《回归分析》

尤进红

上海财经大学统计与管理学院

Email:johnyou07@163.com

教学目的：

《回归分析》是统计学的一个重要分支，在自然科学领域及社会科学领域都有着广泛的应用。本课程的授课对象为统计与管理学院本科实验班学生，以培养学生统计建模和应用能力为主线，重点讲授回归分析理论和方法。通过该课程的学习，应使学生掌握回归分析的基本概念，熟悉其主要理论和方法，培养学生实际数据分析能力，能运用统计方法分析研究现实生活和实际生产中的一些具体问题，有较强动手能力和一定的推理分析能力。教学方法上考虑将多媒体技术、统计分析软件、实践教学等有机结合起来。达到提高课堂教学效率和教学质量的目的，通过案例教学和启发式学习方式，使学生进一步明确如何正确使用所学的统计方法。

课前预习：

由于本课程是关于回归分析的理论和实际应用，要求学生做到课前预习，老师在课堂上将就这些理论和方法进行阐述和解释，如果学生事先阅读有关章节，将有助于理解课程内容。

考核形式：

学生的最后的总分计算方法如下：

平时成绩（考勤 + 作业）	30%
期中考试	20%
期末考试	50%
课后作业	6 次左右课后作业。
	从学期第二次授课开始。
	每次作业要求在一周内完成，独立完成，不接受迟交作业。
期末考核	具体细节待定。

学术诚实：

涉及学生的学术不诚实问题主要包括作业及考试的作弊；抄袭；伪造或不当使用在校学习成绩；未经老师允许获取、利用考试材料。对于学术不诚实的最低惩罚是考试给予 0 分。其它的惩罚包括报告学校相关部门并按照有关规定进行处理。

回归分析教学要点
教学大纲

第一章　模型概论

回归分析是用来研究变量之间相关关系的一种统计方法，它不同于化学、物理、数学中研究的变量之间的关系。回归分析研究的变量之间的关系的特征是：部分确定的关系。

例如：

- 体重 Y 与身高 X 的关系，一般来说当 X 大时，Y 也倾向于大，但 X 不能严格地决定 Y。
- 城市生活用电量 Y 与气温 X 有很大的关系，夏天气温很高或冬天气温很低时用电量就高。相反，在春、秋季气温适量，用电量就相对少，但我们不能由气温 X 准确地决定用电量 Y。
- 数学分析 X_1 和高等代数 X_2 的成绩与回归分析的成绩 Y 有很大的关系。通常数学分析和高等代数学得好，回归分析也学得好，但不能由数学分析和高等代数的成绩来完全决定回归分析的成绩。

在以上的例子中，Y 通常称为因变量或者响应变量，X 称为自变量、预测变量、解释变量或协变量。

Y 的值一部分是由 X 能够决定的部分，它是 X 的函数 f(x)，例：线性关系 $f(x)=\beta_0+\beta_1 x$, 二次多项式关系 $f(x)=\beta_0+\beta_1 x+\beta_2 x^2$, 多个变量 $f(x)=\beta_0+\beta_1 x_1+\beta_2 x_2$, 其中 β_0,β_1,β_2 是未知的，称为未知参数。另一部分是由其它众多未加考虑的因素（包括随机因素）所产生的影响, 称为随机误差，通常用 ε 或 e 来表示，要求 $E\varepsilon=0$ 或 Ee=0。于是，

$$Y=f(X)+e=\begin{cases}\beta_0+\beta_1 X+e\\ \beta_0+\beta_1 X+\beta_2 X^2+e\\ \beta_0+\beta_1 X_1+\beta_2 X_2+e\end{cases}\tag{1.1}$$

(1.1) 称为回归模型，如果因变量与未知参数的关系是线性的，称为线性回归模型; 如果因变量为未知参数是非线性的，比如：$y=exp(\beta_0+\beta_1 X+e)$，称为非线性回归模型；如果 f(x) 是 X 的一个未知函数，则称为非参数回归模型。

本课程我们只涉及线性回归模型，β_0,β_1,β_2 称为未知的回归函数，需要通过观测的数据来估计。

假设自变量 X 分别取值为 $x_1,x_2,...,x_n$ 时因变量 Y 对应的观测值分别为 $y_1,y_2,...,y_n$, 于是我们有 n 组观测值 (x_i,y_i)，如果 Y 与 X 有回归关系 $Y=\beta_0+\beta_1 X+e$, 则这些

(x_i, y_i) 应该满足

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, 2, ..., n$$

e_i 为对应的误差，基于 $(x_i, y_i), i = 1, 2, ..., n$ 和适当的统计方法，我们可以得到 β_0 和 β_1 的估计值：$\widehat{\beta_0}$,$\widehat{\beta_1}$，略去误差项 e_i 得到 $Y = \widehat{\beta_0} + \widehat{\beta_1} X$ 称之为经验回归直线。

例 1.1.1：假设 X 表示身高 (cm)，Y 表示体重 (kg)。假设 Y 与 X 之间具有线性回归关系 $Y = \beta_0 + \beta_1 X + e$,e 表示除了身高 X 之外所有影响体重的其它因素，如：遗传因素，饮食因素和体育锻炼。

测量了 n 个人得到 $(x_i, y_i), i = 1, 2, ..., n$，得到经验回归直线 $Y = -40 + 0.6X$，这个经验回归方程在一定程度上描述了体重与身高的相关关系，给定 X 的一个具体值 x_0，可以算出对应的 Y 值 $y_0 = -40 + 0.6x_0$。

在实际问题中，影响因变量的主要因素往往很多，这涉及多个自变量的回归问题，假设自变量 Y 和 p-1 个自变量 $X_1, ...X_{p-1}$ 之间有如下关系：

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + e$$

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

对 Y,$X_1, ..., X_{P-1}$ 进行了 n 次观测得到 n 组观测值

$$x_{i1}, ..., x_{i,p-1}, y_i, i = 1, 2, ..., n$$

它们满足关系式

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + e_i$$

e_i 为对应的误差，用矩阵记号：

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

则

$$y = X\beta + e$$

y 为 $n \times 1$ 的观测向量，X 为 $n \times p$ 已知矩阵，通常称为设计矩阵，β 为未知的参数向量，e 为 $n \times 1$ 的随机误差向量，其均值为零向量。

关于随机误差 e_i, 除了均值为零外，常用的假设是：

(a) 等方差，即 $Var(e_i) = \sigma^2, i = 1, 2, ..., n$；

(b) 误差是彼此不相关的，即 $Cov(e_i, e_j) = 0, i \neq j$。

通常称以上两条为 Gauss-Markov 假设，即：

$$E(e) = 0, Cov(e) = \sigma^2 I$$

线性回归模型比较简洁，所以有一些模型显然是非线性的，但经过适当变化，可以化为线性模型。

例 1.1.3：在经济学中，著名的 Cobb-Douglas 生产函数为

$$\varphi_t = aL_t^b K_t^c$$

其中 φ_t,L_t,K_t 分别为 t 年的产值，劳动投入量和资金投入量，a,b 和 c 分别为参数，两边取对数，得到：

$$ln(\varphi_t) = ln(a) + bln(L_t) + cln(K_t)$$

若令：$y_t = ln(\varphi_t)$,$x_{t1} = ln(L_t)$,$x_{t2} = ln(K_t)$,$\beta_0 = ln(a)$,$\beta_1 = b$,$\beta_2 = c$，加上误差项，便得到线性关系：

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t$$

把原来的非线性模型化成了线性模型、

例 1.1.4：多项式模型也可以化成线性模型，当代人称为回归分析和回归模型，“回归” Regression” 是应该著名生物学家兼统计学家 Galton(高尔顿) 在研究人类遗传问题时提出的。

他研究了 1078 对父亲及其一子的身高数据，X 表示父亲身高，Y 表示儿子身高，单位：inch(2.54cm)，发现 (x_i, y_i) 大致呈线性关系，x_i 越大，y_i 也越大。Galton 对数据进行了进一步的分析发现了回归效应：

1078 个 x_i 值的算术平均值 $\overline{x} = 68inch$

1078 个 y_i 值的算术平均值 $\overline{y} = 69inch$

子代身高比父代身高增加了一英寸

人们猜测若父亲身高为 x, 他儿子的平均身高大致应为 $y = x + 1$，但 Galton 的结论与此大相径庭，他发现：当父亲身高为 72 英寸时（注意平均身高 $\overline{x} = 68inch$），他们的儿子平均身高仅为 71 英寸，达不到预期的 72+1=73 英寸，反过来，当父亲身高为 64 英寸时，他们的儿子平均身高为 67 英寸，比预期的 64+1=65 英寸高出了 2 英寸。

反映了一个一般规律，即身高超过平均值 $\overline{x} = 68inch$ 的父亲，他们的儿子的平均身高低于父亲的平均身高，反之，身高低于平均身高 $\overline{x} = 68inch$ 的父亲，他们儿子的平均身高将高于父亲的平均身高。

Galton 对这个一般结论的解释为：大自然具有一种约束力，使人类身高的分布在一定时期内相对稳定而不产生两极分化，这就是所谓的回归效应。通过这个例子，Galton 引进了“回归”一词，用他的根据，可以计算出儿子身高 Y 与父亲身高 X 的经验关系：

$$Y = 35 + 0.5X$$

它代表一条直线，人们就把这条直线称为回归直线，说明以上现象不是绝对，有随机性。

第二章　最小二乘估计方法及其理论

设含有 p-1 个自变量的理论线性回归模型的一般形式为

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + e$$

如果对因变量 Y 和自变量 $X_1, X_2, ..., X_{p-1}$ 进行了 n 次观察，得到了 n 次观测数据 $(y_i, x_{i1}, ..., x_{i,p-1}, i = 1, 2, ..., n)$，满足：

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + e_i, i = 1, 2, ..., n$$

记：

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

且假设 $rank(X) = p$,$e_i, (i = 1, 2, ..., n)$ 互不相关，均值皆为零，且有公共方差 σ^2，则得到线性回归模型：

$$y = X\beta + e, E(e) = 0, Cov(e) = \sigma^2 I$$

称 β_0 为常数项，$\beta_I = (\beta_1, ..., \beta_{p-1})^T$ 为回归系数，$\beta = (\beta_0, \beta_I)^T$，y，x 是观测到的值，e 是不可观测的，基于 y,X 来估计 β。

基本思想为：β 的真值应该使误差向量 $e = y - X\beta$ 达到最小，也就是它的长度平方：

$$Q(\beta) = ||e||^2 = ||y - X\beta||^2 = (y - X\beta)^T (y - X\beta)$$

达到最小，这就是最小二乘法，即通过求 $Q(\beta)$ 的最小值来求 β 的估计，注意到

$$Q(\beta) = y^T y - 2y^T X\beta + \beta^T X^T X\beta$$

用矩阵微分公式：

$$\frac{\partial y^T X\beta}{\partial \beta} = X^T y, \frac{\partial \beta^T X^T X\beta}{\partial \beta} = 2X^T X\beta$$

于是：

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X^T y + 2X^T X\beta$$

令其等于 0，得到：

$$X^TX\beta = X^Ty, \tag{2.1}$$

称之为正则方程，由于 X^TX 是 $p \times p$ 方阵，根据假定 X^TX 的秩为 p, 因此正则方程 (2.1) 的解为 $\widehat{\beta} = (X^TX)^{-1}X^Ty$

根据函数极值论，我们知道 $\widehat{\beta}$ 只是 Q 的驻点，但我们还需证明它确实使 Q 达到最小，对任意一个 β

$$\begin{aligned} Q(\beta) &= ||y - X\beta||^2 \\ &= ||y - X\widehat{\beta} + X(\widehat{\beta} - \beta||^2 \\ &= ||y - X\widehat{\beta}||^2 + (\widehat{\beta} - \beta)^T X X^T(\widehat{\beta} - \beta) + 2(\widehat{\beta} - \beta)X^T(y - X\widehat{\beta}) \end{aligned}$$

由于 $X^Ty = X^TX\widehat{\beta}$, 于是上式第三项为 0，而第二项总之是非负的，于是，

$$Q(\beta) \geq ||y - X\widehat{\beta}||^2 = Q(\widehat{\beta}), \tag{2.2}$$

此式表明，$\widehat{\beta}$ 确实使 $Q(\beta)$ 达到最小。

现在证明，使 $Q(\beta)$ 达到最小的必为 $\widehat{\beta}$, 事实上，(2.2) 等式成立，当且仅当 $(\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta) = 0$ 等价地 $X(\widehat{\beta} - \beta) = 0$, 则 $X^TX\beta = X^TX\widehat{\beta}$, 由于 X^TX 是一个正交矩阵，$\beta = \widehat{\beta}$, 因此，使 $Q(\beta)$ 达到最小值的点必为正，则方程的解 $\widehat{\beta} = (X^TX)^{-1}X^Ty$。

定理 2.1

(1) $\widehat{\beta}$ 是 β 的一个无偏估计，即 $E(\widehat{\beta}) = \beta$

(2) 方差最小性（Gauss-Markov 定理）。对任意的 $p \times 1$ 向量 c,$c^T\widehat{\beta}$ 为 $c^T\beta$ 的唯一的最佳线性无偏（BLU）估计。

证明

(1) 显然的

(2) $c^T\widehat{\beta}$ 为 $c^T\beta$ 的无偏估计，而线性是显然的，现证 $c^T\widehat{\beta}$ 的方差最小，首先

$$Var(c^T\widehat{\beta}) = Var(c^T(X^TX)^{-1}X^Ty) = \sigma^2 c^T(X^TX)^{-1}c$$

另一方面，设 a^Ty 为 $c^T\beta$ 的任一无偏估计，于是 a 满足 $a^Ty = c^T$, 则

$$\begin{aligned} Var(a^Ty) - Var(c^T\widehat{\beta}) &= \sigma^2[a^Ta - c^T(X^TX)^{-1}c] \\ &= \sigma^2[a^T - c^T(X^TX)^{-1}X^T][a - X(X^TX)^{-1}c] \\ &= \sigma^2||a - X(X^TX)^{-1}c||^2 \geq 0 \end{aligned}$$

并且等号成立 $\Leftrightarrow a^T = c^T(X^TX)^{-1}X^T \Leftrightarrow a^Ty = c^T\widehat{\beta}$. 证毕

这个重要的定理奠定了 LS 估计在线性模型参数估计理论中的地位。

误差方差 σ^2 的估计也是我们感兴趣的，记 $\widehat{e} = y - X\widehat{\beta} = (I - P_X)y$, 其中 $P_X = X(X^TX)^{-1}X^T$，定义 σ^2 的估计为

$$\widehat{\sigma}^2 = ||y - X\widehat{\beta}||^2/(n-p) = \widehat{e}^T\widehat{e}/(n-p)$$

定理 2.2 $\widehat{\sigma}^2$ 是 σ^2 的无偏估计

证明

$$\begin{aligned} E\widehat{\sigma}^2 &= E\widehat{e}^T\widehat{e}/(n-p) \\ &= E[y^T(I-P_X)y]/(n-p) \\ &= E[e^T(I-P_X)e]/(n-p) \\ &= E[trace(I-P_X)ee^T] \\ &= trace[(I-P_X)Eee^T]/(n-p) \\ &= \sigma^2\frac{trace(I-P_X)}{n-p} \\ &= \sigma^2 \end{aligned}$$

通常称 $\widehat{\sigma}^2$ 为 σ^2 的 LS 估计。

若我们进一步假设误差向量 e 服从多元正态分布，则相应的模型为正态线性模型，记为

$$y = X\beta + e, e \sim N(0, \sigma^2 I)$$

定理 2.3 对于正态线性模型，我们有

(1) LS 估计 $c'\widehat{\beta}$ 是 $c'\beta$ 的极大似然 (ML) 估计，且 $c'\widehat{\beta} \sim N(c'\beta, \sigma^2 c'(X'X)^{-1}c)$;

(2) $\frac{n-p}{n}\widehat{\sigma}^2$ 为 σ^2 的 ML 估计，且 $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$;

(3) $c'\widehat{\beta}$ 与 $\widehat{\sigma}^2$ 相互独立;

(4) $c'\widehat{\beta}$ 是 $c'\beta$ 的最小方差无偏 (MVU) 估计。

证明： 记 $\mu = X\beta$, 考虑 μ 和 σ^2 的似然函数

$$L(\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}e^{-\frac{1}{2\sigma^2}||y-\mu||^2}$$

取对数，略去常数项，得

$$logL(\mu, \sigma^2) = -\frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}||y-\mu||^2$$

对均值向量 μ 的 LS 估计，$\widehat{\mu} = X\widehat{\beta}$, 我们有

$$||y - \widehat{\mu}||^2 = ||y - X\widehat{\beta}||^2 = min||y - X\beta||^2 = min||y - \mu||^2$$

于是，对每一个固定的 σ^2,

$$logL(\widehat{\mu}, \sigma^2) \geq logL(\mu, \sigma^2)$$

而 $logL(\widehat{\mu}, \sigma^2) = -\frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}||y - \widehat{\mu}||^2$，在 $\widetilde{\sigma}^2 = \frac{1}{n}||y - \widehat{\mu}||^2$ 达到最大，于是 $\widehat{\mu} = X\widehat{\beta}$ 和 $\widetilde{\sigma}^2$ 分别为 μ 和 σ^2 的 ML 估计。

由于 $rank(X) = p$, 因此对于任意的 $p \times 1$ 向量，存在 $\alpha \in R^n$ 使得 $C = \alpha X$, 于是 $C'\beta = \alpha' X\beta = \alpha'\mu$，由于 ML 估计的不变性，$C'\beta$ 的 ML 估计为 $\alpha\widehat{\mu} = \alpha X\widehat{\beta} = C'\widehat{\beta}$, 这就证明 LS 估计 $C'\widehat{\beta}$ 为 ML 估计。另 $C'\widehat{\beta} = C'(X'X)^{-1}X'y$ 为 y 的线性函数，而 $y \sim N_n(X\beta, \sigma^2 I)$, 则 $C'\widehat{\beta} \sim N(C'\beta, \sigma^2 C'(X'X)^{-1}C)$

(2) 因为 $P_X X = X$, 所以

$$\begin{aligned}
\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} &= \frac{\widehat{e}'\widehat{e}}{\sigma^2} \\
&= \frac{y'(I - P_X)y}{\sigma^2} \\
&= \frac{e'(I - P_X)e}{\sigma^2} \\
&= z'(I - P_X)z
\end{aligned}$$

其中 $z = e/\sigma \sim N_n(0, I)$，由于 $I - P_X$ 的幂等性，以及 $rank(I - P_X) = tr(I - P_X) = n - rank(P_X) = r - rank(X) = n - p$，$I - P_X$ 特征根只能为 0 或 1，于是存在正交方阵 Q，使得

$$I - P_X = Q' \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix} Q$$

令 $Y = Qz$，则 $Y \sim N_n(0, I_n)$, 对 Y 和 Q 的分块

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix}, Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

其中 $Y_{(1)}$ 为 $(n-p) \times 1$,Q_1 为 $(n-p) \times n$，于是

$$z'(I - P_X)z = Y' \begin{pmatrix} I_{n-p} & 0 \\ 0 & 0 \end{pmatrix} Y = Y'_{(1)}Y_{(1)} \sim \chi^2_{n-p}$$

为证 $c'\widehat{\beta}$ 与 $\widehat{\sigma}^2$ 的独立性，注意到 $c'\widehat{\beta}$ 与 $\widehat{\sigma}^2$ 分别为正态变量 y 的线性型和二次型。为了证明它们的独立性，先讲一个结论：

设 $X \sim N_n(\mu, I)$,A 为 $n \times n$ 对称阵, C 为 $m \times n$ 矩阵, 若 $CA = 0$, 则 CX 和 $X^{'}AX$ 相互独立。

证明: 由 A 的对称性，存在标准化正交阵 P，使得

$$P^{'}AP = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}$$

其中 $\Lambda = diag(\lambda_1, ..., \lambda_r, \lambda_i \neq 0, rank(A) = r$，由 $cA = 0$, 可推得 $cPP^{'}AP = 0$, 等价于

$$cP \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} = 0$$

若记 $D = cP = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$, 可得 $D_{11} = 0, D_{21} = 0$ 于是 D 就写为

$$D = \begin{pmatrix} 0 & D_{12} \\ 0 & D_{22} \end{pmatrix} \triangleq (0; D_1), D_1 : m \times (n - r)$$

将 P 做对应分块: $P = (P_1; P_2)$,P_1 为 $n \times r$，则

$$C = DP^{'} = (0; D_1) \begin{pmatrix} P_1^{'} \\ P_2^{'} \end{pmatrix} = D_1 P_2^{'}$$

$$A = P \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} p^{'} = P_1 \Lambda P_1^{'}$$

记 $Y = P^{'}X$, 则根据正态分布的线性组合仍是正态分布，我们知道

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} = \begin{pmatrix} P_1^{'}X \\ p_2^{'}X \end{pmatrix} \sim N_n(P\mu, I)$$

虽然，$Y_{(1)}$ 和 $Y_{(2)}$ 相互独立，且

$$CX = D_1 P_2 X = D_1 Y_{(2)}$$

$$X^{'}AX = X^{'}P_1 \Lambda P_1^{'}X = Y_{(1)}^{'} \Lambda Y_{(1)}$$

因为 CX 只依赖于 $Y_{(2)}$,$X^{'}AX$ 只依赖于 $Y_{(1)}$, 所以 CX 与 $X^{'}AX$ 独立。

从前文中可以看出，对于 $c^{'}\beta$, 它的 LS 估计和 ML 估计是相同的。但是，对于误差方差 σ^2，两者不同了，很明显 ML 估计 $\widetilde{\sigma}^2$ 是有偏的，$E(\widetilde{\sigma}^2) = \frac{n-p}{n}\sigma^2 < \sigma^2$, 即在平均定义上 ML 估计 $\widetilde{\sigma}^2$ 偏小。

定理 2.4 对于正态线性模型，$c'\widehat{\beta}$ 为 $c'\beta$ 的唯一的最小方差无偏估计 (minimum variance unbiased estimate, 简记为 MVU 估计),$\widehat{\sigma}^2$ 为 σ^2 的唯一 MVU 估计。

在回归分析中，我们的主要兴趣在回归系数 β_I，所以总是把它和常数项分开表示，记

$$X = (1; \widetilde{X}), \beta' = (\beta_0, \beta_I'$$

其中 1 表示由 n 个 1 组成的 n 维列向量，则线性回归模型可以写为

$$y = \beta_0 1 + \widetilde{X}\beta_I + e, E(e) = 0, Cov(e) = \sigma^2 I$$

在实际应用中，有时要对数据中心化，所谓中心化就是把自变量的度量起点移至已在 n 次试验中所取值的中心点处，记 $\overline{X}_j = \frac{1}{n}\sum X_{ij}, j = 1, ..., p-1$, 则:

$$y_i = r_0 + \beta_1(X_{i1} - \overline{X}_1) + ... + \beta_{p-1}(X_{i,p-1} - \overline{X}_{p-1}), i = 1, ..., n$$

其中 $r_0 = \beta_0 + \beta_1\overline{X}_1 + ... + \beta_{p-1}\overline{X}_{p-1} = \beta_0 + \overline{X}'\beta_I, \overline{X}' = (\overline{X}_1', ..., \overline{X}_{P-1}')$ 用矩阵记号，即为:

$$y = r_0 1 + \widetilde{X}_c\beta_I + e, E(e) = 0, Cov(e) = \sigma^2 I, \tag{2.3}$$

其中 $\widetilde{X}_c = (I - \frac{1}{n}11')\widetilde{X}$，称为中心化设计矩阵，它具有性质 $\widetilde{X}'1 = 0$

$$\widetilde{X}'(I - \frac{1}{n}11')1 = \widetilde{X}'(1 - 1) = 0$$

称 (2.3) 为中心化线性回归模型。

对于中心化的线性回归模型（2.3)，正则方程为:

$$\begin{pmatrix} n & 0 \\ 0 & \widetilde{X}_c'\widetilde{X}_c \end{pmatrix}\begin{pmatrix} r_0 \\ \beta_I \end{pmatrix} = \begin{pmatrix} n\overline{y} \\ \widetilde{X}_c'y \end{pmatrix}$$

即:

$$\begin{cases} nr_0 = n\overline{y} \\ \widetilde{X}_c'\widetilde{X}_c\beta_I = \widetilde{X}_c'y \end{cases}$$

其中 $\overline{y} = \frac{1}{n}\sum y_i$, 由此可得 r_o 和 β_I 的 LS 估计为 $\widehat{r}_0 = \overline{y}, \widehat{\beta}_I = (\widetilde{X}_c'\widetilde{X}_c)^{-1}\widetilde{X}_c'y$, 并且 $Cov\begin{pmatrix} \widehat{r}_0 \\ \widehat{\beta}_I \end{pmatrix} = \sigma^2\begin{pmatrix} \frac{1}{n} & 0 \\ 0 & (\widetilde{X}_c'\widetilde{X}_c)^{-1} \end{pmatrix}$

这个事实说明，在中心化线性回归模型中，常数项 r_0 总是用因变量观测值的算术平均值来估计，而回归习俗 β_I 的估计可以从线性回归模型 $y = \widetilde{X}_c\beta_I + e$ 按通常的 LS 公式即得，并且这两个估计总是不相关的。

如果没有中心化，记 $X = (1; \widetilde{X})$, 利用分块矩阵求逆公式有:

$$\begin{aligned}\widehat{\beta} &= \begin{pmatrix} n & 1'\widetilde{X} \\ \widetilde{X}'1 & \widetilde{X}_c'\widetilde{X}_c \end{pmatrix}^{-1} \begin{pmatrix} 1'y \\ \widetilde{X}_c'y \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n} + \overline{x}'(\widetilde{X}_c'\widetilde{X}_c)^{-1}\overline{x} & -\overline{x}'(\widetilde{X}_c'\widetilde{X}_c)^{-1} \\ (\widetilde{X}_c'\widetilde{X}_c)^{-1}\overline{x} & (\widetilde{X}_c'\widetilde{X}_c)^{-1} \end{pmatrix} \begin{pmatrix} 1'y \\ \widetilde{X}_c'y \end{pmatrix} \\ &= \begin{pmatrix} \overline{y} - \overline{x}'(\widetilde{X}_c'\widetilde{X}_c)^{-1}\widetilde{X}_c'y \\ (\widetilde{X}_c'\widetilde{X}_c)^{-1}\widetilde{X}_c'y \end{pmatrix}\end{aligned}$$

上式的后 p-1 个分量即为 $\widehat{\beta}_I$, 这就证明了我们的结论，从第一分量知

$$\widehat{\beta}_0 = \overline{y} - \overline{x}'\widehat{\beta}_I = \widehat{r}_0 - \overline{x}'\widehat{\beta}_I$$

等价地

$$\widehat{r}_0 = \widehat{\beta}_0 + \overline{x}'\widehat{\beta}_I$$

除了中心化，对自变量经常做的另一种处理称为标准化。记

$$s_j^2 = \sum (x_{ij} - \overline{x}_j)^2, j = 1, 2, ..., p-1$$

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$$

令 $Z = (z_{ij})$, 则 Z 就是将原来的设计矩阵 X 经过中心化和标准化后得到的新设计阵，这个矩阵具有如下性质:

(1) $1'Z = 0$,

(2) $R = Z'Z = (r_{ij})$,

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{s_i s_j}, i, j = 1, ..., p-1$$

性质 (1) 是中心化的作用，它使设计阵的每列元素之和都为 0。性质 (2) 是中心化后再施以标准化后的结果。$R = Z'Z$ 的元素 r_{ij} 正是回归自变量 X_i 和 X_j 的样本相关系数。因此，R 是回归自变量的相关阵，于是 $r_{ii} = 1$, 对一切 i 成立。

标准化的好处有二，其一是用 R 可以分析回归自变量之间的相关关系；其二是在一些问题中，诸回归自变量所用的单位可能不相同，取值范围大小也不同，经过标准化消去了单位和取值范围的差异，这便于对回归系数的估计值的统计分析。

如果把线性回归模型既经过中心化，又经过标准化，则 y_i 变形为

$$y_i = r_0 + \frac{x_{i1} - \overline{x}_1}{s_1}\beta_1^0 + ... + \frac{x_{ip-1} - \overline{x}_{p-1}}{s_{p-1}}\beta_{p-1}^0 + e_i, i = 1, ..., n$$

其中 $\beta_i^0 = s_i\beta_i, i = 1, ..., p-1$, 记 $\beta^0 = (\beta_1^0, ..., \beta_{p-1}^0)'$, 用矩阵形式即为

$$y = r_0 1 + Z\beta^0 + e$$

则 r_0 和 β^0 的 LS 估计分别为

$$r_0 = \overline{y}, \widehat{\beta}_i^0 = s_i\widehat{\beta}_i, i = 1, ..., p-1$$

于是对应的经验回归方程分别为:

非中心化：$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + ... + \widehat{\beta}_{p-1} X_{P-1}$

中心化：$\widehat{Y} = \widehat{r}_0 + \widehat{\beta}_1(X_1 - \overline{x}_1) + ... + \widehat{\beta}_{p-1}(X_{P-1} - \overline{x}_{p-1})$

中心化标准化：$\widehat{Y} = \widehat{r}_0 + \widehat{\beta}_1(X_1 - \overline{x}_1)/s_1 + ... + \widehat{\beta}_1(X_1 - \overline{x}_1)/s_{p-1}$

第三章　回归方程和系数的检验

本节主要有回归方程的显著性检验和回归系数的显著性检验。

3.1 模型的检验

对于正态线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + e_i, e_i \sim N(0, \sigma^2), i = 1, 2, ..., n$$

所谓回归方程的显著性检验，就是假设检验，所有的回归系数都等于 0，即检验

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0. \tag{3.1}$$

如果这个假设被拒绝，这就意味着我们接受断言，至少有一个 $\beta_i \neq 0$。换句话说，我们认为 Y 线性依赖于至少某一个自变量 X_i，也可能线性依赖于所有的自变量 $X_1, ..., X_{p-1}$。如果这个假设被接受，这意味着我们接受断言，所有的 $\beta_i = 0$，即我们可以认为，相对于误差而言，所有自变量对因变量 Y 的影响是不重要的。

为了构造检验方法，我们先简单介绍一下似然比检验。

设随机向量 y 服从参数为 $\theta \in \Theta$ 的概率分布族，考虑参数检验问题:

$$H_0 : \theta \in \Theta_0; H_1 : \theta \in \Theta_0^c$$

这里 Θ_0 为 Θ 的一个子集，Θ_0^c 为 Θ_0 的补集。记 $L(\theta|y)$ 为似然函数，$\widehat{\theta}$ 为 θ 的 MLE，$\widehat{\theta}_H$ 为原假设成立时 θ 的 MLE，于是

$$\sup_{\theta \in \Theta} L(\theta|y) = L(\widehat{\theta}|y)$$

$$\sup_{\theta \in \Theta_0} L(\theta|y) = L(\widehat{\theta}_H|y)$$

似然比定义为:

$$\lambda(y) = \frac{\sup\limits_{\theta \in \Theta} L(\theta|y)}{\sup\limits_{\theta \in \Theta_0} L(\theta|y)} = \frac{L(\widehat{\theta}|y)}{L(\widehat{\theta}_H|y)}$$

显然 $\lambda(y) \geq 1$，因为 $L(\widehat{\theta}_H|y)$ 为原假设成立时观测到样本点 y 的可能性的一个度量。当在 $\lambda(y)$ 比较大时，则 $L(\widehat{\theta}_H|y)$ 相对较小，即原假设成立，观测到样本点 y 的可能性较小。自然地，在 $\lambda(y)$ 较大时拒绝原假设，于是取检验的拒绝域形为 $\{y : \lambda(y) > c\}$,c 为一个待定常数，在实际的问题中，为了方便求检验统计量的分布，往往需要求分布。

已知的 $\lambda(y)$ 的单调函数 $G(y)$，这样的检验通常称为似然比检验 (likelihood ratio test)。

对于正态线性回归模型，未知参数 $\theta=(\beta,\sigma^2)$ 的似然函数为

$$L(\theta|y)=L(\beta,\sigma^2|y)=(2\pi)^{\frac{n}{2}}\sigma^{-n}exp(-\frac{1}{2\sigma^2}||y-X\beta||^2)$$

根据定理 2.3,$\widehat{\beta}$ 的 MLE 为 $\beta=(X^{'}X)^{-1}X^{'}y$,$\sigma^2$ 的 MLE 为 $\widetilde{\sigma}^2=\frac{1}{n}||y-X\beta||^2$.

因此似然函数对应的极大值为

$$\sup_{\beta,\sigma^2}L(\beta,\sigma^2|y)=L(\widehat{\beta},\widetilde{\sigma}^2|y)=(2\pi)^{\frac{n}{2}}\widetilde{\sigma}^{-n}exp(-\frac{n}{2})$$

$$\sup_{\beta,\sigma^2\in\Theta_0}L(\beta,\sigma^2|y)=L(\widehat{\beta}_H,\widehat{\sigma}_H^2|y)=(2\pi)^{\frac{n}{2}}\widetilde{\sigma}_H^{-n}exp(-\frac{n}{2})$$

似然比为

$$\lambda(y)=\frac{\sup\limits_{\beta,\sigma^2}L(\beta,\sigma^2|y)}{\sup\limits_{\beta,\sigma^2\in\Theta_0}L(\beta,\sigma^2|y)}=\frac{L(\widehat{\beta},\widetilde{\sigma}^2|y)}{L(\widehat{\beta}_H,\widehat{\sigma}_H^2|y)}=(\frac{||Y-X\widehat{\beta}_H||^2}{||Y-X\widehat{\beta}||^2})^{\frac{n}{2}}$$

记 $SS_e=||Y-X\widehat{\beta}||^2$ 表示模型残差平方和，$SS_{He}=||Y-X\widehat{\beta}_H||^2$ 表示模型在约束 Θ_0 下的残差平方和。

如果考虑齐次线性假设，即 $H\beta=0$,H 为 $m\times P$ 已知矩阵，rk(H)=m,(本来还有约束 $M(H^{'})\subset M(X^{'})$，由于 X 是列满秩，不需要了）.

3.1.1 约束最小二乘估计

假设 $H\beta=d$，H 为 $k\times p$ 的已知矩阵，d 为 $k\times 1$ 的已知向量，且矩阵秩为 k，记

$$H=\begin{pmatrix}h_1^{'}\\...\\h_k^{'}\end{pmatrix},d=\begin{pmatrix}d_1\\...\\d_k\end{pmatrix}$$

则线性约束可以改写为

$$h_i^{'}\beta=d_i,i=1,...,k.$$

我们要在 k 个条件下求 β 使 $Q(\beta)=||y-X\beta||^2$ 达到最小值，为了应用 Lagrange 乘子法，构造辅助函数

$$\begin{aligned}F(\beta,\lambda)&=||y-X\beta||^2+2\sum\lambda_i(h_i^{'}\beta-d_i)\\&=||y-X\beta||^2+2\lambda^{'}(H\beta-d)\end{aligned}$$

$$= (y - X\beta)^{'}(y - X\beta) + 2\lambda^{'}(H\beta - d)$$

其中 $\lambda = (\lambda_1, ..., \lambda_k)^{'}$ 为 Lagrange 乘子法，对函数 $\lambda^{'}(H\beta - d)$ 求对 β 的偏导数，整理并令它们等于零，得到

$$X^{'}X\beta = X^{'}y - H'\lambda, \tag{1}$$

, 然后求解

$$\begin{cases} X^{'}X\beta = X^{'}y - H'\lambda \\ H\beta = d \end{cases}, \tag{2}$$

p+k 个方程，p+k 个参数。

由 (1) 得

$$\widehat{\beta}_H = (X'X)^{-1}X'y - (X'X)^{-1}H'\widehat{\lambda}$$

代入 (2) 得

$$d = H\beta - H(X'X)^{-1}H'\lambda$$

得

$$\widehat{\lambda}_H = (H(X'X)^{-1}H')^{-1}(H\widehat{\beta} - d)$$

将 $\widehat{\lambda}_H$ 代入 $\widehat{\beta}_H$ 得到

$$\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta} - d)$$

下面我们证明 $\widehat{\beta}_H$ 确实是线性约束 $H\beta = d$ 下 β 的最小二乘解。为此我们需要证明如下两点

(a) $H\widehat{\beta}_H = d$

(b) 对一切满足 $H\beta = d$ 的 β，都有

$$||y - X\beta||^2 \geq ||y - X\widehat{\beta}_H||^2$$

(a) 的证明是容易的。为了证明 (b)，我们将平方和 $||y - X\beta||^2$ 作分解

$$\begin{aligned} ||Y - X\beta||^2 &= ||Y - X\widehat{\beta}||^2 + (\widehat{\beta} - \beta)'X'X(\widehat{\beta} - \beta) \\ &= ||Y - X\beta||^2 + (\widehat{\beta} - \widehat{\beta}_H + \widehat{\beta}_H - \beta)'X'X(\widehat{\beta} - \widehat{\beta}_H + \widehat{\beta}_H - \beta) \\ &= ||Y - X\beta||^2 + (\widehat{\beta} - \widehat{\beta}_H)'X'X(\widehat{\beta} - \widehat{\beta}_H) + (\widehat{\beta}_H - \beta)'X'X(\widehat{\beta}_H - \beta) \\ &= ||Y - X\beta||^2 + ||X(\widehat{\beta} - \widehat{\beta}_H)||^2 + ||X(\widehat{\beta}_H - \beta)||^2. \end{aligned} \tag{3}$$

交叉项为 0，因为

$$\begin{aligned}(\widehat{\beta}-\widehat{\beta}_H)'X'X(\widehat{\beta}_H-\beta) &= \widehat{\lambda}_H'H(\widehat{\beta}_H-\beta)\\ &= \widehat{\lambda}_H'(d-d)\\ &= 0\end{aligned}$$

因此 (3) 式证明，对一切满足 $H\beta=d$ 的 β 总有

$$||Y-X\beta||^2 \geq ||Y-X\widehat{\beta}||^2+||X(\widehat{\beta}-\widehat{\beta}_H)||^2. \tag{4}$$

等号成立当且仅当 (3) 式的第三项等于零，也就是 $X\beta=X\widehat{\beta}_H$. 于是在 (4) 式中用 $X\widehat{\beta}_H$ 代替 $X\beta$, 等式成立，即

$$||Y-X\widehat{\beta}_H||^2 = ||Y-X\widehat{\beta}||^2+||X(\widehat{\beta}-\widehat{\beta}_H)||^2, \tag{5}$$

综合 (4) 和 (5)，便证明了结论 (b)。

定理 3.1 对于线性回归模型，设 H 为 $k\times p$ 矩阵，rk(H)=k, 且 $H\beta=d$，则 $\widehat{\beta}_H=\widehat{\beta}-(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta}-d)$ 为 β 的约束 LS 估计，这里 $\beta=(X'X)^{-1}X'Y$。

例 3.1：在天文测量中，最天空中三个星位点构成的三角形 ABC 的三个内角 $\theta_1,\theta_2,\theta_3$ 进行测量，得到的测量值分别为 y_1,y_2,y_3，由于存在测量误差，所以需要它们进行估计，利用线性模型表示有关的量：

$$\begin{cases}y_1=\theta_1+e_1,\\ y_2=\theta_2+e_2,\\ y_3=\theta_3+e_3,\\ \theta_1+\theta_2+\theta_3=\pi,\end{cases}$$

其中 $e_i, i=1,2,3$ 表示测量误差。假设它们满足 Gauss-Markov 假设，这就是一个带有约束条件的线性模型，将它写成矩阵形式

$$\begin{cases}y=X\beta+e,\\ H\beta=b,\end{cases}$$

其中 $y=(y_1,y_2,y_3)', \beta=(\theta_1,\theta_2,\theta_3)', X=I_3$,$I_3$ 表示 3 阶单位阵，$H=(1,1,1)', b=\pi$，则可得到 β 的约束最小二乘估计为

$$\widehat{\beta}_c=\widehat{\beta}-(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta}-d)$$

其中 $\beta=(X'X)^{-1}X'Y$ 是 β 的无约束最小二乘估计，经计算可得

$$\widehat{\beta}_c=\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}-\frac{1}{3}(\sum y_i-\pi)\begin{pmatrix}1\\1\\1\end{pmatrix}$$

其中 $\theta_i=y_i-\frac{1}{3}(y_1+y_2+y_3-\pi), i=1,2,3$ 为 θ_i 的约束最小二乘估计。

和前面类似，我们可以构造 σ^2 的约束 LS 估计如下，

$$\widehat{\sigma}_H^2=\frac{||y-X\widehat{\beta}_H||^2}{n-r+k}$$

定理 3.2 在定理 3.1 的假设下，在参数区域 $H\beta=d$ 上，$\widehat{\sigma}_H^2$ 是 σ^2 的无偏估计，

证明由 $||Y-X\widehat{\beta}_H||^2=||Y-X\widehat{\beta}||^2+||X(\widehat{\beta}-\widehat{\beta}_H)||^2$, 得

$$E||Y-X\widehat{\beta}_H||^2=E||Y-X\widehat{\beta}||^2+E||X(\widehat{\beta}-\widehat{\beta}_H)||^2$$

由前面可知 $E||Y-X\widehat{\beta}||^2=(n-r)\sigma^2$，第二项

$$\begin{aligned}&E||X(\widehat{\beta}-\widehat{\beta}_H)||^2\\=&E(H\widehat{\beta}-d)'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta}-d)\\=&(H\beta-d)'(H(X'X)^{-1}H')^{-1}(H\beta-d)+tr[(H(X'X)^{-1}H')^{-1}Cov(H\widehat{\beta})]\\=&0+tr(\sigma^2I_k)\\=&k\sigma^2\end{aligned}$$

由此得证。

可以证明对于正态线性回归模型，$\widehat{\beta}_H$ 和 $\widehat{\sigma}_H^2=\frac{||Y-X\widehat{\beta}_H||^2}{n}$ 为 β 和 σ^2 的 MLE。

在约束 $H\beta=0$ 下，

$$F=\frac{n-p}{m}((\lambda(Y))^{2/n}-1)=\frac{(SS_{He}-SS_e)/m}{SS_e/(n-p)}$$

显然 F 仅依赖于 $\lambda(y)$ 且为 $\lambda(y)$ 的严增函数。

定理 4.1 对于正态线性回归模型

(1) $SS_e\sim\sigma^2\chi^2_{n-p}$, 其中 p=rk(X);

(2) $SS_{He}-SS_e=(H\widehat{\beta})'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta})\sim\sigma^2\chi^2_{m,\delta}$，其中 δ 为非中心化参数

$$\delta=(H\beta)'(H(X'X)^{-1}H')^{-1}(H\beta)/\sigma^2;$$

(3) $SS_{He}-SS_e$ 与 SS_e 相互独立;

(4) 当线性假设 $H\beta=0$ 为真时，$F\sim F_{m,n-p}$.

证明

(1) 前面已证。

(2) 记 $P_X=X(X'X)^{-1}X'$,$A=X(X'X)^{-1}H'(H'(X'X)^{-1}H')^{-1}H$, 利用 $\widehat{\beta}_H$ 的定义及 $(I-P_X)A=0$，有

$$\begin{aligned}SS_e&=||Y-X\widehat{\beta}_H||^2\\&=||Y-X\widehat{\beta}+A\widehat{\beta}||^2\\&=||Y-X\widehat{\beta}||^2+2(Y-X\widehat{\beta})'A\widehat{\beta}+\widehat{\beta}'A'A\widehat{\beta}\\&=||Y-X\widehat{\beta}||^2+2Y'(I-P_X)A\widehat{\beta}+\widehat{\beta}'A'A\widehat{\beta}\\&=SS_e+\widehat{\beta}'A'A\widehat{\beta}\end{aligned}$$

显然 $A'A=H'(H(X'X)^{-1}H')^{-1}H$, 上式变为

$$SS_{He}=SS_e+(H\widehat{\beta})(H(X'X)^{-1}H')^{-1}(H\widehat{\beta})$$

即 (2) 的前半部分，至于 $SS_{He}-SS_e$ 的分布易从 $H\beta\sim N(H\beta,\sigma^2H'(X'X)^{-1}H')$ $rk(H'(X'X)^{-1}H')=m$ 得到。

(3) 因为 SS_e 和 $SS_{He}-SS_e$ 可以分别表示为

$$SS_e=Y'(I-P_X)Y$$

$$SS_{He}-SS_e=Y'X(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}H(X'X)^{-1}X'Y\triangleq Y'BY$$

且 $(I-P_X)B=0$. 下面介绍一个引理:

Lemma 设 $X\sim N_n(\mu,I)$,A,B 皆 $n\times n$ 对称，若 AB=0，则 $X'AX$ 与 $X'BX$ 相互独立。

根据此引理，立得 SS_e 和 $SS_{He}-SS_e$ 相互独立。

(4) 是 (1) (3) 及 F 分布定义的直接推论。定理证毕。

3.1.2 回归方程的显著性检验

对于正态线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + e_i, e_i \sim N(0, \sigma^2), i = 1, 2, ..., n$$

所谓回归方程的显著性检验，就是假设检验，所有的回归系数都等于 0，即检验

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0, i.e. H\beta = 0, H = (0, I_{p-1})$$

根据前边定理给出的 F 检验统计量可以直接应用到这里。

对于 F 统计量表达式中的 SS_e 和 SS_{He}, 实际多采用下面的计算公式

$$SS_e = ||Y - X\widehat{\beta}||^2 = Y'Y - \widehat{\beta}' X'Y, \tag{$\star$}$$

$$SS_{He} = ||Y - X\widehat{\beta}_H||^2 = Y'Y - \widehat{\beta}'_H X'Y, \tag{$\star\star$}$$

($\star$) 的证明比较简单，下证 ($\star\star$), 我们知道，β 在条件 $H\beta = 0$ 下的约束 LS 解 β_H 满足方程组

$$\begin{cases} X'X\widehat{\beta}_H + H'\widehat{\lambda} = X'Y \\ H\widehat{\beta}_H = 0 \end{cases}$$

其中 λ 为拉氏乘子，利用此事实，有

$$\begin{aligned} SS_{He} &= ||Y - X\widehat{\beta}_H||^2 \\ &= (Y'Y - \widehat{\beta}_H X'Y) + \widehat{\beta}'_H X'X\widehat{\beta}_H - \widehat{\beta}'_H X'Y \\ &= (Y'Y - \widehat{\beta}_H X'Y) + \widehat{\beta}'_H (X'X\widehat{\beta}_H - X'Y) \\ &= (Y'Y - \widehat{\beta}_H X'Y) + \widehat{\beta}'_H H'\widehat{\lambda} \\ &= Y'Y - \widehat{\beta}'_H X'Y \end{aligned}$$

($\star\star$) 得证。

($\star$) 式中 $\widehat{\beta}' X'Y$ 等于未知参数 β 的 LS 解与正则方程右端向量 $X'Y$ 的内积，表示了数据平方和 Y'Y 中能够由因变量 Y 与自变量 $X_1, ..., X_p$ 的线性关系所能解释的部分，称为回归平方和 (regression sum of squares，记为 RSS)，即：记 $RSS(\beta) = \widehat{\beta}' X'Y$, 则

$$SS_e = Y'Y - RSS(\beta)$$

即残差平方和等于总平方和减去回归平方和。类似的，$\widehat{\beta}'_H X'Y$ 称为约束条件 $H\beta = 0$ 下的回归平方和，记为 $RSS_H(\beta)$, 则

$$SS_{He} = Y'Y - RSS_H(\beta)$$

F 统计量可以记为

$$F = \frac{(RSS(\beta) - RSS_H(\beta))/m}{SS_e/(n-r)}$$

下面就现在的特殊情形，导出检验统计量的简单形式。

若 H_0 成立，约简模型为

$$y_i = \beta_0 + e_i, i = 1, 2, ..., n.$$

β_0 的最小二乘估计为 $\widehat{\beta}_0 = \overline{y}$，于是 β_0 对应的回归平方和为

$$RSS_{H_0}(\beta) = RSS(\beta_0) = n\overline{Y}^2$$

而对原模型（无约束），从中心化回归模型知，回归平方和为

$$RSS(\beta) = \widehat{\gamma}_0 ny + \widehat{\beta}_I' \widetilde{X}_c\text{'}y = n\overline{y}^2 + \widehat{\beta}_I' \widetilde{X}_c\text{'}y$$

于是

$$RSS(\beta) - RSS_{H_0}(\beta) = \widehat{\beta}_I' \widetilde{X}_c\text{'}y$$

于是原模型残差平方和为

$$SS_e = y'y - RSS(\beta) = y'y - n\overline{y}^2 - \widehat{\beta}_I' \widetilde{X}_c\text{'}y$$

于是由于 m=p-1, 于是我们可以得到检验假设

$$F = \frac{\widehat{\beta}_I' \widetilde{X}_c\text{'}y/(p-1)}{SS_e/(n-p)}$$

而原假设成立时，$F \sim F_{p-1,n-p}$ 对给定的水平 α, 当 $F > F_{p-1,n-p}(\alpha)$ 时，我们拒绝原假设 H_0，否则就接受 H_0。

需强调的是，如果经过检验，结论是接受原假设 H_0 $\beta_1 = ... = \beta_{p-1} = 0$, 这意思就是说，和模型的各种误差比较起来，诸自变量对 Y 的影响是不重要的。这里可能有两种情况，其一是，模型的各种误差太大，因而即使虎归自变量对 Y 有一定的影响，但与较大的模型误差相比，也不算大。对这种情况，我们就要想办法缩小误差，这包括从分析问题的专业背景入手，检查是否漏掉了重要的自变量，或 Y 对某些自变量有非线性相依关系等。其二是，回归自变量对 Y 的影响确实很小，对这种情况，我们就要放弃建立 Y 对诸自变量的线性回归。

3.2 回归参数的检验

回归方程的显著性检验是对线性回归方程的一个整体性检验。如果我们检验的结果是拒绝原假设，这意味着因变量 Y 线性地依赖于自变量 $X_1, ..., X_{p-1}$ 这个回归自变量的整体。但是，这并不排除 Y 并不依赖于其中某些自变量，即某些 β_i 可能等于零。于是在回归方程显著性检验被拒绝之后，我们还需要对每个自变量逐一做显著性检验，即对固定的 i$(1 \leq i \leq p-1)$, 做如下检验:

$$H_i : \beta_i = 0$$

此假设也是一般线性假设的一种特殊情况，利用前面的定理可以获得所需的检验。对于以上问题，下面我们给出一种直接导出检验统计量的方法。

由于 β 的 LS 估计为 $\widehat{\beta} = (X'X)^{-1}X'Y$，我们有

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

记 $C_{p\times p} = (c_{ij}) = (X'X)^{-1}$, 则有

$$\widehat{\beta_i} \sim (\beta_i, \sigma^2 c_{ii})$$

于是当 H_i 成立时，

$$\frac{\widehat{\beta_i}}{\sigma\sqrt{c_{ii}}} \sim N(0,1)$$

由于 $(n-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$, 且与 $\widehat{\beta_i}$ 相互独立，这里 $\widehat{\sigma}^2 = ||Y - X\widehat{\beta}||^2/(n-p)$，根据 t 分布的定义，有

$$t_i = \frac{\widehat{\beta_i}}{\sqrt{c_{ii}}\widehat{\sigma}} \sim t_{n-p}$$

这里 t_{n-p} 表示自由度为 n-p 的 t 分布。对给定的水平 α，当 $|t_i| > t_{n-p}(\alpha/2)$ 时，我们拒绝原假设 H_i，否则就接受 H_i。

如果我们经过检验，接受原假设 $\beta_i = 0$ 时，我们就认为回归自变量 X_i 对因变量 Y 无显著的影响，因而可以将其从回归方程中剔除。将这个回归自变量从回归方程中剔除后，剩余变量的回归系数的估计也随之发生变化。将 Y 对剩余的回归自变量重新做回归，然后再检验其余回归系数是否为零，再剔除经检验认为对 Y 无显著影响的变量，这样的过程一直继续下去，直到对所有的自变量，经检验都认为对 Y 有显著的影响为止。对回归系数做显著性检验的过程，事实上也是对回归自变量的选择过程。

例 6.2.1 表 6.2.1 给出了煤净化过程的一组数据，Y 为净化后煤溶液张所含杂质的重量；X_1 表示输入净化过程的溶液所含的煤与杂质的比；X_2 是溶液的 pH 值；X_3 表

示溶液流量。通过一组试验数据，建立净化效率 Y 与三个因素 X_1, X_2, X_3 的经验关系，进而据此通过控制某些自变量来提高净化效率 (表 6.2.1).

表 6.2.1　煤净化数据

编号	x_1	x_2	x_3	y
1	1.50	6.00	1315	243
2	1.50	6.00	1315	261
3	1.50	9.00	1890	244
4	1.50	9.00	1890	285
5	2.00	7.50	1575	202
6	2.00	7.50	1575	180
7	2.00	7.50	1575	183
8	2.00	7.50	1575	207
9	2.50	9.00	1315	216
10	2.50	9.00	1315	160
11	2.50	6.00	1890	104
12	2.50	6.00	1890	110

考虑回归模型

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

应用最小二乘法，得到回归系数的估计

$$(\beta_0, \beta_1, \beta_2, \beta_3)' = (397.087, -110.750, 15.583, -0.058)'$$

先考虑回归方程的显著性检验，即 $\beta_1 = \beta_2 = \beta_3 = 0$, 经计算得到

$$\widehat{\beta}_I' \widetilde{X}_c = 31156.02, SS_e = 3486.89.$$

于是 F 统计量为

$$F = \frac{\widehat{\beta}_I' \widetilde{X}_c' y/3}{SS_e/8} = \frac{10385.33}{435.85} = 23.82$$

取 $\alpha = 0.05$, 查表得 $F_{3,8}(0.05) = 4.07$. 因 $F = 23.82 > F_{3,8}(0.05) = 4.07$, 于是我们拒绝原假设 H_0，认为 Y 对 X_1, X_2, X_3 有一定的依赖关系。

进一步考虑回归系数的显著性检验，经计算得

$$c_{11} = 0.49998, c_{22} = 0.05556, c_{33} = 0.0000011.$$

结合上面已经得到的 β 的 LS 估计值及 $\sigma^2 = 435.85$, 容易算得三个回归系数对应的 t_i 值分别为

$$t_1 = -7.5, t_2 = 3.17, t_3 = -2.27$$

对给定的水平 $\alpha = 0.05$，查表得 $t_8(0.025) = 2.306$，对 i=1,2，有 $|t_i| > t_8(0.025) = 2.306$，因此，在水平 $\alpha = 0.05$，对每一个回归系数的单独检验，接受 $\beta_i \neq 0, i = 1, 2.$ 也就是，我们认为 β_1, β_2 是显著的。

3.3 复相关系数

度量随机变量 Y 与随机向量 $X' = (X_1, ..., X_{p-1})$ 相关程度的概念是复相关系数 (Multiple correlation coefficient)，定义为

$$\rho = (\sigma'_{xy}\Sigma_{xx}^{-1}\sigma_{xy})^{1/2}/\sigma_y$$

这里

$$Cov\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma'_{xy} \\ \sigma_{xy} & \Sigma_{xx} \end{pmatrix} \triangleq \Sigma$$

在 Y 与 $X_1, ..., X_{p-1}$ 的联合分布为正态分布的条件下，可以证明

$$\widehat{\sigma}_y^2 = \frac{1}{n}\sum(y_i - \overline{y})^2, \overline{y} = \frac{1}{n}\sum y_i$$

$$\widehat{\Sigma}_{xx} = \frac{1}{n}\widetilde{X}'_c\widetilde{X}_c, \widehat{\sigma}_{xy} = \frac{1}{n}X'_c(y - \overline{y}1)$$

分别为 $\sigma_y^2, \Sigma_{xx}, \sigma_{xy}$ 的 ML 估计，所以可得到 ρ 的估计

$$R \triangleq \widehat{\rho} = (\frac{\widehat{\beta}_I\widetilde{X}'_c y}{\sum(y_i - \overline{y})})^{1/2}$$

称为样本复相关系数。记

$$TSS = \sum(y_i - \overline{y})^2$$

又记

$$RSS(\beta_I) = \widehat{\beta}_I\widetilde{X}'_c y$$

它是回归系数 β_I 的回归平方和, 等价于

$$R^2 = \frac{RSS(\beta_I)}{TSS}$$

即复相关系数的平方等于回归平方和与总平方和之比。若 R=1, 则 $RSS(\beta_I) = TSS$，这说明因变量的总方差完全由回归来解释，所以，Y 与 $X_1, ..., X_{p-1}$ 之间有严格的线性关系。相反，若 R=0，则 $RSS(\beta_I) = 0$，这说明只考虑 Y 与 $X_1, ..., X_{p-1}$ 之间无任何线性关系，无法解释 Y 的变差。所以，Y 与 $X_1, ..., X_{p-1}$ 之间无任何线性关系。在一般情况下，$0 < R < 1$，Y 与 $X_1, ..., X_{p-1}$ 之间有一定的线性关系。一般来说，R 越大，表明 Y 与 $X_1, ..., X_{p-1}$ 之间的线性关系程度越强。因此在应用上，R 也是度量回归方程优劣的一个重要指标。

第四章　回归自变量的选择

在应用回归分析去处理实际问题时，回归自变量的选择是首先要解决的重要问题。通常，在做回归分析时，人们根据问题本身的专业理论及有关经验，常常把各种与因变量有关或可能有关的自变量引进回归模型，其结果是把一些对因变量影响很小的，有些甚至没有影响的自变量也选入了回归模型中，这样一来，不但计算量大，而且估计和预测的精度也会下降。此外，在一些情况下，某些自变量观测数据的获得代价昂贵。如果这些自变量本身对因变量的影响很小或根本就没有影响，但我们不加选择都引进回归模型，势必造成观测数据收集和模型应用的费用不必要的加大。因此，在应用回归分析时，对进入模型的自变量作精心的选择是十分必要的。本节的目的就是对自变量的选择从理论上作一简要地分析，介绍一些变量的选择准则和一些求" 最优" 自变量子集的计算方法。

4.1 变量选择对估计和预测的影响

假设根据经验和专业理论，初步确定一切可能对因变量 Y 有影响的自变量共有 p-1 个，记为 $X_(1), ..., X_{(p-1)}$，它们与因变量一起适合线性回归模型。在获得了 n 组观测数据后，我们有模型

$$y = X\beta + e, E(e) = 0, Cov(e) = \sigma I$$

假设我们根据某些自变量选择标准，剔除了一些对因变量影响较小的自变量，不妨假设剔除了后 p-q 个自变量 $X_q, ..., X_{p-1}$，记 $X = (X_q; X_t)$,$\beta' = (\beta'_q; \beta'_t)$, 则我们得到一个新模型

$$y = X_q\beta_q + e, E(e) = 0, Cov(e) = \sigma I$$

在全模型下，回归系数 β 的 LS 估计为

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

而在选择模型下，β_q 的 LS 估计为

$$\widetilde{\beta}_q = (X'_qX_q)^{-1}X'_qy$$

对 $\widehat{\beta}$ 作相应的分块：$\widehat{\beta}' = (\widehat{\beta}'_q; \widehat{\beta}'_t)$。

定理 4.1 假设全模型正确，则

(1) $E(\widetilde{\beta}_q) = \beta_q + A\beta_t$, 这里 $A = (X'_qX_q)^{-1}X'_qX_t$

(2) $Cov(\widehat{\beta}_q) \geq Cov(\widetilde{\beta}_q)$

证明

(1) $E(\widetilde{\beta}_q) = (X_q'X_q)^{-1}X_q'E(y) = (X_q'X_q)^{-1}X_q'(X_q; X_t)\begin{pmatrix} \beta_q \\ \beta_t \end{pmatrix} = (I; A)\begin{pmatrix} \beta_q \\ \beta_t \end{pmatrix} = \beta_q + A\beta_t$

于是 (1) 得证。

(2) 根据分块矩阵的逆矩阵公式，有

$$(X'X)^{-1} = \begin{pmatrix} X_q'X_q & X_q'X_t \\ X_t'X_q & X_t'X_t \end{pmatrix}^{-1} = \begin{pmatrix} (X_q'X_q)^{-1} + ADA' & -AD \\ -DA' & D \end{pmatrix}$$

这里 $D = (X_t'(I - P_{X_q})X_t)^{-1}$. 又由

$$Cov(\widehat{\beta}) = Cov\begin{pmatrix} \widehat{\beta}_q \\ \widehat{\beta}_t \end{pmatrix} = \sigma^2(X'X)^{-1}$$

推得 $Cov(\widehat{\beta}_q) = \sigma^2((X_q'X_q)^{-1} + ADA')$, 但 $Cov(\widetilde{\beta}_q) = \sigma^2(X_q'X_q)^{-1}$, 所以

$$Cov(\widehat{\beta}_q) - Cov(\widetilde{\beta}_q) = \sigma^2 ADA'$$

因为 $(X'X)^{-1} > 0$, 所以 $D > 0$。于是 $Cov(\widehat{\beta}_q) - Cov(\widetilde{\beta}_q) > 0$, 从而 (2) 得证。

对于未知参数 θ 的有偏估计 $\widetilde{\theta}$, 协方差阵不能作为衡量估计精度之用，更合理的是均方误差矩阵 (mean square error matrix, 简记为 MSEM). 定义为

$$MSEM(\widetilde{\theta}) = E(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)'$$

易得

$$MSEM(\widetilde{\theta}) = Cov(\widetilde{\theta}) + (E\widetilde{\theta} - \theta)(E\widetilde{\theta} - \theta)'$$

定理 4.2 假设全模型正确，则当 $Cov(\widehat{\beta}_t) \geq \beta_t\beta_t'$ 时，

$$MSEM(\widehat{\beta}_q) \geq MSEM(\widetilde{\beta}_q)$$

证明对估计 $\widetilde{\beta}_q$，易得

$$MSEM(\widetilde{\beta}_q) = \sigma^2(X_q'X_q)^{-1} + A\beta_t\beta_t'A'$$

注意到 $\widehat{\beta}_q$ 为无偏估计，所以

$$MSEM(\widehat{\beta}_q) = \sigma^2((X_q'X_q)^{-1} + ADA')$$

又因 $Cov(\widehat{\beta}_t) = \sigma^2 D$. 故当 $Cov(\widehat{\beta}_t) \geq \beta_t\beta_t'$ 时，$MSEM(\widehat{\beta}_q) \geq MSEM(\widetilde{\beta}_q)$。定理得证。

下面来考虑变量选择对因变量的预测的影响。

假设我们欲预测点 $x_0 = (x_{0q}, ..., x_{0t})$ 对应的因变量 y_0 的值。已知

$y_0 = x_0'\beta + e = x_{0q}'\beta_q + x_{0t}'\beta_t + \varepsilon, E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$, ε 与 e 不相关

在全模型下，我们用 $\widehat{y}_0 = x_0'\widehat{\beta}$ 作为 y_0 的预测，预测偏差为 $z = x_0'\widehat{\beta} - y_0$. 而在选模型下，用 $\widetilde{y}_0 = x_{0q}'\widetilde{\beta}_q$ 作为 y_0 的预测，预测偏差为 $z_q = x_{0q}'\widetilde{\beta}_q - y_0$. 显然，若全模型正确，则预测 $\widehat{y}_0$ 是无偏的，即 E(z)=0. 下面讨论预测偏差的性质。

定理 4.3 假设全模型正确，则

(1) $E(z_q) = x_{0q}'A\beta_t - x_{0t}'\beta_t$，这里 $A = (X_q'X_q)^{-1}X_q'X_t$;

(2) $Var(z) \geq Var(z_q)$。

证明

(1) 因 $E(y_0) = x_{0q}'\beta_q + x_{0t}'\beta_t$，依定理 4.1 立得 (1)。

(2) 依假设，ε 与 e 不相关, 故

$$Var(z) = \sigma^2(1 + x_0'(X'X)^{-1}x_0), Var(z_q) = \sigma^2(1 + x_{0q}'(X_q'X_q)^{-1}x_{0q})$$

根据分块矩阵的公式得到

$$\begin{aligned} Var(z) - Var(z_q) &= \sigma^2(x_0' \begin{pmatrix} (X_q'X_q)^{-1} + ADA' & -AD \\ -DA' & D \end{pmatrix} x_0 - x_{0q}'(X_q'X_q)^{-1}x_{0q}) \\ &= \sigma^2(x_{0q}'ADA'x_{0q} - 2x_{0q}'ADx_{0t} + x_{0t}'Dx_{0t}) \\ &= \sigma^2(A'x_{0q} - x_{0t})'D(A'x_{0q} - x_{0t}) \geq 0. \text{式子 } 1 \end{aligned}$$

定理证毕。

这个定理的第一条结论说明，$\widetilde{y}_0$ 不是无偏预测。和估计的情形一样，这时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).$\widetilde{y}_0$ 的 MSEP 定义为:

$$MSEP(\widetilde{y}_0) = E(\widetilde{y}_0 - y_0)^2 = E(z_q^2) = Var(z_q) + (E(z_q))^2$$

定理 4.4 假设全模型正确，则当 $Cov(\widehat{\beta}_t) \geq \beta_t\beta_t'$ 时，

$$MSEP(\widehat{y}_0) \geq (\widetilde{y}_0)$$

证明得

$$MSEP(\widehat{y}_0) = Var(z)$$

根据假设条件及定理 4.1(1), 有

$$(E(z_q))^2 = (x_{0q}'A\beta_t - x_{0t}'\beta_t)^2 = (x_{0q}'A - x_{0t}')\beta_t\beta_t'(A'x_{0q} - x_{0t}) \leq (x_{0q}'A - x_{0t}')Cov(\widehat{\beta}_t)(A'x_{0q} - x_{0t})$$

因为 $Cov(\widehat{\beta}_t) = \sigma^2 D$, 并利用式子 1, 得

$$(E(z_q))^2 \leq Var(z) - Var(z_q)$$

从而有

$$MSEP(\widehat{y}_0) = Var(z) \geq Var(z_q) + (E(z_q))^2 = MSEP(\widetilde{y}_0)$$

综上，我们有如下结论:

(1) 即使全模型正确，剔除一部分自变量之后，可使得剩余的那部分自变量的回归系数的 LS 估计的方差减小，但此时的估计一般为有偏估计。若被剔除的自变量对因变量影响较小，则可使得剩余的那部分自变量的回归系数的 LS 估计的精度提高。

(2) 当全模型正确时，用选择模型作预测，预测一般是有偏的，但预测偏差的方差减小。若被剔除的自变量对因变量影响较小，则剔除掉这些变量后可使得预测的精度提高。

因此，在应用回归分析去处理实际问题时，无论是从回归系数的估计角度看，还是从预测的角度看，对那些与因变量关系不是很大或难于掌握 (用 $Cov(\widehat{\beta}_t) \geq \beta_t\beta_t'$ 来刻画) 的自变量从模型中剔除都是有利的。

4.2 自变量选择准则

统计学家从数据与模型的拟合优劣，预测精度等不同角度出发提出了多种回归自变量的选择准则，它们都是对回归自变量的所有不同子集进行比较，然后从中挑出一个”最优”的，且绝大多数选择准则是基于残差平方和的。

1. 平均残差平方和准则 (RMS_q)

残差平方和 SS_e 的大小刻画了数据与模型的拟合程度，SS_e 越小，拟合的越好。但‘SS_e 越小越好’却不能作为回归自变量的选择准则，因为它将导致全部自变量的入选。事实上，在选模型下，残差平方和为

$$SS_{eq} = ||y - X_q\widehat{\beta}_q||^2 = y'(I - P_{X_q})y$$

如果在选模型中再增加一个变量，设对应的设计阵为 $X_{q+1} = (X_q; b)$，则残差平方和为

$$SS_{eq+1} = y'(I - P_{X_{q+1}})y$$

利用分块矩阵求逆公式，不难说明 $P_{X_{q+1}} \geq P_{X_q}$，故 $SS_{eq+1} \leq SS_{eq}$.

为了防止选取过多的自变量，一种常见的作法是在残差平方和 SS_{eq} 上添加对增加变量的惩罚因子。平均残差平方和 RMS_q 就是其中一例，平均残差平方和 RMS_q 定义为

$$RMS_q = \frac{SS_{eq}}{n-q}$$

这里 q 为选模型设计阵 X_q 的列数。实际上 RMS_q 就是选模型下误差方差的 LS 估计。因子 $(n-q)^{-1}$ 随自变量的个数增加而变大，它体现了对自变量个数的增加所施加的惩罚。依 RMS_q 准则，按’RMS_q 越小越好‘选择自变量子集。

2.C_p 准则

C_p 准则是基于 C.L.Mallows 提出的 C_p 统计量，它是从预测的观点出发提出的。对于选模型 C_p 统计量定义为

$$C_p = \frac{SS_{eq}}{\widehat{\sigma}^2} - (n - 2q)$$

这里 SS_{eq} 为选模型下的残差平方和，$\widehat{\sigma}^2$ 为全模型下 σ^2 的 LS 估计，q 为选模型设计阵的列数。依 C_p 准则，按’C_p 越小越好’选择自变量子集。

C_p 的想法如下，如果采用选模型，那么我们用 $\widetilde{y} = X_q\widetilde{\beta}_q$ 去预测 $y = X\beta + e$, 则

$$d = E(\widetilde{y} - E(y))'(\widetilde{y} - E(y))$$

度量了这种预测的优劣。根据二次型求期望公式易得

$$d = q\sigma^2 + \beta_t' D^{-1}\beta_t$$

这里 D 的定义同前，即 $D^{-1} = X_t'(I - P_{X_q})X_t$, 令

$$\Gamma_q = \frac{d}{\sigma^2} = q + \frac{\beta_t' D^{-1}\beta_t}{\sigma^2}$$

则 Γ_q 是采用选模型时，在 n 个试验点预测优劣的一个总度量，它反映了选模型的好坏。又因

$$E(SS_{eq}/\sigma^2) = (n-q) + \frac{\beta_t' D^{-1}\beta_t}{\sigma^2}$$

于是

$$\Gamma_q = \frac{E(SS_{eq})}{\sigma^2} - (n-2q), \qquad (\star)$$

在 $(\star)$ 中用 SS_{eq} 代替 $E(SS_{eq})$，用 σ^2 咋全模型下的估计 $\widehat{\sigma}^2$ 代替 σ^2，便得到 C_p 准则，可见，C_p 统计量是作为 Γ_q 的一种估计产生的。

3. AIC 准则

极大似然原理是统计学中估计参数的一种重要的方法。Akaike 把此方法加以修正，提出一种较为一般的模型选择准则，成为 Akaike 信息准则 (Akaike information criterion, 简记为 AIC).

对于一般的统计模型，设 $Y_1, ..., Y_n$ 为一组样本，如果它们服从某个含 k 个参数的模型，对应的似然函数的最大值记为 $L_k(Y_1, ..., Y_n)$, 则 AIC 准则是选择使 AIC 统计量

$$AIC = lnL_k(Y_1, ..., Y_n) - k$$

达到最大的模型。下面我们把此准则则应用于回归模型自变量的选择。

在选模型中，假设误差 $e \sim N(0\sigma^2 I)$，则 β_q 和 σ^2 的似然函数为

$$L(\beta_q, \sigma^2|Y) = (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}||Y - X_q\beta_q||^2)$$

容易求得 β_q 和 σ^2 的极大似然估计分别为

$$\widetilde{\beta}_q = (X_q'X_q)^{-1}X_q'y, \widetilde{\sigma}_q^2 = \frac{SS_{eq}}{n} = \frac{y'(I - P_{Xq})y}{n}$$

得到对数似然函数的最大值

$$lnL(\widetilde{\beta}_q, \widetilde{\sigma}_q^2|y) = (ln(\frac{n}{2\pi})^{n/2} - \frac{n}{2}) - \frac{n}{2}ln(SS_{eq})$$

略去与 q 无关的项，得到

$$AIC = -\frac{n}{2}ln(SS_{eq}) - q$$

按 AIC 准则，我们选择使上述式子达到最大的模型，等价地，可取

$$AIC = nln(SS_{eq}) + 2q$$

于是最后 AIC 准则归结为: 选择使 AIC 达到最小的自变量子集。

例：Hald 水泥问题

考察含如下四种化学成分:

$x_1 : 3Cao \cdot Al_2O_3$ 的含量 (%)

$x_2 : 3Cao \cdot SiO_2$ 的含量 (%)

x_3 : $4Cao \cdot Al_2O_3 \cdot Fe_2O_3$ 的含量 (%)

x_4 : $2Cao \cdot SiO_2$ 的含量 (%)

的某种水泥，每一克所释放的热量 Y 与这四种成分含量之间的关系，共有 13 组数据，列在表 6.3.1 中.

表 6.3.1　Hald 水泥问题数据

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

此问题有四个自变量，共有 15 个不同的自变量子集。这 15 个自变量自子集的 LS 估计和 RMS_q, C_p 及 AIC 值列在表 6.3.2 中，从表 6.3.2 中可以看出，子集 $\{x_1, x_2, x_4\}$ 对应的 RMS_q 和 AIC 值都达到最小 (表中用黑色字体表示)，因此若没有别的附加考虑，在 RMS_q 准则或 AIC 准则下，最优子集回归为

$$y = 71.648 + 1.425x_1 + 0.416x_2 - 0.237x_4.$$

但子集 $\{x_1, x_2, x_4\}$ 对应的 C_p 值是所有值中最小的（表中用黑色字体表示），于是若按 C_p 准则选择自变量，最优子集回归为

$$y = 52.577 + 1.468x_1 + 0.662x_2.$$

可见，在不同的选择变量准则下，与之相应的“最优”自变量子集也不尽相同。注意到 $\{x_1, x_2\}$ 对应的 RMS_q 也比较小，所以综合起来看，$\{x_1, x_2\}$ 是最适合采用的子集。

表 6.3.2 Hald 水泥问题参数 LS 估计及 RMS_q, C_p 和 AIC 值

模型中的自变量	β_0	β_1	β_2	β_3	β_4	RMS_q	C_p	AIC
x_1	81.479	1.869				115.0264	202.55	95.9950
x_1x_2	52.577	1.468	0.662			5.7904	**2.68**	58.0033
x_2	57.424		0.789			82.3942	142.49	91.6535
x_2x_3	72.075		0.731	−1.008		41.5443	62.44	83.6205
$x_1x_2x_3$	48.194	1.696	0.657	0.250		5.3456	3.04	57.7252
x_1x_3	72.349	2.312		0.494		122.7073	198.10	97.7001
x_3	110.203			−1.256		176.3029	315.16	102.9394
x_3x_4	131.282			−1.200	−0.724	17.5738	22.37	72.4360
$x_1x_3x_4$	111.684	1.052		−0.410	−0.643	8.2017	3.50	63.2900
$x_1x_2x_3x_4$	62.405	1.551	0.510	0.102	−0.144	5.9829	5.00	59.8197
$x_2x_3x_4$	203.642		−0.923	−1.448	−1.557	5.6485	7.34	58.4417
x_2x_4	94.160		0.311		−0.457	86.8880	138.23	93.2127
$x_1x_2x_4$	71.648	1.452	0.416		−0.237	**5.3303**	3.02	**57.6879**
x_1x_4	103.097	1.440			−0.614	7.4762	5.50	61.3251
x_4	117.568				−0.738	80.8515	138.73	91.4078

第五章　回归诊断

前面讨论了线性回归模型的 LS 估计及检验问题，当进行上述讨论时，我们对模型做了一些假设，其中最主要的假设是 Gauss-Markov 假设，即假定模型误差 e_i 满足下列条件

(1) $Var(e_i) = \sigma^2$（等方差）;

(2) $Cov(e_i, e_j) = 0, i \neq j$（不相关性）;

(3) 有时我们还假设 e_i 服从正态分布，即 $e_i \sim N(0, \sigma^2)$

一个自然地问题，当有了一批数据之后，怎样考察我们的数据基本上满足这些假设，这是我们回归诊断中要研究的第一个问题。因为这些假设都是关于误差项的，所以很自然我们要从分析它们的估计量 (残差) 的角度来解决。

5.1 残差和残差图

考虑线性回归模型

$$y = X\beta + e, E(e) = 0, Cov(e) = \sigma^2 I$$

定义 $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta}$ 为残差向量，其中 $\widehat{Y} = X\widehat{\beta}$ 称为拟合值向量，如果用 $x_1', .., x_n'$ 表示 X 的 n 个行向量，则

$$\widehat{e_i} = y_i - x_i'\widehat{\beta}$$

为第 i 次试验或观测的残差。我们把 $\widehat{e_i}$ 看做误差 e_i 的一次观测值，如果模型正确的话，它应该具有 e_i 的一些性状。因此，我们可以通过这些 $\widehat{e_i}$ 以及基于它们的一些统计量来考察模型假设的合理性。对于 $\widehat{e_i}$ 的性质科归纳为以下定理。

定理 5.1

(1) $E(\widehat{e}) = 0, Cov(\widehat{e}) = \sigma^2(I - P_X)$;

(2) 若 $e \sim N(0, \sigma^2 I)$，则 $\widehat{e} \sim N(0, \sigma^2(I - P_X))$;

(3) $Cov(\widehat{y}, \widehat{e}) = 0$;

(4) $1'\widehat{e} = 0$

从定理 5.1 可以看出 $Var(\widehat{e}_i) = \sigma^2(I - p_{ii})$，这里 p_{ii} 为 P_X 的第 i 个主对角元。可见这个方差与因变量 Y 的度量单位以及 p_{ii} 有关，因袭直接比较残差 $\widehat{e}_i$ 是不适宜的。为此将其标准化，得到

$$\frac{\widehat{e}_i}{\sigma\sqrt{1-pii}}, i = 1, ..., n$$

但其中 σ 未知，用其估计 $\widehat{\sigma} = (||y - X\widehat{\beta}||^2/(n-p))^{1/2}$, 得到所谓学生化残差

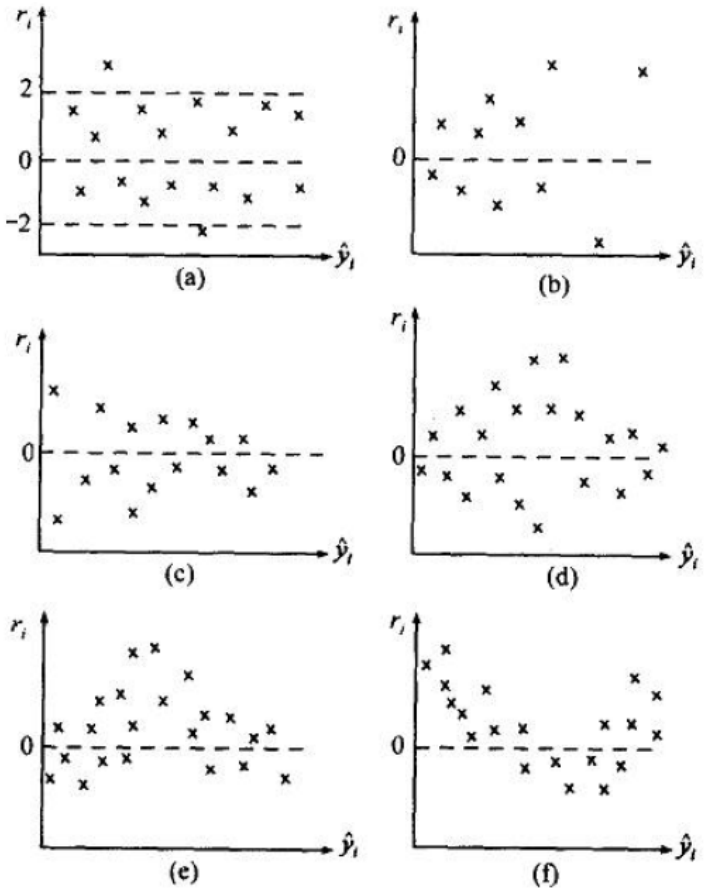
$$r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-p_{ii}}}, i = 1, ..., n$$

这里需要注意的是，即使在 $e \sim N(0, \sigma^2 I)$ 的条件下，r_i 的分布仍然比较复杂，且诸 r_i 彼此不独立。但是在应用上可以近似地认为 r_i 相互独立且服从 $N(0,1)$.

残差图就是以某种残差为纵坐标，以任何其他的量为横坐标的散点图。由于残差作为误差 e_i 的观察值或估计值应该与 e_i 相差不远，故根据残差图的大致形状是否与应有的性质相一致，就可以对假设 $e \sim N(0, \sigma^2 I)$ 的合理性提供一些有益的信息。

下面我们仅就以拟合值 $\widehat{y}_i$ 为横坐标，学生化残差为纵坐标的残差图为例讨论残差图的具体应有。

如果 $e \sim N(0, \sigma^2 I)$ 成立，r_i 近似且服从 $N(0,1)$，且近似相互独立，因此这些 r_i 可以近似看作来自总体 $N(0,1)$ 的一组简单随机样本。根据标准正态分布的性质，大约应有 95% 的 r_i 落在 [-2.2] 中，另外 $\widehat{Y}$ 与残差 $\widehat{e}$ 不相关，因而 $r' = (r_1, ..., r_n)$ 相关性也很小。所以在残差图中，点 $(\widehat{y}_i, r_i), i = 1, ..., n$ 大致应落在宽度为 4 的水平带 $|r_i| \leq 2$ 区域内，且不呈任何的趋势。如图 (a)，这是数据与假设乜有没有不一致的征兆，我们就可以认为 $e \sim N(0, \sigma^2 I)$ 基本上是合理的。图 (b) 显示了误差 e_i 随 $\widehat{y}_i$ 的增大有增加的趋势。图 (c) 所显示的情形正好相反，即误差 e_i 随 $\widehat{y}_i$ 的增大有减小的趋势，而图 (d) 显示对较大或较小的 $\widehat{y}_i$, 误差反而偏小，而对中等大小的 $\widehat{y}_i$，误差反而偏大。(e) 和 (f) 表明回归函数可能是非线性的，或误差 e_i 之间有一定的相关性或漏掉了一个或多个重要的回归自变量。究竟属于何种情况，还需作进一步的诊断。



几种处理方法:

$$\begin{cases} \text{增加自变量} \\ \text{增加自变量的二次项等} \\ \text{考虑加权最小二乘} \\ \text{BOX-COX 变换} \end{cases}$$

例 5.1: 一公司为了研究产品的营销策略，对产品的销售情况进行了调查。设 Y 表示某地区该产品的家庭人均购买量，X 表示家庭人均收入. 表 6.4.1 收集了 53 个家庭的数据，应用 LSE，求得

$$\widehat{Y} = -0.8313 + 0.003683X.$$

相应的残差 $\widehat{e}_i$ 和拟合值 $\widehat{y}_i$ 可作出 $\widehat{e}_i$ 为纵轴的残差图 (下左)。直观上容易看出，残差图从左向右逐渐散开呈漏斗状，这是误差方差不相等的一个征兆。考虑对因变量 Y 作变换，先试变换 $Z = Y^{1/2}$, 得到经验回归方程

$$\widehat{Z} = 0.5822 + 0.000953X$$

计算出新的残差 $\widetilde{e}_i$, 得到新的残差图 (下右)，从图中看出从已无任何明显趋势，这表明我们所用的变换是合适的。最后得到的经验回归方程为

$$\widehat{Y} = \widehat{Z}^2 = (0.5822 + 0.000953X)^2 = 0.339 + 0.001X + 0.00000091X^2$$

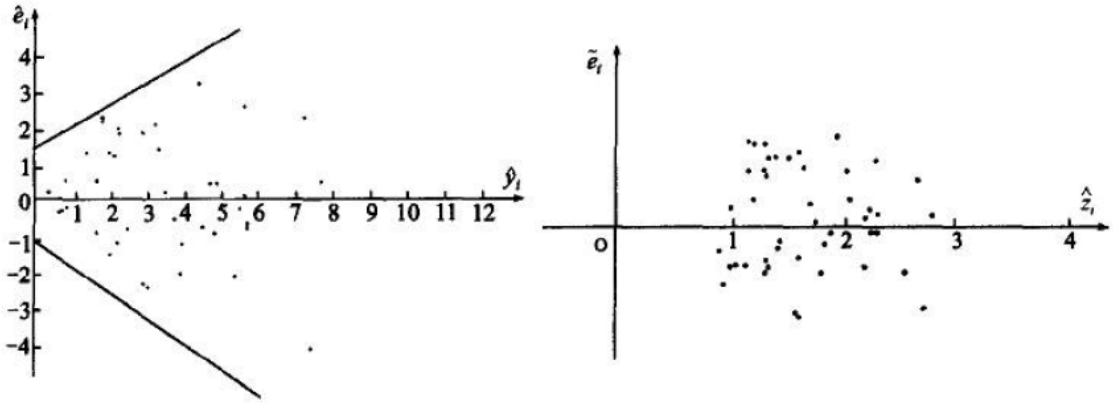


表 6.4.1　家庭人均收入数据

i	X(元)	Y(元)	$\hat{y}_i$	$\hat{e}_i$	$Z=\sqrt{Y}$	$\hat{z}_i'$	$\widetilde{e}_i$
1	679	0.790	1.669	−0.879	0.889	1.229	−0.340
2	292	0.440	0.244	0.196	0.663	0.860	−0.197
3	1012	0.560	2.896	−2.336	0.748	1.547	−0.798
4	493	0.790	0.984	−0.194	0.889	1.052	−0.163
5	582	2.700	1.312	1.388	1.643	1.137	0.506
6	1156	3.640	3.426	0.214	1.908	1.684	0.224
7	997	4.730	2.840	1.890	2.175	1.532	0.643
8	2189	9.500	7.230	2.270	3.082	2.668	0.414
9	1097	5.340	3.209	2.131	2.311	1.628	0.683
10	2078	6.850	6.822	0.028	2.617	2.562	0.055
11	1818	5.840	5.864	−0.024	2.417	2.315	0.102
12	1700	5.210	5.430	−0.220	2.283	2.202	0.080
13	747	3.250	1.920	1.330	1.803	1.294	0.509
14	2030	4.430	6.645	−2.215	2.105	2.517	−0.412
15	1643	3.160	5.220	−2.060	1.778	2.148	−0.370
16	414	0.550	0.693	−0.193	0.707	0.977	−0.270
17	354	0.170	0.472	−0.302	0.412	0.920	−0.507
18	1276	1.880	3.868	−1.988	1.371	1.798	−0.427
19	745	0.770	1.912	−1.142	0.877	1.292	−0.415
20	435	1.390	0.771	0.619	1.179	0.997	0.182
21	540	0.560	1.157	−0.597	0.748	1.097	−0.348
22	874	1.560	2.388	−0.828	1.249	1.415	−0.166
23	1543	5.280	4.851	0.429	2.298	2.052	0.245
24	1029	0.640	2.958	−2.318	0.800	1.563	−0.763
25	710	4.000	1.784	2.216	2.000	1.259	0.741
26	1434	0.310	4.450	−4.140	0.557	1.949	−1.392

续表

i	X(元)	Y(元)	$\hat{y}_i$	$\hat{e}_i$	$Z=\sqrt{Y}$	$\hat{z}_i'$	$\tilde{e}_i$
27	837	4.200	2.251	1.949	2.049	1.380	0.670
28	1255	2.630	3.791	−1.161	1.622	1.778	−0.156
29	1748	4.880	5.606	−0.726	2.209	2.248	−0.039
30	1381	3.480	4.255	−0.775	1.865	1.898	−0.033
31	1428	7.580	4.428	3.152	2.753	1.943	0.810
32	1777	4.990	5.713	−0.723	2.234	2.275	−0.042
33	370	0.590	0.531	0.059	0.768	0.935	−0.167
34	2316	8.190	7.698	0.492	2.862	2.789	0.073
35	1130	4.790	3.330	1.460	2.189	1.659	0.530
36	463	0.510	0.874	−0.364	0.714	1.023	−0.309
37	770	1.740	2.004	−0.264	1.319	1.316	0.003
38	724	4.100	1.835	2.265	2.025	1.272	0.753
39	808	3.940	2.144	1.796	1.985	1.352	0.633
40	790	0.960	2.078	−1.118	0.980	1.335	−0.355
41	783	3.290	2.052	1.238	1.814	1.328	0.486
42	406	0.440	0.664	−0.224	0.663	0.969	−0.306
43	1242	3.240	3.743	−0.503	1.800	1.766	0.034
44	658	2.140	1.592	0.548	1.463	1.209	0.254
45	1746	5.710	5.599	0.111	2.390	2.246	0.144
46	468	0.640	0.892	−0.252	0.800	1.028	−0.228
47	1114	1.900	3.271	−1.371	1.378	1.644	−0.265
48	413	0.510	0.690	−0.180	0.714	0.976	−0.262
49	1787	8.330	5.750	2.580	2.886	2.285	0.601
50	3560	14.940	12.280	2.660	3.865	3.974	−0.109
51	1495	5.110	4.675	0.435	2.261	2.007	0.254
52	2221	3.850	7.348	−3.498	1.962	2.699	−0.736
53	1526	3.930	4.789	−0.859	1.982	2.036	−0.054

5.2 影响分析

回归分析所要研究的另一个重要问题即探查对统计推断有较大影响的数据，这样的数据称为强影响点。

在回归分析中，因变量 Y 的取值 y_i 具有随机性，而自变量 $X_1,..,X_{p-1}$ 的取值 $x_i'=(x_{i1},...,x_{i,p-1}), i=1,...,n$ 也只是所有可能取到的值中的 n 组。我们虚妄每组数据 (x_i', y_i) 对未知参数的估计有一定的影响，但这种影响不能过大。这样，我们得到的经验回归方程具有一定的稳定性。

如果个别一两组的数据对估计有异常大的影响，我们剔除这些数据后，就会得到与原来差异很大的经验回归方程，这样我们就有理由怀疑所建立的经验回归方程是否真正描述了因变量与自变量之间的客观存在的相依关系。

因此，我么在做回归分析时，有必要考察每组数据对参数估计的影响大小，这部分内容在回归诊断中统称为影响分析。

用 $y_{(i)}, X_{(i)}, e_{(i)}$ 分别表示从 y,X,e 中剔除第 i 行后得到的向量或者矩阵。从线性回归模型剔除第 i 组数据后，剩余 n-1 组数据的线性回归模型记为

$$y_{(i)} = X_{(i)} + e_{(i)}, E(e_{(i)}) = 0, Cov(e_{(i)}) = \sigma^2 I_{n-1}$$

依据以上模型得到 β 的 LS 估计为 $\widehat{\beta}_{(i)}$, 则

$$\widehat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}y_{(i)}$$

很显然，向量 $\widehat{\beta} - \widehat{\beta}_{(i)}$ 反映了第 i 组数据对回归系数估计的影响大小，但它是一个向量，不便于定量的比较影响的大小，于是考虑它的某种数量化函数。Cook 统计量就是其中应用最为广泛的一种。

Cook 统计量定义为

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(i)})'X'X(\widehat{\beta} - \widehat{\beta}_{(i)})}{p\widehat{\sigma}^2}, i = 1, ..., n$$

这里 $\widehat{\sigma}^2 = ||Y - X\widehat{\beta}||^2/(n-p)$. 这样我们就可以用数量 D_i 来刻画第 i 组数据对回归系数估计的影响大小了。下面给出一个计算 D_i 的简便公式.

定理 5.2.1

$$D_i = \frac{1}{p}(\frac{p_{ii}}{1-p_{ii}})r_i^2, i = 1, ..., n$$

这里 p_{ii} 为矩阵 $P_X = (X'X)^{-1}X'$ 的第 i 个主对角元，r_i 为学生化残差。

证明设 A 为 $n \times n$ 矩阵，u, v 均为 $n \times 1$ 向量。利用恒等式

$$(A - uv')^{-1} = A^{-1} + \frac{A^{-1}uv'A^{-1}}{1 - u'A^{-1}v}$$

有

$$(X'_{(i)}X_{(i)})^{-1} = (X'X - x_i x'_i)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1 - p_{ii}}$$

这里 x'_i 为 X 的第 i 行，将上式两边右乘 $X'y$，并利用

$$X'Y = X'_{(i)}Y_{(i)} + y_i x_i$$

我们有

$$\widehat{\beta} = \widehat{\beta}_{(i)} + y_i(X'_{(i)}X_{(i)})^{-1}x_i - \frac{(X'X)^{-1}x_i(x'_i\widehat{\beta})}{1 - p_{ii}}$$

另外

$$(X'_{(i)}X_{(i)})^{-1}x_i = \frac{(X'X)^{-1}x_i}{1 - p_{ii}}$$

因此

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{\widehat{e}_i (X'X)^{-1} x_i}{1 - p_{ii}}$$

得证。

此定理告诉我们，在计算 Cook 统计量时，值需要从完全数据的线性回归模型计算出学生化残差 r_i，正交投影阵 P_X 的主对角元就可以了，不必对每一个不完全数据的线性回归模型进行计算。

显然 Cook 统计量 D_i 被分解为两部分，其中一部分为

$$P_i = \frac{p_{ii}}{1 - p_{ii}}$$

它是 p_{ii} 的单调增函数，因为 p_{ii} 度量了第 i 组数据 x_i 到试验中心 $\overline{x} = \sum x_i$ 的距离。因此，本质上 P_i 刻画了第 i 组数据 x_i 距离其他数据的远近。而另一部分为 r_i^2。直观上，如果一组数据距离试验中心很远，并且对应的学生化残差有很大，那么它必定是强影响点。但是，要给 Cook 统计量一个用以判定强影响点的临界值是很困难的。

例 5.2：智力测试数据

表 6.4.2 是教育学家测试的 21 个儿童的记录，其中 X 为儿童的年龄 (以月为单位)，Y 表示某种智力指标。通过这些数据，我们要建立智力随年龄变化的关系。

考虑直线回归 $y = \alpha + \beta X + e$,α, β 的 LS 估计为 $\widehat{\alpha} = 109.87$,$\widehat{\beta} = -1.13$ 于是经验回归方程为 $\widehat{Y} = 109.87 - 1.13X$ 表 6.4.3 给出了各组数据的有关诊断统计量。

表 6.4.2　智力测试数据

序号	x	y	序号	x	y	序号	x	y	序号	x	y
1	15	95	7	18	93	13	10	83	19	17	121
2	26	71	8	11	100	14	11	84	20	11	86
3	10	83	9	8	104	15	11	102	21	10	100
4	9	91	10	20	94	16	10	100			
5	15	102	11	7	113	17	12	105			
6	20	87	12	9	96	18	42	57			

从表 6.4.3 看出，$D_{18} = 0.6781$ 是所有 D_i 中最大的，而其它 D_i 值与 D_{18} 相比也十分小。因此，第 18 号数据时一个对回归估计影响最大的数据，对此数据我们就要格外注意。譬如，检查原始数据的抄录是否有误，如果有误，则需改正后重新计算。不然，需要从原始数据中剔除它。

表 6.4.3　智力测试数据的诊断统计量

序号	$\hat{e}_i$	r_i	p_{ii}	D_i	t_i
1	2.0310	0.1888	0.0479	0.0009	0.1839
2	−9.5721	−0.9444	0.1545	0.0815	0.9416
3	−15.6040	−0.8216	0.0628	0.0717	0.8143
4	−8.7309	−0.8216	0.0705	0.0256	0.8143
5	9.0310	0.8397	0.0479	0.0177	0.8329
6	−0.3341	−0.0315	0.0726	0.0000	0.0307
7	3.4120	0.3189	0.0580	0.0031	0.3112
8	2.5230	0.2357	0.0567	0.0017	0.2298
9	3.1420	0.2972	0.0799	0.0038	0.2899
10	6.6659	0.6280	0.0726	0.0154	0.6177
11	11.0151	1.0480	0.0908	0.0548	1.0508
12	−3.7309	−0.3511	0.0705	0.0047	0.3429
13	−15.6040	−1.4623	0.0628	0.0717	1.5108
14	−13.4770	−1.2588	0.0567	0.0476	1.2798
15	4.5230	0.4225	0.0567	0.0054	0.4131
16	1.3960	0.1308	0.0628	0.0006	0.1274
17	8.6500	0.8060	0.0521	0.0179	0.7982
18	−5.5403	−0.8515	0.6516	**0.6781**	0.8450
19	**30.2850**	2.8234	0.0531	0.2233	3.6071
20	−11.4770	−1.0720	0.0567	0.0345	1.0765
21	1.3960	0.1308	0.0628	0.0006	0.1274

需要指出的是，对已经确认的强影响数据如何处理，这需要具体问题具体分析:

(1) 仔细核查数据，如果强影响数据是由于试验条件失控或记录失误或其它一些过失所致，那么这些数据应该剔除。

(2) 否则，勇敢考虑收集更多的数据或采用一些稳健估计方法以缩小强影响数据对估计的影响，从而获得较稳定的经验回归过程。

5.3 异常点的检验

在回归分析中，一组数据 (x_i', y_i) 如果它的残差 ($\widehat{e_i}$ 或 r_i) 较其它组数据的残差大得多，则称此数据为异常点。

为了讨论探查异常点的一种检验。我们把正态线性回归模型写为如下的分量形式:

$$y_i = x_i'\beta + e_i, e_i \sim N(0, \sigma^2), i = 1, ..., n, \qquad (\star)$$

这里 $e_i (i = 1, ..., n)$ 相互独立。如果第 j 组数据 (x_j', y_j) 是一个异常点，那么它的残差之所以很大时因为它的均值 $E(y_i)$ 发生了非随机漂移 η, 从而 $E(y_j) = x_j'\beta + \eta$. 这样就

产生了一个新模型

$$\begin{cases} y_i = x_i'\beta + e_i, i \neq j, \\ y_j = x_j'\beta + \eta + e_j, e_i \sim N(0, \sigma^2). \end{cases}$$

记 $d_j = (0, ..., 0, 1, 0, ..., 0)'$, 这是一个 n 维向量，它的第 j 个元素为 1，其余元素为零。写成矩阵形式

$$Y = X\beta + d_j\eta + e, e \sim N(0, \sigma^2 I)$$

以上模型称为均值漂移线性回归模型，要判定 (x_j', y_j) 是不是异常点，等价于检验线性假设 $H : \eta = 0$。

为了导出所要的检验统计量，我们先给出漂移模型中参数 β, η 的 LS 估计。分别记为 $\beta^\star, \eta^\star$. 显然，假设 $\eta = 0$ 成立时，β 的 LS 估计就是 $\widehat{\beta} = (x'x)^{-1}X'y$.

定理 5.3.1 对均值漂移线性回归模型，β, η 的 LS 估计分比为

$$\beta^\star = \widehat{\beta}_{(j)}, \eta^\star = \frac{1}{1 - p_{jj}}\widehat{e}_j$$

这里 $\widehat{\beta}_{(j)}$ 为非均值漂移线性回归模型剔除第 j 组数据后得到的 β 的 LS 估计。p_{jj} 为 P_X 的第 j 个主对角元，$\widehat{e}_j$ 为从模型 $(\star)$ 导出的第 j 个残差。

证明显然，$d_j'y = j_j, d_j'd_j = 1$，记 $X = (x_1, ..., x_n)'$, 则 $X'd_j = x_j$。于是根据定义

$$\begin{pmatrix} \beta^\star \\ \eta^\star \end{pmatrix} = [\begin{pmatrix} X' \\ d_j' \end{pmatrix} \begin{pmatrix} X & d_j \end{pmatrix}]^{-1} \begin{pmatrix} X' \\ d_j' \end{pmatrix} y = \begin{pmatrix} X'X & x_j \\ x_j' & 1 \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ y_j \end{pmatrix}$$

根据分块矩阵的求逆公式，以及 $p_{jj} = x_j'(X'X)^{-1}x_j$，有

$$\begin{pmatrix} \beta^\star \\ \eta^\star \end{pmatrix} = \begin{pmatrix} (X'X)^{-1} + \frac{1}{1-p_{jj}}(X'X)^{-1}x_jx_j'(X'X)^{-1} & -\frac{1}{1-p_{jj}}(X'X)^{-1}x_j \\ -\frac{1}{1-p_{jj}}x_j'(X'X)^{-1} & \frac{1}{1-p_{jj}} \end{pmatrix}$$

$$\begin{pmatrix} X'y \\ y_j \end{pmatrix} = \begin{pmatrix} \widehat{\beta} + \frac{1}{1-p_{jj}}(X'X)^{-1}x_jx_j'\widehat{\beta} - \frac{1}{1-p_{jj}}(X'X)^{-1}x_jy_j \\ -\frac{1}{1-p_{jj}}x_j'\widehat{\beta} + \frac{1}{1-p_{jj}}y_j \end{pmatrix} = \begin{pmatrix} \widehat{\beta} - \frac{1}{1-p_{jj}}(X'X)^{-1}x_j\widehat{e}_j \\ \frac{1}{1-p_{jj}}\widehat{e}_j \end{pmatrix}$$

这个定理告诉我们一个很重要的事实：如果因变量的第 j 个观测值发生均值漂移，那么在相应的均值漂移回归模型中，回归系数的 LS 估计等于原来模型中剔除第 j 组数据后，所获得的 LS 估计。

约束条件 $\eta = 0$，即为约简模型 $(\star)$, 于是

$$SS_{He} = \text{模型}(\star)\text{残差平方和} = Y'Y - \widehat{\beta}X'Y$$

而无约束模型的残差平方和为

$$SS_e = Y'Y - \beta^{\star'}X'Y - \eta^\star d_j'Y$$

$$SS_{He} - SS_e = (\beta^\star - \widehat{\beta})'X'Y + \eta^\star d_j'Y = -\frac{1}{1-p_{jj}}\widehat{e}_j x_j'\widehat{\beta} + \frac{1}{1-p_{jj}}\widehat{e}_j y_j = \frac{\widehat{e}_j^2}{1-p_{jj}}$$

这里 $\widehat{e}_j = y_j - x_j'\widehat{\beta}$ 为原模型下第 j 组数据的残差。

利用 $\beta^\star, \eta^\star$ 的具体表达式将 SS_e 作进一步简化:

$$SS_e = Y'Y - \widehat{\beta}'X'Y + \frac{\widehat{e}_j y_j}{1-p_{jj}} = (n-p)\widehat{\sigma}^2 - \frac{\widehat{e}_j^2}{1-p_{jj}}$$

其中 $\widehat{\sigma}^2 = ||Y - X\widehat{\beta}||^2/(n-p)$. 则所求的检验统计量为

$$F = \frac{SS_{He} - SS_e}{SS_e/(n-p-1)} = \frac{(n-p-1)\frac{\widehat{e}_j^2}{1-p_{jj}}}{(n-p)\widehat{\sigma}^2 - \frac{\widehat{e}_j^2}{1-p_{jj}}} = \frac{(n-p-1)r_j^2}{n-p-r_j^2}$$

这里 $r_j = \frac{\widehat{e}_j}{\widehat{\sigma}\sqrt{1-p_{jj}}}$ 为学生化残差。

定理 5.3.2 对于均值漂移线性回归模型，如果假设 $H: \eta = 0$ 成立，则

$$F_j = \frac{(n-p-1)r_j^2}{n-p-r_j^2} \sim F_{1,n-p-1}$$

据此，我们就得到如下检验：对给定的 $\alpha(0 < \alpha < 1)$, 若

$$F_j = \frac{(n-p-1)r_j^2}{n-p-r_j^2} > F_{1,n-p-1}(\alpha)$$

则判定第 j 组数据 (x_j', y_j) 为异常点。

根据 t 分布和 F 分布的关系，我们也可以用 t 检验法完成上面的检验。若定义

$$t_j = F_j^{1/2} = (\frac{(n-p-1)r_j^2}{n-p-r_j^2})^{1/2}$$

则对给定的 α, 当

$$|t_j| > t_{n-p-1}(\alpha/2)$$

时，我们拒绝假设 $H: \eta = 0$，即判定第 j 组数据 (x_j', y_j) 为异常点。

第六章　Box-Cox 变换

对观测得到的试验数据集 $(x'_j, y_j), i = 1, .., n$，若经过回归诊断后得知，它们不满足 Gauss-Markov 条件，我们就要对数据采取" 治疗" 措施，实践证明，数据变换是处理有问题数据的一种好方法。本章介绍最著名的 Box-Cox 变换。它的主要特点是引入一个参数，通过数据本身估计该参数，从而确定应采取的数据变换形式，实践证明，Box-Cox 变换对许多实际数据都是行之有效的。

Box-Cox 变换是对回归因变量的如下变换：

$$Y^{(\lambda} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ lnY, \lambda = 0 \end{cases}$$

这里 λ 是一个待定变换参数。Box-Cox 变换时一族变换，它包括了许多常见的变换，诸如对数变换 $(\lambda = 0)$，倒数变换 $(\lambda = -1)$ 和平方根变换 $(\lambda = 1/2)$ 等等。

对因变量的 n 个观测值 $y_1, ..., y_n$，应用上述变换，得到变换后的向量

$$y^{(\lambda)} = (y_1^{(\lambda)}, ..., y_n^{(\lambda)})'$$

我们要确定变换参数 λ, 使得 $y^{(\lambda)}$ 满足

$$y^{(\lambda)} = X\beta + e, e \sim N(0, \sigma^2 I)$$

即通过因变量的变换，使得变换过的向量 $y^{(\lambda)}$ 与回归自变量之间具有线性相依关系，误差也服从正态分布，误差各分量等方差且相互独立。因此，Box-Cox 变换时通过参数 λ 的选择，达到对原来数据的" 综合治理"，使其满足一个正态线性回归模型的所有假设条件。

我们用极大似然方法来确定 λ，因为 $y^{(\lambda)} \sim N(X\beta, \sigma^2 I)$，所以对固定的 λ, β, σ^2 的似然函数为

$$L(\beta, \sigma^2) = \frac{1}{(\sqrt{1\pi}\sigma)^n} exp - \frac{1}{2\sigma^2}(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)J$$

这里 J 为变换的 Jacobi 行列式

$$J = \prod |\frac{dy_i^{(\lambda)}}{dy_i}| = \prod y_i^{\lambda-1}$$

因此，当 λ 固定时，J 是不依赖于参数 β 和 σ^2 的常数因子。$L(\beta, \sigma^2)$ 的其余部分关于 β 和 σ^2 求导数，令其等于零，可以求得 β 和 σ^2 得极大似然估计

$$\widehat{\beta}(\lambda) = (X'X)^{-1}X'y^{(\lambda)}$$

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n} y^{(\lambda)'}(I - X(X'X)^{-1}X')y^{(\lambda)} = \frac{1}{n} SS_e(\lambda, y^{(\lambda)}),$$

这里残差平方和为

$$SS_e(\lambda, y^{(\lambda)}) = y^{(\lambda)'}(I - X(X'X)^{-1}X')y^{(\lambda)}$$

对应的似然函数最大值为

$$L_{max}(\lambda) = L(\widehat{\beta}(\lambda), \widehat{\sigma}^2(\lambda)) = (2\pi e)^{-n/2} \cdot J \cdot (\frac{SS_e(\lambda, y^{(\lambda)})}{n})^{-n/2}, \qquad (\star)$$

这是 λ 的一元函数。通过求它的最大值来确定 λ, 因 lnx 是 x 的单调函数，我们的问题可以化为求 $lnL_{max}(\lambda)$ 的最大值，对 $(\star)$ 求对数，略去与 λ 无关的常数项，得

$$lnL_max(\lambda) = -\frac{n}{2} lnSS_e(\lambda, y^{(\lambda)}) + lnJ$$

$$\begin{aligned} lnL_{max}(\lambda) &= -\frac{n}{2} lnSS_e(\lambda, y^{(\lambda)}) + lnJ \\ &= -\frac{n}{2} ln(\frac{y^{(\lambda)'}}{J^{1/n}}(I - X(X'X)^{-1}X'\frac{y^{(\lambda)}}{J^{1/n}}) \\ &= -\frac{n}{2} lnSS_e(\lambda, z^{(\lambda)}) \end{aligned}$$

其中

$$SS_e(\lambda, z^{(\lambda)}) = z^{(\lambda)'}(I - X(X'X)^{-1}X')z^{(\lambda)}$$

$$z^{(\lambda)} = (z_1^{(\lambda)}, ..., z_n^{(\lambda)})' = \frac{y^{(\lambda)}}{J^{1/n}}$$

$$z_i^{(\lambda)} = \begin{cases} \frac{y_i^{(\lambda)}}{(\prod y_i)^{(\lambda-1)/n}}, \lambda \neq 0, \\ (lny_i)(\prod y_i)^{1/n}, \lambda = 0. \end{cases}$$

因此求 $lnL_{max}(\lambda)$ 的最大值，我们只需求残差平方和 $SS_e(\lambda, z^{(\lambda)})$ 的最小值。虽然我们很难找出使 $SS_e(\lambda, z^{(\lambda)})$ 达到最小值的 λ 的解析式，但对一系列给定的 λ 值，通过最普通的求 LS 估计的回归程序，我们很容易计算出对应的 $SS_e(\lambda, z^{(\lambda)})$，画出 $SS_e(\lambda, z^{(\lambda)})$ 关于 λ 的曲线，从图中可以近似地找出使 $SS_e(\lambda, z^{(\lambda)})$ 达到最小值的 $\widehat{\lambda}$.

现在我们把 Box-Cox 变换的具体步骤归纳如下：

(1) 对给定的 λ 值，计算出 $z_i^{(\lambda)}$;

(2) 计算残差平方和 $SS_e(\lambda, z^{(\lambda)})$;

(3) 对一系列的 λ 值，重复上述步骤，得到相应的残差平方和 $SS_e(\lambda, z^{(\lambda)})$ 的一串值，以 λ 为横轴，做出相应的曲线。用直观的方法，找出使 $SS_e(\lambda, z^{(\lambda)})$ 达到最小值点的 $\widehat{\lambda}$

(4) 求出 $\widehat{\beta}(\widehat{\lambda})$

例 6.1：在例 5.1 中，我们对因变量 Y 做了平方根变换，这相当于选用变换参数 $\lambda = 0.5$. 应用本节的方法，我们可以证实做这样的变换时合适的。表 6.5.1 给出了 12 个不同的 λ 值对应的残差平方和 $SS_e(\lambda, z^\lambda$，简单比较可以看出当 $\lambda = 0.5$ 时，残差平方和达到最小，因此我们可以近似地认为 0.5 就是变换参数 λ 的最优选择。

表 6.5.1

λ	-2	-1	-0.5	0	0.125	0.25
RSS	34101.04	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	1275.56

第七章　均方误差及多重共线性

LS 估计有许多良好的性质，但在实际应用中，特别是大型线性回归问题中，LS 估计有时表现不理想。例如，有时某些回归系数的估计的绝对值异常大，有时回归系数的估计值的符号与问题的实际意义相违背等。

研究结果表明，产生这些问题的原因之一是回归自变量之间存在着近似线性关系，称为复共线性 (multicolinearity).

本节我们研究复共线性对 LS 估计的影响以及复共线性的诊断和严重程度的度量问题。

首先引进评价一个估计优劣的标准-均方误差（mean squared errors, 简记为 MSE）.

设 θ 为 $p\times 1$ 的未知参数向量，$\widehat{\theta}$ 为 θ 的一个估计。定义 $\widehat{\theta}$ 的均方误差为

$$MSE(\widehat{\theta})=E||\widehat{\theta}-\theta||^2=E(\widehat{\theta}-\theta)'(\widehat{\theta}-\theta).$$

它度量了估计 $\widehat{\theta}$ 与未知参数向量 θ 的平均偏离的大小，一个好的估计应该有较小的均方误差。

定理 7.1

$$MSE(\widehat{\theta})=trCov(\widehat{\theta})+||E\widehat{\theta}-\theta||^2$$

证明

$$\begin{aligned}MSE(\widehat{\theta})&=E(\widehat{\theta}-\theta)'(\widehat{\theta}-\theta)\\&=E[(\widehat{\theta}-E\widehat{\theta})+(E\widehat{\theta}-\theta)]'[(\widehat{\theta}-E\widehat{\theta})+(E\widehat{\theta}-\theta)]\\&=E(\widehat{\theta}-E\widehat{\theta})'(\widehat{\theta}-E\widehat{\theta}))+(E\widehat{\theta}-\theta)'(E\widehat{\theta}-\theta)\\&=\Delta_1+\Delta_2\end{aligned}$$

因为对任意两个矩阵 $A_{m\times n}$ 和 $B_{n\times m}$ 有 trAB=trBA, 于是上式第一项

$$\begin{aligned}\Delta_1&=Etr(\widehat{\theta}-E\widehat{\theta})'(\widehat{\theta}-E\widehat{\theta})\\&=Etr(\widehat{\theta}-E\widehat{\theta})(\widehat{\theta}-E\widehat{\theta})'\\&=trE(\widehat{\theta}-E\widehat{\theta}))(\widehat{\theta}-E\widehat{\theta}))'\\&=trCov(\widehat{\theta})\end{aligned}$$

而第二项 $\Delta_2=(E\widehat{\theta}-\theta)'(E\widehat{\theta}-\theta)=||E\widehat{\theta}-\theta||^2$. 定理证毕.

因此 $\widehat{\theta}$ 的均方误差可以分解为两项之和，其中一项为 $\widehat{\theta}$ 的各分量的方差之和，另一项为 $\widehat{\theta}$ 的各分量的偏差的平方和。因此一个估计的均方误差就是由它的各分量的方差和偏差所决定。一个好的估计应该有较小的方差和偏差。

现在我们用均方误差这个标准来评价 LS 估计。考虑线性回归模型

$$Y = \alpha_0 1 + X\beta + e, E(e) = 0, Cov(e) = \sigma^2 I$$

这里假定 $n \times (p-1)$ 的设计阵 X 已经中心化和标准化，且 rk(X)=p-1. 由于设计阵是中心化的，于是常数项 α_0 和回归系数 β 的 LS 估计能够分离开来，它们分别为

$$\widehat{\alpha}_0 = \overline{y} = \frac{1}{n}\sum y_i$$

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

把 α_0 与 β 的 LS 估计这样分离开来，对研究回归系数的 LS 估计的改进带来了很大的方便，下面我们只讨论回归系数 β 的 LS 估计的改进。

因为 $\widehat{\beta}$ 是 β 的无偏估计，于是在 $MSE(\widehat{\beta})$ 的表达式中，$\Delta_2 = 0$。又因为 $Cov(\widehat{\beta}) = \sigma^2(X'X)^{-1}$, 于是

$$MSE(\widehat{\beta}) = \Delta_1 = \sigma^2 tr(X'X)^{-1}$$

记 $\lambda_1 \geq ... \geq \lambda_{p-1} > 0$ 为 X'X 的特征值，因为 X'X 可逆，所以 $(X'X)^{-1}$ 的特征值为 $\lambda_1^{-1}, ..., \lambda_{p-1}^{-1}$，故上式变为

$$MSE(\widehat{\beta}) = \sigma^2 \sum_{i=1}^{p-1} \frac{1}{\lambda_i}$$

因此，如果 X'X 至少有一个特征值非常小，即非常接近于零，那么 $MSE(\widehat{\beta})$ 就会很大。从均方误差的标准来看，这时的 LS 估计 $\widehat{\beta}$ 就不是一个好的估计。

另一方面

$$MSE(\widehat{\beta}) = E(\widehat{\beta} - \beta)'(\widehat{\beta} - \beta) = E(\widehat{\beta}'\widehat{\beta} - 2\beta'\widehat{\beta} + \beta'\beta) = E||\widehat{\beta}||^2 - \beta'\beta,$$

于是

$$E||\widehat{\beta}||^2 = ||\beta||^2 + MSE(\widehat{\beta}) = ||\beta||^2 + \sigma^2 \sum_{i=1}^{p-1} \frac{1}{\lambda_i}$$

这就是说，当 $X'X$ 至少有一个特征值很小时，LS 估计 $\widehat{\beta}$ 的长度平均说来要比真正的未知向量 β 的长度长的多。这就导致了 $\widehat{\beta}$ 的某些分量的绝对值较大。

因此，当 X'X 至少有一个特征值很小时，LS 估计 $\widehat{\beta}$ 就不再是一个好的估计。

下面我们进一步分析，X'X 至少有一个特征值很小对设计阵 X 本身或回归自变量关系上意味着什么?

记 $X=(x_{(1)},...,x_{(p-1)})$, 即 $x_{(i)}$ 为设计阵 X 的第 i 列，设 λ 为 $X'X$ 的一个特征值，φ 为其对应的特征向量，其长度为 1，即 $\varphi'\varphi=1$. 若 $\lambda\approx 0$, 则

$$X'X\varphi=\lambda\varphi\approx 0$$

用 φ' 左乘上式，得

$$\varphi'X'X\varphi=\lambda\varphi'\varphi=\lambda\approx 0$$

于是，我们有

$$X\varphi\approx 0$$

若记 $\varphi=(c_1,..,c_{p-1})'$, 上式即为

$$c_1x_{(1)}+...+c_{p-1}x_{(p-1)}\approx 0, \qquad (\star)$$

这表明设计阵 X 的列向量 $x_{(1)},...,x_{(p-1)}$ 之间有近似的线性关系。如果用 $X_1,...,X_{p-1}$ 分别表示 p-1 个回归自变量，那么 $(\star)$ 表明，从现有的 n 组数据看，回归自变量之间有近似线性关系

$$c_1X_1+...+c_{p-1}X_{p-1}\approx 0$$

回归设计阵的列向量之间的关系或等价地回归自变量之间的关系，称为复共线关系。相应地，称设计阵 X 或线性回归模型有复共线性，有时也称设计阵 X 是病态的（ill-conditioned）.

显然，"X'X 的特征值很小" 等价于设计阵 X 之间存在复共线性关系，并且 X'X 有几个特征值很小，设计阵 X 就存在几个复共线关系。因此，复共线性是 LS 估计变坏的原因。方阵 X'X 的条件数定义为

$$k=\frac{\lambda_1}{\lambda_{p-1}}$$

即 X'X 的最大特征值和最小特征值之比。条件数刻画了 X'X 的特征值的散布程度，可以用来判断复共线性是否存在以及复共线性严重程度。从实际应用的角度，一般若 $k<100$，则认为复共线性的程度很小；若 $100\le k\le 1000$，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。

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。虽然这样产生的复共线性是非本质的。另一种产生复共线性的重要原因是，自变量之间客观上就有近似的线性关系。比如，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 都看作自变量，那么因为家庭高收入的住房也相应的宽敞一些，在自变量 x_1 和 x_2 之间就有复共线性。

第八章　有偏估计

当设计阵存在复共线关系时，LS 估计的性质不够理想，有时甚至很坏。为此，统计，统计学家做了种种努力，试图改进最小二乘估计。一方面是从模型或数据角度去考虑（变量选择和回归诊断），另一方面就是寻求一些新的估计。Stein 于 1955 年证明了，当维数大于 2 时，正态均值向量的 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。由此为开端，近 30 年来，人们提出了许多新的估计，其中主要有岭估计，主成分估计等。这些估计有一个共同的特点，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。

8.1 岭估计

对于线性回归模型，回归系数 β 的岭估计定义为

$$\widehat{\beta}(k) = (X'X + kI)^{-1}X'Y$$

这里 $k > 0$ 是可选择参数，称为岭参数或偏参数。如果 k 取实验数据 y 无关的常数，则 $\widehat{\beta}(k)$ 为线性估计。否则，$\widehat{\beta}(k)$ 就是非线性估计。当 k 取不同的值，我们得到不同的估计，因此岭估计 $\widehat{\beta}(k)$ 是一个估计类。特别，取 k=0,$\widehat{\beta}(0) = (X'X)^{-1}X'Y$. 于是严格地讲，LS 估计是岭估计类中的一个估计。与 LS 估计相比，岭估计是把 X'X 换成了 X'X+kI 得到的。直观上看这样作的理由也是明显的。因为当 X 呈病态时，X'X 的特征值至少有一个非常接近于零，而 X'X+kI 的特征值 $\lambda_1 + k, .., \lambda_{p-1} + k$ 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 $MSE(\widehat{\beta}(k)) < MSE(\widehat{\beta})$.

为了证明关于岭估计优良性的一个基本定理，我们引进线性回归模型的典则形式。设 $\lambda_1, ..., \lambda_{p-1}$ 为 X'X 的特征值，$\phi_1, ..., \phi_{p-1}$ 为对应的标准正交化特征向量。记 $\Phi = (\phi_1, ..., \phi_{p-1})$, 则 Φ 为 $(p-1) \times (p-1)$ 标准正交阵。再记

$$\Lambda = diag(\lambda_1, ..., \lambda_{p-1})$$

于是 $X'X = \Phi\Lambda\Phi'$, 则线性回归模型可改写为

$$Y = \alpha_0 1 + Z\alpha + e, E(e) = 0, Cov(e) = \sigma^2 I, \qquad (\star)$$

这里 $Z = X\Phi, \alpha = \Phi'\beta$。我们称 $(\star)$ 为线性回归模型的典则形式，α 称为典则回归系数。因为 X 是中心化的，于是 Z 也是中心化的。对典则形式 $(\star)$，α_0 和 α 的 LS 估计分别为

$$\widehat{\alpha}_0 = \overline{Y}, \widehat{\alpha} = (Z'Z)^{-1}Z'Y$$

注意到 $Z'Z = \Phi' X'X\Phi = \Lambda$, 因而

$$\widehat{\alpha} = \Lambda^{-1} Z'Y$$

$$Cov(\widehat{\alpha}) = \sigma^2 \Lambda^{-1} Z'Z\Lambda^{-1} = \sigma^2 \Lambda^{-1}$$

按定义典则回归系数 α 的岭估计为

$$\widehat{\alpha}(k) = (Z'Z + kI)^{-1} Z'Y = (\Lambda + kI)^{-1} Z'Y$$

容易证明

$$\widehat{\alpha} = \Phi' \widehat{\beta}$$

$$\widehat{\alpha}(k) = \Phi' \widehat{\beta}(k)$$

则，典则回归参数 α 的 LS 估计与原来回归参数 β 的 LS 估计之间差一个标准正交阵，因而有

$$MSE(\widehat{\alpha}) = MSE(\widehat{\beta})$$

同样的

$$MSE(\widehat{\alpha}(k)) = MSE(\widehat{\beta}(k))$$

现在我们证明岭估计的优良性的基本定理。

定理 8.1 存在 $k > 0$，使得

$$MSE(\widehat{\beta}(k)) < MSE(\widehat{\beta})$$

即存在 $k > 0$，使得在均方误差意义下，岭估计优于 LS 估计。

证明由于 $MSE(\widehat{\alpha}) = MSE(\widehat{\beta})$ 和 $MSE(\widehat{\alpha}(k)) = MSE(\widehat{\beta}(k))$，所以只需证明存在 $k > 0$，使得

$$MSE(\widehat{\alpha}(k)) < MSE(\widehat{\alpha})$$

由于设计阵 Z 是中心化的，于是 $1'Z = 0$, 所以

$$\begin{aligned} E(\widehat{\alpha}(k)) &= (\Lambda + kI)^{-1} Z'(\alpha_0 1 + Z\alpha) \\ &= (\Lambda + kI)^{-1} Z'Z\alpha \\ &= (\Lambda + kI)^{-1} Z'\Lambda\alpha \end{aligned}$$

另外

$$Cov(\widehat{\alpha}(k)) = \sigma^2 (\Lambda + kI)^{-1} Z'Z(\Lambda + kI)^{-1}$$

$$= (\Lambda + kI)^{-1}\Lambda(\Lambda + kI)^{-1}$$

所以

$$\begin{aligned} & MSE(\widehat{\beta}(k))trCov(\widehat{\beta}(k)) + ||E(\widehat{\beta}(k)) - \alpha||^2 \\ & = \sigma^2 \sum_{i=1}^{p-1} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p-1} \frac{\alpha_i^2}{(\lambda_i + k)^2} \\ & = f_1(k) + f_2(k) = f(k) \end{aligned}$$

对 k 求导，得

$$f_1'(k) = -2\sigma^2 \sum_{i=1}^{p-1} \frac{\lambda_i}{(\lambda_i + k)^3}$$

$$f_2'(k) = 2k \sum_{i=1}^{p-1} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}$$

因为 $f_1'(0) < 0, f_2'(0) = 0$，所以 $f'(0) < 0$. 显然 $f_1'(k)$ 和 $f_2'(k)$ 在 $k \geq 0$ 时都连续，所以 $f'(k)$ 在 $k \geq 0$ 时也连续。因而，当 $k > 0$ 时充分小时 $f'(k) < 0$, 这就是说，$f(k) = MSE(\widehat{\alpha}(k))$ 在 $k > 0$ 充分小时，是 k 的单调函数，因而存在 $k^\star > 0$，当 $k \in (0, k^\star)$ 时，有 $f(k) < f(0)$. 但 $f(0) = MSE(\widehat{\alpha}(k))$。

注 1 这个定理为岭估计的实际应用奠基了理论基础，具有重要的意义。从该定理可以看到 $MSE(\widehat{\beta}(k)) < MSE(\widehat{\beta})$ 成立的 k 依赖于未知参数 β 和 σ^2。因此，对固定的 k, 岭估计 $\widehat{\beta}(k)$ 不是在整个参数空间上一致优于 LS 估计。

注 $2\widehat{\beta}(k) = A_k\widehat{\beta}$，这里 $A_k = (X'X + kI)^{-1}X'X$. 这表明岭估计是 LS 估计的一个线性变换。

注 3 对任意的 $k > 0$ 和 $||\widehat{\beta}|| \neq 0$，总有

$$||\widehat{\beta}(k)|| = ||\widehat{\alpha}(k)|| = ||(\Lambda + kI)^{-1}\Lambda\widehat{\beta}|| \leq ||\widehat{\alpha}|| = ||\widehat{\beta}||$$

这表明，岭估计 $\widehat{\beta}(k)$ 的长度总比 LS 估计 $\widehat{\beta}$ 的长度小。因此 $\widehat{\beta}(k)$ 是对 $\widehat{\beta}$ 向原点一种压缩，所以通常也称之为一种压缩估计。

当设计阵 X 呈病态时，平均来说 LS 估计 $\widehat{\beta}$ 偏长，对它做适当的压缩式应该的，这个结果从一个侧面说明了岭估计的合理性。

在实际应用中，岭参数的选择是一个很重要的问题，定理 8.1 仅说明 $\widehat{\beta}(k)$ 优于 $\widehat{\beta}$ 的 k 的存在性，并没有给出具体的算法，我们自然希望找到使 $MSE(\widehat{\beta}(k))$ 达到最小的 k。显然这个最优值 k 应该在方程

$$f'(k) = f_1'(k) + f_2'(k) = 2\sum_{i=1}^{p-1} \frac{\lambda_i(k\alpha_i^2 - \sigma^2)}{(\lambda_i + k)^3}$$

的根中去找。显然，k 的最优值依赖于未知参数 β, σ^2，从而不可能通过解方程 $f'(k) = 0$ 去获得。

1.Hoerl-Kennard 公式

岭估计是由 Hoerl 和 Kennard 于 1970 年提出的。他们所用的选择 k 的公式是

$$\widehat{k} = \frac{\widehat{\sigma}^2}{max_i \widehat{\alpha}_i^2}$$

这个方法是基于如下的考虑。由于 $f'(k) = f_1'(k) + f_2'(k) = 2\sum_{i=1}^{p-1} \frac{\lambda_i(k\alpha_i^2 - \sigma^2)}{(\lambda_i + k)^3}$，如果 $k\alpha_i^2 - \sigma^2 < 0$ 对 $i = 1, ..., p-1$ 都成立，则 $f'(k) < 0$. 于是取

$$k^\star = \frac{\sigma^2}{max_i \alpha_i^2}$$

当 $0 < k < k^\star$ 时，$f'(k)$ 总是小于 0，因而 f(k) 总是 k 的单调函数，故有 $f(k^\star) < f(0)$, 即 $MSE(\widehat{\beta}(k)) = MSE(\widehat{\beta})$.

2. 岭迹法

岭估计 $\widehat{\beta}(k) = (X'X + kI)^{-1}X'Y$ 是随 k 值改变而变化。若记 $\widehat{\beta}_i(k)$ 为 $\widehat{\beta}(k)$ 的第 i 个分量，它是 k 的一元函数，当 k 在 $[0, +\inf)$ 上变化时，$\widehat{\beta}_i(k)$ 的图形称为岭迹。选择 k 的岭迹法是：将 $\widehat{\beta}_1(k), ..., \widehat{\beta}_{p-1}(k)$ 的岭迹画在同一个图上。根据岭迹的变化趋势选择 k 值，使得各个回归系数的岭估计大体上稳定，并且各个回归系数的岭估计值得符号比较合理。另外，LS 估计是使残差平方和达到最小的估计。k 越大，岭估计与 LS 估计偏离越大。对应的残差平方和也随着 k 的增加而增加。因此，当用岭迹法选择 k 值时，还应考虑使得残差平方和不要上升太多。

例 8.1：外贸数据分析

我们所考虑的因变量 Y 为进口总额，自变量 X_1 为国内总产值，X_2 为存储量，X_3 为总消费量。为了建立 Y 对自变量 X_1, X_2, X_3 之间的依赖关系，收集了 11 组数据，列在表 6.7.1 中。

将原始数据中心化和标准化，计算得到

$$X'X = \begin{pmatrix} 1 & 0.026 & 0.997 \\ 0.026 & 1 & 0.036 \\ 0.997 & 0.036 & 1 \end{pmatrix}$$

再计算它的三个特征值，分别为 $\lambda_1 = 1.999, \lambda_2 = 0.998, \lambda_3 = 0.003$。于是 X'X 的条件数 $\lambda_1/\lambda_3 = 666.333$，可见设计阵存在中等程度的复共线性。$\lambda_3$ 对应的特征向量为 $\phi = (-0.7070, -0.0070, 0.7072)$. 由上一节的讨论知，三个自变量之间存在复共线关系

$$-0.7070X_1 - 0.0070X_2 + 0.7072X_3 \approx 0.$$

注意到，自变量 X_2 的系数绝对值相对非常小，可视为 0，而 X_1, X_3 的系数又近似相等，因此自变量之间的复共线关系可近似地写为 $X_3 = X_1$。注意这里的 X_1, X_3 都是经过中心化和标准化的变量，还原为原来的变量，近似复共线关系为

$$\frac{X_1 - \overline{x}_1}{s_1} = \frac{X_3 - \overline{x}_3}{s_3}.$$

从表 6.7.1 可以算出

$$\overline{x}_1 = 194.59, s_1 = (\sum_{i=1}^{1} 1(x_{i1} - \overline{x}_1)^2)^{1/2} = 94.87.$$

$$\overline{x}_3 = 139.74, s_3 = (\sum_{i=1}^{1} 1(x_{i3} - \overline{x}_3)^2)^{1/2} = 65.25.$$

代入上式得

$$X_3 = 5.905 + 0.688X_1.$$

这就是总消费量和国内总产值之间一个线性依赖关系，因此 X 是中心化和标准化的，于是 X'X 是相关阵，其中 0.997 正是 X_1 和 X_3 的相关系数。可见，X_1 和 X_3 有如此大的相关系数，和我们找出它们之间的复共线关系这一事实是吻合的。既然自变量之间存在中等程度的复相关性，我们就采用岭估计来估计回归系数。对于中心化和标准化的变量，计算出的岭迹列在表 6.7.2，对应的岭迹图画在图 6.7.1. 表 6.7.2 的最后一列是岭估计对应的残差平方和。我们看到，随着 k 的增加，岭估计的残差平方和也随之增加，所以残差平方和是岭参数 k 的单调增函数，这是很自然的，因为 LS 估计是使残差平方和达到最小的估计。随着 k 的增加，岭估计与 LS 估计的偏离就愈大，因此它的残差平方和自然也就愈大。从岭迹图上可以看出，岭迹 $\widehat{\beta}_1$ 随着 k 的增加，很快增加，大约在 k=0.01 处从负值变为正值。而 $\widehat{\beta}_2$ 相对比较稳定，但 $\widehat{\beta}_3$ 随着 k 的增加，骤然减少，大约在 k=0.04 以后就稳定下来。总体来看，我们可以取 k=0.04，对于的岭估计为

$$\widehat{\beta}_1(0.04) = 0.42, \widehat{\beta}_2(0.04) = 0.213, \widehat{\beta}_3(0.04) = 0.525$$

各变量的平均值为

$$\overline{x}_1 = 194.59, \overline{x}_2 = 3.3, \overline{x}_3 = 139.74, \overline{y} = 21.89$$

相应的

$$s_1 = 94.87, s_2 = 5.22, s_3 = 65.26, s_y = 14.37$$

代入经验回归方程，化简后得到如下岭回归方程

$$\widehat{Y} = -8.5537 + 0.0635X_1 + 0.5859X_2 + 0.1156X_3$$

表 6.7.1　外贸数据

序号	国内总产值 (x_1)	存储量 (x_2)	总消费量 (x_3)	进口总额 (y)
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239.0	0.7	167.6	26.3

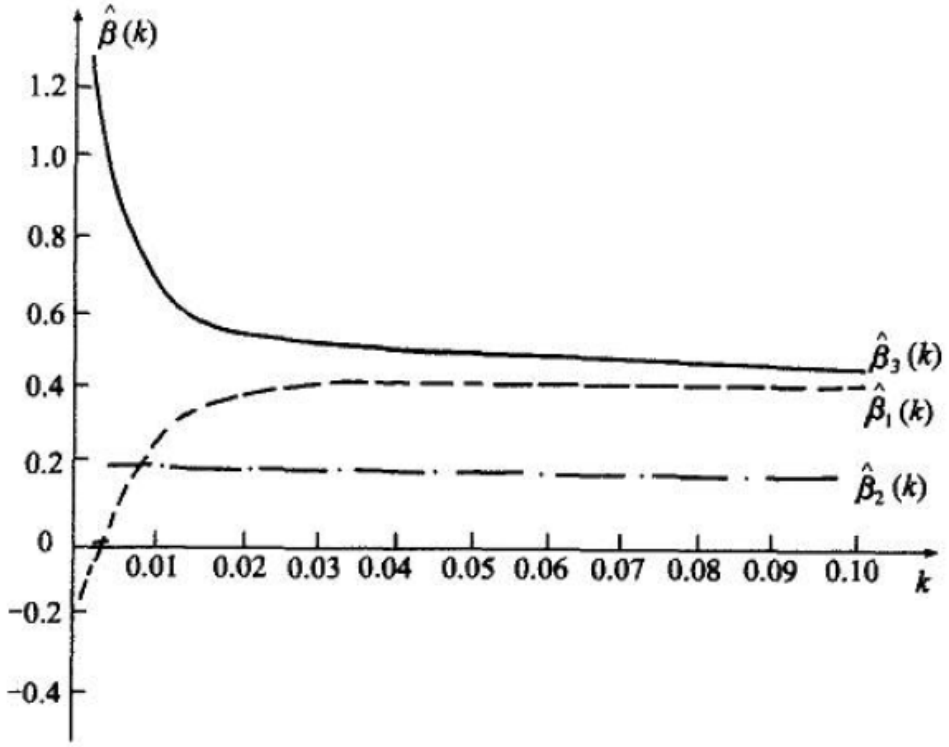


图 6.7.1　外贸数据回归的岭迹

表 6.7.2　外贸数据的岭回归

k	$\hat{\beta}_1(k)$	$\hat{\beta}_2(k)$	$\hat{\beta}_3(k)$	RSS
0.000	-0.339	0.213	1.303	1.673
0.001	-0.117	0.215	1.080	1.728
0.002	0.010	0.216	0.952	1.809
0.003	0.092	0.217	0.870	1.881
0.004	0.150	0.217	0.811	1.941
0.005	0.193	0.217	0.768	1.990
0.006	0.225	0.217	0.735	2.031
0.007	0.251	0.217	0.709	2.066
0.008	0.272	0.217	0.687	2.095
0.009	0.290	0.217	0.669	2.120
0.010	0.304	0.217	0.654	2.142
0.020	0.379	0.216	0.575	2.276
0.030	0.406	0.214	0.543	2.352
0.040	0.420	0.213	0.525	2.416
0.050	0.427	0.211	0.513	2.480
0.060	0.432	0.209	0.504	2.548
0.070	0.434	0.207	0.497	2.623
0.080	0.436	0.206	0.491	2.705
0.090	0.436	0.204	0.486	2.794
0.100	0.436	0.202	0.481	2.890
0.200	0.426	0.186	0.450	4.236
0.300	0.411	0.173	0.427	6.155
0.400	0.396	0.161	0.408	8.489
0.500	0.381	0.151	0.391	11.117
0.600	0.367	0.142	0.376	13.947
0.700	0.354	0.135	0.361	16.911
0.800	0.342	0.128	0.348	19.957
0.900	0.330	0.121	0.336	23.047
1.000	0.319	0.115	0.325	26.149

岭估计的一种推广形式，称为广义岭估计。对于线性回归模型，回归系数 β 的广义岭估计定义为

$$\widehat{\beta}(K) = (X'X + \Phi K \Phi')^{-1} X'Y$$

这里 Φ 的定义同上文，即 Φ 为标准正交阵，使得 $\Phi' X'X \Phi = diag(\lambda_1, ..., \lambda_{p-1}), K = diag(k_1, ..., k_{p-1})$.

8.2 主成分估计

在研究岭估计的优良性时，给出了一般线性回归模型的典则形式。新的设计阵 $Z=(z_{(1)},...,z_{(p-1)}=\ X\phi_1,...,X\phi_{p-1})$, 即

$$z_{(1)}=X\phi_1,...,z_{(p-1)}=X\phi_{p-1}$$

于是 Z 的第 i 列 $z_{(i)}$ 是原来 p-1 个回归自变量的线性组合，其组合系数为 $X'X$ 的第 i 个特征值对应的特征向量 ϕ_i。因此，Z 的 p-1 个列就对应于 p-1 个以原来变量的特殊线性组合 (即以 X'X 的特征向量为组合系数) 构成的新变量。在多元统计学中，称这些新变量为主成分。排在第一排的新变量对应于 X'X 的最大特征值，称为第一主成分，排在第二列的就成为第二主成分，依次类推。因为 X 是中心化的，即 $1'X=0$, 于是 $1'Z=1'X\Phi=0$. 所以 Z 也是中心化的。因而 Z 的各列元的平均值为

$$\overline{z}_j=\frac{1}{n}\sum_{i=1}^{n}z_{ij}=0,j=1,...,p-1$$

另外

$$z'_{(i)}z_{(i)}=\phi'_iX'X\phi_i=\lambda_i$$

因此

$$\sum_{i=1}^{n}(z_{ij}-\overline{z}_j)^2=z'_{(i)}z_{(i)}=\lambda_i,j=1,...,p-1$$

于是 X'X 的第 i 个特征值 λ_i 就度量了第 i 个主成分取值变动大小。当设计阵 X 存在复共线关系时，有一些 X'X 的特征值很小，不妨设 $\lambda_{r+1},...,\lambda_{p-1}\approx 0$. 这时后面的 p-r-1 个主成分取值变动就很小，再结合 $\overline{z}_j=0$. 因此，在用主成分作为新的回归自变量时，这后面的 p-r-1 个主成分对应变量的影响就可以忽略掉，故可将他们从回归模型中剔除。用最小二乘法做剩下的 r 个主成分的回归，然后再变回到原来的自变量就得到了主成分回归。

现将上述思想具体化。记 $\Lambda=diag(\lambda_1,...,\lambda_{p-1})$, 对 Λ,α,Z,Φ 做分块

$$\Lambda=\begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix},\alpha=\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix},Z=(Z_1;Z_2),\Phi=(\Phi_1;\Phi_2)$$

其中 Λ_1 为 $r\times r$ 矩阵，α_1 为 $r\times 1$ 向量，Z_1 为 $n\times r$ 矩阵，Φ_1 为 $(p-1)\times r$ 矩阵代入 $Y=\alpha_0 1+Z\alpha+e,E(e)=0,Cov(e)=\sigma^2I$ 中，并剔除 $Z_2\alpha_2$ 项得到回归模型

$$Y=\alpha_0 1+Z_1\alpha_1+e,E(e)=0,Cov(e)=\sigma^2I$$

这个新的回归模型就是在剔除了后面 p-r-1 个对应变量影响较小的主成分后得到的。因此，事实上我们是利用主成分进行了一次回归自变量的选择。对子模型应用最小二乘法，得到 α_0, α_1 的 LS 估计:

$$\widehat{\alpha}_0 = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\widehat{\alpha}_1 = (Z_1'Z_1)^{-1}Z_1'Y = \Lambda_1^{-1}Z_1'Y$$

这相当于用 $\widetilde{\alpha}_2 = 0$ 去估计 α_2。利用关系 $\beta = \Phi\alpha$，可以获得原来参数 β 的估计

$$\widehat{\beta} = \Phi \begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 \end{pmatrix} \begin{pmatrix} \widehat{\alpha}_1 \\ 0 \end{pmatrix} = \Phi_1\Lambda_1^{-1}Z_1'Y = \Phi_1\Lambda_1^{-1}\Phi_1'X'Y$$

这就是 β 的主成分估计。

显然，

$$E(\widetilde{\beta}) = \begin{pmatrix} \Phi_1 & \Phi_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} = \Phi_1\alpha_1$$

但

$$\beta = \Phi\alpha = \Phi_1\alpha_1 + \Phi_2\alpha_2$$

可见，一般说来 $E(\widetilde{\beta}) \neq \beta$，于是主成分估计也是有偏估计。我们应该用均方误差作为度量其优劣的标准。下面的定理证明了，在一定的条件下主成分估计比 LS 估计有较小均方误差。

定理 8.2 当设计阵存在复共线关系时，适当选择保留的主成分个数可致主成分估计比 LS 估计有较小的均方误差，即

$$MSE(\widetilde{\beta}) < MSE(\widehat{\beta}$$

证明假设 $X'X$ 的后面 p-r-1 个特征值 $\lambda_{r+1}, ..., \lambda_{p-1}$ 很接近于 0，则有

$$MSE(\widetilde{\beta}) = MSE\begin{pmatrix} \widehat{\alpha}_1 \\ 0 \end{pmatrix} = trCov\begin{pmatrix} \widehat{\alpha}_1 \\ 0 \end{pmatrix} + \|E\begin{pmatrix} \widehat{\alpha}_1 \\ 0 \end{pmatrix} - \alpha\|^2 = \sigma^2 tr(\Lambda_1^{-1}) + ||\alpha_2||^2$$

因为

$$MSE(\widehat{\beta}) = \sigma^2 tr(\Lambda^{-1})$$

所以

$$MSE(\widetilde{\beta}) = MSE(\widehat{\beta}) + (||\alpha_2||^2 - \sigma^2 tr(\Lambda_2^{-1}))$$

于是

$$MSE(\widetilde{\beta}) < MSE(\widehat{\beta})$$

当且仅当

$$||\alpha_2||^2 < \sigma^2 tr(\Lambda_2^{-1}) = \sigma^2 \sum_{i=r+1}^{p-1} \frac{1}{\lambda_i}, \qquad (\star)$$

因为假设 $X'X$ 的后面 p-r-1 个特征值接近于 0，于是上式右端很大，故等式 $(\star)$ 成立。

注因为 $\alpha_2 = \Phi_2'\beta$，于是变回到原来参数，$(\star)$ 可变形为

$$(\frac{\beta}{\sigma})'\Phi_2\Phi_2'(\frac{\beta}{\sigma}) < tr\Lambda_2^{-1}, \qquad (\star\star)$$

这就是说，仅当 β, σ^2 满足 $(\star\star)$ 时，主成分估计才能比 LS 估计有较小的均方误差。所以有如下的结论

(1) 对固定的参数 β, σ^2，当 $X'X$ 的后面 p-r-1 个特征值很小时，主成分估计比 LS 估计有较小的均方误差。

(2) 对给定的 $X'X$, 也就是固定的 Λ_2，对相对比较小的 $\frac{\beta}{\sigma}$, 主成分估计比 LS 估计有较小的 junf 误差。

在主成分估计应用中，有一个重要的问题就是如何选择保留主成分的个数。通常有两种方法，其一是保留对应的特征值相对比较大的那些主成分；其二是选择 r, 使得 $\sum_{i=1}^{r} \lambda_i$ 与全部 p-1 个特征值之和 $\sum_{i=1}^{p-1} \lambda_i$ 的比值达到预先给定的值，譬如 75% 或 80% 等。

主成分作为原来变量的线性组合，是一种" 人造变量", 一般并不具有任何实际含义，特别当回归自变量具有不同度量单位时，更是如此。比如 X_1, X_2 分别表示该农作物生长期内平均气温和降雨量，它们的度量单位分别是摄氏度和毫米。

第九章　多元线性模型

前面讨论的线性模型都只包含一个因变量。例如，研究产品的某一项性能指标 Y_1 与原材料含量，加工条件 $X_1,...,X_{p-1}$ 之间的关系，导致了一个因变量 Y_1 对多个自变量 $X_1,...,X_{p-1}$ 的线性模型。但是，实际应用上，人们也常常会遇到含多个因变量的问题。例如，如果我们同时对产品的多个指标 $Y_1,...,Y_q$ 感兴趣，这时就有 q 个因变量，这很自然地导致了对多个因变量与多个自变量的线性模型的研究。

一般，假设研究 q 个因变量 $Y_1,...,Y_q$ 和 p-1 个自变量 $X_1,...,X_{p-1}$ 之间的关系，若 Y_j 与 $X_1,...,X_{p-1}$ 有线性关系：

$$Y_j=\beta_{0j}+\beta_{1j}X_1+...+\beta_{(p-1)j}X_{p-1}+\varepsilon_j,j=1,...,q$$

为了估计系数 β_{ij}，对 $Y_1,...,Y_q$ 和 $X_1,...,X_{p-1}$ 做 n 此观测，得到数据

$$y_{i1},...,y_{iq};x_{i1},...,x_{i(p-1)};i=1,...,n$$

它们满足

$$y_{ij}=\beta_{0j}+\beta_{1j}x_{i1}+...+\beta_{(p-1)j}x_{ip-1}+\varepsilon_{ij},i=1,...,n,j=1,...,q$$

若引进矩阵记号

$$Y_{n\times q}=\begin{pmatrix}y_{11}&y_{12}&...&y_{1q}\\y_{21}&y_{22}&...&y_{2q}\\...&...&...&...\\y_{n1}&y_{n2}&...&y_{nq}\end{pmatrix}=(y_1,...,y_q)$$

$$X_{n\times p}=\begin{pmatrix}1&x_{11}&...&x_{1p-1}\\1&x_{21}&...&x_{2p-1}\\...&...&...&...\\1&x_{n1}&...&x_{np-1}\end{pmatrix}$$

$$B_{p\times q}=\begin{pmatrix}\beta_{01}&\beta_{02}&...&\beta_{0q}\\\beta_{11}&\beta_{12}&...&\beta_{1q}\\...&...&...&...\\\beta_{p-11}&\beta_{p-12}&...&\beta_{p-1q}\end{pmatrix}=(\beta_1,...,\beta_q)$$

$$\varepsilon_{n\times q}=\begin{pmatrix}\varepsilon_{11}&\varepsilon_{12}&...&\varepsilon_{1q}\\\varepsilon_{21}&\varepsilon_{22}&...&\varepsilon_{2q}\\...&...&...&...\\\varepsilon_{n1}&\varepsilon_{n2}&...&\varepsilon_{nq}\end{pmatrix}=(\varepsilon_1,...,\varepsilon_q)$$

这里随机误差矩阵 ε 的不同行对应于不同次观测，我们假定它们不相关，均值为零，用公差协方差矩阵为 $\Sigma > 0$。B 为未知参数矩阵，每个列对应于一个因变量。Y 为因变量随机预测阵，它们不同行对应于不同次观测，每个列对应于一个因变量。假设 rk(X)=p, 于是

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon\text{的行向量互不相关，均值为 } 0\text{，协方差矩阵为}\Sigma \end{cases}$$

为多元线性模型。

9.1 kronecker 积与向量化运算

先来介绍一下 Kronecker 乘积与向量化运算。

定义：设 $A = (a_{ij})$ 和 $B = (b_{ij})$ 分别为 $m \times n$，$p \times q$ 的矩阵，定义矩阵 $C = (a_{ij}B)$。这是一个 $mp \times nq$ 的矩阵，称为 A 和 B 的 Kronecker 乘积，记为 $C = A \otimes B$, 即

$$A \otimes B = \begin{pmatrix} a_{11} \otimes B & a_{12} \otimes B & ... & a_{1n} \otimes B \\ a_{21} \otimes B & a_{22} \otimes B & ... & a_{2n} \otimes B \\ ... & ... & ... & ... \\ a_{m1} \otimes B & a_{m1} \otimes B & ... & a_{mn} \otimes B \end{pmatrix}$$

这种乘积具有以下性质：

(1) $0 \otimes A = A \otimes 0 = 0$,

(2) $(A_1 + A_2) \otimes B = (A_1 \otimes B) + (A_2 \otimes B)$,$A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$,

(3) $(\alpha A) \otimes (\beta B) = \alpha\beta(A \otimes B)$,

(4) $(A \otimes B)' = A' \otimes B'$,

(5) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

定理设 A,B 分别为 $n \times n$,$m \times m$ 的方阵，$\lambda_1, ..., \lambda_n$ 和 $\mu_1, ..., \mu_m$ 分别为 A,B 的特征值，则

(1) $\lambda_i\mu_j, i = 1, ..., n; j = 1, ..., m$ 为 $A \otimes B$ 的特征值，且 $|A \otimes B| = |A|^m|B|^n$;

(2) $tr(A \otimes B) = tr(A)tr(B)$;

(3) $rk(A \otimes B) = rk(A)rk(B)$;

(4) 若 $A \geq 0, B \geq 0$, 则 $A \otimes B \geq 0$.

证明 (1) 记 A,B 的 Jordan 标准形分别为

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & & \\ ... & ... & & \\ 0 & 0 & ... & \lambda_n \end{pmatrix}, \Delta = \begin{pmatrix} \mu_1 & & & \\ 0 & \mu_2 & & \\ ... & ... & & \\ 0 & 0 & ... & \mu_n \end{pmatrix}$$

依 Jordan 分解, 存在可逆矩阵 P 和 Q, 使得 $A = P\Lambda P^{-1}, B = Q\Delta Q^{-1}$, 利用 Kronecker 乘积的性质, 得

$$A \otimes B = P\Lambda P^{-1} \otimes Q\Delta Q^{-1} = (P \otimes Q)(\Lambda \otimes \Delta)(P \otimes Q)^{-1}$$

因此

$$\begin{aligned} |A \otimes B| &= |\Lambda \otimes \Delta| \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m} \lambda_i \mu_j \\ &= (\prod_{i=1}^{n} \lambda_i)^m (\prod_{j=1}^{m} \mu_j)^n \\ &= |A|^m |B|^n \end{aligned}$$

定义: 设 $A_{m\times n} = (a_1, a_2, ..., a_n)$, 定义 $mn \times 1$ 的向量

$$Vec(A) = \begin{pmatrix} a_1 \\ a_2 \\ .... \\ a_n \end{pmatrix}$$

(1) $Vec(A + B) = Vec(A) + Vec(B)$;

(2) $Vec(\alpha A) = \alpha Vec(A)$, 这里 α 为数;

(3) $tr(AB) = (Vec(A'))'Vec(B)$;

(4) $tr(A) = tr(AI) = tr(IA) = (Vec(I_n))'Vec(A)$;

(5) 设 a 和 b 分别为 $n \times 1$,$m \times 1$ 向量, 则 $Vec(ab') = b \otimes a$;

(6) $Vec(ABC) = (C' \otimes A)Vec(B)$.

我们只证明 (6), 设 $C_{m\times n}=(c_{ij})=(c_1,...,c_n), B=(b_1,...,b_m)$, 依定义

$$
\begin{aligned}
(C'\otimes A)Vec(B) &= \begin{pmatrix} c_{11}\otimes A & c_{21}\otimes A & ... & c_{m1}\otimes A \\ c_{12}\otimes A & c_{22}\otimes A & ... & c_{m2}\otimes A \\ ... & ... & ... & ... \\ c_{1n}\otimes A & c_{2n}\otimes A & ... & c_{mn}\otimes A \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_m \end{pmatrix} \\
&= \begin{pmatrix} A\Sigma c_{j1}b_j \\ A\Sigma c_{j2}b_j \\ ... \\ A\Sigma c_{jn}b_j \end{pmatrix} \\
&= \begin{pmatrix} ABc_1 \\ ABc_2 \\ ... \\ ABc_n \end{pmatrix} = Vec(ABC)
\end{aligned}
$$

9.2 多元线性模型

现在讨论未知参数 B 和 Σ 的估计问题，应用 $Vec(ABC)=(C'\otimes A)Vec(B)$，有

$$Vec(Y)=(I\otimes X)Vec(B)+Vec(\varepsilon)$$

因为

$$Cov(y_i,y_j)=\sigma_{ij}I_n, i,j=1,...,q$$

这里 $\Sigma=(\sigma_{ij})_{q\times q}$，再由 $Cov(Vec(\varepsilon))=\Sigma\otimes I_n$, 多元线性模型化为如下一元线性模型

$$
\begin{cases}
Vec(Y)=(I\otimes X)Vec(B)+Vec(\varepsilon), \\
Cov(Vec(\varepsilon))=\Sigma\otimes I_n, \\
E(Vec(\varepsilon))=0
\end{cases}
$$

应用一元线性模型的结果和 Kronecker 乘积的性质，$\beta\triangleq Vec(B)$ 的 BLU 估计为

$$
\begin{aligned}
\beta^\star &= Vec(B^\star) \\
&= [(I\otimes X)'(\Sigma\otimes I_n)^{-1}(I\otimes X)]^{-1}(I\otimes X)(\Sigma\otimes I_n)^{-1}Vec(Y) \\
&= (\Sigma^{-1}\otimes X'X)^{-1}(\Sigma^{-1}\otimes X')Vec(Y) \\
&= (I\otimes (X'X)^{-1}X')Vec(Y)
\end{aligned}
$$

$$= Vec((X'X)^{-1}X'Y)$$

于是 B 的 BLU 估计为

$$B^\star = (X'X)^{-1}X'Y$$

若记 $B^\star = (\beta_1^\star, ..., \beta_q^\star)$, 则

$$\beta_i^\star = (X'X)^{-1}X'y_i, i = 1, ..., q.$$

此即从一元线性模型

$$y_i = X\beta_i + \varepsilon_i, i = 1, ..., q$$

导出的 LS 估计。这个结果表明：q 个因变量的多元线性模型的参数矩阵 B 的 BLU 估计可以从 q 个一元线性模型得到。

容易证明

$$Cov(Vec(B^\star)) = \Sigma \otimes (X'X)^{-1}$$

于是

$$Cov(\beta_i^\star, \beta_j^\star) = \sigma_{ij}(X'X)^{-1}, i, j = 1, ..., q.$$

现在讨论 Σ 的估计，定义

$$Y^\star = X\beta^\star = X(X'X)^{-1}X'Y \overset{\Delta}{P}_X Y$$

$$\widehat{\varepsilon} = Y - \varepsilon^\star = (I - P_X)Y$$

应用事实：$E(x'Ay) = tr[ACov(y, x)] + [E(x)]'A[E(y)]$, 有

$$\begin{aligned} E(y_i'(I - P_X)y_i] &= \sigma_{ij}tr(I - P_X) + \beta_i'X'(I - P_X)X\beta_j \\ &= \sigma_{ij}tr(I - P_X) \\ &= (n - q)\sigma_{ij} \end{aligned}$$

于是

$$E(\widehat{\varepsilon}'\widehat{\varepsilon}) = E[Y'(I - P_X)Y] = (n - p)\Sigma$$

最后，我们得到 Σ 的一个无偏估计

$$\Sigma^\star = \frac{1}{n - p}Y'(I - P_X)Y$$

同样的处理手法也可应用于更一般的多元线性模型：

$$\begin{cases} Y = X_1BX_2 + \varepsilon, \\ \varepsilon\text{的行向量互不相关，均值为 0，协方差阵为}\Sigma, \end{cases}$$

这里 Y 仍为 $n \times q$ 随机观测阵，X_1, X_2 分别为 $n \times p$ 和 $k \times q$ 已知矩阵，B 为 $n \times q$ 的未知参数阵，关于 ε 的假设同前面的模型一样。这类模型的不少例子来自生物生长问题，故得生长曲线模型 (growth-curve model) 之名。

例 9.1：生物学家欲研究白鼠的某个特征随时间变化情况，随机选用 n 只小白鼠做试验。在时刻 $t_1, ..., t_p$ 对每只小白鼠观测该特征的值。设第 i 只小白鼠的 p 此观测值为 $y_{i1}, ..., y_{ip}, i = 1, ..., n$。假定不同白鼠的观测值是不相关的，而同一只白鼠的 p 次观测却是相关的，且协方差阵为 $\Sigma(> 0)$。从理论分析认为，这些观测值与观测时间 t 的关系为 k-1 阶多项式：

$$Y = f(t) = \beta_0 + \beta_1 t + ... + \beta_{k-1} t^{k-1}$$

这就是所谓理论生长曲线。生物学家的目的是估计 $\beta_0, ..., \beta_{k-1}$，以得到经验生长曲线。若以 ε_{ij} 记 y_{ij} 所含的误差，则对观测数据 y_{ij}，我们有模型

$$\begin{pmatrix} y_{11} & y_{12} & ... & y_{1p} \\ y_{21} & y_{22} & ... & y_{2p} \\ ... & ... & ... & ... \\ y_{n1} & y_{n2} & ... & y_{np} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix} \begin{pmatrix} \beta_0 & \beta_1 & ... & \beta_{k-1} \end{pmatrix} \begin{pmatrix} 1 & 1 & ... & 1 \\ t_1 & t_2 & ... & t_p \\ ... & ... & ... & ... \\ t_1^{k-1} & t_2^{k-1} & ... & t_p^{k-1} \end{pmatrix} + (\varepsilon_{ij})$$

应用矩阵向量化方法，有

$$\begin{cases} Vec(Y) = (X_2' \otimes X_1) Vec(B) + Vec(\varepsilon), \\ E(Vec(\varepsilon)) = 0, \\ Cov(Vec(\varepsilon)) = \Sigma \otimes I. \end{cases}$$

得到 $\beta^\star = Vec(B^\star)$ 的 GLS 解为

$$\beta^\star = Vec((X_1'X_1)^{-1} X_1' Y \Sigma^{-1} X_2' (X_2 \Sigma^{-1} X_2')^{-1})$$

等价地

$$B^\star = (X_1'X_1)^{-1} X_1' Y \Sigma^{-1} X_2' (X_2 \Sigma^{-1} X_2')^{-1}$$

另外，不难证明

$$S = \frac{1}{n - rk(X_1)} Y'(I - X_1(X_1'X_1)^{-1} x_1') Y$$

是 Σ 的一个无偏估计。当 $n - rk(X_1) > q$ 时，它以概率为 1 的正定，将 S 代入得到

$$\widetilde{B} = (X_1'X_1)^{-1} X_1' Y S^{-1} X_2' (X_2 \Sigma^{-1} X_2')^{-1}$$

称为 B 的两步 GLS 解。

两步估计：

考虑一般线性模型

$$Y = X\beta + e, E(e) = 0, Cov(e) = \Sigma(\theta)$$

这里 Y 为 $n \times 1$ 观测向量，X 为 $n \times p$ 设计阵，β 为 $p \times 1$ 未知参数向量，e 为 $n \times 1$ 随机误差。$\theta = (\theta_1, ..., \theta_m)$ 也是未知参数向量。设 $\Sigma(\theta) > 0$ 对一切 θ 成立。记

$$\widehat{\beta}(\theta) = (X'\Sigma^{-1}(\theta)X)^{-1}X'\Sigma^{-1}(\theta)Y$$

当 θ 已知时，$\widehat{\beta}(\theta)$ 就是它的 GLS 估计，也是 BLU 估计。如果 θ 是未知的，设 $\widehat{\theta}$ 是它的一个估计，则 $\widehat{\beta}(\widehat{\theta})$ 为 β 的两步估计。那么在什么条件相爱，$\widehat{\beta}(\widehat{\theta})$ 仍然是 β 的无偏估计。

我们先引进一个概念。

设 W 为一空间，若对任一 $y \in W$, 统计量 S(y) 满足 S(-y)=S(y), 则称 S(y) 对 $y \in W$ 为偶函数。若对任一 $y \in W$, 统计量 S(y) 满足 S(-y)=S(y), 则称 S(y) 对 $y \in W$ 为奇函数。另外，若对一切 y 和 β，统计量 S(y) 满足

$$S(y - X\beta) = S(y)$$

则称 S(y) 是交换不变的。

引理 1 设 u 为一随机向量，其分布关于原点是对称的，记为

$$u \overset{d}{=} -u$$

，又 g(u) 是 u 的奇函数，则 g(u) 的分布关于原点也是对称的.

证明因为 $u \overset{d}{=} -u$，于是 $g(u) \overset{d}{=} -g(u)$, 但是 g(u) 为奇函数，故 $g(-u) = -g(u)$，这样就有

$$g(u) \overset{d}{=} g(-u) = -g(u)$$

这就证明了 g(u) 的分布关于原点对称。引理证毕。

例:

(1) 对任意 $\Sigma > 0$，所元正态分布 $N_p(0, \sigma^2\Sigma)$ 都是关于原点对称的。

(2) 自由度为 n 的多元 t 分布也是关于原点对称的。

定理对于线性模型 $Y = X\beta + e, Cov(e) = \Sigma(\theta)$，假设 e 的分布关于原点是对称的。设 $\widehat{\theta} = \widehat{\theta}(Y)$ 是 θ 的一个估计，它是 y 的偶函数且具有变换不变性。若 $E(\widehat{\beta}(\widehat{\theta}))$ 存在，则两步估计 $\widehat{\beta}(\widehat{\theta})$ 是 β 的无偏估计。

证明因为

$$\begin{aligned}\widehat{\beta}(\widehat{\theta})-\beta &= (X'\Sigma^{-1}(\widehat{\theta})X)^{-1}X'\Sigma^{-1}(\widehat{\theta})Y-\beta \\ &= (X'\Sigma^{-1}(\widehat{\theta})X)^{-1}X'\Sigma^{-1}(\widehat{\theta})e\end{aligned}$$

从 θ 的不变性可得:

$$\widehat{\theta}=\widehat{\theta}(y)=\widehat{\theta}(Y-X\beta)=\widehat{\theta}(e)$$

因而

$$\widehat{\beta}(\widehat{\theta})-\beta=(X'\Sigma^{-1}(\widehat{\theta}(e))X)^{-1}X'\Sigma^{-1}(\widehat{\theta}(e))e$$

记 $u(e)=\widehat{\beta}(\widehat{\theta})-\beta$, 因为 $\widehat{\theta}=\widehat{\theta}(Y)=\widehat{\theta}(e)$ 是 e 的偶函数, 从上式容易推出 $u(-e)=-u(e)$, 即 u(e) 为 e 的奇函数。利用引理便知，u(e) 的分布关于原点是对称的，故有

$$E(u(e))=E(\widehat{\beta}(\widehat{\theta})-\beta)=0.$$

9.3 单因子方差分析模型

以前讨论的线性回归模型中，所涉及的自变量一般来说都是连续变量，研究的基本目的则是寻求因变量与自变量之间客观存在的依赖关系，另外一种模型，它的自变量是示性变量这种变量往往表示某种效应的存在与否，因而只能取 0,1 两个值，这种模型是比较两个或多个因素效应大小的一种有力工具。通常把这种模型称为方差分析模型。

比较三种药治疗某种疾病的效果。药效度量指标为 Y。假设现在对每种药各有 n 个人服用，记 y_{ij} 为服用地 i 种药的第 j 个病人的药效测量值。y_{ij} 可表示为

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

称 μ 为总平均，α_i 表示第 i 种药的效应，e_{ij} 表示随机误差。其均值为 0，方差都相等，彼此互不相关。

模型可写为

$$\begin{pmatrix} y_{11} \\ ... \\ y_{1n} \\ y_{21} \\ ... \\ y_{2n} \\ y_{31} \\ ... \\ y_{3n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ .. & .. & .. & .. \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ .. & .. & .. & .. \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ .. & .. & .. & .. \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ ... \\ e_{1n} \\ e_{21} \\ ... \\ e_{2n} \\ e_{31} \\ ... \\ e_{3n} \end{pmatrix}$$

用 Y,X,β 和 e 分别表示式中的四个向量或矩阵，则该模型具有形式

$$Y = X\beta + e$$

不同的是 X 中元素 $x_{ij} = 1$ 或 0，rk(X)=3 小于 X 的列数 4.

9.4 双因子模型

假设在一次生产过程中，影响产品质量指标 Y 的有两个因素 A 和 B。设 A 有 a 个水平，因素 B 有 b 个水平。记 y_{ij} 表示在因子 A 的第 i 个水平，因素 B 的第 j 个水平时生产的产品质量测量值，则 y_{ij} 可分解为

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, 2, 3; j = 1, ..., b$$

这里 μ 为总平均，α_1 为因素 A 的第 i 个水平的效应，β_j 为因素 B 的第 j 个水平的效应，e_{ij} 为随机误差。

9.54 具有交互效应的双因子模型

在前面的例子中因素 A 和因素 B 的效应具有可加性。但是，在一些实际的问题中，这种情况不总是成立的。例如，在化工试验中，若因素 A 表示化学反应的温度，因素 B 表示化学反应的压力，两者对化学反应的质量或产量 Y 的贡献一般不具有可加性。如果对每一个水平组合 (i,j) 重复 c 次试验，这时一个合理模型是

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, i = 1, ..., a; j = 1, ..., b; k = 1, ..., c$$

这里 γ_{ij} 称为因素 A 的第 i 个水平和因素 B 的第 j 个水平的交互效应。

进一步可推广到三因素方差模型。

9.6 协方差分析模型

线性回归模型所涉及的自变量一般是取连续的数量因子。设计阵 X 的元素 x_{ij} 可取连续值。而在方差分析中模型中，自变量是属性因子，设计阵 X 的元素 x_{ij} 只能取 0,1 两个值。

协方差分析模型则是上述两种模型的混合，模型中的自变量既有属性因子又有数量因子。设计阵由两部分组成，一部分由 0,1 两个数为元素，而另一部分的元素可取连续值。它可以看做由方差分析模型和线性回归模型的设计矩阵组拼而成。

例：试验者欲比较两种饲料的催肥效果，用每种饲料喂养三头猪。要考虑的协变量是小猪的初始体重，记 y_{ij} 为喂第 i 种饲料的第 j 头猪的体重增加量，则 y_{ij} 可分解为

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + e_{ij}, i = 1, 2, j = 1, 2, 3$$

这里 μ 为总平均，α_i 为第 i 种饲料的效应，x_{ij} 为喂第 i 种饲料的第 j 头猪的初始体重，γ 为协变量的系数，即回归系数。e_{ij} 的假设同单向分类模型。若记

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}, X = \begin{pmatrix} 1 & 1 & 0 & x_{11} \\ 1 & 1 & 0 & x_{12} \\ 1 & 1 & 0 & x_{13} \\ 1 & 0 & 1 & x_{21} \\ 1 & 0 & 1 & x_{22} \\ 1 & 0 & 1 & x_{23} \end{pmatrix}, \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma \end{pmatrix}, e = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

则，模型具有形式：

$$y = X\beta + e$$

9.7 混合效应模型

混合效应模型的最一般形式为

$$y = X\beta + U_1\xi_1 + ... + U_k\xi_k, \qquad (\star)$$

其中 y 为 $n \times 1$ 观测量，X 为 $n \times p$ 已知设计矩阵，β 为 $p \times 1$ 非随机的参数向量，称为固定效应。U_i 为 $n \times q_i$ 已知设计矩阵，ξ_i 为 $q_i \times 1$ 随机向量，称为随机效应。一般我们假设

$$E(\xi_i) = 0, Cov(\xi_i) = \sigma_i^2 I_q, Cov(\xi_i, \xi_j) = 0, i \neq j$$

于是

$$E(y) = X\beta, Cov(y) = \sum_{i=1}^{k} \sigma_i^2 U_i U_i'$$

σ_i^2 称为方差分量，也称 $(\star)$ 为方差分量模型。

在模型 $(\star)$ 中，最后一个随机效应向量 ξ_k 是通常的随机误差向量 e，而 $U_k = I_n$。

例 1：两向分类混合模型

研究人的血压，在一天内的变化规律。在一天内选择 a 个时间点测量别观测者的血压，假设观测了 b 个人，用 y_{ij} 表示第 i 个时间点的第 j 个人的血压，则 y_{ij} 可表为

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, ..., a; j = 1, ..., b,$$

这里 α_i 为第 i 个时间点的效应，它是非随机的，是固定效应。β_j 为第 j 个人的个体效应。如果这 b 个人是我们感兴趣的特定的 b 个人。那么 β_j 是非随机的。这时以上模型就是固定效应模型。

但是，如果我们要研究的兴趣只是放在比较不同时间点的血压高低上，被观测的 b 个人是随机抽取的，这时，β_j 就是随机变量，于是该模型就是混合效应模型。

一个效应究竟看作随机的还是固定的，这取决于研究的目的和样品取得的方法。如果观测的个体是随机抽取来的，那么它们的效应就是随机的，否则就是固定的。记

$$y = (y_{11}, ..., y_{1b}, ..., y_{a1}, ..., y_{ab})'$$

这是 $ab \times 1$ 的向量。

$$X = (1_{ab} : I_a \otimes 1_b), U = 1_a \otimes I_b, \gamma = (\mu, \alpha_1, ..., \alpha_a)',$$

$$\beta = (\beta_1, ..., \beta_b)', e = (e_{11}, ..., e_{1b}, ..., e_{a1}, ..., e_{ab})',$$

则模型变形为

$$y = X\gamma + U\beta + e$$

$Var(\beta_i) = \sigma_\beta^2, Var(e_{ij}) = \sigma^2$, 则观测向量的协方差矩阵为

$$Cov(y) = \sigma_\beta^2 UU' + \sigma^2 I_{ab} = \sigma_\beta^2 (J_a \otimes I_b) + \sigma^2 I_{ab}$$

其中 $J_n = 1_n 1_n'$,σ_β^2 和 σ^2 是方差分量。

例 2：Panel 数据模型

假设我们对 N 个个体 (如个人，家庭，公司，城市，国家或区域等) 进行了 T 个时刻的观测，观测数据可写为

$$y_{it} = x_{it}'\beta + \xi_i + \varepsilon_{it}, i = 1, ..., N; t = 1, ..., T,$$

其中 y_{it} 表示第 i 个个体第 t 个时刻的某项经济指标，x_{it} 是 $p \times 1$ 已知向量，它刻画了第 i 个个体在时刻 t 的一些自身特征，ξ_i 是第 i 个个体的个体效应，ε_{it} 是随机误差项。

如果我们的目的是研究整个市场的运行规律，而不是关心这特定的 N 个个体，这 N 个个体只不过是从总体中抽取的随机样本，这时个体效应就是随机的，记

$$y = (y_{11}, ..., y_{1T}, y_{21}, ..., y_{NT})'$$

$$X = (x_{11}, ..., x_{1T}, x_{21}, ..., x_{NT})'$$

$$U_1 = I_N \otimes 1_T, \xi = (\xi_1, ..., \xi_N)', \varepsilon = (\varepsilon_{11}, ..., \varepsilon_{1T}, \varepsilon_{21}, ..., \varepsilon_{NT})'$$

则模型可表示为

$$y = X\beta + U_1\xi + \varepsilon$$

如果假设 $Var(\xi_i) = \sigma_\xi^2, Var(\varepsilon_{it}) = \sigma_\varepsilon^2$，所有 ξ_i 和 ε_{it} 都不想关，则

$$Cov(y) = \sigma_\xi^2 U_1 U_1' + \sigma_\varepsilon^2 I_{NT} = \sigma_\xi^2 (I_N \otimes J_T) + \sigma_\varepsilon^2 I_{NT}$$

σ_ξ^2 和 σ_ε^2 就是方差分量。

如果我们把时间效应也考虑进来，则以上模型可以改写为

$$y_{it} = x_{it}'\beta + \xi_i + \lambda_t + \varepsilon_{it}, i = 1, ..., N, t = 1, ..., T.$$

如果时间效应 λ_t 也看成随机的，并且假设 $Var(\lambda_t) = \sigma_\lambda^2$,$\lambda_t$ 与所有的 ξ_i 和 ε_{it} 不相关，记 $U_2 = 1_N \otimes I_T, \lambda = (\lambda_1, ..., \lambda_T)'$, 则我们得到如下模型

$$y = X\beta + U_1\xi + U_2\lambda + \varepsilon$$

此时，观测向量的协方差矩阵为

$$Cov(y) = \sigma_\xi^2 (I_N \otimes J_T) + \sigma_\lambda^2 (J_N \otimes I_T) + \sigma_\varepsilon^2 I_{IT}$$

$\sigma_\xi^2, \sigma_\lambda^2$ 和 σ_ε^2 为方差分量。

第十章　案例分析

10.1 上市公司净资产收益率预测分析报告

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

一、研究目的

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

二、数据来源和相关说明

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

- 公司当年的净资产收益率 (ROEt): 该财务指标直接反映了公司当年的盈利状况。可以预期，当年表现好的公司，其下年度的表现也趋向于较好。
- 资产周转率 v(ATO): 该财务指标综合评价了企业全部资产的利用效率。
- 主营业务利润/主营业务收入 (PM): 该财务指标反映了公司收入的质量。
- 债务资本比率 (LEV): 该财务指标反映了公司的基本债务状况。
- 主营业务收入增长率 (GROWTH): 该增长率指标反映了公司的成长状况。

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

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

三、描述性分析

为了获得对数据的整体了解，我们对数据进行简单的描述性分析，得到表 10-1。

表 10-1　**样本描述**

变量名	均值	最小值	中位数	最大值	标准差
ROEt	0.068	-1.390	0.080	1.421	0.519
ATO	0.430	-0.928	0.438	1.927	0.460
PM	0.211	-0.424	0.218	0.698	0.181
LEV	0.709	-7.941	0.560	9.362	3.182
GROWTH	0.331	-5.962	0.379	6.092	2.120
PB	2.127	-20.816	2.271	32.591	9.513
ARR	0.201	-2.601	0.223	3.187	0.949
INV	0.100	-0.264	0.102	0.431	0.122
ASSET	21.066	18.629	21.056	23.414	0.855
ROE	0.410	-1.161	0.420	5.285	0.545

从表 10-1 中的描述性统计可以看出：

公司当年的净资产收益率 (ROEt) 介于-1.390 与 1.421 之间，其平均水平约为 0.068(平均值) 和 0.080(中位数)，标准差为 0.519。而公司下年度的净资产收益率 (ROE) 介于-1.161 与 5.285 之间，其平均水平约为 0.410(平均值) 和 0.420(中位数)，标准差为 0.545。从中位数可以看出，超过半数的公司有正的净资产收益率 (ROE)；同当年相比，公司下一年的净资产收益率 (ROE) 有一定的增长，而且不同公司之间的差距在扩大。

资产周转率 (ATO) 的均值 (0.430) 和标准差 (0.460) 从一个侧面反映出大多数公司资产的平均利用水平。主营业务利润/主营业务收入 (PM) 的均值 (0.211) 和标准差

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

四、数据建模

1. 全模型分析

在本节中，我们用线性回归的分析方法建立模型，以此来寻找自变量 (公司当年的 9 项财务指标) 和因变量 (公司下一年的净资产收益率，ROE) 之间的关系。通过从图 10-1 中考察观测值的 Cook 距离，我们删除了第 47 个强影响点。

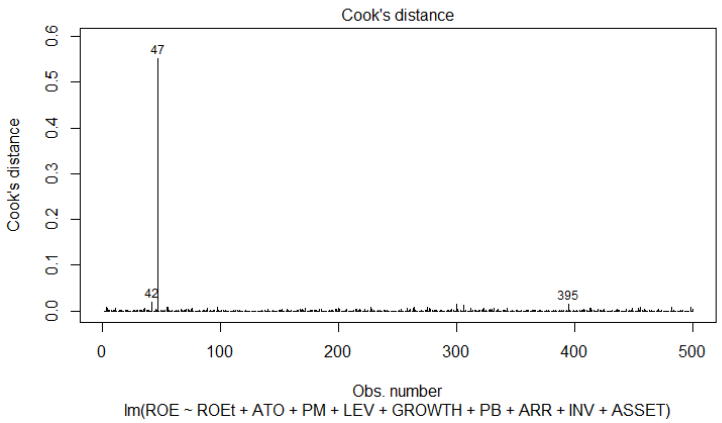


图 10-1　Cook 距离

利用剩余的数据，我们运用包括全部 9 个自变量的全模型对因变量进行估计，得到表 10-2 所示的估计结果。从表 10-2 中我们可以看到，模型 F 检验的 P 值非常小，表明该模型是显著的，即自变量和因变量之间确实存在一定的关系。另外，未调整的判决系数 (R-Square) 为 31.29%，调整后的判决系数为 30.03%，这都表明该模型对自变量和因变量之间的关系有一定的解释能力。通过考察各个自变量对应的 t 检验的 P 值，在 0.05 的置信水平下，我们可以断定下一年的净资产收益率 (ROE) 与当年净资产收益率

(ROEt) 和主营业务收入增长率 (GROWTH) 之间有着显著的正相关关系，与债务资本比率 (LEV) 显著地负相关，而对于其他变量暂时没有定论。

表 10-2　全模型

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

2. 模型选择与预测

从以上全模型的分析结果容易发现，有三个财务指标非常重要，但是我们不能排除其他变量也有预测能力的可能。因此，我们用两种两种最为常用的选择变量的方法，即 AIC 和 BIC，来选择最具有预测能力的模型。

如果使用 AIC 来选择模型，我们可以得到如下的模型估计结果，如表 10-3 所示。

表 10-3　AIC

变量名	系数估计值	标准差	P 值
截距	0.384554	0.018384	0.00
ROEt	0.546030	0.034258	0.00
LEV	-0.028790	0.005556	0.00
GROWTH	0.015699	0.008071	0.05
ARR	-0.034533	0.018101	0.06
残差项标准差	0.3818	模型 F 检验 P 值	< 0.0001
判决系数 (R-Square)	0.4205	调整的判决系数 (R-Square)	0.4158

从表 10-3 中可以看到，AIC 认为第 1 个变量 (ROEt)、第 4 个变量 (LEV)、第 5 个变量 (GROWTH) 以及第 7 个变量 (ARR) 对于预测下一年的净资产收益率 (ROE) 非

常重要。而且，选出的所哟变量都在 0.10 的水平下是显著的，其判断系数 (R-Square) 相对于全模型有所提高。

如果用 BIC 来选择模型，我们可以得到如下的模型估计结果，如表 10-4 所示。从表 10-4 中可以清楚地看到，BIC 认为第 1 个变量 (ROEt) 和第 4 个变量 (LEV) 对于预测来年的净资产收益率非常重要，但是它不认为第 5 个变量 (GROWTH) 和第 7 个变量 (ARR) 也很重要。而且，BIC 选出的所有变量都在 0.01 的水平下是显著的。其判决系数相对于全模型也有所提高，但略微低于 AIC 所选出的模型。

表 10-4　BIC

变量名	系数估计值	标准差	P 值
截距	0.383049	0.017940	0.00
ROEt	0.549988	0.034408	0.00
LEV	-0.029561	0.005576	0.00
残差项标准差	0.384	模型 F 检验 P 值	< 0.0001
判决系数 (R-Square)	0.4116	调整的判决系数 (R-Square)	0.4092

为了从三个不同的模型 (全模型、AIC 选择的最优模型以及 BIC 选择的最优模型) 中选出最具有预测能力的模型，我们用 2003 年的数据来对模型的预测能力进行检验。另外，我们也考虑最简单的预测方法，即仅用 2003 年的盈利预测 2004 年，称之为” 直接预测”。通过计算得到模型的预测结果比较，如表 10-5 所示。

表 10-5　**预测结果比较**

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

从表 10-5 中我们不难发现。所有基于模型的预测都要优于仅仅依靠当年净资产收益率的预测。具体地说，如果我们仅仅利用公司当年的盈利能力来简单预测来年，那么其平均绝对误差为 0.4157；而如果考虑了全模型，那么该预约误差下降为 0.2945。在此基础上，经过 AIC 变量选择后的模型预测精度为 0.2937，而经过 BIC 变量选择后的模型预测精度为 0.2948。

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

五、结论及建议

从上述的分析结果可知，我国上市公司的财务信息为预测下一年的盈利能力提供了重要信息，而且表现出比较好的预测能力。具体来说，从保守和谨慎的角度来看，公司当年的净资产收益率 (ROEt)、债务资本比率 (LEV)、主营业务收入增长率 (GROWTH) 以及应收账款/主营业务收入 (ARR) 这四个财务指标，尤其是前两个，对于预测公司下一年的净资产收益率 (ROE) 非常重要。

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

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

附录　程序及注释

```
rm(list=ls())       #清理当前工作空间
a=read.table("C:/Users/Administrator/Desktop/R/CH1/roe.txt",header=T)
                    #读入以空格为分隔符，并带有标题行的文本文件
round(a[1:10,],4)         #用4位小数点的格式显示a中前10行的数据
a1=a[a$year==2002,-1]     #从a中选出year为2002的数据,并删除第1列,然后赋值给a1
Mean=sapply(a1,mean)      #计算a1中各列的均值
Min=sapply(a1,min)        #计算a1中各列的最小值
Median=sapply(a1,median)  #计算a1中各列的中位数
Max=sapply(a1,max)        #计算a1中各列的最大值
SD=sapply(a1,sd)          #计算a1中各列的标准差
cbind(Mean,Min,Median,Max,SD)
                    #将均值、最小值、中位数、最大值、标准差
```

```
                    #集中在一起展示
round(cor(a),3)          #计算相关系数，用3位小数点的格式展示
plot(a1$ROEt,a1$ROE)     #画出ROEt和ROE之间的散点图
lm1=lm(ROE~ROEt+ATO+PM+LEV+GROWTH+PB+ARR+INV+ASSET,data=a1)
                    #用a1中数据拟合线性回归模型
summary(lm1)             #给出模型lm1中系数估计值、P值等细节
round(a1[1:10,],3)       #用3位小数点的格式显示a1中前10行的数据
par(mfrow=c(2,2))        #设置画图为2x2的格式
plot(lm1,which=c(1:4))   #画出lm1中对应于模型检验的4张图，
                    #包括残差图、QQ图和Cook距离图

a1=a1[-47,]              #删除a1中第47行的观测
lm2=lm(ROE~ROEt+ATO+PM+LEV+GROWTH+PB+ARR+INV+ASSET,data=a1)

                    #用上一行命令得到的新数据a1再次拟合线型回归模型
                    #结果赋值给lm2
plot(lm2,which=c(1:4))   #画出lm2中对应于模型检验的4张图，
                    #包括残差图、QQ图和Cook距离图

library(car)             #载入程序包Car
round(vif(lm2),2)        #计算模型lm2的方差膨胀因子，用2位小数点的格式展示

lm.aic=step(lm2,trace=F) #根据AIC准则选出最优模型，并赋值给lm.aic
summary(lm.aic)          #给出模型lm.aic中系数估计值、P值等细节

lm.bic=step(lm2,k=log(length(a1[,1])),trace=F)
                    #根据BIC准则选出最优模型，并赋值给lm.bic
summary(lm.bic)          #给出模型lm.bic中系数估计值、P值等细节

a2=a[a$year==2003,-1]    #从数据a中选出year为2003的观测，
                    #并删除第一列，赋值给a2
round(a2[1:5,],3)        #用3为小数点的格式展示a2的前5行数据

y1=predict(lm2,a2)       #用全模型lm2对a2进行预测
```

```
y2=predict(lm.aic,a2)     #用模型lm.aic对a2进行预测
y3=predict(lm.bic,a2)     #用模型lm.bic对a2进行预测
y0=a2[,10]              #选出a2中的第10列，即当年的ROE

r0=y0-a2$ROEt            #用当年ROE对下年进行预测的残差
r1=y0-y1                 #用全模型lm2预测的残差
r2=y0-y2                 #用模型lm.aic预测的残差
r3=y0-y3                 #用模型lm.bic预测的残差
resid=abs(as.data.frame(cbind(r0,r1,r2,r3)))
                    #计算残差的绝对值
sapply(resid,mean)       #计算不同模型的平均绝对偏差，
                    #即对残差的绝对值取平均
```

10.2 北京市商品房价格影响因素分析报告

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

一、研究目的

北京市房地产市场是我国房地产市场中最为发达，也是最具有代表性的几个房地产市场之一，而且其商品房的销售价格差异巨大，从最低的 4000 元/平方米一直到最高的 20000 元/平方米。因此，找出使什么样的因素在影响着北京市商品房的销售价格，以及为何会产生价格上的巨大差异，是一件非常有意义的事情。本报告利用北京市商品房的销售价格数据，确定影响商品房销售价格的重要因素，并量化了这些因素的影响。通过了解影响商品房销售价格的因素，购房者可以更加理性地选择房屋，价值评估机构可以通过理论模型来评估房屋价值，房地产项目开发商可以预测项目的销售价格，再结合各个项目的成本，从而科学地选择项目进行开发并制定相关的开发策略。

二、数据来源和相关说明

我们从“搜房网”(www.soufun.com) 上随机选取了北京地区 2003-2004 年度开盘的新楼盘共 506 个。在进行数据清理后，我们最终得到 200 个合格的楼盘样本。我们希望能够基于这样一个公开的实际数据，建立恰当的经济计量模型，从数量上刻画房地产销售价格同各个影响因素之间的关系。具体地说，我们的数据包含了表 10-6 中的信息。

从表 10-6 可以看出，在数据所涵盖的所有变量中，有一个是我们特别感兴趣的，试图通过其他变量来解释的，那就是销售均价。这也就是我们的因变量，而其他的所有变量都是自变量。

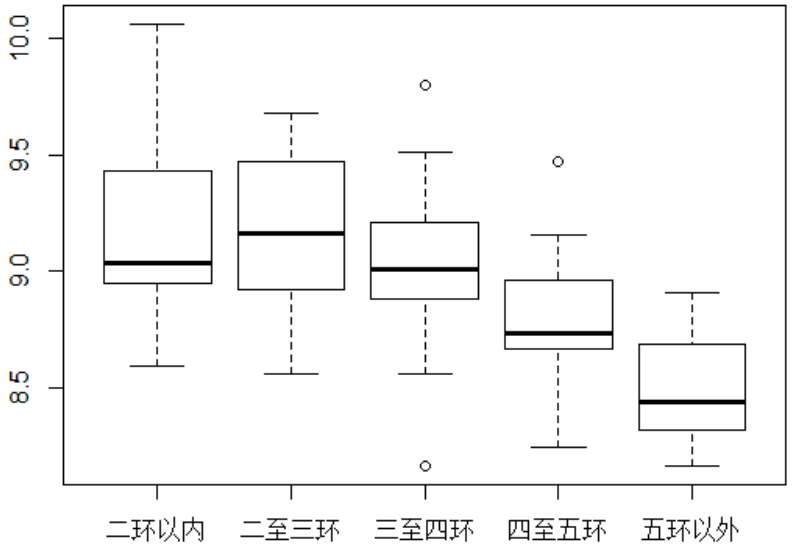
通过对数据的简单分析，我们发现对销售均价进行对数变换对本报告的数据分析是有帮助的，具有稳定方差的作用。在以后的分析中，我们都采用对数变换后的销售均价作为因变量。

三、描述性分析

为了获得对数据的整体概念，我们利用盒装图对数据进行简单的描述性分析，如图 10-2 所示。

表 10-6　变量说明

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



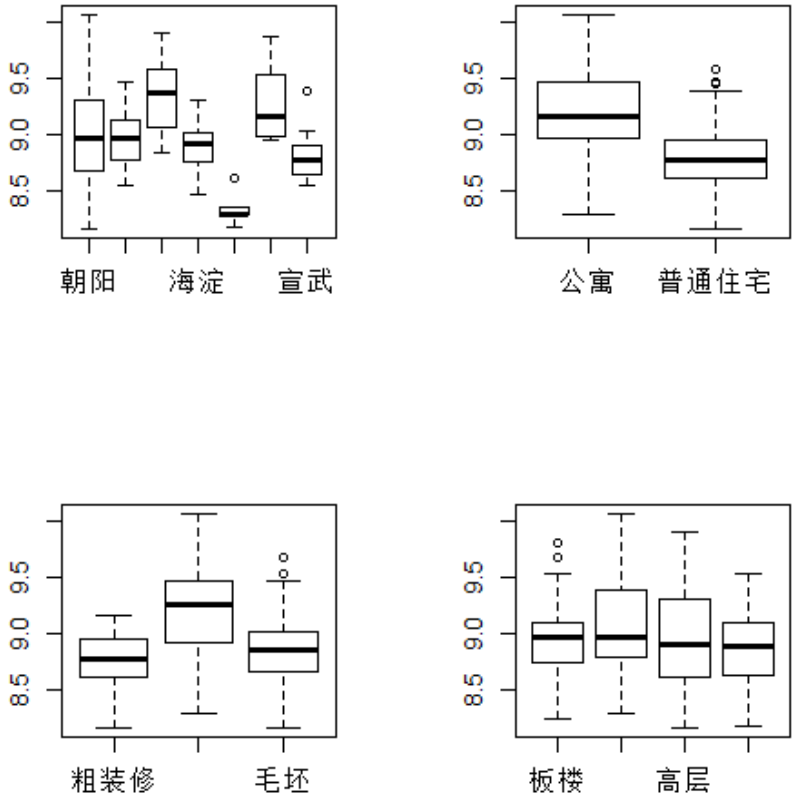


图 10-2　**盒装图**

从图 10-2 可以看出：

- 地理位置对房屋的均价有显著的影响。随着地理位置从内环向外环延伸，商品房均价的中位数呈现出明显的下降趋势。这说明，距离城市中心越远，商品房平均价格越低。
- 城区的不同不仅极大地影响着房屋的均价，而且还影响着房屋价格的波动程度。
- 公寓的平均价格明显高于普通住宅，但波动程度相当。
- 精装修房屋的平均价格明显高于粗装修以及毛坯房，而且价格波动较大。此外，粗装修房屋的均价并没有明显高于毛坯房。
- 不同的房屋类型 (如板楼、塔楼、板塔结合) 对房屋的平均价格及其波动程度影响都很小。

四、数据建模

在本节中，我们考虑的是离散型的自变量对销售价格的影响，因此采用方差分析的方法建立模型，来寻找因变量和自变量之间的关系。具体地说，就是比较不同城区、不同地理位置、不同建筑类别等的商品房价格从平均水平上来说有没有显著差异，并量化这种差异。需要特别指出的是，一般来说，地理位置对房地产价格的影响非常重要，因此我们在模型分析中考虑两个地理位置因素 (即所在区县以及所在环线) 的交互作用，而对其他的自变量我们采用可加模型的形式。利用全部数据估计各个因素对销售价格的具体影响，得到方差分析表 10-7，从表 10-7 中可以发现所在区县、所在环线、物业类别、装修状况以及所在区县和所在环线的交互项的作用均是显著的，建筑类别的作用不显著。

表 10-7　方差分析表

	自由度	平方和	均方和	F 值	P 值
所在区县	6	6.94	1.16	27.20	0.0000
所在环线	4	8.00	2.00	47.02	0.0000
物业类别	1	2.12	2.12	49.90	0.0000
装修状况	2	1.44	0.72	16.91	0.0000
建筑类别	3	0.17	0.06	1.33	0.2671
所在区县: 所在环线	8	0.80	0.10	2.36	0.0196
残差	175	7.44	0.04		

利用全部数据估计各个因素的不同水平对销售价格的具体影响，得到估计结果如表 10-8 所示。从表 10-8 中可以发现，“所在区县”的不同水平的价格之间近似有如下从大到小的关系：海淀、朝阳、石景山、东城、西城、崇文、宣武。需要注意的是，这种近似关系式通过和朝阳区的比较得出的，而且石景山区和朝阳区之间的差异并不显著。根据“所在环线”的估计结果可以看出，地理位置对房屋的均价有显著的影响，随着地理位置从内环向外环延伸，商品房均价的平均水平呈现出显著的下降趋势。还可以发现，对于不同“物业类别”的房屋，公寓的平均价格显著高于普通住宅。“装修状况”的不同水平之间也有区别，精装修房屋的价格显著高于粗装修的房屋，而毛坯房与粗装修房屋之间的区别并不是很显著。另外，所在城区和所在环线的交互作用也是非常显著的。因此，通过上述分析，我们可以判断出显著影响北京市房平均销售价格的因素有所在区县、所在环线、物业类别以及装修状况。

表 10-8　模型估计结果

变量	系数估计值	标准差	t 值	P 值
截距	9.8329	0.2190	44.91	0.0000
所在区县: 崇文	-0.8238	0.2219	-3.71	0.0003
所在区县: 东城	-0.5172	0.2257	-2.29	0.0231
所在区县: 海淀	0.3046	0.1113	2.74	0.0068
所在区县: 石景山	-0.0494	0.1037	-0.48	0.6343
所在区县: 西城	-0.5619	0.2342	-2.40	0.0175
所在区县: 宣武	-0.8599	0.2221	-3.87	0.0002
所在环线: 二至三环	-0.6229	0.2163	-2.88	0.0045
所在环线: 三至四环	-0.8028	0.2159	-3.72	0.0003
所在环线: 四至五环	-1.0186	0.2148	-4.74	0.0000
所在环线: 五环以外	-1.3581	0.2266	-5.99	0.0000
物业类别: 普通住宅	-0.1635	0.0359	-4.56	0.0000
装修状况: 精装修	0.2608	0.0617	4.23	0.0000
装修状况: 毛坯	0.0680	0.0548	1.24	0.2161
装修状况: 板塔结合	-0.0283	0.0542	-0.52	0.6027
装修状况: 高层	-0.0260	0.0438	-0.59	0.5543
装修状况: 塔楼	-0.0391	0.0384	-1.02	0.3096
所在区县: 崇文: 所在环线: 二至三环	0.4794	0.2483	1.93	0.0551
所在区县: 东城: 所在环线: 二至三环	0.5557	0.2460	2.26	0.0251
所在区县: 海淀: 所在环线: 二至三环	-0.4785	0.1379	-3.47	0.0007
所在区县: 西城: 所在环线: 二至三环	0.5026	0.3205	1.57	0.1186
所在区县: 宣武: 所在环线: 二至三环	0.3586	0.2411	1.49	0.1387
所在区县: 崇文: 所在环线: 三至四环	0.5038	0.3057	1.65	0.1011
所在区县: 海淀: 所在环线: 三至四环	-0.3093	0.1289	-2.40	0.0175
所在区县: 海淀: 所在环线: 四至五环	-0.3246	0.1295	-2.51	0.0131

五、结论与建议

从上述分析结果可知，影响北京市商品房平均销售价格的主要因素有所在区县、所在环线、物业类别以及装修状况。而且，所在区县与所在环线之间存在着相互影响。商品房购房者可以利用本报告得到的模型，对价格进行合理的预测，比较不同的楼盘，从而为购房决策提供帮助。本报告的分析结果还可以为相关机构的价值评估提供理论依据。对于房地产开发商，本报告提供了更多有用的信息。开发商可以利用该模型来合理

地选择项目，并结合成本核算的结果来对项目的开发制定最优的策略。例如，由装修类型不同水平的差异可以发现，精装修房屋的销售价格显著高于毛坯房，但粗装修房屋则与毛坯房之间无显著差异，因此开发商就可以考虑不对房屋进行粗装修，而在成本核算之后根据利润最大化的原则选择精装修或者毛坯房。类似地，对于其他因素的分析也可以为开发商提供有价值的参考。

附录　程序及注释

```
rm(list=ls())              #清空当前工作空间
a=read.csv("C:/Users/Administrator/Desktop/R/CH2/real.csv",header=T)
                          #读入csv格式的数据，赋值为a
attach(a)                  #将数据集a中个变量添加到工作空间，便与直接调用
pairs(a[,c(1:6)])          #对a的前6列做散点图

boxplot(price~ring)        #画出price与ring之间的盒状图

log.price=log(price)       #对price对数变化，并赋值给log.price
boxplot(log.price~ring)    #画出log.price与ring之间的盒状图

par(mfrow=c(2,2))          #设置画图模式2x2的格式
boxplot(log.price~dis)     #画出log.price与dis之间的盒状图
boxplot(log.price~wuye)    #画出log.price与wuye之间的盒状图
boxplot(log.price~fitment)#画出log.price与fitment之间的盒状图
boxplot(log.price~contype)#画出log.price与contype之间的盒状图

summary(a[,c(1:5)])        #给出a中前5列的描述性分类统计

lm1=lm(log.price~as.factor(ring))
                          #用离散变量ring做解释性变量做单因子方差分析
library(car)               #载入程序包car
Anova(lm1,type="III")      #对模型lm1做三型方差分析
summary(lm1)               #显示模型lm1的各方面细节，包括参数估计值、P值等

lm2.1=lm(log.price~as.factor(ring)+as.factor(wuye))
                          #不带交互作用的双因子方差分析
```

```
Anova(lm2.1,type="III")  #对模型lm2.1做三型方差分析
summary(lm2.1)           #显示模型lm2.1的各方面细节，包括参数估计值、P值等

lm2.2=lm(log.price~as.factor(ring)*as.factor(wuye))
                         #带交互作用的双因子方差分析
Anova(lm2.2,type="III")  #对模型lm2.2做三型方差分析
summary(lm2.2)           #显示模型lm2.2的各方面细节，包括参数估计值、P值等
dis=as.factor(dis)       #把数据因子化
ring=as.factor(ring)
wuye=as.factor(wuye)
fitment=as.factor(fitment)
contype=as.factor(contype)
lm4=lm(log.price~dis*ring+wuye+fitment+contype)
                         #包括所有变量的全模型方差分析
summary(lm4)             #显示模型lm4的各方面细节，包括参数估计值、P值等
par(mfrow=c(2,2))        #设置画图模式为2x2
plot(lm4,which=c(1:4))   #画出lm4中对应于模型检验的4张图，
                         #包括残差图、QQ图和Cook距离图
par(mfrow=c(1,1))        #设置画图模式为1x1

a0=read.csv("D:/Practical Business Data Analysis/case/CH2/new.csv")
                         #读入新数据，赋值给a0
a0=a0[,c(1:5)]           #取a0的前5列
a0                       #展示a0的数据

y.pred=exp(predict(lm4,a0))#用模型lm4对a0做预测
a0$y.pred=y.pred         #将预测结果赋值给a0中的变量y.pred
a0                       #展示a0的数据,包括预测值
```

10.3 教学评估数据分析报告

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

一、研究目的

在大专院校的教学管理中，教学评估是一种重要的衡量教员教学成绩的手段。如果该评估手段非常准确，那么我们就可以通过简单地比较两门课的教学评估成绩来比较两个教员的教学绩效。因此，如何客观有效地对教员的教学进行评估，是一件非常重要且基本的工作。本报告试图通过对北京大学的光华管理学院教学评估数据的分析，建立一个计量经济学模型，以此来找出影响最终教学评估成绩的因素，并根据数据分析的结果，提出一个合理的绩效考核标准。

二、数据来源和相关说明

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

三、描述性分析

在所有解释性变量中，学生人数是唯一的连续型变量。经验告诉我们，学生人数是一个非常重要的因素，我们利用散点图 (即图 10-3) 来寻找它们之间的关系。

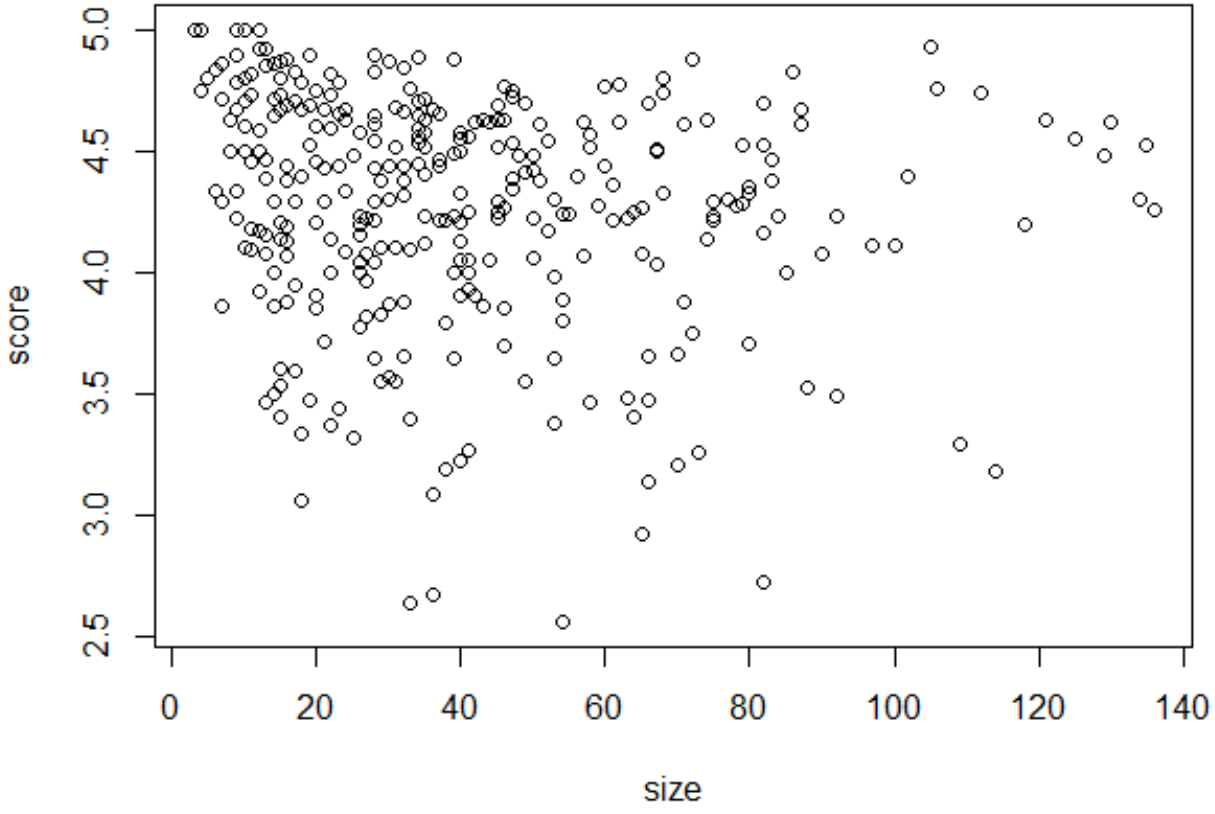


图 10-3　最终评估得分与学生人数的散点图

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

为了从直观上获得对各个离散型变量与因变量之间关系的初步认识，我们利用盒状图对数据进行简单的描述性分析，得到图 10-4。

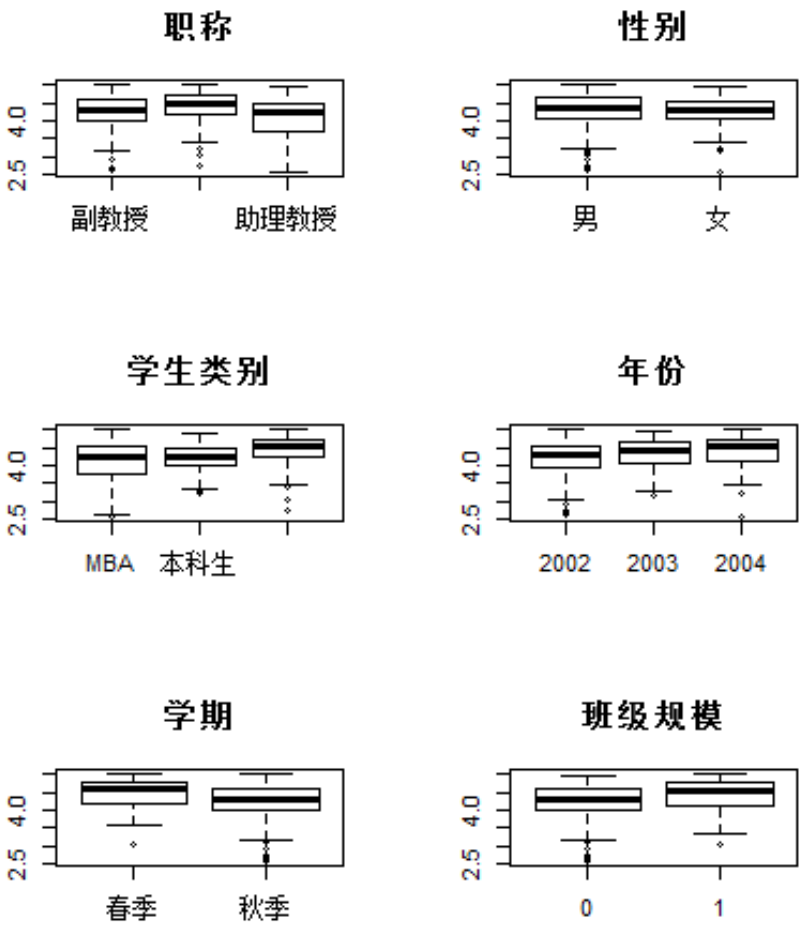


图 10-4　基于盒状图的描述性分析

从图 10-4 中我们可以得到如下直观的印象：

- 教员职称确实能够影响教学评估成绩。随着教员职称的提高 (从助理教授到副教授再到正教授)，平均教学评估成绩 (以中位数计) 依次增高。这在一定程度上反映出积累的教学科研经验对教学评估成绩的影响。
- 教员不同性别的两组之间差别很小，表明性别对教学评估成绩影响甚微。
- 不同的学生类别对教学评估成绩影响很大。研究生给出的平均教学评估成绩明显高于本科生和 MBA，而本科生和 MBA 之间差异不大。
- 随着时间的推移，各组的中位数在依次提高，表明北京大学光华管理学院的教学质量在稳步提高。

- 秋季学期 (即每学年的第一个学期) 的教学评估成绩低于春季学期。
- 小于 20 人的班级教学评估成绩明显高于大于 20 人的班级。

四、数据建模

1. 全模型分析

通过上一节的分析我们可以发现，自变量与因变量之间确实存在相关性。我们用多因素协方差分析的方法建立模型，来寻找自变量和因变量之间的关系。从先前的分析中发现，最终评估得分与班级规模之间有显著的相关性，但是与学生人数之间的关系并不显著，因此我们考虑班级规模与学生人数之间的交互作用。我们单独对其进行的检验也证明了该交互作用是显著的。而对其他的变量，我们考虑可加模型。利用全部数据对此进行估计，我们得到模型估计结果 (如表 10-9 所示)。

表 10-9　全模型方差分析

变量名	P 值
教员职称	< 0.0000
教员性别	0.7565
学生类别	< 0.0000
年份	0.0002
学期	0.6963
班级规模	0.0680
学生人数	0.6690
班级规模: 学生人数	0.0011
残差项标准差 0.4319	模型 F 检验 P 值 < 0.0001
判决系数 (R-square)0.2054	调整的判决系数 (R-square)0.1787

从表 10-9 中我们可以看到，模型 F 检验的 P 值非常小，表明该模型是显著的，即自变量和因变量之间确实存在一定的关系。另外，未调整的判决系数 (R-square) 为 20.54%，调整后的判决系数 (adjusted R-square) 为 17.87%，这都表明该模型对自变量和因变量之间的关系有一定的解释能力。通过对各个自变量所对应的 t 检验的 P 值得考察，在 0.10 的显著水平下，我们可以断定：①教员职称、学生类别、年份、班级规模和学生人数都是重要的影响因素；②教员的性别和学期对最终评估得分影响甚微。从表 10-10 中我们可以看到以上五个重要的影响因素也大致与人们的直觉相一致。教员职称在一定程度上代表了教员的教学能力，教学能力高的教员自然能得到较高的教学评估成绩；MBA 和本科生对教员的期望值一般都较高；随着学院的成长，教员各方面经验的

表 10-10　全模型估计结果

变量	系数估计值	标准差	t 值	P 值
截距	3.9255	0.1068	36.77	0.0000
教员职称：正教授	0.1637	0.0696	2.35	0.0191
教员职称：助理教授	-0.1006	0.0703	-1.43	0.1532
教员性别：女	0.0334	0.0665	0.50	0.6160
学生类别：本科生	0.0945	0.0673	1.40	0.1616
学生类别：研究生	0.2429	0.0583	4.17	0.0000
年份：2003	0.1446	0.0594	2.44	0.0153
年份：2004	0.2404	0.0569	4.22	0.0000
学期：秋季	-0.0214	0.0845	-0.25	0.8005
班级规模：1(小于 20)	0.6018	0.1611	3.74	0.0002
学生人数	0.0009	0.0012	0.75	0.4520
班级规模：1(小于 20): 学生人数	-0.0351	0.0107	-3.28	0.0011

积累在教学上也有显著的体现，这就导致最终教学评估得分逐步递增；班级规模是一个很重要的影响因素，小规模的班级总是比大规模的班级更容易教授，而且教员也能有更多的精力来关注每位同学。

2. 模型选择

我们很容易从以上全模型的分析结果中发现，有五个自变量非常重要，对最终评估得分有显著的影响，但是我们不能排除其他变量也有预测能力的可能。因此，我们用两种最为常用的选择变量的方法，即 AIC 和 BIC，来选择最具有预测能力的模型。通过计算，得到模型如，我们发现 AIC 和 BIC 选择了同样的模型，而且正好是包含上述五个重要因素的模型。因此，我们可以认为该模型是一个科学合理的模型，称为最优模型。对该模型的参数估计结果表示如表 10-11 所示。

表 10-11　最优模型参数估计结果

变量	系数估计值	标准差	t 值	P 值
截距	3.9161	0.0826	47.40	0.0000
教员职称：正教授	0.1748	0.0525	3.33	0.0010
教员职称：助理教授	-0.0962	0.0695	-1.39	0.1670
学生类别：本科生	0.0954	0.0671	1.42	0.1564
学生类别：研究生	0.2416	0.0578	4.18	0.0000
年份：2003	0.1470	0.0590	2.49	0.0132
年份：2004	0.2415	0.0567	4.26	0.0000
班级规模：1	0.5962	0.1603	3.72	0.0002
学生人数：	0.0008	0.0011	0.68	0.4949
班级规模: 学生人数	-0.0349	0.0107	-3.28	0.0012

为了确保模型分析结果的可能性，我们对模型的独立性，正态性以及同方差假定进行检验，得到图 10-5，观察图 10-5 可以发现这些假定基本满足。

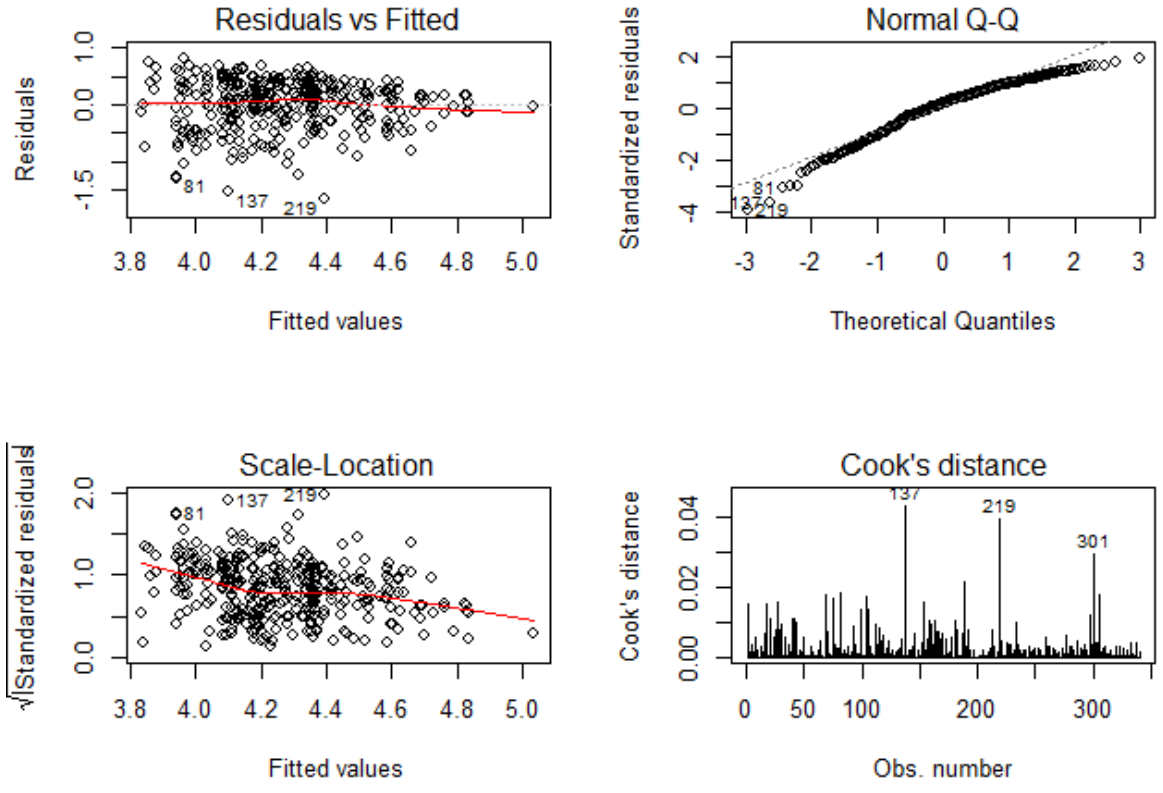


图 10-5　模型诊断图

根据表 10-11 的估计结果，我们可以获得以下重要结论：

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

五、结论与建议

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

附录　程序及注释

```
rm(list=ls())            #清空当前工作空间
a=read.csv("C:/Users/Administrator/Desktop/R/CH3/teaching.csv",header=T)
                         #读入csv格式的数据，并赋值给a
attach(a)                #将a中各变量加入工作空间
a[c(1:5),]               #展示a的前5行数据
plot(size,score)         #画出size与score的散点图

boxplot(score~ceiling(size/20))
                         #画出score与分组的size的盒状图
table(ceiling(size/20))  #计算分组的size的频数
group=1*(size<=20)       #根据size是否大于20生成0、1变量，
```

```
par(mfrow=c(3,2))          #设置画图模式为3x2
boxplot(score~title,main="职称")#画出score与title的盒状图
boxplot(score~gender,main="性别")#画出score与gender的盒状图
boxplot(score~student,main="学生类别")#画出score与student的盒状图
boxplot(score~year,main="年份")#画出score与year的盒状图
boxplot(score~semester,main="学期")#画出score与semester的盒状图
boxplot(score~group,main="班级规模")#画出score与group的盒状图
par(mfrow=c(1,1))          #设置画图模式，还原成1x1

lm1=lm(score~as.factor(group)+size)
                    #用解释性变量group和size拟合线性模型
summary(lm1)              #显示模型lm1的各方面细节，包括参数估计值、P值等
plot(size,score)          #画出size与score的散点图
points(size,lm1$fitted,col=2)
                    #在size与score的散点图上用红色画出size与lm1$fitted的数据点

lm2=lm(score~as.factor(group)*size)
                    #拟合带交互作用的线性模型
library(car)              #载入程序包car
Anova(lm2,type="III")     #对模型lm2做三型方差分析

summary(lm2)              #显示模型lm2的各方面细节，包括参数估计值、P值等
plot(size,score)          #画出size与score的散点图
points(size,lm2$fitted,col=2)
                    #在size与score的散点图上用红色画出size与lm2$fitted的数据点

lm3.1=lm(score~as.factor(title)+as.factor(gender)+as.factor(student)
+as.factor(year)+as.factor(semester)+as.factor(group)*size)
                    #拟合考虑所有变量的全模型
Anova(lm3.1,type="III")   #对模型lm3.1做三型方差分析
lm3.2=lm(score~as.factor(title)+as.factor(student)+as.factor(year)+
as.factor(group)*size)    #删除全模型中不显著的变量，重新拟合
Anova(lm3.2,type="III")   #对模型lm3.2做三型方差分析
```

```
summary(lm3.2)          #显示模型lm3.2的各方面细节，包括参数估计值、P值等

par(mfrow=c(2,2))       #设置画图模式为2x2的格式
plot(lm3.2,which=c(1:4)) #画出模型lm3.2的前4个与模型诊断相关的图,包括残差图、cook距离等
par(mfrow=c(1,1))       #设置画图模式，还原成1x1
Anova(lm3.1,type="III")  #对模型lm3.1做三型方差分析

AIC(lm3.1)              #计算模型lm3.1的AIC值
AIC(lm3.1,k=log(length(score)))#计算模型lm3.1的BIC值

Anova(lm3.2,type="III")  #对模型lm3.2做三型方差分析

AIC(lm3.2)              #计算模型lm3.2的AIC值
AIC(lm3.2,k=log(length(score)))#计算模型lm3.2的BIC值

lm.aic=step(lm3.1,trace=F)#根据AIC准则从lm3.2中选出最优模型
Anova(lm.aic,type="III") #对模型lm.aic做三型方差分析

lm.bic=step(lm3.1,k=log(length(score)),trace=F)
                    #根据BIC准则从lm3.2中选出最优模型
Anova(lm.bic,type="III") #对模型lm.bic做三型方差分析

par(mfrow=c(2,2))
plot(lm.aic,which=c(1:4))
par(mfrow=c(1,1))
a0=read.csv("D:/Practical Business Data Analysis/case/CH3/new.csv",header=T)
                    #读入用作预测的数据，并赋值给a0
a0$group=1*(a0$size<=20) #根据size是否大于20生成新的0、1变量,并赋值给a0$group
a0                      #展示数据a0

score.hat=predict(lm.aic,a0)#利用lm.aic对a0进行预测预测
a0$score.hat=score.hat   #将预测值赋给a0的变量score.hat
a0                      #展示数据a0，此时包括预测值
```

summary(lm.aic)　　#显示模型lm1的各方面细节，包括参数估计值、P值等

a$adj.score=lm.aic$residuals#将模型lm.aic中的残差赋值给a中的变量adj.score
a[c(1:10),]　　#展示数据a0的前10行，此时包括adj.score