hat{\sigma}^2/\lambda_i}, & \text{若 } \hat{\alpha}_i^2 > \hat{\sigma}^2/\lambda_i \\ \infty, & \text{不然} \end{cases}$$

这就证明了所要的结论。

从 $\hat{\alpha}(K) = (\Lambda + K)^{-1} \Lambda \hat{\alpha} = D \hat{\alpha}$, 有关系式

$$\hat{\alpha}_i(k_i) = d_i \hat{\alpha}_i, \quad d_i = \lambda_i / (\lambda_i + k_i) \quad (3.12)$$

其中 $\hat{\alpha}(K) = (\hat{\alpha}_1(k_1), \dots, \hat{\alpha}_p(k_p))'$. 可见选择 k_i 等价于选择 d_i , 下面的 Hemmerle 法就是选择 d_i .

(2) Hemmerle 法

$$d_i = \begin{cases} \frac{1}{2} + \sqrt{\frac{1}{4} - \hat{\tau}_i^{-1}}, & \text{若 } \hat{\tau}_i \geq 4 \\ 0, & \text{若 } \hat{\tau}_i < 4 \end{cases} \quad (3.13)$$

这里 $\hat{\tau}_i$ 由 (3.11) 定义。

这个方法的背景是，对 (3.7) 作迭代，得到 α_i 的广义岭估计序列，当迭代次数 $m \rightarrow \infty$ 时，其极限估计就是 (3.13)。具体地说，就是在迭代的第一步，先用 LS 估计 $\hat{\alpha}_i$ 、 $\hat{\sigma}^2$ 代替 (3.7) 中的 α_i 和 σ^2 ，得到

$$k_i^{(0)} = \hat{\sigma}^2 / \hat{\alpha}_i^2, \quad i=1, \dots, p$$

将此 $k_i^{(0)}$ 代入 (3.2)，又得到典则回归系数 α_i 的广义岭估计

$$\hat{\alpha}_i^{(1)} = u_i / (\lambda_i + k_i^{(0)}), \quad i=1, \dots, p$$

这里 $U = (u_1, \dots, u_p)' = Z'Y = \Phi'X'Y$ 。再将 $\hat{\sigma}^2$ 、 $\hat{\alpha}_i^{(1)}$ 代入 (3.7) 又得到一组 $k_i^{(1)}$ 。而后以 $k_i^{(1)}$ 代替 $k_i^{(0)}$ 重复上面的过程。这样做下去，我们得到一组广义岭估计序列 $\{\alpha_i^{(m)}, i=1, \dots, p\}_{m=1}^\infty$ ：

$$\begin{aligned} \hat{\alpha}_i^{(m)} &= \frac{u_i}{\lambda_i + k_i^{(m-1)}} = \frac{u_i}{\lambda_i + \hat{\sigma}^2 / \hat{\alpha}_i^{(m-1)2}} \\ &= \frac{\hat{\alpha}_i^{(m-1)2} u_i}{\lambda_i \hat{\alpha}_i^{(m-1)2} + \hat{\sigma}^2} \end{aligned} \quad (3.14)$$

为了证明 $\{\hat{\alpha}_i^{(m)}\}$ 的收敛性，先证明如下引理。

引理 3.1 设 $0 < c_0 < 1$, $a > 0$, $b \geq 0$,

$$c_{n+1} = c_n^2 / [c_n^2 + b(1 - c_n)^2 + a], \quad n=0, 1, 2, \dots \quad (3.15)$$

则当 $n \rightarrow \infty$ 时 $\lim c_n = c^*$ 存在有限，且

(1) 若 $4a(b+1) > 1$ ，则 $c^* = 0$ ，

(2) 若 $4a(b+1) \leq 1$ ，则

$$c^* = \begin{cases} c_1^*, & \text{当 } c_0 > c_2^* \\ c_2^*, & \text{当 } c_0 = c_2^* \\ 0, & \text{当 } c_0 < c_2^* \end{cases}$$

此处

$$c_i^* = \frac{(2b+1) \pm \sqrt{1-4a(b+1)}}{2(b+1)}, \quad i=1, 2 \quad (3.16)$$

约定 $i=1$ 取 $+$ 号.

证明 首先指出一个事实, 假定 c_n 收敛于 c^* , 那么对(3.15)两边取极限, 得

$$c^* = \frac{c^{*2}}{c^{*2} + b(1-c^*)^2 + a}$$

由此解得 $c^*=0$ 或 c_i^* , $i=1, 2$. 即 $\{c_n\}$ 的极限只能取这三个值. 下面分情况考虑.

(1) 假定 $4a(b+1) > 1$

记 $g(x) = x^2 + b(1-x)^2 + a - x = (1+b)x^2 - (2b+1)x + (a+b)$, 其判别式为 $(2b+1)^2 - 4(b+1)(a+b) = 1 - 4a(b+1) < 0$, 因此对任何实数 x , 有 $g(x) > 0$. 即对任意实数 x , 有

$$\frac{x}{x^2 + b(1-x)^2 + a} < 1.$$

由此得到

$$c_{n+1} = c_n \frac{c_n}{c_n^2 + b(1-c_n)^2 + a} < c_n$$

即 $\{c_n\}$ 单调下降, 且下界为零, 因此极限 c^* 存在. 因 c^* 必为实数, 在 $4a(b+1) > 1$ 的条件下, c_1^* , c_2^* 都是复数. 故 $c^*=0$.

(2) 假定 $4a(1+b) \leq 1$. 分四种情况来证明:

(a) 设 $c_0 \geq c_1^*$. 由 c_1^* 的表达式易见 $0 < c_1^* < 1$. 令

$$f(x) = \frac{x^2}{x^2 + b(1-x)^2 + a}$$

注意到 c_1^* 为方程 $f(x)=x$ 的解, 即 $f(c_1^*)=c_1^*$. 又直接计算 $f'(x)$, 不难发现在 $0 < x < 1$ 时, $f'(x) > 0$. 因此, 当 $0 < x < 1$ 时, $f(x)$ 是增函数. 再由 $0 < c_1^* \leq c_0 < 1$ 知, $c_1 = f(c_0) \geq f(c_1^*) = c_1^*$. 又由前面

关于判别式的计算及 $4a(b+1) \leq 1$ 的假设可知, 当 $x > c_1^*$ 时, $g(x) > 0$. 故 $c_1 \leq c_0$. 于是我们证明了 $c_1^* \leq c_1 < c_0$. 用归纳法可知: $c_0 > c_1 \geq c_2 \geq \dots \geq c_n \geq \dots \geq c_1^*$. 于是 $\{c_n\}$ 是收敛序列. 因已证其极限只能为 0, c_1^* , c_2^* 中的一个, 故必有 $\lim_{n \rightarrow \infty} c_n = c_1^*$.

(b) 设 $c_2^* < c_0 < c_1^*$. 同样根据 $g(x)$ 的判别式及 $4a(b+1) \leq 1$ 的假设知, $c_2^* < x < c_1^*$ 时, $g(x) < 0$. 因此, 若能证明, 对一切 n , 有 $c_2^* < c_n < c_1^*$, 则由 $g(c_n) < 0$ 及 c_n 与 c_{n+1} 的递推关系式可推得 $\{c_n\}$ 为单调上升序列. 首先, 由 $c_2^* < c_0 < c_1^*$ 以及 $f(x)$ 在 $0 < x < 1$ 是增函数, 可知 $c_1 = f(c_0) < f(c_1^*) = c_1^*$. 又 $c_1 = f(c_0) > c_0$. 这就是说, 由 $c_2^* < c_0 < c_1^*$ 可推出 $c_2^* < c_0 < c_1 < c_1^*$. 用归纳法可证明: $c_2^* < c_n < c_1^*$. 于是 $\{c_n\}$ 为单调上升有界序列, 因而其极限必存在. 又因此极限只能为 0, c_1^* , c_2^* 三者之一, 故必有 $\lim_{n \rightarrow \infty} c_n = c_1^*$.

(c) 设 $c_0 < c_2^*$. 由于当 $x < c_2^*$ 时, $g(x) < 0$, 由前面证明可知, 此时 $\{c_n\}$ 为单调下降序列, 且有下界零. 因此, $\{c_n\}$ 收敛. 易见其极限 c^* 必为 0.

(d) 设 $c_0 = c_2^*$. 这时一切 $c_n = c_2^*$. 因此对这种情况结论显然成立.

引理证毕.

对 $b=0$ 的特殊情况, 为后面应用方便, 写成如下推论形式.

推论 3.1 若 $0 < c_0 < 1$, $a > 0$, $c_{n+1} = c_n^2 / (a_n^2 + c)$, 则 $\lim_{n \rightarrow \infty} c_n = c^*$ 存在有限, 且

$$(1) \quad a > \frac{1}{4} \text{ 时, } c^* = 0$$

$$(2) \quad a \leq \frac{1}{4} \text{ 时, } c^* = \begin{cases} c_1^*, & \text{若 } c_0 > c_2^* \\ c_2^*, & \text{若 } c_0 = c_2^* \\ 0, & \text{若 } c_0 < c_2^* \end{cases}$$

这里 $c_i = \frac{1}{2} \pm \sqrt{\frac{1}{4} - a}$, 并约定 $i=1$ 取 + 号。

现在我们来证明 $\{\hat{a}_i^{(m)}\}$ 的收敛性。

定理3.1 当 $m \rightarrow \infty$ 时, 由(3.14)所定义的广义岭估计序列收敛, 且极限为

$$\hat{a}_i^* = \begin{cases} \left(\frac{1}{2} + \sqrt{\frac{1}{4} - \hat{\tau}_i^{-1}} \right) \hat{a}_i, & \text{若 } \hat{\tau}_i \geq 4 \\ 0 & \text{若 } \hat{\tau}_i < 4 \end{cases}$$

这里 $\hat{\tau}_i = \lambda_i \hat{a}_i^2 / \hat{\sigma}^2$.

证明 因为 $\hat{a}_i = u_i / \lambda_i$, 所以

$$\begin{aligned} \hat{a}_i^{(m)} &= \frac{\hat{a}_i^{(m-1)2} \hat{a}_i}{\hat{a}_i^{(m-1)2} + (\hat{\sigma}^2 / \lambda_i)} \\ &= \frac{(\lambda_i \hat{a}_i^{(m-1)} / \hat{a}_i)^2}{(\lambda_i \hat{a}_i^{(m-1)} / \hat{a}_i)^2 + \hat{\tau}_i^{-1}} \hat{a}_i \end{aligned}$$

视 $c_m = \lambda_i \hat{a}_i^{(m)} / \hat{a}_i$, $a = \hat{\tau}_i^{-1}$, 应用推论3.1, 即得所证。

例3.1 乙炔转化率问题

Marguardt等曾经应用广义岭估计研究了乙炔转化问题. 因变量 Y 为乙炔转化率(%), 自变量有三个: x_1 为反应温度($^{\circ}\text{C}$), x_2 为两种化学原材料用量之比, x_3 为反应时间(秒). 原始数据列在表3.1. 对于这种涉及化学反应过程的问题, 通常采用完全二次响应曲面模型:

$$\begin{aligned} Y = & \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \\ & + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + e. \end{aligned}$$

应用最小二乘法, 所得到的经验回归方程为

$$\begin{aligned} Y = & 35.8971 + 4.0187x_1 + 2.7811x_2 - 8.0311x_3 - 6.4568x_1x_2 \\ & - 26.9818x_1x_3 - 3.7683x_2x_3 - 12.5237x_1^2 - 0.9721x_2^2 \\ & - 11.5943x_3^2. \end{aligned} \quad (3.17)$$

表3·1 乙炔转化率数据(未中心标准化)

数据序号	乙炔转化率	反应温度	某两种材料比	反应时间
1	49.0	1300	7.5	0.0120
2	50.2	1300	9.0	0.0120
3	50.5	1300	11.0	0.0115
4	48.5	1300	13.5	0.0130
5	47.5	1300	17.0	0.0135
6	44.5	1300	23.0	0.0120
7	28.0	1200	5.3	0.0400
8	31.5	1200	7.5	0.0380
9	34.5	1200	11.0	0.0320
10	35.0	1200	13.5	0.0260
11	38.0	1200	17.0	0.0340
12	38.5	1200	23.0	0.0410
13	15.0	1100	5.3	0.0840
14	17.0	1100	7.5	0.0980
15	20.5	1100	11.0	0.0920
16	29.5	1100	17.0	0.0860

复相关系数的平方 $R = 0.998$, $RMS = \hat{\sigma}^2 = 0.8126$. 具体计算表明, 这个方程与原来数据的拟合效果比较好, 但预测比较差。图4·3·1中, A 、 B 、 E 、 F 、 I 、 J 六个点张成了两个自变量: 反应温度和反应时间的取值的凸包。对这六个点拟合值 \hat{Y}_i 和试验值 Y_i 都很相近。可是把这个方程用于 C 、 D 、 G 、 H 四个点的预测(注意这些点仍在反应温度和反应时间的试验取值范围之内), 则四个转化率预测值竟有三个是负的, 可见预测效果较差。计算发现, 自变量 x_1 和 x_2 有较大的负相关系数, 于是 x_1 和 x_2 之间可能有复共线关系。从图4·3·1也容易看出, 就这两个自变量而言, 试验点集中在 AI

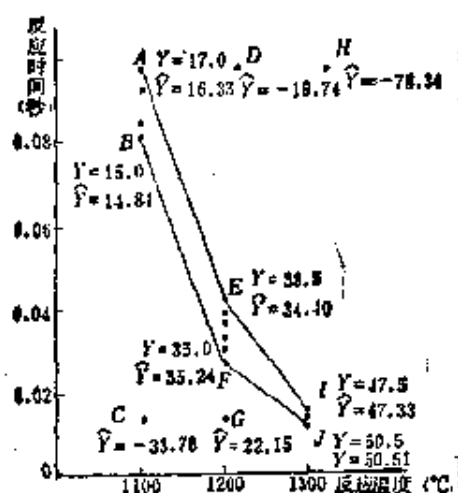


图4.3.1 乙炔转化率的预测

线段附近，而在C和H两点所处的区域没有试验点。如果客观条件容许的话，可以在这些区域补作一些试验，而后再作最小二乘分析。下面我们就现有数据应用岭估计来处理这个问题。

表3.2是对中心标准化数据计算出的方差扩大因子，VIF最大值为6569.91 $\gg 1000$ ，可见有较严重的复共线性。表3.3对中心标准化设计阵给出了 $X'X$ 的标准正交化特征向量即正交阵 Φ ，表3.4是典则形式下的设计阵 Z ，表3.5列出了应用Hemmerle法选择岭参数 K 所作出的广义岭估计。我们看到对典则参数 α_i 而言，有四个LS估计 $\hat{\alpha}_i$ 压缩为零。这个表的最后两列是原回归系数 β_i 的广义岭估计。倒数第2列是由 $\hat{\beta}(K) = \Phi \hat{\alpha}(K)$ 公式算出的原回归系数的广义岭估计，它对应于设计阵 X 已经中心标准化的回归模型。该表最后一列是只对设计阵作中心化所得到的广义岭估计。把这一列与用最小二乘法所导出的经验回归方程(3.17)相比，我们发现，原来较大的LS估计 $\hat{\beta}_{11}$ 、 $\hat{\beta}_{13}$ 、 $\hat{\beta}_{23}$ 压缩得比较厉害。

表3.2 乙炔转化率数据的方差扩大因子(中心标准化数据)

自变量	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
方差扩大因子	374	1.74	679.11	31.03	6565.91	35.60	1762.58	3.17	1158.13

表3.3 乙炔转化率数据的正交阵 Φ

.3387	.1075	.6495	.0073	.1428	-.2438	-.2077	-.5436	.1768
.1324	.3391	-.0068	-.7243	-.5843	.0205	-.0102	-.0295	-.0035
-.4137	-.0978	-.4696	-.0718	-.0182	.0160	-.1468	-.7172	.2390
-.2191	.5403	.0897	.3612	-.1661	.3733	-.5885	.0909	.0003
.4493	.0860	-.2863	.1912	-.0943	.0333	.0575	.1543	.7969
.2524	-.5172	-.0570	-.3447	.2007	.3232	-.6209	.1280	.0061
-.4056	-.0742	.4404	-.2230	.1443	.5393	.3233	.0565	.4087
.0258	.5316	-.2240	-.3417	.7342	-.0705	-.0057	.0761	.0050
-.4667	-.0969	.1421	-.1337	-.0350	-.6299	-.3089	.3631	.3309

表3.4 乙炔转化率问题典则模型的设计阵Z

Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
.5415	-1.0347	1.0487	-.1880	1.7389	-.6593	.6492	.7822	.2402
.4846	-.8830	1.1638	-.0468	.8909	-.3874	.5067	.2045	-.1939
.4048	-.6129	1.2914	.0676	-.0025	-.1631	.2187	-.0898	-.16609
.3388	-.1513	1.3176	.1315	-.7526	.3579	.1269	-1.2150	.9250
.2353	.6905	1.2785	-.0089	-1.0842	.6884	-.4181	-1.2768	1.6754
.0310	2.7455	.9535	-.7783	.2235	.2093	-1.1200	1.3128	-1.1453
.5940	-.0135	-1.0885	1.1554	1.5790	.1926	-1.3363	-.4626	.5064
.6385	-.2399	-.9170	1.0916	.3634	.4238	-1.2453	-.7138	-.3611
.7139	-.3558	-.7151	.8354	-.9374	.3207	-.6525	.5144	-.7716
.7436	-.2228	-.6170	.5668	-1.4297	-.4038	.5657	.25203	1.4085
.7668	.1034	-.3326	-.0706	-1.3472	-.3706	1.5958	-.8815	-1.3485
.8726	1.1054	-1.5272	-1.8442	.8129	-.9285	.8411	-.8981	.7053
-1.7109	.8164	-.3702	1.2052	.8885	1.9123	2.0708	.2251	-.1036
-2.1618	.1360	-.1026	.5619	-.1290	-2.5588	-.3380	-.1080	.8652
-1.6050	-.6784	-.3117	-.3325	-.7456	-.0658	-.8259	-.4662	-1.0012
-.8875	-1.4521	-.6417	-2.3461	-.0690	1.4324	-.6387	.5524	.1699

表3.5 乙炔转化率问题的广义岭回归(Hammerle法)

变量	α_i 的LS估计	λ_i	τ_i	a_i	α_i 的广义岭估计	β_i 的广义岭估计	仅中心化的广义岭估计
z_1	-0.35225	4.20480	1363.71	0.999266	-0.351991	α_0	34.3462
z_2	0.0047813	2.16261	0.13	0.0	0.0	β_1	5.7929
z_3	0.60045	1.13839	1072.39	0.999067	0.599890	β_2	2.4701
z_4	-0.23836	1.04130	154.59	0.993489	-0.236808	β_3	-5.4886
z_5	0.0094903	0.38453	0.09	0.0	0.0	β_4	-4.5715
z_6	0.21713	0.04951	6.10	0.793407	0.172272	β_5	-0.6175
z_7	0.38298	0.01363	5.23	0.742115	0.284215	β_6	-1.3318
z_8	0.52070	0.00513	3.63	0.0	0.0	β_7	2.5761
z_9	-2.4010	0.00010	1.46	0.0	0.0	β_8	-0.3548
						β_9	-0.2350

如果应用狭义岭估计, 所得结论差不多, 应用岭迹法, 选择的 $k=0.032$ 把估计结果对图4·3·1的诸点作拟合和预测, 则效果有相当的改善, 原来三个负转化率都变成了正的。

§4.4 Stein估计

从前面两节的讨论我们知道, 岭估计都是对LS估计 $\hat{\beta}$ 向原点作压缩. 一般说来, 它们是对 $\hat{\beta}$ 各分量的不均匀压缩. 本节我们讨论一种均匀压缩估计, 它是由Stein于1955年提出的, 故文献中称为Stein估计. 这是最简单、提出最早的一种有偏估计. 虽然它的应用远不及岭估计, 但却在有偏估计发展史上占有重要地位。

(一) 定义及性质

对线性回归模型(1.5), 记 $\hat{\beta}$ 为回归系数 β 的LS估计. 称 $\hat{\beta}_c(c) = c \hat{\beta}$ 为 β 的Stein估计, 这里 $0 \leq c \leq 1$ 称为压缩系数. 当 c 在 $[0, 1]$ 区间变化时, 就生成了一个估计类. 为符号简单计, 以下简记为 $\hat{\beta}_c$.

Stein估计有下面的性质:

- (1) 当 $c \neq 1$, 显然 $\hat{\beta}_c$ 为 β 的有偏、压缩估计.
- (2) 存在 $0 < c < 1$, 使得 $\text{MSE}(\hat{\beta}_c) < \text{MSE}(\hat{\beta})$.

事实上, $\hat{\beta}_c$ 的均方误差

$$\begin{aligned} \text{MSE}(\hat{\beta}_c) &= \text{tr COV}(\hat{\beta}_c) + \|E\hat{\beta}_c - \beta\|^2 \\ &= c^2 \sigma^2 \text{tr}(X'X)^{-1} + (c-1)^2 \|\beta\|^2 \\ &= c^2 \sigma^2 \sum_{i=1}^p \lambda_i^{-1} + (c-1)^2 \|\beta\|^2 \\ &\triangleq g(c) \end{aligned}$$

对 c 求导数, 并令其等于零, 解得 c 的最优值

$$c^* = \|\beta\|^2 / \left[\sigma^2 \sum_{i=1}^p \lambda_i^{-1} + \|\beta\|^2 \right] \quad (4.1)$$

容易证明, 在 c^* 处, $g(c) = \text{MSE}(\hat{\beta}_c)$ 达到最小, 且 $c^* \leq c < 1$ 时
 $\text{MSE}(\hat{\beta}_c) < \text{MSE}(\hat{\beta})$

即 Stein 估计比 LS 估计有较小的均方误差

和岭估计一样, 可以证明^[8], 对一切 $0 < c \leq 1$ Stein 估计是 β 的可容许估计, 并且它还是 Bayes 估计.

(二) 压缩系数的选择

从 (4.1) 看到, 压缩系数 c 的最优值依赖于未知参数 β 和 σ^2 , 因此和岭估计一样, 在应用上, c 必须通过数据来选择. 下面我们介绍几种方法.

(1) Stein-James 法. 假设误差 $e \sim N(0, \sigma^2 I)$, 可以证明^[8], 如果取

$$c = \left(1 - \frac{d \hat{\sigma}^2}{\hat{\beta}' X' X \hat{\beta}} \right) \quad (4.2)$$

其中 d 满足

$$0 < d < \frac{2(n-p-1)}{n-p+1} \left(\lambda_p \sum_{i=1}^p \lambda_i^{-1} - 2 \right)$$

则对一切 β 和 σ^2 , Stein 估计比 LS 估计有较小的均方误差. 其中 $\lambda_1 \geq \dots \geq \lambda_p$ 为 $X'X$ 的特征根. 这个事实的最早一种形式是由 Stein 和 James 给出的, 它是 Stein 估计的最坚实的理论支持.

(2) 应用公式

$$c = \begin{cases} \frac{1}{2} + \sqrt{\frac{1}{4} - \hat{\tau}^{-1}}, & \text{当 } \hat{\tau} \geq 4 \\ 0, & \text{当 } \hat{\tau} < 4 \end{cases} \quad (4.3)$$

这里 $\hat{\tau} = \|\hat{\beta}\|^2 / \left(\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} \right)$.

公式(4.3)的背景是, 对(4.1)应用迭代法, 产生一个序列 $\{c_m\}$, 当 $m \rightarrow \infty$ 时, 该序列的极限就是(4.3)。现在我们来证明这个事实。

首先以 β, σ^2 的LS估计 $\hat{\beta}, \hat{\sigma}^2$ 代入(4.1), 得到压缩系数

$$c_0 = \|\hat{\beta}\|^2 / \left(\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} + \|\hat{\beta}\|^2 \right)$$

然后以 $c_0 \hat{\beta}$ 作为 β 的估计再代入(4.1), 又得到新的压缩系数

$$c_1 = c_0^2 \|\hat{\beta}\|^2 / \left(\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} + c_0^2 \|\hat{\beta}\|^2 \right)$$

以 c_1 代替 c_0 重复上述过程。一般在第 m 步迭代可得

$$\begin{aligned} c_m &= c_{m-1}^2 \|\hat{\beta}\|^2 / \left(\hat{\sigma}^2 \sum_{i=1}^p \lambda_i^{-1} + c_{m-1}^2 \|\hat{\beta}\|^2 \right) \\ &= c_{m-1}^2 / (c_{m-1}^2 + \hat{\tau}^{-1}) \end{aligned} \quad (4.4)$$

这里 $\hat{\tau}$ 如(4.3)处所定义。

对 $\{c_m\}$ 应用推论 3.1, 并注意到 $c_0 = (1 + \hat{\tau}^{-1})^{-1} > c_1^* = \frac{1}{2} -$

$\sqrt{\frac{1}{4} - \hat{\tau}^{-1}}$, 立得

$$\lim_{m \rightarrow \infty} c_m = \begin{cases} \frac{1}{2} + \sqrt{\frac{1}{4} - \hat{\tau}^{-1}}, & \text{当 } \hat{\tau} \geq 4 \\ 0, & \text{当 } \hat{\tau} < 4 \end{cases}$$

这就证明了我们的结论。

在实际应用上, Stein估计往往失之过于简单, 因此它的应用与本章所要介绍的其它有偏估计比较起来相对较少。基于这个原因, 其它一些选择压缩系数的方法, 例如Farebrother法等就不再详细讨论了。

§4.5 主成分估计

主成分估计是W.F. Massy于1965年^[61]提出的另一种有偏估计. 这种估计提出的背景与前几节讨论的有偏估计有较大不同, 它主要是基于多元统计中一个重要概念——主成分. 因此, 我们首先引进主成分的概念.

(一) 主成分

假设 x 为 $p \times 1$ 随机向量, $Ex = \mu$, $\text{COV}(x) = \Sigma > 0$, 这里 μ , Σ 皆已知. 记 $\lambda_1 \geq \dots \geq \lambda_p$ 为 Σ 的特征根, $\varphi_1, \dots, \varphi_p$ 为对应的标准正交化特征向量, 即 $\Phi = (\varphi_1, \dots, \varphi_p)$ 为正交阵, 且使

$$\Phi' \Sigma \Phi = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix} \quad (5.1)$$

我们称

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} = \Phi'(x - \mu) \quad (5.2)$$

为随机向量 x 的主成分, 称 $z_i = \varphi_i'(x - \mu)$ 为 x 的第 i 个主成分, $i = 1, \dots, p$. 主成分有许多优良性质, 因而在多元数据分析中有很多应用, 成为多元统计学中一个重要概念. 下面只讨论一些与本节内容有联系的一些性质.

(1) $\text{COV}(z) = \Lambda$. 即任两个主成分都不相关, 且第 i 个主成分的方差为 λ_i .

(2) $\sum_{i=1}^p \text{Var}(z_i) = \sum_{i=1}^p \text{Var}(x_i) = \text{tr}(\Sigma)$. 即主成分的方差之和与原随机向量的方差之和相等.

$$(3) \sup_{a' a = 1} \text{Var}(a' x) = \text{Var}(z_1) = \lambda_1 \quad (5.3)$$

$$\sup_{\substack{\varphi_j' a = 0 \quad j=1, \dots, i-1 \\ a' a = 1}} \text{Var}(a' x) = \text{Var}(z_i) = \lambda_i, \quad i=2, \dots, p \quad (5.4)$$

这个性质说明, 对任意单位向量 a , 在随机变量 $a'x$ 中, 第一主成分 $z_1 = \varphi_1'(x - \mu)$ 的方差最大. 而在与第一主成分不相关的随机变量 $a'x$ 中, 第二主成分 $z_2 = \varphi_2'(x - \mu)$ 方差最大. 一般说来, 在与前 $i-1$ 个主成分不相关的随机变量 $a'x$ 中, 第 i 个主成分 $z_i = \varphi_i'(x - \mu)$ 的方差最大.

现在我们证明这些事实. (1)、(2) 是简单的, 我们只证 (3). 因为 $\text{Var}(a'x) = a' \Sigma a$, 所以问题归结为求 $a' \Sigma a / a' a$ 的最大值. 因为 $\varphi_1, \dots, \varphi_p$ 为 R^p 的一组标准正交基, 所以对任一向量 $a \in R^p$, 存在向量 $t \in R^p$, 使得 $a = \Phi t$. 利用 (5.1), 有

$$\begin{aligned} \sup_{a \neq 0} \frac{a' \Sigma a}{a' a} &= \sup_{t \neq 0} \frac{t' \Phi' \Sigma \Phi t}{t' t} \\ &= \sup_{t \neq 0} \frac{t' \Lambda t}{t' t} \\ &= \sup_{t \neq 0} \frac{\sum \lambda_i t_i^2}{\sum t_i^2} \\ &= \sup_w \sum_i \lambda_i w_i \\ &= \lambda_1 \end{aligned}$$

其中 $w_i = t_i^2 / \sum t_i^2$. 所以 $w_i \geq 0$, $\sum_{i=1}^p w_i = 1$. 上式最大值在 $w_1 = 1$, $w_i = 0$, $i > 1$, 即 $t' = (1, 0, \dots, 0)$ 达到, 也就是 $a = \varphi_1$ 达到. (5.3) 得证.

为证 (5.4), 只要注意到约束条件 $\varphi_j' a = 0$, $j = 1, \dots, i-1$ 等

价于 $a \in \mathcal{M}(\varphi_1, \dots, \varphi_p)$. 所以, 以子空间 $\mathcal{M}(\varphi_1, \dots, \varphi_p)$ 代替 R^p , 用类似方法可证得(5.4).

因为各个主成分互不相关, 第 i 个主成分 z_i 对总方差 $\text{tr}(\Sigma)$ 的贡献为 λ_i , 因此 λ_i 愈大, z_i 对总方差的贡献愈大. 如果 $\lambda_{r+1}, \dots, \lambda_p$ 都等于零, 则主成分 z_{r+1}, \dots, z_p 的方差皆为零, 再加之它们的均值都是零, 所以这些主成分都等于零(严格地说, 它们是以概率为 1 取零), 那么这些主成分就可以去掉. 这样原来 x 是 p 维向量, 现在若考虑主成分的话, 只需处理 r 维向量, 降低了问题的维数. 有时候, 后面的 $p-r$ 个主成分的方差并不严格地等于零, 只是近似地等于零, 这时它们在总方差所占的比例很小, 我们也就把它们略去了.

在应用上, μ 和 Σ 是已知的情况是很少见的. 如果 x_1, \dots, x_n 为一组随机样本, 则用 $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_i x_i$, $\hat{\Sigma} = \sum_i (x_i - \bar{x})(x_i - \bar{x})' / n$ 分别估计 μ 和 Σ . 记 $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$ 和 $\hat{\Phi} = (\varphi_1, \dots, \varphi_p)$ 分别为 $\hat{\Sigma}$ 的特征根和对应的标准正交化特征向量, 那么类似于(5.2), 称向量

$$z_{(i)} = \hat{\Phi}'(x_i - \bar{x}), \quad i = 1, \dots, n \quad (5.5)$$

为样本主成分, 而

$$Z = \begin{bmatrix} z_{(1)}' \\ \vdots \\ z_{(n)}' \end{bmatrix} = \begin{bmatrix} (x_1 - \bar{x})' \\ \vdots \\ (x_n - \bar{x})' \end{bmatrix} \hat{\Phi} \quad (5.6)$$

为样本主成分组成的矩阵. 和总体主成分一样, $\hat{\lambda}_1, \dots, \hat{\lambda}_p$ 度量了各样本主成分对总方差的贡献大小. 如果后面的几个 $\hat{\lambda}_i$ 比较接近于零或它们在总方差中所占的比例很小, 那么它们对应的样本主成分也就可以略去了.

因为以下我们总是讨论样本主成分, 因此略去“样本”二字. 凡提到主成分就是指样本主成分.

(二) 回归系数的主成分估计

讨论线性回归模型

$$Y = a_0 \mathbf{1} + X\beta + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I, \quad (5.7)$$

假设 X 已经中心化, 那么常数项 a_0 用 $\hat{a}_0 = \bar{Y} = \frac{1}{n} \sum_i Y_i$ 来估计. 记 $\lambda_1 \geq \dots \geq \lambda_p$ 为 $X'X$ 的特征根, $\varphi_1, \dots, \varphi_p$ 为对应的标准正交化特征向量. 记 $\Phi = (\varphi_1, \dots, \varphi_p)$, 那么上面模型的典则形式为

$$Y = a_0 \mathbf{1} + Z\alpha + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I \quad (5.8)$$

这里 $Z = X\Phi$, $\alpha = \Phi'\beta$. 如果把原来的 p 个回归自变量 $x' = (x_1, \dots, x_p)$ 视为随机向量, 设计阵 X 的 n 个行作为 x 的 n 个随机样本 (中心化了样本), 那么 $X'X/n$ 就是 x 的协方差阵 Σ 的一个估计. 而 $Z = (Z_1, \dots, Z_p)$ 就是样本主成分组成的设计阵. 可见, 所谓线性回归模型 (5.7) 的典则形式就是以原回归变量 $x' = (x_1, \dots, x_p)$ 的主成分 z_1, \dots, z_p 为新自变量的回归模型. 如果设计阵 X 呈病态, 那么 $X'X$ 的特征根 $\lambda_1, \dots, \lambda_p$ 中有一部分很小, 不妨设后 $p-r$ 个很小, 即 $\lambda_{r+1}, \dots, \lambda_p \approx 0$, 这时后 $p-r$ 个新自变量 (即主成分) z_{r+1}, \dots, z_p 在 n 次试验中取值变化很小. 事实上, 记 $Z'_i = (z_{1i}, \dots, z_{ni})$, $\bar{z}_i = \frac{1}{n} \sum_{j=1}^n z_{ji}$, 因 X 已中心化可推知 $\bar{z}_i = 0$. 则第 i 个新自变量 z_i (即第 i 个主成分) 在 n 次试验中取值波动大小为

$$\begin{aligned} \sum_{j=1}^n (z_{ji} - \bar{z}_i)^2 &= \sum_{j=1}^n z_{ji}^2 \\ &= Z'_i Z_i = \varphi'_i X' X \varphi_i = \lambda_i \approx 0, \quad i \geq r+1. \end{aligned}$$

这些新自变量的作用就可以并入常数项, 也就是说, 新自变量 z_{r+1}, \dots, z_p 可以从模型中剔除.

基于上述思想, 若 $\lambda_{r+1}, \dots, \lambda_p \approx 0$. 将 A 、 a 、 Z 、 Φ 作相应分块,

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \text{ 其中 } A_1: r \times r,$$

$$a = \begin{pmatrix} a_{(1)} \\ a_{(2)} \end{pmatrix}, \text{ 其中 } a_{(1)}: r \times 1$$

$$Z = (Z_{(1)}, Z_{(2)}), \text{ 其中 } Z_{(1)}: n \times r$$

$$\Phi = (\Phi_1, \Phi_2), \text{ 其中 } \Phi_1: p \times r,$$

于是(5.8)变形为

$$Y = a_0 \mathbf{1} + Z_{(1)} a_{(1)} + Z_{(2)} a_{(2)} + e, \quad E(e) = 0, \\ \text{COV}(e) = \sigma^2 I.$$

剔除 $Z_{(2)} a_{(2)}$ 这一项, 即用 $\hat{a}_{(2)} = 0$ 估计 $a_{(2)}$, 然后求得 $a_{(1)}$ 的 LS 估计

$$\hat{a}_{(1)} = A_1^{-1} Z_{(1)}' Y \quad (5.9)$$

最后利用关系 $\beta = \Phi a$, 得到 β 的估计

$$\tilde{\beta} = \Phi \begin{pmatrix} \hat{a}_{(1)} \\ 0 \end{pmatrix} = \Phi_1 \hat{a}_{(1)} = \Phi_1 A_1^{-1} Z_{(1)}' Y \quad (5.10)$$

称为 β 的**主成分估计** (principal components estimate).

与岭估计、Stein估计相类似, 主成分估计具有下列性质:

(1) $\tilde{\beta} = \Phi_1 \Phi_1' \hat{\beta}$, 即主成分估计是LS估计的一个线性变换。

$$\begin{aligned} \text{证明 } \tilde{\beta} &= \Phi_1 A_1^{-1} \Phi_1' X' Y \\ &= \Phi_1 A_1^{-1} \Phi_1' X' X \hat{\beta} \\ &= \Phi_1 A_1^{-1} \Phi_1' \Phi \Lambda \Phi' \hat{\beta} \\ &= \Phi_1 \Phi_1' \hat{\beta} \end{aligned}$$

(2) $E \tilde{\beta} = \Phi_1 \Phi_1' \beta$, 只要 $r < p$, 主成分估计就是有偏估计。

(3) $\|\tilde{\beta}\| < \|\hat{\beta}\|$, 即主成分估计 $\tilde{\beta}$ 是压缩估计。

此因 $\|\tilde{\beta}\| = \|\Phi_1 \Phi_1' \hat{\beta}\|$

$$= \left\| \Phi \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ & & & 0 \end{bmatrix} \Phi' \hat{\beta} \right\| = \left\| \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ & & & 0 \end{bmatrix} \Phi' \hat{\beta} \right\|$$

$$\begin{aligned} &< \|\Phi' \hat{\beta}\| \\ &= \|\hat{\beta}\| \end{aligned}$$

(4) 当设计阵病态时, 适当选择 r , 可使

$$\text{MSE}(\tilde{\beta}) < \text{MSE}(\hat{\beta}) \quad (5.11)$$

证明 由(1.2), 有

$$\begin{aligned} \text{MSE}(\tilde{\beta}) &= \text{MSE} \left(\begin{pmatrix} a_{(1)} \\ 0 \end{pmatrix} \right) \\ &= \sigma^2 \text{tr}(\hat{\alpha}_{(1)}) + \|a_2\|^2 \\ &= \sigma^2 \sum_{i=1}^r \lambda_i^{-1} + \sum_{i=r+1}^p \alpha_i^2 \\ &= \text{MSE}(\hat{\beta}) + \left(\sum_{i=r+1}^p \alpha_i^2 - \sigma^2 \sum_{i=r+1}^p \lambda_i^{-1} \right) \end{aligned} \quad (5.12)$$

因设计阵被假定是病态的, 故有一部分特征根非常接近于零, 不妨设后 $p-r$ 个 λ_i 很接近于零, 此时 $\sum_{i=r+1}^p \lambda_i^{-1}$ 很大, 可致(5.11)第二项为负, 于是(5.11)得证。

和其它有偏估计一样, 可以证明^[8], 主成分估计是可容许估计。关于这个估计的进一步性质, 可参看[52]、[53]及所引文献。

对于主成分估计, 有一个选择保留主成分个数的问题。应用上也要通过数据来确定。通常采用的方法有两种。其一是略去特征根很接近于零的那些主成分; 其二是选择 r , 使得前 r 个特征根之和在 p 个特征根总和中所占的比例达到预先给定的值。譬如, 选

择 r ,使

$$\sum_{i=1}^r \lambda_i / \sum_{i=1}^4 \lambda_i > 75\% \text{ 或 } 80\% \text{ 等等}$$

这里,选择 r 的问题实际上是典则形式(5·8)中选择自变量问题。

一些模拟研究结果^[64]表明,当设计阵 X 呈病态时,主成分估计确能对LS估计作相当的改进。

下面我们举两个实例说明主成分估计的方法和应用。

例5·1 Hald水泥数据(续例3·3·2)

对表3·3·2的原始数据施中心化标准化程序,则 $X'X$ 就是样本相关阵(见表3·5·2),此时 $\text{tr}(X'X) = 4$,它的四个特征根依次为2.2357, 1.5761, 0.1866, 0.0016.最后一个特征根很接近于零,而且前三个特征根之和所占比例达到0.9995.于是我们略去第4个主成分,即选 $r = 3$.对应的三个特征向量分别为

$$\varphi'_1 = (0.4760, 0.5639, -0.3941, -0.5479),$$

$$\varphi'_2 = (-0.5090, 0.4139, 0.6050, -0.4512),$$

$$\varphi'_3 = (0.6755, -0.3144, 0.6377, -0.1954),$$

相应的回归方程为

$$Y = 95.4230 + 9.8831z_1 + 0.1250z_2 + 4.5583z_3.$$

变换到原来的自变量,得到经验回归方程

$$Y = 85.7430 + 1.3119x_1 + 0.2694x_2 - 0.1428x_3 - 0.3801x_4.$$

考虑到前两个特征根之和在总和所占比例已达到0.9529,故也可选 $r = 2$.此时对应的经验回归方程为

$$Y = 85.8603 + 1.3227x_1 + 0.2661x_2 - 0.1546x_3 - 0.3767x_4.$$

例5·2 经济分析问题

表5·1是 Malinvand 于1966年提出的研究法国经济问题的一组实际数据.所考虑的因变量为进口总额,自变量有3个: x_1 = 国内总产值, x_2 = 存储量, x_3 = 总消费量.所有可能子集回归列在表5·2.从此表可以看出,自变量 x_3 进入回归方程对 x_1 的回归系数影响很大,这表明含有 x_1 和 x_3 的复共线关系是存在的,将原始数据中心标准化,求得 $X'X$ 为

$$X'X = \begin{bmatrix} 1 & 0.026 & 0.997 \\ 0.026 & 1 & 0.036 \\ 0.997 & 0.036 & 1 \end{bmatrix}$$

它的三个特征根分别为 $\lambda_1 = 1.999$, $\lambda_2 = 0.998$, $\lambda_3 = 0.003$. 最后一个特征根很小, 由此也可以看出复共线性存在. 再看条件数 $\lambda_1/\lambda_3 = 666.333$, 可见有中等程度的复共线性. $X'X$ 的三个特征向量为

$$\varphi'_1 = (0.7063, 0.0435, 0.7065),$$

$$\varphi'_2 = (-0.0357, 0.9990, -0.0258),$$

$$\varphi'_3 = (-0.7070, -0.0070, 0.7072).$$

三个主成分分别为

$$z_1 = 0.7063x_1 + 0.0435x_2 + 0.7065x_3,$$

$$z_2 = -0.0357x_1 + 0.9990x_2 - 0.0258x_3,$$

$$z_3 = -0.7070x_1 - 0.0070x_2 + 0.7072x_3.$$

因为 $\lambda_3 = 0.003 \approx 0$, 于是 $z_3 \approx 0$ 就是一个复共线关系, 即

$$-0.7070x_1 - 0.0070x_2 + 0.7072x_3 \approx 0$$

表5.1 法国经济分析数据(单位: 十亿法郎)

年	国内总产值 (x_1)	存储量 (x_2)	总消费量 (x_3)	进口总额 (y)
1949	149.3	4.2	108.1	15.9
50	161.2	4.1	114.8	16.4
51	171.5	3.1	123.2	19.0
52	175.5	3.1	126.9	19.1
53	180.8	1.1	132.1	18.8
54	190.7	2.2	137.7	20.4
55	202.1	2.1	146.0	22.7
56	212.4	5.6	154.1	26.5
57	226.1	5.0	162.3	28.1
58	231.9	5.1	164.3	27.6
59	239.0	0.7	167.6	26.3

表5.2 法国经济问题的所有可能子集回归

进入回归的变量	回 归 系 数 的 LS 估 计		
	x_1	x_2	x_3
1	0.146	—	—
2	—	0.691	—
3	—	—	0.214
1, 2	0.145	0.622	—
1, 3	-0.109	—	0.372
2, 3	—	0.596	0.212
1, 2, 3	-0.051	0.587	0.287

为一复共线关系。注意到 x_2 的系数 $-0.0070 \approx 0$ ，而 x_1 和 x_3 的系数绝对值近似相等，于是复共线关系为 $x_1 \approx x_3$ ，这和 x_1 与 x_3 的简单相关系数 $r = 0.997$ 是一致的。请注意 $x_1 \approx x_3$ 是对中心标准化数据而言的，从表5.1可以看出，对原始数据 $x_1 \approx x_3$ 并不成立。但是根据中心标准化，容易把关系 $x_1 \approx x_3$ 还原到原来变量去。据报导，这个信息对回归预报、分析经济政策以及作出新决策都是有益的。

保留前两个主成分，算出主成分回归，还原到原来原量，得到主成分回归方程：

表5.3 法国经济分析问题的三种估计

变 量	常数项	x_1	x_2	x_3	复相关系数的平方
主成分估计 ($r=2$)	-9.1057	0.0727	0.6091	0.1062	0.988
LS估计	-10.1300	-0.0514	0.5859	0.2868	0.992
岭估计 $k=0.04$	-8.5537	0.0635	0.5859	0.1156	0.988

$$Y = -9.1057 + 0.0727x_1 + 0.6091x_2 + 0.1062x_3.$$

表5·3给出了主成分估计、LS估计和岭估计.我们看到,主成分估计和岭估计大体相近.与LS估计相比,复共线关系中所含的自变量 x_1 和 x_3 的系数的两种有偏估计变化较大,且 x_1 的系数 β_1 的符号也发生了变化.即在最小二乘回归看来,自变量 x_1 对 Y 有“负”影响,而在主成分估计和岭估计看来,它却有“正”影响.

§4.6 特征根估计

回归系数的特征根估计是由Webster等^[65]于1974年提出的,它是上节讨论的主成分估计的一种推广.在特征根估计中我们考虑特征根和特征向量时,不象在主成分估计那样只限于自变量部分,而是把因变量也放在一起来考虑.这是特征根估计引人注目的原因之一.

(一) 问题的提出

我们考虑线性回归模型

$$Y = \alpha \mathbf{1} + X\beta + e, \quad E(e) = 0, \quad \text{COV}(e) = \sigma^2 I, \quad (6.1)$$

假设设计阵 X 已经中心化、标准化,则 α, β 的LS估计分别为 $\hat{\alpha} = \bar{Y} = \frac{1}{n} \sum_i Y_i$, $\hat{\beta} = (X'X)^{-1} X'Y$.现在将 Y 也中心化、标准化,

即以

$$\tilde{Y} = (Y - \bar{Y}\mathbf{1})/s_y \quad (6.2)$$

代替 Y , 这里 $s_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. 记 $A = (\tilde{Y} : X)$, 则

$$A'A = \begin{pmatrix} \widetilde{Y}'\widetilde{Y} & \widetilde{Y}'X \\ X'\widetilde{Y} & X'X \end{pmatrix} = \begin{pmatrix} 1 & \widetilde{Y}'X \\ X'\widetilde{Y} & X'X \end{pmatrix} \quad (6.3)$$

记 $\lambda_0, \lambda_1, \dots, \lambda_p$ 为 $A'A$ 的特征根, $\psi_0, \psi_1, \dots, \psi_p$ 为 $A'A$ 的标准正交特征向量, 于是若记

$$\Psi = (\psi_0, \psi_1, \dots, \psi_p),$$

则有

$$\Psi' A' A \Psi = \Lambda = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_p). \quad (6.4)$$

再记

$$\psi'_i = (\psi_{0i}, \psi_{1i}, \dots, \psi_{pi}), \quad i=0, 1, \dots, p,$$

$$\widetilde{\psi}'_i = (\psi_{1i}, \dots, \psi_{pi}), \quad i=0, 1, \dots, p$$

如果 $\lambda_i=0$, 则因 $A'A\psi_i=\lambda_i\psi_i=0$, 左乘 ψ'_i 得

$$\psi'_i A' A \psi_i = \lambda_i \psi'_i \psi_i = \lambda_i = 0,$$

于是 $A\psi_i=0$. 这个事实说明, 从现有 n 组数据看, 因变量 \widetilde{y} (记 y 为因变量, $\widetilde{y} = (y - \bar{Y})/s_y$)和自变量 x_1, x_2, \dots, x_p 之间有严格的线性关系:

$$\psi_{0i} \widetilde{y} + \psi_{1i} x_1 + \dots + \psi_{pi} x_p = 0 \quad (6.5)$$

这时可能发生如下两种情况:

(1) $\psi_{0i} \neq 0$

若 $\psi_{0i} \neq 0$, 则由(6.5)可知, \widetilde{y} 可通过 x_1, x_2, \dots, x_p 线性表出

$$\widetilde{y} = -\frac{1}{\psi_{0i}} (\psi_{1i} x_1 + \dots + \psi_{pi} x_p)$$

将 $\widetilde{y} = (y - \bar{Y})/s_y$ 代入, 得

$$y = \bar{Y} - \frac{s_y}{\psi_{0i}} (\psi_{1i} x_1 + \dots + \psi_{pi} x_p) \quad (6.6)$$

因为 n 组数据严格满足这个方程, 所以残差平方和为零。因而(6.6)是经验回归方程中最理想的候选者, 如把它用作预测, 它应有

较好的预测效果。

$$(2) \psi_{0i} = 0$$

此时从(6.5)知, 自变量 x_1, \dots, x_p 有严格线性关系

$$\psi_{1i}x_1 + \dots + \psi_{pi}x_p = 0. \quad (6.7)$$

在应用上, 如果 $\lambda_i \approx 0$, 则(6.5)只是一个近似关系

$$\psi_{0i}y + \psi_{1i}x_1 + \dots + \psi_{pi}x_p \approx 0$$

再若 $\psi_{0i} \approx 0$, 上式即变为复共线关系

$$\psi_{1i}x_1 + \dots + \psi_{pi}x_p \approx 0. \quad (6.8)$$

把上面的讨论归纳起来就是: 对每一对 $\lambda_i \approx 0, \psi_{0i} \approx 0$, 设计阵 X 就有一个复共线关系(6.8)存在. 如果至少有一对这样的 λ_i, ψ_{0i} 存在, LS估计的性质就显著地变差. 本节要讨论的特征根估计就是针对这个事实设计的。

(二) 回归系数的特征根估计

我们先证明, 通过 $A'A$ 的特征根 λ_i 和 $\psi_i, i=0, 1, \dots, p$ 表示 LS 估计 $\hat{\beta}$ 的一个关系式.

引理6.1 记 $\hat{\beta}' = (\hat{\beta}_1, \dots, \hat{\beta}_p)$, 则

$$\hat{\beta}_j = -s_y \sum_{i=0}^p \psi_{ji} a_i, \quad j=1, \dots, p \quad (6.9)$$

其中
$$a_i = \frac{\psi_{0i}}{\lambda_i \sum_{j=1}^p (\psi_{0j}^2 / \lambda_j)}, \quad i=0, 1, \dots, p \quad (6.10)$$

证明 记

$$A'A = \begin{bmatrix} 1 & b' \\ b & B \end{bmatrix} \stackrel{\Delta}{=} \begin{bmatrix} b_{00} & b_{01} & \dots & b_{0p} \\ b_{10} & b_{11} & \dots & b_{1p} \\ \vdots & \vdots & & \vdots \\ b_{p0} & b_{p1} & \dots & b_{pp} \end{bmatrix}$$

这里 $b_{00}=1$, $b_{0j}=b_{j0}$, $j=1, \dots, p$ 由 (6.4) 得。

$$A' A \psi_i = \lambda_i \psi_i, \quad i=0, 1, \dots, p$$

即

$$b_{l0} \psi_{0l} + b_{l1} \psi_{1l} + \dots + b_{lp} \psi_{pl} = \lambda_l \psi_{ll}, \quad l=0, 1, \dots, p$$

$$i=0, 1, \dots, p$$

两边用 ψ_{0i}/λ_i 去乘, 而后对 $i=0, 1, \dots, p$ 求和, 有

$$b_{l0} \sum_{i=0}^p \frac{\psi_{0i}^2}{\lambda_i} + b_{l1} \sum_{i=0}^p \frac{\psi_{0i} \psi_{1i}}{\lambda_i} + \dots + b_{lp} \sum_{i=0}^p \frac{\psi_{0i} \psi_{pi}}{\lambda_i}$$

$$= \sum_{i=0}^p \psi_{li} \psi_{0i} = 0 \quad l=1, 2, \dots, p$$

用 $\sum_{i=0}^p \psi_{0i}^2/\lambda_i$ 去除上式两边, 得

$$b_{l0} + b_{l1} \frac{\sum_{i=0}^p \left(\frac{\psi_{0i} \psi_{1i}}{\lambda_i} \right)}{\sum_{i=0}^p \left(\frac{\psi_{0i}^2}{\lambda_i} \right)} + \dots + b_{lp} \frac{\sum_{i=0}^p \left(\frac{\psi_{0i} \psi_{pi}}{\lambda_i} \right)}{\sum_{i=0}^p \left(\frac{\psi_{0i}^2}{\lambda_i} \right)} = 0$$

$$l=1, \dots, p \quad (6.11)$$

若令 $u' = (u_1, \dots, u_p)$

$$u_j = \frac{\sum_{i=0}^p \left(\frac{\psi_{0i} \psi_{ji}}{\lambda_i} \right)}{\sum_{i=0}^p \left(\frac{\psi_{0i}^2}{\lambda_i} \right)}, \quad j=1, \dots, p$$

(6.11) 即为

$$b_{l0} + b_{l1} u_1 + \dots + b_{lp} u_p = 0, \quad l=1, \dots, p.$$

此即 $Bu = -b$. 根据 B 、 b 的定义, 我们得到

$$u = -B^{-1}b = -(X'X)^{-1}X'\tilde{Y}$$

$$= -(X'X)^{-1}X'(Y - \bar{Y}\mathbf{1})/s_y$$

$$= -\hat{\beta}/s_y$$

于是

$$\hat{\beta}_j = -s_y u_j$$

$$= -s_y \sum_{i=0}^p \psi_{ji} a_i, \quad i=1, \dots, p$$

引理证毕.

先假设每个 $\psi_{0i} \neq 0, i=0, 1, \dots, p$. 不管 λ_i 的值如何, 对每个 i 我们作预测方程 (6.6). 一般说来, 这些方程有的预测效果好些, 有的预测效果差一些. 我们试图通过加权方法得到一个较好的预测方程. 于是考虑预测方程

$$y = \sum_{i=0}^p w_i \psi_{0i} \left(\bar{Y} - \frac{s_y}{\psi_{0i}} \sum_{j=1}^p \psi_{ji} x_j \right) \quad (6.12)$$

其中 $w_i \psi_{0i}$ 为权, 满足

$$\sum_{i=0}^p w_i \psi_{0i} = 1 \quad (6.13)$$

改写 (6.12) 为

$$y = \bar{Y} - s_y \sum_{j=1}^p \left(\sum_{i=0}^p \psi_{ji} w_i \right) x_j \quad (6.14)$$

下面我们证明, 对 (6.12) 应用最小二乘法所确定的 w_i 一定等于 (6.10) 定义的 a_i . 这就是说, (6.14) 中 x_j 的回归系数的估计就是 $\hat{\beta}_j$.

事实上, 由 (6.14) 得到的在 n 个试验点的拟合值所构成的向量为

$$Y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix} = \bar{Y} \mathbf{1} - s_y X \sum_{i=0}^p w_i \tilde{\psi}_i$$

从而残差平方和为

$$\begin{aligned} \text{RSS} &= \|Y - Y^*\|^2 \\ &= s_y^2 \|\tilde{Y} + X \sum_{i=0}^p w_i \tilde{\psi}_i\|^2 \end{aligned}$$

利用

$$\sum_{i=0}^p w_i \psi_{0i} = 1$$

上式可变形为

$$\begin{aligned} \text{RSS} &= s_y^2 \left\| \sum_{i=0}^p w_i \psi_{0i} \tilde{Y} + X \sum_{i=0}^p w_i \tilde{\psi}_i \right\|^2 \\ &= s_y^2 \left\| \sum_{i=0}^p w_i (\psi_{0i} \tilde{Y} + X \tilde{\psi}_i) \right\|^2 \\ &= s_y^2 \left\| \sum_{i=0}^p w_i A \psi_i \right\|^2 \end{aligned}$$

由(6.4)知, $A\Psi = (A\psi_0, \dots, A\psi_p)$ 的列向量两两正交, 故

$$\begin{aligned} \text{RSS} &= s_y^2 \sum_{i=0}^p w_i^2 \psi_i' A' A \psi_i \\ &= s_y^2 \sum_{i=0}^p w_i^2 \lambda_i \end{aligned} \quad (6.15)$$

我们要在约束条件 $\sum_{i=0}^p w_i \psi_{0i} = 1$ 下求(6.15)的最小值。应用 Lagrange 乘子法, 作辅助函数

$$F(w_0, \dots, w_p, \mu) = s_y^2 \sum_{i=0}^p w_i^2 \lambda_i - 2\mu \left(\sum_{i=0}^p w_i \psi_{0i} - 1 \right)$$

其中 μ 为 Lagrange 乘子。由

$$\frac{\partial F}{\partial w_i} = 0, \quad i=0, \dots, p, \quad \frac{\partial F}{\partial \mu} = 0$$

解得

$$w_i = \psi_{0i} / \lambda_i \sum_{j=0}^p \frac{\psi_{0j}^2}{\lambda_j}, \quad i=0, 1, \dots, p$$

比较(6.10), 知 $w_i = a_i$, $i=0, \dots, 1, \dots, p$, 这就证明了我们的断言。

现在我们假设 $\lambda_i \approx 0$, $\psi_{0i} \approx 0$, $i=0, 1, \dots, k-1$. 那末根据前面的分析, 对这些 i 我们就没有方程(6.6). 于是在推出(6.12)时, 求和就不包括这些 i , 即

$$y = \sum_{i=k}^p w_i \psi_{0i} \left(\bar{Y} - \frac{s_y}{\psi_{0i}} \sum_{j=1}^p \psi_{ji} x_j \right)$$

现假定 $\sum_{i=k}^p w_i \psi_{0i} = 1$, 上式即为

$$y = \bar{Y} - s_y \sum_{i=k}^p \left(\sum_{j=1}^p \psi_{ji} w_i \right) x_j \quad (6.16)$$

这时相应的约束条件为 $\sum_{i=k}^p w_i \psi_{0i} = 1$. 仿照前面的方法, 从现有 n 组数据求 w_k, \dots, w_p , 使得 (6.16) 的残差平方和达到最小, 就得到

$$w_i = \frac{\psi_{0i}}{\lambda_i \sum_{j=k}^p \psi_{0j}^2 / \lambda_j}, \quad i = k, \dots, p \quad (6.17)$$

此时相应的残差平方和为

$$RSS = s_y^2 \left(\sum_{i=k}^p \frac{\psi_{0i}^2}{\lambda_i} \right)^{-1}$$

这样导出预测方程或者说求回归系数估计的方法称为特征根法, 所得到的估计叫做**特征根估计** (Latent root estimate). 从 (6.16) 和 (6.17) 知, 回归系数的特征根估计为

$$\tilde{\beta}_j = -s_y \sum_{i=k}^p \psi_{ji} w_i \quad (6.18)$$

其中 w_i 由 (6.17) 定义。比较 (6.18) 和 (6.9) 我们发现, 特征根估计和 LS 估计不同之处仅在于求和的起点、若前 k 对 $\lambda_i \approx 0, \psi_{0i} \approx 0$, 那么特征根估计 (6.18) 的和式就比 LS 估计 (6.9) 要少 k 项。

综合上面的讨论, 计算回归系数特征根估计的步骤可归纳如下:

(1) 依 (6.3) 计算相关阵 $A'A$,

(2) 计算 $A'A$ 的特征根 $\lambda_0, \lambda_1, \dots, \lambda_p$ 及对应的标准正交化特征向量 $\psi_0, \psi_1, \dots, \psi_p$, 其中 $\psi_i = (\psi_{0i}, \psi_{1i}, \dots, \psi_{pi})$, $i = 0, 1, \dots, p$.

(3) 找出同时都很接近于零的 λ_i , ψ_{0i} . 设 $\lambda_i \approx 0$, $\psi_{0i} \approx 0$, $i = 1, \dots, k-1$, 应用(6.17)和(6.18)计算特征根估计。

当然在应用上, 仍然有一个在何种限度内可以认为 $\lambda_i \approx 0$, $\psi_{0i} \approx 0$ 的问题。Webster等^[55]建议当 $\lambda_i < 0.05$, $|\psi_{0i}| < 0.10$ 时就可认为它们近似等于零了。当然这也只是一种经验上的看法。关于特征根估计的性质的研究所见较少, 很可能由于它与 $A'A$ 而不是 $X'X$ 的特征根和特征向量相联系, 因而困难相对大一些。但是模拟研究结果支持了这种估计。

下面举两个例子说明具体方法

例6.1 Webster-Gunst-Mason数据(续例1.1)

原始数据列在表1.1. 经中心化和标准化后计算得样本相关阵 A' , d 为

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1	0.252	-0.099	0.217	-0.339	0.364	0.811
X_1		1	0.052	-0.343	-0.498	0.417	-0.192
X_2			1	-0.432	-0.371	0.485	-0.317
X_3				1	-0.355	-0.505	0.494
X_4					1	-0.215	-0.087
X_5						1	-0.123
X_6							1

虽然从原始数据的构造知道存在复共线性, 但从上述矩阵是看不出来的, 因为相关系数只度量一对变量之间的相关关系。表6.1列出了 $A'A$ 的全部特征根 λ_i 及 $|\psi_{0i}|$ 。按Webster等的法则, λ_0 和 ψ_{00} 可视为近似等于零。于是公式(6.17)和(6.18)从 $k=1$ 算起, 所得到的

表6.1 $A'A$ 的特征根及 $|\psi_{0i}|$

i	0	1	2	3	4	5	6
λ_i	0.0010	0.0287	0.3115	0.9178	1.1150	2.1816	2.4444
$ \psi_{0i} $	0.0339	0.6987	0.0713	0.0388	0.3406	0.6006	0.1653

表6·2 特征根估计与LS估计的比较

参 数	β_1	β_2	β_3	β_4	β_5	β_6
LS估计	-6.0378	-8.4720	-10.1435	-11.7271	4.0967	9.4505
特征根估计	2.5447	-0.3982	0.2416	-0.3748	4.2125	9.4914
真 值	2	1	0.2	-2	3	10

特征根估计列在表6·2。对于这个问题，数据是用模拟方法算出的，我们知道回归系数的真值(表6·2最后一行)。与回归系数的LS估计相比较，特征根估计比LS估计更接近真值。

Webster等曾经把这个模型的模拟重复了40遍，每次计算出 β_i 的两种估计，最后计算出这40组估计的均值、方差以及均方误差。总的看起来，在 β_1 — β_4 的估计上，特征根估计比LS估计要好得多，对 β_5 和 β_6 的估计，两种方法大体相当。另外，尽管LS估计是无偏估计，但就这40次估计算出的平均值而论，LS估计的偏差反而比特征根估计要大。

例6·2 Hald水泥问题(续例3·3·2)

对Hald水泥数据，前面已经作了多次讨论。这里我们计算它的特征根估计。首先， $A'A$ 阵为

$$\begin{array}{c}
 Y \quad X_1 \quad X_2 \quad X_3 \quad X_4 \\
 \left[\begin{array}{ccccc}
 1 & 0.7307 & 0.8163 & -0.5347 & -0.8213 \\
 & 1 & 0.2286 & -0.8241 & -0.2454 \\
 & & 1 & -0.1392 & -0.9730 \\
 & & & 1 & 0.0295 \\
 & & & & 1
 \end{array} \right]
 \end{array}$$

它的特征根 λ_i 为 $\lambda_0 = 0.0016$, $\lambda_1 = 0.0117$, $\lambda_2 = 0.1990$, $\lambda_3 = 1.5761$, $\lambda_4 = 3.2116$ ，对应的特征向量为

$$\psi'_0 = (0.0408, -0.2617, -0.6523, -0.2657, -0.6586),$$

$$\psi'_1 = (-0.8047, 0.4129, 0.1884, -0.0531, -0.3791),$$

$$\begin{aligned}\psi'_1 &= (0.2112, 0.5809, -0.3899, 0.6747, -0.1039), \\ \psi'_2 &= (-0.0034, 0.5125, -0.4096, -0.6080, 0.4471), \\ \psi'_4 &= (0.5534, 0.4012, 0.4682, -0.3189, -0.4603).\end{aligned}$$

检查 λ_i 及 ψ_{0i} ，依Webster法则， $\lambda_0 = 0.0016 \approx 0$ ， $\lambda_2 = 0.0117 \approx 0$ 。但 ψ_{0i} 中只有 $\psi_{00} = 0.0408 \approx 0$ 。和上例一样，在(6·17)和(6·18)中从 $k=1$ 求和计算特征根估计。计算结果列在表6·3。我们看到，与LS估计相比特征根估计的值改变了不少，特别 β_3 的估计符号也改变了。这说明，自变量 x_3 对因变量 Y 的影响在两种估计中完全相反。

表6·3 Hald水泥问题的特征根估计

参 数	β_1	β_2	β_3	β_4	残差平方和
特征根估计	1.2730	0.2308	-0.1812	-0.4179	48.76
LS估计	1.551	0.5102	0.1019	-0.1441	47.86

§4.7 小结

本章我们讨论了目前被认为最重要的几种有偏估计。除了前面讨论过的性质之外，与LS估计相比，有关这些估计的理论研究结果和实际应用的经验都还很不够。关于这些估计优良性的进一步研究，作得比较多的是计算机模拟试验。例如，McDonald和Galarneau^[46]、Hoerl和Kennard^[45]、Hoerl、Kneard和Baldwin^[49]等对狭义岭估计和LS估计作了比较。Hemmerle和Bran^{tle}^[50]把两种岭估计与LS作了比较。Gunst等^[54]则专门比较了特征根估计和LS估计。Lawless和Wang^[48]对主成分估计、狭义岭估计、广义岭估计和LS估计作了比较。这些大量的模拟研究表明，当设计阵 X 呈病态或者说复共线性存在时，有偏估计确实在均方误差意义下改进了LS估计。但是，在众多的有偏估计中还没

有一个估计被认为优于所有其它估计。一般说来，这些估计的性质好坏与复共线性的严重程度以及回归系数真值在参数空间中的位置有关。

对于复共线性当然还可以应用剔除变量的方法“打破”复共线性。例如，若已经找出回归自变量之间有复共线关系： $x_1 + x_2 + x_3 \approx 1$ ，那么从中解出 $x_3 \approx 1 - x_1 - x_2$ ，代入原来回归模型就消去了自变量 x_3 。但这样做有一个缺点是，复共线关系是对现有 n 组数据而言的，它只是一个经验性结果。如果把模型用于预测，未来的新数据未必就满足这种复共线关系，这样就不能保证有好的预测效果。相比之下，岭估计、主成分估计等不仅排除复共线性对估计的影响，而且对变量之间的关系能够提供更多的信息。这一点在岭迹分析中看得更明显。

关于有偏估计的应用，和其它任何统计方法一样切不可机械地套用现成公式，那种草率粗疏的作法实不可取。如前所述，无论是理论结果还是应用经验都表明，有偏估计仅仅在复共线性存在时优于LS估计。因此，如果经过 $X'X$ 的特征根或条件数的研究并没有发现有复共线性存在，那么鉴于LS估计具有相当丰富的理论结果和应用经验，这时还是宜于使用LS估计。

第五章 其他方法

回归分析是数理统计学中应用最广泛的一个分支。因此，许多年来，对回归分析的理论和方法的研究，也一直很活跃。形形色色的理论模型在文献中被提出，各种带普遍性的方法，适用于某种特定情况下的方法，以至经验性的方法，也不断被理论研究人员和应用工作者提出来。在理论研究工作中，概率极限理论，非参数统计，时间序列分析和随机过程，以至近时发展起来的一些新的统计思想，象“自助”(bootstrap)、“投影追踪”(project pursuit)等，都在日益发挥其作用。

因此，虽然大体上可以说，线性回归模型加最小二乘法，现在仍在回归分析的日常应用上占主导地位，但可以预期，其他模型和方法的应用，必将在回归分析中占到愈来愈重要的地位。例如，上一章所介绍的那些方法，现已日渐为应用工作者所接受和使用。

本章的目的是从回归分析较新近的发展中，挑出一些看来富有实用价值的部分加以介绍。我们将主要限于对基本概念和方法的叙述，有时也提到一些理论结果，但证明大都从略了。因为这些现在大多仍是文献中的研究题目，对本书来说嫌过于专门。

§5.1 权函数估计法

(一) 非参数回归问题

设有 d 维自变量 X 与 1 维因变量 Y 。在本节中, X 、 Y 都假定为随机的。以 $(X_1, Y_1), \dots, (X_n, Y_n)$ 记 (X, Y) 在 n 次观察中所取的值, 总假定它们是独立同分布的。

假定 Y 的均值存在, 即

$$E|Y| < \infty \quad (1.1)$$

则回归函数

$$f(x) = E(Y|X=x) \quad (1.2)$$

存在。对 $f(x)$ 的形状不加任何限制, 就是说, 对 $f(x)$ 一无所知, 要利用样本 $(X_i, Y_i), i=1, \dots, n$, 对指定的 x 值, 去估计 $f(x)$ 。在这里, 由于 f 不是只依赖有限个参数, 故不存在参数估计的问题, 因而也没有相应的参数检验问题。

在统计学上, 这种形式的回归称为**非参数回归**。“参数”和“非参数”构成对统计问题的一种分类; 如果样本分布的数学形式已知, 而只包含若干个未知参数(通常为数不多), 则有关的统计问题(或统计模型)称为**参数性的**, 否则是**非参数性的**。正态线性回归是参数统计模型的一个典型例子。而本节所讨论的问题, 则是非参数统计模型的一个例子。当然, 上述分类的标准并非完全严密, 且人们也不见得总是拘守这种标准。例如, 一元线性回归

$$Y_i = \alpha + \beta x_i + e_i, \quad i=1, \dots, n \quad (1.3)$$

设误差 e_1, \dots, e_n 独立同分布, 有均值 0, 方差 σ^2 非零有限。若进一步假定 $e_i \sim N(0, \sigma^2)$, 则得到正态线性模型。如上所述, 这是

典型的参数模型(其中只包含3个未知参数 α, β, σ^2)。反之,若除了 $E\epsilon_i=0, \text{Var}(\epsilon_i)=\sigma^2, 0<\sigma^2<\infty$ 外,对 ϵ_i 的分布别无所知,则样本 (Y_1, \dots, Y_n) 的分布族并不能用有限个参数去刻画,因而按前述分类标准,(1.3)应列入非参数模型,但习惯上并不这样做。因为在此模型中,人们感兴趣的主要是 $\alpha+\beta x$ 这部分,其中只涉及两个参数。不比在(1.2)中,对 $f(x)$ 毫无所知。即使从它去着眼,也只能把模型视为非参数的。不过,严格地辨明一个统计模型的参数或非参数性并非必要,重要的是模型的确切含义。

讨论非参数统计模型下统计问题的那个数理统计学分支,就称为“非参数统计学”。本节内容可算是这分支的一个组成部分。相对说来,它是非参数统计中较近发展起来的一个方向。

与参数模型相比,非参数模型包罗更为广泛些,或换句话说,在非参数模型中,人为性(或一定程度上的人为性)的假定较少。如(1.3)中,多少是人为地假定了回归为线性,而实际情况可能与此有差距,甚至相去甚远,这就会导致错误。反之,在(1.2)中,对 $f(x)$ 无任何假定,因此也就不会犯这种错误,这是非参数模型的优越性所在。然而,“兴一利必有一弊”。参数统计模型之下往往可以发展有针对性的,因而也是效率较高的统计方法。在非参数模型中,由于假定过于一般,其方法也就难免流于一般,因而效率较低。但是,理论证明:在许多问题中,至少在样本大小很大时(大样本情况),非参数统计方法在效率上的损失并不多,且往往可以和最优良的参数统计方法匹敌。这还不能作为鼓吹以非参数方法代替参数方法的有力理由。因为在现实问题中,使用的样本大小往往不太大,这时,非参数方法的表现如何,人们了解尚不多。

(二) 两种常用的简单估计方法

本段介绍的两种方法,是下一段要介绍的“权函数方法”的简单特例。我们先讨论这两个特例,一则是因为它们都是权函数方法中最常用的,一则是对这两个简单情况的分析,使下一段引进一般的权函数方法显得较为自然和易理解。

先考察一个特例:在样本 $(X_i, Y_i), i=1, \dots, n$ 中,有足够多的 X_i 正好等于指定的 x 。把这些 X_i 记为 X_{i1}, \dots, X_{ir} 。因为它们都等于 x ,故相应的 Y 观察值 Y_{i1}, \dots, Y_{ir} ,是从条件分布 $Y|X=x$ 中抽出的独立同分布样本。此条件分布的均值即 $E(Y|X=x)=f(x)$ 。于是,按“以样本均值估计总体均值”的一般做法,我们可以用

$$m_n^{(1)}(x) = \frac{1}{r}(Y_{i1} + \dots + Y_{ir}) \quad (1.4)$$

去估计 $f(x)$ 。注意, $m_n^{(1)}$ 不仅依赖于指定的 x 值,当然也还与样本 $(X_i, Y_i), i=1, \dots, n$ 有关,故更确切地应记为 $m_n^{(1)}(x, X_1, Y_1, \dots, X_n, Y_n)$ 。为简化记号,在此及以下有关的估计量中,多略去 $(X_i, Y_i), i=1, \dots, n$ 。

这一估计自然而易理解,但它只适用于自变量 X 取离散值的情况,因而意义不大。但这个估计方法启发我们把它稍稍推广一点,以更有实用意义。为此我们这样想:若某个样本 (X_i, Y_i) ,其 X_i 虽不等于 x ,但与 x 差距不大,则这个 X_i 也可选入(即列入前述的 X_{i1}, \dots, X_{ir} 中)。因此,可以在 X 取值的空间 R^d 上引进一个适当的距离 ρ ,而把满足条件 $\rho(x, X_i) \leq \rho_0$ 的 X_i 都收罗进来,此处 ρ_0 是一个适当决定的数。换句话说,若在样本 X_1, \dots, X_n 中,有 X_{i1}, \dots, X_{ir} 与 x 的距离 ρ 不超过 ρ_0 ,则用(1.4)作为 $f(x)=E(Y|X=x)$ 的估计。

距离 ρ 的选择在数学上无多大限制.例如可采用通常的欧氏距离或加权的欧氏距离,即若 $a=(a_1, \dots, a_d)$, $b=(b_1, \dots, b_d)$, 则

$$\rho(a, b) = \left[\sum_{i=1}^d c_i (b_i - a_i)^2 \right]^{1/2} \quad c_i > 0, i=1, \dots, d$$

c_i 为给定常数,其选择可考虑各自变量的单位而定,以免距离基本上取决于个别或少数自变量。也可以用形如 $\max_{1 \leq i \leq d} c_i |b_i - a_i|$ 的距离, c_i 的意义同上。

当选择一定的 ρ_0 时, X_1, \dots, X_n 中满足 $\rho(x, X_i) \leq \rho_0$ 的 X_i 的个数与 x 有关,对某些 x ,个数可能太少。为补救这一点,可考虑不让 ρ_0 固定,而让它与 x 有关,以使上述个数能不少于某个指定的自然数 k ($1 \leq k \leq n$).确切地说,我们对每个 x ,取一个最小可能的 $\rho_0(x)$,使满足 $\rho(x, X_i) \leq \rho_0(x)$ 的 X_i 的个数不小于 k .为方便计,就假定这个数恰好是 k ,且为 X_{i_1}, \dots, X_{i_k} .于是我们用

$$m_n^{(2)}(x) = \frac{1}{k} (Y_{i_1} + \dots + Y_{i_k}) \quad (1.5)$$

去估计 $f(x)$.在这里, X_{i_1}, \dots, X_{i_k} 是 X_1, \dots, X_n 中与 x 点最邻近的 k 个(在距离 ρ 的意义下),称为 x 点的“ k 近邻”。以此之故,估计(1.5)常称为 $f(x)$ 的“ k 近邻估计”。

应当注意的是:两个估计 $m_n^{(1)}(x)$ 和 $m_n^{(2)}(x)$,本质上都是使用与 x 最邻近的那些样本.差别只在于: $m_n^{(1)}(x)$ 是固定距离,因而(1.4)式中的 r 不固定;反之, $m_n^{(2)}(x)$ 是固定个数 k ,因而距离 $\rho_0(x)$ 不固定,而是依赖于 x .

如很多统计方法那样,这种估计方法有其不便的地方,即与(1.4)有关的 ρ_0 ,或(1.5)式中的 k ,该如何选择.统计理论不能提供一种独一无二的选择法,至多只能从大样本角度提供一些一般性的原则,例如当 $n \rightarrow \infty$ 时, ρ_0 应趋于0,但速度不能太快,而 k 应

趋于 ∞ ，速度也不能太快，甚至还可以规定其数量级等。但考虑到在实际问题中， n 并非很大很大，这种原则性的规定并无多大指导意义。因此， ρ_0 与 k 的选择，从根本上说是一个使用经验的问题。从这个意义上看，(1.4)及(1.5)，以及后文提出的更一般的方法，只是提供了一种解决问题的思想而已，而并未形成象最小二乘法那样的有不依人而定的施行格式。这诚然是一个缺点和不便之处，不过也应看到：这是由问题的复杂性而引起的，并非本方法特有的缺点。而且，在方法中规定这点灵活性，有时是有益的。例如，即使最小二乘法，也不能总是硬性地施行，有时要作些修正。采用怎样的修正方法好？这多少也是要凭使用者的经验。

另一个问题牵涉到当维数很高时，样本点在空间 R^d 中的分布很稀疏的现象，即在 ρ_0 不甚大时，(1.4)式中的 ν 与 n 相比很小，以及即使当 k/n 很小时，为使满足条件 $\rho(x, X_i) \leq \rho_0(x)$ 的 X_i 的个数达到 k ， $\rho_0(x)$ 必须很大。这一现象使得当维数 d 很大时，本段（以及下段）的方法实际上无效。因此，本段方法实际上只适用于维数 d 较小的情况。为克服这个困难，有的研究者提出了一些想法，例如所谓“投影追踪法”（PP方法），见[56]。有某些初步证据显示这种方法比传统方法有其优越之处。但到目前为止，关于这类方法的理论成果还不多，它们是否可以作为处理高维数据的有效方法，目前作出定论还为时太早。

（三）一般的权函数估计

由(1.4)式和(1.5)式定义的估计量 $m_n^{(1)}(x)$ 和 $m_n^{(2)}(x)$ ，都是以下的一般形式的特例：

$$m_n(x) = \sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) Y_i \quad (1.6)$$

例如, 在(1.4)式中, 有

$$W_{ni}(x, X_1, \dots, X_n) = \begin{cases} \frac{1}{r}, & \text{当 } \rho(x, X_i) \leq \rho_0 \\ 0, & \text{当 } \rho(x, X_i) > \rho_0 \end{cases} \quad (1.7)$$

而在(1.5)式中, 则有

$$W_{ni}(x, X_1, \dots, X_n) = \begin{cases} \frac{1}{h}, & \text{当 } \rho(x, X_i) \leq \rho_0(x) \\ 0, & \text{当 } \rho(x, X_i) > \rho_0(x) \end{cases} \quad (1.8)$$

也可以这样说: 先按与 x 距离远近, 把 X_1, \dots, X_n 重新排列为 X_{i_1}, \dots, X_{i_n} , 使

$$\rho(x, X_{i_1}) \leq \rho(x, X_{i_2}) \leq \dots \leq \rho(x, X_{i_n}) \quad (1.9)$$

然后令 $W_{ni}(x, X_1, \dots, X_n)$ 等于 $\frac{1}{h}$ 或0, 视 i 等于 j_1, \dots, j_n 中

之一或否而定。

细察(1.6)式, 它是 Y_1, \dots, Y_n 的一个线性式(注意, W_{ni} 不依赖于 Y_1, \dots, Y_n). 因为在实际应用中, 常有

$$W_{ni}(x, X_1, \dots, X_n) \geq 0, \quad \sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) = 1 \quad (1.10)$$

故(1.6)式是 Y_1, \dots, Y_n 的加权平均, 其权 W_{ni} 依赖于 x 和样本 X_1, \dots, X_n . 以此之故, 估计量(1.6)被称为权函数估计。

不难看出: 若假定回归函数为线性的, 即有 $E(Y|X=x) = x'\beta + \alpha$ 的形状, 而用最小二乘法对 α, β 作估计, 得 $\hat{\alpha}$ 和 $\hat{\beta}$. 用 $x'\hat{\beta} + \hat{\alpha}$ 作为回归函数的估计, 则不难验证: $x'\hat{\beta} + \hat{\alpha}$ 可写为(1.6)的形状, 但 $W_{ni}(x, X_1, \dots, X_n)$ 不一定满足(1.10). 例如, 在自变量维数 $d=1$ 时, 有

$$\begin{aligned}
x\hat{\beta} + \hat{a} &= \left\{ \sum_{i=1}^n (X_i - \bar{X}) Y_i / \sum_{i=1}^n (X_i - \bar{X})^2 \right\} \\
&\quad (x - \bar{X}) + \bar{Y} \\
&= \sum_{i=1}^n \left\{ (x - \bar{X})(X_i - \bar{X}) \right. \\
&\quad \left. / \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{n} \right\} Y_i
\end{aligned}$$

因而 $W_{ni}(x, X_1, \dots, X_n) = (x - \bar{X})(X_i - \bar{X}) / \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{n}$, 它

不满足(1·10)的第一式, 但满足第二式。以此之故, 当 C. Stone 在1977年在[57]中引进形如(1·6)的权函数估计类时, 他没有把(1·10)式作为必须的要求。然而, 理论证明: (1·10)的第二式是

重要的。或略放松一点, 也要要求当 $n \rightarrow \infty$ 时, 权的和 $\sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n)$ 在一定意义下收敛于1。不然的话, 就不可能具备最起码的大样本性质(相合性), 因而不可能是一个良好的估计。

从上面与最小二乘估计的对比, 使我们感到这种估计有其合理性。因为从类型上说, 它是最小二乘估计的自然扩张。最小二乘估计相应于一组特殊的权, 它与回归模型的线性假定相应。因此自然地, 当这个线性假定不成立时, 我们应考虑对权函数作更改。如果我们根据某种一般性的考虑(如距离远近)来确定权函数, 则可以预期, 这种权函数估计将不甚依赖回归模型的特殊性, 因而有很广的适用范围。当然, 这种权函数缺乏针对性。这正是我们在第一段中指出的非参数方法的特点。

权函数估计(1·6)可以这样去理解: 在估计 $X=x$ 的条件下 Y 的条件均值时, 让每一个样本都起一定的作用, 但作用的大小与样本 X_i 到 x 点的某种意义下的距离有关。根据这一考虑, 权函数的选定, 就依赖于各样本的作用随距离增加而下降的方式和速度。在

文献中研究较多的有两种类型，分别是(1·4)和(1·5)的推广：

1. 核估计 选定一个定义在 R^d 的实函数 K ，以及适当的常数 $h>0$ ，而令

$$W_n(x, X_1, \dots, X_n) = K\left(\frac{x - X_i}{h}\right) / \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad i=1, \dots, n \quad (1.11)$$

这样定义的权函数 W_n 称为以 K 为核(函数)的核权函数。它显然满足(1·10)的第2式。如果 K 在 R^d 非负(这是常见的情况)，则也满足第1式。

通常选用的核函数 K 具有以下性质： K 在原点处达到最大值，且从原点出发，沿任一方向都是单调下降趋于0，并常是关于原点对称。这保证了(1·11)表达式中的分母不为0，以及当 X_i 与 x 的距离愈远时，权 W_n 愈小。函数 K 的形状决定了权下降的方式，而其下降的速度则可以通过选择 h 来调整： h 愈小，下降愈快。从表面看，表达式 $K\left(\frac{x - X_i}{h}\right)$ 中体现不出 X 的各分量单位之不同在

定义“距离”中的作用。实际上可以这样做：选定了一定的核函数 $K(u) = K(u_1, \dots, u_d)$ 之后，再考虑到 X 的各分量单位不同的情况，将它调整为 $K(c_1 u_1, \dots, c_d u_d)$ ，其中 c_1, \dots, c_d 为适当选定的大于0的常数。我们不妨认为，这步工作已经做了，而 K 就是最后的形式。

估计量(1·4)是(1·11)的特例，其中

$$K(u) = \begin{cases} 1, & \text{当 } \|u\| \leq 1 \\ 0, & \text{当 } \|u\| > 1 \end{cases} \quad (1.12)$$

而取 $h = \rho_0$ 。

与(1·4)相比，这种估计又多了一层不确定性： K 的选择(h 的选择大致相当于(1·4)定义中 ρ_0 的选择)。这是一类估计，如我们

以前讲的，这与其说是提供了一种可以照本施行的估计程序，不如说是提供了一种用以构造估计量的思想。选择 K 的指导性的原则，就是前述关于权随距离下降的方式。(1.12)定义的 K ，所相应的情况是：在一定限度内权不随距离下降，而过此则立降为0。其他可供选择的 K 例如

$$K_1(u) = \begin{cases} 1 - \|u\|, & \text{当 } \|u\| \leq 1, \\ 0, & \text{当 } \|u\| > 1 \end{cases}$$

$$K_2(u) = (1 + \|u\|^r)^{-1}, \quad r > 0 \text{ 为常数}$$

$$K_3(u) = e^{-1/2 \|u\|^2}, \quad r > 0 \text{ 为常数}$$

等等。它们反映的权随距离下降的方式各不相同。至于 h 的选择，如上所述，相当于(1.4)定义中 ρ_0 的选择。一般原则是 h 随 n 增加而单调下降趋于0，但不能太快。大样本理论可以提供我们关于 h 趋于0的数量级的考虑。但这种结论多无补于应用。这是一个主要由情况和经验决定的问题。

2. 近邻型估计 找 n 个常数 $c_{n1}, c_{n2}, \dots, c_{nn}$ ，一般满足

$$c_{n1} \geq c_{n2} \geq \dots \geq c_{nn} \geq 0, \quad \sum_{i=1}^n c_{ni} = 1 \quad (1.13)$$

如在定义估计量(1.5)那样，在 R^d 中引进适当的距离 ρ 。然后按这个距离，把样本 X_1, \dots, X_n 重新排列为 X_{j_1}, \dots, X_{j_n} ，如(1.9)，然后定义

$$W_{nji}(x, X_1, \dots, X_n) = c_{ni}, \quad i=1, \dots, n \quad (1.14)$$

简言之，与 x 最近的样本点，占有最大的权 c_{n1} ，其次近的点，占有次大的权 c_{n2} ，余以次类推。这种定法体现了“与 x 距离愈近的样本，在估计式(1.6)中起的作用愈大”的精神。与核估计不同之处在于，在这里，“最大权”、“次大权”…等都是—定的，并不因样本点的具体位置而异。在核估计则不然。例如估计量(1.4)，在有的情况下，与 x 最近的样本的权与其他样本的权都一样，即为

$\frac{1}{n}$ ，在有的情况下，与 x 最近的样本可占有全部的权，即1，而其他样本的权都是0。

由(1.13)定义的权函数所确定的估计(1.6)称为**近邻型估计**，因为它给与 x 接近的样本以更多的权。如果在(1.13)中有

$$c_{ni}=0, \text{ 当 } i=k+1, k+2, \dots, n \quad (1.15)$$

其中 k 是一个选定的不超过 n 的自然数，则在估计量(1.6)中，只用了与 x 距离最近的 k 个样本，这是估计量(1.5)的推广，在文献中也常把它称为 k 近邻估计。(1.5)相当于 $c_{n1}=\dots=c_{nk}=\frac{1}{k}$ 的特例。

c_{ni} 的选择依据对权下降速度的要求。某些一般性的选择如下：

1. 线性权函数：取 $b_k=k(k+1)/2$,

$$c_{ni} = \begin{cases} (k+1-i)/b_k, & i=1, \dots, k \\ 0, & i=k+1, \dots, n \end{cases}$$

2. 平方权函数：取 $b_k=k(k+1)(4k-1)/6$

$$c_{ni} = \begin{cases} [k^2 - (i-1)^2]/b_k, & i=1, \dots, k \\ 0, & i=k+1, \dots, n \end{cases}$$

等等，这些都满足(1.13)。

(四) 权函数估计的其他应用

以上的讨论是针对估计回归函数的问题。在应用上，有时有必要估计条件分布 $Y|X=x$ 的其他数字特征。例如， $\text{Var}(Y|X=x)$ ， $\text{med}(Y|X=x)$ (条件分布的中位数)之类，权函数估计也可用于处理这种问题。有两种情况：

1. 所要估计的条件数字特征具有 $E(g(Y)|X=x)$ 的形状，或可表为若干个这种量的函数。为估计 $E(g(Y)|X=x)$ ，只须把

$g(Y)$ 看成一个新因变量 Z , 于是 $E(g(Y)|X=x) = E(Z|X=x)$ 就是 Z 对 X 的回归函数。按(1.6), 在选定了适当的权函数 $W_{ni}(x_1, \dots, X_n)$ 后, 它可以用

$$\sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) g(Y_i)$$

去估计之。

例如, 条件方差

$$\text{Var}(Y|X=x) = E(Y^2|X=x) - \{E(Y|X=x)\}^2$$

可以用

$$\sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) Y_i^2 - \left(\sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) Y_i \right)^2$$

去估计。为了保证估计值非负, 应选定权函数满足(1.10)。

2. 如果所要估计的条件数字特征不满足上述情况, 象条件中位数即为一例, 则往往需要先估计条件分布 $Y|X=x$ 本身, 这等于要对每个指定的实数 y 去估计条件概率 $P(Y < y|X=x)$ 。取

$$g(Y) = \begin{cases} 1, & \text{当 } Y < y \\ 0, & \text{当 } Y \geq y \end{cases}$$

则 $P(Y < y|X=x) = E(g(Y)|X=x)$ 。于是按上述, $P(Y < y|X=x)$ 可以用

$$\hat{P}_n(y; x) = \sum_{i=1}^n W_{ni}(x, X_1, \dots, X_n) I_{(Y_i < y)}$$

去估计($I_{(Y_i < y)} = 1$ 或 0 , 视 $Y_i < y$ 或否而定)。

如果选定的权函数 W_{ni} 满足条件(1.10), 则不难看到: 由上式定义的 $\hat{P}_n(y; x)$ 当视为 y 的函数时, 适合分布函数的全部条件。这样, $\hat{P}_n(y; x)$ 就是条件分布 $Y|X=x$ 的分布函数的估计, 我们可以取分布函数 $\hat{P}_n(y; x)$ 的中位数作为条件中位数 $\text{med}(Y|X=x)$ 的估计。

特别, 如果选择形如(1.13)的权函数, 满足条件(1.15), 并

取 $c_{n1} = \cdots = c_{nk} = \frac{1}{k}$, 则 $\hat{P}_n(y; x)$ 就是 Y_{j_1}, \cdots, Y_{j_k} 的经验分布函

数, 即

$$\hat{P}_n(y; x) = Y_{j_1}, \cdots, Y_{j_k} \text{ 中小于 } y \text{ 的个数} \div k$$

其中 j_1, \cdots, j_k 的意义由 (1.9) 式确定。此分布的中位数, 当然也就是 Y_{j_1}, \cdots, Y_{j_k} 的样本中位数。此可称为条件中位数的 k 近邻估计。

我们顺便在此介绍一下权函数方法的另一项应用——用于判别问题, 虽然此问题并不属于本书的范围。

判别问题的提法如下: 按一定标准, 全部对象可以划分为 m 个类, 每个对象必属于某一类且只能属于一个类。如气象预报中, 每日的天气可划入晴、多云、阴、雨等几类之一。但对象所属类别不能直接看出或检出, 而需要通过一些与此有关的, 可以测量的指标去判定之。例如, 判定一个人是否有病, 属于第几期, 是通过一些检查和化验的结果。

把对象所属的类别记为 Y , 则 Y 可以取 $1, 2, \cdots, m$ 这 m 个值。以 X 记该对象的有关指标, X 当然一般是多维的。当给定了 X 之值为 x 时, 并不能唯一决定 Y , 而可以设想它决定了该对象属于各类的(条件)概率, 即

$$p_i(x) = P(Y=i | X=x), \quad i=1, \cdots, m \quad (1.16)$$

引进损失 c_{ij} : 它是当对象确属于第 i 类, 而将其判在第 j 类时, 所遭受的损失。一般有 $c_{ij} = 0$ 当 $i=j$, $c_{ij} > 0$ 当 $i \neq j$ 。据此, 在已测得 $X=x$ 的条件下, 将对象判入第 j 类, 所受的平均损失(条件风险)为

$$l_j(x) = \sum_{i=1}^m p_i(x) c_{ij}, \quad j=1, \cdots, m \quad (1.17)$$

按风险最小的原则, 应取那个 j , 使 $l_j(x)$ 在所有 $l_1(x), \cdots, l_m(x)$

中为最小。这样的判别准则叫做Bayes判别准则。

但一般，(1.16)式定义的 $p_i(x)$ 并非已知。这时(1.17)式中的 $l_i(x)$ 也无法定出，而上述方法无法施行，这就需要有 (X, Y) 的观察样本。也就是说，已有 n 个样品，第 i 个样品之 X 指标值为 X_i ，而其所属类别 Y_i 已确切判定^{*}。这样我们有了样本 $(X_i, Y_i), i=1, \dots, n$ ，通过它们去估计 $p_i(x)$ 。这不难转化为回归函数的估计，只须令 $g(Y)=1$ 或 0 ，视 $Y=i$ 或否而定。则 $p_i(x)=E(g(Y)|X=x)$ ，而 $p_i(x)$ 可以用

$$\hat{p}_{in}(x) = \sum_{j=1}^m W_{nj}(x, X_1, \dots, X_n) I_{(Y_j=i)},$$

$$i=1, \dots, m \quad (1.18)$$

去估计之。此处 $I_{(Y_j=i)}$ 为1或0，视 $Y_j=i$ 或否而定。然后利用 $\hat{p}_{in}(x)$ 对 $l_i(x)$ 作出估计：

$$\hat{l}_{jn}(x) = \sum_{i=1}^m \hat{p}_{in}(x) c_{ij}, \quad j=1, \dots, m \quad (1.19)$$

在 $\hat{l}_{1n}(x), \dots, \hat{l}_{mn}(x)$ 中挑一个最小的。设如是 $\hat{l}_{jn}(x)$ ，则把对象判在 j 类。

这种方法还可以用在其他一些类似的问题，基本思想无甚不同，故不一一细述了。

有的读者或许仍会存在这样的疑问：权函数估计法看来在概念上过于简单粗糙，它在实际问题中到底是否有用，其效率是否过低。人们易于倾向于这样看：方法愈复杂，则愈“高级”，效果也就愈好。其实不尽然，我们可以再举一个例子研究一下。设随机变量 X 有正态分布 $N(a, \sigma^2)$ ， a 和 σ^2 都未知。 c 为一已知常数，要根据 X 的观察样本 X_1, \dots, X_n 去估计概率 $p=P(X < c)$ 。由于 X

*)例如在天气预报问题中，预报日的天气究竟如何，虽在预报时不能确知，但事后可以了然。一个人是否有病在检查时不能确知，日后终见分晓，等等。

的分布已知，在估计 p 时利用这一点是有益的。这个自然的想法是先通过 X_1, \dots, X_n 对 α 和 σ^2 作出估计，然后再用它去估计 p 。经过复杂的论证(见[4]，P85—88)，得到一个形式很复杂的估计量，它在理论上具有某种优越性。然而，直接从“用频率估计概率”这种简单质朴的想法，可得到很易计算的估计量

$$\hat{p} = X_1, \dots, X_n \text{ 中小于 } c \text{ 的个数} \div n$$

从实用的观点看，只要 n 不很小，这个估计的效率，比之前引用复杂方法得出的估计，并无重要差别。而 \hat{p} 除了易于计算外，还有一个很重要的优点：即它不依赖“ X 服从正态分布”这个假定；不论 X 的分布如何， \hat{p} 都是一个可用的估计量，而前引的那个复杂估计则很依赖这个假定。当 X 与正态有差距时，会导致显著的误差。权函数估计比之最小二乘估计，虽则概念上简单得多，但可以预期(这一点可通过模拟来验证)，即使在回归函数真为线性时，只要样本大小不太小，其效果不致比最小二乘估计差多少。而在回归非线性时，此方法仍可用，故实在是一个值得重视的优良方法。

(五) 权函数方法与最小二乘法结合使用

在许多实际问题中，回归函数虽不是很接近于线性，但存在着较强的线性趋势。我们知道，最小二乘法对付线性回归有效，而对与线性差距较大的情形，最小二乘法不适用，权函数方法则是一个可以考虑的代替者。因此就提出如下的想法：把这两个方法结合起来使用，各用其所长而避其所短。具体言之，把回归函数 $f(x) = E(Y|X=x)$ 分解成一个线性部分和一个剩余部分，前者用最小二乘法去估计而后者用权函数方法去估计。然后把二者叠加起来，即得回归函数 $f(x)$ 的估计。从直观上看，这样做应能带来较好的效果。

具体做法如下：设有了样本 $(X_i, Y_i), i=1, \dots, n$ ，用最小二乘法，定出使 $\sum_{i=1}^n (Y_i - \alpha - \beta' X_i)^2$ 达到最小的 α, β ，设为 $\hat{\alpha}, \hat{\beta}$ 。则 $\hat{\alpha} + \hat{\beta}' x$ 就作为回归函数 $f(x)$ 的“线性部分” $\alpha + \beta' x$ 的估计。剩余的部分，即 $f(x) - \alpha - \beta' x$ ，则通过残差

$$\delta_i = Y_i - (\hat{\alpha} + \hat{\beta}' X_i), i=1, \dots, n$$

并利用权函数方法去估计之：选定了权函数 W_{ni} 以后，就用 $\sum_{i=1}^n W_{ni}$
(x, X_1, \dots, X_n) δ_i 去估计 $f(x) - (\alpha + \beta' x)$ 。把二者结合，最后得到 $f(x)$ 的估计量为

$$\hat{\alpha} + \hat{\beta}' x + \sum_{i=1}^n W_{ni}(x, X, X_1, \dots, X_n) \delta_i \quad (1.20)$$

在线性回归之下，残差对于估计回归函数已无用，其作用只在于“诊断”模型有何问题，以及估计误差方差。如果回归并非线性，则残差的形成，不光是由于误差存在，也因为真实的回归函数与线性函数有偏离。因此，用残差去估计这个偏离部分，是自然的事情。

Stone曾提到，模拟的结果显示，(1.20)这种估计比单纯的权函数估计更有效些。这种少量模拟的结果当然不能据以定论。从另一面看，当回归函数接近线性时，这个方法也接近最小二乘法。因此有一些理由希望：这方法在保持最小二乘法优点的同时，增加了稳健性，以使之在回归模型与线性有较大差距时，仍有较好的效果。事情是不是这样还有待进一步的研究。然而，理论上至少证明了这一点：不论回归函数的形状如何，估计量(1.20)当样本大小 $n \rightarrow \infty$ 时，在矩收敛的意义下收敛于 $f(x)$ （见(六)段）。而最小二乘估计就没有这个优点：如果真实的回归函数并非线性，则当 $n \rightarrow \infty$ 时，最小二乘估计在任何意义下都不能收敛回归函数。

拿(1.20)与单纯的权函数估计相比，一则因在实际问题中，

回归函数虽不严格为线性，但不少时候是接近线性，就这一部分而言，用最小二乘法更有效。其二则还有一个心理上的问题。最小二乘法在很长一段时期以来，已牢固地树立了它的地位，为许多实用统计工作者所接受。他们对象权函数方法这种与最小二乘法根本不同的作法，多少会觉得不大放心。而对象(1·20)这样的折衷性质的估计，则更易于理解和接受。

(六) 理论结果

权函数估计虽则在概念上并不复杂，但理论上却有深入发展，当然主要是有关大样本方面的。这些理论虽说并无直接的实用价值，但它澄清了一个问题：即使用这种方法，只要样本大小足够大，总能为所欲为地接近被估计的回归函数(相合性)，这使我们相信：这的确是一种能满足基本要求、有理论根据的合理的估计方法。

这方面的理论，主要是 Stone 1977 年的文章[57]及那以后一些作者的工作。由于本书的性质，我们不打算在此介绍这些工作的细节，尤其是其证明。有兴趣的读者可参考本书所引有关的文献。

1. 矩相合性 一般，设 ξ, ξ_1, ξ_2, \dots 都是随机变量， $r > 0$ 为常数。若 $\lim_{n \rightarrow \infty} E|\xi_n - \xi|^r = 0$ ，则称序列 $\{\xi_n\}$ 依 r 阶矩收敛于 ξ ，记为 $\xi_n \xrightarrow{r} \xi$ (当 $r=2$ 时也常称为均方收敛)。在此，若 ξ 是被估计的量而 ξ_n 是其估计量，则称 ξ_n 为 ξ 的 r 阶矩相合估计。如所周知，若 $r' > r$ ，则由 r' 阶矩相合必推出 r 阶矩相合。又对任何 $r > 0$ ，由 r 阶矩相合必推出弱相合(依概率收敛)。若以概率 1 成立 $\xi_n \rightarrow \xi$ (写为 $\lim_{n \rightarrow \infty} \xi_n = \xi, \text{ a.s.}$)，则称 ξ_n 强收敛于 ξ ，在相应的情况下称 ξ_n 为 ξ 的强相合估计。由强相合必推出弱相合，但强相合与矩相合彼此没有

包含关系。

Stone在[57]中主要研究了由(1.6)式定义的权函数估计 $m_n(x)$ 的矩相合问题。给定 $r > 0$ ，并假定因变量 Y 满足

$$E|Y|^r < \infty \quad (1.21)$$

问在什么条件下有

$$\lim_{n \rightarrow \infty} E|m_n(X) - f(X)|^r = 0 \quad (1.22)$$

此处如前， $f(x) = E(Y|X=x)$ 。Stone相当彻底地解决了这个问题，得到下面的定理：

定理1.1 假定权函数满足条件(1.10)。且设

1°存在有限常数 C ，使对任何非负函数 g (定义于 R^d ， d 是 X 的维数)，当 n 充分大时有

$$E\left(\sum_{i=1}^n W_{ni}(X, X_1, \dots, X_n) g(X_i)\right) \leq CE(g(X))$$

2°对任给 $\varepsilon > 0$ ，当 $n \rightarrow \infty$ 时有

$$\sum_{i=1}^n W_{ni}(X, X_1, \dots, X_n) I_{(|X_i - X| > \varepsilon)} \xrightarrow{P} 0$$

3°当 $n \rightarrow \infty$ 时，有

$$\max_{1 \leq i \leq n} W_{ni}(X, X_1, \dots, X_n) \xrightarrow{P} 0$$

则只要 Y 满足(1.21)，就必成立(1.22)。

有趣的是以上结果的逆也成立(仍在 W_{ni} 满足(1.10)的前提下)：如果由(1.21)总能推出(1.22)(不论 $r > 0$ 如何)，则 W_{ni} 必满足以上三条件1°—3°。

在Stone原来的工作中，没有限制权函数必满足(1.10)，，这时条件要复杂些，除这里列出的三条外还有两条。但这时条件的必要性还不能得到证明，而只能在若干限制下得到证明。此处引述的定理1.1是Stone的一般结果的特例。

以上三条件中，2°和3°易于解释。条件2°表明：当 n 很大时，

与 X 的距离超过一定限度 ε 的那些样本点所占的权的总和,变得很小,且随 $n \rightarrow \infty$ 而趋于0(在依概率的意义下)。这就是说,在估计 $f(x)$ 时,主要依靠的是 X 近傍的那些样本点。因此,权函数估计要有相合性,本质上必是“近邻型”的。条件3°则表明:单个样本的权,在 n 很大时,变得很小。这意思是说:随着样本个数(即 n)的增加,任何一个样本,那怕是与 X 最邻近的那个样本,起的作用也愈来愈小。这两条结合起来,就看出:为了权函数估计有相合性,这估计既只能主要依靠 X 近傍那部分样本点,也不能过分依靠其中任一特定的样本点。这从直观上看是可以理解的。至于条件1°,则没有这么便于解释的意义。

为方便计引进这样一个名词:一个权函数 W_n ,如果能使当(1.21)成立时必能推出(1.22),则称它为**相合权函数**,或说它有相合性。

早在Stone的工作[57]发表之前,有的作者已研究过一些特殊类型的权函数估计。但只在Stone于[57]中得到深刻的结果定理1.1后,这种估计才引起统计界的广泛注意,认为这是一个突破性的进展。不过,就某一具体类型的权函数估计去验证定理中的三个条件,特别是条件1°,并非易事。因此,定理1.1还不能说已经一劳永逸地解决了权函数估计的相合性问题。幸好对两种重要的类型(1.11)和(1.13),问题已有了解决。近邻型权函数(1.13)的问题,也是Stone在[57]中解决的,我们将其表为定理1.2。关于核权函数的问题见[58],这个解决也还只针对很特殊的核函数。

定理1.2 设权函数 W_n 通过(1.9)、(1.13)和(1.14)而定义,则当 c_n 满足条件

$$1^\circ \text{ 对任给 } \varepsilon > 0, \text{ 有 } \lim_{n \rightarrow \infty} \sum_{j > \varepsilon n} c_{nj} = 0$$

$$2^\circ \lim_{n \rightarrow \infty} c_{n1} = 0$$

则 W_{ni} 是相合权函数.

很明显, 本定理的条件 1° 、 2° 分别与定理 1.1 的条件 2° 、 3° 相当. 易验证: k 近邻估计 (1.5) 相应 $c_{n1} = \cdots = c_{nk} = \frac{1}{k}$, $c_{ni} = 0$ 当 $k + 1 \leq i \leq n$, 只要当 $n \rightarrow \infty$ 时 $k \rightarrow \infty$ 而 $k/n \rightarrow 0$, 则定理 1.2 的两个条件满足, 而 (1.5) (当 Y 有 r 阶矩时) 是 $f(x)$ 的 r 阶相合估计.

2. 强相合性 在 Stone 的奠基性工作 [57] 中, 没有讨论这方面的问题. 首先得出一般性结果的是 Devroye 的工作 [59], 其中处理了核权函数和近邻权函数这两种基本情况, 但只限于性质很特殊的核函数. 现引述 Devroye 关于近邻权函数估计的结果如下:

定理 1.3 设因变量 Y 有界, 权函数 W_{ni} 通过 (1.9)、(1.13) 和 (1.14) 定义, 满足以下条件:

$$\lim_{n \rightarrow \infty} k/n = 0, \quad \lim_{n \rightarrow \infty} (\log n)/k = 0$$

$$\{k \max_{1 \leq i \leq k} c_{ni}; n=1, 2, \dots\} \text{ 有界}$$

则有 $\lim_{n \rightarrow \infty} m_n(X) = f(X)$, a. s.

这个结果要求 Y 有界, 这从理论上说是很强的要求. 赵林城和白志东在 [60] 中, 将 Y 的条件降低到只要求某阶 (>1) 矩存在, 但相应地对 k 的要求有所加强. 另外, 本书作者之一在 [61] 中讨论了一般权函数 (不必属于上述两类) 的情况.

这些结果可用于处理种种有关权函数估计的相合性问题. 举两个例子.

首先是 (五) 段中所引进的估计 (1.20), 把这个估计记为 $\hat{m}_n(X)$, 我们可证明下述定理:

定理 1.4 设 X 和 Y 的二阶矩都有限, 又 X 的协方差阵 $A = \text{COV}(X)$ 为 d 阶正定方阵. 则当 W_{ni} 为相合权函数时, 有

$$\hat{m}_n(X) \xrightarrow{P} f(X), \text{ 当 } n \rightarrow \infty \quad (1.23)$$

这就是我们在第(五)段中说过的:不论回归模型是否为线性,由最小二乘法和权函数法结合使用而得到的估计(1·20),总是回归函数 $f(x)$ 的相合估计。

其次是(四)段中所讨论的判别问题。我们曾指出,若取 j ,使由(1·17)定义的 $l_j(x) = \min \{l_1(x), \dots, l_m(x)\}$,则所达到的风险(平均损失)最小。这样的 j 当然依赖于 x ,不妨记为 $d(x)$ 。这个判别量的风险是

$$R = E \left[\sum_{i=1}^m \phi_{i,d(x)}(X) c_{i,d(x)} \right]$$

它称为该判别问题的Bayes风险。

取定权函数 W_{ni} ,而用(1·19)所定义的 $\hat{l}_{jn}(x)$ 去估计(1·17)定义的 $l_j(x)$,其中 $\hat{\phi}_{jn}(x)$ 由(1·18)式定义。取 $d_n(x)$,使

$$\hat{l}_{d_n(x),n}(x) = \min \{ \hat{l}_{1n}(x), \dots, \hat{l}_{mn}(x) \}$$

这个判别的风险是

$$R_n = E \left[\sum_{i=1}^m \phi_{i,d_n(X)}(X) c_{i,d_n(X)} \right]$$

可以证明下面的结果:

定理1·5 如在估计量(1·18)式中, W_{ni} 是相合权函数,则有 $\lim_{n \rightarrow \infty} R_n = R$ 。

这是说,虽然这样决定的判别 d_n 的风险 R_n 一般要大于 R (R 是最小可能的风险),但当 n 很大时, R_n 可任意接近 R 。

§5.2 广义线性模型

(一) 广义线性模型的定义

我们先回忆一下通常线性模型的定义。仍以 X 和 Y 分别记自变

量和因变量, X 可以是多维而 Y 则是一维。设一共进行了 n 次观察, 第 i 次观察时 X 取 x_i 为值, Y 取 Y_i 为值。此处及以下, 把 x_i 视为通常的常数。线性模型的假定包含两方面:

1' 记 $\mu_i = EY_i$, 则 $\mu_i = x_i'\beta$, $i=1, \dots, n$

2' 有关 (Y_1, \dots, Y_n) 之分布的某些要求, 如 Y_1, \dots, Y_n 有等方差, 两两不相关之类。一般, 常要求 Y_1, \dots, Y_n 相互独立, $Y_i \sim N(x_i'\beta, \sigma^2)$, $i=1, \dots, n$.

所谓“广义线性模型”(Generalized Linear Model, 以下简称为 GLM), 是线性模型的推广。其基本假定分别是上面 1', 2' 两条的推广:

1. 存在一个严增可微的函数 g , 使

$$\eta_i = g(\mu_i) = x_i'\beta, \mu_i \text{ 仍为 } EY_i, i=1, \dots, n \quad (2.1)$$

此处的函数 g 称为“联系函数”(link function)

2. Y_1, \dots, Y_n 独立; Y_i 的分布 $f(y_i, \mu_i, \phi)$ 属于指数型。更确切地说, 在固定 ϕ 时, 分布族 $f(y_i, \mu_i, \phi)$ 是带参数 μ_i 的指数型布族。

关于“指数型分布族”的确切意义, 到下一段再讲。现在只指出: 常见的重要分布, 特别是正态、二项、Poisson 等分布, 都属于这种类型。参数 ϕ 的意义, 多与 Y_i 的方差有关。如在通常的正态线性模型中, $Y_i \sim N(x_i'\beta, \sigma^2)$ 。其中 σ^2 就起着此处的 ϕ 的作用。

GLM 产生的实际背景, 我们可以举几个例子。首先, 在应用上常碰到因变量 Y 为属性变量的情况, 即只取 0, 1 这两个值, 或只取 1, \dots, m 这些值之一。在这里, Y 的值其实并无数量的意义, 而只是指示对象所属类别。这时, EY 之值总落在一个有界的范围内(如在上两例中, EY 在 $(0, 1)$ 之内或 $(1, m)$ 之内), 它不可能等于 x 的一个线性函数, 因此拿它作线性模型处理显然不行。但如(取 Y 只取 0, 1 为值的例)引进一个适当的严增函数 g , 把 $(0, 1)$ 区

间变换到 $(-\infty, \infty)$ ，而记 $EY=p$ ，则 $g(p)$ 可取 $(-\infty, \infty)$ 内任何值，因而有可能表成 $x'\beta$ 的形状。一个实际例子如下：以 x 记个体所受的辐射剂量（此处自变量为1维），而 p 记在这个条件下，个体引发某种疾病（如癌症）的概率。在抽样时，一方面记录下所抽个体所受剂量 x ，一方面记录下 Y ， $Y=1$ 或 0 ，视个体发病或否而定。则 $EY=p$ ，而可以考虑采取模型 $g(p)=\alpha+\beta x$ 。例如以下三种函数常被采用：

1. logit: $g(p)=\log\frac{p}{1-p}$

2. Probit: $g(p)=\Phi^{-1}(p)$ ， Φ 为 $N(0, 1)$ 的分布函数

3. log-log: $g(p)=\log(-\log(1-p))$

尤其是前两种。当然，在具体问题中， $g(p)$ 的选定（以及归根到底，GLM是否合用）要参考实际数据。在试验中，把全部试验个体分成 n 组，第 i 组有 m_i 个，施以剂量 x_i ，记录下这 m_i 个中的发病率 Y_i 。则 (x_i, Y_i, m_i) ， $i=1, \dots, n$ ，就是样本。这模型下要处理的统计问题，首要的是估计 α 和 β ，尤其是 β 。它反映发病率与辐射剂量的关系的密切程度。

又如，设想一种工业产品，其质量指标由产品上的缺陷点数 Y 来衡量。 Y 只能取非负整数， $EY \geq 0$ ，不能有 $x'\beta$ 的形式。在一般情况下，有理由假定 Y 服从Poisson分布 $\mathcal{P}(\mu)$ ，而一种简单的模型是 $\log\mu=x'\beta$ ，其中 x 是种种与产品质量有关的因子，如工艺因子，配方因子等。这是所谓“对数线性模型”的一个例子。

由这些例子及类似的例子可以看出，GLM是一个有广泛应用的统计模型。因此，GLM的某些特殊类型及其应用，早在GLM的一般理论发展以前，就在文献中有了大量的反映。这里所描述的这种一般形式的GLM，是Grizzle等于1969年在[62]中提出的。但他们局限于考虑因变量 Y 有Poisson分布的情况。1971年，Dem-

Box在[63]中提出了一般的形式,但局限于考虑所谓“正则联系函数”(见[64])。然后,1972年Nelder和Wedderburn在[65]中考虑了一般情况,并给了GLM这个名称。1983年出版了专著[64],它是目前这方面比较完备的著作(理论方面的讨论较少)。针对GLM的某些特殊情况(如对数线性模型等)也有专著出版。

(二) 指数型分布

我们依照[64],称随机变量 Y 的分布属指数型,若

$$Y \sim f(y, \theta, \phi) d\nu \quad (2.2)$$

而

$$f(y, \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} \quad (2.3)$$

此处 ν 为计数测度或Lebesgue测度。具体说,若 Y 为离散型变量,则 ν 为某个至多为可数点集 A 上的计数测度,而

$$P(Y=y) = \begin{cases} f(y, \theta, \phi), & \text{当 } y \in A \\ 0, & \text{当 } y \notin A \end{cases}$$

若 Y 为连续型变量,则 ν 为Lebesgue测度,而 $f(y, \theta, \phi)$ 就是 Y 的概率密度。

下面是几个常见的例子:

1. $Y \sim N(\mu, \sigma^2)$, 有

$$\theta = \mu, \quad b(\theta) = \theta^2/2, \quad \phi = \sigma^2, \quad a(\phi) = \phi,$$

$$c(y, \phi) = -\frac{1}{2} [y^2/\phi + \log(2\pi\phi)]$$

2. $Y \sim \text{Poisson 分布 } \mathcal{P}(\mu)$, 有

$$\theta = \log \mu, \quad b(\theta) = e^\theta, \quad \phi = 1, \quad a(\phi) = 1,$$

$$c(y, \phi) = -\log y!$$

3. $Y \sim \text{二项分布 } B(n, p)$, 有

$$\theta = \log \frac{p}{1-p}, \quad b(\theta) = n \log(1 + e^\theta), \quad \phi = a(\phi) = 1,$$

$$c(y, \phi) = \log \binom{n}{y}$$

可以证明:

$$EY = \mu = b'(\theta) \quad (2.4)$$

$$\text{Var}(Y) = b''(\theta) a(\phi) \quad (2.5)$$

这里当然要假定有关的导数存在。

证明的关键在于, 等式

$$\int \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} d\nu(y) = 1 \quad (2.6)$$

左边可以在积分号下对 θ 求导。这一点的细节在此从略, 读者可参看[66]§1.2。在此基础上两边对 θ 求导得

$$\int [y - b'(\theta)]/a(\phi) \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} d\nu(y) = 0$$

注意到(2.6)式及

$$EY = \int y \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} d\nu(y)$$

即得(2.4)。在(2.6)的积分号下求两次导数, 即可证得(2.5)。

现引进典则联系函数。设在(2.4)中, 函数 $b'(\theta)$ 为 θ 的严增连续函数。把 $\mu = b'(\theta)$ 的反函数记为 g , 即 $\mu \equiv b'(g(\mu))$, 称 g 为典则联系函数。我们留给读者证明以下几个例子:

1. 正态 $N(\mu, \sigma^2)$: $g(\mu) = \mu$.
2. Poisson $\mathscr{D}(\mu)$: $g(\mu) = \log \mu$.
3. 二项分布 $Y \sim B(n, \mu)$: $g(\mu) = \log \frac{\mu}{n - \mu}$

若取因变量 Y/n , 则 $g(\mu) = \frac{\mu}{1-\mu}$

选用典则联系函数有一个优点, 就是这时存在着简单的充分统计量。事实上, 由(2.3)知, 样本 (Y_1, \dots, Y_n) 的概率密度函数是

$$\exp \left(\sum_{i=1}^n (y_i \theta_i - b(\theta_i)) / a(\phi) \right) \exp \left(\sum_{i=1}^n c(y_i, \phi) \right) \quad (2.7)$$

这里 θ_i 是第 i 次观察时, 参数 θ 所取之值。设第 i 次观察时, 自变量 X 取值 $x'_i = (x_{i1}, x_{i2}, \dots, x_{id})$, 而记 $\beta' = (\beta_1, \dots, \beta_d)$, 则由典则联系, 知

$$\theta_i = g(\mu_i) = x'_i \beta = \sum_{j=1}^d x_{ij} \beta_j, \quad \mu_i = EY_i, \quad i=1, \dots, n \quad (2.8)$$

以此代入(2.7), 得 (Y_1, \dots, Y_n) 的密度为

$$\exp \left\{ \sum_{i=1}^n \beta_j \sum_{i=1}^n x_{ij} y_i / a(\phi) \right\} \exp \left\{ - \sum_{i=1}^n b \left(\sum_{j=1}^d x_{ij} \beta_j \right) / a(\phi) \right\} \exp \left\{ \sum_{i=1}^n c(y_i, \phi) \right\}$$

由此式, 根据因子分解定理, 知在参数 ϕ 值固定的情况下, 线性统计量

$$\left(\sum_{i=1}^n x_{i1} Y_i, \dots, \sum_{i=1}^n x_{id} Y_i \right)$$

是充分统计量, 它也是完备统计量。

当然, 在一个具体问题中是否选用典则联系函数, 主要应根据问题的情况而定, 以上的考虑只是一个方面。

(三) 极大似然估计

我们要根据样本 (Y_1, \dots, Y_n) 对参数 $\beta = (\beta_1, \dots, \beta_d)'$ 作估

计(如果 ϕ 也未知, 则还有估计 ϕ 的问题). 用极大似然法原则上是很简单的: 从(2.7)、(2.8)看出, 要求 β 的极大似然估计 β^* , 归结为在约束(2.8)之下, 使表达式 $\sum_{i=1}^n (y_i \theta_i - b(\theta_i))$ 达到最大值. 这往往只能通过叠代的程序来求近似值. 一个比较简单易行的叠代程序如下:

1° 设 $\eta_1^0, \dots, \eta_n^{(0)}$ 为 η_1, \dots, η_n 的当前值, 算出

$$\mu_i^0 = g^{-1}(\eta_i^0), \quad i=1, \dots, n$$

g^{-1} 表 g 的反函数.

2° 利用样本 Y_1, \dots, Y_n (它在叠代过程中当然是不变的)及上述 η_i^0 和 μ_i^0 之值, 算出

$$z_i^0 = \eta_i^0 + (Y_i - \mu_i^0) g'(\mu_i^0), \quad i=1, \dots, n \quad (2.9)$$

3° 以 x_i 记自变量 X 在第 i 次观察中所取的值. 找 β , 使表达式

$$\sum_{i=1}^n w_i (z_i^0 - x_i' \beta)^2 \quad (w_i = (g'(\mu_i^0))^{-2} (V(\mu_i^0))^{-1})$$

达到最小, 设其解为 $\beta^{(0)}$ (这就是一个以 X 为自变量, Z 为因变量, w_1, \dots, w_n 为权的加权最小二乘估计. 在第一章中已给出过解的公式). 此处 V 的意义是: 由 $\mu = b'(\theta)$ (见(2.4))解出 $\theta = h(\mu)$ (此处要假定 $b'(\theta)$ 为 θ 的严格单调函数). 代入 $b''(\theta)$ 中得 $b''(h(\mu))$, 此函数即记为 $V(\mu)$.

4° 由 $\beta^{(0)}$ 产生 η_1, \dots, η_n 的更新值:

$$\eta_i^{(1)} = x_i' \beta^{(0)}, \quad i=1, \dots, n$$

然后又回到开头1°处. 迭代过程进行到 η 的值变化很小为止. 最后一轮得出的 β 值就是所要求的估计值.

这个叠代程序的根据如下. 记

$$l = y\theta - b(\theta)$$

由于(2.4)式, 也可以把 l 表为 y 和 μ 的函数. 同样, 由于 $\eta=g(\mu)$, l 也可以表为 y 和 η 的函数. 有

$$\begin{aligned}\partial l / \partial \theta &= y - b'(\theta) = y - \mu, \\ \partial l / \partial \mu &= (\partial l / \partial \theta) / (\partial \mu / \partial \theta) = (y - \mu) / V(\mu)\end{aligned}\quad (2.10)$$

后一式是因为 $\partial \mu / \partial \theta = (b'(\theta))' = b''(\theta) = V(\mu)$. 记

$$l_i = Y_i \theta_i - b(\theta_i), \quad L = \sum_{i=1}^n l_i$$

又设与 θ_i 相应的 μ 、 η 值记为 μ_i 、 η_i (具体说, $\mu_i = b'(\theta_i)$, $\eta_i = g(\mu_i)$) . 似然方程

$$\partial L / \partial \beta_j = 0, \quad j=1, \dots, d$$

成为

$$\sum_{i=1}^n ((Y_i - \mu_i) / V_i) \frac{d\mu}{d\eta} \Big|_{\eta=\eta_i} x_{ij} = 0, \quad j=1, \dots, d \quad (2.11)$$

其中 V_i 记 $V(\mu_i)$ (2.11)是根据(2.10), 以及 $\eta_i = \sum_{j=1}^d x_{ij} \beta_j$, 因而

$\partial \eta_i / \partial \beta_j = x_{ij}$. 由于

$$(d\mu/d\eta) \Big|_{\eta=\eta_i} = (g'(\mu_i))^{-1}$$

并注意

$$w_i = [V_i (g'(\mu_i))^2]^{-1}$$

可将(2.11)写为

$$\sum_{i=1}^n w_i (Y_i - \mu_i) g'(\mu_i) x_{ij} = 0, \quad j=1, \dots, d \quad (2.12)$$

把(2.11)的左边表达式再对 β_k 求偏导数, 得

$$\begin{aligned}\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^n (Y_i - \mu_i) \frac{\partial}{\partial \beta_k} (V_i^{-1} \mu'(\eta_i) x_{ij}) \\ &\quad + \sum_{i=1}^n V_i^{-1} \mu'(\eta_i) x_{ij} \frac{\partial}{\partial \beta_k} (y_i - \mu_i)\end{aligned}\quad (2.13)$$

此处 $\mu'(\eta_i)$ 就是 $(d\mu/d\eta)|_{\eta=\eta_i}$ 。现有

$$\begin{aligned}\frac{\partial}{\partial \beta_k} (Y_i - \mu_i) &= -\partial \mu_i / \partial \beta_k \\ &= -\mu'(\eta_i) \partial \eta_i / \partial \beta_k \\ &= -\mu'(\eta_i) x_{ik}\end{aligned}$$

又 $E(Y_i - \mu_i) = 0$ 。故 (2.13) 式两边求均值得

$$\begin{aligned}E\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_k}\right) &= -\sum_{i=1}^n V_i^{-1}(\mu'(\eta_i))^2 x_{ij} x_{ik} \\ &= -\sum_{i=1}^n w_i x_{ij} x_{ik}\end{aligned}\quad (2.14)$$

以 β 记 β 的当前值, b 为其改变量。把 L 考虑为 β 的函数在 β 点展开, 只取到二次项为止。

记 $b = (b_1, \dots, b_d)'$, 有

$$L(\beta + b) \approx L(\beta) + \sum_{j=1}^d \frac{\partial L}{\partial \beta_j} b_j + \frac{1}{2} \sum_{j,k=1}^d \frac{\partial^2 L}{\partial \beta_j \partial \beta_k} b_j b_k$$

并把右端的 $\partial^2 L / \partial \beta_j \partial \beta_k$ 代换为由 (2.14) 决定的 $E(\partial^2 L / \partial \beta_j \partial \beta_k)$, 然后右端对 b_j 求偏导数并命之为 0, $j = 1, \dots, d$, 则由 (2.14), 以及 $\partial L / \partial \beta_j$ 就等于 (2.12) 左端的事实, 将得到决定 b 的方程

$$Ab = C \quad (2.15)$$

其中 $A = (a_{jk})$ 为 d 阶方阵, $a_{jk} = \sum_{i=1}^n w_i x_{ij} x_{ik}$, 而 $C = (c_1, \dots, c_d)'$, c_j 等于 (2.12) 的左边。方程 (2.15) 中的矩阵 A 和向量 C 都可根据观察值 (x_i, Y_i) , $i = 1, \dots, n$ 及 β 的当前值算出 (由 β 决定 η_1, \dots, η_n 的当前值为 $\eta_i = x_i' \beta$, $i = 1, \dots, n$ 。后者决定 $\mu_i = g^{-1}(\eta_i)$, 因而也决定了 w_i)。故解出 (2.15) 得 b , 即找到 β 的下一个值 $\tilde{\beta} = \beta + b$ 。也可直接算 $\tilde{\beta}$,

$$A\tilde{\beta} = A\beta + Ab$$

其第 j 元为

$$\begin{aligned} (A\tilde{\beta})_j &= \sum_{k=1}^d A_{jk} \beta_k + \sum_{i=1}^n w_i (Y_i - \mu_i) g'(\mu_i) x_{ij} \\ &= \sum_{k=1}^d \sum_{i=1}^n w_i x_{ij} x_{ik} \beta_k + \sum_{i=1}^n w_i (Y_i - \mu_i) g'(\mu_i) x_{ij} \end{aligned}$$

注意到 $\sum_{k=1}^d x_{ik} \beta_k = \eta_i$, 即得

$$\begin{aligned} (A\tilde{\beta})_j &= \sum_{i=1}^n w_i x_{ij} (\eta_i + (Y_i - \mu_i) g'(\mu_i)) \\ j &= 1, \dots, d \end{aligned} \quad (2.16)$$

根据第一章关于加权 LS 估计的讨论, 看出: 方程组 (2.16) 的解 $\tilde{\beta}$, 正是在样本 (x_i, z_i) , $i=1, \dots, n$ (其中 $z_i = \eta_i + (Y_i - \mu_i) g'(\mu_i)$) 之下, 以 (w_1, \dots, w_n) 为权的加权最小二乘解, 这正是我们的叠代程序中所规定的.

可以取 Y_i 作为 μ_i 的近似值 μ_i^0 , 算出 $\eta_i^0 = g(\mu_i^0)$, 然后即可由 (2.9) 算出 Z_i^0 .

一个重要的特例是 g 为典则联系函数. 这时, 由

$$\mu = b'(\theta), \quad \eta = \theta$$

知

$$d\mu/d\eta = b''(\theta) = V(\mu) \quad (2.17)$$

这样得到 $V_i^{-1} \mu'(\eta_i) = 1$, 故

$$\partial(V_i^{-1} \mu(\eta_i) x_{ij}) / \partial \beta_k = 0$$

因而由 (2.13) 得

$$\begin{aligned} \partial^2 L / \partial \beta_j \partial \beta_k &= - \sum_{i=1}^n w_i x_{ij} x_{ik} \\ &= E(\partial^2 L / \partial \beta_j \partial \beta_k) \end{aligned}$$

所以在这个情况下, 没有因用 $E(\partial^2 L / \partial \beta_j \partial \beta_k)$ 代替 $\partial^2 L / \partial \beta_j \partial \beta_k$ 而带来误差的问题.

另外, 在这个情况下有

$$\begin{aligned} w_i &= [V_i(g'(\mu_i))^2]^{-1} \\ &= (g'(\mu_i))^{-1} \end{aligned} \quad (2.18)$$

故由(2.12)得似然方程

$$\sum_{i=1}^n x_{ij} Y_i = \sum_{i=1}^n x_{ij} \mu_i, \quad j=1, \dots, d \quad (2.19)$$

上式右边为左边的均值, 这一事实可帮助更直接写出似然方程, 又

注意: $(\sum_{i=1}^n x_{ij} Y_i, j=1, \dots, d)$ 是充分统计量.

(四) 0—1变量情况

这是GLM的一个极重要的例子, 已在(一)中描述过: 所研究的对象可划分两类之一, 有一些指标 X , 与对象所属的类别有关. 将试验单元分成 N 个组, 第 i 组有 n_i 个, 具 X 指标值为 x_i , 试验结果, n_i 个单元中属于第一类的个数记为 Y_i , 以 p_i 记属于第一类的概率, 假定 Y_i 服从二项分布 $B(n_i, p_i)$, $i=1, \dots, N$, 且 Y_1, \dots, Y_N 独立.

p_i 与 x_i 有关, 且假定存在一个连续严增函数 g , 使

$$g(p_i) = x_i' \beta, \quad i=1, \dots, N \quad (2.20)$$

常用的几种函数 g 是(一)中提到的三种情况: logit, probit和log-log. 其中尤以logit最常用. 理由有以下几条: 一是意义清楚, 它反映两类的机会(概率)对比的对数; 二是如后面见到的, 它容许在条件化的基础上进行统计分析; 三是如现在要指出的, 它在先行和后行两种抽样方式的基础上, 得到同一模型.

为说明这最后一点, 先考虑一个简单的情况. 分别以 X 和 \bar{X} 表示对试验对象“施加辐射”和“不施加辐射”, 以 D 和 \bar{D} 表示试验对象“生病”和“不生病”, 四种情况的概率列表如下:

	\bar{D}	D	和
X	p_{00}	p_{01}	$p_{0.}$
X	p_{10}	p_{11}	$p_{1.}$
和	$p_{.0}$	$p_{.1}$	1

不施辐射时，发病与不发病概率比的对数为 $\log \frac{p_{01}}{p_{00}}$ 而施以

辐射时这个量为 $\log \frac{p_{11}}{p_{10}}$ 。二者之差为

$$\Delta = \log(p_{11}p_{00}/p_{01}p_{10})$$

为估计 $\log(p_{01}/p_{00})$ ，可固定不施加辐射的试验对象数，以观察发病与不发病的数目。而为估计 $\log(p_{11}/p_{10})$ ，则要固定施加辐射的试验对象数。

另一方面， Δ 也可表为 $\log(p_{11}/p_{01}) - \log(p_{10}/p_{00})$ 。要估计这两项，就要分别在发病和不发病的对象中，估计其接受辐射的比率。这后一种作法就带有“后行”的性质。

更一般地，在logit的模型中，辐射剂量为 x 的条件下发病概率 $P(D/x)$ 有形式

$$P(D/x) = e^{a+\beta x} / (1 + e^{a+\beta x}) \quad (2.21)$$

在“先行”分析中，是在种种剂量 x 之下观察试验对象中发病的比率。而“后行”分析的作法则是：事后从发病者群中抽出一定的百分率 p_1 ，从不发病者群中抽出一定的百分率 p_0 。在抽出的全部样本中估计一定剂量 x 之下的发病概率 $\tilde{P}(D|x)$ 。按Bayes公式，有

$$\tilde{P}(D|x) = p_1 P(D|x) / (p_1 P(D|x) + p_0 P(\bar{D}|x))$$

利用(2·21), 以及由(2·21)得出的 $P(\bar{D}|x) = 1/(1 + e^{a' + \beta x})$, 得

$$\tilde{P}(D|x) = e^{a' + \beta x} / (1 + e^{a' + \beta x})$$

其中 $a' = a + \log(p_1/p_0)$, 故仍得 logit 模型, 且 β 的值一样, 而常数项相差 $\log(p_1/p_0)$. 若知道 p_1 和 p_0 , 则这个差值可以知道. 特别当 $p_1 = p_0$ 时有 $a' = a$. 又注意上述论证当 x 为多维仍有效.

下面我们就 logit 模型来讨论. 其他模型处理的方法也是一样的.

先讨论 β 的估计问题. (Y_1, \dots, Y_N) 的似然函数, 取对数并去掉与参数无关的加项后, 为

$$\tilde{l}(p, Y) = \sum_{i=1}^N \left(Y_i \log \frac{p_i}{1-p_i} + n_i \log(1-p_i) \right) \quad (2 \cdot 22)$$

利用关系*)

$$\begin{aligned} \log(p_i/(1-p_i)) &= x_i' \beta \\ &= \sum_{j=1}^d x_{ij} \beta_j, \quad i=1, \dots, N \end{aligned} \quad (2 \cdot 23)$$

可以把(2·22)写为

$$l(\beta, Y) = \sum_{i=1}^N \sum_{j=1}^d Y_i x_{ij} \beta_j - \sum_{i=1}^N n_i \log(1 + \exp(\sum_{j=1}^d x_{ij} \beta_j))$$

似然方程组可直接根据(2·19)式和(2·23)式写出(注意, (2·23)为典则联系函数), 为

$$\begin{aligned} \sum_{i=1}^N Y_i x_{ij} &= \sum_{i=1}^N n_i x_{ij} \exp(x_i' \beta) / (1 + \exp(x_i' \beta)) \\ j &= 1, \dots, d \end{aligned} \quad (2 \cdot 24)$$

此方程不易直接求解, 要用在(三)中介绍的叠代法. 叠代程序是:

*) 注意这个写法中没有特别标明常数项, 就是说, 引进了一个恒等于1的自变量

$$z_i = \eta_i + (Y_i - n_i p_i) / (n_i p_i (1 - p_i)), \quad \eta_i = \log \frac{p_i}{1 - p_i}$$

$$i = 1, \dots, N$$

$$w_i = n_i p_i (1 - p_i), \quad i = 1, \dots, N$$

p_i 的初始值可取为 $(Y_i + \frac{1}{2}) / (n_i + 1)$, 这使 $0 < p_i < 1$, 而 η_i 为

有限. 以 $\{w_1, \dots, w_N\}$ 为权, 用 (x_i, z_i) , $i = 1, \dots, N$, 作加权最小二乘法得 β . 由 β 用 $\eta_i = x_i' \beta$ 产生 η_i 的新值, 再开始下一轮的循环.

这个极大似然估计 $\hat{\beta}$ 的大样本性质, 可以在两种情况下去讨论.

1. N 固定, $\min\{n_1, \dots, n_N\} \rightarrow \infty$

若 $X = (x_1 : x_2 : \dots : x_N)$ 有秩 d , 则有

$$(\hat{\beta} - \beta) \sim N_d(0, (XWX')^{-1}) \quad (2.25)$$

这里 W 是一个 N 阶对角形方阵, 主对角线元为 $n_1 p_1 (1 - p_1), \dots, n_N p_N (1 - p_N)$, 而

$$p_i = \exp(x_i' \beta) / (1 + \exp(x_i' \beta)) \quad (2.26)$$

其中 β 是参数真值. 在实际应用中, 可以用上述叠代过程结束时所得到的 β 值来代替之.

2. $N \rightarrow \infty$, n_1, n_2, \dots 有界

这时 (2.25) 仍适用. 但 $\{x_i\}$ 要满足某些从实用的观点看是合理的条件, 包括当 N 充分大时, 上述矩阵 X 有秩 d .

利用 (2.25), 就可以对 β 或其线性函数进行检验及作区间(域)估计, 这当然是近似的. 不存在基于精确分布的小样本方法.

以上的渐近理论可以作两个方面的推广. 一是我们不必假定样本有二项分布, 而只须要求其到二阶为止的矩有性质:

$$EY_i = n_i p_i, \quad \text{Var}(Y_i) = n_i p_i (1 - p_i),$$

$$i=1, \dots, N \quad (2.27)$$

这里只要求 Y_i 的取值限于 $0 \leq Y_i \leq n_i$ ，而不必为整数。另一个推广牵涉到在同一组（指接受同一剂量 x 的那些试验单元）内的观察结果可能不独立。这样， Y_i 当然就不再可能是二项分布，而且其方差一般倾向于比组内观察值独立时为大。例如，同一组内的个体可能很接近，而一种疾病有传染性。这倾向于使 Y_i 取较大和较小之值，从而增大其方差。故一般可设

$$EY_i = n_i p_i, \quad \text{Var}(Y_i) = \sigma^2 n_i p_i (1 - p_i), \quad i=1, \dots, N \quad (2.28)$$

上述第一种情况是第二种情况当 $\sigma^2=1$ 的特例。渐近分布(2.25)可在这种情况下推广为

$$(\hat{\beta} - \beta) \sim N_d(0, \sigma^2 (XW X')^{-1})$$

$\hat{\beta}$ 仍是按前面给出的叠代程序去计算。 σ^2 的估计见后。

（五）续上段：模型的拟合问题

在使用一种统计模型时，总有实际数据与模型拟合程度如何的问题。在GLM中，常用似然比对数的两倍作为数据与模型的拟合优度。对本模型来说，可算出

$$l_1 = \max \left\{ \prod_{i=1}^N \binom{n_i}{Y_i} p_i^{Y_i} (1 - p_i)^{n_i - Y_i} : 0 \leq p_i \leq 1, i=1, \dots, N \right\}$$

$$l_2 = \max \left\{ \prod_{i=1}^N \binom{n_i}{Y_i} p_i^{Y_i} (1 - p_i)^{n_i - Y_i} : p_i = \exp(x_i' \beta) \right.$$

$$\left. / (1 + \exp(x_i' \beta)), i=1, \dots, N, \beta \in R^d \right\}$$

然后算出

$$D(Y) = 2 \log(l_1/l_2)$$

$$= 2 \sum_{i=1}^n \left(Y_i \log \frac{Y_i}{n_i \hat{p}_i} - (n_i - Y_i) \log \frac{n_i - Y_i}{n_i (1 - \hat{p}_i)} \right) \quad (2.29)$$

其中

$$\hat{p}_i = \exp(x_i' \hat{\beta}) / (1 + \exp(x_i' \hat{\beta})), \quad i = 1, \dots, N \quad (2.30)$$

而 $\hat{\beta}$ 是用前述叠代程序求出的, β 的极大似然估计.

可证若 $Y_i \sim B(n_i, p_i)$, $p_i = x_i' \beta / (1 + e^{x_i' \beta})$, $i = 1, \dots, N$, 则当 N 固定, $\min n_i \rightarrow \infty$ 且矩阵 $(x_1 : \dots : x_N)$ 有秩 d 时, $D(Y) \rightarrow \chi^2_{N-d}$. 这可用来检验模型的拟合程度: 指定水平 α , 当 $D(Y) > \chi^2_{N-d}(\alpha)$ 时, 认为数据与模型拟合不好.

可以证明: 在上述模型正确时, 有

$$E(D(Y)) = N - p + \sum_{i=1}^n \frac{1 - p_i(1 - p_i)}{6n p_i(1 - p_i)} + O\left(\frac{1}{n^2}\right) \quad (2.31)$$

Lawley 在 1956 年证明: 若以 $D^0(Y) = D(Y)/(1 + c)$ 代 $D(Y)$, 其中

$$c = (N - p)^{-1} \sum_{i=1}^N [1 - p_i(1 - p_i)] / [6n p_i(1 - p_i)]$$

则 $D^0(Y)$ 的任意阶矩与 χ^2_{N-p} 的同阶矩只相差 $O\left(\frac{1}{n^2}\right)$, 因而有理由期望: 以 $D^0(Y)$ 代替 $D(Y)$ 可改善 χ^2 逼近.

与前面讨论的关于 $\hat{\beta}$ 的极限分布的问题相似, 若不假定诸 Y_i 服从二项分布, 而代之以矩条件 (2.28) (另外, 再加上 Y_1, \dots, Y_N 独立, $0 \leq Y_i \leq n_i$), 则可以证明

$$D(Y) \xrightarrow{\mathcal{L}} \sigma^2 \chi^2_{N-d} \quad (2.32)$$

但由于 σ^2 未知, 这个极限关系已无法用来对数据与模型的拟合程

度进行检验·因为，若 $D(Y)$ 很大，则既可以是由于不拟合，也可以是由于 σ^2 大·但是，(2·32)式启示了 σ^2 的一大样本估计

$$\hat{\sigma}^2 = D(Y)/(N-d) \quad (2\cdot33)$$

但一般更常用的 σ^2 的估计量是

$$\tilde{\sigma}^2 = \frac{1}{N-d} \sum_{i=1}^N (Y_i - n_i \hat{p}_i)^2 / (n_i \hat{p}_i (1 - \hat{p}_i)) \quad (2\cdot34)$$

其中 \hat{p}_i 的计算见(2·30)·

(2·29)定义的 $D(Y)$ 在文献中称为“Deviance”，目前尚未想到合宜的中文译名·这个量可用于模型的比较和选择·简言之，在两个供选择(或一类供选择)的模型中，Deviance小者为佳·举几个例子·自变量 X 的某个分量，例如其第一分量 X_1 ，是与个体的某个指标 t 的值有关， X_1 不一定就取为 t 本身，也可以取为 t^k, e^t 或 $\log t$ 之类·哪一个为好，可由比较Deviance决定·如决定选取 $X_1 = t^k$ ，则 k 值应取为多少，可以通过对一些 k 值算出Deviance，画出 $(k, \text{Deviance}(k))$ 的图形以作选择 k 的依据·又例如联系函数，究竟是取logit形式好，还是取某种别的形式好？也可以通过比较在各种联系函数之下的Deviance来决定(当然，公式(2·29)只适用于logit联系函数·但Deviance的一般定义——似然比对数的两倍——并不局限于这个情况)。

Devicance相当于通常线性模型的RSS(残差平方和)。如我们所知，RSS只是提供了模型拟合优度的一种综合性指标。更细致一些就要使用残差，进行在第二章中所讲的“回归诊断”，其基本思想也可移用于此处讨论的GLM。自然，在细节上更加麻烦。

首先是标准化残差的计算。设观察值 Y_1, \dots, Y_N 要拟合某一特殊类型的GLM。估计出模型中的参数(如0—1 logit模型中的(2·22)一(2·23)中的 β)后，即可算出 EY_i 的估计值 $\hat{\mu}_i$ (如在上述模

型中, $\hat{\mu}_i = n_i \hat{p}_i$, \hat{p}_i 见 (2.30)。则

$$\delta'_i = (Y_i - \hat{\mu}_i) / (\text{Var}(Y_i - \hat{\mu}_i))^{1/2} \quad (2.35)$$

就是 (x_i, Y_i) 这个样本点的标准化残差 (使 $E\delta_i = 0$, $\text{Var}(\delta_i) = 1$)。在通常线性模型的情况, $\text{Var}(y_i - \hat{\mu}_i)$ 可以由 $\text{Var}(Y_i) = \sigma^2$ 经过一个简单的调整而得。对此处 GLM, 我们也使用这一调整, 但在此结果只是近似的。具体做法如下: 得到模型中参数的估值后, 算出权 w_1, \dots, w_N 。如在 logit 模型中就是 $w_i = n_i \hat{p}_i (1 - \hat{p}_i)$, 在一般 GLM 中则按

$$w_i = [V(\hat{\mu}_i)(g'(\hat{\mu}_i))^2]^{-1}$$

计算, 此处 V 的意义是 $V(\mu_i) = \text{Var}(Y_i)$ 。记

$$W^{1/2} = \text{diag}(w_1^{1/2}, \dots, w_N^{1/2})$$

而令 $H = W^{1/2} X' (X W X')^{-1} X W^{1/2}$, 此处仍如前, $X = (x_1 : \dots : x_N)$, 以 h_{ii} 记 H 的 (i, i) 元, 则以 $[(1 - h_{ii}) V(\hat{\mu}_i)]^{1/2}$ 代替 (2.35) 右边的分母, 而将 (2.35) 修正为

$$\delta_i = (Y_i - \hat{\mu}_i) / [(1 - h_{ii}) V(\hat{\mu}_i)]^{1/2}, \quad i = 1, \dots, N \quad (2.36)$$

这就是实际算出的标准化残差, 利用它们, 可以据第二章的方法对模型进行“诊断”。不言而喻, 由于即使在通常线性模型中, 这类做法不少从理论的观点看只是近似, 它们的使用需要有相当的经验, 而且在此处 GLM 的情况下, 所涉及的有关问题比在通常线性模型下要复杂得多。因此, 这一切都无法讲出一个确切的程式, 它是一种实用者可以利用的工具, 当他“善于”使用这一工具时, 可帮助他解决一些与模型的可用性有关的问题, 而不致引入歧途。

(六) 续上段：条件化的方法

在通常的 2×2 列联表中，如样本数较多，则用大样本 χ^2 分布处理。当样本数很少时，Fisher提出了“精确分析”的方法，即在给定表格边缘和的条件下，通过对种种可能的情况计算其条件概率去处理。这个思想也可用于此处带回归自变量的0—1 logit模型。但条件化的作用更主要的是消去我们不感兴趣的多余参数。下面我们将通过具体例子解释这一点

先考虑这样一个概率问题：设 $Y \sim B(n_1, p_1)$, $Z \sim B(n_2, p_2)$, Y, Z 独立。则

$$\begin{aligned} P(Y=y|Y+Z=m) &= \frac{\binom{n_1}{y} p_1^y (1-p_1)^{n_1-y} \binom{n_2}{m-y} p_2^{m-y} (1-p_2)^{n_2-m+y}}{\sum_i \binom{n_1}{i} p_1^i (1-p_1)^{n_1-i} \binom{n_2}{m-i} p_2^{m-i} (1-p_2)^{n_2-m+i}} \\ &= \binom{n_1}{y} \binom{n_2}{m-y} \psi^y / \sum_i \binom{n_1}{i} \binom{n_2}{m-i} \psi^i \end{aligned} \quad (2.37)$$

其中

$$\psi = p_1(1-p_2)/p_2(1-p_1) \quad (2.38)$$

故这个条件分布只依赖于 ψ ，其中就有可能把我们不感兴趣的参数消去掉。

举一个具体例子。设要研究盐的摄入量对患高血压症的影响。在 N 个地区做试验。设在第 i 个地区中，当食盐日摄入量为 x （以某个从卫生的角度看来是合理的值 a 为原点，即实际摄入量为 $a+x$ ）时，高血压发病的概率为

$$p_x = e^{a+\beta x} / (1 + e^{a+\beta x}) \quad (2.39)$$

我们假定 β 的数值与地区无关，它是我们关心的主要对象。由(2.39)知，对摄入标准量($x=0$)的食盐的人来说，高血压发病的概率

为 $e^a/(1+e^a)$ 。因为很可能不同地区高血压的发病率不一样,故此处的 a 与地区有关: 在第 i 地区 a 应改为 a_i 。这样, 在 N 个地区作观察试验的结果, 将产生 $N+1$ 个参数 β, a_1, \dots, a_N , 其数目随 N 增加, 而其实对我们重要的参数只有一个。

现如我们把第 i 地区的接受试验者分为两组: 第一组 n_1 人, 给以食盐量 x_i , 观察到其中患高血压 Y 人。第二组 n_2 人, 给以食盐量 0, 观察到其中患高血压者 Z 人。则按 (2.39) 的记号, 有

$$p_1 = e^{a_i + \beta x_i} / (1 + e^{a_i + \beta x_i}), \quad p_2 = e^{a_i} / (1 + e^{a_i})$$

由 (2.38) 算出 $\psi = e^{\beta x_i}$, 这样就把我们不感兴趣的参数 a_i 消掉了。

有了上述背景, 我们可引进一般的模型: 在自变量取值 x_i 的条件下做试验, 产生符合上述条件的 Y, Z , 记为 Y_i, Z_i , 相应的概率 p_1 和 p_2 则记为 p_{1i} 和 p_{2i} , 而令

$$\psi_i = p_{1i}(1 - p_{2i}) / p_{2i}(1 - p_{1i})$$

又把上文在定义 Y, Z 时的 n_1, n_2 记为 n_{1i}, n_{2i} , 并假定

$$\log \psi_i = x_i' \beta, \quad i = 1, \dots, N \quad (2.40)$$

记 $Y_i + Z_i = m_i$ 。在 $\{Y_i + Z_i = m_i, i = 1, \dots, N\}$ 的条件下, (Y_1, \dots, Y_N) 的对数似然函数为 (去掉与 ψ 无关的加项)

$$\tilde{l}(\psi, Y) = \sum_{i=1}^N Y_i \log \psi_i - \sum_{i=1}^N \log P_{0i}(\psi_i) \quad (2.41)$$

这里我们引进函数

$$P_{ri}(t) = \sum_j j^r \binom{n_{1i}}{j} \binom{n_{2i}}{m_i - j} t^j, \\ r = 0, 1, 2, \dots; i = 1, \dots, N \quad (2.42)$$

以 (2.40) 代入 (2.41), 可将 $\tilde{l}(\psi, Y)$ 写为 β 和 Y 的函数 $l(\beta, Y)$:

$$l(\beta, Y) = \beta' XY - \sum_{i=1}^N \log P_{0i}(e^{x_i' \beta}) \quad (2.43)$$

其中 $X = (x_1 : \dots : x_N)$, $Y = (Y_1, \dots, Y_N)'$

如果引进新参数 $\theta_i = \log \psi_i$, 则拿(2.41)和(2.7)比较, 即可知这时似然函数有标准形式。再按典则联系函数的定义, 即可知 $\theta_i = x_i' \beta$, 也就是 $\log \psi_i = x_i' \beta$, 就是典则联系, 这正与(2.40)符合。因此, 在我们这个(条件化的)模型下, 有典则联系。根据(三)段结尾处所述, 这时似然方程将有形状

$$\sum_{i=1}^N x_{ij} Y_i = E\left(\sum_{i=1}^N x_{ij} Y_i\right), \quad i=1, \dots, N \quad (2.44)$$

据(2.37), Y_i 的(条件)分布为

$$P(Y_i = y | Y_i + Z_i = m_i) = \binom{n_{1i}}{y} \binom{n_{2i}}{m_i - y} \psi_i^y / P_{0i}(\psi_i)$$

其中 $P_{0i}(\psi_i)$ 见(2.42)。再据(2.42), 得

$$\begin{aligned} E(Y_i) &= \frac{P_{1i}(\psi_i)}{P_{0i}(\psi_i)} \\ &= \mu_i \end{aligned} \quad (2.45)$$

故(2.44)写为

$$\sum_{i=1}^N x_{ij} Y_i = \sum_{i=1}^N x_{ij} P_{1i}(\psi_i) / P_{0i}(\psi_i), \quad \psi_i = e^{x_i' \beta} \quad (2.46)$$

这个方程不易直接求解, 应用(三)段中介绍的迭代法, 据(2.15), 得迭代公式:

$$\beta \text{ 的更新值} = \beta \text{ 的当前值} + (X'WX')^{-1} X'(Y - \hat{\mu})$$

此处 $X = (x_1 : \dots : x_N)$, $Y = (Y_1, \dots, Y_N)'$, $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_N)'$, $\hat{\mu}_i$ 是这样算出的: 由 β 的当前值 $\hat{\beta}$ 通过 $\hat{\psi}_i = e^{x_i' \hat{\beta}}$ 算出 $\hat{\mu}_i$, 以 $\hat{\psi}_i$ 代(2.45)中的 ψ_i , 其结果即为 $\hat{\mu}_i$ 。而 $W = \text{diag}(w_1, \dots, w_N)$, 其中

$$w_i = P_{2i}(\hat{\psi}_i) / P_{0i}(\hat{\psi}_i) - P_{1i}^2(\hat{\psi}_i) / P_{0i}^2(\hat{\psi}_i) \quad (2.47)$$

而 $\hat{\psi}_i = e^{x_i' \hat{\beta}}$, $\hat{\beta}$ 为 β 的当前值

为证(2.47)式, 根据此处为典则联系之事实, 用(2.18)式, 得

$$w_i = [g'(\hat{\mu}_i)]^{-1}$$

g 是联系函数^{*}), 据(2.45)及(2.40), 它是由

$$\mu_i = P_{1i}(\psi_i)/P_{0i}(\psi_i), \quad \eta_i = \log \psi_i \quad (2.48)$$

所确定的函数 $\eta_i = g_i(\mu_i)$ 。为了证明由(2.48)确能定出这个函数 g_i , 就必须证明: 函数

$$h(t) = P_{1i}(t)/P_{0i}(t) \quad (2.49)$$

是 $t > 0$ 处的严格增加函数。为此只须证 $h'(t) > 0$ 。但

$$h'(t) = [P_{0i}(t)P'_{1i}(t) - P_{1i}(t)P'_{0i}(t)]/P_{0i}^2(t) \quad (2.50)$$

考察这样一个随机变量 ξ , 其分布为

$$P(\xi = j) = \binom{n_{1i}}{j} \binom{n_{2i}}{m_i - j} t^j / P_{0i}(t), \quad j = 0, 1, \dots$$

则有

$$E\xi = P_{1i}(t)/P_{0i}(t), \quad E\xi^2 = P_{2i}(t)/P_{0i}(t)$$

因而

$$\text{Var}(\xi) = (P_{0i}(t)P_{2i}(t) - P_{1i}^2(t))/P_{0i}^2(t) \quad (2.51)$$

但易见 $P'_{1i}(t) = P_{2i}(t)/t$, $P'_{0i}(t) = P_{1i}(t)/t$ 。由此及(2.50), (2.51), 知

$$h'(t) = \text{Var}(\xi)/t \quad (2.52)$$

因 $\text{Var}(\xi) > 0$, $t > 0$, 知 $h'(t) > 0$, 因而(2.49)所定义的 $h(t)$ 为严增, 这证明了函数 g_i 的存在性。进一步, 利用(2.52), 有

$$d\mu_i/d\psi_i = (\text{Var}(\xi)/t)|_{t=\psi_i}$$

因此

$$g'_i(\mu_i) = d\eta_i/d\mu_i$$

^{*}此处情况较(三)段所讨论的略复杂, 即对每个 i , 联系函数不同。但(三)段处的推理, 包括导致(2.19)式的推理, 显然仍有效。其实, 这个情况在前面早已碰到过了。(在0—1变量logit模型中的联系函数 $\log(p/(1-p))$ 似与 i 无关, 但一写成 Y 均值 $\mu_i = n_i p_i$ 的函数, 就得到 $\log(\mu_i/(n_i - \mu_i))$, 而作为 μ 的函数 $\log(\mu/(n_i - \mu))$ 与 i 有关。

$$\begin{aligned}
&= (d\eta_i/d\psi_i)/(d\mu_i/d\psi_i) \\
&= \frac{1}{\text{Var}(\xi)} \Big|_{t=\psi_i} \\
&= P_{0i}^2(\psi_i)/[P_{0i}(\psi_i)P_{2i}(\psi_i) - P_{1i}^2(\psi_i)]
\end{aligned} \tag{2.53}$$

最后, 由(2.53)及 $w_i = (g_i'(\hat{\mu}))^{-1}$, 即得(2.47)。

这个叠代过程比较复杂。因为在每一个回合, 要计算 $3N$ 个多项式 $P_{ji}(t): j=0, 1, 2; i=1, \dots, N$ 在某点 t 处的值, 当 N 较大时这计算量是很大的。

把 β 的这个估计量记为 $\hat{\beta}$, 它可称为条件极大似然估计。可以证明: 若 N 固定, $\min\{m_i = Y_i + Z_i: i=1, \dots, N\} \rightarrow \infty$ 而 $X = (x_1: \dots: x_N)$ 之秩为 d , 或者当 $N \rightarrow \infty$ 且当充分大时上述矩阵 X 之秩为 d , 则*)

$$(\hat{\beta} - \beta) \sim N_d(0, (XWX')^{-1}) \tag{2.54}$$

其中 $W = \text{diag}(w_1, \dots, w_N)$ 应理解为由(2.47)所决定, 且其中的 $\hat{\psi}_i$ 是叠代終了时之值。

(2.54)式即可用来对 β 或其一线性函数 $c'\beta$ 进行检验, 或作区间(域)估计。这当然是大样本性质的。对个别特殊情况, 也可以采用较方便的方法。例如, 要检验原假设

$$\beta = 0 \tag{2.55}$$

就是说, 个体的指标值 X 对其归类毫无影响。 $\beta = 0$ 相当于

$$\psi_1 = \psi_2 = \dots = \psi_N = 1 \tag{2.56}$$

“指标 X 有效果”一事通常可反映在 Y 值倾向于大或小(相对于(2.55)的情况而言), 于是 $T = \sum_{i=1}^N Y_i$ 可作为检验统计量。在原假

*) 严格地说, (2.54)应理解为在给定 $Y_i + Z_i, i=1, \dots, N$ 的条件下, $(\hat{\beta} - \beta)$ 渐近于 d 维正态分布 $N_d(0, (XWX')^{-1})$ 。

设(2·55)(即(2·56))之下算出 $E(T)$ 和 $\text{Var}(T)$, 分别记为 $E_0(T)$ 和 $\text{Var}_0(T)$ 。由 Y_1, \dots, Y_N 的独立性及(2·45), (2·51)(一切都是在给定 $Y_i + Z_i, i=1, \dots, N$ 的条件下去讨论), 有

$$E_0(T) = \sum_{i=1}^N P_{1i}(1)/P_{0i}(1)$$

$$\text{Var}_0(T) = \sum_{i=1}^N [P_{0i}(1)P_{2i}(1) - P_{1i}^2(1)]/P_{0i}^2(1)$$

因为 $(T - E_0(T))/\sqrt{\text{Var}_0(T)} \sim N(0, 1)$, 可以取

$$|T - E_0(T)| > u_{\alpha/2} \sqrt{\text{Var}_0(T)}$$

为否定域, 此处 $u_{\alpha/2}$ 为 $N(0, 1)$ 的 $\alpha/2$ -上侧分位点。在某种对现象本身的了解的基础上, 也可以使用单边检验。

在通常的线性模型理论中, 当自变量只取少数离散值时, 相应的理论和方法就是方差分析, 它是一种把总变差平方和按各种效应去分解的技术。在GLM中当然也有这样的问题。例如, 设想对概率 $P(Y=1)=p$ 起影响的有两个因子 X_1, X_2 , 分别有 I 和 J 个水平。当 X_1, X_2 分别处在水平 i, j 时, 概率 $P(Y=1)=p_{ij}$ 适合logit线性模型:

$$\log \frac{p_{ij}}{1-p_{ij}} = \mu + \alpha_i + \beta_j \quad (2\cdot57)$$

这里当然可加上约束条件 $\sum \alpha_i = \sum \beta_j = 0$ 。也可以考虑带交互效应的模型。

模型(2·57)是(2·23)的特例。前述的估计参数的方法, 当然也可用于此处估计 μ, α_i, β_j 。为了检验某一效应不存在, 在通常正态线性模型中, 是根据方差分析表中该效应的平方和(及误差平方和)及其自由度, 用F检验法去检验之。在GLM中, 没有这样方便而良好的方法, 但有一种作法, 可比附于通常方差分析中总变差平方和的分解, 叫所谓 **Deviance 分析** (Analysis of Deviance)。拿模型(2·57)来说, 设我们要检验假设

$$\alpha_1 = \cdots = \alpha_j = 0 \quad (2.58)$$

即因子 X_1 没有效应。为此先就模型(2.57)算出其deviance, 记为 D_0 。又在假设(2.58)成立之下, 即对模型

$$\log \frac{p_{ij}}{1 - p_{ij}} = \mu + \beta_j$$

算出其deviance, 记为 D_1 。总有 $D_1 \geq D_0$ 。于是 $D_1 - D_0$ 的大小可以作为假设(2.58)的正确性的度量: $D_1 - D_0$ 愈大, 假设(2.58)愈不象是正确。但是, 由于此处并没有类似于通常方差分析中的F统计量, 无法对给定的检验水平去定出其临界值。所以, 把这个做法直接用于检验的目的还是不行。这种“Deviance分析”的用处, 是一般地考察各种效应的相对重要性。例如, 在一个包含许多因子的模型中, 可以用上述办法分别计算各因子的deviance。凡是deviance大者, 该因子在模型中重要性就大些, 否则就小些。如果希望在所考虑的大量因子中去掉一部分, 则可优先考虑去掉那些其deviance小的因子。关于这类问题, 在文献中有不少讨论, 提出了种种处理方法。由于这些并非基于坚实的理论, 好比残差分析一类的方法, 在一定程度上是凭经验行事的東西。

(七) 其他模型

以上几段对变量取0—1值的模型作了较仔细的讨论。用于估计参数的方法, 就是通常的极大似然估计法。只要给出的模型能使我们写出样本的概率函数, 则极大似然方法就可以用。因此, 其他一些类似的模型, 也可以循着这个途径处理。估计量的渐近分布则可以用参数估计中极大似然估计的一般理论去处理之。近年来还发展了一种所谓拟似然函数法(参看[63]和[64]的第8章), 其中只须对样本的一、二阶矩作适当的假定, 而无须假定它属于指数型分布族。因此, 我们只在下面简略地提一下几种在应用上

重要的模型。

首先是多组分类数据的模型。因变量 Y 只取有限个值 $1, \dots, k$, 这些数只指示对象所属的类。例如人的 ABO 血型有 4 类, 一种工业产品质量可分为 3 个等级, 等等。假定有一些协变量 X , 对所考察对象的所属类别有影响。在 X 取值 $x_i (i=1, \dots, N)$ 的条件下, 对 n_i 个对象观察其所属类别的情况, 结果记为 (Y_{i1}, \dots, Y_{ik}) , 即 n_i 个中有 Y_{ij} 个属于 j 类。假定 (Y_{i1}, \dots, Y_{ik}) 服于多项分布 $M(n_i, p_{i1}, \dots, p_{ik})$, 即

$$P(Y_{i1}=y_{i1}, \dots, Y_{ik}=y_{ik}) = \frac{n_i!}{y_{i1}! \dots y_{ik}!} p_{i1}^{y_{i1}} \dots p_{ik}^{y_{ik}} \quad (2.59)$$

(当 y_{i1}, \dots, y_{ik} 为非负整数, 和为 n_i 。其他情况为 0)

问题是要对 p_{i1}, \dots, p_{ik} 与 x_i 的关系规定一个模型。常采用的模型有以下几种:

1. 记 $\gamma_{ij} = p_{i1} + \dots + p_{ij}$, 令

$$\log(\gamma_{ij}/(1-\gamma_{ij})) = \alpha_i + x_i' \beta_j \quad j=1, \dots, k-1, i=1, \dots, N \quad (2.60)$$

对固定的 i , $\gamma_{ij}/(1-\gamma_{ij})$ 当然是随 j 的增加而上升, 其比为 $e^{\alpha_1} : e^{\alpha_2} : \dots : e^{\alpha_{k-1}} = 1$, 与 x_i 和 β 无关。另一方面, 对固定的 j , $\log(\gamma_{ij}/(1-\gamma_{ij}))$ 是 x_i 的线性函数, 且回归系数 β 与 j 无关。

2. γ_{ij} 如前, 令

$$\log(p_{ij}/(1-\gamma_{ij})) = \alpha_i + x_i' \beta_j, \quad j=1, \dots, k-1, i=1, \dots, N \quad (2.61)$$

这个模型是 Cox 在 1972 年 [67] 中在研究离散时间的寿命数据分析问题时提出的。

当然, 在 (2.60) 和 (2.61) 中, 也可以让 β 依赖于 j 。

3. 对任何 $j \neq j'$, 假定 $\log(p_{ij}/p_{ij'})$ 有线性形状:

$$\log(p_{ij}/\hat{p}_{ij}) = \alpha_{jj'} + x'_i \beta_{jj'}, \quad j, j' = 1, \dots, k, j \neq j' \quad (2.62)$$

由(2.62)不难得到

$$\alpha_{jj'} + \alpha_{j'j''} = \alpha_{jj''}, \quad \beta_{jj'} + \beta_{j'j''} = \beta_{jj''} \quad (2.63)$$

记 $\alpha_{ij} = \alpha_j$, $\beta_{ij} = \beta_j$, 则由(2.63)得到

$$\alpha_{jj''} = \alpha_{j''} - \alpha_j, \quad \beta_{jj''} = \beta_{j''} - \beta_j$$

以此代入(2.62), 并利用关系式 $\sum_{j=1}^k p_{ij} = 1$, 即可得到

$$p_{ij} = \exp(\alpha_j + x'_i \beta_j) / \sum_{r=1}^k \exp(\alpha_r + x'_i \beta_r), \\ j = 1, \dots, k, i = 1, \dots, N$$

在此式中不失普遍性可命 $\alpha_k = 0$, $\beta_k = 0$. 此因上式可改写为

$$p_{ij} = \exp((\alpha_j - \alpha_k) + x'_i(\beta_j - \beta_k)) \\ / \sum_{r=1}^k \exp((\alpha_r - \alpha_k) + x'_i(\beta_r - \beta_k))$$

然后将 $\alpha_j - \alpha_k$ 和 $\beta_j - \beta_k$ 改写为 α_j 和 β_j 即可.

在模型(2.60)、(2.61)中, 需要把这 k 个类排成一个次序. 在不少问题中, 这 k 个类有一个自然的次序. 例如产品质量可分为好、较好, 一般、差等几个等级, 则这时这 k 个类应按这自然的顺序排列. 如果这 k 个类纯粹是名义上的, 没有什么内在联系, 则概率和 y_{ij} 就没有什么意义, 因而不应使用这种模型.

每一个上述这样的模型都容易把 p_{ij} 表为 $x_1, \dots, x_N, \alpha_1, \dots, \alpha_{k-1}, \beta$ (或 $\beta_1, \dots, \beta_{k-1}$) 的函数. 设独立参数的个数为 r , 并记为 $\theta = (\theta_1, \dots, \theta_r)'$, 又记 $x = (x_1, \dots, x_N)'$, 则 p_{ij} 可记为 $p_{ij}(x, \theta)$. 这样可写出 $(Y_{ij}; j = 1, \dots, k, i = 1, \dots, N)$ 的对数似然函数为 (去掉与 θ 无关的加项).

$$L = \sum_{i=1}^N \sum_{j=1}^k Y_{ij} \log p_{ij}(x, \theta)$$

于是得到决定 θ 的极大似然估计的方程:

$$\sum_{i=1}^N \sum_{j=1}^k Y_{ij} (\partial p_{ij}(x, \theta) / \partial \theta_u) / p_{ij}(x, \theta) = 0$$

$$u=1, \dots, r \quad (2.64)$$

对上述几个模型(2.61)–(2.63)来说, 方程(2.64)当然有某种特殊形状, 但总是需要用叠代的方法求解。

可以证明: 在以上几种模型下, 当 N 固定, $\min(n_1, \dots, n_N) \rightarrow \infty$ 且 $(x_1: \dots: x_N)$ 之秩为 r 时, 或 $N \rightarrow \infty$ 而当 N 充分大时上述矩阵的秩为 r 时, 有

$$(\hat{\theta} - \theta) \sim N_r(0, B) \quad (2.65)$$

其中

$$B^{-1} = \sum_{i=1}^N D_i W_i D_i$$

而 $D_i = (d_{iu})$ 为 $k \times r$ 矩阵, $W_i = \text{diag}(w_{i1}, \dots, w_{ik})$, 其中

$$d_{iu} = \partial p_{iu}(x, \theta) / \partial \theta_u |_{\theta = \hat{\theta}}$$

$$w_{iu} = 1 / p_{iu}(x, \hat{\theta})$$

上述模型也可以沿着象在 0—1 变量的情况下往(2.28)的方向推广。这往往是由于同一类之内的观察结果不独立的原因:

$$E Y_{ij} = n_i p_{ij},$$

$$\text{Cov}(Y_{iu}, Y_{iv}) = \begin{cases} \sigma^2 n_i p_{iu}(1 - p_{iu}) \\ -\sigma^2 n_i p_{iu} p_{iv}, & u \neq v \end{cases} \quad (2.66)$$

这里 $\sigma^2 > 0$ 是一个参数。在这个模型下, 前面的估计方法及渐近分布(2.65)仍适用, 但(2.65)中的 B 要改为 $\sigma^2 B$. 这就是我们前面提到的“拟似然函数法”的一个例子。

在“模型正确”的前提下(指(2.66), (2.60), 或(2.61), 或(2.62)), σ^2 可以用

$$\sigma^2 = \chi^2 / (Nk - N - r)$$

$$= \sum_{i=1}^N \sum_{j=1}^k (Y_{ij} - \hat{\mu}_{ij})^2 \hat{\mu}_{ij}^{-1} / (Nk - N - r)$$

去估计。此处

$$\hat{\mu}_{ij} = n_i p_{ij}(x_i, \hat{\theta}), \quad j=1, \dots, k, \quad i=1, \dots, N$$

而当已知 $\sigma^2=1$ 时, χ^2 可以作为模型的拟合优度,在 N 很大时近似地有分布 χ^2_{Nk-N-r} (当模型正确时)。

另一个常见的模型是Poisson模型: Y_1, \dots, Y_N 独立, Y_i 服从Poisson分布 $\mathcal{P}(\mu_i)$, μ_i 与协变量 X 所取的值 x_i 有关。常考虑的是对数线性模型

$$\log \mu_i = x_i' \beta, \quad i=1, \dots, N \quad (2.67)$$

(这里可假定常数项已纳入 β 内)

(2.67)是典则联系。根据(2.19)式, β 的极大似然估计 $\hat{\beta}$ 由方程

$$\sum_{i=1}^N x_i Y_i = \sum_{i=1}^N x_i e^{x_i' \hat{\beta}} \quad (2.68)$$

这个方程用近似法求解的程序,可以按(三)段中的方法。又因此处的 $g(\mu) = \log \mu$, 据(2.18)式, 权 w_i 为 $w_i = \mu_i$ (μ_i 的当前值), 而(2.9)式决定 z_i 的公式为

$$z_i = \eta_i + (Y_i - \mu_i) / \mu_i, \quad \eta_i = x_i' \beta = \log \mu_i \quad (2.69)$$

叠代过程开始时, 取 $\mu_i = Y_i$, $\eta_i = \log Y_i$ (如 $Y_i = 0$, 开始时可舍去这个值)。对数据 (x_i, Y_i) , 以 $w_i = Y_i$ ($i=1, \dots, N$)为权, 作加权最小二乘法得 β 的估值 β_0 。第二步算出 $\eta_i = x_i' \beta_0$, $\mu_i = \exp(x_i' \beta_0)$, 权为 $w_i = \exp(x_i' \beta_0)$ 。这样继续下去, 直到叠代值稳定到某个值 $\hat{\beta}$ 为止。可以证明: 当 $N \rightarrow \infty$, 或 N 固定而 $\min(\mu_1, \dots, \mu_N) \rightarrow \infty$ 时, 有

$$(\hat{\beta} - \beta) \sim N_d(0, (XWX')^{-1})$$

此处 $X = (x_1 : \dots : x_N)$, $W = \text{diag}(\hat{\mu}_1, \dots, \hat{\mu}_N)$, 而 $\hat{\mu}_i = e^{x_i' \hat{\beta}}$, $i =$

1, ..., N, 其中 d 为自变量的维数。

§5.3 截尾回归

(一) 因变量定点截尾

设自变量 X 为 d 维, 因变量 Y 为一维。假定它们符合正态线性模型, 即 $Y = X'\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ 。这模型的统计推断问题已在第一章中仔细讨论过了。但在有些实际问题中, 对因变量的观察受到一定范围的限制。例如在试验时, 仪器只能记录高于(或低于)某点以上(或以下)的 Y 值。或如在社会调查中, 例如调查种种人的收入, 当年收入在10000元以下时, 记录了其收入的具体数字, 而在年收入超过10000元时, 原始记录上只有“超过10000元”。这类例子还可以举出很多。

这样我们就得到模型:

$$\begin{cases} Y_i = x_i' \beta + \varepsilon_i, & i=1, \dots, n \\ \varepsilon_1, \dots, \varepsilon_n \text{ 独立同分布, } \varepsilon_1 \sim N(0, \sigma^2) \\ \text{观察到的样本是 } Z_i = \max(Y_i, 0), & i=1, \dots, n \end{cases} \quad (3.1)$$

这个写法把回归的常数项 α 也纳入 β 中, 即 $\beta = (\beta_0, \beta_1, \dots, \beta_d)'$, $\beta_0 = \alpha$, $x_i' = (1, x_{i1}, \dots, x_{id})$. 由此就容易说明: 若不在 0 点而在另一点 α 截尾, 即

$$Z_i = \max(Y_i, \alpha), \quad i=1, \dots, n$$

也不难化成(3.1)的形状, 只须令 $\tilde{Z}_i = Z_i - \alpha$, $\tilde{Y}_i = Y_i - \alpha$, 则 $\tilde{Z}_i = \max(\tilde{Y}_i, 0)$, 而 $\tilde{Y}_1, \dots, \tilde{Y}_n$ 仍适合(3.1)的前两行, 只是其中的常数项 β_0 改为 $\tilde{\beta}_0 = \beta_0 - \alpha$. 又若截尾的方向是另一边, 即

$$Z_i = \min(Y_i, \alpha), \quad i=1, \dots, n$$

则只须令 $\tilde{Z}_i = a - Z_i$, $\tilde{Y}_i = a - Y_i$, 则得到 $\tilde{Z}_i = \max(\tilde{Y}_i, 0)$, 而 $\tilde{Y}_1, \dots, \tilde{Y}_n$ 仍适合(3.1)前两行, 但 β_0 改为 $\tilde{\beta}_0 = a - \beta_0$, β 改为 $\tilde{\beta} = -\beta$. 甚至也可以考虑每次观察的截尾点 a_i 不同(但已知)的情况, 虽则这种情况在应用中出现较少。事实上, 若 $Z_i = \max(Y_i, \sigma)$, 则令 $\tilde{Z}_i = Z_i - a_i$, $\tilde{Z}_i = Y_i - a_i$, 仍得 $\tilde{Z}_i = \max(\tilde{Y}_i, 0)$, $\tilde{Y}_1, \dots, \tilde{Y}_n$ 仍适合(3.1)的前两行, 但 x_i 要改为 $(-a_i, x_i)$, 而 β' 要改为 $(1, \beta')$. 这里有一个特点是回归系数向量 $(1, \beta')$ 有一个分量已知, 这当然等价于“过原点的回归”的情况。

现来考虑在模型(3.1)之下, β 的极大似然估计。为此, 记

$$R = \{i: 1 \leq i \leq n, Y_i > 0\}, S = \{i: 1 \leq i \leq n, Y_i \leq 0\}$$

并以 $|R|, |S|$ 记 R, S 中所包含的元素数。记

$$f(z, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-z^2/2\sigma^2},$$

$$F(z, \sigma^2) = \int_{-\infty}^z f(u, \sigma^2) du$$

则得 (Z_1, \dots, Z_n) 的似然函数是

$$L = \prod_{i \in S} [1 - F(\beta' x_i, \sigma^2)] \prod_{i \in R} f(z_i - \beta' x_i, \sigma^2)$$

暂记

$$F_i = F(\beta' x_i, \sigma^2), f_i = f(\beta' x_i, \sigma^2)$$

得到

$$\begin{aligned} \log L &= \sum_{i \in S} \log(1 - F_i) - \frac{|R|}{2} \log \sigma^2 \\ &\quad - \sum_{i \in R} \frac{1}{2\sigma^2} (z_i - \beta' x_i)^2 \end{aligned} \quad (3.2)$$

注意到

$$\partial F_i / \partial \beta = f_i x_i, \quad \partial f_i / \partial \beta = -f_i \frac{\beta' x_i}{\sigma^2} x_i \quad (3.3)$$

此处 $\partial F_i/\partial \beta$ 定义为 $(\partial F_i/\partial \beta_1, \dots, \partial F_i/\partial \beta_d)'$ 等等. 又有

$$\begin{aligned}\partial F_i/\partial(\sigma^2) &= (\partial F_i/\partial(\beta'x_i/\sigma))(\partial(\beta'x_i/\sigma)/\partial(\sigma^2)) \\ &= \frac{1}{\sqrt{2\pi}} \exp(-(\beta'x_i/\sigma)^2/2) \beta'x_i (-\sigma^{-3}/2) \\ &= -\frac{1}{2} \sigma^{-2} \beta'x_i f_i\end{aligned}\quad (3.4)$$

$$\begin{aligned}\frac{\partial f_i}{\partial(\sigma^2)} &= -\frac{\sigma^{-3}}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-(\beta'x_i/\sigma)^2/2) \\ &\quad + \frac{1}{\sqrt{2\pi} \sigma} \exp(-(\beta'x_i/\sigma)^2/2) (\beta'x_i)^2 \frac{\sigma^{-4}}{2} \\ &= [(\beta'x_i)^2 - \sigma^2] f_i / (2\sigma^4)\end{aligned}\quad (3.5)$$

由这些关系式算出 $\partial \log L/\partial \beta$ 等, 写出似然方程为

$$\begin{aligned}\partial \log L/\partial \beta &= -\sum_{i \in S} f_i x_i / (1 - F_i) + \sigma^{-2} \sum_{i \in R} (z_i - \beta'x_i) x_i \\ &= 0\end{aligned}\quad (3.6)$$

$$\begin{aligned}\partial \log L/\partial(\sigma^2) &= \sum_{i \in S} (\beta'x_i) f_i / [2\sigma^2(1 - F_i)] \\ &\quad - |R|/(2\sigma^2) + \sum_{i \in R} (z_i - \beta'x_i)^2 / (2\sigma^4) \\ &= 0\end{aligned}\quad (3.7)$$

若以 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 记(3.6)–(3.7)的解, 即 β 、 σ^2 的极大似然估计, 则有

$$\hat{\sigma}^2 = \sum_{i \in R} (z_i - \hat{\beta}'x_i)^2 / [|R| - \sum_{i \in S} (1 - F_i)^{-1} (\hat{\beta}'x_i) f_i]\quad (3.8)$$

此式并没有把 $\hat{\sigma}^2$ 表为 $\hat{\beta}$ 的函数, 因为 F_i, f_i 也与 $\hat{\sigma}^2$ 有关(自然, 在(3.8)右边应以 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 代替 β 和 σ^2). 若 σ^2 假定为已知, 此式提供了一个反复叠代的程序. 但在 β 、 σ^2 都未知时, 上述形式不方便. 1977年, Fair在[68]中导出了一个更方便的公式, 如下: 把(3.6)左乘 $(2\sigma^2)^{-1}\beta'$, 得

$$\begin{aligned}
& - (2\sigma^2)^{-1} \sum_{i \in S} (1 - F_i)^{-1} (\beta' x_i) f_i \\
& + (2\sigma^4)^{-1} \sum_{i \in R} (z_i - \beta' x_i) \beta' x_i = 0
\end{aligned} \quad (3.9)$$

把(3.7)和(3.9)相加, 得

$$- (2\sigma^2)^{-1} |R| + (2\sigma^4)^{-1} \sum_{i \in R} (z_i - \beta' x_i) z_i = 0$$

解出

$$\hat{\sigma}^2 = |R|^{-1} \sum_{i \in R} (z_i - \hat{\beta}' x_i) z_i \quad (3.10)$$

(3.6)乘以 σ^2 , 得

$$\left(\sum_{i \in R} x_i x_i' \right) \beta = \sum_{i \in R} x_i Y_i - \sum_{i \in S} (1 - F_i)^{-1} \sigma^2 f_i x_i \quad (3.11)$$

令 $R = \{i_1, \dots, i_{|R|}\}$, $S = \{j_1, \dots, j_{|S|}\}$, 以及

$$\begin{aligned}
X_1 &= \begin{bmatrix} x_{i_1}' \\ \vdots \\ x_{i_{|R|}}' \end{bmatrix}, \quad Z_1 = \begin{bmatrix} z_{i_1} \\ \vdots \\ z_{i_{|R|}} \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_{j_1}' \\ \vdots \\ x_{j_{|S|}}' \end{bmatrix} \\
\hat{v} &= -\hat{\sigma}^2 \begin{bmatrix} (1 - F_{i_1})^{-1} f_{i_1} \\ \vdots \\ (1 - F_{j_{|S|}})^{-1} f_{j_{|S|}} \end{bmatrix}
\end{aligned} \quad (3.12)$$

则可将(3.11)改写为

$$X_1' X_1 \hat{\beta} = X_1' Z_1 + X_2' v \quad (3.13)$$

得其解(要假定 $(X_1' X_1)^{-1}$ 存在)

$$\hat{\beta} = (X_1' X_1)^{-1} (X_1' Z_1 + X_2' v) \quad (3.14)$$

(3.10)和(3.14)一起构成一个叠代程序: 从 β, σ^2 的一组初始值 $\beta^{(0)}, (\sigma^{(0)})^2$ 出发, 以后者代替(3.10)和(3.14)右边表达式中的 $\hat{\beta}, \hat{\sigma}$, 算出的 $\hat{\beta}$ 和 $\hat{\sigma}^2$, 即(3.14)和(3.10)的左边, 记为 $\beta^{(1)}$ 和 $(\sigma^{(1)})^2$. 然后以后者为初始值, 又得到第二步叠代值。这样下去, 直到叠代值稳定下来为止。

现考察(3.2)定义的函数 $\log L$ 作为 β 和 σ 的函数的性状。由于

我们假定了 $(X_1'X_1)^{-1}$ 存在, 且 Y_1, \dots, Y_n 服从正态分布, 知以概率 1 成立

$$\min \left\{ \sum_{i \in R} (z_i - x_i' \beta)^2 : \beta \in R^d \right\} > 0 \quad (3.15)$$

而对使上式成立的样本值, 显然有

$$\lim_{\sigma \rightarrow 0+} \log L = -\infty, \quad \lim_{\sigma \rightarrow \infty} \log L = -\infty$$

对 $\beta \in R^d$ 一致成立。因此存在一个区间 $[a, b]$ ($0 < a < b < \infty$, a, b 与样本值有关), 使得在寻求函数 $\log L$ 的最大值点时, 只须考虑 $a \leq \sigma \leq b$ 内的 σ 。另一方面, 易见当 $X_1'X_1$ 的逆矩阵存在时, 有

$$\lim_{\|\beta\| \rightarrow \infty} \sum_{i \in R} (z_i - x_i' \beta)^2 = \infty \quad (3.16)$$

因此当 $\|\beta\| \rightarrow \infty$ 时, 对 $[a, b]$ 内的 σ , 一致地有 $\log L \rightarrow -\infty$ 。这说明: $\log L$ 必在某个有限点 (β^*, σ^*) 处达到最大值, 而后者必然是似然方程的一解。这个分析证明了: 以概率 1 极大似然估计存在, 且必为似然方程之一解。但并未证明似然方程只有唯一的一组解 (这一点看来是正确的), 因而也就未能证明: 当上述叠代过程收敛到一组值时, 它就是极大似然估计, 上述分析肯定了: σ^2 的极大似然估计必大于 0。但表达式(3.10)的右边可以为负。因此, 当上述叠代过程由(3.10)算出负值时, 说明初始值取得不当, 而必须中止叠代, 另行选择初始值。一个可以考虑的初始值是 $\beta^{(0)} = (X_1'X_1)^{-1}X_1'Z_1$, $(\sigma^{(0)})^2 = \sum_{i \in R} (z_i - x_i' \beta^{(0)})^2 / (|R| - d)$, 即以 $\{(x_i, z_i) : i \in R\}$ 用通常最小二乘法算出的结果。

Dempster等于[69]中提出了一种所谓EM程序, 它包含两步。第一步在模型中的样本值不完全和有缺落时, 用一定办法补上。第二步则利用这已修补的数据进行统计分析 (这种思想, 在处理方差分析中的缺落值(missing value)时, 早就用过了)。拿模型(3.1)来讨论。令 $Y_{(2)} = (Y_{11}, \dots, Y_{1n})'$ 。若观察因变量 Y 时不在 0

点处截尾，则将得 $Y_{(2)}$ 为

$$Y_{(2)} = X_2\beta + e_{(2)} \quad (3.17)$$

$e_{(2)} = (e_{11}, \dots, e_{j1,1})'$. 经过截尾, $Y_{(2)}$ “失落”了。现在我想把它补上。方法是以 $e_{(2)}$ 的期望值代 $e_{(2)}$ ——当然是在 $Y_{(2)} \leq 0$ 的条件下的条件期望值。这等于要在 $\xi \sim N(0, \sigma^2)$, a 为常数, 要算 $E(\xi | \xi \leq -a)$. 它显然等于

$$\begin{aligned} E(\xi | \xi \leq -a) &= \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{-a} t e^{-t^2/2\sigma^2} dt \\ &\quad / \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{-a} e^{-t^2/2\sigma^2} dt \\ &= -\sigma^2 f(a, \sigma^2) / F(-a, \sigma^2) \\ &= -\sigma^2 f(a, \sigma^2) / (1 - F(a, \sigma^2)) \end{aligned} \quad (3.18)$$

由此, 注意到(3.12)式中的 v 的定义, 有

$$E(e_{(2)} | Y_{(2)} \leq 0) = v$$

在完全模型的决定 β 的方程

$$(X_1'X_1 + X_2'X_2)\beta = X_1'Z_1 + X_2'Y_{(2)} \quad (3.19)$$

中, 以 $X_2\beta + v$ 代替“失落值” $Y_{(2)}$, 消去方程两边的公共项 $X_2'X_2\beta$, 即得(3.14)。由此可见, EM 方法导致与我们上述方法一样的结果。值得注意的是: 此处的“补入值” $X_2\beta + v$ 与未知参数 β 有关, 这与通常在方差分析中以已知数值递补失落值者不同。

(二) 被截尾数据被丢弃时

上段讨论的模型(3.1)属于这种情况: 数据中有被截尾的, 但是, 虽然经过截尾后数据原值已经失落, 但我们还是记下了“这个数据曾被截尾”的情况。另一个情况是: 数据一经截尾就完全被丢弃了。例如, 在研究人的年收入与一些因素的关系时, 不考

愿那些年收入超过 2 万元的人。

一般，仍设自变量 x 与因变量 Y 符合正态线性模型 $Y = x'\beta + e$, $e \sim N(0, \sigma^2)$. 在 x_1, \dots, x_N 处作 N 次观察，结果本来是

$$Y_i = x_i'\beta + e_i, \quad i = 1, \dots, N \quad (3.20)$$

但实际上，我们只是当 $Y_i \leq L_i$ (L_i 为某个预先给定的常数) 时才保留 Y_i ，否则就不要了。把 (3.20) 中经过这个手续保留下来的那 n 个就记为 Y_1, \dots, Y_n ，则似然函数应是

$$\begin{aligned} L &= \prod_{i=1}^n (Y_i \text{ 的密度函数在 } Y_i \text{ 点之值}) / P(Y_i \leq L_i) \\ &= \prod_{i=1}^n \left(\frac{1}{\sigma} \varphi \left(\frac{Y_i - x_i'\beta}{\sigma} \right) \right) / \Phi \left(\frac{L_i - x_i'\beta}{\sigma} \right) \end{aligned} \quad (3.21)$$

此处 φ , Φ 分别为标准正态分布 $N(0, 1)$ 的密度函数和分布函数。

取 $\log L$ ，对 β 和 σ^2 求偏导数，得似然方程

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= \sigma^{-2} \sum_{i=1}^n (Y_i - x_i'\beta) x_i + \sigma^{-1} \sum_{i=1}^n \varphi \left(\frac{L_i - x_i'\beta}{\sigma} \right) x_i \\ &\quad / \Phi \left(\frac{L_i - x_i'\beta}{\sigma} \right) \\ &= \sigma^{-2} \sum_{i=1}^n \left(Y_i - x_i'\beta + \varphi \left(\frac{L_i - x_i'\beta}{\sigma} \right) \right) \\ &\quad / \Phi \left(\frac{L_i - x_i'\beta}{\sigma} \right) x_i \\ &= 0 \end{aligned} \quad (3.22)$$

$$\frac{\partial \log L}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (Y_i - x_i'\beta)^2 + \frac{\varphi \left(\frac{L_i - x_i'\beta}{\sigma} \right)}{\Phi \left(\frac{L_i - x_i'\beta}{\sigma} \right)}$$

$$\frac{L_i - x_i' \beta}{\sigma^2} = 0 \quad (3.23)$$

记

$$\begin{aligned} X' &= (x_1 : \dots : x_n), \quad Y' = (Y_1, \dots, Y_n) \\ u &= (u_1, \dots, u_n)', \quad u_i = \frac{\varphi\left(\frac{L_i - x_i' \beta}{\sigma}\right)}{\Phi\left(\frac{L_i - x_i' \beta}{\sigma}\right)} \sigma \end{aligned} \quad (3.24)$$

由(3.22), (3.23), 得

$$\beta = (X'X)^{-1}(X'Y + X'u) \quad (3.25)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \left[(Y_i - x_i' \beta)^2 + \frac{\varphi((L_i - x_i' \beta)/\sigma)}{\Phi((L_i - x_i' \beta)/\sigma)} \sigma (L_i - x_i' \beta) \right] \quad (3.26)$$

(3.25)和(3.26)组成一个决定 β 和 σ^2 的叠代程序。如前,初始值可取为数据 (x_i, Y_i) 在通常最小二乘法之下对 β 和 σ^2 的估计值。

方程(3.25)有如下的解释,它也就体现了EM算法在本模型中的应用:要是没有截断,第 i 次观察结果 Y_i 的均值应是 $x_i' \beta$ 。有了截断, Y_i 在第 i 次观察(设它没有超过 L_i ,因为在 Y_i 超过 L_i 时它就被丢弃了)的均值应是

$$\begin{aligned} & x_i' \beta + E(e_i | e_i \leq L_i - x_i' \beta) \\ &= x_i' \beta - \sigma \varphi\left(\frac{L_i - x_i' \beta}{\sigma}\right) / \Phi\left(\frac{L_i - x_i' \beta}{\sigma}\right) \\ &= x_i' \beta - u_i \end{aligned}$$

因此,所记录下的 Y_i ,比它应该有的值,“平均说来”,少算了 u_i 这么多。故把数据 Y_i 修改为 $Y_i + u_i$,对经过修改后的数据使用最小二乘法,就得到(3.25)式。与上段的情况一样,在此经过修改后的“数据” $Y_i + u_i$,并不是已知数,而是与被估计的参数有

关。

本段所讨论的模型及其应用,可参看[70]、[71]和[68]等参考文献。

另外,顺便在此提一下截尾点为随机的情况。设自变量 x (为简便计假定为一维)和因变量 Y 满足线性回归模型

$$Y_i = \alpha + \beta x_i + e_i, \quad i=1, \dots, n, \quad e_1, e_2, \dots \text{独立同分布。}$$

另有一串独立同分布的随机变量 T_1, T_2, \dots , 且 $\{e_i\}$ 与 $\{T_i\}$ 也独立。使在第 i 次观察时, Y_i 在 T_i 点被截尾, 即我们只观察到

$$Z_i = \min(Y_i, T_i), \quad i=1, \dots, n$$

记 $\delta_i=1$ 或 0 , 视 $Y_i > T_i$ 或否而定。可分为两个情况: 1° 除知道 Z_i 外也知道 δ_i , 即知道各次观察是否被截尾了。2° 不知道 δ_i , 即只观察到截尾后之值 Z_i , 而不知是否被截尾了。

如果给出了 e_i 和 T_i 的概率密度(例如, $e_i \sim N(0, \sigma^2)$ 而 T_i 有密度 $\lambda^{-1}e^{-t/\lambda}I_{(t>0)}$), 则在以上两种情况下, 都不难写出样本的似然函数, 从而写出似然方程。

如果 e_i, T_i 的密度函数的形式未知, 就属于非参数情况, 极大似然估计法无法使用。这种情况在应用上不多见, 目前也无有效的方法去处理它, 此处就不多说了。读者可参看 Miller [72]、Buckley和James[73], 以及Koul 等[74]等人的工作。

(三) 定数截尾的情况

这种情况常出现在有关元件(或生物体等)的寿命试验中。以 Y 记元件的寿命。 Y 的分布依赖于若干因子(例如, 制造该元件时的工艺和配方因子) X , X 为 d 维。更具体地说, Y 的分布依赖于一个回归项*) $X'\beta$, 可能还有一个反映分布(或 Y 经过某种变换后

*) 此处为书写简便, 仍将常数项纳入 β 中。

的分布)的刻度的参数 σ . Y 的密度函数和分布函数分别记为 $f(y; X'\beta, \sigma)$ 以及 $F(y; X\beta, \sigma)$. 现在 X 的 n 个值 x_1, \dots, x_n 处做试验, 所得结果分别记为 Y_1, \dots, Y_n . 由于这 n 个试验中, 有的可能持续很长时间. 为避免这一点, 常指定一个 r , 使得在得出 r 个试验结果后即停止试验. 换言之, 若以 $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(r)}$ 记 Y_1, Y_2, \dots, Y_n 的次序统计量, 则在观察了 $Y_{(1)}, \dots, Y_{(r)}$ 后即停止. 要利用试验结果去估计 β , 可能还有 σ .

为书写方便计, 不妨设 Y_1, \dots, Y_n 的次序统计量就是 Y_1, \dots, Y_n . 这只要经过把原试验的编号加以适当调整即可实现.

用极大似然法估计 β 和 σ , 原则上很简单: $(Y_{(1)}, \dots, Y_{(r)})$ 的对数似然函数是

$$\begin{aligned} \log L = & \sum_{i=1}^r \log f(Y_i; x_i' \beta, \sigma) \\ & + \sum_{i=r+1}^n \log [1 - F(Y_i; x_i' \beta, \sigma)] \end{aligned} \quad (3.27)$$

以此就不难写出似然方程. 当然, 这种方程一般只能用迭代法求解.

我们来考察两个最重要的例子.

一个例子是寿命 Y 服从指数分布, 有密度函数 $\lambda^{-1} e^{-y/\lambda} (y > 0)$. $\lambda = E(Y)$ 是平均寿命. 假定 $\log \lambda = X' \beta$. 因此这事实上是一个 GLM. 有

$$\begin{aligned} f(y; X\beta) &= \exp(-X'\beta) \exp(-ye^{-X'\beta}) \\ F(y; X\beta) &= 1 - \exp(-ye^{-X'\beta}) \end{aligned}$$

以此代入 (3.27), 计算 $\partial \log L / \partial \beta$, 得似然方程为

$$\begin{aligned} & - \sum_{i=1}^r x_i + \sum_{i=1}^r Y_i x_i \exp(-\sum_{i=1}^r Y_i x_i' \beta) \\ & + Y_r \sum_{i=r+1}^n x_i \cdot \exp(-Y_r \sum_{i=r+1}^n x_i' \beta) = 0 \end{aligned} \quad (3.28)$$

用通常的牛顿法迭代求解，程序是：设 $\beta^{(0)}$ 是 β 的当前值。则 $\beta^{(0)}$ 的一步改变量 b 由

$$b = -H_0^{-1}h_0 \quad (3.29)$$

决定。其中 h_0 为在(3.28)的左边用 $\beta^{(0)}$ 代替 β 所得向量，而 H_0 的 (j, k) 元为

$$\begin{aligned} & -\exp\left(-\sum_{i=1}^r Y_i x_i' \beta^{(0)}\right) \sum_{i=1}^r Y_i x_{ij} \sum_{i=1}^r Y_i x_{ik} \\ & -\exp\left(-Y_r \sum_{i=r+1}^n x_i \beta^{(0)}\right) Y_r^2 \sum_{i=r+1}^n x_{ij} \sum_{i=r+1}^n x_{ik}, \quad j, k=1, \dots, d \end{aligned}$$

另一个例子是寿命 Y 服从对数正态分布，即 $\log Y \sim N(X\beta, \sigma^2)$ 。仍设 Y_1, \dots, Y_n 的次序统计量就是 Y_1, \dots, Y_n ，并把后者转换到 $Z_i = \log Y_i, i=1, \dots, n$ 。则 (Z_1, \dots, Z_r) 的似然函数是

$$L = \sigma^{-r} \prod_{i=1}^r \varphi((z_i - x_i' \beta)/\sigma) \prod_{i=r+1}^n [1 - \Phi((z_i - x_i' \beta)/\sigma)]$$

此处 φ 和 Φ 是 $N(0, 1)$ 的密度函数和分布函数。似然方程是

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= \sigma^{-2} \sum_{i=1}^r (z_i - x_i' \beta) x_i + \sum_{i=r+1}^n \sigma^{-1} (1 - \Phi_i)^{-1} \varphi_i x_i \\ &= 0 \end{aligned} \quad (3.30)$$

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma} &= -\frac{r}{\sigma} + \sigma^{-3} \sum_{i=1}^r (z_i - x_i' \beta)^2 + \sum_{i=r+1}^n \sigma^{-2} (1 - \Phi_i)^{-1} \\ &\quad \varphi_i (z_i - x_i' \beta) \\ &= 0 \end{aligned} \quad (3.31)$$

此处

$$\varphi_i = \varphi((z_i - x_i' \beta)/\sigma), \quad \Phi_i = \Phi((z_i - x_i' \beta)/\sigma)$$

把(3.30)式写成行向量的形式，右乘 $-\beta$ ，并将所得结果加到乘以 σ 后的(3.31)式，即得

$$-r + \sigma^{-2} \sum_{i=1}^r (z_i - x_i' \beta) z_i + \sum_{i=r+1}^n \sigma^{-1} (1 - \Phi_i)^{-1} \varphi_i z_i = 0 \quad (3.32)$$

由(3.32)解得

$$\sigma^2 = \frac{1}{r} \sum_{i=1}^r (z_i - x_i' \beta) z_i + \frac{\sigma}{r} \sum_{i=r+1}^n (1 - \Phi_i)^{-1} \varphi_i z_i \quad (3.33)$$

又把(3.30)写为

$$\left(\sum_{i=1}^r x_i x_i' \right) \beta = \sum_{i=1}^r z_i x_i + \sigma \sum_{i=r+1}^n (1 - \Phi_i)^{-1} \varphi_i x_i \quad (3.34)$$

一如在(一)段中讨论的情况那样, (3.33)、(3.34)构成一个决定 (β, σ^2) 的迭代程序: 由 (β, σ^2) 的当前值 $(\beta^{(0)}, \sigma_0^2)$ 出发, 以此代入(3.33)、(3.34)的右边。由(3.33)定出 σ^2 的新值 σ_1^2 , 而由(3.34)定出 β 的新值 $\beta^{(1)}$ 。

不难看出: (3.34)也是EM算法的一个例子。事实上, 若 $i > r$, 则在无截尾时, z_i 应为 $x_i' \beta + e_i$ 。可现在 z_i 必 $\geq z_r$, 故算出 $E(e_i | x_i' \beta + e_i \geq z_r)$ (x_i, z_r 都视作常数)不难得到

$$E(e_i | x_i' \beta + e_i \geq z_r) = \sigma(1 - \Phi_i)^{-1} \varphi_i, \quad i = r+1, \dots, n$$

若无截尾, 则决定 β 的方程应为

$$\left(\sum_{i=1}^n x_i x_i' \right) \beta = \sum_{i=1}^r z_i x_i + \sum_{i=r+1}^n z_i x_i$$

现通过截尾, z_{r+1}, \dots, z_n 已失落, 以 $x_i' \beta + E(e_i | x_i' \beta + e_i \geq z_r) = x_i' \beta + \sigma(1 - \Phi_i)^{-1} \varphi_i, i = r+1, \dots, n$ 代替, 得

$$\begin{aligned} & \sum_{i=1}^r x_i x_i' \beta + \sum_{i=r+1}^n x_i x_i' \beta \\ &= \sum_{i=1}^r z_i x_i + \sum_{i=r+1}^n x_i (x_i' \beta + \sigma(1 - \Phi_i)^{-1} \varphi_i) \end{aligned}$$

两边消去 $\sum_{i=r+1}^n x_i x_i' \beta$, 即得(3.34)。

在寿命试验时, 常作定时截尾。即试验到指定时刻即终止。这时, n 组试验中有 r 组已做完(其元件在指定时刻前损坏)。所得

数据与我们这里所考虑的情况，除一点以外都相同：不同之处是在定时截尾时 r 为随机的。虽然如此，在应用上仍用定数截尾一样的方法去处理。

(四) 有重复的情况

上段所讨论的截尾试验有一个缺点：有时，从设计的观点考虑，希望在指定的 n 个点 x_1, \dots, x_n 处都取得数据，而经过截尾，则某些 x_i 处的值被“截”掉了。但如能在每个 x_i 处重复试验若干次，而截尾则是在每个 x_i 处的观察值去进行，则仍可缩短试验时间，并弥补上面提到的缺点。

具体说，设在 $X=x_i$ 时进行 m_i 次试验，所得结果记为 Y_{i1}, \dots, Y_{im_i} ， $i=1, \dots, n$ 。假定这 $m_1 + \dots + m_n$ 个结果全体独立。现对每一组 Y_{i1}, \dots, Y_{im_i} 实行截尾（定数或定时）。为方便计，设截尾所得结果是 Y_{i1}, \dots, Y_{in} ， $i=1, \dots, n$ 。

如果知道 Y 的密度 $f(y; X'\beta, \sigma)$ ，则不难写出 (Y_{i1}, \dots, Y_{in}) ， $i=1, \dots, n$ 的似然函数，因而可以写出似然方程，以决定 β 和 σ 的极大似然估计。

这里我们介绍另外一种做法。它有一个要求：即将 Y 经过某种变换（例如 $Z = \log Y$ ）变到 Z 后， $(Z - X'\beta)/\sigma$ 的分布为一个与 β 、 X 和 σ 都无关且均值为0的已知分布 F 。设由 Y 到 Z 的变换 $Z = g(Y)$ 是 Y 的严增函数（严降的情况可类似处理），则 Z_{i1}, \dots, Z_{im_i} 就是 Z_{i1}, \dots, Z_{im_i} 的截尾。

设 $r_1 + \dots + r_n$ 个随机变量 e_{i1}, \dots, e_{ir_i} ， $i=1, \dots, n$ 独立同分布，公共分布为 F 。以 $e_{i1} \leq \dots \leq e_{ir_i}$ 记 e_{i1}, \dots, e_{ir_i} 的次序统计量。则从概率分布的角度，可以把样本 Z_{ij} 等表为

$$Z_{ij} = x_i'\beta + \sigma e_{ij}, \quad j=1, \dots, r_i, \quad i=1, \dots, n \quad (3.35)$$

对固定的 i ， Z_{i1}, \dots, Z_{ir_i} 构成线性模型，由之可以对 $x_i'\beta$ 作一个伊

计。按第一章的方法，作出其方差最小的线性无偏估计。为此，以 V_i 记 $(\tilde{e}_{i1}, \dots, \tilde{e}_{ir_i})$ 的协方差阵。由于分布 F 已知， V_i 在原则上是可以算出的。记 $Z_i = (Z_{i1}, \dots, Z_{ir_i})'$, $1 = (1, \dots, 1)'$ 。找 a ，使 $(Z_i - al)'V_i^{-1}(Z_i - al)$ 达到最小。这个 a 记为 U_i ， U_i 有

$$U_i = c_{i1} Z_{i1} + \dots + c_{ir_i} Z_{ir_i} \quad (3.36)$$

的形状， c_{i1}, \dots, c_{ir_i} 为常数。 $\text{Var}(U_i)$ 有 $h_i^2 \sigma^2$ 的形状，其中 h_i^2 可以由 (3.36) 式中的系数 c_{ij} 及已知的分布 F 算出（事实上， $h_i^2 = c_i' V_i c_i$ ， $c_i = (c_{i1}, \dots, c_{ir_i})'$ ）。

这样，我们有

$$U_i = x_i' \beta + \varepsilon_i, \quad i = 1, \dots, n \quad (3.37)$$

其中 $\varepsilon_1, \dots, \varepsilon_n$ 独立，有均值 0，而 $\text{Var}(\varepsilon_i) = h_i^2 \sigma^2$ 。这是一个在第一章中已讨论过的那种线性模型，由之可得到 β 的无偏估计 $\hat{\beta}$ ，后者可表为 U_1, \dots, U_n 的线性函数，因而可表为 $\{Z_{ij}; j = 1, \dots, r_i, i = 1, \dots, n\}$ 的线性函数。同样可得到 σ^2 的无偏估计。

§5.4 M估计法

（一）稳健性的一般概念

本节要介绍的 M 估计法是一种所谓稳健方法，因此先对这个概念作点论述。在理论上探讨一种统计方法的性质时，是从一定模型出发。例如线性模型

$$Y_i = x_i' \beta + e_i, \quad i = 1, \dots, n \quad (4.1)$$

假定 e_1, \dots, e_n 独立且各服从正态分布 $N(0, \sigma^2)$ 。在这个假定下，我们就可以论证某种统计方法有怎样的优良性——例如， $c' \beta$ 的最小二乘估计是在一切估计类中，方差一致最小的估计； $c' \beta = 0$ 的假设可通过具有 F 分布的统计量去检验等等。然而，在实际问题

中，模型与理论上的要求严格符合的情况，可以说是没有的。就(4.1)来说，虽然在不少问题中， ϵ_i 的分布确是近似于正态，但与正态多少会有些差距，它们的方差不见得完全一样，也不见得是严格地独立等等。

这种情况就自然地引起一个要求：即所使用的统计方法，应具备一定的“抗干扰性”。就是说，当实际模型与理论上的假定略有背离时，该方法仍能保持较好的性能。不然的话，该方法在理论假定下的优良性就完全是纸面上的，不仅没有实际意义，还可能把使用者引入歧途，这是统计方法稳健性意义的一个方面。这个意思多少有些象力学中“稳定平衡”与“不稳定平衡”之间的差别，虽则这比喻不完全贴切。

统计方法的稳健性还有另一方面的意思，这与数据有关。在统计理论上探讨一个方法的性质时，我们假定：数据确是从理论模型所规定的分布中的随样抽样。但在实际问题中，特别在数据量较多时，偶尔会发生少量的所谓“过失误差”(gross error).例如写错一位数字，把小数点的位置打错了等等。即使一个性质很优良的统计方法，但如数据中有受到过失误差影响的成份，则其使用效果也会受到影响。稳健性的另一方面的意思就在于：当数据中只有少量受到过失误差的影响时，使用效果不致受到太大的干扰。

以上的描述是定性的。在稳健理论中，也定义了某些定量的指标来衡量稳健性。但从实用观点看，应该说这主要是一个相比较而存在的概念：不存在什么“最稳健的”统计方法。但是，给了两个处理同一统计推断问题的统计方法，可以从种种角度去考察谁的稳健性更大一些。例如，要从样本 $X_1, \dots, X_n \sim N(\alpha, \sigma^2)$ 去估计均值 α .两个常用的方法是样本均值 \bar{X} 和样本中位数 m 。设样本中有个别值因有过失误差而变得异常的大，则它对 \bar{X} 有相当

的影响,而使估计结果偏高,但对 m 则没有影响。因此从“抗过失误差”这个角度看, m 的稳健性优于 \bar{X} 。而且,稳健性还必须与这方法的总的性能结合起来考察。如在本例中,取常数作为估计值(与样本无关),则它抗过失误差的能力最强。但它的总的性能很差——实际上是一个无用的估计,因而也就不能认为它有稳健性。样本中位数 m 则不然:它作为正态均值的估计量,一般性能也堪称良好,因而可算是一个稳健估计。

稳健性的概念萌芽于本世纪初,但其发展,则主要是自五十年代后期以来,经过Tukey, Huber, Hampel等一批学者的工作。Huber在1981年出版了这方面的专著[75]。在统计学近代发展的早期,有一个有兴趣的争论问题。自1916年以来,伟大的统计学家R. A. Fisher曾就 $N(a, \sigma^2)$ 中 σ 的估计问题,与天文学家Eddington发生过争论: Fisher主张用 S (S^2 是样本方差),而Eddington主张用“平均绝对差” $d = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$ 。到1920

年, Fisher提出了一个支持使用 S 的有力论据: S 是 σ 的充分统计量(当时尚没有这个词, Fisher的论据中包含这个意思),因而这场争论以有利于Fisher而结束。但到五十年代, Tukey从稳健性的角度重新考察了这个问题,发现当总体分布与正态略有偏离时, S 受的影响颇大而 d 就小些,即 d 的稳健性优于 S ,而这又构成支持使用 d 的一个有力论点。

在线性回归中,其所以有必要考虑稳健方法,是因为最小二乘法受到异常值的影响较大,而此则是由于 x^2 是一个随 $|x|$ 增长很快的函数。这里所谓“异常值”,其产生除了通常的过失误差外,还由于当(4.1)中的误差 ϵ_i 不是正态分布,而是属于一种所谓“重尾型”(heavy-tailed)分布,即尾部概率较大时,也会倾向于易产生异常大的观察值。在目前文献中出现的一些稳健性的回归方

法中，只有Huber提出的M方法较有实用价值，且较易实施。下面我们只限于介绍这个方法。

(二) M估计的定义

最小二乘法要求找 β ，使 $\sum_{i=1}^n (Y_i - x_i'\beta)^2$ 达到最小。如上所述，正因为平方函数增长太快，使这个和，从而使最小二乘估计 β ，受个别异常值的影响很大。为了减轻这种影响，可改用增长较慢的函数 ρ 代替平方函数，例如，取 $\rho(x) = |x|$ ，则得到下述估计 β 的方法：找 β ，使 $\sum_{i=1}^n |Y_i - x_i'\beta|$ 达到最小。现已有一些文献讨论这种估计。顺便说一句：在估计分布的对称中心的问题中，这个方法引出的估计量就是样本中位数。我们前已提到：这是一个稳健性较好的估计。

由以上的考虑，引进如下的一般定义：设函数 $\rho(x)$ 定义于 $-\infty < x < \infty$ ，满足以下的条件：

- 1° $\rho(x)$ 在 $(-\infty, \infty)$ 处处连续。
- 2° $\rho(0) = 0$ ， $\rho(x)$ 在 $(-\infty, 0)$ 非增，在 $(0, \infty)$ 非降
- 3° $\lim_{x \rightarrow \infty} x^{-2}\rho(x) = 0$

记

$$D(\beta) = \sum_{i=1}^n \rho(Y_i - x_i'\beta) \quad (4.2)$$

若 β^* (β^* 当然与 $x_i, Y_i, i=1, \dots, n$ 有关) 满足

$$D(\beta^*) = \min_{\beta} D(\beta) \quad (4.3)$$

则称 β^* 为 β 的一个M估计。

由此可知，M估计不是指一个确定的估计，而是指一类估计，与所选的函数 ρ 有关。

M估计名称的来由,是因为这种估计在形式上与极大似然估计(缩写为MLE)有相似之处。事实上,若模型(4.1)中误差 e_1, \dots, e_n 独立同分布,且 e_1 有已知的概率密度函数 f ,则 (Y_1, \dots, Y_n) 的对数似然函数为 $\sum_{i=1}^n \log f(Y_i - x_i' \beta)$ 。记 $\rho = \log f$,则 β 的MLE就是在取函数 ρ 时的M估计。

在 x 的维数 $d=1$,且 $x_1=x_2=\dots=1$ 的特例,模型(4.1)就是估计位置参数。对这个特例,M估计的概念是1964年Huber在[76]中引进的。Huber在[76]中研究了这种情况下M估计的大样本性质。直到现在为止,这仍是稳健方法研究得最深入且最有实用价值的一种情况。

不言而喻,M估计的性质,与 ρ 的选择有很大关系。从根本上说, ρ 的良好选择取决于误差的分布,但误差的分布一般未知,故只能在若干种有代表性的类型中去挑选。后面我们将列出几个例子。一般说来,在文献中并没有关于 ρ 应满足的条件公认的表述。但看来任何合理的定义,都应该满足我们上面提出的三个条件1°—3°。

如果 $\psi(x) = \rho'(x)$ 处处存在,则为使(4.3)成立, $\beta^* = (\beta_1^*, \dots, \beta_d^*)'$ 必须满足下述方程:

$$\sum_{i=1}^n \psi(\delta_i) x_{ij} = 0, \quad j=1, \dots, d \quad (4.4)$$

此处已记 $x_i' = (x_{i1}, \dots, x_{id})$,而

$$\delta_i = Y_i - \sum_{j=1}^d x_{ij} \beta_j^*, \quad i=1, \dots, n \quad (4.5)$$

为了证明方程组(4.4)有解,且解 β^* 满足(4.3),还需对 ρ 作一些限制。有下面的定理:

定理4.1 假定 $\rho(x)$ 为定义在 $(-\infty, \infty)$ 的非负凸函数, $\rho'(x) = \psi(x)$ 处处存在,且

$$\lim_{\|x\| \rightarrow \infty} \rho(x) = \infty \quad (4.6)$$

又设矩阵

$$A = \begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \cdot & \cdot & \cdots & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

之秩为 d , 则方程组 (4.4) — (4.5) 必有解, 其解必满足 (4.3)。反过来, 任何满足 (4.3) 的 β^* 必是 (4.4) — (4.5) 的解。若进一步假定 $\rho(x)$ 为严格凸函数, 则方程组 (4.4) — (4.5) 的解唯一。

证 先证明以下两点预备事实:

1° 当 $\|\beta\|^2 = \sum_{i=1}^d \beta_i^2 \rightarrow \infty$ 时有 $D(\beta) \rightarrow \infty$. $D(\beta)$ 由 (4.2) 定义. 事

实上, 记 $C = \{\beta: \beta \in R^d, \|\beta\| = 1\}$, 而

$$K(\beta) = \|A\beta\|^2 = \beta' A' A \beta$$

则因 A 之秩为 d , 知 $A' A$ 为 d 阶正定方阵, 故有 $\min\{K(\beta): \beta \in C\} = q > 0$. 因而当 $\|\beta\| \rightarrow \infty$ 时有 $K(\beta) = \|\beta\|^2 K(\beta/\|\beta\|) \geq q \|\beta\|^2 \rightarrow \infty$.

又因

$$\sum_{i=1}^n |Y_i - x'_i \beta| \geq \left(\sum_{i=1}^n (Y_i - x'_i \beta)^2 \right)^{1/2} \geq \|A\beta\| - \|Y\|$$

由 $\lim_{\|\beta\| \rightarrow \infty} K(\beta) = \infty$ 知

$$\lim_{\|\beta\| \rightarrow \infty} \sum_{i=1}^n |Y_i - x'_i \beta| = \infty$$

这与 (4.6) 结合, 即得到所要的结果。

2° $D(\beta)$ 为 β 的凸函数. 若 ρ 为严凸, 则 D 也是严凸. 事实上, 设 $c_1 > 0$, $c_2 > 0$, $c_1 + c_2 = 1$, 而 $\beta^{(i)} \in R^d$, $i = 1, 2$, $\beta^{(1)} \neq \beta^{(2)}$. 记 $\beta = c_1 \beta^{(1)} + c_2 \beta^{(2)}$, 则

$$D(\beta) = \sum_{i=1}^n \rho(Y_i - x'_i \beta)$$

$$\begin{aligned}
&= \sum_{i=1}^n \rho[c_1(Y_i - x_i' \beta^{(1)}) + c_2(Y_i - x_i' \beta^{(2)})] \\
&\leq \sum_{i=1}^n \{c_1 \rho(Y_i - x_i' \beta^{(1)}) + c_2 \rho(Y_i - x_i' \beta^{(2)})\} \\
&= c_1 D(\beta^{(1)}) + c_2 D(\beta^{(2)}) \quad (4.7)
\end{aligned}$$

这证明了 $D(\beta)$ 为凸的。因 $\beta^{(1)} \neq \beta^{(2)}$, 故必存在 i , 使 $x_i' \beta^{(1)} \neq x_i' \beta^{(2)}$ (这由矩阵 A 之秩为 d 推出), 故当 ρ 为严凸时, 对这个 i , 有

$$\rho[c_1(Y_i - x_i' \beta^{(1)}) + c_2(Y_i - x_i' \beta^{(2)})] < \sum_{i=1}^n c_i \rho(Y_i - x_i' \beta^{(1)})$$

因此 (4.7) 式成为严格不等号。这证明了 $D(\beta)$ 为严凸的。

现转到定理的证明。由 $D(\beta)$ 为 β 的连续函数及预备事实 1°, 知 $D(\beta)$ 的最小值必在有限点处达到。再由 ρ 有导数知 $D(\beta)$ 对 β 有一阶偏导数。由此可知, 使 $D(\beta)$ 达到最小的 β^* , 必满足方程组 (4.4) — (4.5) (这部分证明不需要 ρ 的凸性)。

其次证明: 方程组 (4.4) — (4.5) 的任一组解 $\beta^* = (\beta_1^*, \dots, \beta_d^*)'$ 必是 $D(\beta)$ 的最小值点。因若不然, 则存在 $\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_d)'$ 使 $D(\tilde{\beta}) < D(\beta^*)$. 考虑函数

$$h(t) = D(\tilde{\beta} + t(\beta^* - \tilde{\beta}))$$

由 $D(\beta)$ 为 β 的凸函数, 易知 $h(t)$ 为 t 的凸函数。又因 $h(1) = D(\beta^*) > D(\tilde{\beta}) = h(0)$, 即知 $h'(1) > 0$. 但另一方面, 根据已证部分, 由 β^* 为 $D(\beta)$ 的最小值点, 知 β^* 为 (4.4) — (4.5) 的解。即有

$\partial D(\beta) / \partial \beta_j |_{\beta=\beta^*} = 0, j=1, \dots, d$. 由此将得

$$h'(1) = \sum_{j=1}^d (\partial D(\beta) / \partial \beta_j) |_{\beta=\beta^*} (\beta_j^* - \tilde{\beta}_j) = 0$$

与 $h'(1) > 0$ 矛盾。这证明了所说的结论, 即 (4.4) — (4.5) 的解必是 $D(\beta)$ 的最小值点。

综合起来, 证明了 $D(\beta)$ 的最小值点与方程组 (4.4) — (4.5) 的解等价。由于 $D(\beta)$ 有最小值点, 也就证明了方程组 (4.4) — (4.5)

必有解。

最后，由预备事实2*，知当 ρ 为严凸时， D 也为严凸，故只能有唯一的最小值点（存在性已证明如上），定理证毕。

不难举例说明：当 ρ 不为严凸时， $D(\beta)$ 的最小值点可以不唯一。例如，取 $\rho(x) = |x|$ ，而模型为 $Y_i = \beta + e_i$ ， $i = 1, \dots, n$ 。这相应于(4.1)中 $d = 1$ ， $x_1 = \dots = x_n = 1$ 的情况，有 $D(\beta) = \sum_{i=1}^n |Y_i - \beta|$ ，若 n 为偶数，而 $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ 为 Y_1, \dots, Y_n 按由小到大的排列，则如所周知，区间 $[Y_{(n/2)}, Y_{(n/2+1)}]$ 内任一个 β 值都使 $D(\beta)$ 达到最小。

若不假定 ρ 为凸函数，则使 $D(\beta)$ 达到最小的点 β^* 仍满足(4.4)–(4.5)。但在这个情况下，当这个方程组的解不唯一时，还不能肯定其解必为 $D(\beta)$ 的最小值点。

在常用的函数 ρ 中， $\psi(x)$ 在个别点不连续，特别是在0点不连续，例如， $\rho(x) = |x|$ 。对这种情况，方程组(4.4)在经过修改后仍适用。修改的内容是：当 $\delta_i = 0$ 时，要把 $\psi(\delta_i)x_{ij}$ 改为 $|cx_{ij}|$ ，其中 $c = \min(\rho'(0+), \rho'(0-))$ ，且(4.4)式中的“=”号要改为“ \geq ”。经过这一修改， $D(\beta)$ 的最小值点与方程组(4.4)–(4.5)的解仍能保持一致。读者可以就 $\sum_{i=1}^n |Y_i - \beta|$ 这一特例来验证一下这个结论。

因为M估计是通过方程组(4.4)–(4.5)确定的，故也可以不先给定函数 ρ ，而直接给定函数 ψ ，并把M估计定义为方程组(4.4)–(4.5)的解。根据前面所提出的对函数 ρ 的要求，函数 ψ 应满足以下的条件：

1. $\psi(x) \geq 0$ 当 $x > 0$ ， $\psi(x) \leq 0$ 当 $x < 0$ ， $\psi(x)$ 在 $|x| > 0$ 时连续。
2. $\lim_{|x| \rightarrow \infty} \psi(x)/x = 0$ 。
3. $\psi(0+)$ ， $\psi(0-)$ 存在

$\psi(x)$ 在 $x=0$ 处不必有定义, 但在方程组(4.4)中, 碰到 $\delta_i=0$ 时 $\psi(\delta_i)x_i$ 要改为 $\min\{-\psi(0-), \psi(0+)\}|x_i|$. 以下是几个这样的例子:

1. (Huber[76])

$$\psi(x) = \begin{cases} -K, & x < -K, \\ x, & |x| \leq K, \\ K, & x > K \end{cases} \quad (K > 0 \text{ 为常数})$$

2. (Hampel 见[77])

$$\psi(x) = \begin{cases} \text{Sign}(x)|x|, & |x| \leq a \\ \text{Sign}(x)a, & a \leq |x| \leq b \\ \text{Sign}(x)(c - |x|)/(c - b), & b < |x| \leq c \\ 0, & |x| > c \end{cases}$$

($0 < a < b < c < \infty$, a, b, c 为常数)

3. (Andrews[77]).

$$\psi(x) = \begin{cases} \sin(x/c), & |x| \leq c\pi \\ 0, & |x| > c\pi \end{cases} \quad (c > 0 \text{ 为常数})$$

前已提过, M估计的一个主要作用是对付异常值。在[77]中, Andrews通过分析Daniel和Wood的一个实例, 说明这一点。在该例中, 有三个自变量 X_1, X_2, X_3 和因变量 Y , 样本大小 $n=21$. 用最小二乘法配出线性回归后, 算出21个样本点的残差是

$$\begin{array}{cccccccc} 3.24 & -1.92 & 4.56 & 5.70 & -1.71 & -3.01 & -2.39 \\ -1.39 & -3.14 & 1.27 & 2.64 & 2.78 & -1.43 & -0.05 \\ 2.36 & 0.91 & -1.52 & -0.46 & -0.60 & 1.41 & -7.24 \end{array}$$

(4.8)

其中 -7.24 很显著。因此, 具有这个残差的第21号样本点, 有较大的可能为异常值。其余各样本点中, 看上去只有第4、3、1、

9号, 其残差较大, 把第1、3、4、和21号样本点弃置不用, 重用最小二乘法配出线性回归方程, 并针对所配出的方程, 就全部21个样本点算出残差值, 所得结果如下:

$$\begin{array}{ccccccc} 6.08 & 1.15 & 6.44 & 8.18 & -0.67 & -1.25 & -0.42 \\ 0.58 & -1.06 & 0.35 & 0.96 & 0.47 & -2.51 & -1.34 \\ 1.34 & 0.14 & -0.37 & 0.10 & 0.59 & 1.93 & -8.63 \end{array} \quad (4.9)$$

在这里, 第1、3、4和21号样本的残差比前更大, 更象是“异常值”。一般讲, 当有异常数据存在时, 最小二乘法起一种调和的作用, 即在牺牲无异常的数据的拟合程度的代价下, 使异常数据的“异常程度”降低。这两者都是有害的, 因为它掩盖了事情的真象。由此可见, 最小二乘法不是发现和处理异常数据的恰当工具。

另一方面, 若采用Andrews的函数 $\psi(x)$, 并取其中的 $c=1.5$, 作出 β 的M估计, 则由此算出21个样本点的残差是

$$\begin{array}{ccccccc} 6.11 & 1.04 & 6.31 & 8.24 & -1.24 & -0.71 & -0.33 \\ 0.67 & -0.97 & 0.14 & 0.79 & 0.24 & -2.71 & -1.44 \\ 1.33 & 0.11 & -0.42 & 0.08 & 0.63 & 1.87 & -8.91 \end{array} \quad (4.10)$$

在此, 不仅第21号, 而且第1、3、4号样本点, 其残差都很大, 因而异常性很突出。而在最初使用最小二乘法而得到的残差(4.8)中, 第1、3、4号的残差虽较大, 但并不很突出, 因而它们是否应判为异常值, 不易决定。除此以外, 从(4.10)看, 第9号样本点并无异常之处, 而在(4.8)中, 这样本点则是一个被怀疑的对象。然而, 在去掉第1、3、4、21号样本点而配出的回归方程之下, 从(4.9)看出9号样本点并不突出, 这支持了9号样本点并非异常点的看法, 与用M估计得出的结论一致。这一例子启示我们: 一个样本点是否为异常值, 不能单凭它在最小二乘法之下

计算出的残差去判定。除了经验以及对这样本是如何获得的了解外，用M估计方法也能有一些帮助。

拿M估计和最小二乘法比较，M估计在缩小残差上并无作用。这是理所当然的，因为最小二乘法已做到了使残差平方和达到最小，别的任何估计方法，其残差平方和只有增大而无缩小，因此总的说残差是增大。然而，用M估计代替最小二乘估计的目的，并不在于缩小残差，而在于得到回归系数的较好估计。就本例而言，情况如下：

1. 全部21个样本点，用最小二乘法：

$$y = -89.9 + 0.72x_1 + 1.30x_2 - 0.15x_3$$

2. 去掉第1、3、4和21号点，用最小二乘法：

$$y = -37.6 + 0.80x_1 + 0.58x_2 - 0.07x_3$$

3. 全部21个样本点，用前面提到的 ψ 作M估计：

$$y = -37.2 + 0.82x_1 + 0.52x_2 - 0.07x_3$$

第二、三两个估计基本一样。可是，当用最小二乘法时，为得到第二个估计，需要判定出第1、3、4和21等号样本点为异常数据，这需要使用者有相当的经验。而用M方法，则不需经过这个排除异常值的程序，就达到了同样的效果。

(三) M估计的计算

为决定回归系数的M估计，需要解方程

$$\sum_{i=1}^n \psi(Y_i - \sum_{j=1}^d x_{ij}\beta_j) x_{ij} = 0, \quad j=1, \dots, d$$

这方程一般只能用迭代法求解。一种简易的迭代步骤如下：对 β 的当前值 $\beta^{(0)}$ 算出

$$d_i^{(0)} = Y_i - x_i' \beta^{(0)}, \quad i=1, \dots, n$$

然后由线性方程组

$$\sum_{k=1}^d \left(\sum_{i=1}^n \psi'(\delta_i^{(0)}) x_{ik} x_{ik} \right) b_k = \sum_{i=1}^n \psi(\delta_i^{(0)}) x_{ij}, \quad j=1, \dots, d$$

决定 $\beta^{(0)}$ 的改变量 $b=(b_1, \dots, b_d)'$, 而得到 β 的新值 $\beta^{(0)}+b$.

在这一迭代计算中, 初始值的选择很重要。一个可以考虑的作法, 是使用 β 的最小二乘估计 $\hat{\beta}$ 作为 β 的初始值。1974年, Andrews在[77]中提出了另一种取法。他认为, 当模型(4.1)中误差项 e_i 的分布与正态分布有较大偏离时, 最小二乘估计 $\hat{\beta}$ 可能与 $D(\beta)$ 的最小值点相去甚远, 而迭代结果有可能收敛到一个局部极值点, 后者不一定是 $D(\beta)$ 的最小值点。在估计位置参数这个简单情况(相当于(4.1)中 x 的维数 $d=1$, $x_1=\dots=x_n=1$), 样本中位数是一个较稳健的估计, 以之作为初始值想必较好。Andrews把这个想法推广到多元线性回归的一般情况。

Andrews方法的实质, 是将线性回归中逐次消去的方法, 即“向前法”, 作为一种稳健性的修正。现介绍他的方法。作矩阵

$$M = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} & Y_1 \\ x_{21} & x_{22} & \dots & x_{2d} & Y_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nd} & Y_n \end{bmatrix}$$

以 M_i 记矩阵 M 的第 i 列, $i=1, \dots, d+1$. 对任何 $i < j$, 以 R_{ij} 记运算

$$M_i - b_{ij}M_j \Rightarrow M_i$$

即把 M 的第 j 列换为 $M_i - b_{ij}M_j$, 其中

$$b_{ij} = M_i' M_j / \|M_j\|^2 \quad (4.11)$$

由 M 出发, 施以运算 R_{12} , 得 $R_{12}M$ 。对 $R_{12}M$ 施以运算 R_{13} , 得 $R_{13}R_{12}M$ 。这样依次作完下式所示的全部运算:

$$R_{d,d+1}R_{d-1,d+1}R_{d-2,d+1}\dots R_{1,d+1}R_{1d}\dots R_{12}M \quad (4.12)$$

将所得结果记为 M_p 。应当注意的是: 当用(4.11)式算 b_{ij} 时, M_i ,

M_i 都是当前值。

M_p 的最后一列，就是用最小二乘法时， Y_1, \dots, Y_n 各数据的残差。如果除残差以外，还需要回归系数的估计值，则需要执行运算(4.12)时，同时也将每步运算施加于 $d+1$ 阶单位阵 I_{p+1} ，不过 b_{ij} 是从矩阵 M 算出，而非从 I_{p+1} 算出。当对 I_{p+1} 施行了这一运算后，所得结果最末一列从第二元开始，依次是 $\beta_1, \beta_2, \dots, \beta_d$ 的最小二乘估计。

以上不过是把最小二乘估计的算法重述了一遍。但这个叙述使我们看出：最小二乘法之缺乏稳健性，就在于 b_{ij} 的计算公式(4.11)，它受异常值的影响太大。Andrews保留上述程序，而修改 b_{ij} 的定义，使之有更好的稳健性。作法如下：设在某一阶段要作变换 R_{ij} ，而当时矩阵的第 i, j 列分别为 $(x_1, \dots, x_n)'$ 和 $(y_1, \dots, y_n)'$ 。指定相当小的数 $p_1 > 0, p_2 > 0$ ，把 x_1, \dots, x_n 中最大的 $p_1 n$ 个和最小的 $p_1 n$ 个割弃不用，把与去掉的诸 x_i 相应的 y_i 也去掉。又算出 x_1, \dots, x_n 的样本中位数 μ ，把位于 μ 的左右邻近的各 $p_2 n$ 个割弃不用，相应的诸 y_i 也去掉。对剩下来的 x_i ，按在 μ 的左、右，分为 L (下部) 和 H (上部)。这两部分的样本中位数分别记为 μ_L 和 μ_H 。与 L 中的 x_i 相应的 y_i 的全体记为 L' ，其样本中位数记为 $\mu_{L'}$ 。类似地决定 $\mu_{H'}$ 。最后令

$$b_{ij} = (\mu_{H'} - \mu_{L'}) / (\mu_H - \mu_L)$$

以之代替(4.11)的 b_{ij} 去施行上述程序。

(四) M估计的大样本性质

这是一个在文献中受到不少注意的题目。这里只不加证明地引述 Huber 1973 年在 [78] 中得出的下述结果：

定理4.2 假定以下条件成立

1° $\rho(x)$ 的二阶导数 $\rho''(x)$ 在 $(-\infty, \infty)$ 处处存在且有界，

又 $\lim_{|x| \rightarrow \infty} \rho(x) = M < \infty$.

2° e_1, e_2, \dots 独立同分布, $E\psi(e_1) = 0, \psi = \rho'$.

3° 当 n 充分大时, 方阵 $S_n = \sum_{i=1}^n x_i x_i'$ 为满秩方阵, 且

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} x_i' S_n^{-1} x_i = 0 \quad (4.13)$$

则方程组 (4.4) — (4.5) 有一解 $\hat{\beta}^{(n)}$, 满足

$$B_n^{-1}(\hat{\beta}^{(n)} - \beta) \xrightarrow{\mathcal{M}} N_d(0, I_d) \quad (4.14)$$

其中

$$B_n = (\sqrt{E\psi^2(e_1)} / |E\psi'(e_1)|) S_n^{-1/2} \quad (4.15)$$

本定理条件中苛刻之处在于要求 ρ'' 处处存在且有界. 在应用上碰到的一些情况, 例如在第(二)段中提到的几种情况, 甚至 $\rho'(x)$ 也可以在个别点不存在, 对估计位置参数这个简单情, 可参看 Huber 的工作 [76].

不难看到, 在 (4.13) 成立时必有

$$\lim_{n \rightarrow \infty} S_n^{-1} = 0$$

由此及 (4.14) 和 (4.15) 可知, 在本定理条件下, $\hat{\beta}^{(n)}$ 是 β 的弱相合估计。

对 $\rho(x) = x^2$ 的情况, $\hat{\beta}^{(n)}$ 就是最小二乘估计. 因而本定理 (在相应的条件下) 给出最小二乘估计的渐近正态性的结果。

(五) 刻度参数的估计

在线性模型 (4.1) 中, 除回归系数 β 外, 误差 e_i 的方差也是一个很重要的参数. 方差, 或其平方根即标准差, 是用来作为误差大小的一种综合性指标的. 但有一些原因, 使得用方差或标准差作为这种指标, 不一定总是恰当或可行. 例如, e_i 的方差可以不存在, 还有, 方差和标准差的估计, 稳健性也比较差. 这主要是

它们受分布的尾部的影响太大，就是说，对分布的远尾部作一点很细微的改变，就可以大大改变方差之值。由于这一点，人们也常说，方差(及标准差)这个指标的稳健性差。

因此就提出问题：用一种更具有稳健性的、能反映误差大小的指标，来代替标准差，并用富于稳健性的方法去估计它。为寻找这样一种指标，我们先提出：若以 $G(e)$ 表示某种反映误差 e 的散布程度的指标，则对 $G(e)$ 作下面一些要求，看来是合理的：

1. $G(e+c) = G(e)$ ，对任何常数 c

2. $G(ce) = |c|G(e)$ ，对任何常数 c 。

3. $G(e) \geq 0$ 。等号当且仅当 e 退化(以概率1等于一常数)时成立。

任何一个满足以上三个条件，且由随机变量 e 的分布决定的指标 $G(e)$ ，都称为是 e 的**刻度参数**。准此，一个随机变量 e 可以有許多刻度参数，这一点不难理解：刻度参数是描述 e 的散布程度的。对“散布程度”，可以从种种不同的角度去理解和描述。

标准差 $\sigma = \sqrt{\text{Var}(e)}$ 显然满足以上三条件，是最常用的刻度参数之一。另外几个刻度参数的例子如下： $\sigma_1 = E|e - Ee|$ ， $\sigma_2 = \text{med}(|e - \text{med}(e)|)$ ($\text{med}(e)$ 表 e 的中位数)， $\sigma_3 = \xi_{0.75}(e) - \xi_{0.25}(e)$ ，其中 $\xi_p(e)$ 表 e 的100 p %分位数。为要 σ_1 存在，只需 $E|e|$ 存在，而 σ_2 ， σ_3 对任何随机变量都存在。

估计刻度参数的一般方法，可以举 σ_2 的估计为例。先估计出回归系数 β 得 β^* ，算出残差 $\delta_i = Y_i - x_i'\beta^*$ ， $i=1, \dots, n$ 。用 $\text{med}(\delta_1, \dots, \delta_n)$ 估计 $\text{med}(e)$ ($\text{med}(\delta_1, \dots, \delta_n)$ 表 $\delta_1, \dots, \delta_n$ 的样本中位数)。算出 $\eta_i = \delta_i - \text{med}(\delta_1, \dots, \delta_n)$ ， $i=1, \dots, n$ ，最后用 $\sigma_2 = \text{med}(\eta_1, \dots, \eta_n)$ 去估计 σ_2 。若事先假定了误差 e_i 的分布关于原点，对称，则 $\text{med}(e) = 0$ ， $\sigma_2 = \text{med}(|e|)$ ，而我们就用 $\text{med}(|\delta_1|, \dots, |\delta_n|)$ 去估计 σ_2 。

也可以把刻度参数 ν 与回归系数 β 放在一起同时去估计。方法如下：以 $\beta^{(0)}$, $\nu^{(0)}$ 记 β 、 ν 的当前值。算出 $\delta_i^{(0)} = Y_i - x_i' \beta^{(0)}$, $i=1, \dots, n$ 。由方程组

$$\sum_{k=1}^d \sum_{i=1}^n \psi'(\delta_i^{(0)} / \nu^{(0)}) x_{ik} x_{ij} b_k = \sum_{i=1}^n \psi(\delta_i^{(0)} / \nu^{(0)}) x_{ij}$$

$$j=1, \dots, d \quad (4.16)$$

决定 $\beta^{(0)}$ 的改变量 $b = (b_1, \dots, b_d)'$ ，而得 β 的下一步值 $\beta^{(1)}$ 。算出 $\delta_i^{(1)} = Y_i - x_i' \beta^{(1)}$, $i=1, \dots, n$ ，并用

$$\nu^{(1)} = \text{med} \{ |\delta_i^{(1)} - \text{med}(\delta_1^{(1)}, \dots, \delta_n^{(1)})|, i=1, \dots, n \}$$

$$(4.17)$$

作为 $\nu^{(0)}$ 的下一步值。这里是把上文的 σ_2 取为刻度参数。若取另外的刻度参数，则只须用相应的公式代替公式(4.17)式即可。

这个叠代过程要求 ψ' 存在。当然，这对原来的(不考虑刻度参数的)叠代程序(见(三))也是一样的。

参 考 文 献

- [1] 华东师范大学数学系，回归分析及其试验设计，上海教育出版社，1978.
- [2] 朱伟勇等，最优设计理论及应用，辽宁人民出版社，1981.
- [3] Stigler, S.M. Gauss and the invention of least squares, Ann. Statist. 9(1981)465-74 (张尚志译文载《数学译林》，1982).
- [4] 成平，陈希孺，陈桂景，吴传义，参数估计，上海科技出版社，1985.
- [5] 陈希孺，陈桂景，吴启光，赵林城，线性模型参数的估计理论，科学出版社，1985.

- [6] Lehmann, E. L., Testing Statistical Hypothesis, John Wiley, 1959.
- [7] Miller, R. G. Jr, Simultaneous Statistical Inference, Mcgraw-Hill, 2nd ed, 1976.
- [8] 王松桂, 线性模型的理论及其应用, 安徽教育出版社, 1987.
- [9] Dunn, O.J. Confidence interval for the means of dependent, normally distributed variables, J. Amer. Statist. Assoc., 54, (1959)613-21.
- [10] Lieberman, G.J., Prediction regions for several predictions from a single regression line, Technometrics, 3(1961)21-7.
- [11] Kendall, M. et al, The Advanced Theory of Statistics, Vol.2, 4th ed, Charles Griffin & Company Limit, 1979.
- [12] Gleser, L.J., Estimation in a multivariate "error-in-variables" regression model, Large sample results, Ann. Statist., 9(1981)24-44.
- [13] Anderson, T.W., Estimation of linear relationships, approximate distributions and connection with simultaneous equations in econometrics, J. Roy. Statist. Soc. Ser. B., 38(1976)1-36.
- [14] 张尧庭, 方开泰, 多元统计分析引论, 科学出版社, 1983.
- [15] Rao, C. R., Linear Statistical Inferences and its Applications, John Wiley, 1973.
- [16] Loeve, M., Probability Theory, D. Van Nostrand, 1963.
- [17] Belsley, D.A. et al, Regression Diagnostics, John

Wiley, 1980.

- [18] 王松桂, 线性回归诊断, 数理统计与管理, 6(1985)38—49
和1(1986)40—47.
- [19] Seber, G.A.F., Linear Regression Analysis, John
Wiley, 1977.
- [20] Cook, D. R. and Weisberg, S., Residuals and
Inference in regression, New York, 1982.
- [21] Durbin, J. and Wantson, G.S., Testing for serial
correlation in least squares regression, (I), Biometri-
ka, 37(1950)409—35; (II), 38(1951)159—78; (III), 58
(1971)1—19.
- [22] Montgomery, D.C. and Peck, E.A., Introduction to
Linear Regression Analysis, John Wiley, 1982.
- [23] Box, G.E.P. and Cox, D.R., An analysis of trans-
formations (with discussion) J. Roy. Statist. Soc.
Ser. B., 26(1964)211—46.
- [24] Cook, R.D. and Weisberg, S., Characterizations of
an empirical influence function for detecting influen-
tial cases in regression, Technometrics, 22(1980)495—
508.
- [25] Johnson, W. and Geisser, S., A predictive view of
the detection and characterization of influential ob-
servations in regression analysis, J. Amer. Statist.
Assoc., 76(1983)137—44.
- [26] Andrews, D.F. and Pregibon, D., Finding outliers
that matter, J. Roy. Statist. Soc. Ser. B., 40(1978)
85—93.

- [27] Beckman, R.J. and Cook, R.D., Outlier...s, Technometrics, 25(1983)119-63.
- [28] Mallows, C.L., Choosing variables in a linear regression; a graphical aid. Presented at the Central Regional Meeting of the IMA, Manhan, Kansas, 1964.
- [29] Mallows, C. L., Some Comments on C_p , Technometrics, 15(1973)661-75.
- [30] McDonald, G. C. and Schwing, R.C., Instabilities of regression estimates relating air pollution to mortality, Technometrics, 15(1973)463-81.
- [31] Allen, D.M., Mean square error of prediction as a criterion for selecting variables, Technometrics, 13 (1971)469-75.
- [32] Allen, D.M., The relationship between variable selection and data augmentation and a method for prediction, Technometrics, 16(1974)125-7.
- [33] Aitkin, M.A., Simultaneous inference and the choice of variable subsets, Technometrics, 16(1974)221-7.
- [34] Akaike, H. A new look at the statistical identification model, IEEE Trans. Auto. Control, 19(1974) 716-23.
- [35] Schwarz, G., Estimating the dimension of a model, Ann. Statist. 6(1978)461-4.
- [36] Smith, A.F. M., et al, Bayes factors and choice criteria for linear models, J. Roy. Statist. Soc.

Ser. B, 42(1980)213-20.

- [37] Furnival, G.M., All possible regression with less computation, *Technometrics*, 13(1971)403-8.
- [38] Furnival, G.M. and Wilson, R.W. M., Regression by leaps and bounds, *Technometrics*, 16(1974)499-511.
- [39] Garside, M. J., The best subset in multiple regression analysis, *Appli. Statist.*, 14(1965)196-200.
- [40] Schatzoff, M.R. et al, Efficient calculation of all possible regression, *Technometrics*, 10 (1968) 769-79.
- [41] 方开泰, 王东谏、吴国富, 一类带约束的回归——配方回归, *计算数学*(1982), 58-69.
- [42] 中国科学院数学所, 回归分析方法, 科学出版社, 1975.
- [43] Webster, J.T. Gunst, R.T. and Manson, R.L., Latent root regression, *Technometrics*, 16(1974)513-22.
- [44] Hoerl, A.E. and Kennard, R.W., Ridge regression, biased estimation for non-orthogonal problems, *Technometrics*, 12(1970)55-88.
- [45] Hoerl, A.E. and Kennard, R.W., Ridge regression, application for non-orthogonal problems. *Technometrics* 12(1970)69-72.
- [46] McDonald, G.C. and Galarneau, D.I., A Monte-Carlo evaluation of some ridge-type estimators, *J. Amer. Statist. Assoc.*, 70(1975)407-16.
- [47] Vinod, H.D. and Ullah, A., *Recent Advances in*

Regression Methods, Marcel, Dekker, 1981.

- [48] Lawless, J.F. and Wang, P., A simulation of ridge and other regression estimators, Commun. Statist. A5(1976)307-23.
- [49] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F., Ridge regression, some simulations, Commun. Statist., 4(1975)105-23.
- [50] Hemmerle, W.J. and Brantle, T.F., Explicit and constrained generalized ridge regression, Technometrics, 20(1978)109-20.
- [51] Massy, W.F., principle components regression in exploratory statistical research, J. Amer. Statist. Assoc., 60(1965)234-66.
- [52] 王松桂, 主成分的最优性及广义主成分估计类, 应用概率统计, 1(1985)23-30.
- [53] 王松桂, 林春土, 主成分的最优性, 科学通报, 8(1984)449-51.
- [54] Gaust, R.F. and Mason, R.L. Biased estimation in regression, An evaluation using mean squared error. J. Amer. Statist. Assoc., 72(1977)616-28.
- [55] Webster, J.T. et al, Latent root regression analysis, Technometrics, 16(1974)513-22.
- [56] Huber, P.J., Projection pursuit, Ann. Statist., 13(1986)435-74.
- [57] Stone, C., Consistent nonparametric regression, Ann. Statist., 10(1977)595-645.
- [58] Yakowitz, S. et al, Contribution to the theory of

- nonparametric regression, with application to system identification, *Ann. Statist.*, 7(1979)139-49.
- [59] Devroye, L., On the almost everywhere convergence of nonparametric regression function estimates, *Ann. Statist.*, 14(1981)1310-9.
- [60] 赵林城, 白志东, 非参数回归函数最近邻估计的强相合性, *中国科学, A*, 5(1984)387-93.
- [61] 陈希孺, 非参数回归估计的几乎处处收敛性, *数学年刊, B*, 1(1985)103-8.
- [62] Grizzle, J.E. et al, Analysis of categorical data by linear models, *Biometrics*, 25(1969)489-504.
- [63] Dempster, A.P., An overview of multivariate data analysis, *J. Multi. Anal.*, 1(1971)316-46.
- [64] McCullagh, P. and Nelder, J.A., *Generalized Linear Models*, Chapman and Hall, 1983.
- [65] Nelder, J.A. and Wedderburn, H. W. M., *Generalized Linear Models*, *J. Roy. Statist. Soc. A.* (1972) 370-84.
- [66] 陈希孺, *数理统计引论*, 科学出版社, 1981.
- [67] Cox, D.R., Regression models and life tables (with discussion) *J. Roy. Statist. Soc. B*, 34(1972)187-220.
- [68] Fair, R.C., A note on the computation of the Tobit estimator, *Econometrica*, 45(1977)1723-8.
- [69] Dempster, A.P., et al, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc.*, B.39(1977)1-22.
- [70] Amemiya, T., Regression analysis when the depen.

dent variable is truncated normal, *Econometrica*, 41 (1973)997-1016.

- [71] Tobin, J. Estimation of relationships for limited dependent variables, *Econometrica*, 26(1958)24-36.
- [72] Miller, R.G., Least squares regression With censored data, *Biometrika*, 63(1976)449-64.
- [73] Buckley, J. et al, Linear regression with censored data, *Biometrika*, 66(1979)429-36.
- [74] Koul, H. et al, Regression analysis with randomly right-censored data, *Ann. Statist.* 9(1981)1276-88.
- [75] Huber, P., *Robust Statistics*, John wiley, 1981.
- [76] Huber, P., Robust estimation of a location parameter, *Ann. Math. Statist.*, 35(1964)73-102.
- [77] Andrews, D., A robust method for multiple linear regression, *Technometrics*, 16(1974)523-31.
- [78] Huber, P., Robust statistics: A review, *Ann. Math. Statist.*, 43(1973)1041-6.

名 词 索 引

A

AIC准则 193

B

Bayes风险 299

Bayes判别准则 292

Bonferrons法 51

Box-Cox变换 129

不相关残差 104

C

残差 19, 95

残差平方和 19, 272

残差图 106

C_p -统计量 173

D

大样本方法 22

Deviance分析 322

典则回归系数 227,

典则联系函数 303

定点截尾 328

定数截尾 336

Durbin-Watson检验 113

E

EM程序 332

F

F-检验 26

F-统计量 26

非参数回归 280

方差扩大因子 223,

方差稳定化变换 122

复共线性 220,

G

Gauss-Markov定理 15

Gauss-Markov假定 14

Gauss消去法 202

广义岭估计 242

广义线性模型 300

高杠杆点 99

H

核估计 287

Hoerl-Kennard-Baldwin公式 233,

Hoerl-Kennard公式 231

回归系数 12

回归树 207

回归平方和 33

回归诊断 91

回归显著性检验 25, 32

J

结构关系模型 65, 77

加权最小二乘 18

经验影响函数 135

校准问题 60

截尾回归 329

近邻型估计 288

均方误差 218

均方误差矩阵 156

K

k-近邻估计 283

L

Lawless-Wang公式 233

联系函数 300

岭估计 226

岭迹 230, 235,

岭迹分析 235,

岭参数 226

M

M-估计 344

帽子矩阵 95

N

拟似然函数法 323

内插 57

P

偏差残差 106

平均残差平方和 164

平均预测均方误差 166

普通残差 95

Q

强影响点 92, 136, 140

权函数估计 285,

区间预测 57

S

扫描运算 196

Stein估计 256,

双h公式 233,

双h类岭估计 233,

T

特征根估计 268, 274

条件数 221,

同时区间估计 45

同时预测区间 57

t-直接法 215

W

外推 57

稳健性 341

X

线性回归模型 5, 91

相合性 23, 84

相合估计 84

相关比 73

相合权函数 297

向前法 212

向后法 213

小样本方法 22

学生化残差 100

Y

一致最小方差无偏估计 21

样本复相关系数 33

有偏估计 217,

异常点 93, 137, 146

预测 54

预测平方和准则 189

预测残差 103

Z

正态概率纸 126

正态化变换 129

指数型分布 302

逐步法 214

主成分 259

主成分估计 259, 263

最小二乘估计 10, 14

最佳线性逼近的剩余方差 70

置信区域 42

总平方和 32