第三章 模型的估计

估计回归参数的最基本的方法是最小二乘法. 本章前三节讨论如何应用最小二乘法求回归参数的最小二乘估计, 并研究这种估计的基本性质. 第四和第五节讨论回归模型的基本假设的适用性以及当这些假设不适用时, 对数据应该做的变换. 第六节讨论广义最小二乘估计. 第七节讨论一类特殊的自变量, 并分析它是如何给最小二乘估计带来危害的. 第八节和第九节分别讨论两种新的估计方法: 岭估计和主成分估计.

假定自变量不是随机变量, 因为它的取值往往可以被人为控制. 当然, 实际情况中自变量也可以是随机变量, 这时关于模型的理论分析需要借助概率极限理论, 本课程不涉及这些内容. 此外, 我们约定: 一维的自变量用小写字母表示, 一维的因变量是随机变量, 也用小写字母表示(有时也表示样本观测值).

3.1 最小二乘估计

用y表示因变量, $x_1,\cdots,x_p$表示(可能)对y有解释能力的p个自变量. 假设它们来自如下的多元线性回归模型:

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+e, \tag{3.1.1}$$

其中e是随机误差(模型误差), $\beta_0,\cdots,\beta_p$为待估的未知参数. 称β_0为回归常数, $\beta_1,\cdots,\beta_p$为(多元)回归系数, 有时把它们统称为回归系数. 记$\boldsymbol{x}=(x_1,\cdots,x_p)'$. 回归分析的首要任务是估计回归函数:

$$\mathsf{E}(y|\boldsymbol{x})=\beta_0+\beta_1x_1+\cdots+\beta_px_p,$$

它定量地刻画了因变量在平均意义下与自变量之间的相依关系. 显然, 对于线性回归模型(3.1.1), 估计回归函数等价于估计回归系数.

假定已经有样本$\{(x_{i1},\cdots,x_{ip},y_i),\ \ i=1,\cdots,n\}$, 则它们满足

$$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+e_i,\ \ i=1,\cdots,n. \tag{3.1.2}$$

假设模型误差$\{e_i,i\geq1\}$满足Gauss-Markov假设. 若采用矩阵符号, 则(3.1.2)可写为

$$\boldsymbol{Y}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}1&x_{11}&\cdots&x_{1p}\\1&x_{21}&\cdots&x_{2p}\\\vdots&\vdots&\ddots&\vdots\\1&x_{n1}&\cdots&x_{np}\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix}+\begin{pmatrix}e_1\\e_2\\\vdots\\e_n\end{pmatrix},$$

或

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}. \tag{3.1.3}$$

其中$\boldsymbol{Y}$是$n\times1$的观测向量(因变量向量), $\boldsymbol{X}$是$n\times(p+1)$的设计矩阵(假设$p+1<n$), $\boldsymbol{\beta}$是$(p+1)\times1$的回归系数向量, $\boldsymbol{e}$是$n\times1$的随机误差向量. 将Gauss-Markov假设也写成矩阵形式:

$$\mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n. \tag{3.1.4}$$

把(3.1.3)和(3.1.4)合写在一起, 即可得完整的线性回归模型:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},\ \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n. \tag{3.1.5}$$

用最小二乘方法寻找$\boldsymbol{\beta}$的估计, 这个估计因此被称为最小二乘估计(least squares estimator, LSE). 这个方法是寻找一个$\boldsymbol{\beta}$的估计, 使得误差向量$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$的长度之平方$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$(即模型的误差平方和$\sum_{i=1}^n e_i^2$)达到最小. 在1个自变量的情形下, 最小二乘方法就是寻找一条直线使得样本点与直线的竖直距离之和达到最小(也可以考虑垂直距离, 但竖直距离比垂直距离更容易计算且表达式更简单, 所以采用竖直距离), 参考图3.1.1; 在2个自变量的情形下, 最小二乘方法就是寻找一个二维曲面使得样本点与曲面的竖直距离之和达到最小; 对于更多个自变量的情形, 可类似理解.

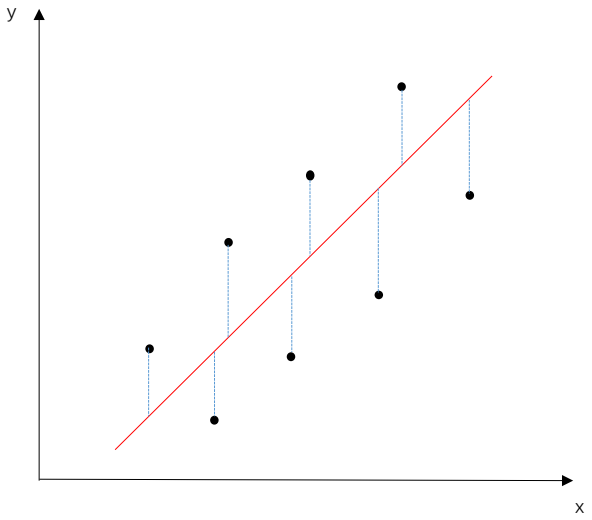


图3.1.1 最小二乘法的几何解释(以1个自变量为例): 样本点与直线的竖直距离之和达到最小

记

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

对$\boldsymbol{\beta}$求导, 令其等于零, 可得方程组

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}. \tag{3.1.6}$$

称这个方程组为正规方程组或正则方程组. 这个方程组有唯一解的充要条件是$\boldsymbol{X}'\boldsymbol{X}$的秩是$p+1$, 这等价于$\boldsymbol{X}$的秩是$p+1$(即$\boldsymbol{X}$是列满秩的). 因为$\boldsymbol{X}$是可以人为控制的, 所以总假定$\boldsymbol{X}$是列满秩的. 于是得到(3.1.6)的唯一解

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.1.7}$$

以上的讨论只能说明$\hat{\boldsymbol{\beta}}$是$Q(\boldsymbol{\beta})$的一个驻点, 但未必就是最小值点. 下面来说明$\hat{\boldsymbol{\beta}}$确实是$Q(\boldsymbol{\beta})$的最小值点. 对任意的$\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, 有

$$\begin{aligned}
&\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \\
=&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2 \\
=&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2+2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}). \qquad (3.1.8)
\end{aligned}$$

因为$\hat{\boldsymbol{\beta}}$满足正规方程组(3.1.6), 所以$\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\boldsymbol{0}$. 这就证明了对任意的$\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, 有

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2=\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2. \qquad (3.1.9)$$

所以,

$$Q(\boldsymbol{\beta})=\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \geq \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2=Q(\hat{\boldsymbol{\beta}}).$$

记$\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)'$, 可得回归方程

$$\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}} \quad 或 \quad \hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p. \qquad (3.1.10)$$

这个方程是不是在平均意义下刻画了y与$x_1,\cdots,x_p$的真实的相依关系, 还需作进一步的统计分析, 留待以后处理.

称$\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=:\boldsymbol{H}\boldsymbol{Y}$为$\boldsymbol{Y}$的拟合值向量或投影向量, 其中

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

被称为帽子矩阵(hat matrix). 此外, 称$\hat{\boldsymbol{e}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}$为残差向量. 投影向量与残差向量的几何解释见图3.1.2.

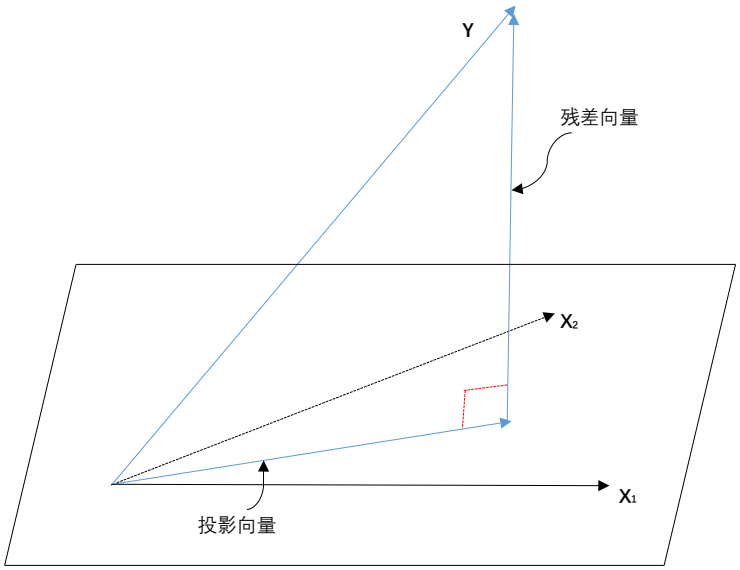


图3.1.2 投影向量和残差向量的几何解释

注: (1) $\boldsymbol{H}$是对称幂等矩阵. $\boldsymbol{I}_n-\boldsymbol{H}$也是对称幂等矩阵. $\boldsymbol{H}$其实是由$\boldsymbol{X}$的$p+1$个列向量所张成的线性空间(即$\boldsymbol{X}$的列空间)的投影矩阵, 而$\boldsymbol{I}_n-\boldsymbol{H}$是这个线性空间的正交补空间的投影矩阵. $\hat{\boldsymbol{Y}}=\boldsymbol{H}\boldsymbol{Y}$是$\boldsymbol{Y}$在这个线性空间上的投影, $\hat{\boldsymbol{e}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}$是$\boldsymbol{Y}$在正交补空间上的投影.

(2) $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$也可由几何方法得到. 注意到$\boldsymbol{Y}-\hat{\boldsymbol{Y}}$与$\boldsymbol{X}$的列向量所张成的空间正交, 所以有$\boldsymbol{X}'(\boldsymbol{Y}-\hat{\boldsymbol{Y}})=0$, 即$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{Y}$.

例3.1.1 一元线性回归. 假设自变量只有一个, 记为x. 样本为$\{(x_i,y_i),\ i=1,\cdots,n\}$. 于是有线性回归模型

$$y_i=\beta_0+\beta_1x_i+e_i,\quad i=1,\cdots,n.$$

这时的正规方程组为

$$\begin{pmatrix} n & \sum\limits_{i=1}^n x_i \\ \sum\limits_{i=1}^n x_i & \sum\limits_{i=1}^n x_i^2 \end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=\begin{pmatrix}\sum\limits_{i=1}^n y_i\\ \sum\limits_{i=1}^n x_iy_i\end{pmatrix}.$$

当设计矩阵$\boldsymbol{X}$是列满秩时, 即$x_i,i=1,\cdots,n$, 不全相等时, $\sum_{i=1}^n(x_i-\bar{x})^2\neq 0$, 于是$\beta_0$和$\beta_1$的LSE为

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x},\\ \hat{\beta}_1=\dfrac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2}=:\dfrac{S_{xy}}{S_{xx}}\end{cases}$$

回归系数的LSE的含义是什么? 先假设模型中没有截距且只有一个自变量, 即

$$y_i=\beta x_i+e_i,\quad i=1,\cdots,n.$$

容易看出β的LSE以及残差向量分别是

$$\hat{\beta}=\frac{\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n x_i^2}=:\frac{\langle\boldsymbol{x},\boldsymbol{Y}\rangle}{\langle\boldsymbol{x},\boldsymbol{x}\rangle},\quad \hat{\boldsymbol{e}}=\boldsymbol{Y}-\boldsymbol{x}\hat{\beta},\tag{3.1.11}$$

其中$\langle\cdot,\cdot\rangle$表示内积运算, $\boldsymbol{x}=(x_1,\cdots,x_n)'$. 若模型中有截距, 此时的模型为

$$y_i=\beta_0+\beta_1x_i+e_i,\quad i=1,\cdots,n.$$

那么β_1的LSE是

$$\hat{\beta}_1=\frac{\langle\boldsymbol{x}-\bar{x}\mathbf{1}_n,\boldsymbol{Y}\rangle}{\langle\boldsymbol{x}-\bar{x}\mathbf{1}_n,\boldsymbol{x}-\bar{x}\mathbf{1}_n\rangle}.$$

可以把$\hat{\beta}_1$看成是进行两次无截距回归(3.1.11)后的结果:

(1) x关于“自变量”$x_0\equiv 1$进行回归, 得到残差变量z, 相应的残差向量为$\boldsymbol{z}=\boldsymbol{x}-\bar{x}\mathbf{1}_n$;

(2) y关于z进行回归, 得到$\hat{\beta}_1$.

把变量b关于变量a进行(无截距)回归称为“b关于a正交化”或者“b被a调整”. 在上面的过程中, 经过第一步后, 残差向量$\boldsymbol{z}$和$\boldsymbol{x}_0=\mathbf{1}_n$(它是“自变量”$x_0\equiv 1$的

观测向量)是相互正交的, 即$\mathbf{1}_n'\boldsymbol{z}=0$. 带截距项的一元线性回归可看成是因变量$y$关于两个自变量$x_0\equiv 1$和$x_1$进行回归. 图3.1.3 描述了$y$关于$x_0$和$x_1$进行回归的过程. x_1关于x_0正交化后得到残差变量z, y关于z进行回归得到x_1的回归系数. 把向量$\boldsymbol{Y}$关于$\boldsymbol{x}_0=\mathbf{1}_n$和$\boldsymbol{z}$分别进行投影, 然后求和, 就得到了$\boldsymbol{Y}$的最小二乘拟合$\hat{\boldsymbol{Y}}$. 需要注意的是, 正交化过程不改变自变量空间, 但$\boldsymbol{x}_0$和$\boldsymbol{z}$构成了这个空间的一组正交基.

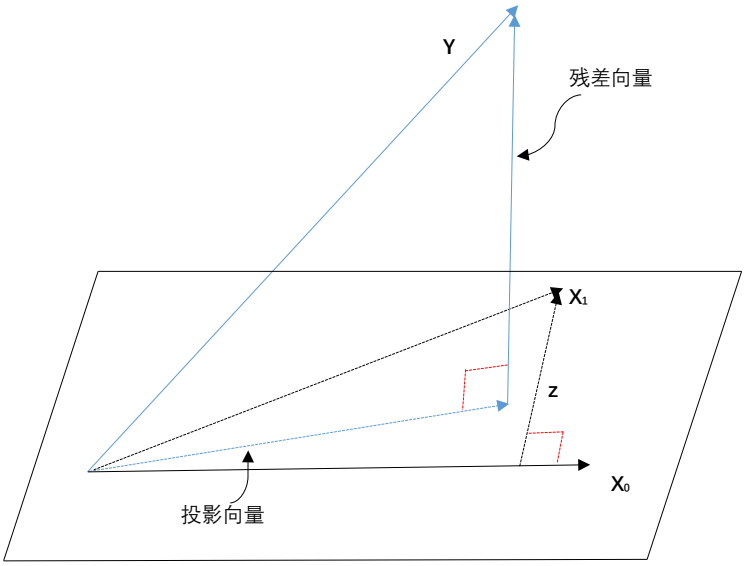


图3.1.3 x_2关于x_1正交化

现在假设有p个自变量: $x_1,\cdots,x_p$. 把上面的过程进行推广, 则得到下面的Gram-Schmidt回归过程:

(1) 令$z_0=x_0\equiv 1$, 相应的观测向量$\boldsymbol{z}_0=\boldsymbol{x}_0=\mathbf{1}_n$;

(2) 对于$j=1,\cdots,p$依次进行以下操作: x_j关于$z_0,z_1,\cdots,z_{j-1}$进行(无截距)回归, 得到回归系数$\hat{\gamma}_{kj}=\langle\boldsymbol{z}_k,\boldsymbol{x}_j\rangle/\langle\boldsymbol{z}_k,\boldsymbol{z}_k\rangle$, $k=0,\cdots,j-1$, 以及残差变量z_j, 相应的残差向量$\boldsymbol{z}_j=\boldsymbol{x}_j-\sum_{k=0}^{j-1}\hat{\gamma}_{kj}\boldsymbol{z}_k$;

(3) y关于残差变量z_p进行(无截距)回归, 得到$\hat{\beta}_p=\frac{\langle\boldsymbol{z}_p,\boldsymbol{Y}\rangle}{\langle\boldsymbol{z}_p,\boldsymbol{z}_p\rangle}$.

在步骤(2)中, 每一个$\boldsymbol{x}_j$是残差向量$\{\boldsymbol{z}_k,k=0,\cdots,j\}$的线性组合. $\boldsymbol{z}_0,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_p$是相互正交的, 构成了自变量空间的一组正交基, 因此$\boldsymbol{Y}$在由$\boldsymbol{z}_0,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_p$所张成的空间上的投影就是$\hat{\boldsymbol{Y}}$.

若对自变量重新排序, 可知任何一个自变量x_j都可成为x_p, 因此多元线性回归方程中的$\hat{\beta}_j$其实是y关于$z_{j:01\cdots(j-1)(j+1)\cdots p}$的(无截距)回归系数估计, 其中$z_{j:01\cdots(j-1)(j+1)\cdots p}$表示$x_j$关于$x_0\equiv 1,x_1,\cdots,x_{j-1},x_{j+1},\cdots,x_p$进行回归后得到的残差变量.

最后, 得到以下的结论: (1) $\hat{\beta}_j$衡量了当x_j关于$x_0\equiv 1,x_1,\cdots,x_{j-1},x_{j+1},\cdots,x_p$正交化后, x_j对y的额外贡献大小; (2) $\hat{\beta}_j$的大小与所有自变量$x_0\equiv 1,x_1,\cdots,x_p$有

关.

由Gram-Schmidt回归过程的步骤(2)可看出

$$x_{ij} = \begin{pmatrix} z_{i0} & z_{i1} & \cdots & z_{i,j-1} & z_{ij} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \hat{\gamma}_{0j} \\ \hat{\gamma}_{1j} \\ \vdots \\ \hat{\gamma}_{j-1,j} \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

因此, 设计矩阵$\boldsymbol{X}$可写为

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma},$$

其中$\boldsymbol{Z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_p)$, 而$\boldsymbol{\Gamma} = (\hat{\gamma}_{kj})$是一个上三角矩阵. 引入$(p+1)\times(p+1)$的对角矩阵$\boldsymbol{D}$, 对角线元素分别为$\|\boldsymbol{z}_j\|$, $j = 0, 1, \cdots, p$. 则

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{\Gamma} =: \boldsymbol{Q}\boldsymbol{R},$$

这就是$\boldsymbol{X}$的所谓QR分解. 这里, $\boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{D}^{-1}$是$n \times (p+1)$的正交矩阵, $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{\Gamma}$是$(p+1) \times (p+1)$的上三角矩阵.

QR分解给出了$\boldsymbol{X}$的列空间的一组非常有用的正交基(即$\boldsymbol{Q}$的各列), 并为最小二乘拟合带来方便. 例如, 容易看出

$$\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{Y}, \ \hat{\boldsymbol{Y}} = \boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{Y}.$$

因此求解$\hat{\boldsymbol{\beta}}$就变得容易了(因为$\boldsymbol{R}$是上三角矩阵).

在回归分析中, 有时会把原始数据进行中心化和标准化. 令

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, \cdots, p.$$

将(3.1.2)改写为

$$y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_p(x_{ip} - \bar{x}_p) + e_i, \quad i = 1, \cdots, n. \tag{3.1.12}$$

这里, $\alpha = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_p\bar{x}_p$. 称(3.1.12)为中心化模型. 记

$$\boldsymbol{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}.$$

则(3.1.12)可改写为

$$\boldsymbol{Y} = \boldsymbol{1}_n\alpha + \boldsymbol{X}_c\boldsymbol{\beta} + \boldsymbol{e} = \begin{pmatrix} \boldsymbol{1}_n & \boldsymbol{X}_c \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{e}. \tag{3.1.13}$$

这里, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)'$. 注意到

$$\mathbf{1}_n'\boldsymbol{X}_c=\mathbf{0},$$

因此正规方程组可写为

$$\begin{pmatrix} n & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_c'\boldsymbol{X}_c \end{pmatrix}\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix}=\begin{pmatrix}\mathbf{1}_n'\boldsymbol{Y}\\ \boldsymbol{X}_c'\boldsymbol{Y}\end{pmatrix}.$$

因此,

$$\begin{cases} n\alpha=\mathbf{1}_n'\boldsymbol{Y},\\ \boldsymbol{X}_c'\boldsymbol{X}_c\boldsymbol{\beta}=\boldsymbol{X}_c'\boldsymbol{Y}. \end{cases}$$

于是回归系数的LSE为

$$\begin{cases} \hat{\alpha}=\bar{y},\\ \hat{\boldsymbol{\beta}}=(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}. \end{cases} \tag{3.1.14}$$

可以看出, 对于中心化线性回归模型(3.1.12), 回归常数的LSE是因变量的样本均值, 而$\boldsymbol{\beta}$的LSE $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}$等价于从线性回归模型$\boldsymbol{Y}=\boldsymbol{X}_c\boldsymbol{\beta}+\boldsymbol{e}$中计算$\boldsymbol{\beta}$的LSE. 在实际应用中, 计算$(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}$总是比计算$(\boldsymbol{X}'\boldsymbol{X})^{-1}$要方便一点, 且我们总是特别关心回归系数, 所以中心化是有好处的.

例3.1.2 一元线性回归(续). 将例3.1.1中的一元线性回归模型进行中心化, 得

$$y_i=\alpha+\beta_1(x_i-\bar{x})+e_i,\quad i=1,\cdots,n. \tag{3.1.15}$$

由公式(3.1.14)得LSE

$$\begin{cases} \hat{\alpha}=\bar{y},\\ \hat{\beta}_1=\dfrac{\sum_{i=1}^n(x_i-\bar{x})y_i}{\sum_{i=1}^n(x_i-\bar{x})^2}=\dfrac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2}. \end{cases}$$

另外, 还可以对自变量做标准化处理. 记

$$\begin{aligned} s_j^2&=\sum_{i=1}^n(x_{ij}-\bar{x}_j)^2,\quad j=1,\cdots,p,\\ z_{ij}&=\frac{x_{ij}-\bar{x}_j}{s_j},\quad i=1,\cdots,n;j=1,\cdots,p. \end{aligned}$$

令$\boldsymbol{Z}=(z_{ij})_{n\times p}$, 它具有性质: (1) $\mathbf{1}_n'\boldsymbol{Z}=\mathbf{0}$; (2) $\boldsymbol{R}=\boldsymbol{Z}'\boldsymbol{Z}=(r_{ij})$, 其中

$$r_{ij}=\frac{\sum_{k=1}^n(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{s_is_j},\quad i,j=1,\cdots,p$$

为自变量x_i与x_j的样本相关系数. 所以$\boldsymbol{R}$是自变量的样本相关系数矩阵. 经过标准化后的线性回归模型为

$$y_i=\gamma+\frac{x_{i1}-\bar{x}_1}{s_1}\beta_1+\cdots+\frac{x_{ip}-\bar{x}_p}{s_p}\beta_p+e_i,$$

或写成矩阵形式:

$$\boldsymbol{Y} = \mathbf{1}_n\gamma + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{e} = \begin{pmatrix}\mathbf{1}_n & \boldsymbol{Z}\end{pmatrix}\begin{pmatrix}\gamma \\ \boldsymbol{\beta}\end{pmatrix} + \boldsymbol{e}.$$

回归参数的LSE为

$$\hat{\gamma} = \bar{y}, \quad \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)' = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.$$

相应的回归方程为

$$\hat{y} = \hat{\gamma} + \frac{x_1 - \bar{x}_1}{s_1}\hat{\beta}_1 + \cdots + \frac{x_p - \bar{x}_p}{s_p}\hat{\beta}_p. \tag{3.1.16}$$

对自变量进行标准化有以下好处: (1) 可以用来分析回归自变量之间的相关关系; (2) 消除了量纲的影响后, 便于对回归系数的估计值进行统计分析.

例3.1.3 一个试验容器靠蒸汽供应热量, 使其保持恒温. 在表3.1.1中, 自变量x表示容器周围空气在单位时间内的平均温度(℃), y表示单位时间内消耗的蒸汽量(L), 共观测了25个时间单位. 图3.1.4是这些数据的散点图, 对这组数据, 应用中心化线性回归模型(3.1.15), 计算得到:

$$\bar{y} = 9.424, \ \bar{x} = 52.6,$$

回归参数的LSE为

$$\hat{\alpha} = \bar{y} = 9.424, \quad \hat{\beta}_1 = -0.080.$$

所以回归方程为

$$\hat{y} = 9.424 - 0.080(x - 52.6),$$

或写成

$$\hat{y} = 13.623 - 0.080x.$$

序号	$y(L)$	x(℃)	序号	$y(L)$	x(℃)
1	10.98	35.3	14	9.57	39.1
2	11.13	29.7	15	10.94	46.8
3	12.51	30.8	16	9.58	48.5
4	8.40	58.8	17	10.09	59.3
5	9.27	61.4	18	8.11	70
6	8.73	71.3	19	6.83	70
7	6.36	74.4	20	8.88	74.5
8	8.5	76.7	21	7.68	72.1
9	7.82	70.7	22	8.47	58.1
10	9.14	57.5	23	8.86	44.6
11	8.24	46.4	24	10.36	33.4
12	12.19	28.9	25	11.08	28.6
13	11.88	28.1			

表3.1.1 蒸汽数据

R代码及分析结果如下:

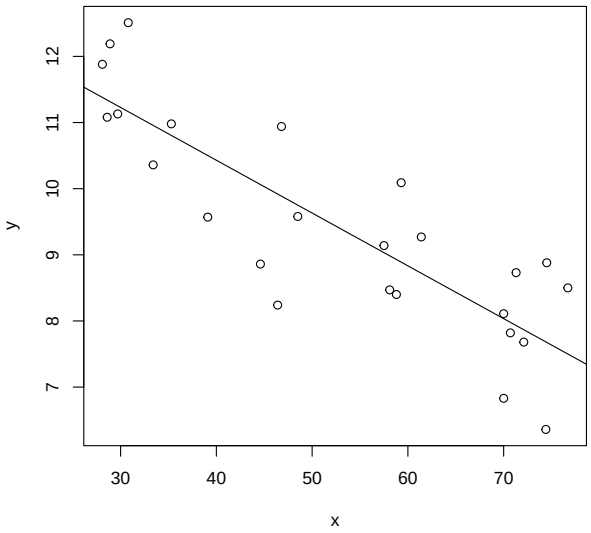


图3.1.4 散点图与回归直线

```
> yx=read.table(***.txt)   #读取数据到yx
> y=yx[,1]   #把yx的第1列赋值给y
> x=yx[,2]   #把yx的第2列赋值给x
> steam=data.frame(y,x)   #建立数据框
> plot(y~x)   #画散点图，等价于plot(x,y)
> lm.sol=lm(y~x,data=steam)   #对数据steam进行线性回归
> abline(lm.sol)   #画回归直线
> summary(lm.sol)   #提取线性回归结果

Call:
lm(formula = y ~ x, data = steam)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6789 -0.5291 -0.1221  0.7988  1.3457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
x           -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’ . 0.1 ‘ ’ 1

```

```
Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-squared:  0.7144,    Adjusted R-squared:  0.702
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07
```

3.2 最小二乘估计的性质

最小二乘估计具有一些良好的性质.

定理3.2.1 对于线性回归模型(3.1.5), $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$具有下列性质:

(a) $\mathsf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$;

(b) $\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

证明: (a) 易知$\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, 所以

$$\mathsf{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \cdot \mathsf{E}(\boldsymbol{Y}) = \boldsymbol{\beta}.$$

(b) 因为$\mathsf{Cov}(\boldsymbol{Y}) = \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$, 所以

$$\begin{aligned}\mathsf{Cov}(\hat{\boldsymbol{\beta}}) =& \mathsf{Cov}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}) \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

证毕. □

设$\boldsymbol{c}$是p+1维的常数向量, 对于线性函数$\boldsymbol{c}'\boldsymbol{\beta}$(这是一个未知参数), 称$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$为$\boldsymbol{c}'\boldsymbol{\beta}$的LSE.

推论3.2.1 对于线性回归模型(3.1.5), $\boldsymbol{c}'\boldsymbol{\beta}$具有下列性质:

(a) $\mathsf{E}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \boldsymbol{c}'\boldsymbol{\beta}$;

(b) $\mathsf{Cov}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$.

即对任意的线性函数$\boldsymbol{c}'\boldsymbol{\beta}$, $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$为$\boldsymbol{c}'\boldsymbol{\beta}$的无偏估计, 方差为$\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$. 因为$\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$是$y_1, \cdots, y_n$的线性函数, 所以$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$为$\boldsymbol{c}'\boldsymbol{\beta}$的一个线性无偏估计.[1] 还可以构造出$\boldsymbol{c}'\boldsymbol{\beta}$的其它线性无偏估计. 这构成了$\boldsymbol{c}'\boldsymbol{\beta}$的线性无偏估计类.

定理3.2.2(Gauss-Markov) 对于线性回归模型(3.1.5), 在$\boldsymbol{c}'\boldsymbol{\beta}$的所有线性无偏估计中, 最小二乘估计$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$是唯一的最小方差线性无偏估计(best linear unbiased estimator, BLUE).

证明: 设$\boldsymbol{a}'\boldsymbol{Y}$为$\boldsymbol{c}'\boldsymbol{\beta}$的一个线性无偏估计. 于是对任意的$\boldsymbol{\beta} \in \mathbb{R}^{p+1}$,

$$\boldsymbol{c}'\boldsymbol{\beta} = \mathsf{E}(\boldsymbol{a}'\boldsymbol{Y}) = \boldsymbol{a}'\boldsymbol{X}\boldsymbol{\beta}.$$

因此

$$\boldsymbol{a}'\boldsymbol{X} = \boldsymbol{c}'. \tag{3.2.1}$$

注意到$\mathsf{Var}(\boldsymbol{a}'\boldsymbol{Y}) = \sigma^2\boldsymbol{a}'\boldsymbol{a} = \sigma^2\|\boldsymbol{a}\|^2$, 故对$\|\boldsymbol{a}\|^2$做分解:

$$\|\boldsymbol{a}\|^2 = \|\boldsymbol{a} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|^2$$

[1] 线性估计是指估计量是$y_1, \cdots, y_n$的线性函数.

$$
\begin{aligned}
=&\|\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|^2+\|\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|^2\\
&+2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}).
\end{aligned}\tag{3.2.2}
$$

由(3.2.1)可知

$$
2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c})=2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{a}-\boldsymbol{c})=0.
$$

由推论3.2.1(b)可知

$$
\begin{aligned}
\sigma^2\|\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|^2=&\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\\
=&\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\\
=&\mathsf{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}).
\end{aligned}
$$

于是, 对$\boldsymbol{c}'\boldsymbol{\beta}$的任一个线性无偏估计$\boldsymbol{a}'\boldsymbol{Y}$有

$$
\mathsf{Var}(\boldsymbol{a}'\boldsymbol{Y})=\mathsf{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})+\sigma^2\|\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|^2\geq\mathsf{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}),\tag{3.2.3}
$$

等号成立当且仅当$\|\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\|=0$, 即$\boldsymbol{a}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$, 此时$\boldsymbol{a}'\boldsymbol{Y}=\boldsymbol{c}'\hat{\boldsymbol{\beta}}$. 定理得证. □

在线性回归模型(3.1.5)中还有一个未知参数σ^2(误差方差), 它的大小反映了模型误差对因变量的影响大小, 在回归分析中起着重要的作用. 现在来估计σ^2.

$\boldsymbol{e}=\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}$是误差向量, 不可观测. 用$\hat{\boldsymbol{\beta}}$代替$\boldsymbol{\beta}$, 称

$$
\hat{\boldsymbol{e}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}\tag{3.2.4}
$$

为残差(residual)向量. 记$\boldsymbol{x}_i'$为设计矩阵$\boldsymbol{X}$的第i行, 则

$$
\hat{e}_i=y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}=y_i-\hat{y}_i,\quad i=1,\cdots,n\tag{3.2.5}
$$

为第i个残差. 称$\hat{y}_i$为第i次观测的拟合值, $\hat{\boldsymbol{Y}}$为拟合值向量. 自然地, 将$\hat{\boldsymbol{e}}$看作$\boldsymbol{e}$的近似. 下面用

$$
\mathrm{RSS}:=\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}=\sum_{i=1}^{n}\hat{e}_i^2\tag{3.2.6}
$$

来构造σ^2的一个无偏估计量. RSS表示残差平方和(residual sum of squares). 它从整体上反映了观测数据与回归直线的偏离程度, 也可以说是度量了观测数据与线性模型(3.1.5)的吻合程度. RSS越小表示数据与模型的吻合度越高.

定理3.2.3 对于线性回归模型(3.1.5),

(a) $\mathrm{RSS}=\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}=:\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}$;

(b) $\hat{\sigma}^2=\mathrm{RSS}/(n-\mathrm{rk}(\boldsymbol{X}))$是$\sigma^2$的无偏估计量.

证明: (a) 因为$\hat{\boldsymbol{e}}=(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}$, 所以

$$
\begin{aligned}
\mathrm{RSS}=&\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}=\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{H})'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}=\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}\\
=&\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{e}.
\end{aligned}
$$

(b) 注意到$\mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n$, 由定理2.2.1可知

$$
\mathsf{E}(\mathrm{RSS})=\mathsf{E}[\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{e}]
$$

$$=0+\mathrm{tr}[(\boldsymbol{I}_n-\boldsymbol{H})\cdot\sigma^2\boldsymbol{I}_n]$$
$$=\sigma^2(n-\mathrm{tr}(\boldsymbol{H})).$$

根据对称幂等矩阵的秩与迹相等这一性质,

$$\mathrm{tr}(\boldsymbol{H})=\mathrm{rk}(\boldsymbol{H})=\mathrm{rk}(\boldsymbol{X}).$$

于是$\mathsf{E}(\mathrm{RSS})=\sigma^2(n-\mathrm{rk}(\boldsymbol{X}))$, 即$\mathrm{RSS}/[n-\mathrm{rk}(\boldsymbol{X})]$是$\sigma^2$的一个无偏估计量. □

如果误差向量$\boldsymbol{e}$服从正态分布, 即$\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, 那么可以得到关于$\hat{\boldsymbol{\beta}}$和$\hat{\sigma}^2$更好的性质.

定理3.2.4 对于线性回归模型(3.1.5), 若误差向量$\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, 则
(a) $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$;
(b) $\mathrm{RSS}/\sigma^2\sim\chi^2(n-\mathrm{rk}(\boldsymbol{X}))$;
(c) $\hat{\boldsymbol{\beta}}$与RSS相互独立.

证明: (a) 注意到$\boldsymbol{Y}\sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$以及$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$是$\boldsymbol{Y}$的线性变换, 那么由定理2.3.4便可推得(a).

(b) 由于

$$\frac{\mathrm{RSS}}{\sigma^2}=\left(\frac{\boldsymbol{e}}{\sigma}\right)'(\boldsymbol{I}_n-\boldsymbol{H})\left(\frac{\boldsymbol{e}}{\sigma}\right),$$

并注意到$\boldsymbol{e}/\sigma\sim N(\boldsymbol{0},\boldsymbol{I}_n)$, 所以根据定理2.4.3, 只需证明$\boldsymbol{I}_n-\boldsymbol{H}$的秩为$n-\mathrm{rk}(\boldsymbol{X})$. 利用对称幂等矩阵的秩与迹相等这一性质, 可得

$$\mathrm{rk}(\boldsymbol{I}_n-\boldsymbol{H})=\mathrm{tr}(\boldsymbol{I}_n-\boldsymbol{H})=n-\mathrm{rk}(\boldsymbol{H})=n-\mathrm{rk}(\boldsymbol{X}).$$

(c) 因为$\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$, 而$\mathrm{RSS}=\boldsymbol{e}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{e}$, 注意到

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\cdot\sigma^2\boldsymbol{I}_n\cdot(\boldsymbol{I}_n-\boldsymbol{H})=\boldsymbol{0},$$

所以由推论2.4.10可知$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$与RSS相互独立, 即$\hat{\boldsymbol{\beta}}$与RSS相互独立. □

当$\boldsymbol{\beta}$的第一个分量为β_0时, 取$\boldsymbol{c}=(0,\cdots,0,1,0,\cdots,0)'$, 其中1在$\boldsymbol{c}$的第$i+1$个位置, 则$\boldsymbol{c}'\boldsymbol{\beta}=\beta_i$. 再记$\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\cdots,\hat{\beta}_p)'$, 于是$\boldsymbol{c}'\hat{\boldsymbol{\beta}}=\hat{\beta}_i$. 用$(\boldsymbol{A})_{ij}$表示矩阵$\boldsymbol{A}$的第$(i,j)$元素, 那么有如下推论.

推论3.2.2 对于线性回归模型(3.1.5), 若$\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, 则
(a) $\hat{\beta}_i\sim N(\beta_i,\sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1})_{i+1,i+1})$;
(b) 在β_i的一切线性无偏估计中, $\hat{\beta}_i$是唯一的方差最小者, $i=1,\cdots,p$.

将定理3.2.1和定理3.2.4应用于中心化模型(3.1.13), 则有下面的推论.

推论3.2.3 对于中心化模型(3.1.13), 注意这里的$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)'$, 有
(a) $\mathsf{E}(\hat{\alpha})=\alpha,\mathsf{E}(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$, 这里$\hat{\alpha}=\bar{y}$, $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}$;
(b)

$$\mathsf{Cov}\begin{pmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}=\sigma^2\begin{pmatrix}\frac{1}{n} & \boldsymbol{0}\\ \boldsymbol{0} & (\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\end{pmatrix};$$

(c) 若进一步假设$\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, 则

$$\hat{\alpha}\sim N(\alpha,\frac{\sigma^2}{n}),\quad \hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}),$$

且$\hat{\alpha}$与$\hat{\boldsymbol{\beta}}$相互独立.

定义

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}}, \tag{3.2.7}$$

其中

$$\mathrm{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{y})'(\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{y})$$

被称为回归平方和(或解释平方和: explained sum of squares),

$$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (\boldsymbol{Y} - \mathbf{1}_n\bar{y})'(\boldsymbol{Y} - \mathbf{1}_n\bar{y})$$

被称为总偏差平方和(或总平方和: total sum of squares). 称R^2为判定系数或测定系数, $R = \sqrt{R^2}$为复相关系数.

前面还出现过一个平方和, RSS. 为了了解TSS, ESS, RSS三者之间的关系, 考察正规方程组$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$的一个等价写法. 对目标函数

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

关于$\beta_0, \beta_1, \cdots, \beta_p$分别求偏导数, 并令这些导函数为0, 可得

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) = 0, \\ \sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})x_{i1} = 0, \\ \qquad\qquad \vdots \\ \sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})x_{ip} = 0. \end{cases}$$

这个方程组与$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$等价, 所以也是正规方程组. 由于最小二乘估计$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$是正规方程组的解, 所以

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \\ \sum\limits_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})x_{i1} = 0, \\ \qquad\qquad \vdots \\ \sum\limits_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})x_{ip} = 0. \end{cases}$$

即,

$$\begin{cases} \sum_{i=1}^{n}(y_i-\hat{y}_i)=0, \\ \sum_{i=1}^{n}(y_i-\hat{y}_i)x_{i1}=0, \\ \qquad\vdots \\ \sum_{i=1}^{n}(y_i-\hat{y}_i)x_{ip}=0. \end{cases} \tag{3.2.8}$$

由(3.2.8)的第一个等式可知:

$$\sum_{i=1}^{n}\hat{e}_i=0,\quad \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\bar{y}.$$

所以ESS $=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$其实是$\hat{y}_1,\cdots,\hat{y}_n$的偏差平方和. 而$\hat{y}_1,\cdots,\hat{y}_n$是通过回归方程计算出来的, 这也是ESS被称为回归平方和的原因.

应用(3.2.8)可以证明

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.\end{aligned}$$

即,

$$\text{TSS}=\text{ESS}+\text{RSS}.$$

因此, 容易看出: $0\le R^2\le 1$. R^2有什么用呢?

若模型中没有任何自变量, 即$y_i=\beta_0+e_i$, $i=1,\cdots,n$, 则$\bar{y}$为β_0的LSE, TSS为模型的残差平方和. 若在模型中引入自变量$x_1,\cdots,x_p$, 此时的残差平方和为平方和分解式TSS = ESS + RSS中的RSS. 即, ESS的大小衡量了当模型中引入p个自变量后, 残差平方和的减少量. 因此, R^2 = ESS/TSS衡量了当模型中引入p个自变量后, 残差平方和减少的比例. 也可以说: R^2衡量了自变量$x_1,\cdots,x_p$对y的解释能力.

R^2多大才可接受? 这需要具体问题具体分析. 在生物学以及社会科学领域, 变量之间的相关性往往较弱, 模型里的噪音很大, 因此我们预期R^2的值较低. 例如, $R^2=0.6$可能被认为是可接受的. 而在物理学和工程学中, 大部分数据是来自严格控制的实验, 我们通常期望能得到更大的R^2. 此时$R^2=0.6$将被视为过小了. 在特定领域有一定的经验是必要的, 这对你判断R^2是否可接受是有帮助的.

若使用中心化模型(3.1.13), 那么ESS可通过下列公式计算:

$$\text{ESS}=\hat{\boldsymbol{\beta}}'\boldsymbol{X}_c'\boldsymbol{Y}=\boldsymbol{Y}'\boldsymbol{X}_c(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y},$$

这里$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$. 事实上, 由$\hat{\boldsymbol{Y}} = \mathbf{1}_n\hat{\alpha} + \boldsymbol{X}_c\hat{\boldsymbol{\beta}}$及公式(3.1.14)可知

$$\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{y} = \hat{\boldsymbol{Y}} - \mathbf{1}_n\hat{\alpha} = \boldsymbol{X}_c\hat{\boldsymbol{\beta}}.$$

所以,

$$\begin{aligned}\text{ESS} =&(\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{y})'(\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{y})\\ =&(\boldsymbol{X}_c\hat{\boldsymbol{\beta}})'\boldsymbol{X}_c\hat{\boldsymbol{\beta}}\\ =&\hat{\boldsymbol{\beta}}'\boldsymbol{X}_c'\cdot\boldsymbol{X}_c(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}\\ =&\hat{\boldsymbol{\beta}}'\boldsymbol{X}_c'\boldsymbol{Y}.\end{aligned}$$

对于一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n,$$

可以证明$R^2 = r^2$, 其中r为自变量x与因变量y的样本相关系数. 事实上, 因为$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$, 所以

$$\text{ESS} = \sum_{i=1}^n(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2\sum_{i=1}^n(x_i - \bar{x})^2 = \frac{\left[\sum_{i=1}^n(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n(x_i - \bar{x})^2}.$$

所以

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\left[\sum_{i=1}^n(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^n(x_i - \bar{x})^2\sum_{i=1}^n(y_i - \bar{y})^2} = r^2.$$

对于一般的线性回归模型,

$$R^2 = \text{corr}^2(\boldsymbol{Y}, \hat{\boldsymbol{Y}}),$$

这里$\text{corr}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$表示观测向量$\boldsymbol{Y}$与拟合值向量$\hat{\boldsymbol{Y}}$的样本相关系数. 事实上,

$$\begin{aligned}\text{corr}^2(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) &= \frac{[\sum_{i=1}^n(y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_{i=1}^n(y_i - \bar{y})^2\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}\\ &= \frac{[\sum_{i=1}^n(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_{i=1}^n(y_i - \bar{y})^2\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}\\ &= \frac{[\sum_{i=1}^n(\hat{y}_i - \bar{y})^2]^2}{\sum_{i=1}^n(y_i - \bar{y})^2\sum_{i=1}^n(\hat{y}_i - \bar{y})^2}\\ &= \frac{\text{ESS}^2}{\text{TSS}\cdot\text{ESS}} = \frac{\text{ESS}}{\text{TSS}} = R^2.\end{aligned}$$

例3.2.1　根据经验知, 在人的身高相等的条件下, 其血压的收缩压y与体重x_1、年龄x_2有关. 现收集了13名男子的测量数据, 见表3.2.1. 试建立y关于x_1和x_2的线性回归方程.

序号	x_{i1}	x_{i2}	y_i	序号	x_{i1}	x_{i2}	y_i
1	152	50	120	8	158	50	125
2	183	20	141	9	170	40	132
3	171	20	124	10	153	55	123
4	165	30	126	11	164	40	132
5	158	30	117	12	190	40	155
6	161	50	125	13	185	20	147
7	149	60	123				

表3.2.1 血压数据

利用中心化模型

$$y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + e_i, \quad i = 1, \cdots, 13,$$

经计算可得

$$\bar{x}_1 = \frac{1}{13}\sum_{i=1}^{13} x_{i1} = 166.8,\ \bar{x}_2 = \frac{1}{13}\sum_{i=1}^{13} x_{i2} = 38.85,\ \bar{y} = \frac{1}{13}\sum_{i=1}^{13} y_i = 130,$$

$$\boldsymbol{X}_c = \begin{pmatrix} -14.08 & 11.15 \\ 16.92 & -18.85 \\ 4.92 & -18.85 \\ -1.08 & -8.85 \\ -8.08 & -8.85 \\ -5.08 & 11.15 \\ -17.08 & 21.15 \\ -8.08 & 11.15 \\ 3.92 & 1.15 \\ -13.08 & 16.15 \\ -2.08 & 1.15 \\ 23.92 & 1.15 \\ 18.92 & -18.85 \end{pmatrix},$$

正规方程组$\boldsymbol{X}_c'\boldsymbol{X}_c\boldsymbol{\beta} = \boldsymbol{X}_c'\boldsymbol{Y}$为

$$\begin{cases} 2078.92\beta_1 - 1533.85\beta_2 = 1607, \\ -1533.85\beta_1 + 2307.69\beta_2 = -715. \end{cases}$$

解得$\hat{\beta}_1 = 1.068,\ \hat{\beta}_2 = 0.4$, 而$\hat{\alpha} = \bar{y} = 130$. 所以回归方程为

$$\begin{aligned} \hat{y} =& \hat{\alpha} + \hat{\beta}_1(x_1 - \bar{x}_1) + \hat{\beta}_2(x_2 - \bar{x}_2) \\ =& 130 + 1.068 \times (x_1 - 166.8) + 0.4 \times (x_2 - 38.85) \\ =& -62.963 + 1.068x_1 + 0.4x_2. \end{aligned}$$

此外, 还可算得

$$\mathrm{ESS} = \hat{\boldsymbol{\beta}}' \boldsymbol{X}_c' \boldsymbol{Y} = 1430.276,\ \mathrm{TSS} = 1512,$$

所以$R^2 = 1430.276/1512 = 0.9459$.

若是使用统计软件进行回归分析, 则没有必要进行中心化或者标准化. R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x1=yx[,1]
> x2=yx[,2]
> y=yx[,3]
> blood_pressure=data.frame(y,x1,x2)
> lm.sol=lm(y~x1+x2,data=blood_pressure)
> summary(lm.sol)

Call:
lm(formula = y ~ x1 + x2, data = blood_pressure)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0404 -1.0183  0.4640  0.6908  4.3274

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.96336   16.99976  -3.704 0.004083 **
x1            1.06828    0.08767  12.185 2.53e-07 ***
x2            0.40022    0.08321   4.810 0.000713 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’ . 0.1 ‘ ’ 1

Residual standard error: 2.854 on 10 degrees of freedom
Multiple R-squared:  0.9461,    Adjusted R-squared:  0.9354
F-statistic: 87.84 on 2 and 10 DF,  p-value: 4.531e-07
```

由输出结果得回归方程与R^2:

$$\hat{y} = -62.963 + 1.068x_1 + 0.400x_2, \quad R^2 = 0.9461.$$

在这个回归方程中, $\hat{\beta}_1 = 1.068$是回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{3.2.9}$$

中β_1的最小二乘估计, 这个大小有什么含义? 为了理解它, 先让y关于x_1进行回归建模:

$$y = \beta_0 + \beta_1 x_1 + e. \tag{3.2.10}$$

用R进行分析,

```
> summary(lm(y~x1,data=blood_pressure))

Call:
lm(formula = y ~ x1, data = blood_pressure)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8055 -2.0815  0.8814  3.1084  6.5075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6231    18.0911   0.090     0.93
x1            0.7730     0.1086   7.117 1.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '' . 0.1 ' ' 1

Residual standard error: 4.952 on 11 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.8053
F-statistic: 50.65 on 1 and 11 DF,  p-value: 1.951e-05
```

可以发现x_1的回归系数的最小二乘估计不再是1.068, 而是0.773了. 所以我们不能说: 当用x_1和x_2作为自变量进行回归建模时(即, 对于模型(3.2.9)), x_1每变动一个单位, y将大约变动1.068个单位. 正确的说法是: 当x_2固定不变时, x_1每变动一个单位, y将大约变动1.068个单位. 而对于模型(3.2.10), 因为只有一个自变量, 才可以直接说: x_1每变动一个单位, y将大约变动0.773个单位.

3.3 约束最小二乘估计

对$\boldsymbol{\beta}$不加任何约束条件的情形下, 前面讨论了它的LSE以及它的基本性质. 但在一些特殊场合, 例如假设检验问题(把“假设”看成“约束”), 需要求带有约束条件的LSE.

假设

$$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b} \tag{3.3.1}$$

是一个相容线性方程组(即方程组有解), 其中$\boldsymbol{A}$是$k \times (p+1)$的已知矩阵, 秩为k, $\boldsymbol{b}$是$k \times 1$的已知向量.

下面用Lagrange乘子法求线性回归模型(3.1.5)在满足线性约束(3.3.1)时的LSE. 即在(3.3.1)这个条件下寻找$\boldsymbol{\beta}$使得

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

达到最小. 构造辅助函数

$$F(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\lambda}'(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{b})$$

$$=(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+2\boldsymbol{\lambda}'(\boldsymbol{A}\boldsymbol{\beta}-\boldsymbol{b}),$$

这里$\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_k)'$为Lagrange乘子向量. 对函数$F(\boldsymbol{\beta},\boldsymbol{\lambda})$关于$\boldsymbol{\beta}$求导并令它等于$\mathbf{0}$, 得

$$-\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{A}'\boldsymbol{\lambda}=\mathbf{0}. \tag{3.3.2}$$

现在, 需要解由(3.3.1)和(3.3.2)组成的联立方程组. 用$\hat{\boldsymbol{\beta}}_c$和$\hat{\boldsymbol{\lambda}}_c$表示这个方程组的解. 由(3.3.2)得

$$\hat{\boldsymbol{\beta}}_c=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c=\hat{\boldsymbol{\beta}}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c. \tag{3.3.3}$$

代入(3.3.1)得

$$\boldsymbol{b}=\boldsymbol{A}\hat{\boldsymbol{\beta}}_c=\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c,$$

等价地写成

$$\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c=\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{b}. \tag{3.3.4}$$

由于$\boldsymbol{A}$的秩为k, 所以$\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$是$k\times k$可逆矩阵, 因此(3.3.4)有唯一解

$$\hat{\boldsymbol{\lambda}}_c=(\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{b}).$$

再将它代入(3.3.3)得

$$\hat{\boldsymbol{\beta}}_c=\hat{\boldsymbol{\beta}}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{b}). \tag{3.3.5}$$

接下来说明$\hat{\boldsymbol{\beta}}_c$确实是线性约束$\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$下$\boldsymbol{\beta}$的LSE. 这只需证明:

(a) $\boldsymbol{A}\hat{\boldsymbol{\beta}}_c=\boldsymbol{b}$;

(b) 对一切满足$\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$的$\boldsymbol{\beta}$, 都有$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\geq\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c\|^2$.

由(3.3.5)可推得(a). 为证(b), 将$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$分解得

$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2=&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\\=&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c+\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta})\|^2\\=&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c)\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta})\|^2,\end{aligned} \tag{3.3.6}$$

这里的推导用到了下述关系: $(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X}=\mathbf{0}$以及对一切满足$\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$的$\boldsymbol{\beta}$,

$$(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta})=\hat{\boldsymbol{\lambda}}_c'\boldsymbol{A}(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta})=\hat{\boldsymbol{\lambda}}_c'(\boldsymbol{A}\hat{\boldsymbol{\beta}}_c-\boldsymbol{A}\boldsymbol{\beta})=\mathbf{0}.$$

(3.3.6)表明: 对一切满足$\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$的$\boldsymbol{\beta}$, 总有

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\geq\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c)\|^2, \tag{3.3.7}$$

等号成立当且仅当$\boldsymbol{X}(\hat{\boldsymbol{\beta}}_c-\boldsymbol{\beta})=\mathbf{0}$, 即$\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_c$(因为$\boldsymbol{X}$列满秩). 因此在(3.3.7)中用$\hat{\boldsymbol{\beta}}_c$代替$\boldsymbol{\beta}$, 等号成立. 即,

$$\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_c\|^2=\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_c)\|^2. \tag{3.3.8}$$

现在, 结合(3.3.7)和(3.3.8)便可推得(b). □

把$\hat{\boldsymbol{\beta}}_c$称为$\boldsymbol{\beta}$的约束最小二乘估计, 于是有下面的定理.

定理3.3.1 对于线性回归模型(3.1.5)中的$\boldsymbol{\beta}$, 满足(3.3.1)的约束LSE为

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'(\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{b}),$$

其中$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$是无约束条件下的LSE.

例3.3.1 在天文测量中, 对天空中的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, 样本为$y_1, y_2, y_3$. 由于存在测量误差, 所以需要对$\theta_1, \theta_2, \theta_3$进行估计, 利用线性模型表示有关的量:

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

把它写成矩阵形式:

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \\ \boldsymbol{A}\boldsymbol{\beta} = b, \end{cases}$$

这里$\boldsymbol{Y} = (y_1, y_2, y_3)'$, $\boldsymbol{\beta} = (\theta_1, \theta_2, \theta_3)'$, $\boldsymbol{X} = \boldsymbol{I}_3$, $\boldsymbol{A} = (1, 1, 1)$, $b = \pi$. 注意到$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{Y}$, 应用定理3.3.1,

$$\hat{\boldsymbol{\beta}}_c = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} - \frac{1}{3}(\sum_{i=1}^{3} y_i - \pi) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

即

$$\hat{\theta}_i = y_i - \frac{1}{3}(\sum_{i=1}^{3} y_i - \pi)$$

为θ_i的约束LSE.

3.4 回归诊断

对于线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},\ \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n, \tag{3.4.1}$$

前面在讨论未知参数的最小二乘估计以及估计量的分布时, 作了如下的基本假设:

(a) 线性假设: 因变量与自变量具有线性相关关系;
(b) 方差齐性假设: $\mathsf{Var}(e_i) = \sigma^2,\ i = 1, \cdots, n$;
(c) 不相关性假设: $\mathsf{Cov}(e_i, e_j) = 0,\ i \neq j$;
(d) 正态性假设: $e_i \sim N(0, \sigma^2),\ i = 1, \cdots, n$.

如果这些假设不成立, 那么前面讨论的最小二乘估计以及它的统计性质就有可

能是不成立的. 因此, 在实际问题中, 当有了一批数据后, 需要考察手上的数据是否满足或者基本满足这些假设. 这是模型诊断的内容.

因为这些假设都与随机误差$\boldsymbol{e}$有关, 而残差向量$\hat{\boldsymbol{e}}$可看成是$\boldsymbol{e}$的近似, 因此可以通过残差来分析上述四个基本假设是否成立. 正因为这个原因, 这部分内容也被称为残差分析.

除了需要对模型进行诊断, 还需要对数据本身进行诊断, 它包含异常点诊断和强影响点诊断. 在回归分析中, 所谓异常点(又叫离群点)是指对既定模型偏离很大的数据点. 给它下一个准确的定义是困难的, 至今还没有统一的定义. 目前, 对异常点较为流行的看法是: (1) 异常点是指那些与绝大多数数据点明显不协调的数据点; (2) 异常点就是那些污染点, 即与绝大多数数据点不是来自同一分布的个别数据点. 异常点的混入将对参数的估计造成影响, 因此需要检测出异常点并将它删除.

数据集中的强影响点是指对统计分析结果(参数估计、假设检验等)产生较大影响的数据点, 在这里, 特指对回归参数$\boldsymbol{\beta}$的最小二乘估计有较大影响的数据点. 对每一组数据点$(\boldsymbol{x}_i', y_i)$, 我们希望它对回归参数的估计有一定的影响, 但又希望这种影响不能太大. 这样, 得到的回归方程就有一定的稳定性. 否则, 如果个别一两组数据对估计有异常大的影响, 当剔除这些数据之后, 就将得到与原来差异很大的回归方程. 因此在回归分析中, 需要考察每组数据对参数估计的影响大小, 这部分内容被称为影响分析. 检测出强影响点后, 需要核查这些数据是否正常, 若不正常则删除之, 否则考虑收集更多的数据或采用一些稳健估计方法以稀释/缩小强影响点对估计的影响.

异常点和强影响点是两个不同的概念. 从后面的分析可以看出, 它们之间既有一定的联系也有一定的区别. 强影响点可能同时是异常点也可能不是. 反之, 异常点可能同时又是强影响点也可能不是.

因此, 回归诊断包含数据的诊断与模型的诊断这两部分内容. 其中, 数据的诊断包括异常点诊断和强影响点诊断, 模型的诊断包括线性诊断、方差齐性诊断、不相关性诊断和正态性诊断.

先来讨论残差分析. 对于线性回归模型(3.4.1), 在前文中用

$$\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \tag{3.4.2}$$

表示第i个残差, 并把它看成e_i的一个近似. 若模型(3.4.1)是正确的, 那么$\hat{e}_i$应具有e_i的一些性状. 回忆$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$为拟合值向量, $\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$为第$i$个拟合值,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}, \tag{3.4.3}$$

其中$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$是帽子矩阵, 它是对称幂等矩阵. 残差向量$\hat{\boldsymbol{e}}$可被表示为

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e}. \tag{3.4.4}$$

$\boldsymbol{I}_n - \boldsymbol{H}$也是一个对称幂等矩阵.

定理3.4.1 对于线性回归模型(3.4.1)成立,

(a) $\mathsf{E}(\hat{\boldsymbol{e}}) = \boldsymbol{0},\ \mathsf{Cov}(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})$;

(b) $\mathsf{Cov}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{e}}) = \boldsymbol{0}$.

(c) 若进一步假设$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, 则$\hat{\boldsymbol{e}} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}))$.

证明: 留给读者自行证明. □

因为$\mathsf{Var}(\hat{e}_i) = \sigma^2(1-h_{ii})$($h_{ii}$表示矩阵$\boldsymbol{H}$的第$i$个对角线元素), 非齐性, 这有碍于$\hat{e}_i$的实际应用. 因此考虑所谓的学生化残差

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \quad i = 1, \cdots, n, \tag{3.4.5}$$

这里$\hat{\sigma}^2 = \mathrm{RSS}/(n - \mathrm{rk}(\boldsymbol{X})) = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n - \mathrm{rk}(\boldsymbol{X}))$.

即使在$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$的条件下, r_i的分布仍然比较复杂, 但可近似地认为r_i相互独立且服从$N(0,1)$(见陈希孺与王松桂的《近代回归分析》). 由定理3.4.1可知$\{r_i, i \geq 1\}$与$\{\hat{y}_i, i \geq 1\}$独立($\{\hat{e}_i, i \geq 1\}$与$\{\hat{y}_i, i \geq 1\}$也独立).

残差图是以某种残差(学生化残差r_i或普通残差$\hat{e}_i$)为纵坐标, 以任何其它的变量为横坐标的散点图. 前已指出残差$\hat{e}_i$作为误差e_i的近似应与e_i相差不远, 故根据残差图的性状是否与应有的性质相一致, 就可以对模型假设的合理性提供一些有用的判断信息.

下面以拟合值$\hat{y}_i$为横坐标, 学生化残差r_i为纵坐标的残差图为例讨论残差图的具体应用. 值得一提的是, 通常情况下, 以普通残差$\hat{e}_i$为纵坐标和以学生化残差r_i为纵坐标的残差图形状大致相同, 以某个自变量x_j为横坐标或者以序号i为横坐标和以拟合值$\hat{y}_i$为横坐标的残差图形状也大致相同.

线性诊断. 若线性假设成立, 那么e_i不包含来自自变量的任何信息, 因此残差图不应呈现任何有规则的形状, 否则有理由怀疑线性假设不成立.

方差齐性诊断. 若方差齐性, 那么残差图上的点是“均匀”散布的. 否则, 残差图通常将呈现“喇叭型”或“倒喇叭型”或两者兼而有之等形状.

不相关性诊断. 若不相关性成立, 那么残差图上的点不呈现规则性, 否则, 散点图通常将呈现“集团性”或“剧烈交错性”等形状.

正态性诊断. 若正态性成立, 那么学生化残差r_i可近似看成相互独立且服从$N(0,1)$. 所以, 在以r_i为纵坐标$\hat{y}_i$为横坐标的残差图上, 平面上的点$(\hat{y}_i, r_i), i = 1, \cdots, n$, 大致应落在宽度为4的水平带$|r_i| \leq 2$区域内(这个频率应在95%左右), 且不呈现任何趋势.

也可以用学生化残差的QQ图来做正态性诊断. 一组容量为n的数据关于某个分布$F(x)$的QQ图就是以数据的i/n分位数为纵坐标, 以$F(x)$的i/n分位数为横坐标的散点图. 如果数据确实是来自$F(x)$的一个简单随机样本, 则这些散点应大致散布在一条斜率为1的直线上或其附近. 此外, 还可以用Shapiro-Wilk方法来做正态性检验, 其原假设是“残差来自正态分布”).

来了解一下来自正态分布、对数正态分布(非对称分布的代表)、柯西分布(重尾分布的代表)和均匀分布(轻尾分布的代表)的随机数所展示的QQ图和Shapiro-Wilk正态性检验结果. R程序如下:

```
> n=50
> x=rnorm(n); qqnorm(x); qqline(x)
> shapiro.test(x)
> x=exp(rnorm(n)); qqnorm(x); qqline(x)
> shapiro.test(x)
> x=rcauchy(n); qqnorm(x); qqline(x)
> shapiro.test(x)
> x=runif(n); qqnorm(x); qqline(x)
> shapiro.test(x)
```

输出的四张QQ图见图3.4.1. 由此我们可以了解不同特征的随机变量它的QQ图大致是什么样子的, 这对今后利用QQ图判断数据是否来自正态分布是有帮助的. 上述的四个正态性检验的结果如下:

```
        Shapiro-Wilk normality test
data:  x
W = 0.98895, p-value = 0.9186

        Shapiro-Wilk normality test
data:  x
W = 0.78587, p-value = 4.206e-07

        Shapiro-Wilk normality test
data:  x
W = 0.53299, p-value = 2.222e-11

        Shapiro-Wilk normality test
data:  x
W = 0.94898, p-value = 0.03099
```

若取显著性水平为0.05, 那么根据p-value, 只接受第一组随机数的正态性假设.

下面来看图3.4.2中的6张残差图, 并做相应的分析. 在图3.4.2中, 对于图(a), 判断为其性状与正态性假设基本一致, 因此可认为正态性假设$\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$是合理的; 对于图(b), 由于它呈现喇叭型, 因此认为方差齐性假设不合理; 对于图(c), 由于它也呈现喇叭型, 因此认为方差齐性假设不合理; 对于图(d), 由于它呈现有规则的形状, 因此认为线性假设不合理, 残差中应含有自变量的信息; 对于图(e), 由于它呈现集团性, 因此认为不相关性假设不合理; 对于图(f), 由于它呈现剧烈交错性, 因此认为不相关性假设不合理.

从残差图诊断出来“疾病”(也就是某些假设条件不成立)后, 需要对问题“对症下药”. 如果有症状使我们怀疑线性假设不成立, 那么可以考虑在回归自变量中增加某些自变量的二次项, 如x_1^2, x_2^2或交叉项x_1x_2等, 具体增加哪些自变量的哪些项, 需视实际效果而定; 如果有症状使我们怀疑方差齐性假设不成立, 那么可以考虑对因变量作适当的变换使新变量具有近似相等的方差(方差稳定性变换), 或者采用后面介绍的广义最小二乘估计. 如果有症状使我们怀疑不相关性假设不成立, 那么可以考虑对因变量作“差分”, 使新变量具有近似独立性, 或者采用后面介绍的广义最小二乘估计. 如果有症状使我们怀疑正态性假设不成立, 那么可以考虑对因变量作正态性变换——Box-Cox 变换. 值得一提的是, Box-Cox变换是一种综合治理方案, 下一节再详细介绍.

先介绍方差稳定性变换. 设随机变量Y的均值为μ, 方差为σ^2, 假设方差为均值的函数, 即假设$\sigma^2 = g(\mu)$, 其中g是一个已知的函数. 例如, 若$Y \sim B(n,p)$, 则$\mathsf{E}(Y) = np = \mu$, $\mathsf{Var}(Y) = np(1-p) = \mu(1-\mu/n)$. 现需设法做一个变换$Z = f(Y)$, 使得$\mathsf{Var}(Z)$为常数. 这需要找出函数$f$的表达式. 这里, 为了不引起混淆, 用大写字母$Y$和$Z$表示随机变量, 小写字母$y$和$z$表示非随机的变量. 记$z = f(y)$, 并令它在$y = \mu$处做Taylor展开, 取近似式

$$f(y) = f(\mu) + f'(\mu)(y - \mu).$$

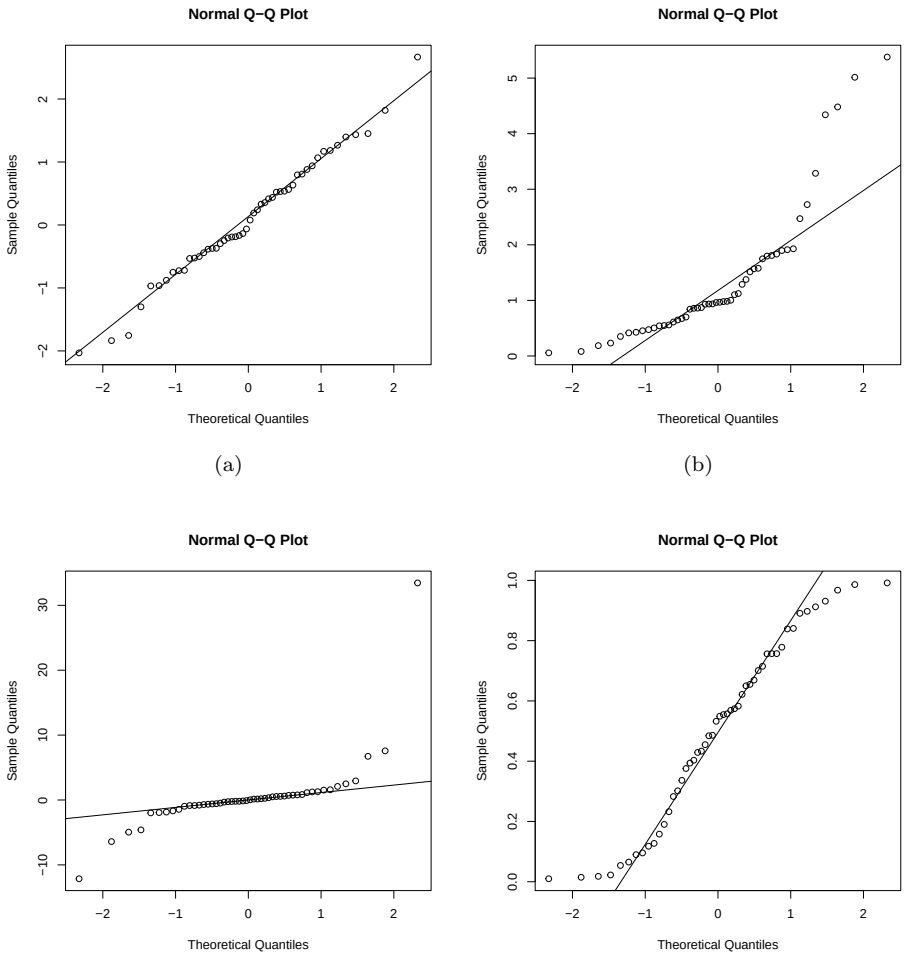
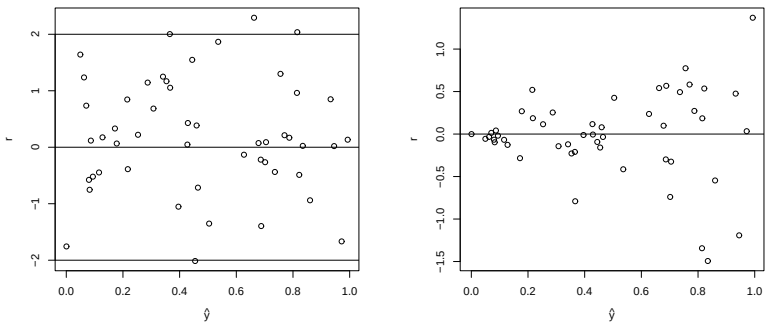
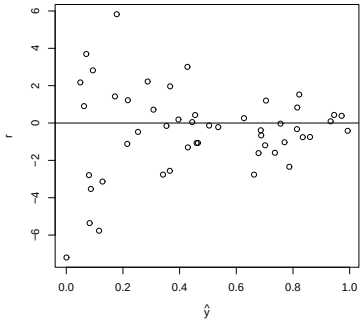


图3.4.1　QQ图.　图(a)是正态分布的QQ图, 图(b)是对数正态分布的QQ图, 图(c)是柯西分布的QQ图, 图(d)是均匀分布的QQ图

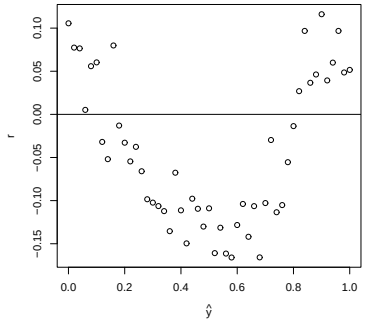


(a)

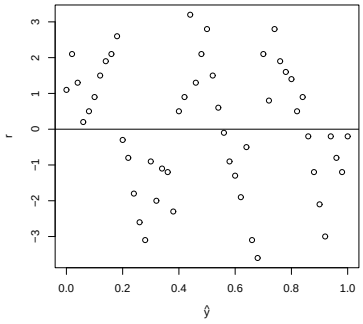
(b)



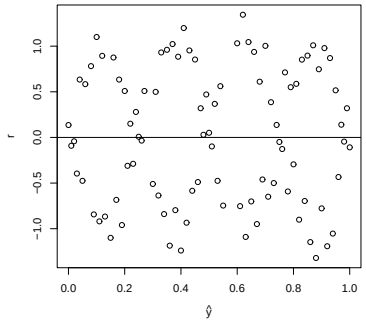
(c)



(d)



(e)



(f)

图3.4.2 残差图

将y改成随机变量, 有

$$Z = f(Y) = f(\mu) + f'(\mu)(Y - \mu),$$

其方差$c = \mathsf{Var}(Z) = [f'(\mu)]^2\mathsf{Var}(Y) = [f'(\mu)]^2 g(\mu)$. 取$f'(\mu) = \sqrt{c/g(\mu)}$, 则得到

$$f(y) = \int \sqrt{c/g(y)}\mathrm{d}y.$$

下面是几个特殊情形:

(1) $g(\mu)$与μ成正比时, 记$g(\mu) = a\mu$, 则

$$\int \sqrt{\frac{c}{ay}}\mathrm{d}y = 2\sqrt{\frac{c}{a}y} + c',$$

略去常数不计, 可作平方根变换: $Z = \sqrt{Y}$.

(2) 当$g(\mu)$与μ^2成正比时, 记$g(\mu) = a\mu^2$, 设$y > 0$, 则

$$\int \sqrt{\frac{c}{ay^2}}\mathrm{d}y = \sqrt{\frac{c}{a}}\ln y + c'.$$

略去常数不计, 可作对数变换: $Z = \ln Y$.

(3) 当$g(\mu)$与μ^4成正比时, 记$g(\mu) = a\mu^4$, 则

$$\int \sqrt{\frac{c}{ay^4}}\mathrm{d}y = -\sqrt{\frac{c}{a}}\frac{1}{y} + c'.$$

略去常数不计, 可作倒数变换: $Z = 1/Y$.

在应用上, 首先从残差图粗略地考察一下σ^2与μ可能存在的几种关系(即估计函数$g(\cdot)$), 然后从公式

$$f(y) = \int \sqrt{c/g(y)}\mathrm{d}y$$

确定对应的变换. 对几种变换过的数据分别作最小二乘处理, 画出新的残差图, 看哪一种变换的残差图无方差非齐性的征兆, 从中选出最好的方差稳定性变换.

例3.4.1 为研究用电高峰每小时的用电量y与每月总用电量x的关系, 现收集了某月53户数据, 见表3.4.1.

先应用最小二乘法, 得回归方程

$$\hat{y} = -0.83130 + 0.00368x.$$

其R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x=yx[,1]
> y=yx[,2]
> electricity=data.frame(y,x)
> lm.sol=lm(y~x,data=electricity)
> summary(lm.sol)
```

序号	x	y	序号	x	y
1	679	0.790	28	1748	4.880
2	292	0.440	29	1381	3.480
3	1012	0.560	30	1428	7.580
4	493	0.790	31	1255	2.630
5	582	2.700	32	1777	4.990
6	1156	3.640	33	370	0.590
7	997	4.730	34	2316	8.190
8	2189	9.500	35	1130	4.790
9	1097	5.340	36	463	0.510
10	2078	6.850	37	770	1.740
11	1818	5.840	38	724	4.100
12	1700	5.210	39	808	3.940
13	747	3.250	40	790	0.960
14	2030	4.430	41	783	3.290
15	1643	3.160	42	406	0.440
16	414	0.500	43	1242	3.240
17	354	0.170	44	658	2.140
18	1276	1.880	45	1746	5.710
19	745	0.770	46	468	0.640
20	435	1.390	47	1114	1.900
21	540	0.560	48	413	0.510
22	874	1.560	49	1787	8.330
23	1543	5.280	50	3560	14.940
24	1029	0.640	51	1495	5.110
25	710	4.000	52	2221	3.850
26	1434	0.310	53	1526	3.930
27	837	4.200			

表3.4.1　用电量数据

```
7
8  Call:
9  lm(formula = y ~ x, data = electricity)
10
11 Residuals:
12     Min      1Q  Median      3Q     Max
13 -4.1399 -0.8275 -0.1934  1.2376  3.1522
14
15 Coefficients:
16              Estimate Std. Error t value Pr(>|t|)
17 (Intercept) -0.8313037  0.4416121  -1.882   0.0655 .
18 x            0.0036828  0.0003339  11.030 4.11e-15 ***
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '' . 0.1 ' ' 1
21
22 Residual standard error: 1.577 on 51 degrees of freedom
23 Multiple R-squared:  0.7046,    Adjusted R-squared:  0.6988
24 F-statistic: 121.7 on 1 and 51 DF,  p-value: 4.106e-15
```

接下来做残差分析, 其R代码如下:

```
1 > y.fit=predict(lm.sol)
2 > e.hat=residuals(lm.sol)  #或者e.hat=y-y.fit
3 > e.std=rstandard(lm.sol)
4 > plot(e.hat~y.fit)
5 > plot(e.std~y.fit)
6 > plot(e.hat~x)
```

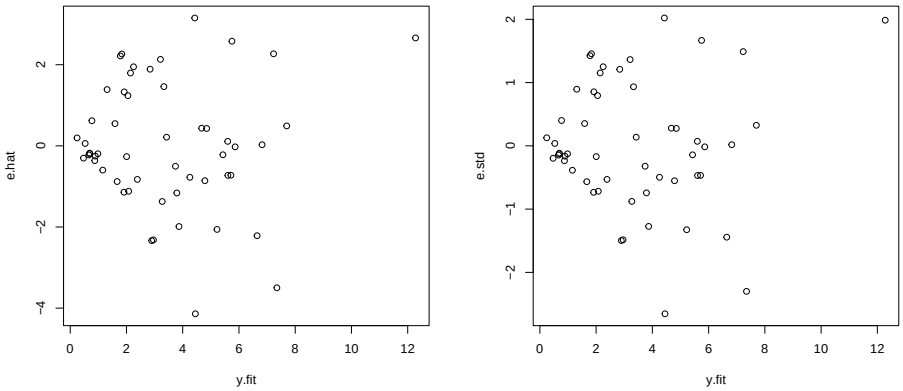
所得图像见图3.4.3的(a)-(c). 普通残差图也可以用命令plot(lm.sol,which=1)得到, 图像见图3.4.3的(d).

从残差图可看出, 这是一个喇叭型残差图, 是方差齐性不被符合的一个征兆. 考虑对因变量y作变换, 尝试$z=\sqrt{y}$, 得回归方程

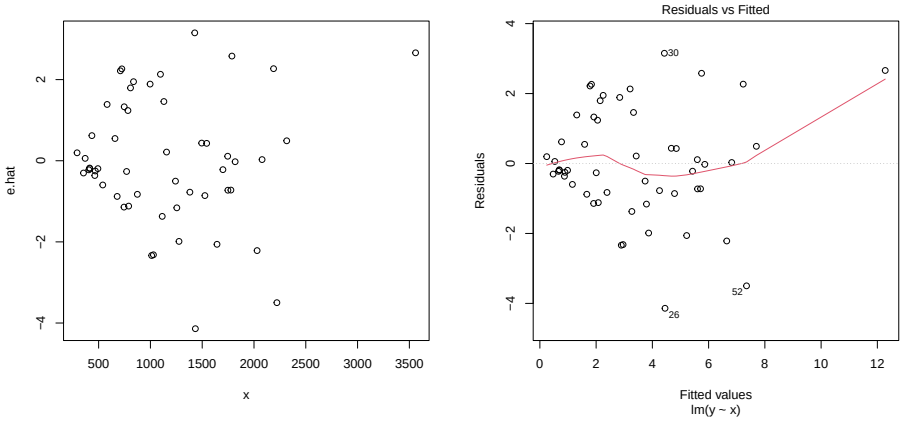
$$\hat{z}=0.5822+0.000953x.$$

其R代码及分析结果如下:

```
1  > z=sqrt(y)
2  > electricity2=data.frame(z,y,x)
3  > lm.sol2=lm(z~x,data=electricity2)
4  > summary(lm.sol2)
5  > z.fit=predict(lm.sol2)
6  > e.hat=residuals(lm.sol2)
7  > plot(e.hat~z.fit)
8
9  Call:
10 lm(formula = z ~ x, data = electricity2)
11
12 Residuals:
```



(a) (b)



(c) (d)

图3.4.3 残差图

```
     Min       1Q   Median       3Q      Max
-1.39185 -0.30576 -0.03875  0.25378  0.81027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.822e-01  1.299e-01   4.481 4.22e-05 ***
x           9.529e-04  9.824e-05   9.699 3.61e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '' . 0.1 ' ' 1

Residual standard error: 0.464 on 51 degrees of freedom
Multiple R-squared:  0.6485,    Adjusted R-squared:  0.6416
F-statistic: 94.08 on 1 and 51 DF,  p-value: 3.614e-13
```

输出的残差图见图3.4.4. 可以看到新的残差图不呈现任何明显规则性, 这表明所

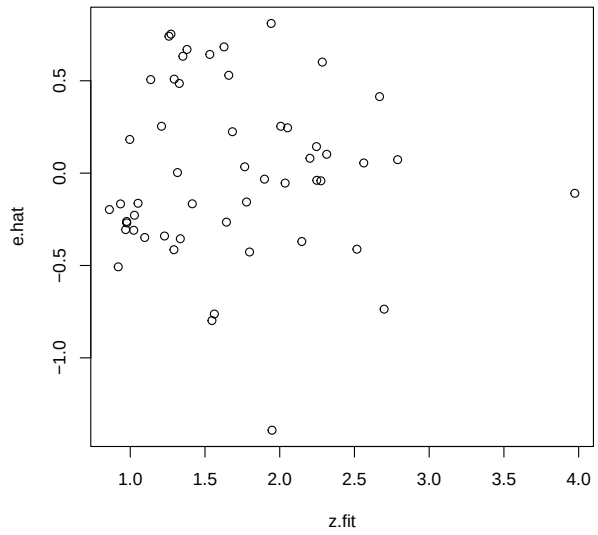


图3.4.4　残差图

用的变换是合适的. 最后得到回归方程为

$$\hat{y} = \hat{z}^2 = (0.5822 + 0.000953x)^2 = 0.339 + 0.0011x + 0.00000091x^2.$$

接下来做数据的诊断: 异常点诊断和强影响点诊断.

异常点诊断. 由于学生化残差r_i可近似看成相互独立且服从$N(0,1)$, 那么$|r_i| > 2$是个小概率事件, 发生的概率约为0.05. 因此, 若有某个$|r_i| > 2$, 就有理由怀疑相应的样本点$(\boldsymbol{x}_i', y_i)$是异常点.

强影响点诊断. 先引进一些记号, 用$\boldsymbol{Y}_{(i)}$, $\boldsymbol{X}_{(i)}$和$\boldsymbol{e}_{(i)}$分别表示从$\boldsymbol{Y}$, $\boldsymbol{X}$和$\boldsymbol{e}$中剔除第i行后所得到的向量或矩阵. 剔除第i组数据后, 剩下的$n-1$组数据的线性回归模型为

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{e}_{(i)},\ \mathsf{E}(\boldsymbol{e}_{(i)}) = \boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e}_{(i)}) = \sigma^2\boldsymbol{I}_{n-1}. \tag{3.4.6}$$

把从这个模型求得的$\boldsymbol{\beta}$的LSE记为$\hat{\boldsymbol{\beta}}_{(i)}$, 则

$$\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)}. \tag{3.4.7}$$

向量$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$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, 不便于应用分析, 应考虑它的某种数量化函数. Cook距离是其中应用最广泛的一个数量化函数. 首先来求$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$的精确表达式. 为此, 需先介绍一个恒等式.

引理3.4.1 设$\boldsymbol{A}$为$n \times n$可逆矩阵, $\boldsymbol{u}$和$\boldsymbol{v}$均为$n \times 1$向量, 那么有

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

记$\boldsymbol{x}_i'$为设计矩阵$\boldsymbol{X}$的第i行. 那么$\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)'$. 利用上述引理, 可知

$$\begin{aligned}(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1} =& (\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i')^{-1} \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1 - h_{ii}},\end{aligned} \tag{3.4.8}$$

其中$h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$为帽子矩阵$\boldsymbol{H}$的第$i$个对角线元素, 被称为杠杆点. 若某$h_{ii}$值较大, 则称$\boldsymbol{x}_i$为高杠杆点. 又因为

$$\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)} = \sum_{j\neq i}\boldsymbol{x}_j y_j = \sum_{j=1}^{n}\boldsymbol{x}_j y_j - \boldsymbol{x}_i y_i = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}_i y_i,$$

所以

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(i)} =& (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)} \\ =& \Big[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1 - h_{ii}}\Big](\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}_i y_i) \\ =& \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{1 - h_{ii}} - \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i h_{ii} y_i}{1 - h_{ii}} \\ =& \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i}{1 - h_{ii}} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{1 - h_{ii}} \\ =& \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{e}_i}{1 - h_{ii}},\end{aligned} \tag{3.4.9}$$

由此可得

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{e}_i}{1 - h_{ii}}.$$

Cook引进下列的距离:

$$D_i(\boldsymbol{M},c)=\frac{(\hat{\boldsymbol{\beta}}_{(i)}-\hat{\boldsymbol{\beta}})'\boldsymbol{M}(\hat{\boldsymbol{\beta}}_{(i)}-\hat{\boldsymbol{\beta}})}{c},$$

其中$\boldsymbol{M}$是给定的正定矩阵, c是给定的正常数. 容易看出

$$D_i(\boldsymbol{M},c)=\frac{\hat{e}_i^2}{c(1-h_{ii})^2}\cdot\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{M}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i.$$

取$\boldsymbol{M}=\boldsymbol{X}'\boldsymbol{X}$, $c=(p+1)\hat{\sigma}^2$, 则

$$D_i=\frac{\hat{e}_i^2}{(p+1)\hat{\sigma}^2(1-h_{ii})^2}\cdot h_{ii}=\frac{1}{p+1}\cdot\frac{h_{ii}}{1-h_{ii}}\cdot r_i^2.$$

定理3.4.2 Cook距离

$$\begin{aligned}D_i=&\frac{(\hat{\boldsymbol{\beta}}_{(i)}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{(i)}-\hat{\boldsymbol{\beta}})}{(p+1)\hat{\sigma}^2}\\=&\frac{\hat{e}_i^2}{(p+1)\hat{\sigma}^2(1-h_{ii})^2}\cdot h_{ii}\\=&\frac{1}{p+1}\cdot\frac{h_{ii}}{1-h_{ii}}\cdot r_i^2,\ i=1,\cdots,n,\end{aligned}\tag{3.4.10}$$

其中h_{ii}为帽子矩阵$\boldsymbol{H}$的第i个对角线元素, r_i是学生化残差.

这个定理表明, 在计算Cook距离的时候, 只需要从完全数据的线性回归模型算出学生化残差r_i和帽子矩阵的对角线元素h_{ii}就可以了, 而不必对任何一个不完全数据的线性回归模型(3.4.6)进行计算.

关于Cook距离的构造有另一解释. 向量$\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{(i)}-\hat{\boldsymbol{\beta}})=\hat{\boldsymbol{Y}}_{(i)}-\hat{\boldsymbol{Y}}$反映了第$i$组数据对模型拟合的影响, 而Cook距离$D_i$正是向量$\hat{\boldsymbol{Y}}_{(i)}-\hat{\boldsymbol{Y}}$的一个数量化函数.

下面分析h_{ii}的含义. h_{ii}其实度量了第i组自变量数据$\boldsymbol{x}_i$到它的中心$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i$的距离. 考虑模型

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e},\ \mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n,$$

假设自变量已中心化, 则

$$\boldsymbol{X}=\begin{pmatrix}1&x_{11}-\bar{x}_1&\cdots&x_{1p}-\bar{x}_p\\\vdots&\vdots&\ddots&\vdots\\1&x_{n1}-\bar{x}_1&\cdots&x_{np}-\bar{x}_p\end{pmatrix}=:\begin{pmatrix}1&(\boldsymbol{x}_1-\bar{\boldsymbol{x}})'\\\vdots&\vdots\\1&(\boldsymbol{x}_n-\bar{\boldsymbol{x}})'\end{pmatrix},$$

这里$\bar{\boldsymbol{x}}=(\bar{x}_1,\cdots,\bar{x}_p)'$, $\boldsymbol{x}_i=(x_{i1},\cdots,x_{ip})'$. 经简单计算可得

$$\boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix}n&\boldsymbol{0}\\\boldsymbol{0}&\boldsymbol{L}\end{pmatrix},$$

其中$\boldsymbol{L}=\sum_{i=1}^n(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$, 为$p$阶方阵. 进一步可知

$$h_{ii}=\frac{1}{n}+(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'\boldsymbol{L}^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}).$$

因此, h_{ii}度量了$\boldsymbol{x}_i$到它的中心$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$的距离. 若是一元中心化线性回归模型, 则

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}.$$

此时, h_{ii}如下的表达式:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

在公式(3.4.10)中, $1/(p+1)$与i无关, $P_i = \frac{h_{ii}}{1-h_{ii}}$是$h_{ii}$的单调增函数. 因为$h_{ii}$度量了$\boldsymbol{x}_i$到中心$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$的距离, 因此$P_i$是第$i$组数据$\boldsymbol{x}_i$距离其它数据的远近的一个度量. Cook距离被$P_i$和$r_i^2$的大小所决定. 此外可看出, h_{ii}越大, 则$\mathsf{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$越小, $\hat{y}_i$越接近y_i.

定理3.4.2告诉我们, 高杠杆点可能是强影响点, 也可能不是; 异常点可能是强影响点, 也可能不是; 但如果一组数据既是高杠杆点又是异常点, 那么它就是强影响点. 要给Cook距离一个用以判定强影响点的临界值是比较困难的.

应用置信椭球可以对Cook距离推导过程中的$\boldsymbol{M}$和c的选取给予一定的理论支持. 假设线性模型含有截距项. 在误差的正态假设下,

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

利用推论2.4.1知

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2(p+1).$$

另一方面

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

且与$\hat{\boldsymbol{\beta}}$相互独立. 所以

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(p+1)\hat{\sigma}^2} \sim F(p+1, n-p-1). \tag{3.4.11}$$

称集合

$$S = \left\{ \boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(p+1)\hat{\sigma}^2} \le F_\alpha(p+1, n-p-1) \right\}$$

为$\boldsymbol{\beta}$的置信水平为$1-\alpha$的置信椭球. (3.4.10)与(3.4.11)中的统计量很相似, 但前者并不服从F分布. 然而借助后者可以对D_i值的大小给出概率解释. 例如, 若$D_i = F_{0.50}(p+1, n-p-1)$, 则表明第$i$组数据$(\boldsymbol{x}_i', y_i)$被剔除后, $\boldsymbol{\beta}$的估计$\hat{\boldsymbol{\beta}}_{(i)}$移动到了$\boldsymbol{\beta}$的置信水平为0.5的置信椭球边界上; 若$D_j = F_{0.80}(p+1, n-p-1)$, 则$\hat{\boldsymbol{\beta}}_{(j)}$移动到了$\boldsymbol{\beta}$的置信水平为0.2的置信椭球边界上. 可知第$i$组数据对估计的影响比第$j$组数据来得大(因为$\hat{\boldsymbol{\beta}}_{(i)}$比$\hat{\boldsymbol{\beta}}_{(j)}$更远离$\boldsymbol{\beta}$了).

对于例3.4.1, 做如下的数据诊断分析. R代码如下:

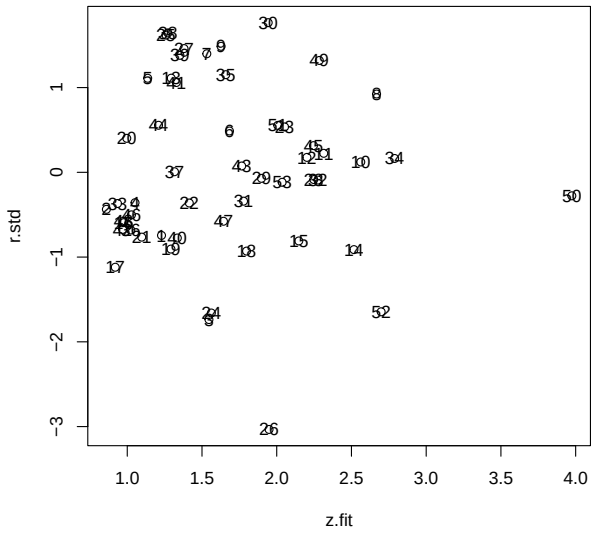


图3.4.5　残差图

```
> yx=read.table("electricity_data.txt")
> x=yx[,1]
> y=yx[,2]
> z=sqrt(y)
> electricity2=data.frame(z,x)
> lm.sol2=lm(z~x, data=electricity2)
> summary(lm.sol2)
> z.fit=predict(lm.sol2)
> r.std=rstandard(lm.sol2)
> plot(r.std~z.fit)
> text(z.fit, r.std, labels=seq(1,53,1))
> r.std
> cook=cooks.distance(lm.sol2)
> cook
> library(faraway)
> halfnorm(cook,2,ylab="Cook's distance")  #2表示找出2个潜在的强影响点
```

残差图见图3.4.5. 计算得到的Cook距离为

```
> cook=cooks.distance(lm.sol2)
> cook
           1            2            3            4            5
8.250076e-03 5.227243e-03 3.041009e-02 2.564379e-03 2.132165e-02
```

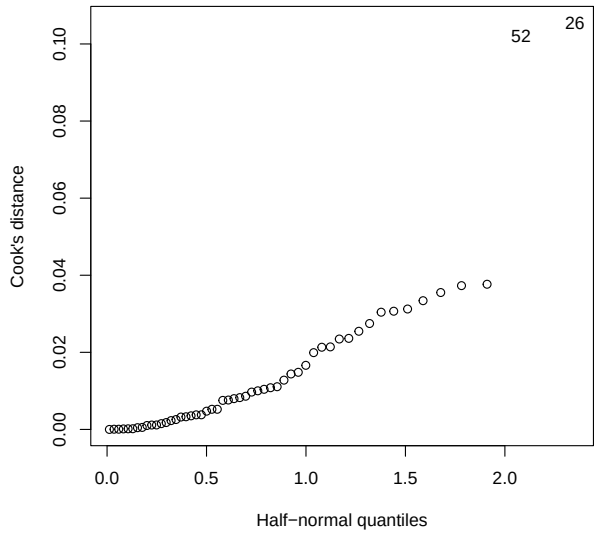


图3.4.6　半正态图

```
           6            7            8            9           10
2.286658e-03 1.991863e-02 3.066520e-02 2.141277e-02 4.522198e-04
          11           12           13           14           15
1.013856e-03 5.184643e-04 1.663042e-02 2.345137e-02 1.001708e-02
          16           17           18           19           20
7.986568e-03 3.123646e-02 8.608331e-03 1.107690e-02 3.524540e-03
          21           22           23           24           25
1.081715e-02 1.496214e-03 3.784084e-03 2.747336e-02 3.730115e-02
          26           27           28           29           30
1.055235e-01 2.546339e-02 1.300886e-04 5.478818e-05 3.551254e-02
          31           32           33           34           35
1.141198e-03 1.574666e-04 3.284202e-03 1.154314e-03 1.278325e-02
          36           37           38           39           40
9.684229e-03 6.219617e-07 3.766526e-02 2.361819e-02 7.625211e-03
          41           42           43           44           45
1.438682e-02 1.041041e-02 5.467630e-05 4.734775e-03 1.781595e-03
          46           47           48           49           50
5.227580e-03 3.215142e-03 7.532806e-03 3.339572e-02 1.481864e-02
          51           52           53
3.788335e-03 1.019629e-01 1.780708e-04
```

从图3.4.6的半正态图上可以找到两个强影响点.

由上面的输出结果, 可以看出: 第26号样本点为异常点, 需要检查数据来源是

否有过失, 若无过失则删除此样本点; 此外, 第26号和52号样本点为强影响点.

例3.4.2　智力测试数据. 表3.4.2是教育学家测试的21个儿童的记录, 其中x是儿童的年龄(单位: 月), y是某种智力指标, 通过这些数据, 建立智力随年龄变化的相依关系.

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

表3.4.2　智力测试数据

R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x=yx[,1]
> y=yx[,2]
> lm.sol=lm(y~x)
> summary(lm.sol)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-15.604  -8.731   1.396   4.523  30.285

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.8738     5.0678  21.681 7.31e-15 ***
x            -1.1270     0.3102  -3.633  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’ . 0.1 ‘ ’ 1

Residual standard error: 11.02 on 19 degrees of freedom
Multiple R-squared:   0.41,     Adjusted R-squared:  0.3789
F-statistic:  13.2 on 1 and 19 DF,  p-value: 0.001769
```

接下来做数据的诊断, R代码如下:

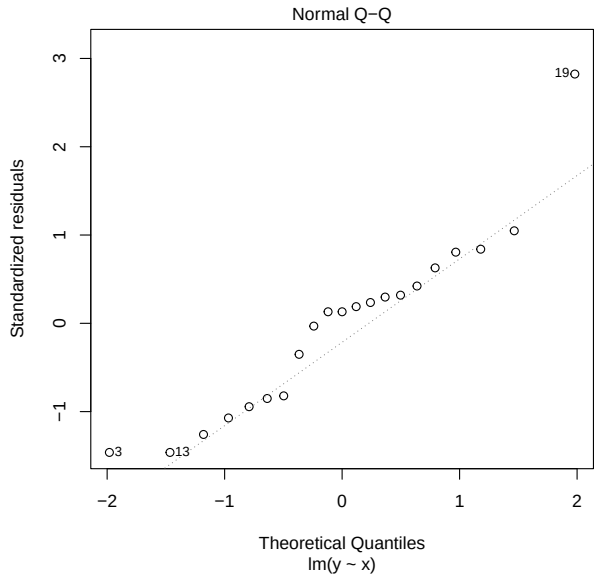


图3.4.7　QQ图

```
> plot(lm.sol,which=2)  #画QQ图
> y.fit=predict(lm.sol)
> r.std=rstandard(lm.sol)
> plot(r.std~y.fit,xlab=expression(hat(y)),ylab="r")
> text(y.fit,r.std,labels=seq(1,21,1))
> cook=cooks.distance(lm.sol)
> cook
> library(faraway)
> halfnorm(cook,2,ylab="Cook's distance")
```

输出的QQ图见图3.4.7, 残差图见图3.4.8.　计算得到的Cook距离如下:

```
> cook=cooks.distance(lm.sol)
> cook
          1           2           3           4           5
8.974064e-04 8.149796e-02 7.165814e-02 2.561596e-02 1.774366e-02
          6           7           8           9          10
3.877627e-05 3.130575e-03 1.668209e-03 3.831949e-03 1.543952e-02
         11          12          13          14          15
5.481014e-02 4.677623e-03 7.165814e-02 4.759781e-02 5.361216e-03
         16          17          18          19          20
5.735845e-04 1.785650e-02 6.781120e-01 2.232883e-01 3.451889e-02
         21
```

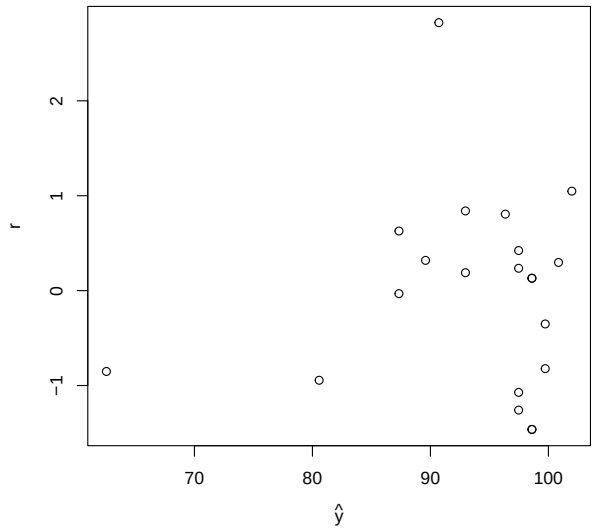


图3.4.8　残差图

```
12 5.735845e-04
```

半正态图见图3.4.9.

基于以上的输出结果, 可以认为: 第19号样本点为异常点; 第18样本点是强影响点. 若想找出两个强影响点的话, 可认为另一强影响点是第19号样本点. 由QQ图和残差图, 接受线性假设、方差齐性假设、不相关性假设和正态性假设. (因这里的样本量较少, 不同的人可能有不同的观点)

3.5　Box-Cox变换

对于观测数据, 若经过回归诊断后判断它们不满足线性假设、方差齐性假设、不相关性假设和正态性假设中的一个或若干个, 那么就要对数据采取“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. 本节介绍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.

Box-Cox变换是对因变量的如下变换:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \ln y, & \lambda = 0, \end{cases} \tag{3.5.1}$$

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

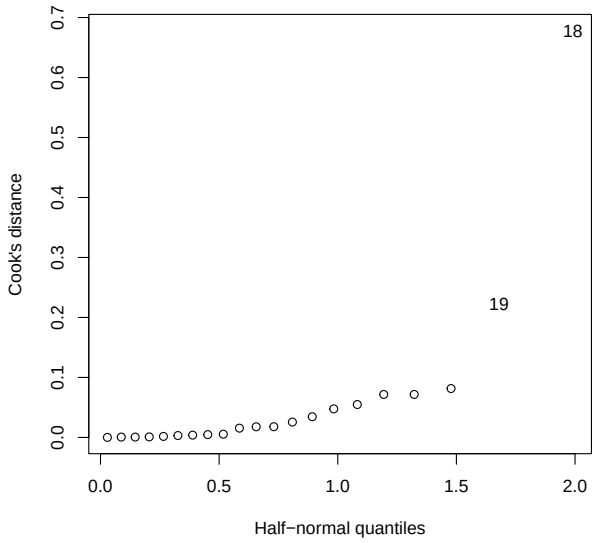


图3.4.9　半正态图

对因变量的n个观测值$y_1, \cdots, y_n$应用上述变换, 得到变换后的观测向量

$$\boldsymbol{Y}^{(\lambda)} = (y_1^{(\lambda)}, \cdots, y_n^{(\lambda)})'.$$

确定变换参数λ使得$\boldsymbol{Y}^{(\lambda)}$满足一个理想的线性回归模型:

$$\boldsymbol{Y}^{(\lambda)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{3.5.2}$$

即要求变换后的观测向量$\boldsymbol{Y}^{(\lambda)}$与回归自变量之间具有线性相关关系, 误差满足正态分布、方差齐性、相互独立. 因此, Box-Cox变换是通过对参数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.

通常用极大似然方法来确定λ. 对固定的λ、$\boldsymbol{\beta}$和σ^2, $\boldsymbol{Y}^{(\lambda)}$的似然函数为

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y}^{(\lambda)} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}^{(\lambda)} - \boldsymbol{X}\boldsymbol{\beta}) \Big\}.$$

所以$\boldsymbol{Y}$的似然函数为

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

这里J为变换的Jacobi行列式

$$J = \prod_{i=1}^{n} \frac{\mathrm{d}y_i^{(\lambda)}}{\mathrm{d}y_i} = \prod_{i=1}^{n} y_i^{\lambda - 1}.$$

对$\ln L(\boldsymbol{\beta},\sigma^2)$关于$\boldsymbol{\beta}$和$\sigma^2$求导并令其等于零, 可得$\boldsymbol{\beta}$和$\sigma^2$的极大似然估计:

$$\begin{cases} \hat{\boldsymbol{\beta}}(\lambda)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{(\lambda)}, \\ \hat{\sigma}^2(\lambda)=\frac{1}{n}\boldsymbol{Y}^{(\lambda)'}(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}^{(\lambda)}=:\frac{1}{n}\mathrm{RSS}(\lambda,\boldsymbol{Y}^{(\lambda)}). \end{cases} \tag{3.5.3}$$

对应的似然函数最大值为

$$L_{\max}(\lambda)=L(\hat{\boldsymbol{\beta}}(\lambda),\hat{\sigma}^2(\lambda))=(2\pi e)^{-n/2}\cdot|J|\cdot\Big(\frac{\mathrm{RSS}(\lambda,\boldsymbol{Y}^{(\lambda)})}{n}\Big)^{-n/2}. \tag{3.5.4}$$

它是λ的函数, 通过求它的最大值来确定λ. 这等价于通过求$\ln L_{\max}(\lambda)$的最大值来确定λ. 首先, 有

$$\begin{aligned} \ln L_{\max}(\lambda) &= -\frac{n}{2}\cdot\ln\Big[\mathrm{RSS}(\lambda,\boldsymbol{Y}^{(\lambda)})\Big]+\ln|J|+C \\ &= -\frac{n}{2}\ln\Big[\frac{\boldsymbol{Y}^{(\lambda)'}}{|J|^{1/n}}(\boldsymbol{I}_n-\boldsymbol{H})\frac{\boldsymbol{Y}^{(\lambda)}}{|J|^{1/n}}\Big]+C \\ &=: -\frac{n}{2}\ln\Big[\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})\Big]+C, \end{aligned} \tag{3.5.5}$$

这里$\boldsymbol{Z}^{(\lambda)}=(z_1^{(\lambda)},\cdots,z_n^{(\lambda)})'=\boldsymbol{Y}^{(\lambda)}/|J|^{1/n}$, 而

$$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})=\boldsymbol{Z}^{(\lambda)'}(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Z}^{(\lambda)}. \tag{3.5.6}$$

(3.5.5)式对Box-Cox变换在计算机上的实现带来很大方便, 因为为了求$\ln L_{\max}(\lambda)$的最大值, 只需求$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$的最小值. 虽然很难找到使$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$达到最小值的$\lambda$的解析表达式, 但对一系列给定的$\lambda$值, 通过求最小二乘估计的回归程序, 容易计算出对应的$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$. 然后画出$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$关于$\lambda$的图像, 从图上可以近似地找出使$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$达到最小值的$\lambda$(记作$\hat{\lambda}$).

Box-Cox变换的具体步骤如下:

(1) 对给定的λ值, 计算$\boldsymbol{Z}^{(\lambda)}$;

(2) 按(3.5.6)式计算残差平方和$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$;

(3) 对一系列给定的λ值, 重复上述步骤, 得到相应的残差平方和$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$的一串值. 以$\lambda$为横坐标, $\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$为纵坐标, 画出相应的曲线. 用直观方法, 找出使$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$达到最小值的$\lambda$, 记作$\hat{\lambda}$.

例3.5.1　在例3.4.1中, 对因变量y作了平方根变换, 这相当于使用了$\lambda=0.5$的Box-Cox变换. 现在来证实这样的变换是合适的. 下表给出了12个不同的λ值和对应的残差平方和$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$的大小, 简单比较后可发现当$\lambda=0.5$时残差平方和$\mathrm{RSS}(\lambda,\boldsymbol{Z}^{(\lambda)})$达到最小. 因此近似地认为0.5就是变换参数$\lambda$的最优选择.

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

表3.5.1

使用的R代码如下:

```
> yx=read.table("***.txt")
> x=yx[,1]
> y=yx[,2]
> lm.sol=lm(y~x)
> library(MASS)
> boxcox(lm.sol,plotit=T,lambda=seq(-2,2,by=0.05))
```

输出的Box-Cox变换图见图3.5.1.

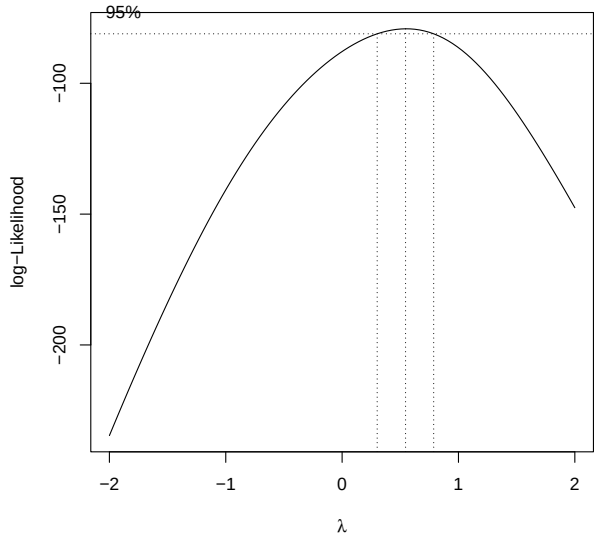


图3.5.1　Box-Cox变换图

3.6　广义最小二乘估计

前面的讨论总是假设线性回归模型的误差是方差齐性且不相关的, 即$\mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$. 但是在许多实际问题中, 数据往往不满足这个假设(可通过残差图判断). 这时需假设误差向量的协方差矩阵为$\mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}$, 这里$\boldsymbol{\Sigma}$是一个对称正定矩阵. $\boldsymbol{\Sigma}$包含未知参数, 但这里假设$\boldsymbol{\Sigma}$是完全已知的.

要讨论的线性回归模型为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \ \ \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \ \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}. \tag{3.6.1}$$

主要目的是估计$\boldsymbol{\beta}$. 因为$\boldsymbol{\Sigma}$是对称正定矩阵, 所以存在$n \times n$的正交阵$\boldsymbol{P}$使得

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}',$$

这里$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$, $\lambda_i>0$, $i=1,\cdots,n$, 是$\boldsymbol{\Sigma}$的特征根. 记

$$\boldsymbol{\Sigma}^{\frac{1}{2}}=\boldsymbol{P}\mathrm{diag}(\lambda_1^{\frac{1}{2}},\cdots,\lambda_n^{\frac{1}{2}})\boldsymbol{P}',$$

则可知$(\boldsymbol{\Sigma}^{\frac{1}{2}})^2=\boldsymbol{\Sigma}$, 称$\boldsymbol{\Sigma}^{\frac{1}{2}}$是$\boldsymbol{\Sigma}$的平方根阵. 用符号$\boldsymbol{\Sigma}^{-\frac{1}{2}}$表示$\boldsymbol{\Sigma}^{\frac{1}{2}}$的逆矩阵.

对模型(3.6.1)进行线性变换: 用$\boldsymbol{\Sigma}^{-\frac{1}{2}}$左乘(3.6.1). 记

$$\boldsymbol{Z}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y},\ \boldsymbol{U}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X},\ \boldsymbol{\varepsilon}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}.$$

因为$\mathsf{Cov}(\boldsymbol{\varepsilon})=\boldsymbol{\Sigma}^{-\frac{1}{2}}\sigma^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}}=\sigma^2\boldsymbol{I}_n$, 于是得到如下的线性回归模型:

$$\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{\beta}+\boldsymbol{\varepsilon},\ \ \mathsf{E}(\boldsymbol{\varepsilon})=\boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}_n. \tag{3.6.2}$$

这是前面已讨论过的模型. 在这个新模型中, $\boldsymbol{\beta}$的LSE为

$$\boldsymbol{\beta}^*=(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Z}=(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}. \tag{3.6.3}$$

称$\boldsymbol{\beta}^*$为$\boldsymbol{\beta}$的广义最小二乘估计(generalized least squares estimator, GLSE), 注意它与σ^2无关. 这个估计具有良好的统计性质.

定理3.6.1 对于线性回归模型(3.6.1), 下列结论成立:

(a) $\mathsf{E}(\boldsymbol{\beta}^*)=\boldsymbol{\beta}$;

(b) $\mathsf{Cov}(\boldsymbol{\beta}^*)=\sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$;

(c) 对任意的$p+1$维列向量$\boldsymbol{c}$, $\boldsymbol{c}'\boldsymbol{\beta}^*$为$\boldsymbol{c}'\boldsymbol{\beta}$的唯一最小方差线性无偏估计.

证明: (a)

$$\mathsf{E}(\boldsymbol{\beta}^*)=(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathsf{E}(\boldsymbol{Y})=(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{\beta}.$$

(b) 利用定理2.1.3,

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{\beta}^*)=&\mathsf{Cov}[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}]\\ =&(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathsf{Cov}(\boldsymbol{Y})((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})'\\ =&\sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})'\\ =&\sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}.\end{aligned}$$

(c) 设$\boldsymbol{b}'\boldsymbol{Y}$是$\boldsymbol{c}'\boldsymbol{\beta}$的任一线性无偏估计. 对于模型(3.6.2),

$$\boldsymbol{c}'\boldsymbol{\beta}^*=\boldsymbol{c}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Z},\ \boldsymbol{b}'\boldsymbol{Y}=\boldsymbol{b}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y}=\boldsymbol{b}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z},$$

即$\boldsymbol{c}'\boldsymbol{\beta}^*$为$\boldsymbol{c}'\boldsymbol{\beta}$的LSE, 且它是$\boldsymbol{c}'\boldsymbol{\beta}$的线性无偏估计, 而$\boldsymbol{b}'\boldsymbol{Y}=\boldsymbol{b}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$也是$\boldsymbol{c}'\boldsymbol{\beta}$ 的线性无偏估计. 所以对模型(3.6.2)应用Gauss-Markov定理可知$\boldsymbol{c}'\boldsymbol{\beta}^*$为$\boldsymbol{c}'\boldsymbol{\beta}$的唯一最小方差线性无偏估计. □

定理3.6.1(c)就是一般情形下的Gauss-Markov定理, 它表明在一般线性回归模型(3.6.1)中, GLSE $\boldsymbol{\beta}^*$是最优的(若$\boldsymbol{\Sigma}=\boldsymbol{I}_n$, 则GLSE退化到LSE $\hat{\boldsymbol{\beta}}$). 对于模型(3.6.1), 容易证明$\hat{\boldsymbol{\beta}}$仍是无偏估计, 但未必是最优的线性无偏估计, 因为$\mathsf{Var}(\boldsymbol{c}'\boldsymbol{\beta}^*)\le\mathsf{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})$. 这就是说, 对于一般线性回归模型(3.6.1), GLSE总是优于LSE.

模型(3.6.1)的最简单的例子是因变量的不同观测具有不等方差的情形, 即

$$\mathsf{Cov}(\boldsymbol{e}) = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2),$$

这里的$\sigma_i^2, i = 1, \cdots, n$, 不全相等. 记$\boldsymbol{x}_1', \cdots, \boldsymbol{x}_n'$分别是设计矩阵$\boldsymbol{X}$的$n$个行向量. 容易推出:

$$\boldsymbol{\beta}^* = \Big(\sum_{i=1}^n \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2}\Big)^{-1} \Big(\sum_{i=1}^n \frac{\boldsymbol{x}_i y_i}{\sigma_i^2}\Big). \tag{3.6.4}$$

两个和式分别是$\boldsymbol{x}_i\boldsymbol{x}_i'$和$\boldsymbol{x}_i y_i$的加权和(权重都为$1/\sigma_i^2$), 因此也称$\boldsymbol{\beta}^*$为加权最小二乘估计(weighted least squares estimator, WLSE). σ_i^2往往是未知的, 这时需设法求得它们的估计$\hat{\sigma}_i^2$, 然后在(3.6.4)中用$\hat{\sigma}_i^2$代替σ_i^2. 这种估计方法称为两步估计(two-stage estimate)方法.

例3.6.1 假设用一种精密仪器在两个实验室对同一个量μ分别进行了n_1次和n_2次测量, 记这些测量值分别为$y_{11}, \cdots, y_{1n_1}$和$y_{21}, \cdots, y_{2n_2}$. 把它们写成线性回归模型的形式:

$$\begin{cases} y_{1i} = \mu + e_{1i}, & i = 1, \cdots, n_1, \\ y_{2i} = \mu + e_{2i}, & i = 1, \cdots, n_2. \end{cases}$$

由于两个实验室的客观条件及仪器的精度不同, 故它们的测量误差的方差不等. 设

$$\mathsf{Var}(e_{1i}) = \sigma_1^2, \ \mathsf{Var}(e_{2i}) = \sigma_2^2, \ \sigma_1^2 \neq \sigma_2^2.$$

记$\boldsymbol{e} = (e_{11}, \cdots, e_{1n_1}, e_{21}, \cdots, e_{2n_2})'$, 则

$$\mathsf{Cov}(\boldsymbol{e}) = \begin{pmatrix} \sigma_1^2 \boldsymbol{I}_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_2^2 \boldsymbol{I}_{n_2} \end{pmatrix} = \sigma_2^2 \begin{pmatrix} \theta \boldsymbol{I}_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_2} \end{pmatrix} =: \sigma_2^2 \boldsymbol{\Sigma},$$

这里$\theta = \sigma_1^2/\sigma_2^2$. 假设$\theta$已知, 则$\boldsymbol{\Sigma}$已知. 注意到这里的设计矩阵$\boldsymbol{X} = (1, \cdots, 1)' = \boldsymbol{1}_{n_1+n_2}$, 于是$\mu$的GLSE为

$$\mu^* = \big(\frac{n_1}{\theta} + n_2\big)^{-1} \Big(\frac{1}{\theta}\sum_{i=1}^{n_1} y_{1i} + \sum_{i=1}^{n_2} y_{2i}\Big).$$

记

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}, \qquad \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i},$$
$$\omega_1 = \frac{1}{\mathsf{Var}(\bar{y}_1)} = \frac{n_1}{\sigma_1^2}, \qquad \omega_2 = \frac{1}{\mathsf{Var}(\bar{y}_2)} = \frac{n_2}{\sigma_2^2}.$$

则μ^*可改写为

$$\mu^* = \frac{\omega_1}{\omega_1 + \omega_2}\bar{y}_1 + \frac{\omega_2}{\omega_1 + \omega_2}\bar{y}_2.$$

即μ^*是两个实验室观测值均值的加权平均, 它们的权重$\frac{\omega_1}{\omega_1+\omega_2}$和$\frac{\omega_2}{\omega_1+\omega_2}$与各实验室测量的误差方差和测量次数有关. 误差方差大的, 测量次数少的, 对应的权重就小.

μ^*包含未知参数σ_1^2和σ_2^2, 不能付诸实际应用. 可以设法构造σ_1^2和σ_2^2的估计. 事实上, 这两个实验室的观测数据分别构成了线性回归模型:

$$\boldsymbol{Y}_i = \mu \mathbf{1}_{n_i} + \boldsymbol{e}_i, \quad i = 1, 2,$$

这里$\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{in_i})'$, $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{in_i})'$. 因为$\mathsf{Cov}(\boldsymbol{e}_i) = \sigma_i^2 \boldsymbol{I}_{n_i}$, 所以$\boldsymbol{e}_i, i = 1, 2$, 满足Gauss-Markov假设. 所以$\sigma_i^2$的LSE为

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \|\boldsymbol{Y}_i - \bar{y}_i \mathbf{1}_{n_i}\|^2, \quad i = 1, 2.$$

用$\hat{\sigma}_i^2, i = 1, 2$, 代替$\mu^*$中的$\sigma_i^2, i = 1, 2$, 即可得到$\mu$的两步估计.

3.7 多重共线性

回归系数的LSE有许多优良的性质, 其中最重要的是Gauss-Markov定理, 它表明在线性无偏估计类中, LSE是唯一的具有最小方差的估计. 正是这一优点, 使得LSE在线性统计模型的估计理论和实际应用中占有绝对重要的地位.

以前讨论的LSE需要假设设计矩阵$\boldsymbol{X}$是列满秩的, 即要求矩阵$\boldsymbol{X}$的列向量之间是线性无关的. 然而, 在实际问题中, 由于经常要处理含有较多自变量的大型回归问题, 且经济变量之间往往不是孤立的而是相互联系的, 这些都会导致设计矩阵$\boldsymbol{X}$的列向量之间不可能完全线性无关. 很多情况下, 设计矩阵$\boldsymbol{X}$的列向量之间存在着所谓的多重共线性/复共线性(multi-collinearity)关系.

定义3.7.1(完全的多重共线性/完全的复共线性) 若存在不全为0的$p+1$个常数$c_0, c_1, \cdots, c_p$使得

$$c_0 + c_1 x_{i1} + \cdots + c_p x_{ip} = 0, \quad i = 1, \cdots, n,$$

则称自变量$x_1, \cdots, x_p$之间存在着完全的多重共线性/完全的复共线性关系.

在实际问题中, 完全的多重共线性/完全的复共线性关系并不多见, 一般出现的是一定程度的共线性.

定义3.7.2(多重共线性/复共线性) 若存在不全为0的$p+1$个常数$c_0, c_1, \cdots, c_p$使得

$$c_0 + c_1 x_{i1} + \cdots + c_p x_{ip} \approx 0, \quad i = 1, \cdots, n,$$

则称自变量$x_1, \cdots, x_p$之间存在着多重共线性/复共线性关系.

对经济数据建模时, 多重共线性的情形很常见. 多重共线性会给多元线性回归分析带来什么影响、如何诊断自变量之间的多重共线性关系以及如何克服多重共线性的影响等问题将是本节要讨论的主要内容.

先引入一个概念: 均方误差(mean squared errors, MSE), 它是评价一个估计优劣性的标准之一.

定义3.7.3 设$\boldsymbol{\theta}$为一列向量, $\hat{\boldsymbol{\theta}}$为$\boldsymbol{\theta}$的一个估计. 定义$\hat{\boldsymbol{\theta}}$的均方误差为

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathsf{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \mathsf{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})].$$

定理3.7.1

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{tr}[\mathsf{Cov}(\hat{\boldsymbol{\theta}})] + \|\mathsf{E}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2.$$

证明: 不难看出

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\theta}}) =&\mathsf{E}[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})]\\ =&\mathsf{E}[(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})+(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})]'[(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})+(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})]\\ =&\mathsf{E}(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})+\mathsf{E}(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\ =&:\Delta_1+\Delta_2.\end{aligned}$$

利用迹的性质,

$$\begin{aligned}\Delta_1 =&\mathsf{E}\{\mathrm{tr}[(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})]\}\\ =&\mathsf{E}\{\mathrm{tr}[(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})']\}\\ =&\mathrm{tr}[\mathsf{E}(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}-\mathsf{E}\hat{\boldsymbol{\theta}})']\\ =&\mathrm{tr}[\mathsf{Cov}(\hat{\boldsymbol{\theta}})].\end{aligned}$$

$\Delta_2=\mathsf{E}[(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'(\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})]=\|\mathsf{E}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2$是显然的. 证毕. □

若记$\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,\cdots,\hat{\theta}_{p+1})'$, 则

$$\Delta_1=\sum_{i=1}^{p+1}\mathsf{Var}(\hat{\theta}_i),$$

它是$\hat{\boldsymbol{\theta}}$各分量的方差之和. 而

$$\Delta_2=\sum_{i=1}^{p+1}(\mathsf{E}\hat{\theta}_i-\theta_i)^2,$$

它是$\hat{\boldsymbol{\theta}}$各分量的偏倚平方之和. 所以, 一个估计的均方误差由它的方差和偏差平方所决定. 一个好的估计应同时具有较小的方差和偏差平方.

定理3.7.2 在线性回归模型(3.1.5)中, 对$\boldsymbol{\beta}$的LSE $\hat{\boldsymbol{\beta}}$, 有
(a) $\mathrm{MSE}(\hat{\boldsymbol{\beta}})=\sigma^2\sum_{i=1}^{p+1}\frac{1}{\lambda_i}$;
(b) $\mathsf{E}\|\hat{\boldsymbol{\beta}}\|^2=\|\boldsymbol{\beta}\|^2+\sigma^2\sum_{i=1}^{p+1}\frac{1}{\lambda_i}$,
其中$\lambda_1,\cdots,\lambda_{p+1}>0$为$\boldsymbol{X}'\boldsymbol{X}$的特征根.

证明: (a) 因为LSE $\hat{\boldsymbol{\beta}}$是无偏估计, 所以$\Delta_2=0$,

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}})=\Delta_1=\mathrm{tr}[\mathsf{Cov}(\hat{\boldsymbol{\beta}})]=\sigma^2\mathrm{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}].$$

因为$\boldsymbol{X}'\boldsymbol{X}$是对称正定矩阵, 所以存在正交阵$\boldsymbol{P}$使得

$$\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{P}\mathrm{diag}(\lambda_1,\cdots,\lambda_{p+1})\boldsymbol{P}',$$

这里$\lambda_1,\cdots,\lambda_{p+1}>0$为$\boldsymbol{X}'\boldsymbol{X}$的特征根. 所以

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}=\boldsymbol{P}\mathrm{diag}(\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_{p+1}})\boldsymbol{P}'.$$

利用迹的性质可得

$$\mathrm{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}] = \mathrm{tr}\Big(\mathrm{diag}(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_{p+1}})\Big) = \sum_{i=1}^{p+1} \frac{1}{\lambda_i}.$$

所以$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{i=1}^{p+1} \frac{1}{\lambda_i}$.

(b) 因为

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\beta}}) =& \mathsf{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ =& \mathsf{E}(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}'\boldsymbol{\beta}) \\ =& \mathsf{E}\|\hat{\boldsymbol{\beta}}\|^2 - \boldsymbol{\beta}'\boldsymbol{\beta},\end{aligned}$$

于是

$$\mathsf{E}\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{i=1}^{p+1} \frac{1}{\lambda_i}.$$

证毕. □

结论(a)告诉我们, 如果$\boldsymbol{X}'\boldsymbol{X}$至少有一个特征根非常小, 即非常接近于零, 那么$\mathrm{MSE}(\hat{\boldsymbol{\beta}})$就会很大. 从均方误差的标准来看, 最小二乘估计$\hat{\boldsymbol{\beta}}$不是一个好的估计. 这和Gauss-Markov定理并不矛盾, 因为Gauss-Markov定理仅仅保证了最小二乘估计在线性无偏估计类中的方差最小性. 但在$\boldsymbol{X}'\boldsymbol{X}$至少有一个特征根非常小时, 这个最小的方差值本身就很大, 因而导致了很大的均方误差. 结论(b)告诉我们, 如果$\boldsymbol{X}'\boldsymbol{X}$至少有一个特征根非常小, 那么最小二乘估计$\hat{\boldsymbol{\beta}}$的长度平均说来要比真正的$\boldsymbol{\beta}$的长度长很多. 这就导致了$\hat{\boldsymbol{\beta}}$的某些分量的绝对值过大. 总之, 当$\boldsymbol{X}'\boldsymbol{X}$至少有一个特征根非常小时, 最小二乘估计不再是一个好的估计了.

“至少有一个特征根非常小”在设计矩阵$\boldsymbol{X}$或者回归自变量上意味着什么呢? 记$\boldsymbol{X} = (\mathbf{1}_n, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$, 即$\boldsymbol{x}_i$为$\boldsymbol{X}$的第$i+1$列. 设$\lambda$为$\boldsymbol{X}'\boldsymbol{X}$的一个特征根, $\boldsymbol{\phi}$为对应的特征向量, 其长度为1, 即$\boldsymbol{\phi}'\boldsymbol{\phi} = 1$. 若$\lambda \approx 0$, 则

$$\|\boldsymbol{X}\boldsymbol{\phi}\|^2 = \boldsymbol{\phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\phi} = \lambda\boldsymbol{\phi}'\boldsymbol{\phi} = \lambda \approx 0.$$

于是$\boldsymbol{X}\boldsymbol{\phi} \approx \mathbf{0}$. 记$\boldsymbol{\phi} = (c_0, c_1, \cdots, c_p)'$, 则

$$c_0\mathbf{1}_n + c_1\boldsymbol{x}_1 + \cdots + c_p\boldsymbol{x}_p \approx \mathbf{0}, \tag{3.7.1}$$

即设计矩阵$\boldsymbol{X}$的列向量之间近似线性相关. 反之, 若设计矩阵$\boldsymbol{X}$的列向量之间近似线性相关, 即(3.7.1) 成立, 此时$\boldsymbol{X}'\boldsymbol{X}$ 仍是正定矩阵, 但$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$. 由此可知

$$\prod_{i=1}^{p+1} \lambda_i = |\boldsymbol{X}'\boldsymbol{X}| \approx 0,$$

所以至少有一个特征根非常小, 接近于零. 也就是说至少有一个特征根非常小与$\boldsymbol{X}$的列向量之间近似线性相关是等价的. 此时称设计矩阵$\boldsymbol{X}$是病态矩阵. 基于当前的数据, $\boldsymbol{X}$的列向量之间近似线性相关与自变量之间具有多重共线性关系是等价的.

下面介绍多重共线性的诊断.

(1) 方差膨胀因子(variance inflation factor, VIF)诊断法

记R_j^2为自变量x_j对其余$p-1$个自变量的判定系数, 定义

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}, \quad j = 1, \cdots, p.$$

由于R_j^2度量了自变量x_j与其余$p-1$个自变量之间的线性相关程度, $x_1, \cdots, x_p$之间的多重共线性越严重, R_j^2越接近于1, VIF_j也就越大. 因此用VIF来度量多重共线性的程度是合理的. 度量的准则是: 当有某个

$$\mathrm{VIF}_j \geq 10$$

时, 认为自变量之间存在严重的多重共线性.

注: 也有人认为当

$$\overline{\mathrm{VIF}} = \frac{1}{p}\sum_{i=1}^{p} \mathrm{VIF}_j \gg 1$$

时, 判断认为自变量之间存在严重的多重共线性. 但这种标准看起来还不够客观.

(2) 特征根与条件数(condition index, CI)诊断法

为了消除量纲的影响(否则, 特征根具有量纲), 假设自变量与因变量的观测值均已标准化. 此时可认为线性回归模型没有截距项, 设计矩阵$\boldsymbol{X}$是$n \times p$矩阵, $\boldsymbol{X}'\boldsymbol{X}$是$p$个自变量的样本相关系数矩阵.

特征根诊断法: 如果$\boldsymbol{X}'\boldsymbol{X}$有$m$个特征根近似为零, 那么$\boldsymbol{X}$就有$m$个多重共线性关系, 并且这$m$个多重共线性关系的系数向量就是这$m$个接近于零的特征根所对应的标准正交化特征向量.

条件数诊断法: 假设$\boldsymbol{X}'\boldsymbol{X}$的$p$个特征根分别为$\lambda_1, \cdots, \lambda_p$, 其中最大特征根为$\lambda_{\max}$, 最小特征根为$\lambda_{\min}$, 称

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j}, \quad j = 1, \cdots, p$$

为特征根λ_j的条件数. 记

$$\kappa = \max_j \kappa_j = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

它度量了矩阵$\boldsymbol{X}'\boldsymbol{X}$的特征根的散布程度, 也可用来近似衡量最小特征根接近零的程度, 因此可以用来判断多重共线性是否存在以及多重共线性的严重程度. 条件数的判定准则如下:

(1) 若$0 < \kappa \leq 100$, 则认为不存在多重共线性;

(2) 若$100 < \kappa \leq 1000$, 则认为存在较强的多重共线性;

(3) 若$\kappa > 1000$, 则认为存在严重的多重共线性.

例3.7.1 考虑一个有六个回归自变量的线性回归问题, 原始数据见表3.7.1.

序号	y	x_1	x_2	x_3	x_4	x_5	x_6
1	10.006	8	1	1	1	0.541	-0.099
2	9.737	8	1	1	0	0.130	0.070
3	15.087	8	1	1	0	2.116	0.115
4	8.422	0	0	9	1	-2.397	0.252
5	8.625	0	0	9	1	-0.046	0.017
6	16.289	0	0	9	1	0.365	1.504
7	5.958	2	7	0	1	1.996	-0.865
8	9.313	2	7	0	1	0.228	-0.055
9	12.960	2	7	0	1	1.380	0.502
10	5.541	0	0	0	10	-0.798	-0.399
11	8.756	0	0	0	10	0.257	0.101
12	10.937	0	0	0	10	0.440	0.432

表3.7.1

数据分析的R代码如下:

```
> yx=read.table("***.txt")
> y=yx[,1]
> x1=yx[,2]
> x2=yx[,3]
> x3=yx[,4]
> x4=yx[,5]
> x5=yx[,6]
> x6=yx[,7]
> colinearity=data.frame(y,x1,x2,x3,x4,x5,x6)
> lm.sol=lm(y~.,data=colinearity)
> summary(lm.sol)
> library(DAAG)
> vif(lm.sol)
```

输出的VIF值见表3.7.2.

x_1	x_2	x_3	x_4	x_5	x_6
182.0500	161.3600	266.2600	297.7100	1.9200	1.4553

表3.7.2　VIF值

因为前四个VIF值都远大于10, 所以认为自变量之间存在严重的多重共线性. 接下来进行特征根与条件数诊断, R代码如下:

```
> X=cbind(x1,x2,x3,x4,x5,x6)
> rho=cor(X)
> rho
> eigen(rho)
```

```
> kappa(rho,exact=TRUE)  #默认是exact=FALSE, 这时有较大的计算误差
```

输出的特征根和特征向量为

```
> rho=cor(X)
> rho
            x1          x2         x3          x4         x5          x6
x1  1.00000000  0.05230658 -0.3433818 -0.49761095  0.4172974 -0.19209942
x2  0.05230658  1.00000000 -0.4315953 -0.37069641  0.4845495 -0.31673965
x3 -0.34338179 -0.43159531  1.0000000 -0.35512135 -0.5051579  0.49437941
x4 -0.49761095 -0.37069641 -0.3551214  1.00000000 -0.2145543 -0.08690551
x5  0.41729739  0.48454950 -0.5051579 -0.21455429  1.0000000 -0.12295400
x6 -0.19209942 -0.31673965  0.4943794 -0.08690551 -0.1229540  1.00000000
> eigen(rho)
eigen() decomposition
$values
[1] 2.428787365 1.546152096 0.922077664 0.793984690 0.307892134
  0.001106051

$vectors
           [,1]        [,2]        [,3]        [,4]       [,5]
  [,6]
[1,] -0.3907189  0.33968212  0.67980398 -0.07990398  0.2510370
  -0.447679719
[2,] -0.4556030  0.05392140 -0.70012501 -0.05768633  0.3444655
  -0.421140280
[3,]  0.4826405  0.45332584 -0.16077736 -0.19102517 -0.4536372
  -0.541689124
[4,]  0.1876590 -0.73546592  0.13587323  0.27645223 -0.0152087
  -0.573371872
[5,] -0.4977330  0.09713874 -0.03185053  0.56356440 -0.6512834
  -0.006052127
[6,]  0.3519499  0.35476494 -0.04864335  0.74817535  0.4337463
  -0.002166594
```

输出的条件数为

```
> kappa(rho,exact=TRUE)  #默认是exact=FALSE, 这时有较大的计算误差
[1] 2195.908
```

由于$\lambda_{\min} = 0.001106051$, 对应的特征向量为

$$(0.44768, 0.42114, 0.541689, 0.57337, 0.00605, 0.00217)',$$

所以标准化自变量$x_i^*, i = 1, \cdots, 6$, 之间存在如下的多重共线性关系:

$$0.44768x_1^* + 0.42114x_2^* + 0.541689x_3^* + 0.57337x_4^* + 0.00605x_5^* + 0.00217x_6^* \approx 0.$$

这说明对于原始自变量$x_i, i = 1, \cdots, 6$, 存在常数c_0使得

$$c_0 + 15.249x_1 + 16.112x_2 + 16.042x_3 + 15.967x_4 + 0.583x_5 + 0.452x_6 \approx 0.$$

条件数$\kappa = 2195.908 > 1000$, 也说明自变量之间存在严重的多重共线性关系.

消除多重共线性的方法有: (1) 增大样本容量, 消除或缓解设计矩阵列向量的近似线性相关性(注意: 多重共线性可能是数据的问题, 不一定是自变量的问题); (2) 牺牲无偏性, 寻找有偏估计(将介绍岭估计和主成分估计).

3.8 岭估计

当自变量之间具有多重共线性时, 为了克服LSE明显变坏的问题, Hoerl于1962年提出了一种改进的最小二乘估计方法, 即岭估计(ridge estimate). 之后, Hoerl和Kennard在1970年对该估计作了进一步的详细讨论.

岭估计的思想: 当自变量之间存在多重共线性时, 设计矩阵$\boldsymbol{X}$是病态的, 即$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$, 从而$(\boldsymbol{X}'\boldsymbol{X})^{-1}$接近奇异. 为避免这一现象, 给$\boldsymbol{X}'\boldsymbol{X}$ 加上一个正常数对角矩阵$k\boldsymbol{I}(k > 0)$, 则矩阵

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}$$

就有可能远离奇异性. 因此用

$$\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{3.8.1}$$

作为未知参数$\boldsymbol{\beta}$的估计应该要比LSE更稳定一些.

定义3.8.1 对给定的$k > 0$, 称$\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$为回归系数$\boldsymbol{\beta}$的岭估计. 根据岭估计建立的回归方程称为岭回归方程. 称k为岭参数. 对于$\hat{\boldsymbol{\beta}}(k)$ 的分量$\hat{\beta}_j(k)$, 把在平面直角坐标系中$\hat{\beta}_j(k)$随k变化所表现出来的曲线称为岭迹(ridge trace).

注: k不同, 得到不同的估计. 因此岭估计$\hat{\boldsymbol{\beta}}(k)$是一个估计类. 当$k = 0$时, $\hat{\boldsymbol{\beta}}(k)$就是通常的LSE. 一般情况下, 提起岭估计, 是不包括LSE的.

在进行岭估计之前, 为了消除量纲的影响($k\boldsymbol{I}$是没有量纲的, 所以$\boldsymbol{X}'\boldsymbol{X}$也应没有量纲), 总假设自变量与因变量均已标准化, 因此这里的设计矩阵$\boldsymbol{X}$是$n \times p$矩阵.

下面讨论岭估计的性质.

性质1: $\hat{\boldsymbol{\beta}}(k)$是$\boldsymbol{\beta}$的有偏估计, 即对任意的$k > 0$, $\mathsf{E}(\hat{\boldsymbol{\beta}}(k)) \neq \boldsymbol{\beta}$.

证明: 容易, 略. □

有偏性是岭估计与最小二乘估计之间的一个重要的不同之处. 一个估计的均方误差由方差之和和偏差的平方和组成. 当存在多重共线性时, 最小二乘估计虽然保持偏差部分为零, 但它的方差部分却很大, 最终导致它的均方误差很大. 引进岭估计的目的是牺牲无偏性, 换取方差部分的大幅度减少, 最终降低其均方误差, 使点估计具有一定的稳定性.

性质2: $\hat{\boldsymbol{\beta}}(k)$是最小二乘估计$\hat{\boldsymbol{\beta}}$的一个线性变换.

证明: 只需注意到

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) =& (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ =& (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X} \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\end{aligned}$$

$$=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

即可. □

性质3: 对任意的$k>0$, 若$\|\hat{\boldsymbol{\beta}}\|\neq 0$, 则我们总有$\|\hat{\boldsymbol{\beta}}(k)\|<\|\hat{\boldsymbol{\beta}}\|$. 即岭估计是把最小二乘估计$\hat{\boldsymbol{\beta}}$向原点作适度的压缩而得到的, 岭估计是一个压缩有偏估计.

证明: 考虑多元线性回归模型$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$, 令

$$\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{P},\ \boldsymbol{\alpha}=\boldsymbol{P}'\boldsymbol{\beta},$$

其中$\boldsymbol{P}$为正交矩阵, 满足

$$\boldsymbol{P}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{P}=\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_p),$$

$\lambda_1,\lambda_2,\cdots,\lambda_p>0$为$\boldsymbol{X}'\boldsymbol{X}$的特征根. 这时, 多元线性回归模型可写为

$$\boldsymbol{Y}=\boldsymbol{Z}\boldsymbol{\alpha}+\boldsymbol{e},\ \mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n. \tag{3.8.2}$$

称(3.8.2)为线性回归模型的典则形式, 称$\boldsymbol{\alpha}$为典则回归系数. 注意到$\boldsymbol{Z}'\boldsymbol{Z}=\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P}=\boldsymbol{\Lambda}$, 所以

$$\hat{\boldsymbol{\alpha}}=(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}=\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{Y}.$$

而

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{Y}=\boldsymbol{P}\hat{\boldsymbol{\alpha}}.$$

它们相应的岭估计分别为

$$\begin{aligned}\hat{\boldsymbol{\alpha}}(k)=&(\boldsymbol{Z}'\boldsymbol{Z}+k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{Y}=(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{Y},\\ \hat{\boldsymbol{\beta}}(k)=&(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}\\ =&\boldsymbol{P}\boldsymbol{P}'(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{Y}\\ =&\boldsymbol{P}\hat{\boldsymbol{\alpha}}(k).\end{aligned}$$

因此

$$\|\hat{\boldsymbol{\beta}}(k)\|=\|\hat{\boldsymbol{\alpha}}(k)\|=\|(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}\|<\|\hat{\boldsymbol{\alpha}}\|=\|\hat{\boldsymbol{\beta}}\|,$$

得证. □

注: 易知典则回归系数的最小二乘估计(或岭估计)和原回归系数的最小二乘估计(或岭估计)有相同的均方误差, 即

$$\mathrm{MSE}(\hat{\boldsymbol{\alpha}})=\mathrm{MSE}(\hat{\boldsymbol{\beta}}),\ \ \mathrm{MSE}(\hat{\boldsymbol{\alpha}}(k))=\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)). \tag{3.8.3}$$

岭估计的另一重要性质是下面的岭估计存在性定理.

定理3.8.1(岭估计存在性定理) 存在$k>0$使得

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k))<\mathrm{MSE}(\hat{\boldsymbol{\beta}}).$$

即存在$k>0$, 使得在均方误差意义下, 岭估计优于最小二乘估计.

证明: 由(3.8.3), 只需证明存在$k>0$使得

$$\mathrm{MSE}(\hat{\boldsymbol{\alpha}}(k))<\mathrm{MSE}(\hat{\boldsymbol{\alpha}}). \tag{3.8.4}$$

记$f(k)=\mathrm{MSE}(\hat{\boldsymbol{\alpha}}(k)),\ k\geq 0$. 注意$f(0)=\mathrm{MSE}(\hat{\boldsymbol{\alpha}})$. 若能证明$f(k)$在$[0,\infty)$上是连续函数且$f'(0)<0$, 则必存在一个较小的$k>0$使得(3.8.4)成立.

来讨论$f(k)$. 注意到

$$\begin{aligned}\mathsf{E}(\hat{\boldsymbol{\alpha}}(k)) =&(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{Z}'\mathsf{E}(\boldsymbol{Y})\\ =&(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\alpha}\\ =&(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\boldsymbol{\alpha},\end{aligned}$$

且

$$\begin{aligned}\mathsf{Cov}(\hat{\boldsymbol{\alpha}}(k)) =&\sigma^2(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\\ =&\sigma^2(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}.\end{aligned}$$

所以

$$\begin{aligned}f(k) =&\mathrm{MSE}(\hat{\boldsymbol{\alpha}}(k))=\mathrm{tr}\Big[\mathsf{Cov}(\hat{\boldsymbol{\alpha}}(k))\Big]+\|\mathsf{E}(\hat{\boldsymbol{\alpha}}(k))-\boldsymbol{\alpha}\|^2\\ =&\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}+k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^2}\\ =&: f_1(k)+f_2(k).\end{aligned}\tag{3.8.5}$$

显然$f(k)$是$[0,\infty)$上的连续函数. 又

$$f_1'(k)=-2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3},\quad f_1'(0)=-2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}<0\tag{3.8.6}$$

以及

$$f_2'(k)=2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_i^2}{(\lambda_i+k)^3},\quad f_2'(0)=0,\tag{3.8.7}$$

所以$f'(0)=f_1'(0)+f_2'(0)=-2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}<0$. 证毕. □

岭估计存在性定理在理论上证明了存在某个岭估计优于最小二乘估计, 但要找出这个岭参数k是不容易的. 容易看出, 理论上, 最优的k是下列方程的解:

$$\begin{aligned}f'(k) =&-2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}+2k\sum_{i=1}^{p}\frac{\lambda_i\alpha_i^2}{(\lambda_i+k)^3}\\ =&2\sum_{i=1}^{p}\frac{\lambda_i(k\alpha_i^2-\sigma^2)}{(\lambda_i+k)^3}=0.\end{aligned}\tag{3.8.8}$$

这个解的显式表达式不易写出, 且它依赖于未知参数$\alpha_i, i=1,\cdots,p$. 统计学家从其它途径提出了选择岭参数k的方法.

下面介绍岭参数选择方法

(1) Hoerl-Kennard公式

Hoerl和Kennard提出的选择k的公式是

$$\hat{k} = \frac{\hat{\sigma}^2}{\max\limits_i \hat{\alpha}_i^2}. \tag{3.8.9}$$

获得这个公式的想法如下: 由(3.8.8)知, 若$k\alpha_i^2 - \sigma^2 < 0$对所有的$i = 1, \cdots, p$都成立, 则$f'(k) < 0$. 于是取

$$k^* = \frac{\sigma^2}{\max\limits_i \alpha_i^2},$$

当$0 < k < k^*$时, $f'(k)$总是小于零, 因而$f(k)$在$(0, k^*)$上是单调递减函数, 再由$f(k)$在$[0, \infty)$的连续性得$f(k^*) < f(0)$. 然后再用$\hat{\alpha}_i$和$\hat{\sigma}^2$分别代替α_i和σ^2, 便得(3.8.9).

(2) 岭迹法

将$\hat{\beta}_1(k), \cdots, \hat{\beta}_p(k)$的岭迹画在一张图上, 根据岭迹的变化趋势选择$k$. 以下是几条选择$k$的准则:

(a) 各回归系数的岭估计大致比较稳定;

(b) 用最小二乘估计时符号不合理的回归系数, 其岭估计的符号将变得合理;

(c) 残差平方和不要上升太多.

一般情况下, 选择能使得各条岭迹都开始趋于稳定的最小k值(注意到$k = 0$时, 最小二乘估计是无偏的).

注: 若我们建模的主要目标是为了预测, 那么可用交叉验证的方法选择岭参数.

例3.8.1 法国经济工作者希望通过国内总产值x_1, 存储量x_2, 总消费量x_3去预测进口总额y, 以上变量的单位均为十亿法郎. 为此收集了1949-1959共11年的数据, 见表3.8.1.

年份	x_1	x_2	x_3	y
1949	149.3	4.2	108.1	15.9
1950	161.2	4.1	114.8	16.4
1951	171.5	3.1	123.2	19.0
1952	175.5	3.1	126.9	19.1
1953	180.8	1.1	132.1	18.8
1954	190.7	2.2	137.7	20.4
1955	202.1	2.1	146.0	22.7
1956	212.4	5.6	154.1	26.5
1957	226.1	5.0	162.3	28.1
1958	231.9	5.1	164.3	27.6
1959	239.0	0.7	167.6	26.3

表3.8.1 法国经济数据

数据分析的R代码如下:

```
1 > yx=read.table("***.txt")
```

```
> x1=yx[,2]
> x2=yx[,3]
> x3=yx[,4]
> y=yx[,5]
> mean(x1);mean(x2);mean(x3);mean(y)
> sd(x1);sd(x2);sd(x3);sd(y)
> economy=data.frame(x1,x2,x3,y)
> lm.sol=lm(y~.,data=economy)
> summary(lm.sol)
```

首先获得各变量的样本均值、样本标准差:

```
> mean(x1);mean(x2);mean(x3);mean(y)
[1] 194.5909
[1] 3.3
[1] 139.7364
[1] 21.89091
> sd(x1);sd(x2);sd(x3);sd(y)
[1] 29.99952
[1] 1.649242
[1] 20.6344
[1] 4.543667
```

最小二乘回归的输出结果如下:

```
> summary(lm.sol)

Call:
lm(formula = y ~ ., data = economy)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52367 -0.38953  0.05424  0.22644  0.78313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.12799    1.21216  -8.355  6.9e-05 ***
x1           -0.05140    0.07028  -0.731 0.488344
x2            0.58695    0.09462   6.203 0.000444 ***
x3            0.28685    0.10221   2.807 0.026277 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’ . 0.1 ‘ ’ 1

Residual standard error: 0.4889 on 7 degrees of freedom
Multiple R-squared:  0.9919,    Adjusted R-squared:  0.9884
F-statistic: 285.6 on 3 and 7 DF,  p-value: 1.112e-07
```

由此得到回归方程:

$$\hat{y} = -10.128 - 0.051x_1 + 0.587x_2 + 0.287x_3.$$

回归分析的结果表明: x_1的回归系数的估计是负数, 这不符合其经济意义, 因为法国是一个原材料进口国, 当国内总产值x_1增加时, 进口总额y也应增加. 所以, 回归系数的符号与实际不符. 其原因是三个自变量之间存在着多重共线性, 这可简单地从x_1, x_2, x_3的样本相关系数矩阵看出.

```
> X=cbind(x1,x2,x3)
> rho=cor(X)
> rho
           x1         x2         x3
x1 1.00000000 0.02585067 0.99726069
x2 0.02585067 1.00000000 0.03567322
x3 0.99726069 0.03567322 1.00000000
```

可以看出: 基于当前数据, x_1与x_3存在高度的线性相关性(样本自相关系数为0.997). 进行多重共线性诊断, 发现自变量之间的确存在着多重共线性.

```
> vif(lm.sol)
      x1       x2       x3
186.0000   1.0189 186.1100
> eigen(rho)
eigen() decomposition
$values
[1] 1.999154934 0.998154176 0.002690889

$vectors
           [,1]        [,2]         [,3]
[1,] 0.70633041  0.03568867  0.706982083
[2,] 0.04350059 -0.99902908  0.006970795
[3,] 0.70654444  0.02583046 -0.707197102

> kappa(rho,exact=TRUE)
[1] 742.9346
```

接下来用岭估计方法寻找岭回归方程. 先对数据进行标准化, R代码如下:

```
> yxs=scale(economy)
> x1=yxs[,1]
> x2=yxs[,2]
> x3=yxs[,3]
> y=yxs[,4]
> economy2=data.frame(x1,x2,x3,y)
```

接下来进行岭估计, 并画出岭迹图:

```
> library(MASS)
> rr.sol=lm.ridge(y~0+x1+x2+x3,data=economy2,lambda=seq(0,1,by=0.05))
> rr.sol  #显示岭估计
             x1        x2        x3
0.00 -0.3393426 0.2130484 1.3026815
0.05  0.1745782 0.2171918 0.7864459
0.10  0.2911301 0.2174252 0.6677096
0.15  0.3421956 0.2170209 0.6144891
0.20  0.3705608 0.2164006 0.5839853
0.25  0.3884063 0.2156844 0.5640136
0.30  0.4005218 0.2149193 0.5497830
0.35  0.4091733 0.2141277 0.5390267
0.40  0.4155718 0.2133213 0.5305332
0.45  0.4204234 0.2125070 0.5235962
0.50  0.4241675 0.2116891 0.5177761
0.55  0.4270918 0.2108702 0.5127851
0.60  0.4293927 0.2100524 0.5084265
0.65  0.4312087 0.2092367 0.5045619
0.70  0.4326404 0.2084241 0.5010906
0.75  0.4337625 0.2076154 0.4979377
0.80  0.4346318 0.2068108 0.4950464
0.85  0.4352922 0.2060109 0.4923728
0.90  0.4357779 0.2052158 0.4898824
0.95  0.4361164 0.2044258 0.4875479
1.00  0.4363296 0.2036409 0.4853472
> matplot(rr.sol$lambda,t(rr.sol$coef),type="l",lwd=2,xlab=expression(
  lambda),ylab=expression(hat(beta)(lambda)))  #直接用命令plot(rr.sol)可
  得到一张稍粗糙的岭迹图
```

输出的岭迹图见图3.8.1. 根据岭迹图选择岭参数$k = 0.4$, 标准化变量的岭回归方程为

$$\hat{v} = 0.416u_1 + 0.213u_2 + 0.531u_3.$$

最后, 需要转换成原始变量的岭回归方程:

$$\begin{aligned}\frac{\hat{y} - 21.891}{4.544} =& 0.416 \times \frac{x_1 - 194.591}{30.000} + 0.213 \times \frac{x_2 - 3.300}{1.649} \\ &+ 0.531 \times \frac{x_3 - 139.736}{20.634},\end{aligned}$$

即,

$$\hat{y} = -8.655 + 0.063x_1 + 0.587x_2 + 0.117x_3.$$

注: 当所有变量的量纲都一致时, 可以直接对原始数据进行岭估计. 例如, 在这个例子里直接进行岭估计, 根据岭迹图选择岭参数$k = 0.4$, 可直接得到岭回归方程: $\hat{y} = -8.621 + 0.063x_1 + 0.588x_2 + 0.117x_3$.

```
> rr.sol=lm.ridge(y~.,data=economy,lambda=seq(0,1,by=0.05))
> rr.sol
> plot(rr.sol)
```

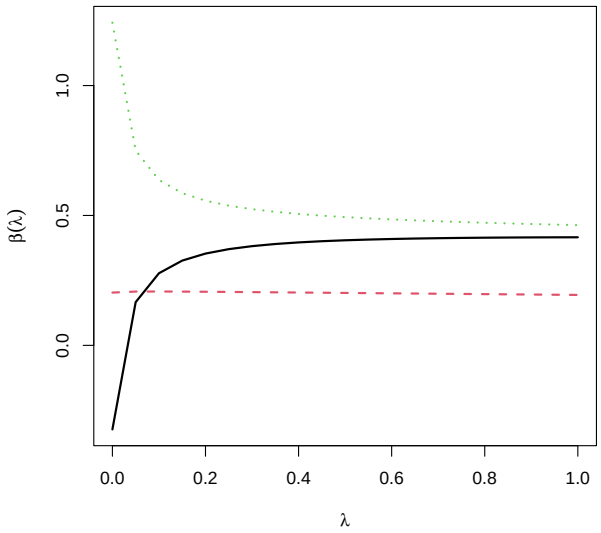


图3.8.1　岭迹图

来分析岭估计的几何意义. 前面已证明岭估计$\hat{\boldsymbol{\beta}}(k)$是最小二乘估计$\hat{\boldsymbol{\beta}}$的一种压缩. 如果现在已经有了$\hat{\boldsymbol{\beta}}$, 希望将它的长度压缩到原来的$c$倍$(0<c<1)$, 并使残差平方和的上升尽可能小, 那么可以证明, 这样的估计就是岭估计. 接下来就来说明这件事情.

设$\boldsymbol{b}$为$\boldsymbol{\beta}$的任一估计, 相应的残差平方和为

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{b}) =&\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|^2\\ =&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{b})\|^2\\ =&|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+(\hat{\boldsymbol{\beta}}-\boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{b}).\end{aligned}$$

将$\hat{\boldsymbol{\beta}}$的长度压缩到原来的c倍且使残差平方和的上升最小, 等价于解下列的极值问题:

$$\min_{\boldsymbol{b}}(\boldsymbol{b}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{\beta}}),\quad \text{s.t.}\quad \|\boldsymbol{b}\|^2=c^2\|\hat{\boldsymbol{\beta}}\|^2. \tag{3.8.10}$$

设$\boldsymbol{P}$为正交矩阵, 满足

$$\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P}=\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_p),$$

其中, $\lambda_1,\lambda_2,\cdots,\lambda_p>0$为$\boldsymbol{X}'\boldsymbol{X}$的特征根. 记

$$\boldsymbol{\alpha}=\boldsymbol{P}'\boldsymbol{\beta},\quad \boldsymbol{d}=\boldsymbol{P}'\boldsymbol{b},\quad \hat{\boldsymbol{\alpha}}=\boldsymbol{P}'\hat{\boldsymbol{\beta}}.$$

则(3.8.10)等价于

$$\min_{\boldsymbol{d}}(\boldsymbol{d}-\hat{\boldsymbol{\alpha}})'\boldsymbol{\Lambda}(\boldsymbol{d}-\hat{\boldsymbol{\alpha}}),\ \ \text{s.t.}\ \ \|\boldsymbol{d}\|^2=c^2\|\hat{\boldsymbol{\alpha}}\|^2. \tag{3.8.11}$$

应用Lagrange乘子法, 构造辅助函数

$$F(\boldsymbol{d},k)=(\boldsymbol{d}-\hat{\boldsymbol{\alpha}})'\boldsymbol{\Lambda}(\boldsymbol{d}-\hat{\boldsymbol{\alpha}})+k(\boldsymbol{d}'\boldsymbol{d}-c^2\|\hat{\boldsymbol{\alpha}}\|^2),$$

其中k为Lagrange乘子($k\neq 0$). 对上式关于$\boldsymbol{d}$求导, 得

$$\frac{\partial F(\boldsymbol{d},k)}{\partial \boldsymbol{d}}=2(\boldsymbol{\Lambda}+k\boldsymbol{I})\boldsymbol{d}-2\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}.$$

令其等于$\mathbf{0}$, 解得

$$\boldsymbol{d}=(\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}. \tag{3.8.12}$$

所以

$$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

剩下的问题是证明(3.8.12)中的$k>0$(k是c的函数). 将(3.8.12)代入(3.8.11)的目标函数, 记之为$Q(k)$. 于是

$$\begin{aligned}
Q(k)=&(\boldsymbol{d}-\hat{\boldsymbol{\alpha}})'\boldsymbol{\Lambda}(\boldsymbol{d}-\hat{\boldsymbol{\alpha}})\\
=&\hat{\boldsymbol{\alpha}}'[((\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}-\boldsymbol{I})'\boldsymbol{\Lambda}((\boldsymbol{\Lambda}+k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}-\boldsymbol{I})]\hat{\boldsymbol{\alpha}}\\
=&k^2\cdot\hat{\boldsymbol{\alpha}}'\mathrm{diag}\left(\frac{\lambda_1}{(\lambda_1+k)^2},\cdots,\frac{\lambda_p}{(\lambda_p+k)^2}\right)\hat{\boldsymbol{\alpha}}\\
=&k^2\sum_{i=1}^{p}\frac{\lambda_i\hat{\alpha}_i^2}{(\lambda_i+k)^2}.
\end{aligned}$$

由于$\lambda_1,\cdots,\lambda_p$都为正数, 因此对$k>0$, 有$(\lambda_i+k)^2>(\lambda_i-k)^2$, $i=1,\cdots,p$. 所以, $Q(k)<Q(-k)$. 这说明$Q(k)$的极小值不会在$(-\infty,0)$上达到. 这就证明了我们所要的结论.

从几何上来说, (3.8.10)的约束条件$\|\boldsymbol{b}\|^2=c^2\|\hat{\boldsymbol{\beta}}\|^2=:h^2$是一个中心在原点, 半径为$h$的球面. 对目标函数$(\boldsymbol{b}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{\beta}})$作椭球

$$(\boldsymbol{b}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{\beta}})=\delta^2. \tag{3.8.13}$$

因$0<c<1$, 所以$\hat{\boldsymbol{\beta}}$在$\|\boldsymbol{b}\|^2=c^2\|\hat{\boldsymbol{\beta}}\|^2=h^2$的球面之外. 故总可找到$\delta>0$使得球$\|\boldsymbol{b}\|^2=h^2$和椭球(3.8.13)相切于某点$\tilde{\boldsymbol{\beta}}$. 显然这个$\tilde{\boldsymbol{\beta}}$就是极值问题(3.8.10)的解, 也就是岭估计$\hat{\boldsymbol{\beta}}(k)$, 如图3.8.2所示(以二维为例).

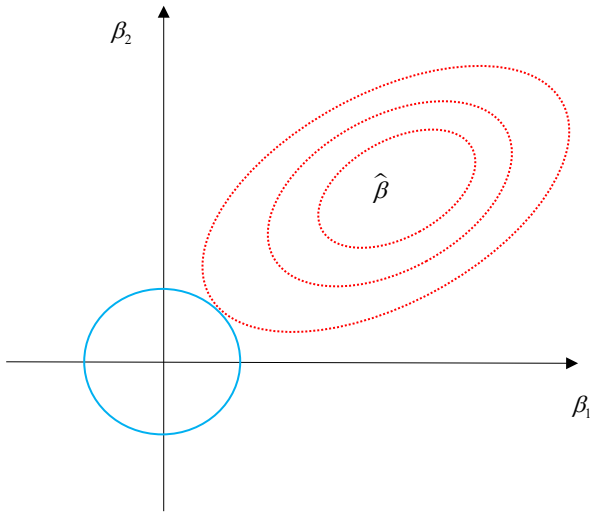


图3.8.2　岭估计的几何意义

容易看出(3.8.10)与下述优化问题等价:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\},$$

其中$\lambda \geq 0$为调节参数(tuning parameter). 在上述优化问题中, $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$为损失函数(loss function), $\|\boldsymbol{\beta}\|^2$为惩罚函数(penalty function), λ在损失和惩罚之间控制权衡.

3.9　主成分估计

主成分(principle component)估计是由Massy于1965年提出的另一种有偏估计, 目的是为了克服设计矩阵$\boldsymbol{X}$为病态矩阵时最小二乘估计的稳定性将变得很差这一缺陷.

主成分估计的基本思想是: (1) 首先借助正交变换将回归自变量变为对应的主成分(要求主成分的观测向量是正交的, 且某些观测向量近似为$\mathbf{0}$向量); (2) 从所有的主成分中删去观测向量近似为$\mathbf{0}$的那些主成分(起到消除多重共线性以及降维的双重作用); (3) 将保留下来的主成分作为新的回归自变量建立回归模型, 用最小二乘法估计模型中的回归系数并得到主成分回归方程. 基于得到的主成分回归方程再将它们转换为原始变量的回归方程.

为了消除量纲的影响, 假设自变量与因变量均已标准化. 考虑回归模型:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \mathsf{E}(\boldsymbol{e}) = \mathbf{0}, \quad \mathsf{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n, \tag{3.9.1}$$

其中$\boldsymbol{X}$是$n \times p$设计矩阵. 记$\lambda_1 \geq \cdots \geq \lambda_p > 0$为$\boldsymbol{X}'\boldsymbol{X}$的特征根, $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_p$为对

应的标准正交化特征向量. 则

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_p)$$

为$p \times p$正交矩阵且

$$\boldsymbol{\Phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Phi} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p) =: \boldsymbol{\Lambda}.$$

再记$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Phi}$, $\boldsymbol{\alpha} = \boldsymbol{\Phi}'\boldsymbol{\beta}$, 则模型(3.9.1)可改写为

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{e}, \quad \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \quad \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n. \tag{3.9.2}$$

在上述的线性回归模型的典则形式中, 新的设计矩阵

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_p) = (\boldsymbol{X}\boldsymbol{\phi}_1, \cdots, \boldsymbol{X}\boldsymbol{\phi}_p).$$

若记$\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$, 则新自变量的观测向量与原始自变量的观测向量的关系为

$$\boldsymbol{z}_j = \phi_{1j}\boldsymbol{x}_1 + \cdots + \phi_{pj}\boldsymbol{x}_p, \quad j = 1, \cdots, p.$$

这是对原始自变量的观测向量的一个线性变换, 变换的系数向量是特征根λ_j所对应的标准正交化特征向量.

统计上, 称观测向量$\boldsymbol{z}_j, j = 1, \cdots, p$, 对应的新自变量$z_j, j = 1, \cdots, p$, 为$p$个主成分. 每个主成分都是原始自变量的线性组合:

$$z_j = \phi_{1j}x_1 + \cdots + \phi_{pj}x_p, \quad j = 1, \cdots, p.$$

主成分具有良好的性质.

性质1 任意两个主成分的观测向量是正交的, 且第j个主成分的偏差平方和$\sum_{i=1}^n (z_{ij} - \bar{z}_j)^2 = \lambda_j$.

证明: 因为$\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{\Phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Phi} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$, 所以

$$\boldsymbol{z}_j'\boldsymbol{z}_k = 0, \quad \forall j \neq k$$

且$\boldsymbol{z}_j'\boldsymbol{z}_j = \lambda_j$, $j = 1, \cdots, p$. 又因为$\boldsymbol{X}$是标准化设计矩阵, 所以

$$\bar{z}_j = \frac{1}{n}\sum_{i=1}^n z_{ij} = \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^p \phi_{kj}x_{ik} = \frac{1}{n}\sum_{k=1}^p \phi_{kj}\sum_{i=1}^n x_{ik} = 0.$$

因此有

$$\sum_{i=1}^n (z_{ij} - \bar{z}_j)^2 = \sum_{i=1}^n z_{ij}^2 = \boldsymbol{z}_j'\boldsymbol{z}_j = \lambda_j, \quad j = 1, \cdots, p.$$

证毕. □

λ_j度量了第j个主成分z_j的取值变动大小. 因为$\lambda_1 \geq \cdots \geq \lambda_p > 0$, 所以称$z_1$为第一主成分, z_2为第二主成分, $\cdots$. 这p个主成分的观测向量是相互正交的. 由上述性质可知, z_1对因变量的解释能力最强, z_2次之, $\cdots$, z_p最弱.

若设计矩阵$\boldsymbol{X}$是病态矩阵, 那么有一些$\boldsymbol{X}'\boldsymbol{X}$的特征根很小, 不妨假设

$$\lambda_{r+1}, \cdots, \lambda_p \approx 0.$$

这时, 后面的$p-r$个主成分的取值变动很小且均在零附近取值. 所以这$p-r$个主成分对因变量的影响可以忽略, 可将它们从回归模型中剔除. 剩下的主成分$z_1,\cdots,z_r$就不存在多重共线性问题了. 用最小二乘法对剩下的r个主成分(即r个新的自变量)关于因变量作回归即可. 最后再变回到原始变量的回归方程.

所以, 主成分回归的建模步骤包括: (1) 正交变换; (2) 降维(消除多重共线性, 减少计算量); (3) 建立回归方程. 前面已描述了如何完成步骤(1)和步骤(2), 接下来分析步骤(3).

对$\boldsymbol{\Lambda},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{\Phi}$作分块,

$$\boldsymbol{\Lambda}=\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_2\end{pmatrix},\ \boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{pmatrix},\ \boldsymbol{Z}=(\boldsymbol{Z}_1\ \ \boldsymbol{Z}_2),\ \boldsymbol{\Phi}=(\boldsymbol{\Phi}_1\ \ \boldsymbol{\Phi}_2),$$

其中$\boldsymbol{\Lambda}_1$为$r\times r$矩阵, $\boldsymbol{\alpha}_1$为$r\times 1$向量, $\boldsymbol{Z}_1$为$n\times r$矩阵, $\boldsymbol{\Phi}_1$为$p\times r$矩阵. 因为$\boldsymbol{Z}_2$近似是$\boldsymbol{0}$矩阵, 所以剔除$\boldsymbol{Z}_2\boldsymbol{\alpha}_2$, 模型(3.9.2)变为

$$\boldsymbol{Y}=\boldsymbol{Z}_1\boldsymbol{\alpha}_1+\boldsymbol{e},\ \ \mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\ \ \mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n. \tag{3.9.3}$$

$\boldsymbol{Z}_1$不是病态矩阵(因为$\boldsymbol{Z}_1'\boldsymbol{Z}_1$的特征根为$\lambda_1,\cdots,\lambda_r$, 均远离0), 所以可直接应用最小二乘法求得$\boldsymbol{\alpha}_1$的LSE

$$\hat{\boldsymbol{\alpha}}_1=(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Y}=\boldsymbol{\Lambda}_1^{-1}\boldsymbol{Z}_1'\boldsymbol{Y}.$$

从模型中剔除了后面的$p-r$个主成分, 这相当于用$\hat{\boldsymbol{\alpha}}_2=\boldsymbol{0}$去估计$\boldsymbol{\alpha}_2$. 利用关系$\boldsymbol{\beta}=\boldsymbol{\Phi}\boldsymbol{\alpha}$, 得$\boldsymbol{\beta}$的主成分估计

$$\tilde{\boldsymbol{\beta}}=\boldsymbol{\Phi}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_1\\ \hat{\boldsymbol{\alpha}}_2\end{pmatrix}=(\boldsymbol{\Phi}_1\ \boldsymbol{\Phi}_2)\begin{pmatrix}\hat{\boldsymbol{\alpha}}_1\\ \boldsymbol{0}\end{pmatrix}=\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{Z}_1'\boldsymbol{Y}=\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{X}'\boldsymbol{Y},$$

相应的主成分回归方程为$\hat{\boldsymbol{Y}}=\boldsymbol{X}\tilde{\boldsymbol{\beta}}$.

总结一下主成分回归建模的过程:

(1) 做正交变换$\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{\Phi}$, 获得新的自变量, 称为主成分;

(2) 做回归自变量选择, 剔除特征根比较小的对应主成分;

(3) 将剩余的主成分对(标准化后的)y做最小二乘回归, 得到主成分回归方程, 再把这个回归方程转换成原始变量的回归方程.

主成分估计有以下的一些性质:

性质2 $\tilde{\boldsymbol{\beta}}=\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'\hat{\boldsymbol{\beta}}$, 即主成分估计是最小二乘估计的一个线性变换.

证明: 根据

$$\boldsymbol{\Phi}_1'\boldsymbol{\Phi}_1=\boldsymbol{I}_r,\ \boldsymbol{\Phi}_1'\boldsymbol{\Phi}_2=\boldsymbol{0}$$

及

$$\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}'=\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1'+\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2'$$

可知

$$\begin{aligned}\tilde{\boldsymbol{\beta}}=&\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{X}'\boldsymbol{Y}\\ =&\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\\ =&\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1'\hat{\boldsymbol{\beta}}+\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2'\hat{\boldsymbol{\beta}}\end{aligned}$$

$$
\begin{aligned}
&=\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1'\hat{\boldsymbol{\beta}}\\
&=\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'\hat{\boldsymbol{\beta}},
\end{aligned}
$$

这就是所要证明的结论. □

性质3 $\mathsf{E}(\tilde{\boldsymbol{\beta}})=\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'\boldsymbol{\beta}$. 即只要$r<p$, 主成分估计就是有偏估计.

证明: 只需注意到$\mathsf{E}(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$即可. □

性质4 $\|\tilde{\boldsymbol{\beta}}\|\leq\|\hat{\boldsymbol{\beta}}\|$, 即主成分估计是压缩估计.

证明: 构造$p\times p$矩阵$\tilde{\boldsymbol{I}}=\mathrm{diag}(\boldsymbol{I}_r,\boldsymbol{0})$, 则由$\boldsymbol{\Phi}$的定义知

$$
\boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'=\boldsymbol{\Phi}\tilde{\boldsymbol{I}}\boldsymbol{\Phi}'.
$$

从而有

$$
\|\tilde{\boldsymbol{\beta}}\|=\|\boldsymbol{\Phi}\tilde{\boldsymbol{I}}\boldsymbol{\Phi}'\hat{\boldsymbol{\beta}}\|=\|\tilde{\boldsymbol{I}}\boldsymbol{\Phi}'\hat{\boldsymbol{\beta}}\|\leq\|\boldsymbol{\Phi}'\hat{\boldsymbol{\beta}}\|=\|\hat{\boldsymbol{\beta}}\|.
$$

证毕. □

定理3.9.1 当原始自变量存在足够严重的多重共线性时, 适当选择保留的主成分个数可使主成分估计比最小二乘估计有较小的均方误差, 即

$$
\mathrm{MSE}(\tilde{\boldsymbol{\beta}})<\mathrm{MSE}(\hat{\boldsymbol{\beta}}).
$$

证明: 假设$\boldsymbol{X}'\boldsymbol{X}$的后$p-r$个特征根$\lambda_{r+1},\cdots,\lambda_p$接近于零. 不难看出

$$
\begin{aligned}
\mathrm{MSE}(\tilde{\boldsymbol{\beta}})=&\mathrm{MSE}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_1\\ \boldsymbol{0}\end{pmatrix}\\
=&\mathrm{tr}\Big[\mathsf{Cov}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_1\\ \boldsymbol{0}\end{pmatrix}\Big]+\Big\|\mathsf{E}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_1\\ \boldsymbol{0}\end{pmatrix}-\boldsymbol{\alpha}\Big\|^2\\
=&\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}_1^{-1})+\|\boldsymbol{\alpha}_2\|^2.
\end{aligned}
$$

因为

$$
\mathrm{MSE}(\hat{\boldsymbol{\beta}})=\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}^{-1})=\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}_1^{-1})+\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}_2^{-1}),
$$

所以

$$
\mathrm{MSE}(\tilde{\boldsymbol{\beta}})=\mathrm{MSE}(\hat{\boldsymbol{\beta}})+(\|\boldsymbol{\alpha}_2\|^2-\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}_2^{-1})).
$$

于是

$$
\mathrm{MSE}(\tilde{\boldsymbol{\beta}})<\mathrm{MSE}(\hat{\boldsymbol{\beta}})
$$

当且仅当

$$
\|\boldsymbol{\alpha}_2\|^2<\sigma^2\mathrm{tr}(\boldsymbol{\Lambda}_2^{-1})=\sigma^2\sum_{i=r+1}^{p}\frac{1}{\lambda_i}. \tag{3.9.4}
$$

当多重共线性足够严重的时候, $\lambda_{r+1},\cdots,\lambda_p$中的某一个可以充分接近于零. 因此上式右端可以足够大使得不等式(3.9.4)成立. □

注意到$\boldsymbol{\alpha}_2=\boldsymbol{\Phi}_2'\boldsymbol{\beta}$, 那么(3.9.4)可写为

$$
\Big(\frac{\boldsymbol{\beta}}{\sigma}\Big)'\boldsymbol{\Phi}_2\boldsymbol{\Phi}_2'\Big(\frac{\boldsymbol{\beta}}{\sigma}\Big)<\mathrm{tr}(\boldsymbol{\Lambda}_2^{-1}). \tag{3.9.5}
$$

也就是说, 当$\boldsymbol{\beta}$和σ满足(3.9.5)时, 主成分估计才比最小二乘估计有较小的均方误差. (3.9.5)表示参数空间中(视$\boldsymbol{\beta}/\sigma$为参数)一个中心在原点的椭球. 于是从(3.9.5)可得如下结论:

(1) 对固定的参数$\boldsymbol{\beta}$和σ^2, 当$\boldsymbol{X}'\boldsymbol{X}$的后$p-r$个特征根比较小时, 主成分估计比最小二乘估计有较小的均方误差;

(2) 对给定的$\boldsymbol{X}'\boldsymbol{X}$, 即固定的$\boldsymbol{\Lambda}_2$, 对相对较小(指绝对值)的$\boldsymbol{\beta}/\sigma$, 主成分估计比最小二乘估计有较小的均方误差.

如何选取保留的主成分个数r呢? 下面介绍两种常用的选取方法.

(1) 略去特征根接近于零的对应主成分;

(2) 选择r使得前r个特征根之和在p个特征根总和中所占的比例(称为累计贡献率)达到预先给定的值. 譬如, 选择最小的r使得

$$\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{p}\lambda_i} > 0.85.$$

在上述两种选取方法中, 第一种方法有一定的主观性, 而第二种方法显得更客观一些.

例3.9.1 (续例3.8.1)法国经济数据分析问题. 在上一节中, 已说明自变量之间具有多重共线性. 接下来进行主成分回归分析. R代码及分析结果如下:

```
> economy.pr=princomp(~x1+x2+x3,data=economy,cor=TRUE)
> summary(economy.pr,loadings=TRUE)
Importance of components:
                          Comp.1    Comp.2       Comp.3
Standard deviation     1.413915 0.9990767 0.0518737839
Proportion of Variance 0.666385 0.3327181 0.0008969632
Cumulative Proportion  0.666385 0.9991030 1.0000000000

Loadings:
   Comp.1 Comp.2 Comp.3
x1  0.706         0.707
x2        -0.999
x3  0.707        -0.707
```

可以看到第三个特征根$\lambda_3 = 0.0518737839^2 = 0.00269 \approx 0$. 三个标准正交化特征向量分别是:

$$\begin{aligned}\boldsymbol{\phi}_1 =&(0.706, 0, 0.707)',\\ \boldsymbol{\phi}_2 =&(0, -0.999, 0)',\\ \boldsymbol{\phi}_3 =&(0.707, 0, -0.707)'.\end{aligned}$$

三个主成分分别是:

$$\begin{aligned}z_1 =&0.706x_1^* + 0.707x_3^*,\\ z_2 =&-0.999x_2^*,\\ z_3 =&0.707x_1^* - 0.707x_3^*.\end{aligned}$$

因为第一个特征根的累计贡献率为0.666385 $\leq$ 0.85, 前两个特征根的累计贡献率0.9991030 $>$ 0.85, 所以删去第三个主成分, 只保留前两个主成分.

下面计算主成分得分(即主成分的观测向量), R代码及计算结果如下:

```
> pre=predict(economy.pr)
> pre
       Comp.1      Comp.2      Comp.3
1  -2.2296493 -0.66983032  0.02173374
2  -1.6979452 -0.58265445  0.07458412
3  -1.1695976  0.07654175  0.02279070
4  -0.9379462  0.08639036 -0.01134096
5  -0.6756511  1.37046303 -0.07612514
6  -0.1996423  0.69131968 -0.02784852
7   0.3771746  0.77997236 -0.04486935
8   1.0192344 -1.42014882 -0.06593076
9   1.6354243 -1.01109953 -0.02472510
10  1.8532401 -1.06476864  0.04718400
11  2.0253583  1.74381457  0.08454728
```

进行主成分估计, R代码及分析结果如下:

```
> z1=pre[,1];z2=pre[,2]
> u=economy2[,4]
> economy3=data.frame(u,z1,z2)
> pc.sol=lm(u~0+z1+z2,data=economy3)
> summary(pc.sol)

Call:
lm(formula = u ~ 0 + z1 + z2, data = economy3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19772 -0.05733  0.01857  0.07852  0.14716

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
z1  0.65787    0.02434  27.032 6.28e-10 ***
z2 -0.18240    0.03444  -5.296 0.000497 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’ . 0.1 ‘ ’ 1

Residual standard error: 0.1141 on 9 degrees of freedom
Multiple R-squared:  0.9883,    Adjusted R-squared:  0.9857
F-statistic: 379.4 on 2 and 9 DF,  p-value: 2.044e-09
```

得主成分回归方程:

$$\hat{u} = 0.65787 z_1 - 0.1824 z_2$$

$$
\begin{aligned}
&=0.65787 \times (0.706x_1^* + 0.707x_3^*) - 0.1824 \times (-0.999x_2^*) \\
&=0.46446x_1^* + 0.18222x_2^* + 0.46511x_3^*.
\end{aligned}
$$

注意以上为标准化变量的回归方程. 转化为原始变量的回归方程, 得

$$
\begin{aligned}
\frac{\hat{y} - 21.891}{4.544} =& 0.46446 \times \frac{x_1 - 194.591}{30.000} + 0.18222 \times \frac{x_2 - 3.300}{1.649} \\
&+ 0.46511 \times \frac{x_3 - 139.736}{20.634},
\end{aligned}
$$

即

$$
\hat{y} = -7.768 + 0.070x_1 + 0.502x_2 + 0.102x_3.
$$

这里的y, x_1, x_2, x_3表示原始变量.

表3.9.1给出了最小二乘估计、岭估计和主成分估计的比较. 总的来说, 岭估计和主成分估计比较相近. 跟最小二乘估计相比, 岭估计和主成分估计都消除或缓解了多重共线性所带来的影响, 所以x_1的回归系数的点估计的符号也发生了变化.

方法	β_0	β_1	β_2	β_3
最小二乘估计	-10.128	-0.051	0.587	0.287
岭估计($k = 0.4$)	-8.655	0.063	0.587	0.117
主成分估计($r = 2$)	-7.768	0.070	0.502	0.102

表3.9.1　最小二乘估计、岭估计与主成分估计的比较

作业

1. 假设要求出4个物体的重量$\mu_1, \cdots, \mu_4$. 一种方法是将每个物体称k次, 然后求平均. 假设称重误差的均值和方差分别为0和σ^2. 现取$k = 5$, 用y_{ij}表示第i个物体第j次称重时得到的重量, $i = 1, \cdots, 4; j = 1, \cdots, 5$.
(1) 试写出相应的线性模型;
(2) 求出μ_i的最小二乘估计$\hat{\mu}_i$, $i = 1, \cdots, 4$;
(3) 计算$\mathsf{Var}(\hat{\mu}_i)$, $i = 1, \cdots, 4$.

2. 设有回归模型

$$
\begin{cases}
y_i = \theta + e_i, & i = 1, \cdots, m, \\
y_{m+i} = \theta + \phi + e_{m+i}, & i = 1, \cdots, m, \\
y_{2m+i} = \theta - 2\phi + e_{2m+i}, & i = 1, \cdots, n,
\end{cases}
$$

其中θ, ϕ是未知参数, 各e_i相互独立, 且服从$N(0, \sigma^2)$.
(1) 写出设计矩阵$\boldsymbol{X}$;
(2) 求θ和ϕ的最小二乘估计$\hat{\theta}$和$\hat{\phi}$;

(3) 证明当$m = 2n$时, $\hat{\theta}$和$\hat{\phi}$不相关.

3. 对于下列的线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

证明: $\boldsymbol{\beta}$的最小二乘估计与极大似然估计是一致的.

4. 设

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + e_i, \quad i = 1, \cdots, n,$$

其中$\{e_i, i = 1, \cdots, n\}$独立同分布, 服从$N(0, \sigma^2)$. 记$\hat{\beta}_1$为$\beta_1$的最小二乘估计, 证明:

$$\mathsf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 (1 - r_{12}^2)},$$

其中r_{12}是$\{(x_{i1}, x_{i2}), i = 1, \cdots, n\}$的样本相关系数.

5. 在动物学研究中, 有时需要找出某种动物的体积与重量的关系. 重量相对容易测量, 而测量体积比较困难. 可以考虑用重量预测体积. 下面是某种动物的18个关于体重x(单位: 公斤)与体积y(单位: 10^{-3}立方米)的已有测量值.

x	y	x	y
17.1	16.7	15.8	15.2
10.5	10.4	15.1	14.8
13.8	13.5	12.1	11.9
15.7	15.7	18.4	18.3
11.9	11.6	17.1	16.7
10.4	10.2	16.7	16.6
15.0	14.5	16.5	15.9
16.0	15.8	15.1	15.1
17.8	17.6	15.1	14.5

(1) 画出散点图;
(2) 求回归直线$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, 并画出回归直线的图像;
(3) 对体重$x_0 = 15.3$的这个动物, 预测它的体积y_0.

6. 设$y_1, \cdots, y_n$是来自$N(\theta, \sigma^2)$的独立同分布样本, 求θ的最小方差线性无偏估计$\hat{\theta}$, 并求$\mathsf{Var}(\hat{\theta})$.

7. 设$y_i \sim N(i\theta, i^2\sigma^2), i = 1, \cdots, n$, 且相互独立, 求$\theta$的最小方差线性无偏估计$\hat{\theta}$, 并求$\mathsf{Var}(\hat{\theta})$.

8. 考虑线性回归模型:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}, \quad \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \quad \mathsf{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n,$$

其中$\boldsymbol{\beta}_1$和$\boldsymbol{\beta}_2$分别是k_1维和k_2维向量. 依据下列的约束条件, 找出$\boldsymbol{\beta}=(\boldsymbol{\beta}_1',\boldsymbol{\beta}_2')'$的约束最小二乘估计:
(1) $\boldsymbol{\beta}_2=\boldsymbol{c}$, 这里$\boldsymbol{c}$是一个已知的$k_2$维向量;
(2) $\boldsymbol{\beta}_1+\boldsymbol{\beta}_2=\boldsymbol{0}$, 其中$k_1=k_2$.
提示: 设$\boldsymbol{A}$为非奇异的对称矩阵, 将其分块为

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{B} & \boldsymbol{C}\\ \boldsymbol{C}' & \boldsymbol{D}\end{pmatrix},$$

则当$\boldsymbol{B}^{-1},\boldsymbol{D}^{-1}$都存在时有

$$\begin{aligned}\boldsymbol{A}^{-1}&=\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{C}_1\\ \boldsymbol{C}_1' & \boldsymbol{D}_1\end{pmatrix}\\ &=\begin{pmatrix}(\boldsymbol{B}-\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}')^{-1} & -\boldsymbol{B}_1\boldsymbol{C}\boldsymbol{D}^{-1}\\ -\boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{B}_1 & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{B}_1\boldsymbol{C}\boldsymbol{D}^{-1}\end{pmatrix}\\ &=\begin{pmatrix}\boldsymbol{B}^{-1}+\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1} & -\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\\ -\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1} & (\boldsymbol{D}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\end{pmatrix}.\end{aligned}$$

9. 设$\boldsymbol{Y}=\boldsymbol{\beta}+\boldsymbol{e},\mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n$. 试用Lagrange乘子法证明: 在约束条件$\boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{0}$下, 使$\|\boldsymbol{Y}-\boldsymbol{\beta}\|^2$达到最小的$\boldsymbol{\beta}$为

$$\hat{\boldsymbol{\beta}}_c=[\boldsymbol{I}_n-\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}]\boldsymbol{Y},$$

其中$\boldsymbol{A}$是已知的$q\times n$矩阵, 其秩为q.

10. 设$\boldsymbol{A}$是$n\times n$的可逆矩阵, $\boldsymbol{u}$和$\boldsymbol{v}$为n维列向量, 试证明:

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

并利用此结论证明以下结论: 设$\boldsymbol{X}_n$, $\boldsymbol{Y}_n$, $\hat{\boldsymbol{\beta}}_n$分别是$p$元(即有$p$个自变量)线性回归模型中基于$n$组观测的设计矩阵、观测向量及$\boldsymbol{\beta}$的最小二乘估计. 现获得了第$n+1$组观测

$$(x_{n+1,1},\cdots,x_{n+1,p},y_{n+1})=(\boldsymbol{x}_{n+1}',y_{n+1}),$$

又记

$$\boldsymbol{X}_{n+1}=\begin{pmatrix}\boldsymbol{X}_n\\ \boldsymbol{x}_{n+1}'\end{pmatrix},\boldsymbol{Y}_{n+1}=\begin{pmatrix}\boldsymbol{Y}_n\\ y_{n+1}\end{pmatrix},$$

$\hat{\boldsymbol{\beta}}_{n+1}$表示基于$n+1$组观测所得到的$\boldsymbol{\beta}$的最小二乘估计, 则有

$$\hat{\boldsymbol{\beta}}_{n+1}=\hat{\boldsymbol{\beta}}_n+\frac{(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}(y_{n+1}-\boldsymbol{x}_{n+1}'\hat{\boldsymbol{\beta}}_n)}{1+\boldsymbol{x}_{n+1}'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}}.$$

11. 在线性回归模型$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e},\mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n$中, 删除第$i$组数据前后的因变量拟合值分别记为$\hat{y}_i$和$\hat{y}_{(i)}$. 证明:

$$\hat{y}_i=h_{ii}y_i+(1-h_{ii})\hat{y}_{(i)},$$

其中h_{ii}是帽子矩阵的第i个对角线元素.

12. 设$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e},\mathsf{E}(\boldsymbol{e})=\boldsymbol{0},\mathsf{Cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n$, $\boldsymbol{X}$是$n\times(p+1)$列满秩设计矩阵. 将$\boldsymbol{X}\boldsymbol{\beta}$写成

$$\boldsymbol{X}\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}.$$

(1) 证明$\boldsymbol{\beta}_2$的最小二乘估计$\hat{\boldsymbol{\beta}}_2$由下式给出:

$$\hat{\boldsymbol{\beta}}_2=[\boldsymbol{X}_2'\boldsymbol{X}_2-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{Y}-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}];$$

(2) 求$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_2)$.

13. 对于线性回归模型$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$, 假设$\boldsymbol{X}$的第一列元素全为1, 证明:
(1) $\sum_{i=1}^n(y_i-\hat{y}_i)=0$;
(2) $\sum_{i=1}^n\hat{y}_i(y_i-\hat{y}_i)=0$,
其中$\hat{y}_i$是拟合值向量$\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$的第$i$个分量, $\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的最小二乘估计.

14. 为了研究水的耗氧量与周围环境的关系, 在实验室条件下, 对连续放置220天的水进行不断的测试, 共作了20次观测. 选取如下变量进行观察: 水的日耗氧量的对数(y), 生物耗氧量(x_1), 总的耗氧量(x_2), 固定物质含量(x_3), 挥发性固定物质含量(x_4), 化学物质耗氧量(x_5). 数据见下表, 试进行回归诊断(包括模型的诊断和数据的诊断).

编号	x_1	x_2	x_3	x_4	x_5	y
1	1125	232	7160	85.9	8905	1.5563
2	920	268	8804	86.5	7388	0.8976
3	835	271	8108	85.2	5348	0.7482
4	1000	237	6370	83.8	8056	0.716
5	1150	192	6441	82.1	6960	0.313
6	990	202	5154	79.2	5690	0.3617
7	840	184	5896	81.2	6932	0.1139
8	650	200	5336	80.6	5400	0.1139
9	640	180	5041	78.4	3177	-0.2218
10	583	165	5012	79.3	4461	-0.1549
11	570	151	4825	78.7	3901	0.0000
12	570	171	4391	78.0	5002	0.0000
13	510	243	4320	72.3	4665	-0.0969
14	555	147	3709	74.9	4642	-0.2218
15	460	286	3969	74.4	4840	-0.3979
16	275	198	3558	72.5	4479	-0.1549
17	510	196	4361	57.7	4200	-0.2218
18	165	210	3301	71.8	3410	-0.3919
19	244	327	2964	72.5	3360	-0.5229
20	79	334	2777	71.9	2599	-0.0458

15. 对于线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},\ \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0},\ \mathsf{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2),$$

记$\boldsymbol{\beta}$的广义最小二乘估计为$\boldsymbol{\beta}^*$, $\{\hat{e}_i, i = 1, \cdots, n\}$为相应的残差. 记

$$\boldsymbol{H}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-\frac{1}{2}}.$$

删去第i组观测$(\boldsymbol{x}_i', y_i)$后, 记$\boldsymbol{\beta}$的广义最小二乘估计为$\boldsymbol{\beta}^*_{(i)}$. 证明:

$$\boldsymbol{\beta}^*_{(i)} = \boldsymbol{\beta}^* - \frac{(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{e}_i/\sigma_i^2}{1 - (\boldsymbol{H}_{\boldsymbol{\Sigma}})_{ii}},$$

这里的$(\boldsymbol{H}_{\boldsymbol{\Sigma}})_{ii}$表示矩阵$\boldsymbol{H}_{\boldsymbol{\Sigma}}$的第$i$个对角线元素.

16. 对于线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},\ \ \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0},\ \ \mathsf{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V},$$

其中$\boldsymbol{V}$为正定矩阵, $\boldsymbol{X}$为$n \times (p+1)$矩阵,
(1) 证明$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$是$\boldsymbol{\beta}$的一个无偏估计;
(2) 证明$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$;
(3) 记$\hat{\sigma}^2 = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}/(n-p-1)$, 证明

$$\mathsf{E}(\hat{\sigma}^2) = \frac{\sigma^2}{n-p-1}\mathrm{tr}[\boldsymbol{V}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')].$$

17. 假设有以下的10个观测数据, 若以x_1, x_2为回归自变量, 判断他们之间是否存在多重共线性关系?

y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5

18. 假设有以下的10个观测数据, 试用岭迹法求y关于x_1, x_2的岭回归方程, 并画出岭迹图.

y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5

19. 对某种商品的销量y进行调查, 并考虑有关的四个因素: x_1表示居民可支配收入, x_2表示该商品的平均价格指数, x_3表示该商品的社会保有量, x_4表示其它消费品平均价格指数. 调查数据见下表, 利用主成分方法建立y与x_1, x_2, x_3, x_4的回归方程.

序号	x_1	x_2	x_3	x_4	y
1	82.9	92.0	17.1	94.0	8.4
2	88.0	93.0	21.3	96.0	9.6
3	99.9	96.0	25.1	97.0	10.4
4	105.3	94.0	29.0	97.0	10.4
5	117.7	100.0	34.0	100.0	12.2
6	131.0	101.0	40.0	101.0	14.2
7	148.2	105.0	44.0	104.0	15.8
8	161.8	112.0	49.0	109.0	17.9
9	174.2	112.0	51.0	111.0	19.6
10	184.7	112.0	53.0	111.0	20.8