第三章　回归参数的估计

Tianxiao Pang

Zhejiang University

September 28, 2016

内容

内容

内容

内容

内容

内容

内容

内容

内容

内容

估计回归参数的最基本方法是最小二乘法(Least Squares Method). 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. 在第四和第五节, 我们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(Box-Cox变换). 在第六节我们将讨论广义最小二乘估计. 在第七节我们将讨论一种特殊的自变量, 并讨论它如何给最小二乘估计带来危害. 第八节和第九节我们将讨论两种新的估计方法: 岭估计和主成分估计. 最后一节简单介绍Stein压缩估计.

从这里开始, 我们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. 同时我们约定: 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).

最小二乘估计

用y表示因变量, $x_1, \cdots, x_p$表示(可能)对y有影响的p个自变量. 假设它们之间满足如下的线性关系式:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e, \tag{3.1.1}$$

其中e是随机误差, $\beta_0, \cdots, \beta_p$表示待估的未知参数. 称β_0为回归常数, 称$\beta_1, \cdots, \beta_p$为回归系数. 有时把它们统称为回归系数. 假定已经有样本

$$(x_{i1}, \cdots, x_{ip}, y_i), \quad i = 1, \cdots, n,$$

则它们满足

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \cdots, n. \tag{3.1.2}$$

假设误差项$e_i, i = 1, \cdots, n$满足Gauss-Markov假设.

若用矩阵表示, 则可写为

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

或写为

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

其中$\boldsymbol{Y}$是$n \times 1$的随机观测向量(因变量向量), $\boldsymbol{X}$是$n \times (p+1)$的已知设计矩阵(假设$p+1 \leq n$), $\boldsymbol{\beta}$是$(p+1) \times 1$的未知参数向量, $\boldsymbol{e}$是$n \times 1$的随机误差向量.

将Gauss-Markov假设也写成矩阵形式:

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

将(3.1.3)和(3.1.4)合写在一起, 即可得到最基本的线性回归模型:

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

我们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(LSE: Least Squares Estimator). 这个方法是寻找一个$\boldsymbol{\beta}$的估计, 使得偏差向量$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$的长度之平方$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$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, 所以我们总假定$\boldsymbol{X}$是列满秩的. 于是我们得到(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. 下证$\hat{\boldsymbol{\beta}}$确实是$Q(\boldsymbol{\beta})$的最小值点.

对任意的$\boldsymbol{\beta}$, 有

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

因为$\hat{\boldsymbol{\beta}}$满足正规方程组(3.1.6), 所以$\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\boldsymbol{0}$. 这证明了对任意的$\boldsymbol{\beta}$, 有

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

由于$\boldsymbol{X}'\boldsymbol{X}$是非负定矩阵, 所以$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\geq 0$. 于是

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\geq\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2.$$

得证.

记$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)'$, 我们可以得到(经验)回归方程:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \text{ 或 } \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p. \tag{3.1.10}$$

这个方程是不是描述了y与$x_1, \cdots, x_p$的真实关系, 还需作进一步的统计分析, 留待以后处理.

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

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

这时的正规方程组为

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix},$$

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

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

在回归分析中, 我们有时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

$$y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_p(x_{ip} - \bar{x}_p) + e_i, \quad i = 1, \cdots, n. \tag{3.1.11}$$

这里, $\alpha = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_p\bar{x}_p$. 称(3.1.11)为中心化模型. 记

$$\boldsymbol{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}. \tag{3.1.12}$$

则(3.1.11)可改写为

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

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

$$\mathbf{1}_n' \boldsymbol{X}_c = \mathbf{0} \tag{3.1.14}$$

因此正规方程组可写为

$$\begin{pmatrix} n & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_c' \boldsymbol{X}_c \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_n' \boldsymbol{Y} \\ \boldsymbol{X}_c' \boldsymbol{Y} \end{pmatrix}, \tag{3.1.15}$$

等价地写成

$$\begin{cases} n\alpha = \mathbf{1}_n' \boldsymbol{Y}, \\ \boldsymbol{X}_c' \boldsymbol{X}_c \boldsymbol{\beta} = \boldsymbol{X}_c' \boldsymbol{Y} \end{cases} \tag{3.1.16}$$

于是回归参数的LSE为

$$\begin{cases} \hat{\alpha} = \bar{y}, \\ \hat{\boldsymbol{\beta}} = (\boldsymbol{X}_c' \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c' \boldsymbol{Y}. \end{cases} \tag{3.1.17}$$

因此, 对于中心化线性回归模型(3.1.11), 回归常数的LSE总是因变量的样本均值, 而回归系数$\boldsymbol{\beta}$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

$$y_i = \alpha + \beta_1(x_i - \bar{x}) + e_i, \quad i = 1, \cdots, n. \tag{3.1.18}$$

由公式(3.1.17)得LSE

$$\begin{cases} \hat{\alpha} = \bar{y}, \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{cases} \tag{3.1.19}$$

另外, 我们还可以对自变量做标准化处理. 记

$$s_j^2 = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad j = 1, \cdots, p,$$

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, \cdots, n; j = 1, \cdots, p.$$

令$\boldsymbol{Z} = (z_{ij})_{n\times p}$, 它具有性质:

- $\mathbf{1}_n' \boldsymbol{Z} = \mathbf{0}$,
- $\boldsymbol{R} = \boldsymbol{Z}'\boldsymbol{Z} = (r_{ij})$,

其中

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{s_i s_j}, \quad i, j = 1, \cdots, p \qquad (3.1.20)$$

为自变量x_i与x_j的样本相关系数. 所以$\boldsymbol{R}$是自变量的相关系数矩阵.

对自变量进行标准化的好处: (1)可以用来分析回归自变量之间的相关关系; (2)标准化消除了量纲的影响, 便于对回归系数估计值的统计分析.

经过标准化后的线性回归模型为

$$y_i = \alpha + \frac{x_{i1} - \bar{x}_1}{s_1}\beta_1 + \cdots + \frac{x_{ip} - \bar{x}_p}{s_p}\beta_p + e_i, \qquad (3.1.21)$$

或写成矩阵形式

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

未知参数的LSE为

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

回归方程为

$$\hat{y} = \hat{\alpha} + \frac{x_1 - \bar{x}_1}{s_1}\hat{\beta}_1 + \cdots + \frac{x_p - \bar{x}_p}{s_p}\hat{\beta}_p. \qquad (3.1.22)$$

例3.1.3 一个试验容器靠蒸汽供应热量, 使其保持恒温, 下表中, 自变量x表示容器周围空气单位时间的平均温度(°C), 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3.1.1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(3.1.18), 我们得到

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

回归参数的LSE为

$$\hat{\alpha} = \bar{y} = 9.424,\ \hat{\beta}_1 = -0.0798.$$

所以回归方程为

$$\hat{y} = 9.424 - 0.0798(x - 52.6),$$

或写成

$$\hat{y} = 13.623 - 0.0798x.$$

表3.1.1: 蒸汽数据

序号	$y(L)$	$x(^{\circ}\mathrm{C})$	序号	$y(L)$	$x(^{\circ}\mathrm{C})$
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			

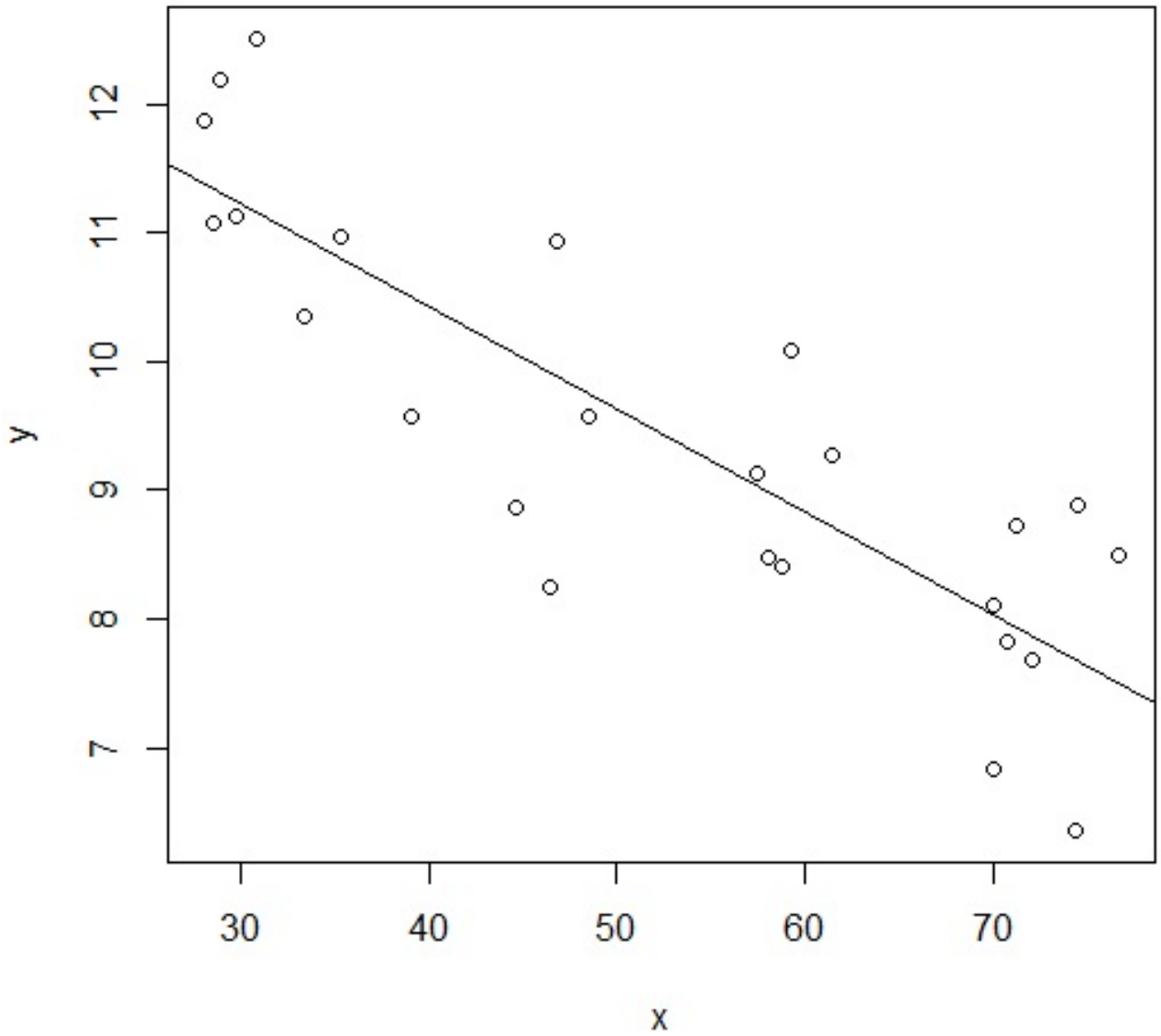


图3.1.1 散点图

R程序:

```
yx=read.table("ex_p33_data.txt")
y=yx[,1]
x=yx[,2]
mydata=data.frame(y,x)
plot(y~x)
lm.sol=lm(y~x,data=mydata)
abline(lm.sol)
summary(lm.sol)
```

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

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
```

直接得到回归方程:

$$\hat{y} = 13.623 - 0.0798x.$$

最小二乘估计的性质

最小二乘估计(LSE)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)*, *LSE* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 具有下列性质:
(a) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$;
(b) $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

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

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

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

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

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), 我们称$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$为$\boldsymbol{c}'\boldsymbol{\beta}$的LSE.

推论 (3.2.1)

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

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$\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$为$y_1, \cdots, y_n$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(线性估计指的是观测$y_1, \cdots, y_n$的线性函数). 我们还可以构造出$\boldsymbol{c}'\boldsymbol{\beta}$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*(BLUE: best linear unbiased estimator)*.

证明: 设$\boldsymbol{a}'\boldsymbol{Y}$为$\boldsymbol{c}'\boldsymbol{\beta}$的一个线性无偏估计. 于是对一切$p+1$维列向量$\boldsymbol{\beta}$,

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

因此

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

因为$\mathsf{Var}(\boldsymbol{a}'\boldsymbol{Y}) = \sigma^2\boldsymbol{a}'\boldsymbol{a} = \sigma^2\|\boldsymbol{a}\|^2$, 我们对$\|\boldsymbol{a}\|^2$做分解:

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

由(3.2.1)可推知

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

由推论3.2.1(b)可推知

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

于是, 我们证明了对$\boldsymbol{c}^{'}\boldsymbol{\beta}$的任一个线性无偏估计$\boldsymbol{a}^{'}\boldsymbol{Y}$有

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{a}^{'}\boldsymbol{Y}) &= \mathrm{Var}(\boldsymbol{c}^{'}\hat{\boldsymbol{\beta}})+\sigma^2\|\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{c}\|^2 \\
&\geq \mathrm{Var}(\boldsymbol{c}^{'}\hat{\boldsymbol{\beta}}), \qquad (3.2.3)
\end{aligned}$$

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)中还有一个参数σ^2(误差方差), 它反映了模型误差的大小, 在回归分析中起着重要的作用. 我们现在来估计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

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

为第i次观测的残差. 自然地, 我们可以将$\hat{\boldsymbol{e}}$看作$\boldsymbol{e}$的一个估计. 我们将用

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

来构造σ^2的无偏估计量.

RSS=Residual Sum of Squares, 表示残差平方和. 它反映了实际数据与理论模型(3.1.5)的偏离程度或者说拟合程度. RSS越小表示数据与模型拟合得越好.

定理 (3.2.3)

(a) $RSS = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y} =: \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$;
(b) $\hat{\sigma}^2 = RSS/(n-p-1)$是$\sigma^2$的无偏估计量.

注: 称$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$为帽子(hat)矩阵, 它是一个对称幂等矩阵.

证明: (a)

$$\begin{aligned}\mathrm{RSS} = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= [(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}]'[(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}] \\ &= \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}.\end{aligned}$$

(b) 由$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}, \mathrm{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$以及定理2.2.1知

$$\begin{aligned}\mathrm{E}(\mathrm{RSS}) &= \mathrm{E}[\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}] \\ &= \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{X}\boldsymbol{\beta} \\ &\quad + \sigma^2\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\ &= \sigma^2[n - \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')].\end{aligned}$$

根据迹的性质$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$知

$$\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{I}_{p+1}) = p + 1.$$

于是$\mathrm{E}(\mathrm{RSS}) = \sigma^2(n - p - 1)$.

如果假设误差向量$\boldsymbol{e}$服从正态分布, 即$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, 那么我们可以得到$\hat{\boldsymbol{\beta}}$和$\hat{\sigma}^2$更多的重要性质.

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) $RSS/\sigma^2 \sim \chi^2(n-p-1)$;
(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) 根据定义

$$\mathsf{RSS} = \boldsymbol{Y}^{'}(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{'}(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{'}\boldsymbol{N}\boldsymbol{Y},$$

这里$\boldsymbol{N} = \boldsymbol{I}_n - \boldsymbol{H}$, 它是一个对称幂等矩阵. 注意到$\boldsymbol{N}\boldsymbol{X} = \boldsymbol{0}$, 所以

$$\mathsf{RSS} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})^{'}\boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) = \boldsymbol{e}^{'}\boldsymbol{N}\boldsymbol{e}. \tag{3.2.7}$$

注意到$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, 所以根据定理2.4.3我们只需证明$\boldsymbol{N}$的秩为$n - p - 1$. 利用幂等矩阵的秩等于它的迹这一性质, 我们可得

$$\begin{aligned}
\mathsf{rk}(\boldsymbol{N}) &= \mathsf{tr}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}) \\
&= n - \mathsf{tr}(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}) \\
&= n - \mathsf{tr}((\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}) \\
&= n - p - 1.
\end{aligned}$$

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

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{N} = \boldsymbol{0},$$

所以由定理2.4.5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$\boldsymbol{\beta}$的第一个分量为常数项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$\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \hat{\beta}_i$. 再用$(\boldsymbol{A})_{ii}$表示矩阵$\boldsymbol{A}$的第$(i,i)$元素, 那么我如下推论:

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$\beta_i, i = 1, \cdots, p$的一切线性无偏估计中, $\hat{\beta}_i, i = 1, \cdots, p$是唯一方差最小者.

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

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) $E(\hat{\alpha}) = \alpha, E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, 这里$\hat{\alpha} = \bar{y}$, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}$;
(b)

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

(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

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}}, \tag{3.2.8}$$

其中

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

称为回归平方和(或解释平方和: Explained Sum of Squares),

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

称为总偏差平方和(或称为总平方和: Total Sum of Squares).
称R^2为判定系数或测定系数, 称$R = \sqrt{R^2}$为复相关系数.

由正规方程组可以证明

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

即

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

因此R^2度量了回归自变量$x_1,\cdots,x_p$对因变量y的拟合程度的好坏. $0\le R^2\le 1$, 它的值越大, 表明y与诸自变量有较大的相依关系.

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$\hat{\boldsymbol{Y}} = \hat{\alpha} \mathbf{1}_n + \boldsymbol{X}_c \hat{\boldsymbol{\beta}}$及公式(3.1.17)可知

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

所以

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

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为模型中自变量与因变量的样本相关系数.

事实上, 若我们把模型中心化:

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

那么可知

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

从而

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

所以

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

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, 见下表. 试建立y关于x_1, x_2的线性回归方程.

表3.2.1: 血压数据

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				

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

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

所以$R^2 = 1430.276/1512 = 0.9459.$

R程序:

```
yx=read.table("ex_p39_data.txt")
x1=yx[,1]
x2=yx[,2]
y=yx[,3]
mydata=data.frame(y,x1,x2)
lm.sol=lm(y~x1+x2,data=mydata)
summary(lm.sol)
```

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

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R运行结果得: $\hat{y} = -62.963 + 1.068x_1 + 0.4x_2$, $R^2 = 0.9461$.

约束最小二乘估计

对$\boldsymbol{\beta}$不加任何约束条件的情形下, 我们讨论了它的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, 我们需要求带有一定约束条件的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$k \times (p+1)$已知矩阵, 秩为k, $\boldsymbol{b}$是$k \times 1$已知向量. 我们用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)这个条件下求$\boldsymbol{\beta}$使$Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$达到最小.

构造辅助函数

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

其中$\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)'$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0, 得

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

现在, 我们需要解由(3.3.1)和(3.3.2)组成的联立方程组.

我们用$\hat{\boldsymbol{\beta}}_c$和$\hat{\boldsymbol{\lambda}}_c$表示这个方程组的解. 由(3.3.2)得

$$\begin{aligned}\hat{\boldsymbol{\beta}}_c &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c \\ &= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\lambda}}_c. \qquad (3.3.3)\end{aligned}$$

代入(3.3.1)得

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

可等价地写成

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

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

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

下证$\hat{\boldsymbol{\beta}}_c$确实是线性约束$\boldsymbol{A\beta} = \boldsymbol{b}$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$\boldsymbol{A\beta} = \boldsymbol{b}$的$\boldsymbol{\beta}$, 都有$\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_c\|^2$.

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

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

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

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

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

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

等号成立当且仅当$\boldsymbol{X}(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) = \boldsymbol{0}$, 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

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

现在, 结合(3.3.7)和(3.3.8)便可推得(b).

我们把$\hat{\boldsymbol{\beta}}_c$称为$\boldsymbol{\beta}$的约束最小二乘估计, 于是有下列的定理:

定理 (3.3.1)

对于线性回归模型*(3.1.5)*, 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, 我们利用线性模型表示有关的量

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

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, 经计算得

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

即

$$\hat{\theta}_i = y_i - \frac{1}{3}\left(\sum_{i=1}^{3} y_i - \pi\right),\ i = 1, 2, 3$$

为θ_i的约束LSE.

回归诊断

对于线性回归模型(3.1.5):

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

我们在讨论未知参数的最小二乘估计以及估计量的分布时, 作了如下的最基本的假设:

(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, 那么我们以前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, 当我们有了一批数据后, 需要考察我们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$\boldsymbol{e}$的一个估计, 因此我们可以通过残差来分析4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, 我们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相当困难的, 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. 这时异常点可理解为所假定分布中的极端点;(2)异常点就是那些污染点, 即是指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, 因此我们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推断(参数估计、假设检验等)产生较大影响的数据点, 在此课程中, 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, 我们得到的经验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, 当我们剔除这些数据之后, 就能得到与原来差异很大的经验回归方程.

因此在回归分析中, 我们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, 我们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缩小/稀释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, 模型的诊断包含线性诊断、方差齐性诊断、不相关性诊断和正态性诊断4 部分内容.

先来讨论残差分析:

对于线性回归模型:

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

我们用

$$\hat{e}_i = y_i - \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}},\ \ i = 1, \cdots, n \tag{3.4.2}$$

表示第i次观测的残差, 并把它看成是e_i的一个估计. 若模型(3.4.1)正确, 那么$\hat{e}_i$应具有e_i的一些性状.

记$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, 称$\hat{\boldsymbol{Y}}$为拟合值向量, 称$\hat{y}_i = \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}$为第$i$个拟合值. 易知

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

其中$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$被称为帽子矩阵, 它是对称幂等矩阵.

残差$\hat{e}$可被表示为

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

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

定理 (3.4.1)

若*(3.4.1)*成立, 则
(a) $E(\hat{e}) = \mathbf{0},\ Cov(\hat{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$;
(b) $Cov(\hat{\boldsymbol{Y}}, \hat{e}) = \mathbf{0}$.
(c) 若进一步假设$e \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$, 则

$$\hat{e} \sim N(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})).$$

证明: 容易, 略.

因为$\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$($h_{ii}$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. 因此我们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$\hat{\sigma}^2 = \mathrm{RSS}/(n - p - 1) = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n - p - 1)$.

即使在$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$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(1987)). 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, 以任何其它的量为横坐标的散点图. 前已指出残差作为误差e_i的估计应该与e_i相差不远, 故根据残差图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信息.

下面我们以拟合值$\hat{y}_i$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, 残差图通常会呈现“喇叭型”或“倒喇叭型”或两者兼而有之等形状.

不相关性诊断: 若独立性成立, 那么残差图上的点不呈现规则性, 否则, 散点图将呈现"集团性"或“剧烈交错性”等形状.

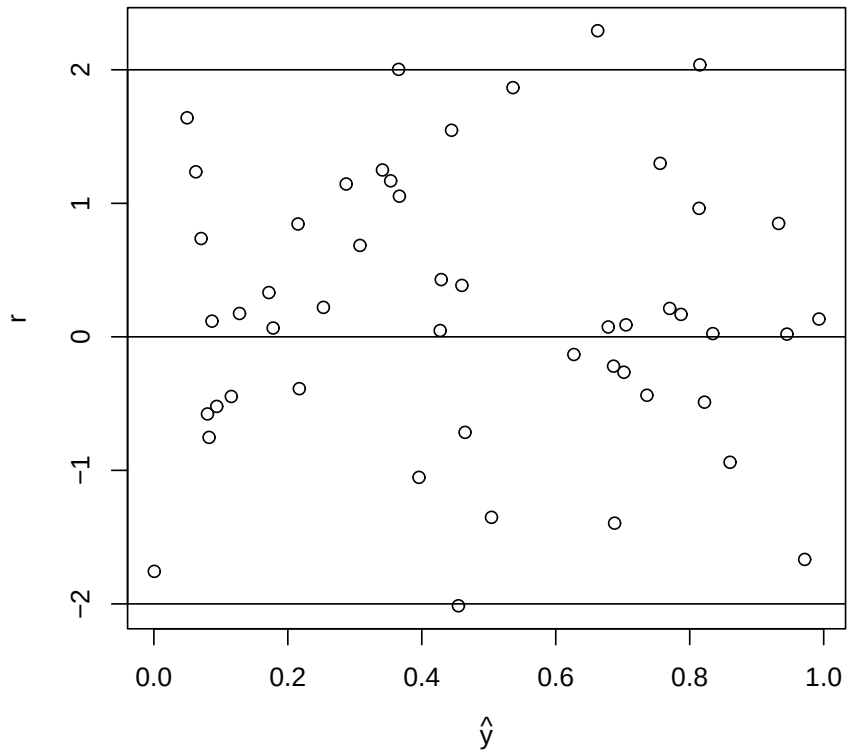
正态性诊断: 若正态性成立, 那么学生化残差r_i可近似看成是相互独立且服从$N(0,1)$. 所以, 在以r_i为纵坐标$\hat{y}_i$为横坐标的残差图上, 平面上的点$(\hat{y}_i, r_i), i = 1, \cdots, n$ 大致应落在宽度为4的水平带$|r_i| \leq 2$区域内(这个频率应在95%左右), 且不呈现任何趋势.

我们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.

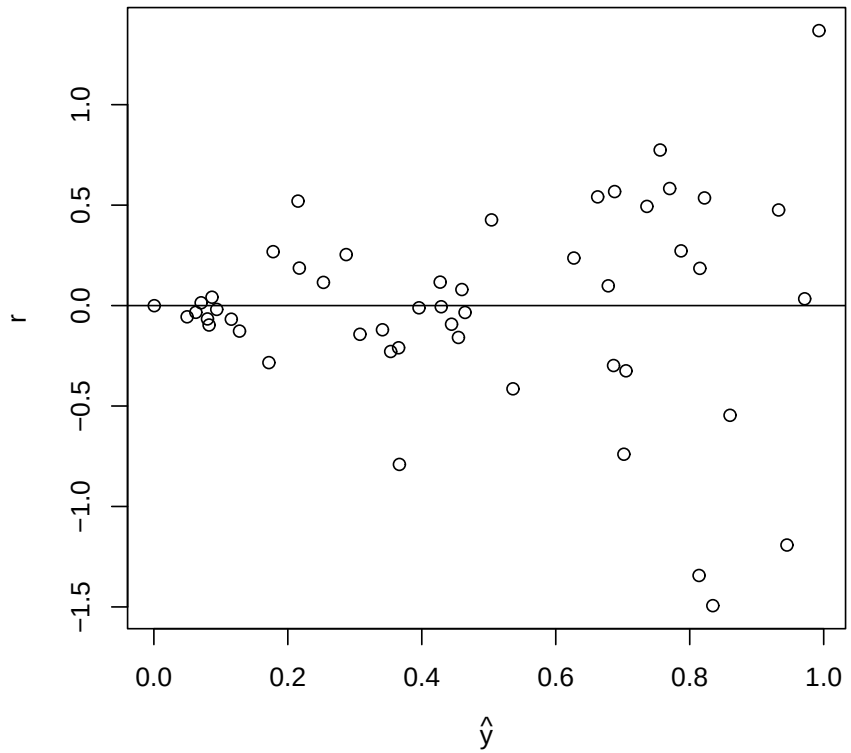
正态性诊断: 若正态性成立, 那么学生化残差r_i可近似看成是相互独立且服从$N(0,1)$. 所以, 在以r_i为纵坐标$\hat{y}_i$为横坐标的残差图上, 平面上的点$(\hat{y}_i, r_i), i = 1, \cdots, n$ 大致应落在宽度为4的水平带$|r_i| \leq 2$区域内(这个频率应在95%左右), 且不呈现任何趋势.

我们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.如果数据是来自$F(x)$的一个简单随机样本, 则这些散点应大致在一条直线上, 因此通过判断QQ图是否近似在一条直线上可得到所给的数据集是否服从所指定分布的一些统计依据.

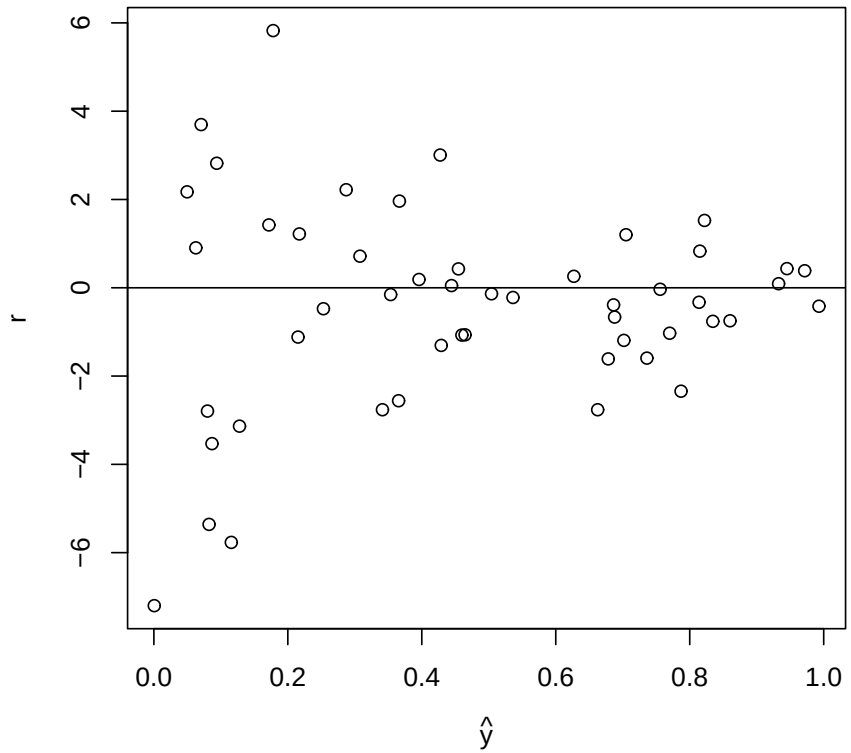
此外, 我们还可以Shapiro-Wilk方法来做正态性检验.



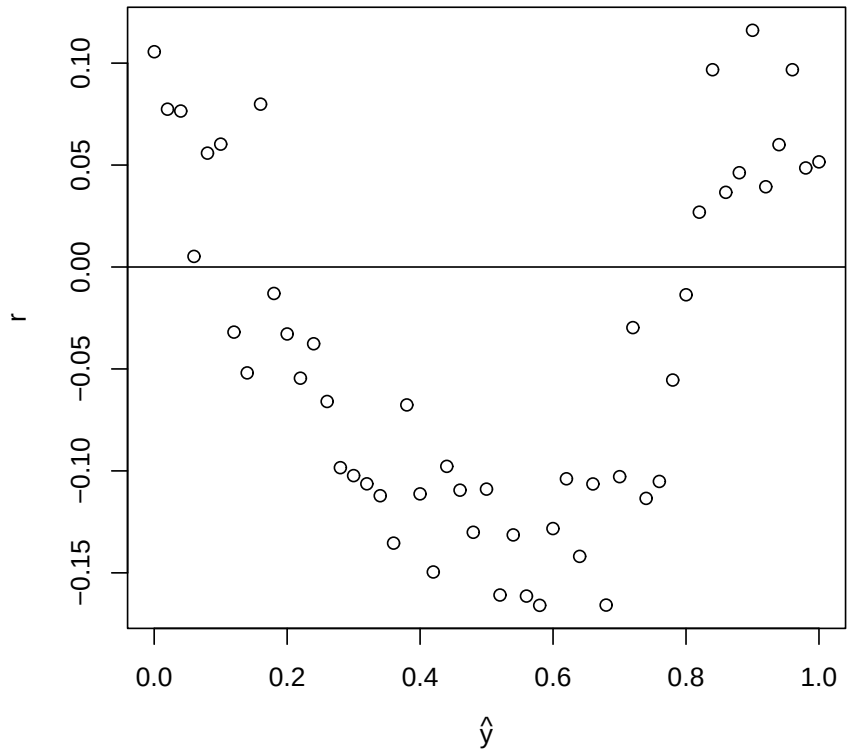
分析: 其性状与正态性假设基本一致, 因此可认为正态性假设$\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$是合理的.



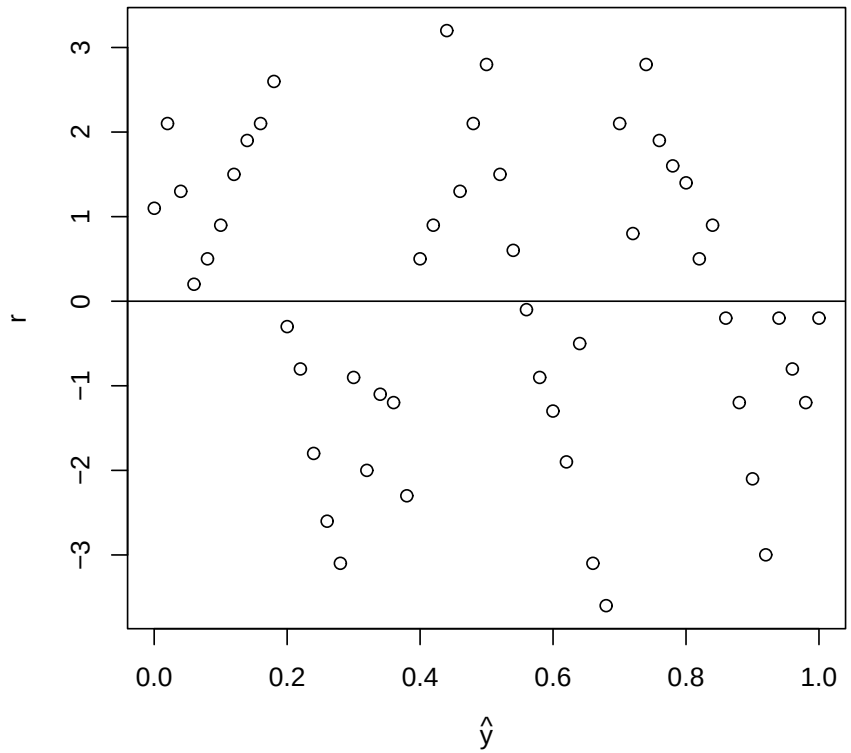
分析: 残差图呈现喇叭型, 因此认为方差齐性假设不合理.



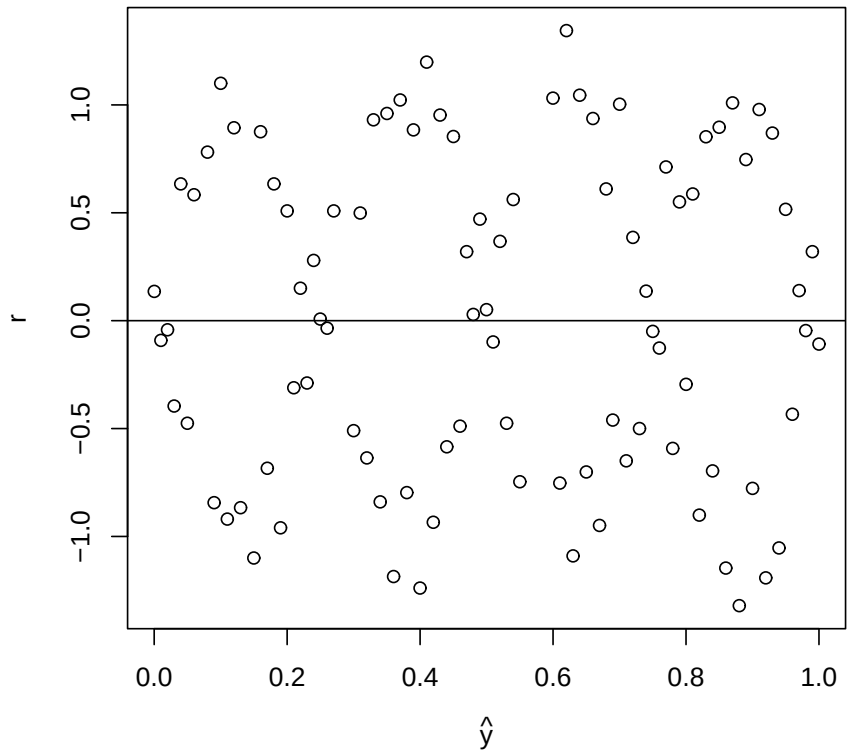
分析: 残差图呈现倒喇叭型, 因此认为方差齐性假设不合理.



分析: 图形呈现有规则的形状, 因此认为线性假设不合理, 残差中应含有自变量的信息.



分析: 残差图呈现集团性, 因此认为不相关性不合理.



分析: 残差图呈现剧烈交错性, 因此认为不相关性不合理.

从残差图诊断出来可能的“疾病”，也就是某些假设条件不成立，我们需要对问题“对症下药”：

如果有症状使我们怀疑线性假设不成立，那么我们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；

如果有症状使我们怀疑正态性假设不成立，那我们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.

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$\mathsf{E}(Y) = np$, $\mathsf{Var}(Y) = np(1-p) = \mu(1-\mu/n)$. 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, 我们用大写字母$Y$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Taylor展开, 取近似式

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

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

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

下面是几个特例:

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

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

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

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

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

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

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

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

求出对应的变换. 对几种变换过的数据分别作最小二乘处理, 作新的残差图, 看哪一种变换的残差图无方差非齐性的征兆. 从中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.

表3.4.1: 用电量数据

i	x	y	$\hat{y}_i$	$\hat{e}_i$	$z=\sqrt{y}$	$\hat{z}_i$	$\tilde{e}_i$
1	679	0.790	1.669	-0.879	0.889	1.229	-0.340
2	292	0.440	0.244	0.196	0.663	0.860	-0.197
3	1012	0.560	2.896	-2.336	0.748	1.547	-0.798
4	493	0.790	0.984	-0.194	0.889	1.052	-0.163
5	582	2.700	1.312	1.388	1.643	1.137	0.506
6	1156	3.640	3.426	0.214	1.908	1.684	0.224
7	997	4.730	2.840	1.890	2.175	1.532	0.643
8	2189	9.500	7.230	2.270	3.082	2.668	0.414
9	1097	5.340	3.209	2.131	2.311	1.628	0.683
10	2078	6.850	6.822	0.028	2.617	2.562	0.055
11	1818	5.840	5.864	-0.024	2.417	2.315	0.102
12	1700	5.210	5.430	-0.220	2.283	2.202	0.080
13	747	3.250	1.920	1.330	1.803	1.294	0.509
14	2030	4.430	6.645	-2.215	2.105	2.517	-0.412

i	x	y	$\hat{y}_i$	$\hat{e}_i$	$z=\sqrt{y}$	$\hat{z}_i$	$\tilde{e}_i$
15	1643	3.160	5.220	-2.060	1.778	2.148	-0.370
16	414	0.500	0.693	-0.193	0.707	0.977	-0.270
17	354	0.170	0.472	-0.302	0.412	0.920	-0.507
18	1276	1.880	3.868	-1.988	1.371	1.798	-0.427
19	745	0.770	1.912	-1.142	0.877	1.292	-0.415
20	435	1.390	0.771	0.619	1.179	0.997	0.182
21	540	0.560	1.157	-0.597	0.748	1.097	-0.348
22	874	1.560	2.388	-0.828	1.249	1.415	-0.166
23	1543	5.280	4.851	0.429	2.298	2.052	0.245
24	1029	0.640	2.958	-2.318	0.800	1.563	-0.763
25	710	4.000	1.784	2.216	2.000	1.259	0.741
26	1434	0.310	4.450	-4.140	0.557	1.949	-1.392
27	837	4.200	2.251	1.949	2.049	1.380	0.670
28	1748	4.880	5.606	-0.726	2.209	2.248	0.039
29	1381	3.480	4.255	-0.775	1.865	1.898	-0.033
30	1428	7.580	4.428	3.152	2.753	1.943	0.810
31	1255	2.630	3.791	-1.161	1.622	1.778	-0.156
32	1777	4.990	5.713	-0.723	2.234	2.275	-0.042
33	370	0.590	0.531	0.059	0.768	0.935	-0.167
34	2316	8.190	7.698	0.492	2.862	2.789	0.073

i	x	y	$\hat{y}_i$	$\hat{e}_i$	$z=\sqrt{y}$	$\hat{z}_i$	$\tilde{e}_i$
35	1130	4.790	3.330	1.460	2.189	1.659	0.530
36	463	0.510	0.874	-0.364	0.714	1.023	-0.309
37	770	1.740	2.004	-0.264	1.319	1.316	0.003
38	724	4.100	1.835	2.265	2.025	1.272	0.753
39	808	3.940	2.144	1.796	1.985	1.352	0.633
40	790	0.960	2.078	-1.118	0.980	1.335	-0.355
41	783	3.290	2.052	1.238	1.814	1.328	0.486
42	406	0.440	0.664	-0.224	0.663	0.969	-0.306
43	1242	3.240	3.743	-0.503	1.800	1.766	0.034
44	658	2.140	1.592	0.548	1.463	1.209	0.254
45	1746	5.710	5.599	0.111	2.390	2.246	0.144
46	468	0.640	0.892	-0.252	0.800	1.028	-0.228
47	1114	1.900	3.271	-1.371	1.378	1.644	-0.265
48	413	0.510	0.690	-0.180	0.714	0.976	-0.262
49	1787	8.330	5.750	2.580	2.886	2.285	0.601
50	3560	14.940	12.280	2.660	3.865	3.974	-0.109
51	1495	5.110	4.675	0.435	2.261	2.007	0.254
52	2221	3.850	7.348	-3.498	1.962	2.699	-0.736
53	1526	3.930	4.789	-0.859	1.982	2.036	-0.054

应用最小二乘法, 得回归方程

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

R程序:

```
yx=read.table("ex_p47_data.txt")
x=yx[,1]
y=yx[,2]
mydata=data.frame(y,x)
lm.sol=lm(y~x,data=mydata)
summary(lm.sol)
```

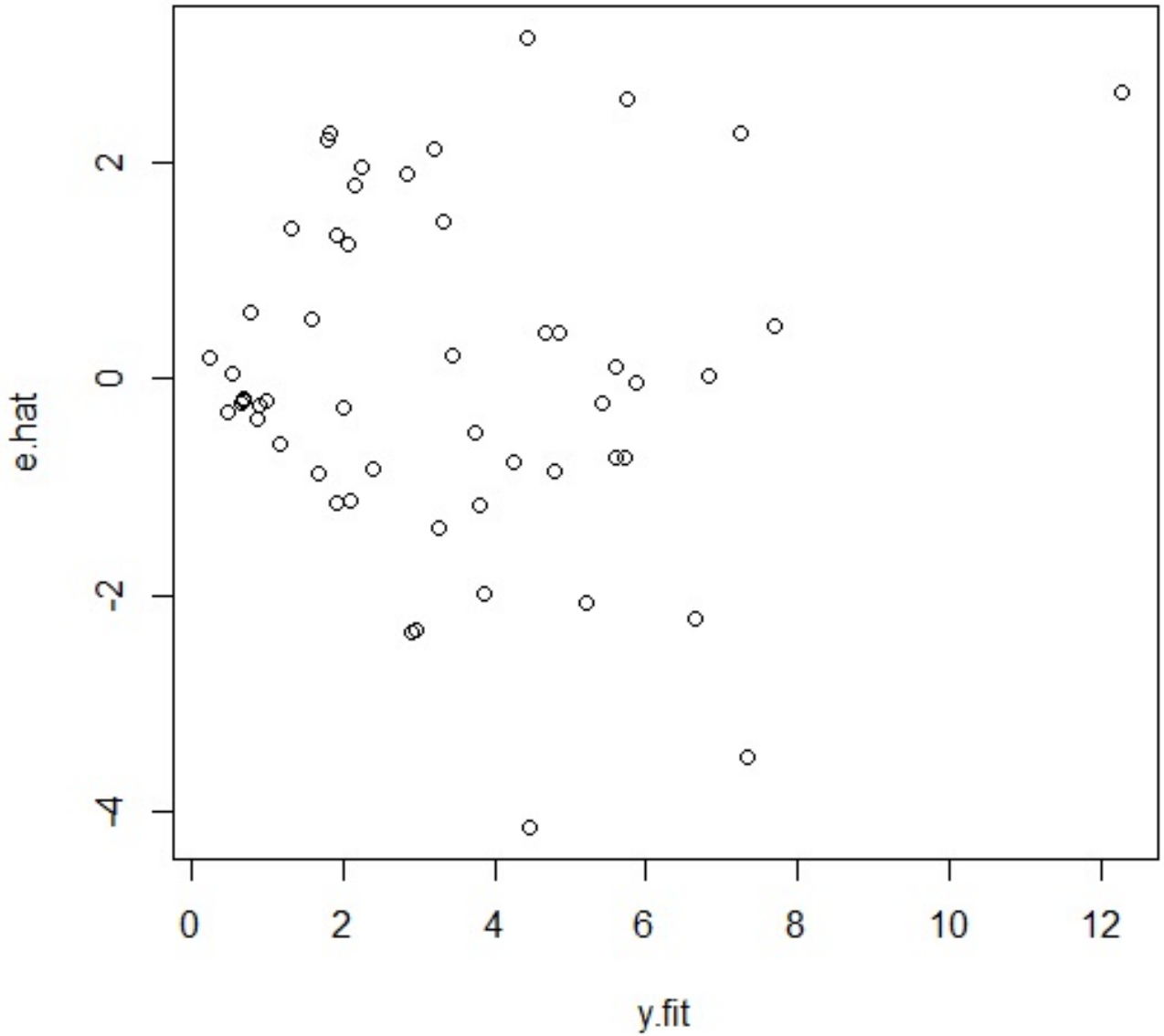
```
Call:
lm(formula = y ~ x, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1399 -0.8275 -0.1934  1.2376  3.1522

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8313037  0.4416121  -1.882   0.0655 .
x            0.0036828  0.0003339  11.030 4.11e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.577 on 51 degrees of freedom
Multiple R-squared:  0.7046,    Adjusted R-squared:  0.6988
F-statistic: 121.7 on 1 and 51 DF,  p-value: 4.106e-15
```

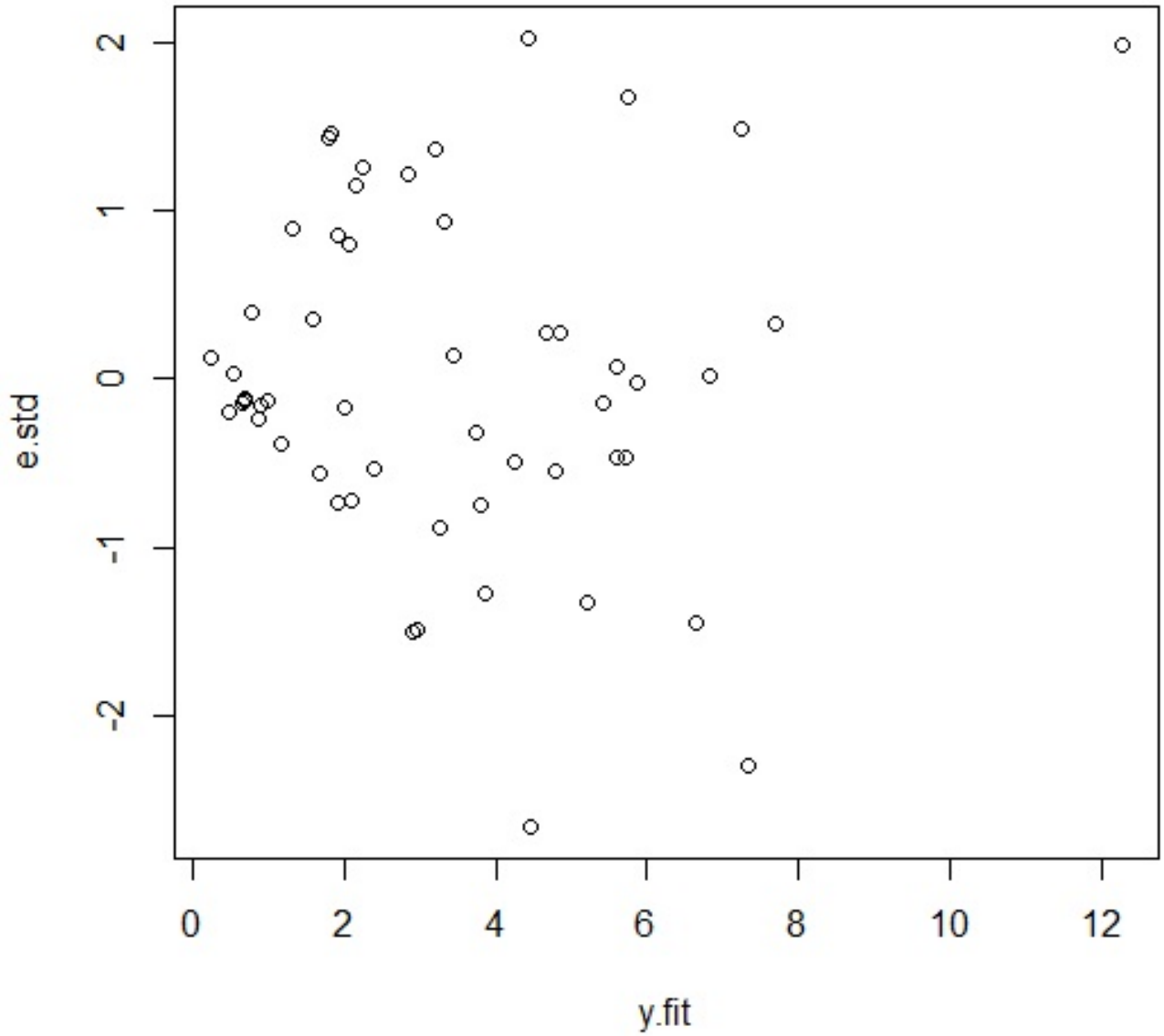
做残差分析:

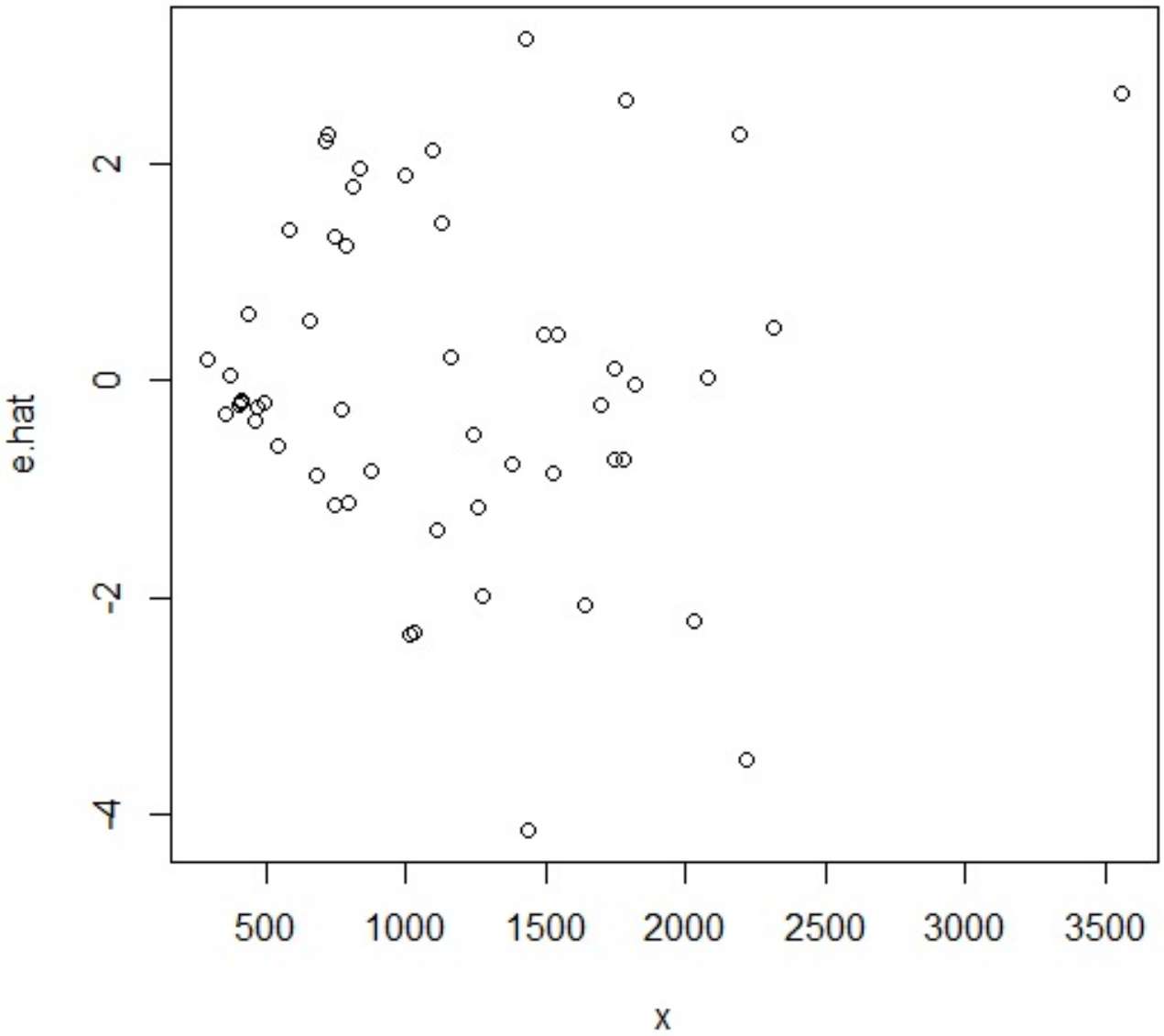
```
y.fit=predict(lm.sol)
e.hat=y-y.fit  /*或者e.hat=residuals(lm.sol)*/
e.std=rstandard(lm.sol)
plot(e.hat~y.fit)
plot(e.std~y.fit)
plot(e.hat~x)
```

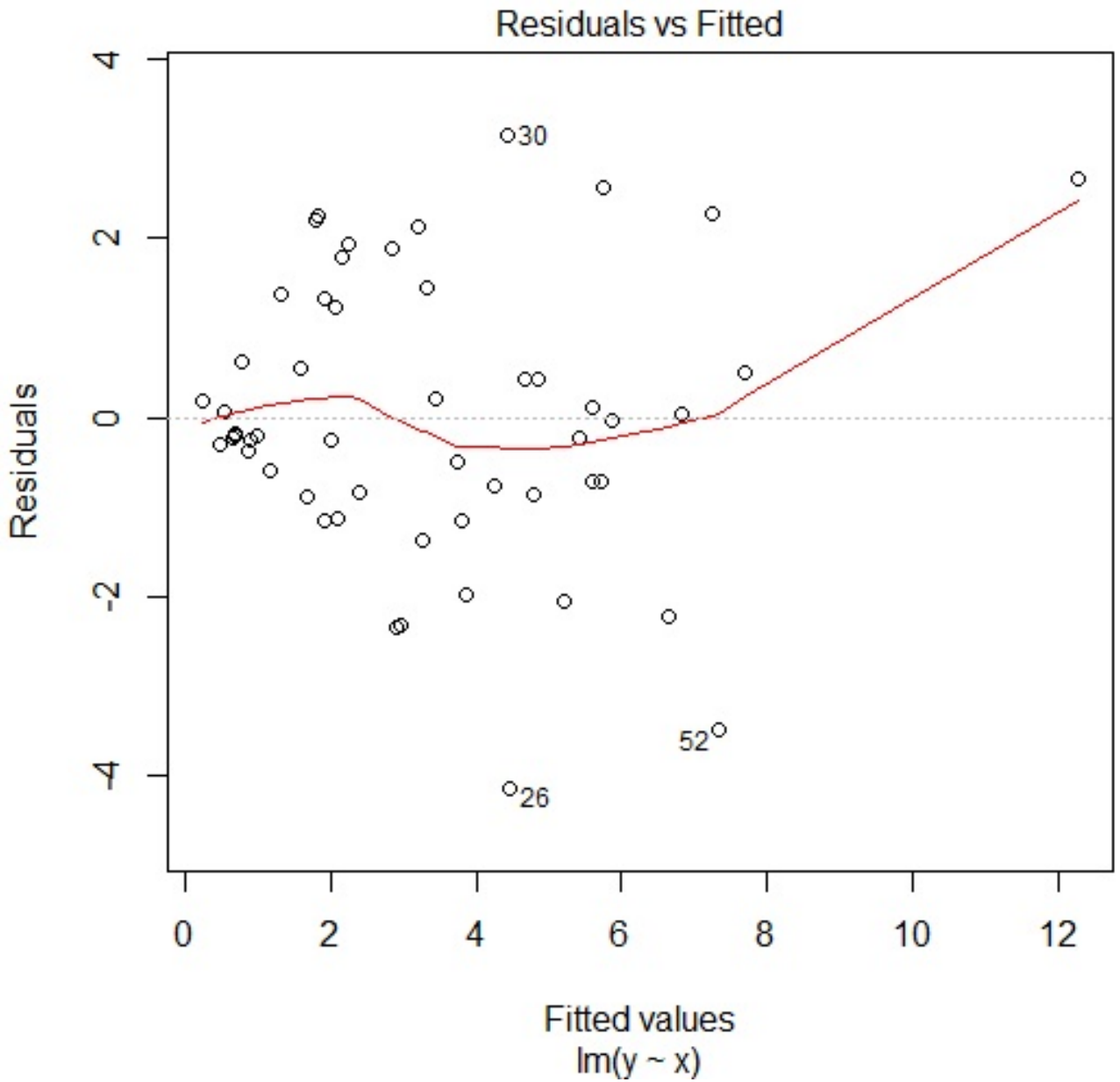
普通残差图也可以通过如下方式得到:

```
plot(lm.sol,which=1)
```


e.hat
y.fit
2
0
-2
-4
0
2
4
6
8
10
12


e.std
2
1
0
-1
-2
0
2
4
6
8
10
12
y.fit


e.hat
2
0
-2
-4
500
1000
1500
2000
2500
3000
3500
x


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x)
30
52
26
4
2
0
-2
-4
0
2
4
6
8
10
12

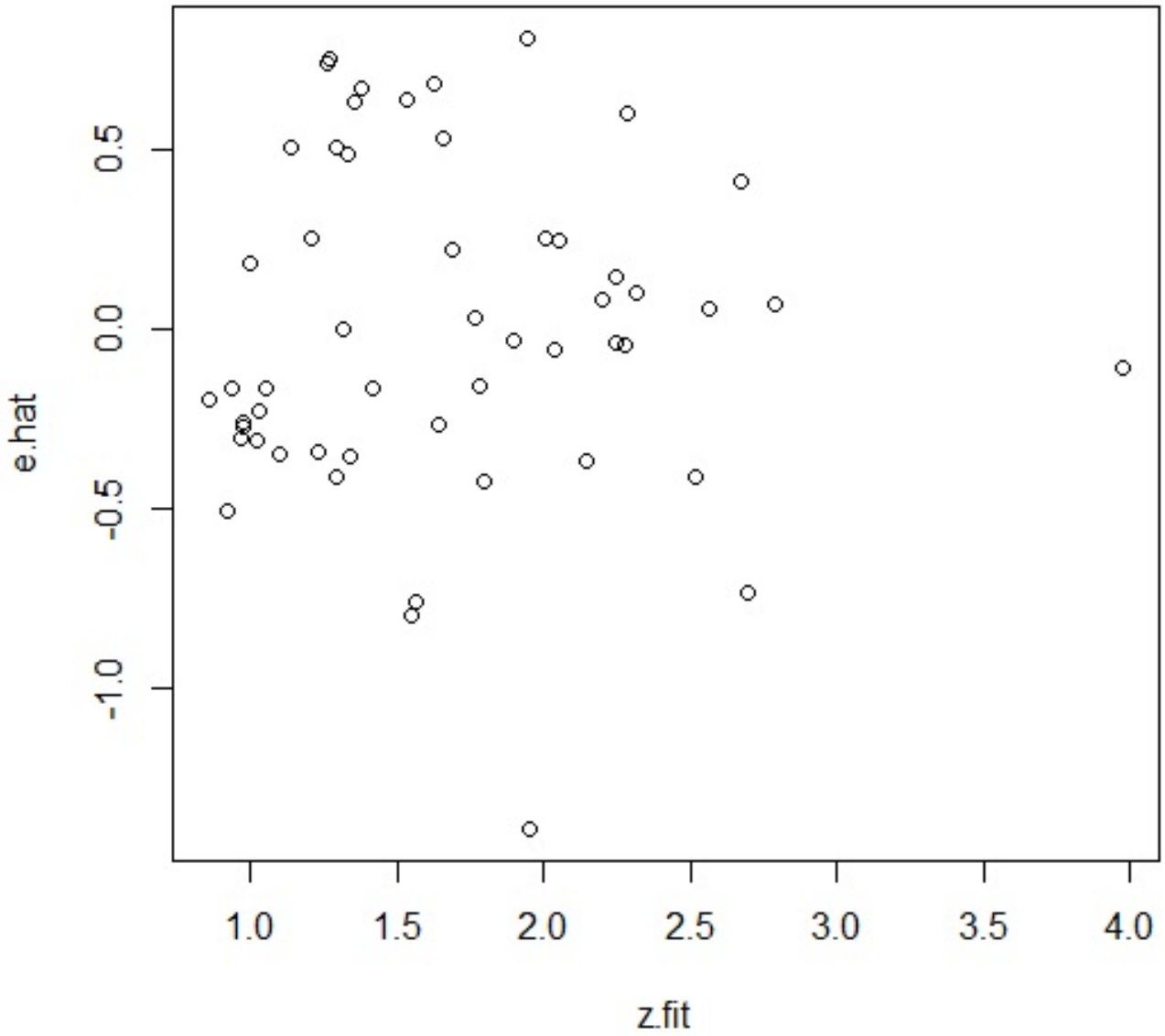
从残差图可看出, 这是一个喇叭型残差图, 是方差齐性不被符合的一个症状. 考虑对因变量y作变换, 尝试$z = \sqrt{y}$, 得回归方程

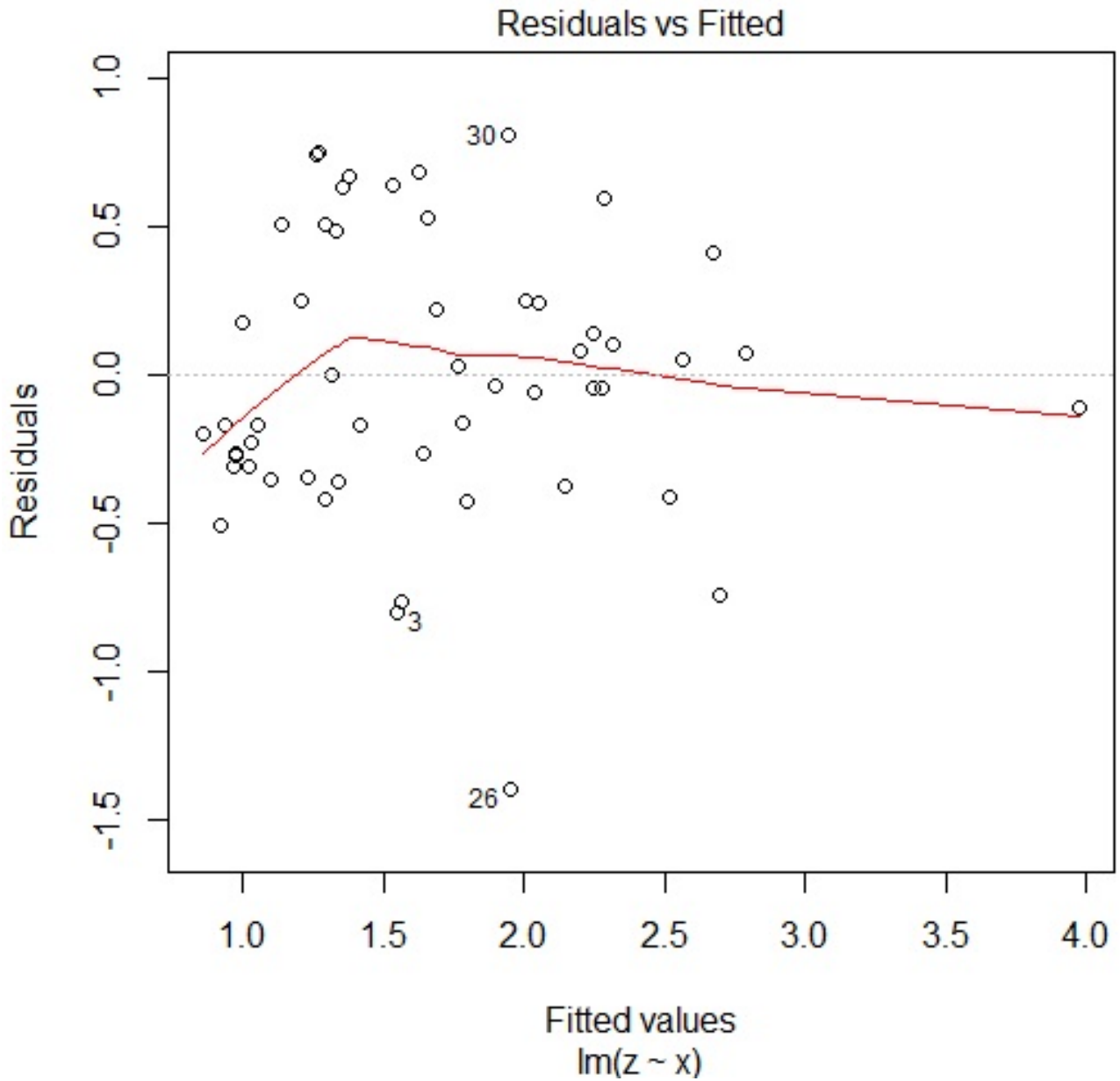
$$\hat{z} = 0.5822 + 0.000953x.$$

R程序:

```
z=sqrt(y)
mydata2=data.frame(z,y,x)
lm.sol2=lm(z~x,data=mydata2)
summary(lm.sol2)
z.fit=predict(lm.sol2)
e.hat=z-z.fit  /*或者e.hat=residuals(lm.sol2)*/
plot(e.hat~z.fit) /*或者用plot(lm.sol2,which=1)*/
```

```
Call:
lm(formula = z ~ x, data = mydata2)

Residuals:
     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
```


e.hat
0.5
0.0
-0.5
-1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
z.fit


Residuals vs Fitted
Residuals
Fitted values
lm(z ~ x)
30
3
26
1.0
0.5
0.0
-0.5
-1.0
-1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

新的残差图不呈现任何明显规则性, 这表明我们所用的变换是合适的. 最后得回归方程为

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

接下来我们做数据的诊断: 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可近似看成是相互独立且服从$N(0,1)$, 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$|r_i| > 2$, 我们就有理由怀疑对应的样本点$(\boldsymbol{x}_i', y_i)$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

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

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

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

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种.

首先我们来求$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$的精确表达式. 为此, 需先介绍一个恒等式:

引理

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, 则被称为高杠杆点. 又因为

$$\boldsymbol{X}_{(i)}^{'}\boldsymbol{Y}_{(i)}=\sum_{j\neq i}\boldsymbol{x}_jy_j=\sum_{j=1}^{n}\boldsymbol{x}_jy_j-\boldsymbol{x}_iy_i=\boldsymbol{X}^{'}\boldsymbol{Y}-\boldsymbol{x}_iy_i,$$

所以

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

所以

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

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

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

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

这个定理表明, 在计算Cook距离的时候, 我们只需要从完全数据的线性回归模型算出学生化残差r_i 和帽子矩阵的对角元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.

h_{ii}的含义: h_{ii}度量了第i组数据$\boldsymbol{x}_i$到中心$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$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

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

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

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

若是一元中心化线性回归模型, 则

$$\boldsymbol{X}=\begin{pmatrix}1 & x_1-\bar{x}\\ \vdots & \vdots\\ 1 & x_n-\bar{x}\end{pmatrix},\quad \boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix}n & 0\\ 0 & S_{xx}\end{pmatrix},\quad h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}.$$

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$h_{ii}$度量了第$i$组数据$\boldsymbol{x}_i$到中心$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$的距离. 因此$P_i$刻画了第$i$组数据距离其它数据的远近. Cook距离被$P_i$和$r_i^2$的大小所决定. 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很困难的.

应用置信椭球可以对Cook距离推导中的$\boldsymbol{M}, c$的选取给予一定的理论支持. 在误差正态假设下,

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

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

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

且二者相互独立. 所以

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

集合

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

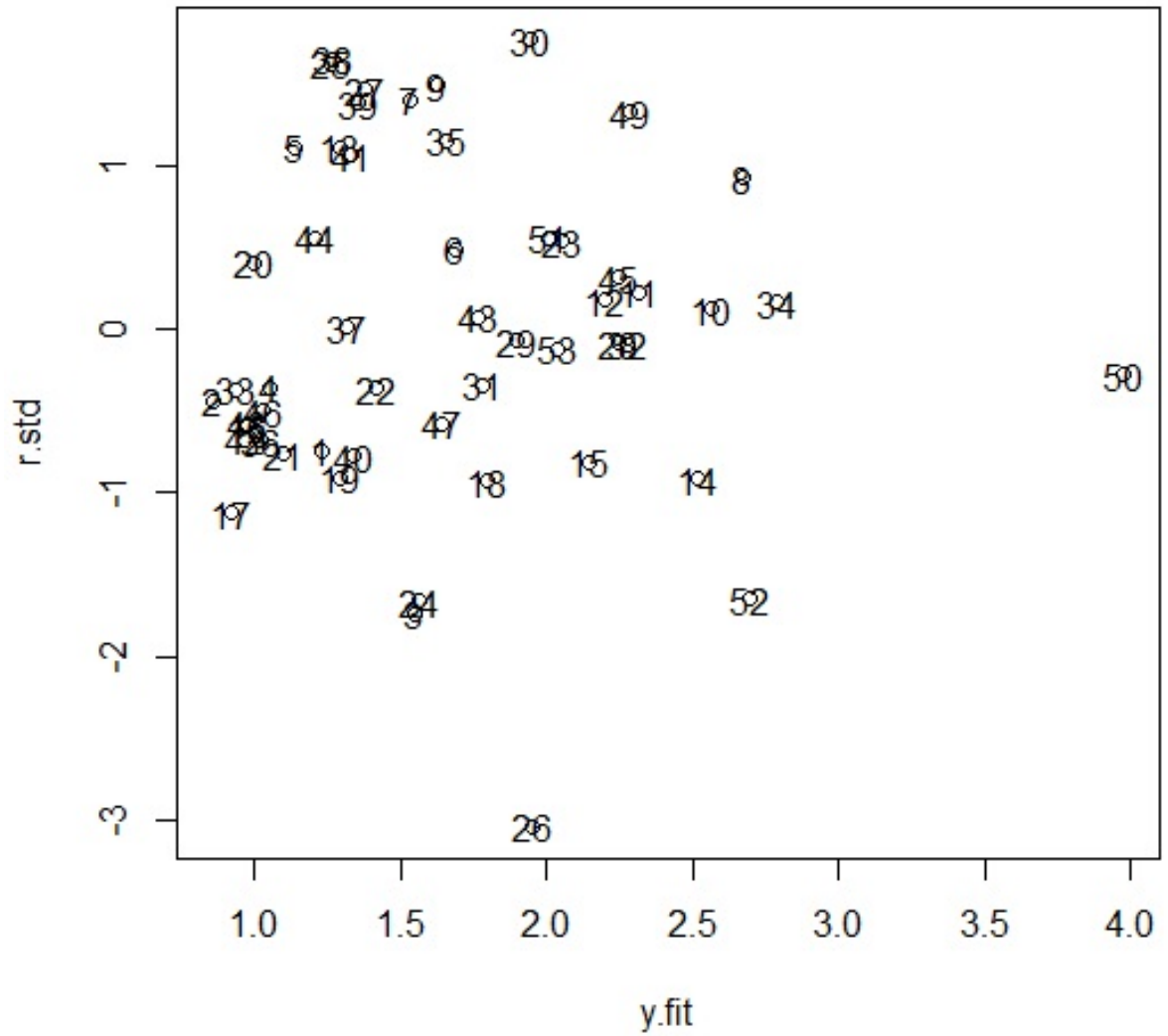
称为$\boldsymbol{\beta}$的置信水平为$1-\alpha$的置信椭球. (3.4.10)与(3.4.11)中的统计量很相似, 但前者并不服从F分布. 然而借助于后者我们可以对D_i值的大小给出以下的概率解释.

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.

对于例3.4.1, 我们做如下的数据诊断分析:

R程序:

```
yx=read.table("ex_p47_data.txt")
x=yx[,1]
y=yx[,2]
z=sqrt(y)
mydata2=data.frame(z,x)
lm.sol2=lm(z~x,data=mydata2)
summary(lm.sol2)
z.fit=predict(lm.sol2)
r.std=rstandard(lm.sol2)
plot(r.std~z.fit)
text(z.fit,r.std,type="1:53")
r.std
influence.measures(lm.sol2)
```



分析: 第26号样本点为异常点. 若有"模棱两可"的点, 则列出学生化残差的值进行进一步确认.

```
> r.std
           1            2            3            4            5
-0.744387230 -0.436329220 -1.737311843 -0.358561912  1.109952186
           6            7            8            9           10
 0.487653003  1.398871568  0.924027947  1.486753696  0.122032935
          11           12           13           14           15
 0.224377723  0.176250615  1.111042097 -0.912031877 -0.809721629
          16           17           18           19           20
-0.593990581 -1.119948730 -0.929176183 -0.905460421  0.401275884
          21           22           23           24           25
-0.764638750 -0.361843861  0.535622776 -1.659938750  1.619903744
          26           27           28           29           30
-3.033586687  1.460144197 -0.084997845 -0.071114186  1.765881261
          31           32           33           34           35
-0.340205437 -0.091377545 -0.367783599  0.163452688  1.152314671
          36           37           38           39           40
-0.680242594  0.006905113  1.644585899  1.380475911 -0.775074198
          41           42           43           44           45
 1.059604108 -0.673843765  0.074686957  0.554983018  0.315052889
          46           47           48           49           50
-0.501789889 -0.577216155 -0.576409293  1.320134935 -0.277001493
          51           52           53
 0.553607956 -1.645551991 -0.117565277
```

```
> influence.measures(lm.sol2)
Influence measures of
         lm(formula = z ~ x, data = mydata2) :

      dfb.1_      dfb.x   dffit cov.r   cook.d    hat inf
1  -0.116369  0.075387 -0.1279 1.048 8.25e-03 0.0289
2  -0.100529  0.080987 -0.1014 1.089 5.23e-03 0.0521
3  -0.167136  0.053283 -0.2518 0.939 3.04e-02 0.0198
4  -0.068530  0.050612 -0.0710 1.076 2.56e-03 0.0384
5   0.195355 -0.136676  0.2070 1.025 2.13e-02 0.0335
6   0.032611  0.000361  0.0671 1.051 2.29e-03 0.0189
7   0.137074 -0.046971  0.2015 0.982 1.99e-02 0.0200
8  -0.118280  0.209606  0.2473 1.078 3.07e-02 0.0670
9   0.117960 -0.017857  0.2095 0.971 2.14e-02 0.0190
10 -0.012863  0.024386  0.0298 1.103 4.52e-04 0.0573
11 -0.012555  0.031940  0.0446 1.080 1.01e-03 0.0387
12 -0.005965  0.020568  0.0319 1.074 5.18e-04 0.0323
13  0.160473 -0.096884  0.1828 1.017 1.66e-02 0.0262
14  0.088426 -0.173844 -0.2162 1.064 2.35e-02 0.0534
15  0.018921 -0.085062 -0.1411 1.045 1.00e-02 0.0297
16 -0.122866  0.094337 -0.1256 1.073 7.99e-03 0.0433
17 -0.246989  0.194476 -0.2506 1.039 3.12e-02 0.0474
18 -0.041800 -0.024499 -0.1310 1.026 8.61e-03 0.0196
19 -0.130582  0.079021 -0.1486 1.034 1.11e-02 0.0263
20  0.081212 -0.061757  0.0833 1.079 3.52e-03 0.0419
21 -0.139874  0.100551 -0.1465 1.054 1.08e-02 0.0357
22 -0.043090  0.021392 -0.0542 1.059 1.50e-03 0.0223
23 -0.002506  0.044542  0.0864 1.056 3.78e-03 0.0257
24 -0.153888  0.044624 -0.2386 0.950 2.75e-02 0.0196
25  0.248879 -0.156457  0.2777 0.963 3.73e-02 0.0276
26 -0.051873 -0.199977 -0.5025 0.715 1.06e-01 0.0224   *
```

27	0.187694	-0.099823	0.2283	0.978	2.55e-02	0.0233	
28	0.003638	-0.010800	-0.0160	1.078	1.30e-04	0.0348	
29	-0.001796	-0.003443	-0.0104	1.063	5.48e-05	0.0212	
30	0.030230	0.106435	0.2723	0.938	3.55e-02	0.0223	
31	-0.016509	-0.007389	-0.0474	1.056	1.14e-03	0.0193	
32	0.004408	-0.012186	-0.0176	1.079	1.57e-04	0.0363	
33	-0.079062	0.061856	-0.0804	1.085	3.28e-03	0.0463	
34	-0.024846	0.041561	0.0476	1.129	1.15e-03	0.0795	*
35	0.083499	-0.005562	0.1604	1.006	1.28e-02	0.0189	
36	-0.134386	0.100812	-0.1384	1.064	9.68e-03	0.0402	
37	0.000955	-0.000561	0.0011	1.068	6.22e-07	0.0254	
38	0.248429	-0.153900	0.2793	0.959	3.77e-02	0.0271	
39	0.184666	-0.102859	0.2193	0.988	2.36e-02	0.0242	
40	-0.104956	0.059993	-0.1230	1.042	7.63e-03	0.0248	
41	0.145639	-0.084047	0.1698	1.021	1.44e-02	0.0250	
42	-0.140573	0.108317	-0.1435	1.069	1.04e-02	0.0438	
43	0.003799	0.001415	0.0104	1.061	5.47e-05	0.0192	
44	0.088758	-0.058583	0.0966	1.059	4.73e-03	0.0298	
45	-0.013379	0.039930	0.0592	1.074	1.78e-03	0.0347	
46	-0.098441	0.073661	-0.1015	1.073	5.23e-03	0.0399	
47	-0.043123	0.004720	-0.0797	1.047	3.22e-03	0.0189	
48	-0.119317	0.091653	-0.1219	1.073	7.53e-03	0.0434	
49	-0.067307	0.182048	0.2604	1.008	3.34e-02	0.0369	
50	0.121751	-0.164710	-0.1706	1.438	1.48e-02	0.2786	*
51	0.002340	0.040354	0.0864	1.053	3.79e-03	0.0241	
52	0.225242	-0.392754	-0.4595	1.003	1.02e-01	0.0700	
53	0.000179	-0.009323	-0.0187	1.067	1.78e-04	0.0251	

总结: 第26号样本点为异常点, 我们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号样本点为强影响点, 需引起注意.

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, 我们要建立智力随年龄变化的关系.

表3.4.2 智力测试数据

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			

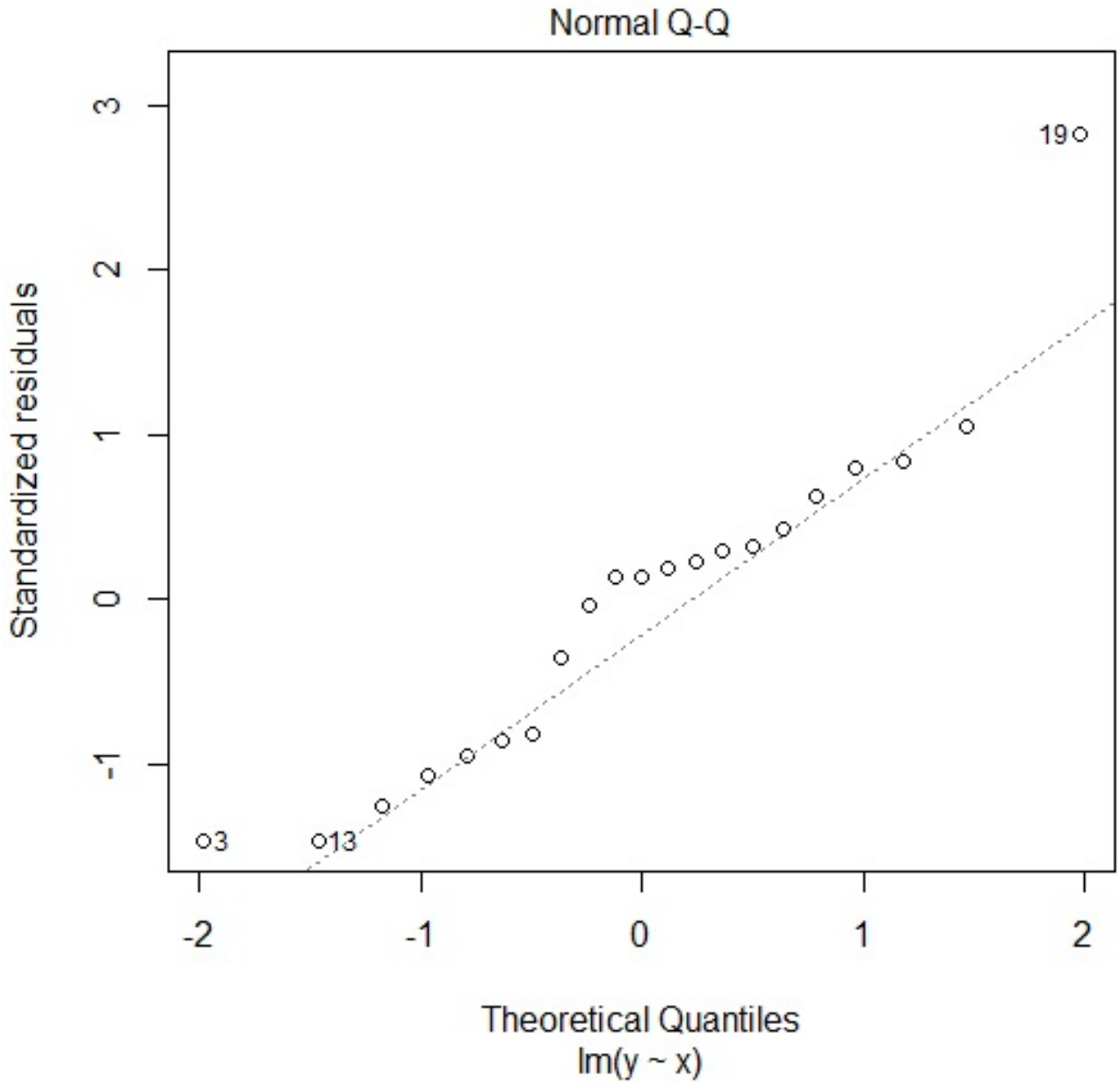
R 程序:

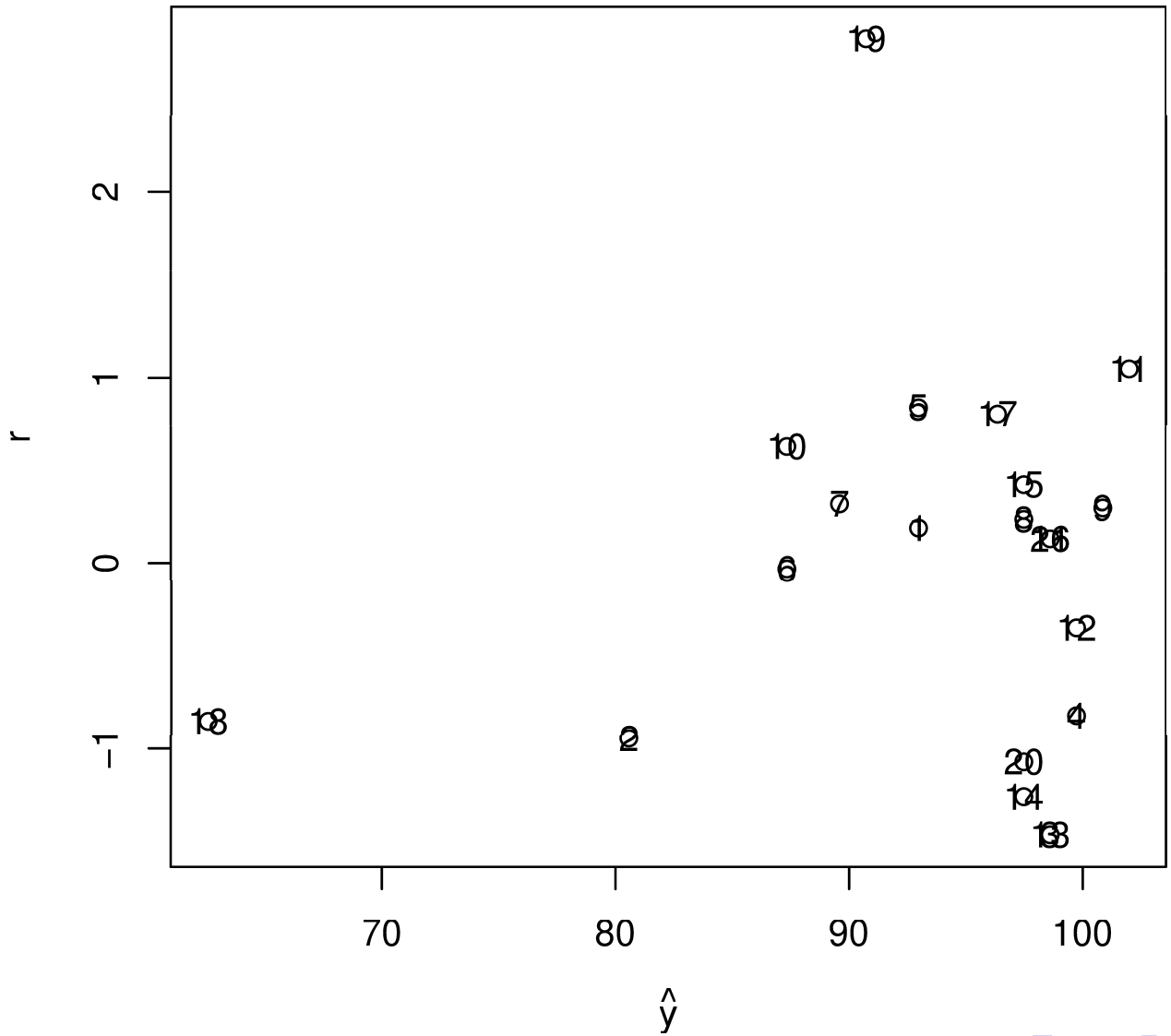
```
x=c(15,26,10,9,15,20,18,11,8,20,7,9,10,11,11,10,12,42,17,11,10)
y=c(95,71,83,91,102,87,93,100,104,94,113,96,83,84,102,100,
  105,57,121,86,100)
lm.sol=lm(y~x)
summary(lm.sol)
influence.measures(lm.sol)
```

```
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
```

```
> influence.measures(lm.sol)
Influence measures of
         lm(formula = y ~ x) :

     dfb.1_    dfb.x    dffit cov.r   cook.d    hat inf
1   0.01664  0.00328  0.04127 1.166 8.97e-04 0.0479
2   0.18862 -0.33480 -0.40252 1.197 8.15e-02 0.1545
3  -0.33098  0.19239 -0.39114 0.936 7.17e-02 0.0628
4  -0.20004  0.12788 -0.22433 1.115 2.56e-02 0.0705
5   0.07532  0.01487  0.18686 1.085 1.77e-02 0.0479
6   0.00113 -0.00503 -0.00857 1.201 3.88e-05 0.0726
7   0.00447  0.03266  0.07722 1.170 3.13e-03 0.0580
8   0.04430 -0.02250  0.05630 1.174 1.67e-03 0.0567
9   0.07907 -0.05427  0.08541 1.200 3.83e-03 0.0799
10 -0.02283  0.10141  0.17284 1.152 1.54e-02 0.0726
11  0.31560 -0.22889  0.33200 1.088 5.48e-02 0.0908
12 -0.08422  0.05384 -0.09445 1.183 4.68e-03 0.0705
13 -0.33098  0.19239 -0.39114 0.936 7.17e-02 0.0628
14 -0.24681  0.12536 -0.31367 0.992 4.76e-02 0.0567
15  0.07968 -0.04047  0.10126 1.159 5.36e-03 0.0567
16  0.02791 -0.01622  0.03298 1.187 5.74e-04 0.0628
17  0.13328 -0.05493  0.18717 1.096 1.79e-02 0.0521
18  0.83112 -1.11275 -1.15578 2.959 6.78e-01 0.6516   *
19  0.14348  0.27317  0.85374 0.396 2.23e-01 0.0531   *
20 -0.20761  0.10544 -0.26385 1.043 3.45e-02 0.0567
21  0.02791 -0.01622  0.03298 1.187 5.74e-04 0.0628
>
```

```
plot(lm.sol,which=2)  /*QQ图*/
y.fit=predict(lm.sol)
e=y-y.fit
r.std=rstandard(lm.sol)
plot(r.std~y.fit,xlab=expression(hat(y)),ylab="r")
text(y.fit,r.std,type="1:25")
```


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
19
3
13
-2
-1
0
1
2
3


r
2
1
0
−1
70
80
90
100
ŷ

总结: 第19号样本点为异常点, 第18,19号的样本点是强影响点. 由QQ图和残差图, 我们接受线性假设、方差齐性假设、不相关性假设和正态性假设.

Box-Cox变换

对于观测数据, 若经过回归诊断后得知, 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, 那么我们就要对数据采取"治疗"措施. 实践证明, 数据变换是处理有问题数据的一种好方法.

本节介绍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. 实践证明, Box-Cox变换对许多实际数据都是行之有效的.

Box-Cox变换是对因变量的如下变换:

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

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, 诸如对数变换($\lambda = 0$), 倒数变换($\lambda = -1$)和平方根变换($\lambda = 1/2$)等等.

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

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

我们要确定变换参数λ使得$\boldsymbol{Y}^{(\lambda)}$满足

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

即要求变换后的向量$\boldsymbol{Y}^{(\lambda)}$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.

我们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

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

$\ln L(\boldsymbol{\beta},\sigma^2)$关于$\boldsymbol{\beta}$和$\sigma^2$求导并令其等于零, 可得$\boldsymbol{\beta}$和$\sigma^2$的MLE为

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

对应的似然函数最大值为

$$\begin{aligned} L_{\mathsf{max}}(\lambda) &= L(\hat{\boldsymbol{\beta}}(\lambda), \hat{\sigma}^2(\lambda)) \\ &= (2\pi e)^{-n/2} \cdot |J| \cdot \left(\frac{\mathsf{RSS}(\lambda, \boldsymbol{Y}^{(\lambda)})}{n}\right)^{-n/2}. \end{aligned} \tag{3.5.4}$$

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

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

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

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

(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, 从图上可以近似的找出使$\mathrm{RSS}(\lambda, \boldsymbol{Z}^{(\lambda)})$ 达到最小值的$\hat{\lambda}$.

Box-Cox变换的具体步骤:

1. 对给定的λ值,计算$z_i^{(\lambda)}$, $i = 1 \cdots, n$.

2. 按(3.5.6)式计算残差平方和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RSS$(\lambda, \boldsymbol{Z}^{(\lambda)})$的一串值. 以$\lambda$为横轴, RSS$(\lambda, \boldsymbol{Z}^{(\lambda)})$ 为纵轴, 画出相应的曲线. 用直观方法, 找出使RSS$(\lambda, \boldsymbol{Z}^{(\lambda)})$达到最小值的$\hat{\lambda}$.

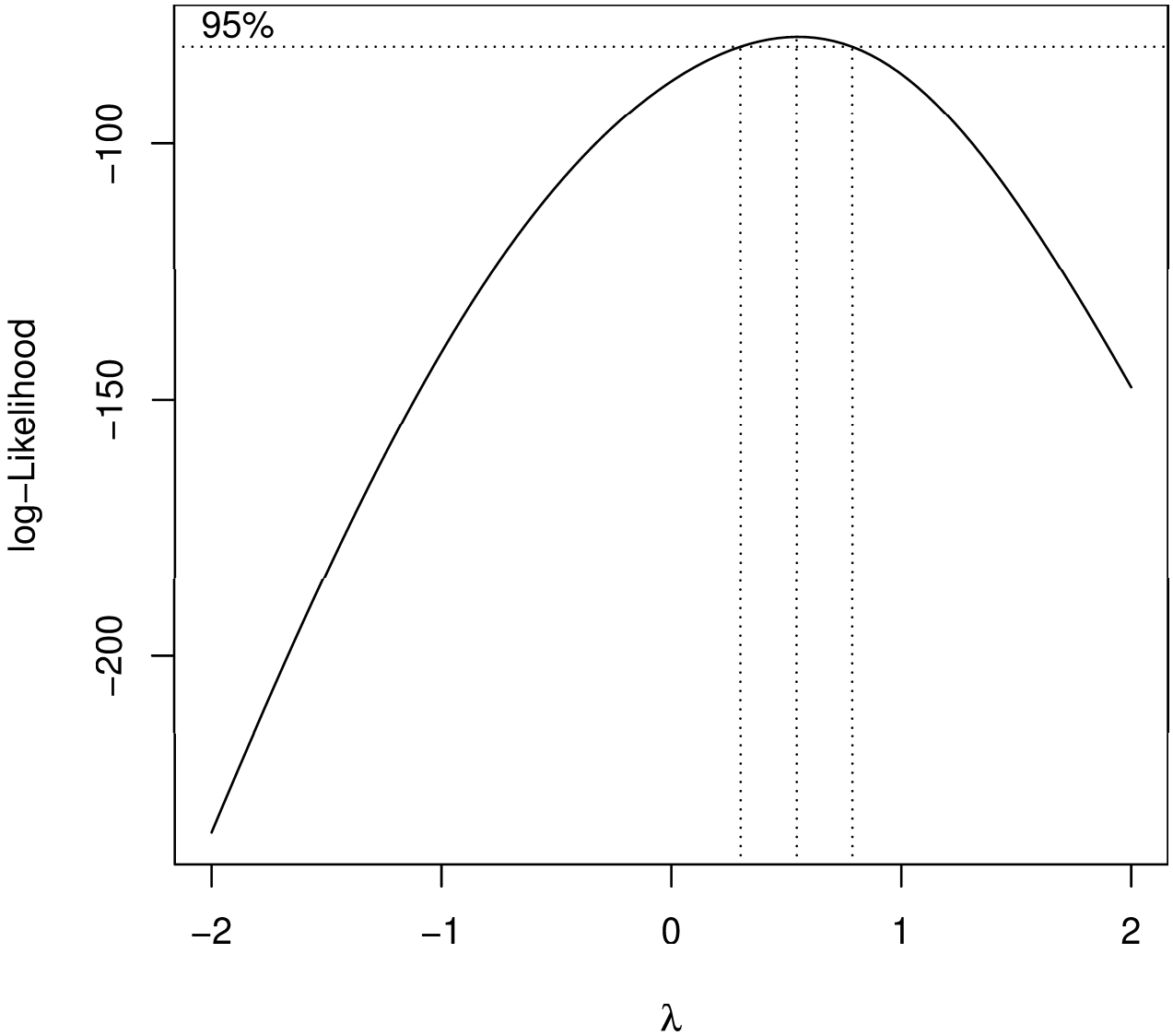
例3.5.1 在例3.4.1中, 我们对因变量y作了平方根变换, 这相当于使用$\lambda = 0.5$的Box-Cox变换. 我们现在来证实这样的变换是合适的. 下表给出了12个不同的λ值对应的残差平方和$\mathrm{RSS}(\lambda, \boldsymbol{Z}^{(\lambda)})$, 简单比较后可发现当$\lambda = 0.5$时残差平方和$\mathrm{RSS}(\lambda, \boldsymbol{Z}^{(\lambda)})$达到最小. 因此近似地认为0.5就是变换参数$\lambda$的最优选择.

表3.5.1

λ	-2	-1	-0.5	0	0.125	0.25
RSS	34101.04	986.04	291.59	134.10	119.20	107.21
λ	0.375	0.5	0.625	0.75	1	2
RSS	100.26	96.95	97.29	101.69	127.87	1275.56

R程序:

```
x<-
c(679,292,1012,493,582,1156,997,2189,1097,2078,1818,1700,747,
2030,1643,414,354,1276,745,435,540,874,1543,1029,710,1434,837,1748,
1381,1428,1255,1777,370,2316,1130,463,770,724,808,790,783,406,1242,
658,1746,468,1114,413,1787,3560,1495,2221,1526)
y<-c(0.79,0.44,0.56,0.79,2.70,3.64,4.73,9.50,5.34,6.85,5.84,5.21,
3.25,4.43,3.16,0.50,0.17,1.88,0.77,1.39,0.56,1.56,5.28,0.64,4.00,0.31,
4.20,4.88,3.48,7.58,2.63,4.99,0.59,8.19,4.79,0.51,1.74,4.10,3.94,0.96,
3.29,0.44,3.24,2.14,5.71,0.64,1.90,0.51,8.33,14.94,5.11,3.85,3.93)
lm.sol<-lm(y~x)
library(MASS)
boxcox(lm.sol,plotit=T,lambda=seq(-2,2,by=0.05))
```


95%
log−Likelihood
−100
−150
−200
−2
−1
0
1
2
λ

广义最小二乘估计

在前面的讨论中, 我们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). 这时我们需假设误差向量的协方差矩阵为$\mathsf{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{\Sigma}$, 这里$\boldsymbol{\Sigma}$是一个正定矩阵. 这个$\boldsymbol{\Sigma}$包含未知参数, 但我们这里假设$\boldsymbol{\Sigma}$ 是完全已知的.

我们要讨论的线性回归模型为

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

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

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

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. $\boldsymbol{\Sigma}^{-\frac{1}{2}}$表示$\boldsymbol{\Sigma}^{\frac{1}{2}}$的逆矩阵.

我们把线性回归模型(3.6.1)进行正交变换. 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$\mathrm{Cov}(\varepsilon) = \boldsymbol{\Sigma}^{-\frac{1}{2}}\sigma^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \sigma^2\boldsymbol{I}_n$, 于是我们得如下的线性回归模型

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

这是我们已讨论过的模型. 在这新模型中, 可得$\boldsymbol{\beta}$的LSE为

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

我们称$\boldsymbol{\beta}^*$为$\boldsymbol{\beta}$的广义最小二乘估计(GLSE). 这个估计具有良好的统计性质.

定理 (3.6.1)

(a) $E(\boldsymbol{\beta}^*) = \boldsymbol{\beta}$;
(b) $Cov(\boldsymbol{\beta}^*) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$;
(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, 而$\boldsymbol{b}'\boldsymbol{Y} = \boldsymbol{b}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$为$\boldsymbol{c}'\boldsymbol{\beta}$的线性无偏估计. 所以对模型(3.6.2)应用Gauss-Markov定理知

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

等号成立当且仅当$\boldsymbol{c}'\boldsymbol{\beta}^* = \boldsymbol{b}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} = \boldsymbol{b}'\boldsymbol{Y}$.

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$\mathrm{Var}(\boldsymbol{c}'\boldsymbol{\beta}^*) \leq \mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})$. 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)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

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

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(WLSE).

σ_i^2往往是未知的, 这时我们需要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).

例3.6.1 假设我们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

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

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$\boldsymbol{X} = (1, \cdots, 1)'$, 于是$\mu$的GLSE为

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

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$\frac{\omega_1}{\omega_1+\omega_2}$ 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, 对应的权就小.

μ^*包含未知参数σ_1^2和σ_2^2, 不能付诸实际应用. 我们可以设法构造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$\boldsymbol{Y}_i = (y_{1i}, \cdots, y_{1n_i})'$, $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{in_i})'$. 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.

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.

但是我们以前讨论的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, 这些都致使设计矩阵$\boldsymbol{X}$的列向量之间不可能完全线性无关. 很多情况下, 设计矩阵$\boldsymbol{X}$的列向量之间存在多重共线性/复共线性(multi-collinearity)关系.

定义

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/复共线性关系.

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

定义 (多重共线性/复共线性关系)

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多, 多重共线性情形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.

我们先引入一个概念: 均方误差(MSE: Mean Squared Errors), 它是用来评价一个估计优劣的标准的.

定义

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

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

定理 (3.7.1)

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

证明: 不难看出

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

利用迹的性质,

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

$\Delta_2 = \mathrm{E}[(\mathrm{E}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\mathrm{E}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \|\mathrm{E}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$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所决定. 一个好的估计应有较小的方差和偏差.

定理

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

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

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

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

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

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

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

利用迹的性质马上可得

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

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

(b) 因为

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

于是

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

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却很大, 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.

我们来分析"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.

记$\boldsymbol{X}=(\mathbf{1},\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p)$, 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, $\boldsymbol{\phi}$为其对应的特征向量, 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

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

即设计矩阵$\boldsymbol{X}$的列向量之间(即自变量之间)有多重共线性关系.

反之, 若设计矩阵$\boldsymbol{X}$的列向量之间有多重共线性关系, 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$\boldsymbol{X}$的列向量之间有多重共线性关系是等价的. 这时称设计矩阵$\boldsymbol{X}$是病态矩阵.

多重共线性的诊断:

(1) 方差膨胀因子(VIF: variance inflation factor)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对其余$p-1$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:

当有某个

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

时, 认为自变量之间存在严重的多重共线性; 或者当

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

时, 认为自变量之间存在严重的多重共线性.

(2)特征根与条件数(CI: Condition Index)诊断法

为了消除量纲的影响, 我们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$n \times p$的矩阵, $\boldsymbol{X}'\boldsymbol{X}$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$\lambda_{\mathsf{max}}$, 最小特征根为$\lambda_{\mathsf{min}}$, 称

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

为特征根λ_j的条件数.

记

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

它可以用来度量矩阵$\boldsymbol{X}'\boldsymbol{X}$的特征根近似为零的程度, 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.

条件数判定准则:

- 若$0 < \kappa < 100$, 则认为不存在多重共线性;
- 若$100 \leq \kappa \leq 1000$, 则认为存在较强的多重共线性;
- 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, 原始数据见下表.

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

R程序:

```
yx=read.table("ex_p62_data.txt")
y=yx[,1]
x1=yx[,2]
x2=yx[,3]
x3=yx[,4]
x4=yx[,5]
x5=yx[,6]
x6=yx[,7]
mydata=data.frame(y,x1,x2,x3,x4,x5,x6)
lm.sol=lm(y~ x1+x2+x3+x4+x5+x6,data=mydata)
summary(lm.sol)
library(DAAG)
vif(lm.sol)
```

VIF输出结果

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

因为前四个VIP都不小于10, 所以认为自变量之间存在严重的多重共线性.

特征根与条件数诊断法:

R程序:

```
X=cbind(x1,x2,x3,x4,x5,x6)
X
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*/
```

特征根、特征向量与条件数:

```
> 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)
$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

> kappa(rho,exact=TRUE)
[1] 2195.908
>
```

$\lambda_{\min} = 0.001106051$, 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

$$c_0 + 0.44768x_1 + 0.42114x_2 + 0.541689x_3 + 0.57337x_4 + 0.00605x_5 + 0.00217x_6 \approx 0.$$

或写为

$$c_0 + 0.44768x_1 + 0.42114x_2 + 0.541689x_3 + 0.57337x_4 \approx 0.$$

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:

- 增大样本容量, 消除或缓解自变量的线性相关性.
- 牺牲无偏性, 寻找有偏估计, 我们将介绍岭估计和主成分估计.

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

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

接近奇异的可能性要比$(\boldsymbol{X}'\boldsymbol{X})^{-1}$接近奇异的可能性小得多. 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$\boldsymbol{\beta}$的估计应该比最小二乘估计要稳定一些.

定义

对给定的$0 < k < \infty$, 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. 由岭估计所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, 我们得到不同的估计. 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, 我们提起岭估计, 是不包括LSE的.

在进行岭估计之前, 为了消除量纲的影响, 我们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.

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$0 < k < \infty$, $\mathsf{E}(\hat{\boldsymbol{\beta}}(k)) \neq \boldsymbol{\beta}$.

有偏性是岭估计与最小二乘估计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. 我们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.

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

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

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

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

我们称(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

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

注: 易知均方误差在估计和参数的正交变换下保持不变, 所以典则回归系数的最小二乘估计(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:

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

定理 (3.8.1, 岭估计存在性定理)

存在$k > 0$使得

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

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$f(0) = \mathrm{MSE}(\hat{\boldsymbol{\alpha}})$. 若我们能证明$f(k)$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

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

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, 这个k是下列方程的解:

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

这个解依赖于未知参数$\alpha_i, i = 1, \cdots, p$和$\sigma^2$, 所以不可能从解方程的角度获得岭参数$k$. 统计学家们从其它途径提出了选择岭参数$k$的方法.

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$i = 1, \cdots, p$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$\hat{\sigma}^2$ 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:

- 各回归系数的岭估计大致比较稳定;
- 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;
- 回归系数没有不合理的符号;
- 残差平方和不要上升太多.

一般情况下, 我们选择最小的k值能使得各条岭迹都开始趋于稳定.

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, 见下表.

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

R程序:

```
library(MASS)
yx=read.table("ex_p68_data.txt")
x1=yx[,1]
x2=yx[,2]
x3=yx[,3]
y=yx[,4]
mean(x1);mean(x2);mean(x3);mean(y)
sd(x1);sd(x2);sd(x3);sd(y)
mydata=data.frame(y, x1, x2, x3)
lm.sol=lm(y~x1+x2+x3)
summary(lm.sol)
```

获得各变量的样本均值、样本标准差数据:

```
> 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
>
```

使用最小二乘方法的结果:

```
> summary(lm.sol)

Call:
lm(formula = y ~ x1 + x2 + x3)

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
```

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, 这是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
>
```

因此x_1与x_3存在高度的相关性.

作多重共线性诊断, 可以明确发现自变量之间存在着多重共线性:

```
> library(DAAG)
> vif(lm.sol)
      x1       x2       x3
186.0000   1.0189 186.1100
> eigen(rho)
$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
>
```

用岭估计方法寻找岭回归方程:

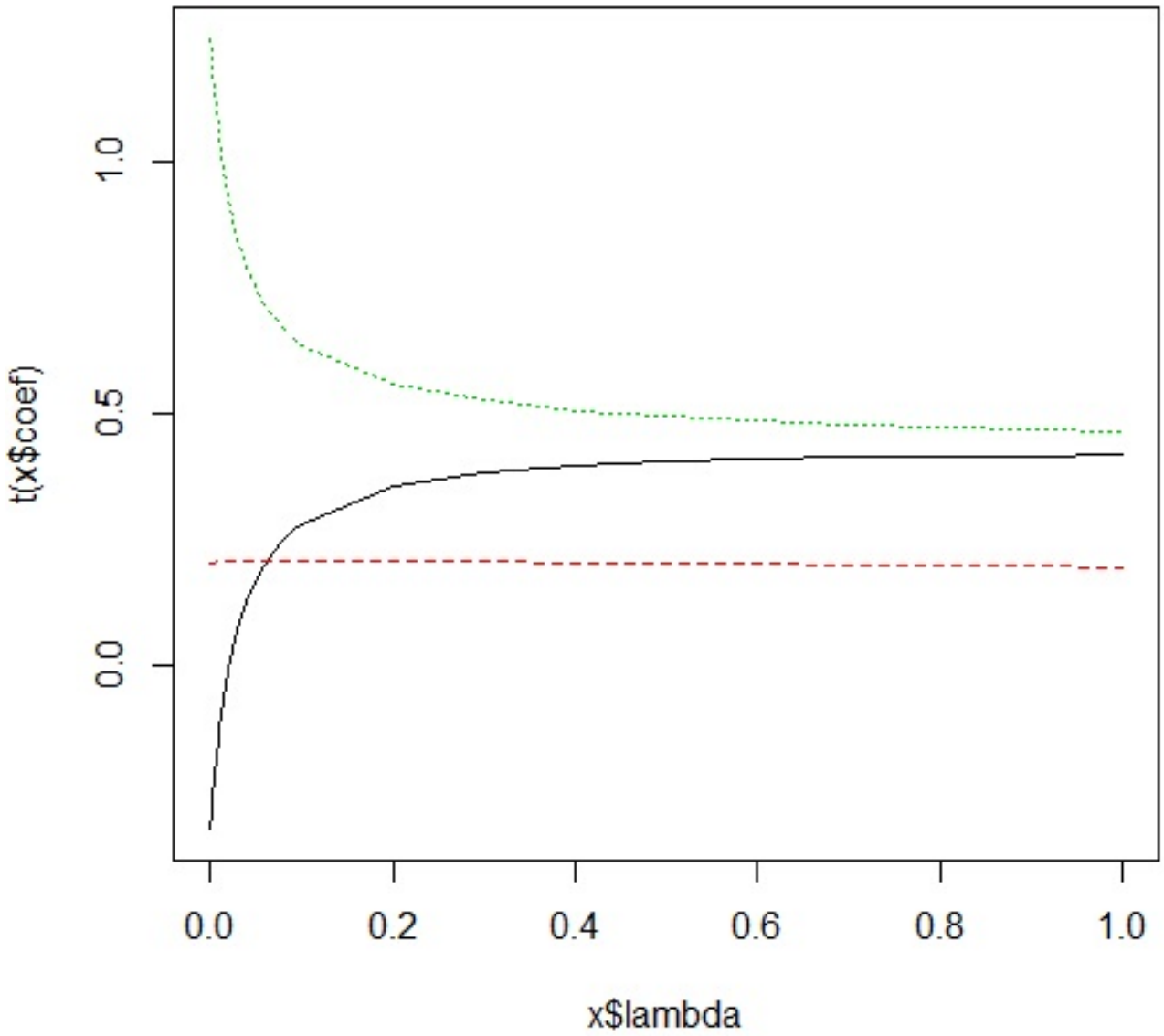
先对数据进行标准化, R程序:

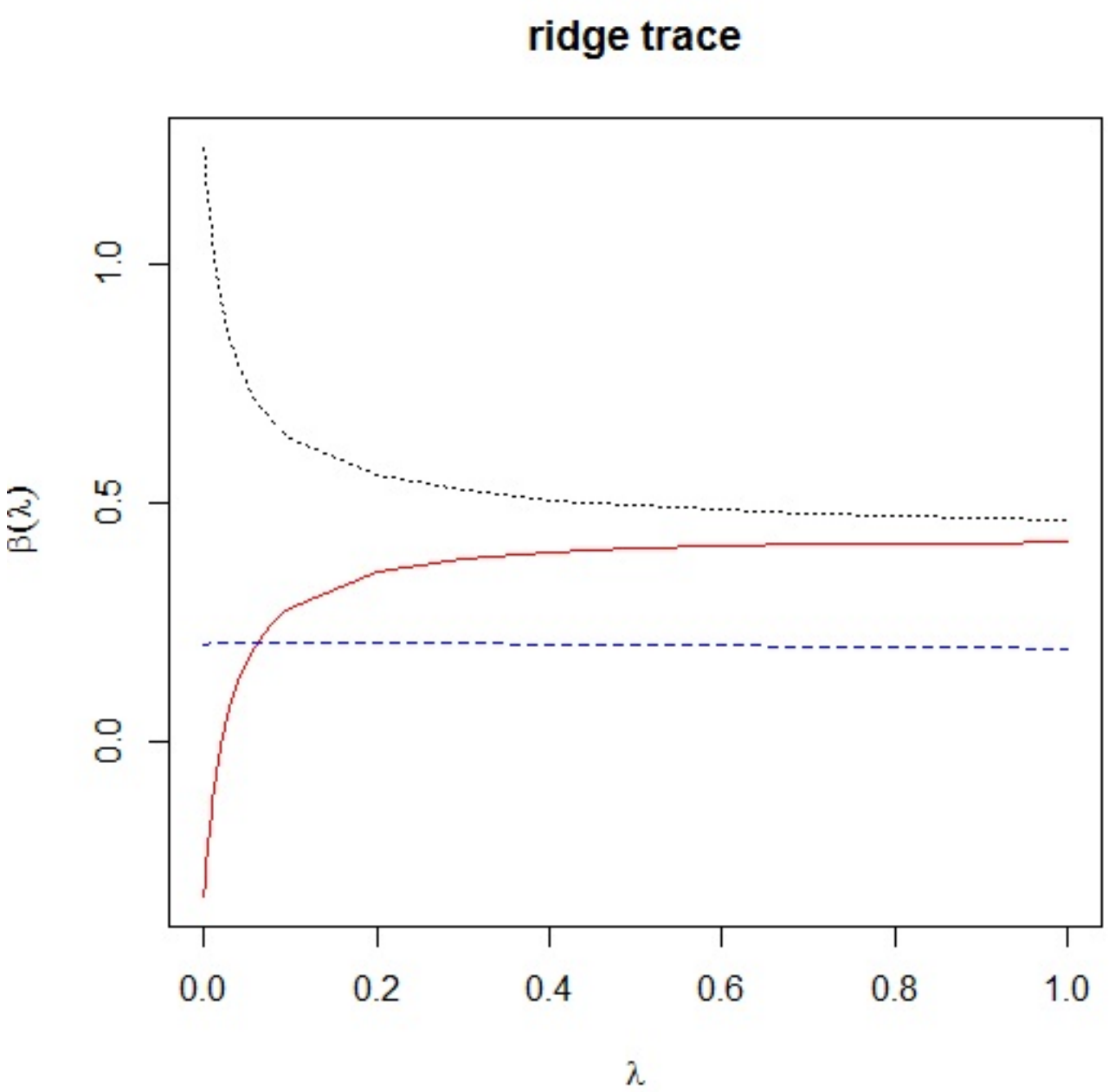
```
yxs=scale(yx)
x1=yxs[,1]
x2=yxs[,2]
x3=yxs[,3]
y=yxs[,4]
mydata2=data.frame(x1,x2,x3,y)
mydata2
```

```
rr.sol=lm.ridge(y~0+x1+x2+x3,data=mydata2,lambda=
c(seq(0,0.01,by=0.001),seq(0.02,0.1,by=0.01),seq(0.2,1,by=0.1)))
rr.sol  /*显示岭估计*/
plot(rr.sol)  /*作岭迹图*/
matplot(rr.sol$lambda,t(rr.sol$coef),type="l",col=c("red","blue",
  "black"),main="ridge trace",xlab=expression(lambda),
ylab=expression(hat(beta)(lambda)))  /*作岭迹图*/
```

岭估计数据:

```
> rr.sol
                x1         x2        x3
0.000 -0.339342628 0.2130484 1.3026815
0.001 -0.312570821 0.2132939 1.2758583
0.002 -0.287494805 0.2135225 1.2507314
0.003 -0.263958525 0.2137361 1.2271447
0.004 -0.241824502 0.2139358 1.2049607
0.005 -0.220971154 0.2141228 1.1840578
0.006 -0.201290557 0.2142983 1.1643279
0.007 -0.182686584 0.2144632 1.1456751
0.008 -0.165073332 0.2146184 1.1280132
0.009 -0.148373803 0.2147645 1.1112654
0.010 -0.132518774 0.2149023 1.0953623
0.020 -0.009178625 0.2159333 0.9715526
0.030  0.072700065 0.2165559 0.8892173
0.040  0.130988751 0.2169461 0.8304795
0.050  0.174578235 0.2171918 0.7864459
0.060  0.208390199 0.2173414 0.7521931
0.070  0.235369741 0.2174239 0.7247752
0.080  0.257387107 0.2174578 0.7023213
0.090  0.275687037 0.2174554 0.6835864
0.100  0.291130095 0.2174252 0.6677096
0.200  0.370560751 0.2164006 0.5839853
0.300  0.400521805 0.2149193 0.5497830
0.400  0.415571800 0.2133213 0.5305332
0.500  0.424167477 0.2116891 0.5177761
0.600  0.429392660 0.2100524 0.5084265
0.700  0.432640361 0.2084241 0.5010906
0.800  0.434631830 0.2068108 0.4950464
0.900  0.435777879 0.2052158 0.4898824
1.000  0.436329637 0.2036409 0.4853472
>
```


t(x$coef)
0.0
0.5
1.0
0.0
0.2
0.4
0.6
0.8
1.0
x$lambda


ridge trace
β(λ)
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
λ

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:

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

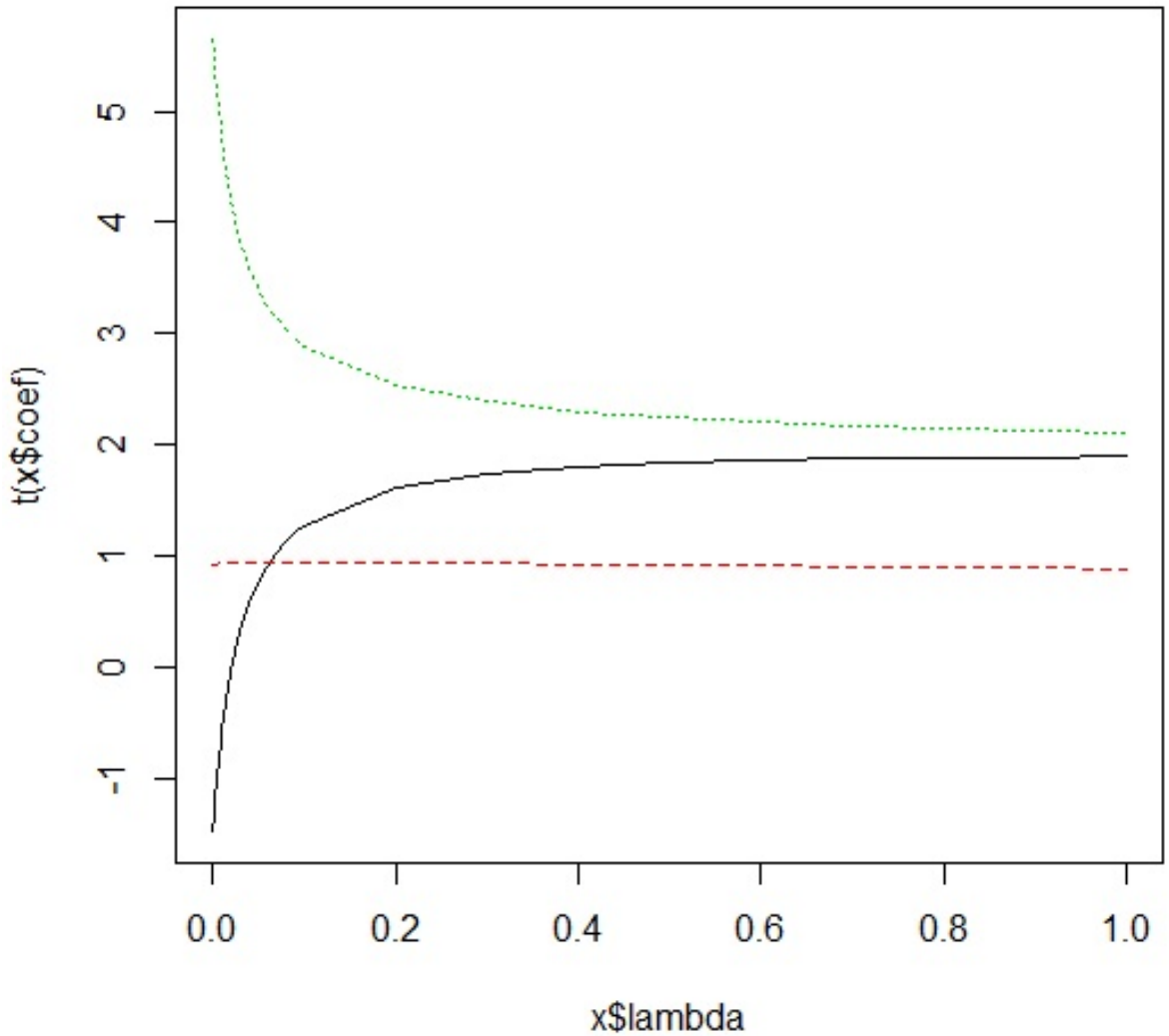
即

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

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:

```
rr.sol=lm.ridge(y~x1+x2+x3,data=mydata,lambda=
 c(seq(0,0.01,by=0.001),seq(0.02,0.1,by=0.01),seq(0.2,1,by=0.1)))
rr.sol
plot(rr.sol)
```

```
> rr.sol
                              x1          x2          x3
0.000 -10.127988 -0.051396160 0.5869490 0.2868487
0.001 -10.093904 -0.047341355 0.5876252 0.2809422
0.002 -10.061883 -0.043543392 0.5882552 0.2754093
0.003 -10.031737 -0.039978633 0.5888434 0.2702156
0.004 -10.003298 -0.036626258 0.5893936 0.2653307
0.005  -9.976418 -0.033467852 0.5899090 0.2607279
0.006  -9.950965 -0.030487068 0.5903925 0.2563834
0.007  -9.926823 -0.027669347 0.5908468 0.2522761
0.008  -9.903886 -0.025001679 0.5912742 0.2483870
0.009  -9.882062 -0.022472401 0.5916767 0.2446991
0.010  -9.861266 -0.020071030 0.5920564 0.2411973
0.020  -9.696169 -0.001390176 0.5948969 0.2139345
0.030  -9.581550  0.011011007 0.5966120 0.1958044
0.040  -9.495657  0.019839296 0.5976872 0.1828705
0.050  -9.427670  0.026441272 0.5983640 0.1731743
0.060  -9.371598  0.031562365 0.5987761 0.1656319
0.070  -9.323856  0.035648633 0.5990035 0.1595945
0.080  -9.282169  0.038983339 0.5990970 0.1546502
0.090  -9.245021  0.041755011 0.5990904 0.1505248
0.100  -9.211364  0.044093985 0.5990070 0.1470287
0.200  -8.966880  0.056124395 0.5961843 0.1285928
0.300  -8.784039  0.060662236 0.5921033 0.1210615
0.400  -8.620758  0.062941679 0.5877007 0.1168227
0.500  -8.466721  0.064243563 0.5832040 0.1140136
0.600  -8.318156  0.065034959 0.5786948 0.1119548
0.700  -8.173345  0.065526849 0.5742091 0.1103395
0.800  -8.031391  0.065828473 0.5697644 0.1090085
0.900  -7.891774  0.066002051 0.5653702 0.1078714
1.000  -7.754169  0.066085619 0.5610313 0.1068728
>
```



选择岭参数$k = 0.4$, 得岭回归方程:

$$\hat{y} = -8.621 + 0.063x_1 + 0.588x_2 + 0.117x_3.$$

广义岭估计:

记$\boldsymbol{K} = \mathsf{diag}(k_1, \cdots, k_p)$, $k_i \geq 0$, $i = 1, \cdots, p$. 称

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

为$\boldsymbol{\beta}$的广义岭估计.

主成分估计

主成分(Principle Component)估计是由W.F.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: 首先借助于正交变换将回归自变量变为对应的主成分(主成分的观测向量是互不相关的, 从而消除了多重共线性问题), 然后从所有的主成分中选取一部分重要的主成分(起到降维的作用)并以它们作为新的回归自变量建立新的回归模型, 用最小二乘方法估计新模型中的回归系数并得到回归方程. 基于得到的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:

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

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

$$\boldsymbol{\Phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Phi} = \mathsf{diag}(\lambda_1, \cdots, \lambda_p) \stackrel{\Delta}{=} \boldsymbol{\Lambda}.$$

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

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

在上述的线性回归典则模型(3.9.2)中, 新的设计矩阵

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

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量, 变换的系数向量为特征根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:

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

主成分的性质: 任意两个的主成分的观测向量都是互不相关的, 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$\boldsymbol{Z}'\boldsymbol{Z}=\boldsymbol{\Phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Phi}=\boldsymbol{\Lambda}=\mathsf{diag}(\lambda_1,\cdots,\lambda_p)$, 所以

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

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

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

λ_j度量了第j个主成分z_j的取值变动大小. 因为$\lambda_1 \geq \cdots \geq \lambda_p > 0$, 所以我们称$z_1$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互不相关的, 所以新的自变量$z_1, \cdots, z_p$(或设计矩阵$\boldsymbol{Z}$)不存在多重共线性问题.

由性质1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就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就可以忽略掉, 可将它们从回归模型中剔除. 用最小二乘法对剩下的r个主成分(即r个新的自变量)作回归即可. 最后再变回到原始变量的回归方程.

注: 主成分回归的主要目的: 正交(消除多重共线性), 降维(减少计算量), 然后建立回归方程.

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

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

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. 剔除$\boldsymbol{Z}_2\boldsymbol{\alpha}_2$, 模型(3.9.2)变为

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

$\boldsymbol{Z}_1$不是病态矩阵, 所以可直接应用最小二乘估计得$\boldsymbol{\alpha}_1$的LSE

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

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

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

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

主成分估计的过程:

Step1: 做正交变换$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Phi}$, 获得新的自变量, 称为主成分.

Step2: 做回归自变量选择, 剔除对应的特征根比较小的那些主成分.

Step3: 将剩余的主成分对y做最小二乘回归, 再返回到原来的参数估计, 得到关于原始变量的主成分回归方程.

主成分估计的性质:

性质1. $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'\hat{\boldsymbol{\beta}}$, 即主成分估计是最小二乘估计的一个线性变换.

证明: 根据下列关系:

$$\boldsymbol{Z} = (\boldsymbol{Z}_1 \vdots \boldsymbol{Z}_2) = (\boldsymbol{X}\boldsymbol{\Phi}_1 \vdots \boldsymbol{X}\boldsymbol{\Phi}_2),\ \boldsymbol{\Phi}_1'\boldsymbol{\Phi}_1 = \boldsymbol{I}_r,\ \boldsymbol{\Phi}_1'\boldsymbol{\Phi}_2 = \boldsymbol{0}$$

及

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

并注意到$\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$($\boldsymbol{H}$为帽子矩阵), 可知

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

性质2. $\mathsf{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\Phi}_1\boldsymbol{\Phi}_1'\boldsymbol{\beta}$. 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3. $\|\tilde{\boldsymbol{\beta}}\| \leq \|\hat{\boldsymbol{\beta}}\|$, 即主成分估计是压缩估计.

证明: 令$\tilde{\boldsymbol{I}} = \mathsf{diag}(\boldsymbol{I}_r, \boldsymbol{0})$, 则由$\boldsymbol{\Phi}$的定义知

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

从而有

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

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

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

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

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

因为

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

所以

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

于是

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

当且仅当

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

当多重共线性足够严重的时候, $\lambda_{r+1}, \cdots, \lambda_p$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.

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

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

这就是说, 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)表示了参数空间中(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:

(a) 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.

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

主成分个数r的选取:

(1)略去特征根接近于零的那些主成分.

(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

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

例3.9.1 (续例3.8.1)法国经济数据分析问题.

先进行例行的最小二乘估计与多重共线性诊断:

```
yx=read.table("ex_p68_data.txt")
x1=yx[,1]
x2=yx[,2]
x3=yx[,3]
y=yx[,4]
economy=data.frame(x1,x2,x3,y)
economy
lm.sol=lm(y~x1+x2+x3,data=economy)
summary(lm.sol)
```

```
X=cbind(x1,x2,x3)
rho=cor(X)
rho
library(DAAG)
vif(lm.sol)
eigen(rho)
kappa(rho,exact=TRUE)
```

为消除多重共线性的影响, 做主成分回归:

```
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
```

第三个特征根$\lambda_3 = 0.0518737839^2 = 0.00269 \approx 0.$

对应的三个标准正交化特征向量为

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

三个主成分分别为

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

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$0.9991030 > 0.85$, 所以我们删去第三个主成分, 只保留前两个主成分.

计算主成分得分(即新变量的观测值向量):

```
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:

```
z1=pre[,1];z2=pre[,2]
yxs=scale(yx)
y=yxs[,4]
mydata=data.frame(y,z1,z2)
pc.sol=lm(y~0+z1+z2,data=mydata)
summary(pc.sol)
```

```
Call:
lm(formula = y ~ 0 + z1 + z2, data = mydata)

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:

$$\begin{aligned}\hat{u} &= -0.65787z_1 - 0.1824z_2 \\ &= -0.65787 \times (-0.706x_1^* - 0.707x_3^*) - 0.1824 \times (-0.999x_2^*) \\ &= 0.46446x_1^* + 0.18222x_2^* + 0.46511x_3^*.\end{aligned}$$

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

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

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.

下表给出了最小二乘估计、岭估计和主成分估计的比较:

方法	常数项	x_1	x_2	x_3
最小二乘估计	-10.128	-0.051	0.587	0.287
岭估计$(k = 0.04)$	-8.655	0.063	0.587	0.117
主成分估计$(r = 2)$	-7.768	0.070	0.502	0.102

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的回归系数的符号也发生了变化.

Stein压缩估计

岭估计和主成分估计这两种有偏估计是对最小二乘估计$\hat{\beta}$向原点作适当的压缩. 但他们对参数估计$\hat{\beta}$的各分量作的是非均匀压缩. 本节将基于Stein的压缩思想讨论最小二乘估计$\hat{\beta}$的均匀压缩估计, 该估计是Stein在1955年提出来的.

定义 (Stein估计)

在线性回归模型*(3.1.5)*中, 其回归系数$\boldsymbol{\beta}$的最小二乘估计为$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. 我们称

$$\hat{\boldsymbol{\beta}}_s(c) = c\hat{\boldsymbol{\beta}}$$

为*Stein*估计, 其中$0 \leq c \leq 1$被称为压缩系数.

显然, $\hat{\boldsymbol{\beta}}_s(c)$是对$\hat{\boldsymbol{\beta}}$的每一分量作了一样的压缩, 所以Stein估计是一种均匀压缩估计.

Stein压缩估计的性质:

性质1. 当$c \neq 1$时, $\hat{\boldsymbol{\beta}}_s(c)$是$\hat{\boldsymbol{\beta}}$的有偏、压缩估计.

性质2. 存在$0 < c < 1$, 使得$\mathsf{MSE}(\hat{\boldsymbol{\beta}}_s(c)) < \mathsf{MSE}(\hat{\boldsymbol{\beta}})$.

证明: $\hat{\boldsymbol{\beta}}_s(c)$的均方误差为

$$\begin{aligned}
\mathsf{MSE}(\hat{\boldsymbol{\beta}}_s(c)) &= \mathsf{tr}[\mathsf{Cov}(\hat{\boldsymbol{\beta}}_s(c))] + \|\mathsf{E}(\hat{\boldsymbol{\beta}}_s(c)) - \boldsymbol{\beta}\|^2 \\
&= c^2\sigma^2 \mathsf{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}] + (c-1)^2\|\boldsymbol{\beta}\|^2 \\
&= c^2\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} + (c-1)^2\|\boldsymbol{\beta}\|^2 \\
&\triangleq g(c).
\end{aligned}$$

这里, 为了消除量纲的影响, 我们假设自变量和因变量均已标准化, 因此设计矩阵$\boldsymbol{X}$为$n \times p$矩阵.

对$g(c)$关于c求导并令其等于零可解得c的最优值为

$$c^* = \frac{\|\boldsymbol{\beta}\|^2}{\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} + \|\boldsymbol{\beta}\|^2} < 1. \tag{3.10.1}$$

由于$g(c)$关于c的二阶导数等于

$$2\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} + 2\|\boldsymbol{\beta}\|^2 > 0,$$

因此$g(c) = \mathrm{MSE}(\hat{\boldsymbol{\beta}}_s(c))$在$c^*$处达到最小, 并且当$c^* \le c < 1$时, 有$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_s(c)) < \mathrm{MSE}(\hat{\boldsymbol{\beta}})$.