第6章　线性回归模型(一)

吴密霞

北京工业大学统计与数据科学系

E-mail:　*wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论（第2版）, 科学出版社.

一元线性回归模型

一元线性回归模型的一个例子.

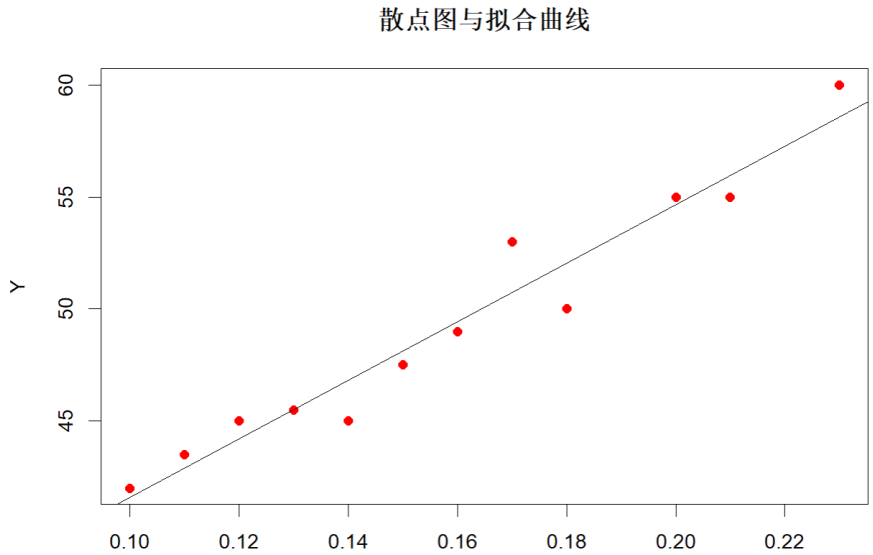
例6.1.1

由专业知识知道，合金的强度$Y(\mathrm{kg/mm^2})$与合金中碳含量$X(\%)$有关. 为了了解它们间的关系，从生产中收集了一批数据(x_i, y_i), $i = 1, 2, \cdots, n$, 具体数据见下表.

X	0.10	0.11	0.12	0.13	0.14	0.15
Y	42.0	43.5	45.0	45.5	45.0	47.5
X	0.16	0.17	0.18	0.20	0.21	0.23
Y	49.0	53.0	50.0	55.0	55.0	60.0

1 数学模型

画$X-Y$ 散点图(scatter plot)



从散点图上发现: n个点基本在一条直线附近，故可认为Y与X的关系基本上是线性的.

引例：一元线性回归模型

- 模型假定

模型假定

$$Y = \beta_0 + \beta_1 X + e,$$

其中$\beta_0 + \beta_1 X$表示Y随X的变化而线性变化的部分. e是随机误差.

- 通常假定$e \sim N(0,\ \sigma^2)$　　（σ^2 未知）
- 称函数$f(X) = \beta_0 + \beta_1 X$为一元线性回归函数
- $\beta_0,\ \beta_1$为回归系数　　（未知）
- 称X为回归自变量(或回归因子, 解释变量)
- 称Y为回归因变量(或响应变量)

2 回归参数的估计

- 设$(x_1, y_1), \ldots, (x_n, y_n)$ 是(X, Y)的一组观测值, 则

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

其中$E(e_i) = 0$, $\mathrm{var}(e_i) = \sigma^2$, $e_1, \cdots, e_n$ 不相关.

估计未知参数的一种直观想法:

要求图中的点(x_i, y_i)与直线上的点$(x_i, \hat{y}_i)$的偏离越小越好

这里

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

称其为回归值或拟合值

回归参数的估计

最小二乘法（Least Square, LS）的思想：

找β_0和β_1使得离差平方和

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

达到最小.

最小二乘估计（LS Estimator）

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$$

如何求极值？

最小二乘估计

令

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0, \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0, \end{cases}$$

经计算可得

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

其中$\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i,\ S_{xx} = \sum\limits_{i=1}^{n}(x_i - \bar{x})^2,\ S_{xy} = \sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$

最小二乘估计

- 经验回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- 方差σ^2的LS估计

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2}{n-2}$$

称$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \cdots, n$ 为残差

称

$$\text{SSE} = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

为残差平方和

最小二乘估计

- 回归系数的LS估计$\hat{\beta}_1$与X和Y的样本相关系数同号

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{yy}}{S_{xx}} = \hat{\rho}_{XY}\sqrt{\frac{S_{xy}}{S_{xx}}}$$

- $\hat{\beta}_0$, $\hat{\beta}_1$ 和$\hat{\sigma}^2$ 是β_0, β_1 和σ^2的无偏估计, 即

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1, \quad E\hat{\sigma}^2 = \sigma^2.$$

- 若$e_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, 则$\hat{\beta}_0$和$\hat{\beta}_1$也是极大似然(Maximum likelihood, ML)估计; σ^2的ML估计为$(n-2)\hat{\sigma}^2/n$. (留给学生练习)

3 回归方程的显著性检验

- 若Y与X具有较强的线性关系, 则可按照LS方法给出参数估计, 否则计算的结果无意义
- $\beta_1 = 0$ 表示$E(Y)$不随X作线性变化;
- $\beta_1 \neq 0$ 表示$E(Y)$随X作线性变化

在一元线性回归模型下, 回归方程的显著性检验等价于回归系数的显著性检验，即检验

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0.$$

回归方程的显著性检验

(1) t 检验法

当H_0成立时，统计量

$$T = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\hat{\sigma}} \quad \sim \quad t_{n-2},$$

对于给定的显著性水平α, 检验的拒绝域为

$$|T| \geq t_{n-2}(\alpha/2).$$

回归方程的显著性检验

(2) F 检验法

当H_0成立时，统计量

$$F = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} \quad \sim \quad F(1, n-2),$$

对于给定的显著性水平α, 检验的拒绝域为

$$F \geq F_\alpha(1, n-2).$$

4 预测

- 当给定$X = x_0$时，$y_0 = \beta_0 + \beta_1 x_0$的点预测:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

- 对给定$X = x_0$, 求$y_0 = \beta_0 + \beta_1 x_0$的的置信度为$1 - \alpha$的预测区间:

$$(\hat{y}_0 - l,\ \ \hat{y}_0 + l),$$

其中, $P\{\hat{y}_0 - l < y_0 < \hat{y}_0 + l\} = 1 - \alpha,$

$$l = t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}.$$

一元线性回归模型: 预测

- 在实际问题中，当样本容量n很大时，则对于在$\bar{x}$ 附近 的x_0, 有

$$\sqrt{1+\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{S_{xx}}} \approx 1, \quad t_{\alpha/2}(n-2) \approx Z_{\alpha/2}.$$

于是，此时y_0的置信度为$1-\alpha$的预测区间可近似地简化为

$$(\hat{y}_0 - \hat{\sigma} Z_{\alpha/2},\ \hat{y}_0 + \hat{\sigma} Z_{\alpha/2}).$$

例

求例6.1.1 的回归方程，并对相应的方程作检验. 并求$X = 0.16$ 时相应Y预测值和置信系数为0.95的预测区间.

解 利用R软件中的lm() 求出回归参数$\hat{\beta}_0, \hat{\beta}_1$ 和作相应的检验.

```
> x<-c(0.10,0.11,0.12,0.13,0.14,0.15,0.16,0.17,0.18,0.20,0.21,0.23)
> y<-c(42.0,43.5,45.0,45.5,45.0,47.5,49.0,53.0,50.0,55.0,55.0,60.0)
> lm.sol<-lm(y~1+x)
> summary(lm.sol)

Call:
lm(formula = y ~ 1 + x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0431 -0.7056  0.1694  0.6633  2.2653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.493      1.580   18.04 5.88e-09 ***
x            130.835      9.683   13.51 9.50e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.319 on 10 degrees of freedom
Multiple R-squared:  0.9481,    Adjusted R-squared:  0.9429
F-statistic: 182.6 on 1 and 10 DF,  p-value: 9.505e-08
```

一元线性回归模型

```
> anova(lm.sol)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 317.82  317.82  182.55 9.505e-08 ***
Residuals 10  17.41    1.74
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> new<-data.frame(x=0.16)
> lm.pred<-predict(lm.sol,new,interval="prediction")
> lm.pred
       fit      lwr      upr
1 49.42639 46.36621 52.48657
```

一般线性回归模型

假设Y为因变量和$X_1, \cdots, X_{p-1}$为自变量,

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

其中β_0为截距项, $\beta_1, \ldots, \beta_{p-1}$为回归系数, e为随机误差.

n次独立观察数据: $(y_i, x_{i1}, \cdots, x_{ip-1}), \quad i = 1, \cdots, n,$ 有

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

记

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

线性回归模型的估计

矩阵形式：

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I_n}, \tag{6.1}$$

- $\mathbf{X} = (\mathbf{1}_n, \tilde{\mathbf{X}})$, $\tilde{\mathbf{X}}$ 为$p-1$个自变量的n次观测矩阵.
- 假设$e_1, \cdots, e_n$互不相关, 均值皆为零, 方差皆为σ^2.
- 假设$\mathbf{X}$ 列满秩, 即$\mathrm{rk}(\mathbf{X}) = p \Longrightarrow \boldsymbol{\beta}$ 可估.

考虑$\boldsymbol{\beta}$ 和σ^2 最小二乘(LS)估计.

LS估计

极小化离差平方和

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

令

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

解得$\boldsymbol{\beta}$的LS估计:

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

从而得σ^2的LS估计

$$\widehat{\sigma}^2 = \frac{\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p} = \frac{\boldsymbol{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{y}}{n-p},$$

其中$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, 并称$\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$为残差平方和, 记作SSE.

LS估计的性质

定理6.1.1

在线性回归模型(6.1) 下, β和σ^2的LS估计满足如下性质:

(1) 无偏性: $E(\widehat{\beta}) = \beta$.

(2) 方差最小性: 对任意$p \times 1$ 向量$\boldsymbol{c}$, $\boldsymbol{c}'\widehat{\beta}$为$\boldsymbol{c}'\beta$的唯一最佳线性无偏 (best linear unbiased, BLU) 估计.

(3) $\widehat{\sigma}^2 = \|\boldsymbol{y} - \mathbf{X}\widehat{\beta}\|^2/(n-p)$为$\sigma^2$的无偏估计.

- 因为假设$\mathbf{X}$ 列满秩（$\mathrm{rk}\mathbf{X} = p$），$\mathbf{X}'\mathbf{X}$可逆，(1)立证. 由定理4.1.2和定理4.1.3立证得(2)和(3). (详细证明回归见下面)

知识回顾*

- 设$\boldsymbol{y} = (y_1, \cdots, y_n)'$为$n \times 1$ 随机向量, 则随机向量$\boldsymbol{y}$的均值向量为

$$E(\boldsymbol{y}) = (E(y_1), \cdots, E(y_n))'.$$

- 设$\boldsymbol{z} = (z_1, \cdots, z_m)'$和$\boldsymbol{y} = (y_1, \cdots, y_n)'$为随机向量, 则

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{y}) &= E[(\boldsymbol{z} - E(\boldsymbol{z}))(\boldsymbol{y} - E(\boldsymbol{y}))'] \\ &= \begin{pmatrix} \mathrm{Cov}(z_1, y_1) & \mathrm{Cov}(z_1, y_2) & \cdots & \mathrm{Cov}(z_1, y_n) \\ \mathrm{Cov}(z_2, y_1) & \mathrm{Cov}(z_2, y_2) & \cdots & \mathrm{Cov}(z_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(z_m, y_1) & \mathrm{Cov}(z_m, y_2) & \cdots & \mathrm{Cov}(z_m, y_n) \end{pmatrix}. \end{aligned}$$

知识回顾*

- 设$\mathbf{A}$为$m \times n$非随机矩阵, $\boldsymbol{y}$和$\boldsymbol{b}$分别为$n \times 1$和$m \times 1$随机向量, 则

$$E(\mathbf{A}\boldsymbol{y} + \boldsymbol{b}) = \mathbf{A}E(\boldsymbol{y}) + E(\boldsymbol{b}).$$

- 设$\mathbf{A}$为$m \times n$阵, $\boldsymbol{y}$为$n \times 1$随机向量, 则

$$\mathrm{Cov}(\mathbf{A}\boldsymbol{y}) = \mathbf{A}\mathrm{Cov}(\boldsymbol{y})\mathbf{A}'.$$

- $\boldsymbol{y}$为$n \times 1$ 随机向量, 则

$$E(\boldsymbol{y}'\mathbf{B}\boldsymbol{y}) = E(\boldsymbol{y})'\mathbf{C}E(\boldsymbol{y}) + \mathrm{tr}(\mathrm{Cov}(\boldsymbol{y})\mathbf{B}).$$

定理6.1.1的证明

证明 由随机向量线性变换后向量的均值和协方差的性质, 立得

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

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Cov}(\boldsymbol{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

由随机向量二次型的期望公式，立得

$$\begin{aligned} E(\widehat{\sigma}^2) &= \frac{1}{n-p}E(\boldsymbol{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{y}) \\ &= \frac{1}{n-p}\left((\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})(\mathbf{X}\boldsymbol{\beta}) + \sigma^2\mathrm{tr}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\right) \\ &= \frac{\sigma^2(n - \mathrm{rk}(\mathbf{X}))}{n-p} = \sigma^2. \end{aligned}$$

定理6.1.1的证明

- 证明$\boldsymbol{c}'\boldsymbol{\beta}$的线性无偏估计.

  由无偏性得证: 对一切$\boldsymbol{\beta} \in \mathbb{R}^p$，都有

$$E(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{c}'\boldsymbol{\beta} \Longleftrightarrow \mathbf{X}'\boldsymbol{a} = \boldsymbol{c}.$$

  于是

$$\begin{aligned} \mathrm{Var}(\boldsymbol{a}'\boldsymbol{y}) - \mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{a}'\boldsymbol{a} - \boldsymbol{c}'(\mathbf{X}'\mathbf{X})^{-1})\boldsymbol{c}) \\ &= \sigma^2\boldsymbol{a}'(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{a} \\ &\geq 0. \end{aligned}$$

  最后不等式应用了对称幂等阵一定是半正定阵，因为它的非零特征根都是1，大于零

正态假设下LS估计的性质

定理6.1.2

假设线性回归模型(6.1) 中 $\boldsymbol{e} \sim N_n(\mathbf{0},\ \sigma^2\mathbf{I_n})$, 则

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

(2) $\boldsymbol{c}'\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{c}'\boldsymbol{\beta}, \sigma^2\boldsymbol{c}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{c})$, $\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$ 是$\boldsymbol{c}'\boldsymbol{\beta}$的唯一最小方差无偏(minimum variance unbiased, MVU)估计.

(3) $(n-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$, 且与$\widehat{\boldsymbol{\beta}}$相互独立.

- 特别地,

$$\widehat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}), \quad i = 0, 1, \cdots, p-1,$$

这里c_{ii} 为$(\mathbf{X}'\mathbf{X})^{-1}$的第$(i+1)$个对角元.

知识回顾

正态随机向量

设n维随机向量$\boldsymbol{y} = (y_1, \cdots, y_n)'$具有密度函数

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

其中$\boldsymbol{y} = (y_1, \cdots, y_n)'$, $-\infty < y_i < +\infty$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)'$, $\boldsymbol{\Sigma} > \mathbf{0}$, 则称$\boldsymbol{y}$为$n$维正态随机向量, 记为$N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

χ^2 分布

设$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \mathbf{I}_n)$. 随机变量$z = \boldsymbol{x}'\boldsymbol{x}$的分布称为自由度为$n$, 非中心参数为$\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}$的$\chi^2$分布, 记为$z \sim \chi^2_{n,\lambda}$. 当$\lambda = 0$时, 称$z$的分布为中心$\chi^2$分布, 记为$z \sim \chi^2_n$.

知识回顾

- 设$\boldsymbol{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \geq \mathbf{0}$, $\mathbf{B}$为$m \times n$ 任意实矩阵, 则

$$\mathbf{B}\boldsymbol{y} \sim N_m(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}').$$

- 设$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \mathbf{I}_n)$, $\mathbf{A}$对称, 则

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} \sim \chi^2_{r,\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}} \Longleftrightarrow \mathbf{A} \text{ 幂等}, \mathrm{rk}(\mathbf{A}) = r.$$

- 设$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \mathbf{I})$, $\mathbf{A}$为$n \times n$对称阵, $\mathbf{C}$为$m \times n$矩阵. 若

$$\mathbf{C}\mathbf{A} = \mathbf{0},$$

则$\mathbf{C}\boldsymbol{x}$和$\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$相互独立.

定理6.1.2的证明

证明 (1) $\boldsymbol{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ 是$\boldsymbol{y}$的线性组合, 故

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

(2) 因为$\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$ 是$\widehat{\boldsymbol{\beta}}$的线性组合, 故

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

其MVU性质由定理4.1.5可得.

(3) 注意到 $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ 为对称幂等阵, $\boldsymbol{e}/\sigma \sim N(\mathbf{0}, \mathbf{I}_n)$, 故

$$\boldsymbol{y}(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\boldsymbol{y} = \boldsymbol{e}(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\boldsymbol{e} \sim \sigma^2\chi^2_{n-p} \Longrightarrow (n-p)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$$

由于$\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_\mathbf{X}) = \mathbf{0}$, 得证

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} \text{ 与 } \widehat{\sigma}^2 = \boldsymbol{y}(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\boldsymbol{y}/(n-p) \text{独立}$$

中心化后的LS估计

- 在回归分析中, 主要兴趣回归系数$\boldsymbol{\beta}_I = (\beta_1, \beta_2, \cdots, \beta_{p-1})'$
- 线性回归模型的中心化形式：

$$\boldsymbol{y} = \mathbf{1}\alpha_0 + \tilde{\mathbf{X}}_c\boldsymbol{\beta}_I + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I}_n, \tag{6.2}$$

其中$\tilde{\mathbf{X}}_c$为中心化的设计矩阵，

$$\tilde{\mathbf{X}}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1(p-1)} - \bar{x}_{p-1} \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2(p-1)} - \bar{x}_{p-1} \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{n(p-1)} - \bar{x}_{p-1} \end{pmatrix}$$

中心化后的LS估计

由于$\tilde{\mathbf{X}}_c'\mathbf{1}_n = \mathbf{0}$, 故$(\alpha_0, \boldsymbol{\beta}_I)$的LS估计为

$$\widehat{\alpha}_0 = \bar{y}, \quad \widehat{\boldsymbol{\beta}}_I = (\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1}\tilde{\mathbf{X}}_c'\boldsymbol{y},$$

$$\text{Cov}\begin{pmatrix} \widehat{\alpha}_0 \\ \widehat{\boldsymbol{\beta}}_I \end{pmatrix} = \sigma^2 \begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1} \end{pmatrix}.$$

这个事实说明: 在中心化线性回归模型中,

- 常数项α_0 总是用因变量观测值的算术平均值$\bar{y}$来估计
- 回归系数$\boldsymbol{\beta}_I$的LS估计与$\boldsymbol{y} = \tilde{\mathbf{X}}_c\boldsymbol{\beta}_I + \boldsymbol{e}$ 的LS估计相同
- 这两个估计总是不相关的

数据标准化

数据标准化的模型

$$\boldsymbol{y} = \mathbf{1}\alpha_0 + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I}_n,$$

其中$\mathbf{Z} = (z_{ij})$, $z_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j}$, $i = 1, \cdots, n$, $j = 1, \cdots, p-1$,

$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, 为第$j$个自变量$X_j$的样本方差.

(1) $\mathbf{1}_n'\mathbf{Z} = \mathbf{0}$;

(2) $\mathbf{Z}'\mathbf{Z} = (r_{ij})$ 为$\boldsymbol{X} = (X_1, \cdots, X_{p-1})'$的样本相关系数矩阵$\mathbf{R}_\mathbf{X}$,

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

LS估计

- 非中心化: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_{p-1} X_{p-1}$;
- 中 心 化: $\widehat{Y} = \widehat{\alpha}_0 + \widehat{\beta}_1 (X_1 - \bar{x}_1) + \cdots + \widehat{\beta}_{p-1} (X_{p-1} - \bar{x}_{p-1})$;
- 中心化标准化: $\widehat{Y} = \widehat{\alpha}_0 + \widehat{\alpha}_1 \frac{(X_1 - \bar{x}_1)}{s_1} + \cdots + \widehat{\alpha}_{p-1} \frac{(X_{p-1} - \bar{x}_{p-1})}{s_{p-1}}$.

中心化和中心化标准化后所得的经验回归方程都与原数据下所得的经验回归方程等价.

$$\widehat{\alpha}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \cdots + \widehat{\beta}_{p-1} \bar{x}_{p-1},$$

$$\widehat{\alpha}_i = \widehat{\beta}_i \cdot s_i, \quad i = 1, \cdots, p-1.$$

数据标准化

标准化的好处:

- 用$\mathbf{R_X}$可以分析回归自变量之间的相关关系;
- 在一些问题中, 诸回归自变量所用的单位可能不相同, 取值范围大小也不同, 经过标准化消去了单位和取值范围的差异, 这便于对回归系数的估计值的统计分析

- 中心化的目的是将回归系数和截距项的估计分离开, 简化运算和相应估计的统计性质(两部分估计不相关);
- 标准化的主要目的是消除量纲影响, 使得标准化后模型的回归系数可直接反映各自变量对因变量预测的重要程度.

回归方程和系数的检验

正态线性回归模型

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i(p-1)}\beta_{p-1} + e_i, \quad e_i \sim N(0,\ \sigma^2), \quad i = 1, \cdots, n, \tag{6.3}$$

- 回归方程的显著性检验(或称回归方程的拟合优度)

$$H_0: \quad \beta_1 = \cdots = \beta_{p-1} = 0 \longleftrightarrow H_1: \quad \beta_1, \cdots, \beta_{p-1}\text{不全为 0.}$$

  - 如果$H_0: \quad \beta_1 = \cdots = \beta_{p-1} = 0$被拒绝, 则可认为$Y$线性依赖于自变量$X_1, \cdots, X_{p-1}$中的至少一个自变量.
  - 如果H_0被接受, 则可认为所有$\beta_i = 0$, 即相对于误差而言, 所有自变量对因变量Y的影响是不重要的.

回归方程和系数的检验

为简单记，采用模型的中心化的矩阵形式:

$$\boldsymbol{y} = \mathbf{1}\alpha_0 + \tilde{\mathbf{X}}_c\boldsymbol{\beta}_I + \boldsymbol{e}, \quad \boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

这里$\boldsymbol{\beta}_I = (\beta_1, \beta_2, \cdots, \beta_{p-1})'$. 原假设$H_0$ 等价于$\boldsymbol{\beta}_I = \mathbf{0}$.
在原模型(6.3)下, α_0 和$\boldsymbol{\beta}_I$ 的LS 估计为

$$\alpha_0 = \bar{y}, \quad \widehat{\boldsymbol{\beta}}_I = (\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1}\tilde{\mathbf{X}}_c'\boldsymbol{y}.$$

将H_0代入原模型，得约简模型:

$$y_i = \beta_0 + e_i, \quad e_i \sim N(0,\ \sigma^2), \quad i = 1, \cdots, n.$$

该模型下, β_0 的LS估计为$\bar{y}$.

回归方程的显著性检验

约简模型的残差平方和就等于总平方和:

$$\mathrm{SSE}_{H_0} = \mathrm{SST} = \|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

原模型(6.3)的残差平方和:

$$\mathrm{SSE} = \|\boldsymbol{y} - \hat{\alpha}_0\mathbf{1}_n - \tilde{\mathbf{X}}_c\hat{\boldsymbol{\beta}}_I\|^2 = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2 - \hat{\boldsymbol{\beta}}_I'\tilde{\mathbf{X}}_c'\boldsymbol{y} = \mathrm{SST} - \mathrm{SSR}.$$

原模型(6.3)的回归平方和:

$$\mathrm{SSR} = \mathrm{SSE}_{H_0} - \mathrm{SSE} = \hat{\boldsymbol{\beta}}_I'\tilde{\mathbf{X}}_c'\boldsymbol{y} = \|\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_I\|^2.$$

当H_0成立时, 两个残差平方和SSE和SSE_{H_0}相对而言应该很接近.

回归方程的显著性检验

由正态向量二次型的理论，易证

- $\mathrm{SSE} = \boldsymbol{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{y} = \boldsymbol{y}'(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n - \mathbf{P}_{\tilde{\mathbf{X}}_c})\boldsymbol{y} \sim \sigma^2\chi^2_{n-p}$,
- SSR和SSE独立,
- H_0成立时, $\mathrm{SSR} = \widehat{\boldsymbol{\beta}}_I'\tilde{\mathbf{X}}_c'\boldsymbol{y} = \boldsymbol{y}'\mathbf{P}_{\tilde{\mathbf{X}}_c}\boldsymbol{y} \sim \sigma^2\chi^2_{p-1}$.

F检验统计量

$$F = \frac{\mathrm{SSR}/(p-1)}{\mathrm{SSE}/(n-p)} = \frac{(\widehat{\boldsymbol{\beta}}_I'\tilde{\mathbf{X}}_c'\boldsymbol{y})/(p-1)}{\mathrm{SSE}/(n-p)}. \tag{6.4}$$

当原假设H_0成立时, $F \sim F_{p-1,\,n-p}$. 给定的水平α, 若$F > F_{p-1,\,n-p}(\alpha)$, 则拒绝原假设$H_0$, 认为回归方程显著, 否则接受$H_0$, 认为回归方程不显著.

回归方程的显著性检验

总平方和SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ 可分解为回归平方和SSR 与残差平方和SSE 两部分, 即

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE},$$

SSR 反映了协变量X对响应变量Y变动平方和的贡献

SSE 反映了随机误差的变动对总平方和的贡献

检验统计量F 就是把SSR 与SSE 进行比较

当回归平方和SSR相对残差平方和SSE比较大时, 就拒绝原假设, 认为回归直线与样本观测值的拟合效果是显著的.

回归方程的显著性检验

对线性回归的显著性检验（或拟合优度检验）使用下面的方差分析表进行解释.

Table: 方差分析表

方差来源	平方和	自由度	均方	F比值
回归	SSR	$p-1$	$\mathrm{MSR}=\mathrm{SSR}/(p-1)$	$F=\dfrac{\mathrm{MSR}}{\mathrm{MSE}}$
误差	SSE	$n-p$	$\mathrm{MSE}=\mathrm{SSE}/(n-p)$	
总和	SST	$n-1$		

回归系数的显著性检验

注 回归方程的显著性检验是对线性回归方程的一个整体性检验.

注 如果检验结果是拒绝原假设, 则这意味着因变量Y线性地依赖于回归自变量$X_1, \cdots, X_{p-1}$整体. 但并不排除某些β_i可能等于零.

因此，在拒绝$H_0: \beta_1 = \cdots = \beta_{p-1} = 0$后, 还需要对每个自变量逐一做显著性检验, 即对固定的i, 考虑

- 回归系数的显著性检验

$$H_{i0}: \quad \beta_i = 0 \longleftrightarrow H_{i1}: \quad \beta_i \neq 0. \tag{6.5}$$

回归系数的显著性检验

根据定理6.1.2知：

- $\widehat{\beta}_i \sim N(\beta_i,\ \sigma^2 c_{ii}), \quad i = 1, \cdots, p-1$, 这里$c_{ii}$ 为$(\mathbf{X}'\mathbf{X})^{-1}$ 的第$i+1$个对角元(即$(\mathbf{X}_c'\mathbf{X}_c)^{-1}$的第$i$个对角元).
- σ^2的LS估计$\widehat{\sigma}^2 = \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p) \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$.
- $\widehat{\sigma}^2$与$\widehat{\beta}_i$相互独立.

当原假设H_{i0}成立时, 有

$$t_i = \frac{\widehat{\beta}_i}{\widehat{\sigma}\sqrt{c_{ii}}} \sim t_{n-p}, \tag{6.6}$$

这里t_{n-p}表示自由度为$n-p$的t分布.

回归系数的显著性检验

- 回归系数的显著性检验

$$H_{i0}: \quad \beta_i = 0 \longleftrightarrow H_{i1}: \quad \beta_i \neq 0,$$

检验统计量为

$$t_i = \frac{\widehat{\beta}_i}{\widehat{\sigma}\sqrt{c_{ii}}}.$$

对给定的检验水平α, 当

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

或P值

$$\mathrm{p}_i = P(t_{n-p} \geq |t_i|) < \alpha/2,$$

则拒绝原假设H_{i0}，认为$\beta_i \neq 0$. 否则就接受H_{i0}.

显著性检验的案例

煤净化问题案例(Mallows,1964)

Y为净化后煤溶液中所含杂质的重量, X_1表示输入净化过程的溶液所含的煤与杂质的比; X_2是溶液的pH值; X_3表示溶液流量. 试验目的: 通过一组试验数据, 建立净化效率Y与因素X_1, X_2和X_3的经验关系, 据此通过控制某些自变量来提高净化效率. 煤净化数据如下：

编号	X_1	X_2	X_3	Y	编号	X_1	X_2	X_3	Y
1	1.50	6.00	1315	243	7	2.00	7.50	1575	183
2	1.50	6.00	1315	261	8	2.00	7.50	1575	207
3	1.50	9.00	1890	244	9	2.50	9.00	1315	216
4	1.50	9.00	1890	285	10	2.50	9.00	1315	160
5	2.00	7.50	1575	202	11	2.50	6.00	1890	104
6	2.00	7.50	1575	180	12	2.50	6.00	1890	110

显著性检验案例

采用线性回归模型:

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

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ 的LS估计:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = (397.087, -110.750, 15.583, -0.058)'$$

- σ^2的LS估计: $\widehat{\sigma}^2 = \mathrm{SSE}/(n-p) = 435.86.$

- $$(\mathbf{X}'\mathbf{X})^{-1} = 10^{-4}\begin{pmatrix} 90359.20696 & -10000 & -4166.6667 & -0024.02251 \\ -10000.0000 & 5000 & 0.0000 & 0.000 \\ -4166.6667 & 0.0000 & 0555.5556 & 0.0000 \\ -0024.0225 & 0.0 & 0.0000 & 0000.0151 \end{pmatrix}$$

显著性检验案例

- 考虑线性回归方程的显著性检验, 即检验假设

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0 \longleftrightarrow H_1: \quad \beta_1, \beta_2, \beta_3 \text{ 不全为0.}$$

计算得$\bar{y} = 199.5833$. 进一步可得各平方和与响应的自由度:

$$\text{SST} = \sum_{i=1}^{12} (y_i - 199.5833)^2 = 31156.02, \quad f_T = 12 - 1 = 11,$$

$$\text{SSE} = (\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 3486.89, \quad f_E = 12 - 4 = 8,$$

$$\text{SSR} = \text{SST} - \text{SSE} = 34642.91 - 3486.89 = 31156.02, \quad f_R = 3.$$

取显著性水平$\alpha = 0.05$. 由

$$F = \frac{\text{SSR}/f_R}{\text{SSE}/f_E} = \frac{10385.33}{435.86} = 23.83 > F_{3,\,8}(0.05) = 4.07$$

显著性检验案例

或

$$P(F_{3,\ 8} > 23.83) = 0.00024 < \alpha = 0.05,$$

故拒绝H_0, 认为回归方程显著.

Table: 方差分析表

方差来源	平方和	自由度	均方	F 比值	p值
回归	31156.02	3	10385.33	23.83	0.00024
误差	3486.89	8	435.86		
总和	34642.91	11			

显著性检验案例

- 考虑回归系数的显著性检验, 即检验假设

$$H_0: \quad \beta_i = 0 \longleftrightarrow H_1: \quad \beta_i \neq 0.$$

$\widehat{\beta}_1$, $\widehat{\beta}_2$和$\widehat{\beta}_3$的标准差的估计为

$$s(\widehat{\beta}_1) = \sqrt{\widehat{\sigma}^2 c_{11}} = \sqrt{435.86 \times 0.49998} = \sqrt{217.9230} = 14.7625,$$

$$s(\widehat{\beta}_2) = \sqrt{\widehat{\sigma}^2 c_{22}} = \sqrt{435.86 \times 0.05556} = \sqrt{24.21444} = 4.9208,$$

$$s(\widehat{\beta}_3) = \sqrt{\widehat{\sigma}^2 c_{33}} = \sqrt{435.86 \times 0.0000011} = \sqrt{0.00066} = 0.0256.$$

显著性检验案例

三个回归系数显著性检验对应的统计量t_i 的值分别为

$$t^{(1)} = \frac{\widehat{\beta}_1}{s(\widehat{\beta}_1)} = \frac{-110.750}{14.7625} = -7.502, \qquad \text{p值} = 2P(t_8 > |t^{(1)}|) = 0.0000691$$

$$t^{(2)} = \frac{\widehat{\beta}_2}{s(\widehat{\beta}_2)} = \frac{15.583}{4.9208} = 3.167, \qquad \text{p值} = 2P(t_8 > |t^{(2)}|) = 0.013258$$

$$t^{(3)} = \frac{\widehat{\beta}_3}{s(\widehat{\beta}_3)} = \frac{-0.058}{0.0256} = -2.274, \qquad \text{p值} = 2P(t_8 > |t^{(3)}|) = 0.052565$$

对给定的水平$\alpha = 0.05$, 查表得$t_8(0.025) = 2.3060$,

$$|t^{(1)}| > t_8(0.025), \quad |t^{(2)}| > t_8(0.025), \quad |t^{(3)}| < t_8(0.025),$$

故拒绝假设$\beta_1 = 0$和$\beta_2 = 0$, 但不能拒绝假设$\beta_3 \neq 0$.

显著性检验案例

删除变量X_3，用R语言中的函数lm()计算结果：

```
y<-c(243, 261, 244, 285, 202, 180, 183, 207, 216, 160, 104, 110)
X1<-c(1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.0, 2.0, 2.5, 2.5, 2.5, 2.5 )
X2<-c(6.0, 6.0, 9.0, 9.0, 7.5, 7.5, 7.5, 7.5, 9.0, 9.0, 6.0, 6.0)
X3<-c(1315, 1315, 1890, 1890, 1575, 1575, 1575, 1575, 1315, 1315, 1890, 1890)
reg<-lm(Y ~ X1+X2 )
summary(reg)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  304.208     57.638   5.278 0.000509 ***
X1          -110.750     17.858  -6.202 0.000159 ***
X2            15.583      5.953   2.618 0.027911 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 

Residual standard error: 25.26 on 9 degrees of freedom
Multiple R-squared:  0.8343,    Adjusted R-squared:  0.7975
F-statistic: 22.66 on 2 and 9 DF,  p-value: 0.0003069
```

显著性检验案例

- 删除变量X_3后，重新计算得经验回归方程为

$$Y = 304.208 - 110.750X_1 + 15.583X_2.$$

- R语言的运行结果表明：F-statistic =22.66, P值=0.0003069. 因此, 在水平$\alpha = 0.05$下，该模型的回归方程显著；
- 回归系数β_1和β_2显著, 因为对应的t 检验的P-value=0.000159, 0.027911. 据此可通过控制自变量X_1和X_2来提高净化效率.

思考：

当多个回归系数被检验为不显著时，是否可以全部删除？

复相关系数

定义: 复相关系数(multiple correlation coefficient)

用来度量随机变量与随机向量相关程度的量. 记Y和$\boldsymbol{X} = (X_1, \cdots, X_{p-1})'$分别为随机变量和$(p-1)\times 1$的随机向量, 且

$$\mathrm{Cov}\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_{XY}' \\ \boldsymbol{\sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} \triangleq \boldsymbol{\Sigma}_0.$$

则称

$$\rho = \left(\boldsymbol{\sigma}_{XY}' \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} \boldsymbol{\sigma}_{XY}\right)^{1/2} \Big/ \sigma_Y$$

为Y和$\boldsymbol{X}$的复相关系数.

复相关系数

- 复相关系数ρ刻画了随机变量Y与随机向量$\boldsymbol{X}$间的线性相关程度, 即Y与$\boldsymbol{X}$的线性组合$\boldsymbol{a}'\boldsymbol{X}$的最大相关系数. 由Cauchy-Schwarz不等式, 立证

$$\rho = \max_{\boldsymbol{a}} \rho_{Y,\boldsymbol{a}'\boldsymbol{X}} = \max_{\boldsymbol{a}} \frac{\boldsymbol{a}'\sigma_{\boldsymbol{X}Y}}{\sigma_Y\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{a}}} = \left(\boldsymbol{\sigma}_{\boldsymbol{X}Y}'\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{X}Y}\right)^{1/2} \Big/ \sigma_Y,$$

其中$\boldsymbol{a}$为$(p-1)$实数空间上的任意一非零向量. $0 \le \rho \le 1$.

Cauchy-Schwarz不等式

设$\boldsymbol{a}$和$\boldsymbol{b}$是两个维数相同的实数向量, 则$(\boldsymbol{a}'\boldsymbol{b})^2 \le (\boldsymbol{a}'\boldsymbol{a})(\boldsymbol{b}'\boldsymbol{b})$, 其等号成立当且仅当存在非零常数$c$使得$\boldsymbol{a} = c\boldsymbol{b}$.

样本复相关系数

- 样本复相关系数

$$R = \left(\frac{\widehat{\boldsymbol{\beta}}_I' \tilde{\mathbf{X}}_c' \boldsymbol{y}}{\sum_i (y_i - \bar{y})^2} \right)^{1/2},$$

其中,

$$\widehat{\sigma}_Y^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2, \qquad \text{其中 } \bar{y} = \frac{1}{n} \sum_i y_i,$$

$$\widehat{\boldsymbol{\Sigma}}_{XX} = \frac{1}{n} \tilde{\mathbf{X}}_c' \tilde{\mathbf{X}}_c, \qquad \widehat{\boldsymbol{\sigma}}_{XY} = \frac{1}{n} \tilde{\mathbf{X}}_c' (\boldsymbol{y} - \bar{y}\mathbf{1}) = \frac{1}{n} \tilde{\mathbf{X}}_c' \boldsymbol{y}.$$

- 复相关系数的平方(Multple R-Squares)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{6.7}$$

复相关系数

注1 若$R^2 = 1$, 则SSR $=$ SST, 因变量Y与自变量$X_1, \cdots, X_{p-1}$之间有严格的线性关系. 若$R^2 = 0$, 则SSR $= 0$, Y与$X_1, \cdots, X_{p-1}$之间无任何线性关系.

注2 通常$0 < R^2 < 1$. R^2越大, 表明Y与自变量$X_1, \cdots, X_{p-1}$ 之间的线性关系程度越强.

复相关系数的平方R^2也是度量回归方程对观测值

$$\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$$

拟合程度的好坏或者回归方程解释能力的大小的一个重要指标, 文献中也称其为 判定系数

调整的复相关系数

注意到

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

其中残差平方和SSE 会随着模型中引入的自变量的增加而减少，因此，将其中SSE和SST比换成它们的均方比, 便得到

调整的复相关系数的平方, Adjusted R-Squares

$$R_A^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2). \tag{6.8}$$

回归方程的显著性检验和回归系数的显著性检验，以及判定系数R^2 都可用R语言的函数lm()计算，函数summary()提取信息.

回归常见几类问题

- 多个回归系数不显著, 如何处理?
- 回归方程显著, 但所有的回归系数都不显著, 怎么解释?
- 如果几组数据导出的回归系数的估计、检验、回归方程的显著性检验统计量, 以及判定系数都近似相等，是否能说明这几组数据可用同一个模型作预测?

分别用三个例子来展示问题.

例6.2.2 预测人体吸入氧气的效率问题

31名中年男性7项指标:吸氧效率(Y)、年龄(X_1)、体重(X_2)、跑1.5千米所需时间(X_3)、休息时心率(X_4)、跑步时心率(X_5) 和最高心率(X_6).

编号	Y	X_1	X_2	X_3	X_4	X_5	X_6	编号	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	44.609	44	89.47	11.37	62	178	182	17	40.836	51	69.63	10.95	57	168	172
2	45.313	40	75.05	10.07	62	185	185	18	46.672	51	77.91	10.00	48	162	168
3	54.297	44	85.84	8.65	45	156	168	19	46.774	48	91.63	10.25	48	162	164
4	59.571	42	68.15	8.17	40	166	172	20	50.388	49	73.37	10.08	67	168	168
5	49.874	38	89.02	9.22	55	178	180	21	39.407	57	73.37	12.63	58	174	176
6	44.811	47	77.45	11.63	58	176	176	22	46.080	54	79.38	11.17	62	156	165
7	45.681	40	75.98	11.95	70	176	180	23	45.441	56	76.32	9.63	48	164	166
8	49.091	43	81.19	10.85	64	162	170	24	54.625	50	70.87	8.92	48	146	155
9	39.442	44	81.42	13.08	63	174	176	25	45.118	51	67.25	11.08	48	172	172
10	60.055	38	81.87	8.63	48	170	186	26	39.203	54	91.63	12.88	44	168	172
11	50.541	44	73.03	10.13	45	168	168	27	45.790	51	73.71	10.47	59	186	188
12	37.388	45	87.66	14.03	56	186	192	28	50.545	57	59.08	9.93	49	148	155
13	44.754	45	66.45	11.12	51	176	176	29	48.673	49	76.32	9.40	56	186	188
14	47.273	47	79.15	10.60	47	162	164	30	47.920	48	61.24	11.50	52	170	176
15	51.855	54	83.12	10.33	50	166	170	31	47.467	52	82.78	10.50	53	170	172
16	49.156	49	81.42	8.95	44	180	185								

```
health.data = read.table("health.txt", header = TRUE)
lm.reg = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = health.data)
summary(reg)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = health.data)
Residuals:
    Min      1Q  Median      3Q     Max
-5.3904 -0.9853  0.0743  1.0220  5.4072
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.86282   12.12765   8.647 7.76e-09 ***
X1           -0.24072    0.09460  -2.545  0.01779 *
X2           -0.07452    0.05328  -1.399  0.17468
X3           -2.62443    0.37251  -7.045 2.77e-07 ***
X4           -0.02532    0.06467  -0.391  0.69889
X5           -0.35992    0.11757  -3.061  0.00536 **
X6            0.28766    0.13438   2.141  0.04267 *
---
Signif.codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.267 on 24 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8189
F-statistic: 23.62 on 6 and 24 DF,  p-value: 5.823e-09
```

能否同时删除变量X_2和X_4？

例6.2.3 Hald 水泥问题

考察含如下四种化学成分:

X_1： $3CaO\cdot Al_2O_3$的含量(%), X_2： $3CaO\cdot SiO_2$的含量(%),

X_3： $4CaO\cdot Al_2O_3\cdot Fe_2O_3$的含量(%), X_4： $2CaO\cdot SiO_2$的含量(%)

的某种水泥, 每一克所释放的热量Y与这四种成分含量之间的关系.

序号	X_1	X_2	X_3	X_4	Y	序号	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5	8	1	31	22	44	72.5
2	1	29	15	52	74.3	9	2	54	18	22	93.1
3	11	56	8	20	104.3	10	21	47	4	26	115.9
4	11	31	8	47	87.6	11	1	40	23	34	83.8
5	7	52	6	33	95.9	12	11	66	9	12	113.3
6	11	55	9	22	109.2	13	10	68	8	12	109.4
7	3	71	17	6	102.7						

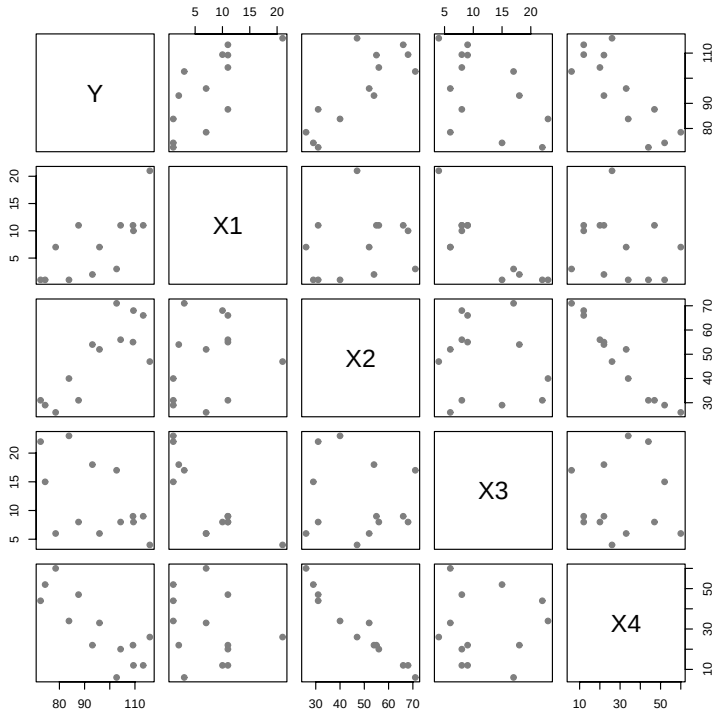
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
X1            1.5511     0.7448   2.083   0.0708 .
X2            0.5102     0.7238   0.705   0.5009
X3            0.1019     0.7547   0.135   0.8959
X4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

运行的结果显示:

- 回归方程显著（F检验的p 值=$4.756e-07 < 0.05$）；
- Y与(X_1, X_2, X_3, X_4)间存在强的线性关系（$R_A^2 = 0.9736$）；
- 4个回归系数都不能排除为0假设（每个回归系数的t检验的p值都大于显著性水平$\alpha = 0.05$）. 这与回归方程显著是否矛盾?

回归方程显著而回归系数都不显著的原因探究：



- 回归方程显著的原因: Y和每个变量的相关性都比较强, 尤其与X_4和X_2;
- 每个回归系数都被检验不显著的原因: 自变量X_1与X_3以及X_2与X_4高度相关.

回归方程显著而回归系数都不显著的原因

- 协变量与因变量本身就不线性相关；
- 自变量之间存在较强的线性相关(即复共线问题).

解决办法

针对第一种情形，可借助于变量选择方法来确定最终的模型的回归自变量(见6.3节)；

针对第二种情形, 除了可以变量选择方法外, 还可通过岭回归、主成分回归等方法来改善模型的估计(见6.6节);

当两种情况都存在，带惩罚函数的变量选择方法通常更有效(6.3节).

Anscombe数据

Anscombe(1973)构造了两个变量Y和X的四组数据，为方便表述，记(Y_j, X_j)为第j组数据对应的因变量和自变量，每组数据有11对观测值，其中前三组数据自变量的值相同

编号	1组		2组		3组		4组	
	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
3	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.5
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.44	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

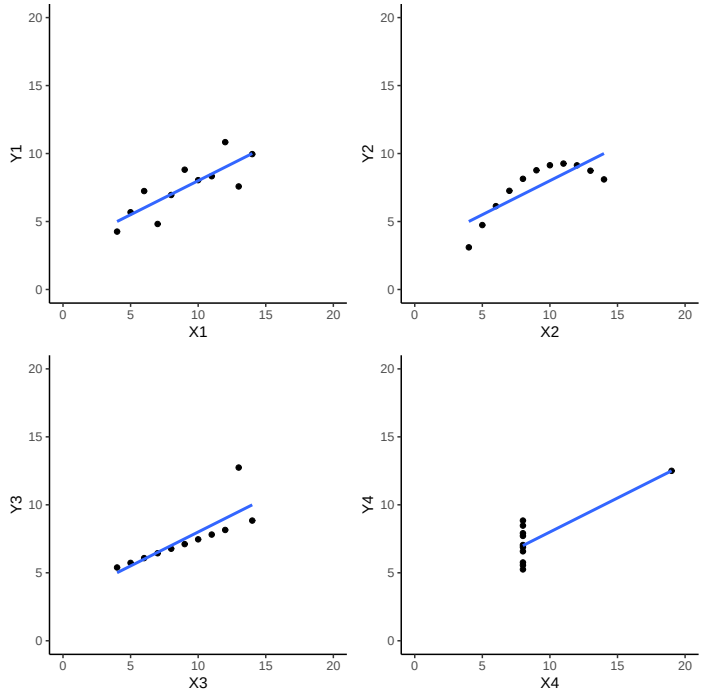
- 证实四组数据得到近似相同的经验回归方程$Y = 3.0 + 0.5X$, 回归方程(回归系数显著), p值≈ 0.0022, $R_A^2 \approx 0.67$.
- 能否认为模型$Y = \beta_0 + \beta_1 X + e$ 对这四组数据适合程度一样?

###四个模型回归系数

		Estimate	Std. Error	t value	Pr(>|t|)
组1					
	(Intercept)	3.0000909	1.1247468	2.667348	0.025734051
	X1	0.5000909	0.1179055	4.241455	0.002169629
组2					
	(Intercept)	3.0009091	1.1253024	2.666758	0.025758941
	X2	0.5000000	0.1179637	4.238590	0.002178816
组3					
	(Intercept)	3.0075455	1.1243638	2.674886	0.025418131
	X3	0.4993636	0.1178654	4.236730	0.002184806
组4					
	(Intercept)	3.0017273	1.1239211	2.670763	0.025590425
	X4	0.4999091	0.1178189	4.243028	0.002164602

###四个模型回归方程的F检验统计量

	组 1	组 2	组 3	组 4
F	17.989942968	17.965648492	17.949878082	18.003288209
p	0.002169629	0.002178816	0.002184806	0.002164602
R2	0.666542460	0.666242034	0.666046727	0.666707257



借助于散点图，可知：

- 第1组数据: 一元线性回归确实适当的
- 第2组数据: 或许缺失了X的二次项或更高次项；
- 第3组数据: 大多数数据点合适, 唯独第三个观测值远离回归直线
- 第4组数据: 不能认为Y和X存在某种线性关系.

变量选择

- 全模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n, \tag{6.11}$$

这里$\boldsymbol{y}$为$n \times 1$观测向量, $\mathbf{X} = (\mathbf{1}_n, \tilde{\mathbf{X}})$为$n \times p$的列满秩设计阵.

- 选模型

$$\boldsymbol{y} = \mathbf{X}_q \boldsymbol{\beta}_q + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n, \tag{6.12}$$

这里$\mathbf{X}_q$中包含了常数项, 且由$\mathbf{X}$的q列组成. 不失一般性，记

$$\mathbf{X} = (\mathbf{X}_q, \mathbf{X}_t), \quad \boldsymbol{\beta} = (\boldsymbol{\beta}_q', \boldsymbol{\beta}_t')'$$

变量选择对估计的影响

- 在全模型下, 回归系数β的LS估计为

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_q', \widehat{\boldsymbol{\beta}}_t')' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y},$$

- 在选模型下, β_q的LS估计为

$$\tilde{\boldsymbol{\beta}}_q = (\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q'\boldsymbol{y}.$$

定理6.3.1

假设全模型正确, 则

(1) $E(\tilde{\boldsymbol{\beta}}_q) = \boldsymbol{\beta}_q + \mathbf{A}\boldsymbol{\beta}_t$, 这里$\mathbf{A} = (\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q'\mathbf{X}_t$;

(2) $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_q) \geq \mathrm{Cov}(\tilde{\boldsymbol{\beta}}_q)$.

定理6.3.1的证明

证明 (1)可由如下推导得证,

$$E(\tilde{\boldsymbol{\beta}}_q) = (\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q'E(\boldsymbol{y}) = (\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q'(\mathbf{X}_q, \mathbf{X}_t)\begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix}$$

$$= (\mathbf{I}_q, \mathbf{A})\begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix} = \boldsymbol{\beta}_q + \mathbf{A}\boldsymbol{\beta}_t.$$

(2) $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. 由分块矩阵的逆矩阵公式(定理2.2.4)

$$\begin{pmatrix} \mathbf{X}_q'\mathbf{X}_q & \mathbf{X}_q'\mathbf{X}_t \\ \mathbf{X}_t'\mathbf{X}_q & \mathbf{X}_t'\mathbf{X}_t \end{pmatrix}^{-1} = \begin{pmatrix} (\mathbf{X}_q'\mathbf{X}_q)^{-1} + \mathbf{A}\mathbf{D}\mathbf{A}' & -\mathbf{A}\mathbf{D} \\ -\mathbf{D}\mathbf{A}' & \mathbf{D} \end{pmatrix},$$

定理6.3.1的证明

立得

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_q) = \sigma^2((\mathbf{X}_q'\mathbf{X}_q)^{-1} + \mathbf{ADA}'),$$

这里$\mathbf{D} = (\mathbf{X}_t'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_q})\mathbf{X}_t)^{-1}$. 由于$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_q) = \sigma^2(\mathbf{X}_q'\mathbf{X}_q)^{-1}$, 故

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_q) - \mathrm{Cov}(\tilde{\boldsymbol{\beta}}_q) = \sigma^2\mathbf{ADA}'.$$

因为$(\mathbf{X}'\mathbf{X})^{-1} > \mathbf{0}$, 所以$\mathbf{D} > \mathbf{0}$. 于是证得

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_q) \geq \mathrm{Cov}(\tilde{\boldsymbol{\beta}}_q).$$

有偏估计的比较准则

均方误差矩阵(mean square error matrix，MSEM)

$$\mathrm{MSEM}(\tilde{\boldsymbol{\theta}}) = E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'.$$

注 有偏估计的估计精度可采用MSEM衡量.

- MSEM与估计的协方差阵和偏差向量的关系

$$\mathrm{MSEM}(\tilde{\boldsymbol{\theta}}) = \mathrm{Cov}(\tilde{\boldsymbol{\theta}}) + (E\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(E\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'.$$

变量选择对估计的影响

定理6.3.2

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

$$\mathrm{MSEM}(\widehat{\boldsymbol{\beta}}_q) \geq \mathrm{MSEM}(\tilde{\boldsymbol{\beta}}_q).$$

证明 依定理6.3.1 立得

$$\mathrm{MSEM}(\tilde{\boldsymbol{\beta}}_q) = \sigma^2(\mathbf{X}_q'\mathbf{X}_q)^{-1} + \mathbf{A}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\mathbf{A}'.$$

注意到$\widehat{\boldsymbol{\beta}}_q$为无偏估计, 所以

$$\mathrm{MSEM}(\widehat{\boldsymbol{\beta}}_q) = \sigma^2((\mathbf{X}_q'\mathbf{X}_q)^{-1} + \mathbf{A}\mathbf{D}\mathbf{A}').$$

又因$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_t) = \sigma^2\mathbf{D}$, 故当$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_t) \geq \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$时, $\mathrm{MSEM}(\widehat{\boldsymbol{\beta}}_q) \geq \mathrm{MSEM}(\tilde{\boldsymbol{\beta}}_q)$. 定理得证.

变量选择对预测的影响

假设

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon = \boldsymbol{x}_{0q}'\boldsymbol{\beta}_q + \boldsymbol{x}_{0t}'\boldsymbol{\beta}_t + \varepsilon,$$

这里, $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$, ε 与$\boldsymbol{e}$ 不相关.

应用经验回归模型预测自变量 $\boldsymbol{x}_0 = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')'$ 对应的因变量 y_0 的值.

- 在全模型下, y_0的点预测 $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \boldsymbol{x}_{0q}'\hat{\boldsymbol{\beta}} + \boldsymbol{x}_{0t}'\hat{\boldsymbol{\beta}}_t$, 预测偏差 $z = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - y_0$.
- 在选模型下, y_0的点预测 $\tilde{y}_0 = \boldsymbol{x}_{0q}'\tilde{\boldsymbol{\beta}}_q$, 预测偏差 $z_q = \boldsymbol{x}_{0q}'\tilde{\boldsymbol{\beta}}_q - y_0$.
- 若全模型(6.11)正确, 则预测$\hat{y}_0$是无偏的, 即$E(z) = 0$.

变量选择对预测的影响

定理6.3.3

假设全模型(6.11)正确, 则

(1) $E(z_q) = \boldsymbol{x}'_{0q}\mathbf{A}\boldsymbol{\beta}_t - \boldsymbol{x}'_{0t}\boldsymbol{\beta}_t$, 这里$\mathbf{A} = (\mathbf{X}'_q\mathbf{X}_q)^{-1}\mathbf{X}'_q\mathbf{X}_t$;

(2) $\mathrm{Var}(z) \geq \mathrm{Var}(z_q)$.

证明　由$E(y_0) = \boldsymbol{x}'_{0q}\boldsymbol{\beta}_q + \boldsymbol{x}'_{0t}\boldsymbol{\beta}_t$ 和定理6.3.1, 立得(1).

(2) 依假设, ε与$\boldsymbol{e}$不相关, 故

$$\mathrm{Var}(z) = \sigma^2(1 + \boldsymbol{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_0), \quad \mathrm{Var}(z_q) = \sigma^2(1 + \boldsymbol{x}'_{0q}(\mathbf{X}'_q\mathbf{X}_q)^{-1}\boldsymbol{x}'_{0q}).$$

应用$(\mathbf{X}'\mathbf{X})^{-1}$的分块矩阵的逆, 得

定理6.3.3的证明

$$
\begin{aligned}
& \operatorname{Var}(z) - \operatorname{Var}(z_q) \\
= \quad & \sigma^2 \left(\boldsymbol{x}_0' \begin{pmatrix} (\mathbf{X}_q'\mathbf{X}_q)^{-1} + \mathbf{ADA}' & -\mathbf{AD} \\ -\mathbf{DA}' & \mathbf{D} \end{pmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_{0q}'(\mathbf{X}_q'\mathbf{X}_q)^{-1}\boldsymbol{x}_{0q} \right) \\
= \quad & \sigma^2(\boldsymbol{x}_{0q}'\mathbf{ADA}'\boldsymbol{x}_{0q} - 2\boldsymbol{x}_{0q}'\mathbf{AD}\boldsymbol{x}_{0t} + \boldsymbol{x}_{0t}'\mathbf{D}\boldsymbol{x}_{0t}) \\
= \quad & \sigma^2(\mathbf{A}'\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t})'\mathbf{D}(\mathbf{A}'\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t}) \geq 0. \quad \text{定理证毕.}
\end{aligned}
$$

这个定理的第一条结论说明, 当全模型正确时, 由选模型得到的预测$\tilde{y}_0$ 不是无偏的. 因此方差不能度量预测的优劣.

变量选择对预测的影响

预测均方误差(mean square error of prediction, MSEP)

$\tilde{y}_0$的预测均方误差定义为

$$\mathrm{MSEP}(\tilde{y}_0) = E(\tilde{y}_0 - y_0)^2 = E(z_q^2) = \mathrm{Var}(z_q) + (E(z_q))^2. \tag{6.13}$$

定理6.3.4

假设全模型(6.11)正确, 则当$\mathrm{Cov}(\widehat{\beta}_t) \geq \beta_t \beta_t'$时,

$$\mathrm{MSEP}(\widehat{y}_0) \geq \mathrm{MSEP}(\tilde{y}_0).$$

定理6.3.4的证明

证明 依公式(6.13), 得

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

根据假设条件及定理6.3.1(1), 有

$$(E(z_q))^2 = (\boldsymbol{x}'_{0q}\mathbf{A}\boldsymbol{\beta}_t - \boldsymbol{x}'_{0t}\boldsymbol{\beta}_t)^2 = (\boldsymbol{x}'_{0q}\mathbf{A} - \boldsymbol{x}'_{0t})\boldsymbol{\beta}_t\boldsymbol{\beta}'_t(\mathbf{A}'\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t})$$

$$\leq (\boldsymbol{x}'_{0q}\mathbf{A} - \boldsymbol{x}'_{0t})\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_t)(\mathbf{A}'\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t}).$$

因为$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_t) = \sigma^2\mathbf{D}$, 得$(E(z_q))^2 \leq \mathrm{Var}(z) - \mathrm{Var}(z_q)$, 从而有

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

定理证毕.

变量选择对预测的影响结论

全模型正确时, 则

- 基于选模型的估计和预测通常都是有偏的;
- 剔除一些对因变量影响较小的自变量, 可使剩余部分自变量的回归系数的LS估计精度和因变量的预测的精度提高.
- 刻画与因变量关系不是很大的协变量的一种度量:

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_t) \geq \boldsymbol{\beta}_t \boldsymbol{\beta}_t'.$$

其在全模型下的估计

$$\widehat{\sigma}^2 (\mathbf{X}_t' (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_q}) \mathbf{X}_t)^{-1} \geq \widehat{\boldsymbol{\beta}}_t \widehat{\boldsymbol{\beta}}_t'$$

评价回归方程的准则

选模型的残差平方和

$$\mathrm{SSE}_q = \|\boldsymbol{y} - \mathbf{X}_q\tilde{\boldsymbol{\beta}}_q\|^2 = \boldsymbol{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_q})\boldsymbol{y}$$

反映了数据与模型的拟合程度. 但SSE_q 随着q增加而减少.

为了防止过拟合，一种办法是添加对增加变量的惩罚因子.

常见的准则：

- R_A^2准则(或RMS_q准则)

$$R_A^2 = 1 - (n-1)\frac{\mathrm{MSE}_q}{\mathrm{SST}}, \quad \mathrm{MSE}_q = \frac{\mathrm{SSE}_q}{n-q}.$$

按“R_A^2越大越好"(或“MSE_q越小越好")选择自变量子集.

评价回归方程的准则

- C_p准则(Mallows,1973)

$$C_p = \frac{\mathrm{SSE}_q}{\hat{\sigma}^2} - (n - 2q).$$

按“C_p愈小愈好”选择自变量子集.

统计量C_p的来源

C_p 是对基于选模型的预测值$\tilde{\boldsymbol{y}}$ 与期望值$E(\boldsymbol{y})$的之间差异的衡量,

$$\Gamma_q = \frac{E(\tilde{\boldsymbol{y}} - E(\boldsymbol{y}))'(\tilde{\boldsymbol{y}} - E(\boldsymbol{y}))}{\sigma^2} = q + \frac{\boldsymbol{\beta}_t' \mathbf{D}^{-1} \boldsymbol{\beta}_t}{\sigma^2}.$$

由于$E\left(\frac{\mathrm{SSE}_q}{\sigma^2}\right) = (n-q) + \frac{\boldsymbol{\beta}_t' \mathbf{D}^{-1} \boldsymbol{\beta}_t}{\sigma^2}$, 于是$\Gamma_q = E\left(\frac{\mathrm{SSE}_q}{\sigma^2}\right) - (n - 2q)$

评价回归方程的准则

- PRESS准则

  预测误差平方和(prediction sum of squares, PRESS)

$$\mathrm{PRESS}_q = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)}),$$

  其中$\hat{y}_{i(i)} = \boldsymbol{x}'_{qi}\hat{\boldsymbol{\beta}}_{q(i)}$为$y_i$在选模型下的预测值, 这里, $\boldsymbol{x}_{qi}$为$\mathbf{X}_q$的第i行向量, $\hat{\boldsymbol{\beta}}_{(i)}$为模型去掉第$i$个观测值后所得的$\boldsymbol{\beta}_q$的LS估计. 按“PRESS愈小愈好”选择自变量子集.

PRESS值也是选模型交叉验证（cross validation, CV）的预测误差平方和，体现了模型的泛化能力.

评价回归方程的准则

- AIC准则

  是由日本统计学家Akaike(1973)提出的, 用于平衡模型选择中模型复杂度（模型自由参数的个数）和精度（拟合度）. 该准则建立在信息熵的概念基础上, 是对极大似然原理的一个扩展.

AIC 统计量

$$\text{AIC} = -2\ln L(\widetilde{\boldsymbol{\theta}}; \boldsymbol{y}) + 2q.$$

AIC准则就是选择使AIC 达到最小的自变量子集.

假设$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, 略去与$q$无关项，得

$$\text{AIC} = n\ln(\text{SSE}_q) + 2q.$$

AIC 统计量

由$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$得似然函数

$$L(\boldsymbol{\beta}_q, \sigma^2; \boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathbf{X}_q\boldsymbol{\beta}_q\|^2\Big).$$

容易求得$\boldsymbol{\beta}_q$和σ^2的极大似然估计分别为

$$\widetilde{\boldsymbol{\beta}}_q = (\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q'\boldsymbol{y}, \quad \widetilde{\sigma}_q^2 = \frac{\mathrm{SSE}_q}{n} = \frac{\boldsymbol{y}'(\mathbf{I}_q - \mathbf{X}_q(\mathbf{X}_q'\mathbf{X}_q)^{-1}\mathbf{X}_q')\boldsymbol{y}}{n}.$$

将其带入对数似然函数, 得

$$\begin{aligned}\ln L(\widetilde{\boldsymbol{\beta}}_q, \widetilde{\sigma}_q^2; \boldsymbol{y}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\widetilde{\sigma}_q^2) - \frac{n}{2} \\ &= -\frac{n}{2}\ln(\mathrm{SSE}_q) + \frac{n}{2}\ln\left(\frac{n}{2\pi}\right) - \frac{n}{2}\end{aligned}$$

略去与q无关的常数项，得$\mathrm{AIC} = n\ln(\mathrm{SSE}_q) + 2q$.

评价回归方程的准则

当样本量n很大时, 由AIC准则易多选协变量.

- BIC准则

  Schwarz(1978) 从贝叶斯观点出发提出了BIC准则.

BIC 统计量

$$\text{BIC} = -2\ln L(\widetilde{\boldsymbol{\theta}}; \boldsymbol{y}) + q\ln n.$$

BIC准则就是选择使AIC 达到最小的自变量子集

假设$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, 略去与$q$无关项，得

$$\text{BIC} = n\ln(\text{SSE}_q) + q\ln(n)$$

AIC准则和BIC准则

- 与AIC相比, BIC考虑了样本数量的影响, 其惩罚项为$q\ln(n)$.
- 当$n > 7$时,

$$q\ln(n) > 2q,$$

故在样本量大时BIC相比AIC对模型参数惩罚更多, 倾向于选择参数少的模型.

文献中还有多种对AIC和BIC准则的扩展, 如

- SBIC准则(Sawa, 1978)
- 风险膨胀准则(RIC) (Foster & George, 1994)
- 扩展的BIC准则(EBIC) (Foygel & Drton, 2010) 等.

模型所有可能回归子模型

例6.2.3中共包含4个自变量. 由于回归方程以检验是显著的, 因此所有可能的回归子模型共有$2^4 - 1 = 15$个, (即去掉所有变量都不选的情形).

Table: 所有可能回归子模型的R^2, R_A^2, C_p, PRESS, AIC和BIC的值

编号	模型中的变量	R^2	R_A^2	C_p	PRESS	AIC	BIC
1	X_4	0.675	0.645	138.731	0.560	97.744	99.439
2	X_2	0.666	0.636	142.486	0.557	98.070	99.765
3	X_1	0.534	0.492	202.549	0.374	102.412	104.107
4	X_3	0.286	0.221	315.154	**0.037**	107.960	109.655
5	X_1, X_2	0.979	0.974	**2.678**	0.965	64.312	**66.572**
6	X_1, X_4	0.972	0.967	5.496	0.955	67.634	69.894
7	X_3, X_4	0.935	0.922	22.373	0.892	78.745	81.005
8	X_2, X_3	0.847	0.816	62.438	0.742	89.930	92.189
9	X_2, X_4	0.680	0.616	138.226	0.462	99.522	101.782
10	X_1, X_3	0.548	0.458	198.095	0.183	104.009	106.269
11	X_1, X_2, X_4	0.9823	**0.9764**	3.018	0.969	**63.866**	66.691
12	X_1, X_2, X_3	0.9823	0.9763	3.041	0.967	63.904	66.728
13	X_1, X_3, X_4	0.981	0.975	3.497	0.965	64.620	67.445
14	X_2, X_3, X_4	0.973	0.964	7.337	0.946	69.468	72.293
15	X_1, X_2, X_3, X_4	**0.9824**	0.974	5.000	0.959	65.837	69.226

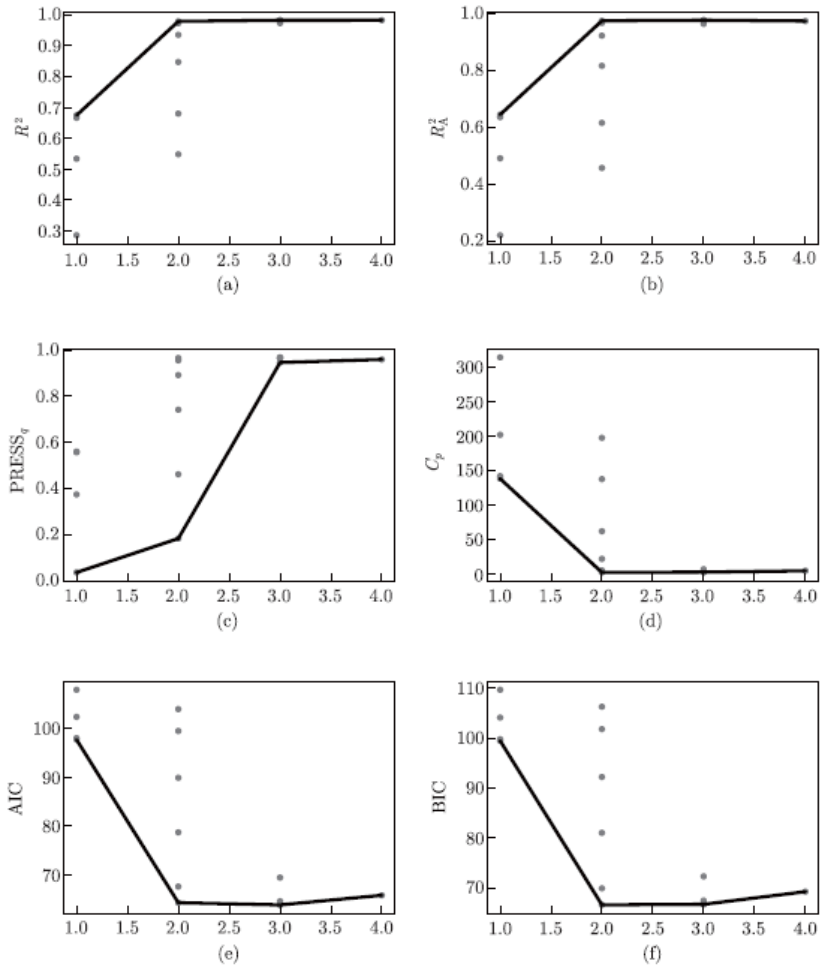


图 6.3.1　子模型的变量个数与 R^2, R_A^2, C_p, PRESS, AIC 和 BIC 各值的散点图

- 全模型(编号为15的模型) 的R^2值最大;　子模型11的R_A^2值最大;
- C_p 和PRESS值分别在编号为5和4的子模型下最小;
- AIC和BIC值最小的子模型分别于对应于R_A^2值最大和C_p值最小的编号为11和5的子模型.

变量选择的自动搜索方法

按照某变量选择准则，选择变量可采用以下方法：

- 最优子集选择(best subset selection) 法

  思想：对2^{p-1}个可能的回归模型进行比较.

- 前向选择(forward) 法

  思想：逐一引入变量, 直到没有可引变量为止.

- 后向剔除(backward) 法

  思想：初始模型为包含全体自变量的模型, 然后逐一剔除.

- 逐步回归(stepwise) 法

  思想：有进有出, 即在前向选择法/后向剔除法的每一步增加了附加条件, 考虑对现有变量的剔除/模型外变量的引入问题

变量选择的自动搜索R语言

在R 语言中, 可以使用函数step() 进行变量选择.

使用格式为

step(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = NULL, steps = 1000, k = 2, ...)

其中object 是lm()或glm()函数分析的结果

scope 是确定逐步搜索的区域

direction 是确定逐步搜索的方向：

"both" 是"逐步回归"，"backward"是后向法(只减少变量)

"forward" 是前向法(只增加变量), 默认值为"backward".

k为正数, 表示自由度数目的倍数, k=2（默认）为AIC准则;

k=log(n)为BIC准则.

例6.2.3 变量选择(向后)

```
 > lm.aic = step(lm.reg)     ##AIC
####输出结果
Start:  AIC=26.94
Y ~ X1 + X2 + X3 + X4

       Df Sum of Sq    RSS    AIC
- X3    1    0.1091 47.973 24.974
- X4    1    0.2470 48.111 25.011
- X2    1    2.9725 50.836 25.728
<none>               47.864 26.944
- X1    1   25.9509 73.815 30.576

Step:  AIC=24.97
Y ~ X1 + X2 + X4

       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- X4    1      9.93  57.90 25.420
- X2    1     26.79  74.76 28.742
- X1    1    820.91 868.88 60.629
```

例6.2.3 变量选择(逐步回归)

R Console

```
> lm.aic = step(lm.reg,  direction ="both")
Start:  AIC=26.94
Y ~ X1 + X2 + X3 + X4

       Df Sum of Sq    RSS    AIC
- X3    1    0.1091 47.973 24.974
- X4    1    0.2470 48.111 25.011
- X2    1    2.9725 50.836 25.728
<none>               47.864 26.944
- X1    1   25.9509 73.815 30.576

Step:  AIC=24.97
Y ~ X1 + X2 + X4

       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- X4    1      9.93  57.90 25.420
+ X3    1      0.11  47.86 26.944
- X2    1     26.79  74.76 28.742
- X1    1    820.91 868.88 60.629
```

AIC准则下的选模型回归结果

```
> summary(lm.aic) ##输出结果
Call:
lm(formula = Y ~ X1 + X2 + X4, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
X1            1.4519     0.1170  12.410 5.78e-07 ***
X2            0.4161     0.1856   2.242 0.051687 .
X4           -0.2365     0.1733  -1.365 0.205395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

BIC变量选择结果

```
> lm.bic <- step(lm.reg,k = log(length(data1[,1])),trace = 0) ##BIC准则
> summary(lm.bic)

Call:
lm(formula = Y ~ X1 + X2, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
-2.893 -1.574 -1.302  1.363  4.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
X1           1.46831    0.12130   12.11 2.69e-07 ***
X2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

带L_0惩罚函数的变量选择方法

带L_0惩罚函数的最小二乘目标函数

$$\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|^2+n\lambda\sum_{j=1}^{p-1}I(|\beta_i|\neq 0), \tag{6.14}$$

其中λ为调节参数(turning parameter), $I(\cdot)$ 是示性函数.

信息准则选择变量的方法大都可转化为最小化(6.14),

- AIC准则： $\lambda=\sigma\sqrt{2/n}$
- BIC准则： $\sigma\sqrt{\ln(n)/n}$

带压缩惩罚函数的变量选择方法

带压缩惩罚函数的最小二乘目标函数

$$\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_{j=1}^{p-1}p_\lambda(|\beta_i|), \tag{6.15}$$

其中$p_\lambda(\cdot)$ 为惩罚函数, λ 为调节参数或截断参数, 是用来控制模型的复杂度.

λ 的选取，可采用数据驱动的准则：

- 交叉验证(CV)
- 广义用交叉验证(GCV)
- BIC准则

压缩估计所需满足的性质

Fan & Li (2001) 建议

一个好的惩罚函数将使导出的统计量具有三条性质:

- 无偏性：

  当真参数很大时，得到的估计量是渐近无偏的，以避免不必要的建模偏差;

- 稀疏性：

  所得到的估计量是一个门限值，自动把小的参数分量估计成0，以便减少模型的复杂性;

- 连续性：

  所得估计量在数据点处是连续的, 避免模型预测的不稳定性.

压缩惩罚函数

以简单目标函数$\frac{1}{2}(z-\theta)^2+p_\lambda(|\theta|)$为例，三种常见的惩罚函数$p_\lambda(|\theta|)$：

- Lasso 惩罚函数(Tibshirani, 1996):

$$p_\lambda(|\theta|)=\lambda|\theta|;$$

- 硬门限惩罚函数(Antoniadis, 1997):

$$p_\lambda(|\theta|)=\lambda^2-(|\theta|-\lambda)^2I(|\theta|<\lambda);$$

- SCAD惩罚函数(Fan, 1997): 其导数为

$$p'_\lambda(|\theta|)=\lambda\left\{I(|\theta|\le\lambda)+\frac{(a\lambda-|\theta|)_+}{(a-1)\lambda}I(|\theta|>\lambda)\right\},$$

其中$a>2$. Fan 和Li (2001) 从Bayes角度建议取$a=3.7$.

压缩惩罚估计

极小化目标函数 $\frac{1}{2}(z-\theta)^2+p_\lambda(|\theta|)$, 得三种惩罚函数$p_\lambda(|\theta|)$下的最小二乘估计：

- Lasso 惩罚最小二乘估计（有偏）

$$\widehat{\theta}=\mathrm{sgn}(z)(|z|-\lambda)_+;$$

- 硬门限惩罚最小二乘估计（不连续）

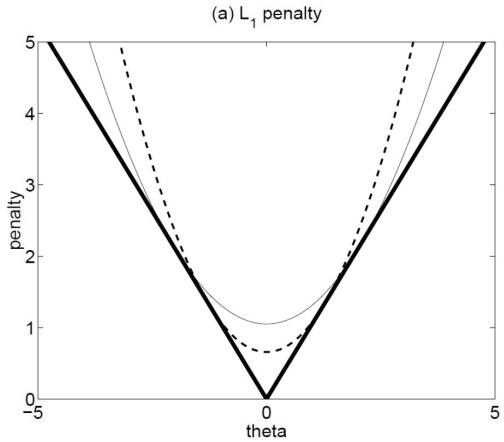
$$\widehat{\theta}=zI(|z|>\lambda).$$

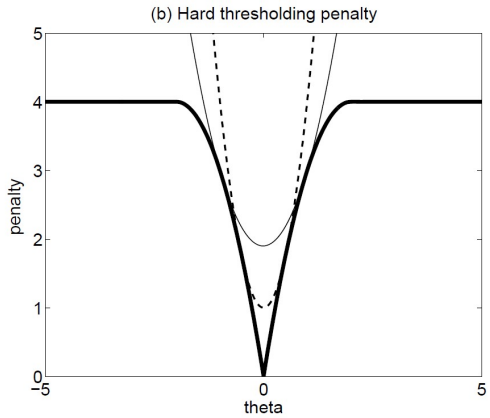
- SCAD惩罚最小二乘估计(无偏，稀疏，连续)

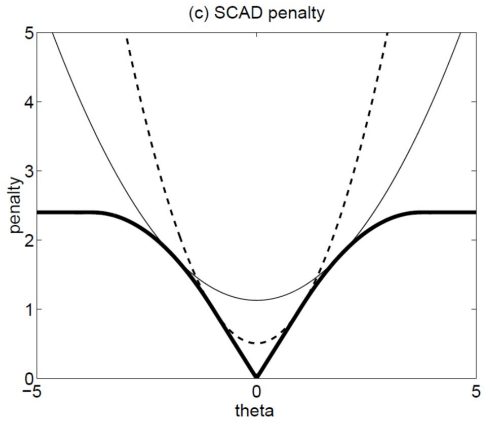
$$\widehat{\theta}=\begin{cases}\mathrm{sgn}(z)(|z|-\lambda)_+, & |z|<2\lambda,\\ \{(a-1)z-\mathrm{sgn}(z)a\lambda/(a-2)\}, & 2\lambda\le|z|\le a\lambda,\\ z, & |z|>a\lambda.\end{cases}$$

压缩惩罚函数、相应的估计

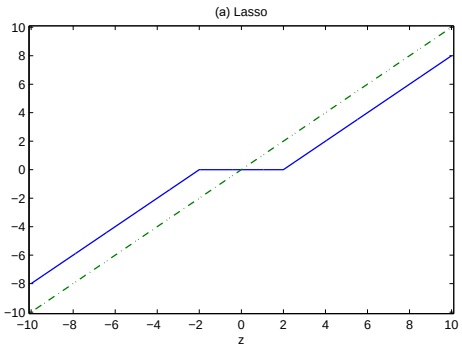
Lasso 惩罚、硬门限惩罚、SCAD惩罚函数和他们的二次逼近图：

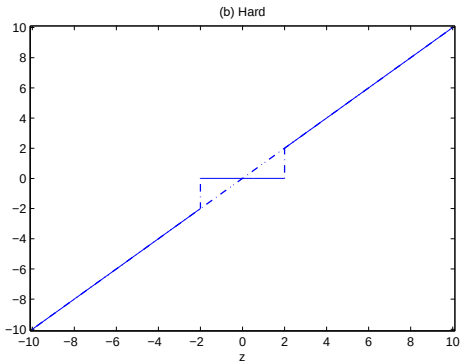


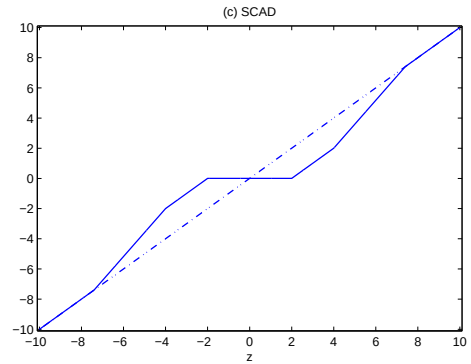




极小化目标函数，对应的Lasso、硬门限惩罚、SCAD惩罚下的解图：







自适应Lasso (adaptive Lasso)

Zou (2006)提出了自适应Lasso (adaptive Lasso)惩罚函数：

$$p_\lambda(|\beta_i|) = \lambda \hat{w}_i |\beta_i|,$$

这里$\hat{w}_i = 1/|\widehat{\beta}_i|^\gamma$, 其中$\widehat{\beta}_i$ 为模型(6.11) 的LS估计, γ为调节参数.

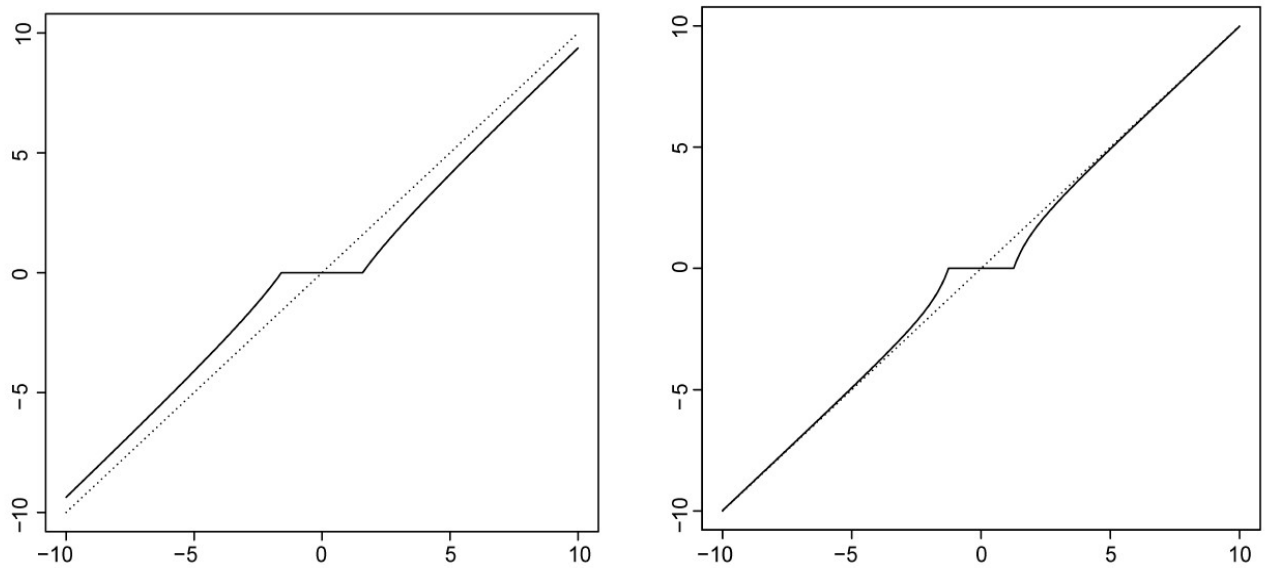
自适应Lasso 估计

$$\widehat{\boldsymbol{\beta}}_{\text{aLasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p-1} \widehat{w}_j |\beta_j| \right\}. \tag{6.16}$$

自适应Lasso 的基本思想：对于LS估计中大的回归系数不进行惩罚, 而对于接近于0 的回归系数给尽量大的惩罚, 并压缩到0.

自适应Lasso (adaptive Lasso)

当调节参数取$\lambda = 2$时，下图分别给出了$\gamma = 0.5$ 和$\gamma = 2$的自适应Lasso的门限函数图：



从图中可以看出，如果选取合适的γ, 自适应Lasso 的解将满足无偏性、稀疏性和连续性.

压缩惩罚方法的应用

注 不同于前面的压缩惩罚方法，压缩惩罚方法涉及阈值, 需要对y中心化, 对$\mathbf{X}$为$p-1$个协变量$X_1, \cdots, X_{p-1}$标准化.

在R语言中, 现成的函数.

- Lasso回归分析： glmnet、gcdnet、lars
- SCAD回归分析： ncvreg
- 自适应Lasso分析： msgps

例子

例6.3.2 （续例6.2.2 预测人体吸入氧气的效率问题）

对例6.2.2的数据, 分别采用R语言中相应的程序包进行Lasso、自适应Lasso 和SCAD回归分析.

解

- 取参数alpha=1, 用函数glmnet()进行Lasso分析;
- 用函数path.plot() 绘制Lasso估计的路径图;
- 用10折交叉验证方法的函数cv.glmnet()选取最优的调节参数λ，并绘制交叉验证误差图.

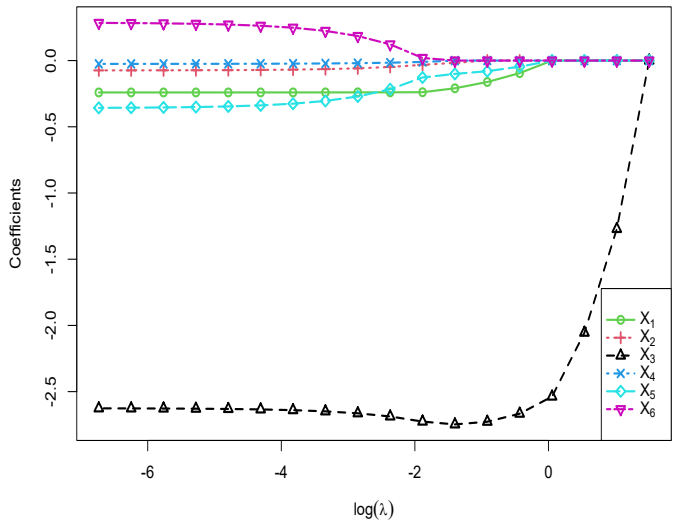
R程序

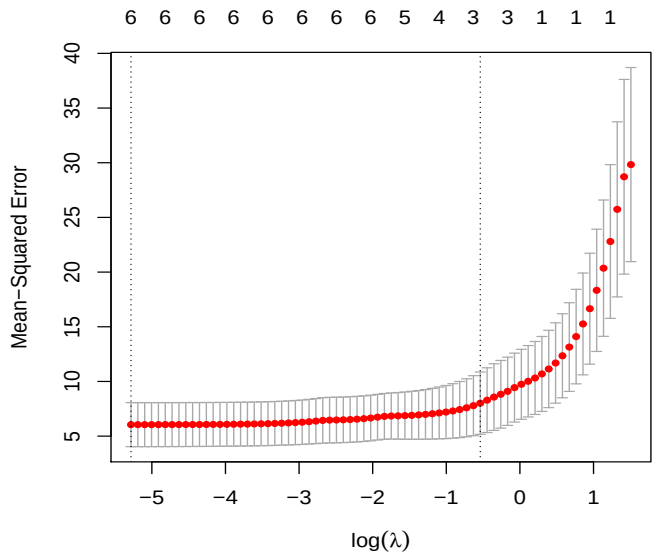
程序和输出结果如下:

```
library(glmnet);
library(latex2exp)
data<- read.table("D:/线性模型/codes/622Health.txt",header = T)
x<-model.matrix(Y~.,data)[,-1]; y=data$Y
fit_lasso = glmnet(x, y, alpha = 1, nlambda = 20)
lam = fit_lasso$lambda
beta.hat = as.matrix(fit_lasso$beta)
path.plot(lam, beta.hat)      ## 绘制Lasso估计的路径图
#### 用函数cv.glmnet()选择最优的lambda
set.seed(2021)
cv.lasso = cv.glmnet(x, y, alpha = 1)
plot(cv.lasso)                ## 绘制交叉验证误差图
```

```
> cv.lasso$lambda.min
[1] 0.005075651

> cv.lasso$lambda.1se
[1] 0.5835766
```





根据不同λ 的取值, 可以得到包含不同变量的模型, 说明Lasso方法具有筛选变量的功能.

- 图(a) 显示: 当$\lambda = 0$时, Lasso估计=最小二乘估计;

  当λ足够大时, 回归系数的Lasso估计均为0;

  Lasso估计最后被压缩成0的系数变量依次为X_3, X_1, X_5, X_6;

  而变量X_2和X_4的系数随着λ变大, 几乎很快同时被压缩成0.

- 图(b) 展示: 可见使$\mathrm{CV}(\widehat{\lambda})$ 最小化的λ 为$\widehat{\lambda} \approx 0.0051$,

  利用“一个标准差"准则选取的λ为$\widetilde{\lambda} \approx 0.5836$.

超高维下的变量选择

sure independence screening (SIS)方法

Fan 和Lv (2008) 提出了基于因变量和自变量的相关系数筛选重要变量的方法, 并给出了SIS、Lasso、自适应Lasso、SCAD方法各自适用的自变量维数范围.

- Fan, J. Q., Lv, J. C. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2008, 70(5): 849-911.

第6章第一部分总结

- 线性回归模型一般、中心化、标准化形式
- 参数LS估计的性质
- 回归方程、回归系数的显著性检验
- 变量选择