# 第五章 自变量的选择

第三章和第四章分别讨论了线性回归模型的参数估计和假设检验问题, 但应用回归分析处理实际问题时, 首先要解决的问题是模型的选择(model selection). 模型的选择包含两方面的内容. 一是选择回归模型的类型, 即选择线性回归模型还是非线性回归模型来处理实际问题, 统计学上称之为回归模型的线性检验. 在有重复试验的情形下可以使用卡方拟合优度检验来处理这个问题, 本课程不讨论这部分内容. 二是在选定模型的类型后, 自变量的选择问题(variable selection). 自变量选择过少或选择不当, 会使所建立的模型与实际有较大的偏离而无法使用. 自变量选择过多, 其后果是计算量增大、估计和预测的精度也会下降.

## 5.1 自变量选择的后果

假设根据经验和专业知识, 初步确定可能对因变量$y$有影响的自变量共有$p$个, 记为$x_1, \cdots, x_p$. 相应的(矩阵形式)线性回归模型为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \quad \mathsf{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n, \tag{5.1.1}$$

这里$\boldsymbol{X}$为$n \times (p+1)$的列满秩设计矩阵, 第一列元素全为1. 称(5.1.1)为全模型.

假设根据某些自变量选择的准则, 剔除了(5.1.1)中的一些对因变量影响较小的自变量, 不妨假设剔除了后面的$p-q$个自变量$x_{q+1}, \cdots, x_p$. 记

$$\begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \boldsymbol{X} = (\boldsymbol{X}_q, \boldsymbol{X}_t) = \begin{pmatrix} \boldsymbol{x}_{1q}' & \boldsymbol{x}_{1t}' \\ \vdots & \vdots \\ \boldsymbol{x}_{nq}' & \boldsymbol{x}_{nt}' \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix}.$$

现在得到一个新的模型

$$\boldsymbol{Y} = \boldsymbol{X}_q\boldsymbol{\beta}_q + \boldsymbol{e}, \quad \mathsf{E}(\boldsymbol{e}) = \boldsymbol{0}, \quad \mathsf{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n, \tag{5.1.2}$$

这里的$\boldsymbol{X}_q$为$n \times (q+1)$的列满秩设计矩阵, $\boldsymbol{\beta}_q$为$q+1$维的列向量. 称(5.1.2)为选模型.

在全模型中, 回归系数$\boldsymbol{\beta}$和$\sigma^2$的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \quad \hat{\sigma}^2 = \frac{\boldsymbol{Y}'[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}}{n-p-1}, \tag{5.1.3}$$

在点$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$. 在选模型中, 回归系数$\boldsymbol{\beta}_q$和$\sigma^2$的最小二乘估计为

$$\widetilde{\boldsymbol{\beta}}_q = (\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'\boldsymbol{Y}, \quad \widetilde{\sigma}_q^2 = \frac{\boldsymbol{Y}'[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{Y}}{n-q-1}, \tag{5.1.4}$$

在点$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$的预测为$\widetilde{y}_{0q} = \boldsymbol{x}_{0q}'\widetilde{\boldsymbol{\beta}}_q$. 因$\hat{\boldsymbol{\beta}}$和$\widetilde{\boldsymbol{\beta}}_q$的维数不同, 故对$\hat{\boldsymbol{\beta}}$作分块:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_q \\ \hat{\boldsymbol{\beta}}_t \end{pmatrix},$$

使得$\hat{\boldsymbol{\beta}}_q$和$\widetilde{\boldsymbol{\beta}}_q$有相同的维数.

若一个估计量$\widetilde{\boldsymbol{\theta}}$是未知参数$\boldsymbol{\theta}$的有偏估计, 那么协方差矩阵不能作为衡量估计精度之用, 一个更合理的度量标准为均方误差矩阵(mean square error matrix, MSEM).

**定义**5.1.1　设$\boldsymbol{\theta}$是一未知参数向量, $\widetilde{\boldsymbol{\theta}}$为$\boldsymbol{\theta}$的一个估计. 定义$\widetilde{\boldsymbol{\theta}}$的均方误差矩阵为

$$\mathrm{MSEM}(\widetilde{\boldsymbol{\theta}}) = \mathsf{E}[(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'].$$

不难看出

$$\mathrm{MSEM}(\widetilde{\boldsymbol{\theta}}) = \mathsf{Cov}(\widetilde{\boldsymbol{\theta}}) + (\mathsf{E}(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\theta})(\mathsf{E}(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{\theta})'. \tag{5.1.5}$$

回忆分块矩阵求逆公式.

**引理**5.1.1(分块矩阵求逆公式)　设$\boldsymbol{A}$为非奇异的对称矩阵, 将其分块为

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

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

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

首先讨论模型选择不当对估计的影响, 我们有下面的结论.

**定理**5.1.1(模型选择不当对估计的影响)　假设全模型(5.1.1)正确, 则

(1) $\mathsf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, $\mathsf{E}(\widetilde{\boldsymbol{\beta}}_q) = \boldsymbol{\beta}_q + \boldsymbol{G}\boldsymbol{\beta}_t$, 这里$\boldsymbol{G} = (\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'\boldsymbol{X}_t$. 所以除了$\boldsymbol{\beta}_t = \boldsymbol{0}$或者$\boldsymbol{X}_q'\boldsymbol{X}_t = \boldsymbol{0}$外, $\mathsf{E}(\widetilde{\boldsymbol{\beta}}_q) \neq \boldsymbol{\beta}_q$;

(2) $\mathsf{Cov}(\hat{\boldsymbol{\beta}}_q) - \mathsf{Cov}(\widetilde{\boldsymbol{\beta}}_q)$为非负定矩阵;

(3) 当$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵时, $\mathrm{MSEM}(\hat{\boldsymbol{\beta}}_q) - \mathrm{MSEM}(\widetilde{\boldsymbol{\beta}}_q)$为非负定矩阵;

(4) $\mathsf{E}(\widetilde{\sigma}_q^2) \geq \mathsf{E}(\hat{\sigma}^2) = \sigma^2$, 且仅当$\boldsymbol{\beta}_t = \boldsymbol{0}$时等号成立.

**证明**　(1) $\mathsf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$是显然的. 现来考察$\widetilde{\boldsymbol{\beta}}_q$的均值. 根据(5.1.4),

$$\begin{aligned} \mathsf{E}(\widetilde{\boldsymbol{\beta}}_q) =& (\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'\mathsf{E}(\boldsymbol{Y}) \\ =& (\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'(\boldsymbol{X}_q, \boldsymbol{X}_t)\begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix} \\ =& (\boldsymbol{I}_{q+1}, \boldsymbol{G})\begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix} \\ =& \boldsymbol{\beta}_q + \boldsymbol{G}\boldsymbol{\beta}_t, \end{aligned}$$

不难看出, 除了$\boldsymbol{\beta}_t = \boldsymbol{0}$或者$\boldsymbol{X}_q'\boldsymbol{X}_t = \boldsymbol{0}$外, $\mathsf{E}(\widetilde{\boldsymbol{\beta}}_q) \neq \boldsymbol{\beta}_q$.

(2) 记

$$\boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_q'\\ \boldsymbol{X}_t'\end{pmatrix}(\boldsymbol{X}_q,\boldsymbol{X}_t)=\begin{pmatrix}\boldsymbol{X}_q'\boldsymbol{X}_q & \boldsymbol{X}_q'\boldsymbol{X}_t\\ \boldsymbol{X}_t'\boldsymbol{X}_q & \boldsymbol{X}_t'\boldsymbol{X}_t\end{pmatrix}=:\begin{pmatrix}\boldsymbol{B} & \boldsymbol{C}\\ \boldsymbol{C}' & \boldsymbol{D}\end{pmatrix},$$

这里$\boldsymbol{B}=\boldsymbol{X}_q'\boldsymbol{X}_q$, $\boldsymbol{C}=\boldsymbol{X}_q'\boldsymbol{X}_t$, $\boldsymbol{D}=\boldsymbol{X}_t'\boldsymbol{X}_t$. 又记

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

由

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}})=\mathsf{Cov}\begin{pmatrix}\hat{\boldsymbol{\beta}}_q\\ \hat{\boldsymbol{\beta}}_t\end{pmatrix}=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}=\sigma^2\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{C}_1\\ \boldsymbol{C}_1' & \boldsymbol{D}_1\end{pmatrix}$$

知$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_q)=\sigma^2\boldsymbol{B}_1$. 又$\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}_q)=\sigma^2(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}=\sigma^2\boldsymbol{B}^{-1}$, 所以由分块矩阵求逆公式知

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_q)-\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}_q)=\sigma^2(\boldsymbol{B}_1-\boldsymbol{B}^{-1})=\sigma^2\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1},$$

该矩阵显然是一个非负定矩阵.

(3) 由公式(5.1.5)以及结论(1)可知

$$\begin{aligned}&\mathrm{MSEM}(\widetilde{\boldsymbol{\beta}}_q)=\sigma^2(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}+\boldsymbol{G}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\boldsymbol{G}'=\sigma^2\boldsymbol{B}^{-1}+\boldsymbol{G}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\boldsymbol{G}',\\&\mathrm{MSEM}(\hat{\boldsymbol{\beta}}_q)=\sigma^2\boldsymbol{B}_1.\end{aligned}$$

注意到$\boldsymbol{G}=\boldsymbol{B}^{-1}\boldsymbol{C}$, 所以当$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵时,

$$\begin{aligned}\mathrm{MSEM}(\hat{\boldsymbol{\beta}}_q)-\mathrm{MSEM}(\widetilde{\boldsymbol{\beta}}_q)=&\sigma^2\boldsymbol{B}_1-\sigma^2\boldsymbol{B}^{-1}-\boldsymbol{G}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\boldsymbol{G}'\\=&\sigma^2\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1}-\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\boldsymbol{C}'\boldsymbol{B}^{-1}\\=&\boldsymbol{B}^{-1}\boldsymbol{C}(\sigma^2\boldsymbol{D}_1-\boldsymbol{\beta}_t\boldsymbol{\beta}_t')\boldsymbol{C}'\boldsymbol{B}^{-1}\\=&\boldsymbol{B}^{-1}\boldsymbol{C}(\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t')\boldsymbol{C}'\boldsymbol{B}^{-1}\end{aligned}$$

为非负定矩阵.

(4) 第三章已经证明$\mathsf{E}(\hat{\sigma}^2)=\sigma^2$. 此外, 利用迹的性质, 有

$$\begin{aligned}\mathsf{E}(\widetilde{\sigma}_q^2)=&\frac{1}{n-q-1}\mathsf{E}\Big\{\boldsymbol{Y}'[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{Y}\Big\}\\=&\frac{1}{n-q-1}\mathrm{tr}\Big\{[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}')\Big\}\\=&\frac{1}{n-q-1}\mathrm{tr}\Big\{[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'](\sigma^2\boldsymbol{I}_n+\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{X}')\Big\}\\=&\frac{1}{n-q-1}\Big\{(n-q-1)\sigma^2+\boldsymbol{\beta}'\boldsymbol{X}'[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{X}\boldsymbol{\beta}\Big\}\\=&\sigma^2+\frac{1}{n-q-1}\boldsymbol{\beta}_t'\boldsymbol{X}_t'[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{X}_t\boldsymbol{\beta}_t\end{aligned}$$

$$
\begin{aligned}
=&\sigma^2+\frac{1}{n-q-1}\boldsymbol{\beta}_t'(\boldsymbol{D}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})\boldsymbol{\beta}_t\\
=&\sigma^2+\frac{1}{n-q-1}\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t\\
\geq&\sigma^2=\mathsf{E}(\hat{\sigma}^2),
\end{aligned}
$$

且等号成立当且仅当$\boldsymbol{\beta}_t=\boldsymbol{0}$. 证毕. □

**注** 实际上, 当推导得到$\mathsf{E}(\widetilde{\sigma}_q^2)=\sigma^2+\frac{1}{n-q-1}\boldsymbol{\beta}_t'\boldsymbol{X}_t'[\boldsymbol{I}_n-\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{X}_t\boldsymbol{\beta}_t$时, 已可知$\mathsf{E}(\widetilde{\sigma}_q^2)\geq\sigma^2$.

记全模型的预测偏差为

$$z_0=y_0-\hat{y}_0=y_0-\boldsymbol{x}_0'\hat{\boldsymbol{\beta}},$$

选模型的预测偏差为

$$z_{0q}=y_0-\widetilde{y}_{0q}=y_0-\boldsymbol{x}_{0q}'\widetilde{\boldsymbol{\beta}}_q.$$

接下来讨论模型选择不当对预测的影响, 我们有下面的结论.

**定理**5.1.2(模型选择不当对预测的影响) 假设全模型(5.1.1)正确, 则

(1) $\mathsf{E}(z_0)=0$, $\mathsf{E}(z_{0q})=\boldsymbol{x}_{0t}'\boldsymbol{\beta}_t-\boldsymbol{x}_{0q}'\boldsymbol{G}\boldsymbol{\beta}_t$. 所以一般情形下, $\widetilde{y}_{0q}$为有偏预测;

(2) $\mathsf{Var}(z_0)\geq\mathsf{Var}(z_{0q})$;

(3) 当$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵时, $\mathrm{MSE}(\hat{y}_0)-\mathrm{MSE}(\widetilde{y}_{0q})\geq 0$.

证明: (1) $\mathsf{E}(z_0)=0$是显然的. 现考察$\mathsf{E}(z_{0q})$. 由定理5.1.1中的结论(1)可知

$$\mathsf{E}(z_{0q})=\boldsymbol{x}_0'\boldsymbol{\beta}-\boldsymbol{x}_{0q}'\mathsf{E}(\widetilde{\boldsymbol{\beta}}_q)=\boldsymbol{x}_0'\boldsymbol{\beta}-\boldsymbol{x}_{0q}'(\boldsymbol{\beta}_q+\boldsymbol{G}\boldsymbol{\beta}_t)=\boldsymbol{x}_{0t}'\boldsymbol{\beta}_t-\boldsymbol{x}_{0q}'\boldsymbol{G}\boldsymbol{\beta}_t,$$

所以一般情形下$\widetilde{y}_{0q}$是有偏预测.

(2) 首先, 容易看出

$$\mathsf{Var}(z_0)=\mathsf{Var}(y_0-\hat{y}_0)=\mathsf{Var}(y_0)+\mathsf{Var}(\hat{y}_0)=\sigma^2(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

而

$$
\begin{aligned}
\mathsf{Var}(z_{0q})=&\mathsf{Var}(y_0-\widetilde{y}_{0q})=\mathsf{Var}(y_0)+\mathsf{Var}(\widetilde{y}_{0q})\\
=&\sigma^2(1+\boldsymbol{x}_{0q}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{0q}).
\end{aligned}
$$

注意到

$$\boldsymbol{x}_{0q}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{0q}=\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}$$

以及

$$
\begin{aligned}
\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0=&(\boldsymbol{x}_{0q}',\boldsymbol{x}_{0t}')\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{C}_1\\ \boldsymbol{C}_1' & \boldsymbol{D}_1\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_{0q}\\ \boldsymbol{x}_{0t}\end{pmatrix}\\
=&\boldsymbol{x}_{0q}'\boldsymbol{B}_1\boldsymbol{x}_{0q}+\boldsymbol{x}_{0q}'\boldsymbol{C}_1\boldsymbol{x}_{0t}+\boldsymbol{x}_{0t}'\boldsymbol{C}_1'\boldsymbol{x}_{0q}+\boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{x}_{0t},
\end{aligned}
$$

可推得

$$\mathsf{Var}(z_0)-\mathsf{Var}(z_{0q})$$

$$
\begin{aligned}
&=\sigma^2[\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0-\boldsymbol{x}_{0q}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{0q}]\\
&=\sigma^2[\boldsymbol{x}_{0q}'(\boldsymbol{B}_1-\boldsymbol{B}^{-1})\boldsymbol{x}_{0q}+\boldsymbol{x}_{0q}'\boldsymbol{C}_1\boldsymbol{x}_{0t}+\boldsymbol{x}_{0t}'\boldsymbol{C}_1'\boldsymbol{x}_{0q}+\boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{x}_{0t}]\\
&=\sigma^2[\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{x}_{0t}-\boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}+\boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{x}_{0t}]\\
&=\sigma^2[\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})-\boldsymbol{x}_{0t}'\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})]\\
&=\sigma^2(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})'\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})\geq 0.
\end{aligned}
$$

(3) 首先, 容易看出

$$
\begin{aligned}
&\mathrm{MSE}(\hat{y}_0)=\mathsf{E}(\hat{y}_0-y_0)^2=\mathsf{E}(z_0^2)=\mathsf{Var}(z_0),\\
&\mathrm{MSE}(\widetilde{y}_{0q})=\mathsf{E}(\widetilde{y}_{0q}-y_0)^2=\mathsf{E}(z_{0q}^2)=\mathsf{Var}(z_{0q})+[\mathsf{E}(z_{0q})]^2.
\end{aligned}
$$

由(1)的证明可得

$$
\begin{aligned}
[\mathsf{E}(z_{0q})]^2=&(\boldsymbol{x}_{0t}'\boldsymbol{\beta}_t-\boldsymbol{x}_{0q}'\boldsymbol{G}\boldsymbol{\beta}_t)^2\\
=&(\boldsymbol{x}_{0t}'-\boldsymbol{x}_{0q}'\boldsymbol{G})\boldsymbol{\beta}_t\boldsymbol{\beta}_t'(\boldsymbol{x}_{0t}'-\boldsymbol{x}_{0q}'\boldsymbol{G})'\\
=&(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})'\boldsymbol{\beta}_t\boldsymbol{\beta}_t'(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t}).
\end{aligned}
$$

所以当$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵时, 根据(2)的证明过程可知

$$
\begin{aligned}
&\mathrm{MSE}(\hat{y}_0)-\mathrm{MSE}(\widetilde{y}_{0q})\\
=&\mathsf{Var}(z_0)-\mathsf{Var}(z_{0q})-[\mathsf{E}(z_{0q})]^2\\
=&(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})'(\sigma^2\boldsymbol{D}_1-\boldsymbol{\beta}_t\boldsymbol{\beta}_t')(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})\\
=&(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})'(\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t')(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q}-\boldsymbol{x}_{0t})\geq 0.
\end{aligned}
$$

证毕. □

总结: (1) 即使全模型正确, 剔除一部分自变量后, 可使得剩余的那部分自变量的回归系数的LSE的方差减少, 但此时的估计一般为有偏估计. 若被剔除的自变量对因变量影响较小或难于掌握(用$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵来刻画), 则剔除这些自变量后可使得剩余自变量的回归系数的LSE的精度(用均方误差来刻画)有所提高. (2) 当全模型正确时, 用选模型作预测, 则预测一般是有偏的, 但预测偏差的方差减小. 若被剔除的自变量对因变量影响较小或难于掌握(用$\mathsf{Cov}(\hat{\boldsymbol{\beta}}_t)-\boldsymbol{\beta}_t\boldsymbol{\beta}_t'$为非负定矩阵来刻画), 则剔除这些自变量后可使得预测的精度(用均方误差来刻画)有所提高.

因此在应用回归分析去处理实际问题时, 无论从回归系数估计的角度看, 还是从预测的角度看, 对那些与因变量关系不大或难于掌握的自变量从模型中剔除都是有利的. 回归模型中自变量的选择要做到少而精.

## 5.2 基于准则的自变量选择

统计学家从数据与模型的拟合程度、预测精度等不同角度出发提出了多种回归自变量的选择准则, 它们都是对回归自变量的所有不同子集进行比较, 然后从中挑选出一个“最优”的, 且绝大部分选择的准则都是与残差平方和有关. 但是我

们不能直接把“残差平方和越小越好”当成自变量选择的一个准则, 理由如下: 记选模型(5.1.2)的残差平方和为$\mathrm{RSS}_q$, 则

$$\mathrm{RSS}_q = \min_{(\beta_0,\beta_1,\cdots,\beta_q)'\in\mathbb{R}^{q+1}} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_q x_{iq})^2.$$

当在选模型(5.1.2)中增加自变量$x_{q+1}$后, 相应的残差平方和

$$\mathrm{RSS}_{q+1} = \min_{(\beta_0,\beta_1,\cdots,\beta_{q+1})'\in\mathbb{R}^{q+2}} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{q+1} x_{i,q+1})^2.$$

改写$\mathrm{RSS}_q$如下:

$$\mathrm{RSS}_q = \min_{\substack{(\beta_0,\beta_1,\cdots,\beta_{q+1})'\in\mathbb{R}^{q+2}\\ \beta_{q+1}=0}} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{q+1} x_{i,q+1})^2.$$

则可知$\mathrm{RSS}_{q+1} \le \mathrm{RSS}_q$. 因此当自变量子集扩大时, 残差平方和随之减少, 如果按“RSS越小越好”的原则选择自变量, 则选入回归模型的自变量将越来越多, 最后将把所有自变量选入回归模型. 可见, 残差平方和不能直接用作选择自变量的准则. 由于$R^2 = \mathrm{ESS}/\mathrm{TSS} = 1 - \mathrm{RSS}/\mathrm{TSS}$, 所以$R^2$也不能直接用作选择自变量的准则.

上述的分析有重要的意义. RSS和$R^2$可用于自变量数目相等的线性回归模型的比较, 但不适用于比较自变量数目不等的线性回归模型. 因为模型的自变量数目越多, RSS会越小($R^2$会越大), 即使新增加的自变量对因变量没有真正的解释能力, RSS也会变小($R^2$也会变大).

那么, 对于线性回归模型, 如果建模的目的是因变量的预测, 什么是合适的模型选择准则呢? 通常存在很多备选的自变量可用于预测因变量, 但没有必要在回归模型中包含所有的自变量. 这里存在权衡关系: 一方面, 自变量越多, 模型的系统偏差越小, 如果所有的参数估计都不存在误差, 那么该模型的预测能力是最优的. 但另一方面, 在样本容量给定的条件下, 参数越多, 参数估计的准确性会越差. 统计学里有一个重要的思想叫做“KISS(keep it sophistically simple)原则”, 就是尽量用简单的模型去刻画数据所包含的重要信息.

下面介绍自变量选择的几个常见准则.

(1) 平均残差平方和准则($\mathrm{RMS}_q$)

由于$\mathrm{RSS}_q$随$q$的增大而下降, 为了防止选取过多的自变量, 一个常见的做法是对$\mathrm{RSS}_q$乘上一个随$q$增加而上升的函数, 作为惩罚因子. 于是定义

$$\mathrm{RMS}_q = \frac{\mathrm{RSS}_q}{n-q-1},$$

分母$n-q-1$其实是$\mathrm{RSS}_q$的自由度. 我们按$\mathrm{RMS}_q$越小越好的原则选择自变量, 并称其为平均残差平方和准则或$\mathrm{RMS}_q$准则.

(2) 调整后的$R^2$准则

判定系数$R_q^2 = \mathrm{ESS}_q/\mathrm{TSS}$度量了数据与模型的拟合程度, 自然希望它越大越好. 但根据定义$R_q^2 = 1 - \mathrm{RSS}_q/\mathrm{TSS}$, 我们不能直接把$R_q^2$作为选择自变量的准则,

否则将把所有的自变量(不管它们是否对因变量有真正的解释能力)选入模型. 为了克服以上缺点, 引入调整后的判定系数

$$\bar{R}_q^2 = 1 - \frac{\mathrm{RSS}_q/(n-q-1)}{\mathrm{TSS}/(n-1)} = 1 - \frac{n-1}{n-q-1}\frac{\mathrm{RSS}_q}{\mathrm{TSS}} = 1 - \frac{n-1}{n-q-1}(1-R_q^2).$$

易知$\bar{R}_q^2 \le R_q^2$, 且$\bar{R}_q^2$并不一定随着自变量个数的增加而增加. 这是因为, 尽管$1-R_q^2$随着自变量的个数的增加而减少, 但是$(n-1)/(n-q-1)$随着$q$的增加而增加, 这就使得$\bar{R}_q^2$并不一定随$q$的增加而增加. 我们选择使$\bar{R}_q^2$达到最大的自变量子集.

(3) $C_p$准则

$C_p$准则是Mallows于1964年提出的, 它是从预测的观点出发提出来的. 对于选模型(5.1.2), $C_p$统计量定义为

$$C_p = \frac{\mathrm{RSS}_q}{\hat{\sigma}^2} - [n - 2(q+1)], \tag{5.2.1}$$

这里$\mathrm{RSS}_q$是选模型(5.1.2)的残差平方和, $\hat{\sigma}^2$为全模型(5.1.1)中$\sigma^2$的最小二乘估计. 我们按“$C_p$越小越好”的准则来选择自变量.

获得(5.2.1)的想法如下: 假设全模型为真, 但为了提高预测的精度, 用选模型(5.1.2)去做预测. 很自然地, 要求$n$个预测值与期望值的相对偏差平方和的期望值

$$\Gamma_q := \mathsf{E}\Big\{\sum_{i=1}^n \Big(\frac{\widetilde{y}_{iq} - \mathsf{E}(y_i)}{\sigma}\Big)^2\Big\} = \mathsf{E}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^n (\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big\}$$

达到最小. 写

$$\begin{aligned}
&\mathsf{E}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^n (\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big\}\\
=&\frac{1}{\sigma^2}\sum_{i=1}^n \mathsf{E}\Big\{[\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q - \mathsf{E}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q)] + [\mathsf{E}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q) - \boldsymbol{x}'\boldsymbol{\beta}]\Big\}^2\\
=&\frac{1}{\sigma^2}\sum_{i=1}^n \Big\{\mathsf{E}[\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q - \mathsf{E}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q)]^2 + [\mathsf{E}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q) - \boldsymbol{x}_i'\boldsymbol{\beta}]^2\Big\}\\
=:&\frac{1}{\sigma^2}(I_1 + I_2).
\end{aligned}$$

易知

$$\begin{aligned}
I_1 =&\sum_{i=1}^n \mathsf{Var}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q) = \sigma^2\sum_{i=1}^n \boldsymbol{x}_{iq}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{iq}\\
=&\sigma^2\mathrm{tr}\Big[(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\sum_{i=1}^n \boldsymbol{x}_{iq}\boldsymbol{x}_{iq}'\Big]\\
=&(q+1)\sigma^2.
\end{aligned}$$

利用定理5.1.1中的结论(1)以及结论(4)的证明过程, 得

$$\begin{aligned}
I_2 =&\sum_{i=1}^{n}[\boldsymbol{x}_{iq}'(\boldsymbol{\beta}_q+\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{\beta}_t)-(\boldsymbol{x}_{iq}'\boldsymbol{\beta}_q+\boldsymbol{x}_{it}'\boldsymbol{\beta}_t)]^2\\
=&\sum_{i=1}^{n}(\boldsymbol{x}_{iq}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{\beta}_t-\boldsymbol{x}_{it}'\boldsymbol{\beta}_t)^2\\
=&\sum_{i=1}^{n}\boldsymbol{\beta}_t'(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{iq}-\boldsymbol{x}_{it})(\boldsymbol{x}_{iq}'\boldsymbol{B}^{-1}\boldsymbol{C}-\boldsymbol{x}_{it}')\boldsymbol{\beta}_t\\
=&\sum_{i=1}^{n}\boldsymbol{\beta}_t'[\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{iq}\boldsymbol{x}_{iq}'\boldsymbol{B}^{-1}\boldsymbol{C}-\boldsymbol{x}_{it}\boldsymbol{x}_{iq}'\boldsymbol{B}^{-1}\boldsymbol{C}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{iq}\boldsymbol{x}_{it}'+\boldsymbol{x}_{it}\boldsymbol{x}_{it}']\boldsymbol{\beta}_t\\
=&\boldsymbol{\beta}_t'[\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{C}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C}+\boldsymbol{D}]\boldsymbol{\beta}_t\\
=&\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t\\
=&(n-q-1)[\mathsf{E}(\widetilde{\sigma}_q^2)-\sigma^2].
\end{aligned}$$

所以

$$\begin{aligned}
\Gamma_q =&\mathsf{E}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(\boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q-\boldsymbol{x}_i'\boldsymbol{\beta})^2\Big\}\\
=&\frac{1}{\sigma^2}\Big\{(q+1)\sigma^2+(n-q-1)[\mathsf{E}(\widetilde{\sigma}_q^2)-\sigma^2]\Big\}\\
=&\frac{\mathsf{E}(\mathrm{RSS}_q)}{\sigma^2}-[n-2(q+1)].
\end{aligned}$$

因为$\mathsf{E}(\mathrm{RSS}_q)$与$\sigma^2$未知, 所以用$\mathrm{RSS}_q$代替$\mathsf{E}(\mathrm{RSS}_q)$, 以及用$\hat{\sigma}^2$代替$\sigma^2$, 即可得到$C_p$统计量.

**注** 由上述证明过程可知:

$$\Gamma_q=q+1+\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}.$$

鉴于$C_p$统计量的重要性, 下面阐述$C_p$统计量的一些性质, 它们对于应用$C_p$统计量作自变量选择, 提供了理论依据.

**定理**5.2.1 假设随机向量$\boldsymbol{e}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, 则对选模型(5.1.2)的$C_p$统计量, 有

$$\mathsf{E}(C_p)=q+1-t+\frac{n-p-1}{n-p-3}\Big(t+\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}\Big),$$

这里$\boldsymbol{D}_1=(\boldsymbol{D}-\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}$, $\boldsymbol{B}=\boldsymbol{X}_q'\boldsymbol{X}_q$, $\boldsymbol{C}=\boldsymbol{X}_q'\boldsymbol{X}_t$, $\boldsymbol{D}=\boldsymbol{X}_t'\boldsymbol{X}_t$.

**证明** 问题归结为计算$\mathsf{E}(\mathrm{RSS}_q/\hat{\sigma}^2)$. 对于全模型(5.1.1), 残差平方和$\mathrm{RSS}=(n-p-1)\hat{\sigma}^2$且

$$\frac{\mathrm{RSS}}{\sigma^2}\sim\chi^2(n-p-1).$$

选模型(5.1.2)中的残差平方和$\mathrm{RSS}_q$可看成是在假设$H:\boldsymbol{\beta}_t=\mathbf{0}$下模型的残差平方和. 因此, $\eta:=\mathrm{RSS}_q-\mathrm{RSS}$与RSS相互独立(根据上一章的最小二乘法基本定理), 且

$$\mathsf{E}\Big(\frac{\mathrm{RSS}_q}{\hat{\sigma}^2}\Big)=(n-p-1)\mathsf{E}\Big(\frac{\mathrm{RSS}_q}{\mathrm{RSS}}\Big)=(n-p-1)\Big[1+\mathsf{E}(\eta)\cdot\mathsf{E}\Big(\frac{1}{\mathrm{RSS}}\Big)\Big].$$

记$k=n-p-1$, 由$\mathrm{RSS}/\sigma^2\sim\chi^2(k)$得

$$\begin{aligned}\mathsf{E}\Big(\frac{\sigma^2}{\mathrm{RSS}}\Big)=&2^{-\frac{k}{2}}\Big[\Gamma\Big(\frac{k}{2}\Big)\Big]^{-1}\int_0^\infty x^{-1}\cdot e^{-\frac{x}{2}}x^{\frac{k}{2}-1}\mathrm{d}x\\=&2^{-\frac{k}{2}}\Big[\Gamma\Big(\frac{k}{2}\Big)\Big]^{-1}2^{\frac{k}{2}-1}\Gamma\Big(\frac{k}{2}-1\Big)\\=&\frac{1}{k-2}=\frac{1}{n-p-3}.\end{aligned}$$

因此

$$\mathsf{E}\Big(\frac{1}{\mathrm{RSS}}\Big)=\frac{1}{\sigma^2}\cdot\frac{1}{n-p-3}.$$

由于假设$H:\boldsymbol{\beta}_t=\mathbf{0}$可以等价地写成$H:\boldsymbol{A}\boldsymbol{\beta}=\mathbf{0}$, 其中$\boldsymbol{A}=(\mathbf{0},\boldsymbol{I}_t)$, 这里的$\mathbf{0}$是$t\times(q+1)$的零矩阵. 显然, $\mathrm{rk}(\boldsymbol{A})=t$. 同时注意到

$$\hat{\boldsymbol{\beta}}_t\sim N(\boldsymbol{\beta}_t,\sigma^2\boldsymbol{D}_1),\quad \frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\hat{\boldsymbol{\beta}}_t\sim N(\frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\boldsymbol{\beta}_t,\boldsymbol{I}_t).$$

所以由第四章的(4.1.8)及非中心卡方分布的定义可知

$$\frac{\eta}{\sigma^2}=\frac{1}{\sigma^2}\hat{\boldsymbol{\beta}}_t'\boldsymbol{D}_1^{-1}\hat{\boldsymbol{\beta}}_t=(\frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\hat{\boldsymbol{\beta}}_t)'(\frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\hat{\boldsymbol{\beta}}_t)\sim\chi^2(t,\delta),$$

其中, 非中心参数

$$\delta=(\frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\boldsymbol{\beta}_t)'(\frac{1}{\sigma}\boldsymbol{D}_1^{-\frac{1}{2}}\boldsymbol{\beta}_t)=\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}.$$

因此

$$\mathsf{E}(\eta)=\sigma^2\Big(t+\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}\Big).$$

现在, 已可推知(注意$p=q+t$)

$$\begin{aligned}\mathsf{E}(C_p)=&\mathsf{E}\Big(\frac{\mathrm{RSS}_q}{\hat{\sigma}^2}\Big)-[n-2(q+1)]\\=&(n-p-1)\Big[1+\frac{1}{n-p-3}\Big(t+\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}\Big)\Big]-[n-2(q+1)]\\=&q+1-t+\frac{n-p-1}{n-p-3}\Big(t+\frac{\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t}{\sigma^2}\Big).\end{aligned}$$

证毕. □

这个定理说明$C_p$统计量不是

$$\Gamma_q = \frac{1}{\sigma^2}(I_1 + I_2) = q + 1 + \frac{\boldsymbol{\beta}_t' \boldsymbol{D}_1^{-1} \boldsymbol{\beta}_t}{\sigma^2}$$

的无偏估计. 但如果$n-p$较大, 使得

$$\frac{n-p-1}{n-p-3} \approx 1, \tag{5.2.2}$$

则$\mathsf{E}(C_p) \approx \Gamma_q$. 即$C_p$统计量是$\Gamma_q$的渐近无偏估计量. 根据$\Gamma_q$的意义, $\Gamma_q$越小越好, 所以应该选择具有最小$C_p$值的自变量子集.

**推论**5.2.1　在定理5.2.1的条件下, 若$\boldsymbol{\beta}_t = \mathbf{0}$, 则

$$C_p = (q + 1 - t) + tu,$$

或等价地写成

$$C_p - (q+1) = t(u-1),$$

其中$u \sim F(t, n-p-1)$.

**证明**　记

$$u = \frac{\mathrm{RSS}_q - \mathrm{RSS}}{t\hat{\sigma}^2} = \frac{(\mathrm{RSS}_q - \mathrm{RSS})/t}{\mathrm{RSS}/(n-p-1)}.$$

易见$u$为假设$H: \boldsymbol{\beta}_t = \mathbf{0}$的$F$检验统计量. 所以, 若$\boldsymbol{\beta}_t = \mathbf{0}$, 则由最小二乘法基本定理知$u \sim F(t, n-p-1)$. 借助$u$, $C_p$可表示为

$$\begin{aligned} C_p =& \Big(\frac{\mathrm{RSS}_q - \mathrm{RSS}}{t\hat{\sigma}^2} + \frac{\mathrm{RSS}}{t\hat{\sigma}^2}\Big)t - [n - 2(q+1)] \\ =& tu + \frac{(n-p-1)\hat{\sigma}^2}{\hat{\sigma}^2} - [n - 2(q+1)] \\ =& (q + 1 - t) + tu. \end{aligned}$$

证毕. □

来解释一下上述性质如何应用于自变量选择. 若$\boldsymbol{\beta}_t = \mathbf{0}$, 即选模型(5.1.2)是正确的, 那么从定理5.2.1知

$$\mathsf{E}(C_p) = q + 1 - t + \frac{n-p-1}{n-p-3}t,$$

若$n-p$较大使得(5.2.2)成立, 那么有

$$\mathsf{E}(C_p) \approx q + 1.$$

注意$q+1$其实是选模型的设计矩阵的秩. 这说明, 对于正确的选模型, 在平面直角坐标系中, 点$(q+1, C_p)$落在第一象限角平分线附近. 如果选模型不正确, 即$\boldsymbol{\beta}_t \neq \mathbf{0}$, 那么在条件(5.2.2)下有

$$\mathsf{E}(C_p) \approx q + 1 + \frac{\boldsymbol{\beta}_t' \boldsymbol{D}_1^{-1} \boldsymbol{\beta}_t}{\sigma^2} > q + 1,$$

此时点$(q+1, C_p)$将会向第一象限角平分线上方移动.

最后, 关于$C_p$统计量, 可以得到如下的自变量选择准则: 选择使得点$(q+1, C_p)$尽可能接近第一象限角平分线且$C_p$值最小的选模型. 称直角坐标系中$(q+1, C_p)$的散点图为$C_p$图.

(4) AIC准则

极大似然原理是统计学中估计参数的一种重要方法. Akaike把此方法加以修正, 提出了一种较为一般的模型选择准则, 称为Akaike信息量准则(Akaike information criterion, AIC).

对于一般的统计模型, 设$y_1, \cdots, y_n$是因变量的一个样本, 如果它们来自某个含$k$个参数的模型, 对应的似然函数的最大值记为$L_k(y_1, \cdots, y_n)$, 则选择使

$$\ln L_k(y_1, \cdots, y_n) - k \tag{5.2.3}$$

达到最大的模型. 对于线性回归模型, $\ln L_k(y_1, \cdots, y_n)$是$k$的增函数, 所以上式中的$-k$起着惩罚的作用. 下面把此准则应用于回归模型的自变量选择.

在选模型(5.1.2)中, 假设误差向量$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, 则$\boldsymbol{\beta}_q$与$\sigma^2$的似然函数为

$$L(\boldsymbol{\beta}_q, \sigma^2 | \boldsymbol{Y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \|\boldsymbol{Y} - \boldsymbol{X}_q \boldsymbol{\beta}_q\|^2 \Big\}. \tag{5.2.4}$$

容易求得$\boldsymbol{\beta}_q$和$\sigma^2$的极大似然估计为

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_q &= (\boldsymbol{X}_q' \boldsymbol{X}_q)^{-1} \boldsymbol{X}_q' \boldsymbol{Y}, \\ \tilde{\sigma}_q^2 &= \frac{\mathrm{RSS}_q}{n} = \frac{\boldsymbol{Y}'[\boldsymbol{I}_n - \boldsymbol{X}_q (\boldsymbol{X}_q' \boldsymbol{X}_q)^{-1} \boldsymbol{X}_q'] \boldsymbol{Y}}{n}. \end{aligned}$$

代入(5.2.4)得到对数似然函数的最大值

$$\ln L(\tilde{\boldsymbol{\beta}}_q, \tilde{\sigma}_q^2 | \boldsymbol{Y}) = \frac{n}{2} \ln\Big(\frac{n}{2\pi}\Big) - \frac{n}{2} - \frac{n}{2} \ln(\mathrm{RSS}_q).$$

略去与$q$无关的项, 按照(5.2.3)得统计量

$$-\frac{n}{2} \ln(\mathrm{RSS}_q) - (q+1).$$

按AIC准则, 选择自变量子集使上式达到最大. 等价地, 记

$$\mathrm{AIC} = n \ln(\mathrm{RSS}_q) + 2(q+1),$$

则应选择使上式达到最小的自变量子集.

Akaike(1976)和Haman(1979)基于Bayes方法提出了Bayes信息准则BIC:

$$\mathrm{BIC} = n \ln(\mathrm{RSS}_q) + (q+1) \ln n.$$

与AIC相比, BIC的惩罚加强了, 从而在选择变量进入模型上更加谨慎. BIC倾向于选择更简单的线性回归模型, 在大样本情形下, BIC更接近真实模型. 实际上, 在一定的正则条件下, 当样本容量$n \to \infty$时, BIC具有变量选择的相合性. 而对AIC来说, 不管样本容量多大, 它都会倾向于接受过多参数的模型. 在统计学

中, 模型中包含过多参数时称为模型过拟合(overfitting), 而模型中包含过少参数时则称为模型欠拟合(underfitting). 在小样本情形下, 上述的描述不一定正确. 在实践中, 最优的AIC模型往往也接近于最优的BIC模型, 它们常常会给出同一最优模型.

(5) $J_p$统计量准则

利用选模型进行预测, 预测偏差$y_0 - \boldsymbol{x}_{0q}'\widetilde{\boldsymbol{\beta}}_q$的方差为

$$[1 + \boldsymbol{x}_{0q}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{0q}]\sigma^2.$$

因而在$n$个样本点$\{(\boldsymbol{x}_i, \tilde{y}_i),\ i = 1, \cdots, n\}$上(这里, $\tilde{y}_i$与$y_i$独立同分布), 这些预测偏差的方差之和为

$$\begin{aligned}\sum_{i=1}^{n}\mathsf{Var}(\tilde{y}_i - \boldsymbol{x}_{iq}'\widetilde{\boldsymbol{\beta}}_q) =&\sigma^2\sum_{i=1}^{n}[1 + \boldsymbol{x}_{iq}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{iq}]\\ =&n\sigma^2 + \sigma^2\mathrm{tr}[(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_{iq}\boldsymbol{x}_{iq}']\\ =&(n + q + 1)\sigma^2.\end{aligned}$$

由于$\sigma^2$未知, 所以用选模型中$\sigma^2$的估计$\widetilde{\sigma}_q^2$代入就得到

$$J_p = (n + q + 1)\widetilde{\sigma}_q^2 = \frac{n + q + 1}{n - q - 1}\mathrm{RSS}_q.$$

这里的$(n + q + 1)/(n - q - 1)$起着惩罚的作用. 我们选择使$J_p$达到最小的自变量子集.

(6) 预测残差平方和$\mathrm{PRESS}_q$(predicted residual sum of squares)准则

为了给出PRESS的定义和表达式, 我们略去$q$, 对全模型作推导. 考虑在建立回归方程时略去第$i$组数据, 此时记

$$\boldsymbol{Y}_{(i)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix},\quad \boldsymbol{X}_{(i)} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_{i-1}' \\ \boldsymbol{x}_{i+1}' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix},\quad \boldsymbol{e}_{(i)} = \begin{pmatrix} e_1 \\ \vdots \\ e_{i-1} \\ e_{i+1} \\ \vdots \\ e_n \end{pmatrix}.$$

相应的模型为

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{e}_{(i)}.$$

此时$\boldsymbol{\beta}$的最小二乘估计为

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

用$\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$去预测$y_i$, 预测偏差记为$\hat{e}_{(i)}$, 即

$$\hat{e}_{(i)} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(i)}.$$

定义预测残差平方和为

$$\text{PRESS} = \sum_{i=1}^{n} [\hat{e}_{(i)}]^2.$$

第三章已证明

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

所以

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

这里的$\hat{e}_i$是全数据情形下的第$i$个残差, $h_{ii}$是全数据情形下帽子矩阵的第$i$个对角线元素. 因此

$$\text{PRESS} = \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1-h_{ii})^2}.$$

若要计算$\text{PRESS}_q$, 只要将$\hat{e}_i$换成全数据情形下选模型的第$i$个残差, $h_{ii}$换成$\boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'$(选模型的帽子矩阵)的第$i$个对角线元素即可. 我们选择使得$\text{PRESS}_q$达到最小的自变量子集.

**例**5.2.1(Hald水泥问题)　下面这组数据来自Hald的著作《Statistical Theory with Engineering Application》(1952). 问题是考察含有如下四种化学成分

$x_1$ : $3CaO \cdot Al_2O_3$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(%)

的水泥, 寻找每一克所释放出的热量$y$与这四种成分含量之间的关系.

| 序号 | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|
| 1 | 7 | 26 | 6 | 60 | 78.5 |
| 2 | 1 | 29 | 15 | 52 | 74.3 |
| 3 | 11 | 56 | 8 | 20 | 104.3 |
| 4 | 11 | 31 | 8 | 47 | 87.6 |
| 5 | 7 | 52 | 6 | 33 | 95.9 |
| 6 | 11 | 55 | 9 | 22 | 109.2 |
| 7 | 3 | 71 | 17 | 6 | 102.7 |
| 8 | 1 | 31 | 22 | 44 | 72.5 |
| 9 | 2 | 54 | 18 | 22 | 93.1 |
| 10 | 21 | 47 | 4 | 26 | 115.9 |
| 11 | 1 | 40 | 23 | 34 | 83.8 |
| 12 | 11 | 66 | 9 | 12 | 113.3 |
| 13 | 10 | 68 | 8 | 12 | 109.4 |

表5.2.1　Hald水泥数据

先考虑用调整后的$R^2$进行自变量选择, R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x1=yx[,1]
> x2=yx[,2]
> x3=yx[,3]
> x4=yx[,4]
> y=yx[,5]
> cement=data.frame(x1,x2,x3,x4,y)
> X=matrix(c(x1,x2,x3,x4),nrow=13,byrow=F)
> library(faraway)
> library(leaps)
> adjr=leaps(X,y,int=T,method="adjr2")
> adjr
$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE
$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
$adjr2
 [1] 0.6449549 0.6359290 0.4915797 0.2209521 0.9744140 0.9669653
 [7] 0.9223476 0.8164305 0.6160725 0.4578001 0.9764473 0.9763796
[13] 0.9750415 0.9637599 0.9735634
> adjr$which[which.max(adjr$adjr2),]
    1     2     3     4
 TRUE  TRUE FALSE  TRUE
```

可以看到, 共有$2^4 - 1 = 15$个自变量子集, 其中, 使调整后的$R^2$达到最大的自变量子集是$\{x_1, x_2, x_4\}$.

接下来考虑用$C_p$统计量进行自变量的选择, R代码及分析结果如下:

```
> cp=leaps(X,y,int=T,method="Cp")
> cp
```

```
$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE
$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
$Cp
 [1] 138.730833 142.486407 202.548769 315.154284   2.678242   5.495851
 [7]  22.373112  62.437716 138.225920 198.094653   3.018233   3.041280
[13]   3.496824   7.337474   5.000000
> cp$which[which.min(cp$Cp),]
    1     2     3     4
 TRUE  TRUE FALSE FALSE
> Cpplot(cp)  # 画Cp图
```

可以看到, 使$C_p$统计量达到最小的自变量子集是$\{x_1, x_2\}$. 画出$C_p$图, 见图5.2.1. 有些模型的$C_p$没有显示在图中, 这是因为它们的$C_p$值太大. 若选择最接近第一象限角平分线且$C_p$值最小的选模型, 则仍选择自变量子集$\{x_1, x_2\}$.

程序包leaps中的参数method只有三个选项: Cp, adjr2, r2. 若使用其它准则选择自变量, 编写R代码如下:

```
> search.results=regsubsets(y~x1+x2+x3+x4,data=cement,method="exhaustive"
  ,nbest=15)
> selection.criteria=summary(search.results)
> selection.criteria
Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement, method = "
  exhaustive", nbest = 15)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
```

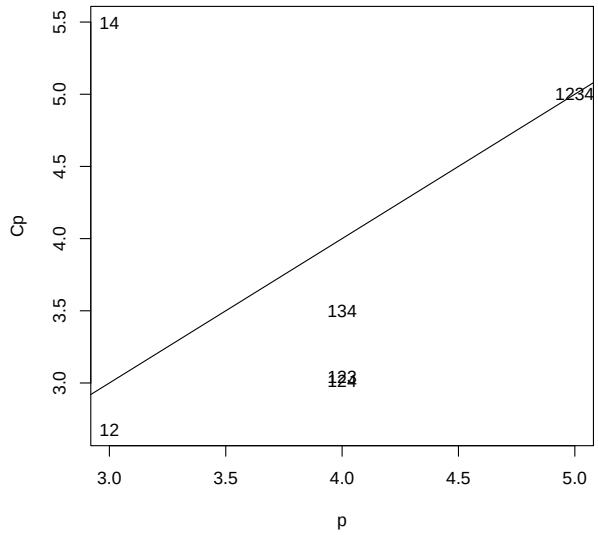


图5.2.1　$C_p$图

```
15 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
1  ( 2 ) " " "*" " " " "
1  ( 3 ) "*" " " " " " "
1  ( 4 ) " " " " "*" " "
2  ( 1 ) "*" "*" " " " "
2  ( 2 ) "*" " " " " "*"
2  ( 3 ) " " " " "*" "*"
2  ( 4 ) " " "*" "*" " "
2  ( 5 ) " " "*" " " "*"
2  ( 6 ) "*" " " "*" " "
3  ( 1 ) "*" "*" " " "*"
3  ( 2 ) "*" "*" "*" " "
3  ( 3 ) "*" " " "*" "*"
3  ( 4 ) " " "*" "*" "*"
4  ( 1 ) "*" "*" "*" "*"
```

summary(search.results)给出了所有15种自变量选择的情况. 其中, 只选择1个自变量进入模型的情况有4种, 只选择2个自变量进入模型的情况有6种, 选择3个自变量进入模型的情况有4 种, 选择4个自变量进入模型的情况有1种.

```
> names(selection.criteria)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
> selection.criteria$which
   (Intercept)    x1    x2    x3    x4
1         TRUE FALSE FALSE FALSE  TRUE
1         TRUE FALSE  TRUE FALSE FALSE
1         TRUE  TRUE FALSE FALSE FALSE
1         TRUE FALSE FALSE  TRUE FALSE
2         TRUE  TRUE  TRUE FALSE FALSE
2         TRUE  TRUE FALSE FALSE  TRUE
2         TRUE FALSE FALSE  TRUE  TRUE
2         TRUE FALSE  TRUE  TRUE FALSE
2         TRUE FALSE  TRUE FALSE  TRUE
2         TRUE  TRUE FALSE  TRUE FALSE
3         TRUE  TRUE  TRUE FALSE  TRUE
3         TRUE  TRUE  TRUE  TRUE FALSE
3         TRUE  TRUE FALSE  TRUE  TRUE
3         TRUE FALSE  TRUE  TRUE  TRUE
4         TRUE  TRUE  TRUE  TRUE  TRUE
> n=length(cement[,1])
> q=as.integer(row.names(selection.criteria$which))
> R.sq=selection.criteria$rsq
> AdjR.sq=selection.criteria$adjr2
> rms=selection.criteria$rss/(n-q-1)
> Cp=selection.criteria$cp
> Jp=selection.criteria$rss*(n+q+1)/(n-q-1)
> aic.f=n*log(selection.criteria$rss)+2*(q+1)
> bic.f=n*log(selection.criteria$rss)+(q+1)*log(n)
> var=as.matrix(selection.criteria$which[,2:5])
> criteria.table=data.frame(cbind(q,rms,R.sq,AdjR.sq,Cp,Jp,aic.f,bic.f,
  var[,1],var[,2],var[,3],var[,4]),row.names=NULL)
> names(criteria.table)=c("q","RMS","Rsq","aRsq","Cp","Jp","AIC","BIC","
  x1","x2","x3","x4")
> round(criteria.table,2)  #保留2位小数
    q    RMS  Rsq aRsq     Cp      Jp    AIC    BIC x1 x2 x3 x4
1   1  80.35 0.67 0.64 138.73 1205.27  92.20  93.33  0  0  0  1
2   1  82.39 0.67 0.64 142.49 1235.91  92.52  93.65  0  1  0  0
3   1 115.06 0.53 0.49 202.55 1725.94  96.86  97.99  1  0  0  0
4   1 176.31 0.29 0.22 315.15 2644.64 102.41 103.54  0  0  1  0
5   2   5.79 0.98 0.97   2.68   92.65  58.76  60.46  1  1  0  0
6   2   7.48 0.97 0.97   5.50  119.62  62.09  63.78  1  0  0  1
7   2  17.57 0.94 0.92  22.37  281.18  73.20  74.89  0  0  1  1
8   2  41.54 0.85 0.82  62.44  664.71  84.38  86.08  0  1  1  0
9   2  86.89 0.68 0.62 138.23 1390.21  93.97  95.67  0  1  0  1
10  2 122.71 0.55 0.46 198.09 1963.32  98.46 100.16  1  0  1  0
11  3   5.33 0.98 0.98   3.02   90.62  58.32  60.58  1  1  0  1
12  3   5.35 0.98 0.98   3.04   90.88  58.36  60.62  1  1  1  0
13  3   5.65 0.98 0.98   3.50   96.02  59.07  61.33  1  0  1  1
14  3   8.20 0.97 0.96   7.34  139.43  63.92  66.18  0  1  1  1
15  4   5.98 0.98 0.97   5.00  107.69  60.29  63.11  1  1  1  1
```

在上述代码框的最后部分, 给出了每一个选模型所包含的自变量个数、平均残差平方和、$R^2$、调整后的$R^2$、$C_p$统计量、$J_p$统计量、AIC、BIC, 以及入选模型的自变量子集. 可据此结果进行自变量选择, 例如, 若用AIC进行自变量选择, 应选择自变量子集$\{x_1, x_2\}$.

若用PRESS统计量进行自变量选择, 则需使用程序包DAAG中的press命令, R代码及分析结果如下:

```
> lm.sol=lm(y~.,data=cement)
> library(DAAG)
> press(lm.sol)
[1] 110.3466
```

可以看出, 对于所有4个自变量都进入模型的情形, PRESS值等于110.3466. press命令只能对一个指定的模型给出PRESS值, 若要逐一计算15个选模型的PRESS值, 那就有点麻烦了. 如果自变量个数更多, 那就更糟糕了. 一种方法是自己编写函数, 计算所有选模型的PRESS值. 例如, 函数multipress的R代码及分析结果如下:

```
> multipress<- function(x, y)
>  {
>  nvar = length(x)
>  combColnames = sapply(1:nvar, function(i) combn(colnames(x), i))
>  df = data.frame(x,y)
>  mods = c()
>  for (i in c(1:nvar))
>  {
>  if (length(colnames(y)) == 0)
>  tmp = 'y~'
>  else
>  tmp = paste0(colnames(y),'~')
>  for (j in c(1:i))
>  {
>  if (j==1)
>  tmp = paste0(tmp, combColnames[[i]][j,])
>  else
>  tmp = paste0(tmp,'+',combColnames[[i]][j,])
>  }
>  mods = c(mods, tmp)
>  }
>  P = sapply(1:length(mods), function(x) press(lm(mods[x], df)))
>  return(data.frame(mods, P))
>  }
> X=data.frame(x1,x2,x3,x4)
> library(DAAG)
> multipress(X,y)
           mods          P
```

```
1           y~x1 1699.61160
2           y~x2 1202.08675
3           y~x3 2616.36385
4           y~x4 1194.21820
5        y~x1+x2   93.88255
6        y~x1+x3 2218.11831
7        y~x1+x4  121.22439
8        y~x2+x3  701.74318
9        y~x2+x4 1461.81421
10       y~x3+x4  294.01387
11    y~x1+x2+x3   90.00001
12    y~x1+x2+x4   85.35112
13    y~x1+x3+x4   94.53706
14    y~x2+x3+x4  146.85269
15 y~x1+x2+x3+x4  110.34656
```

可以看出, 若用PRESS进行自变量的选择, 则应选择自变量子集$\{x_1, x_2, x_4\}$.

此外, 程序包leaps中的regsubsets命令有图示法的变量选择功能:

```
> library(leaps)
> subsets=regsubsets(y~.,data=cement)
> summary(subsets)
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
2  ( 1 ) "*" "*" " " " "
3  ( 1 ) "*" "*" " " "*"
4  ( 1 ) "*" "*" "*" "*"
> plot(subsets)
> plot(subsets,scale="Cp")
> plot(subsets,scale="adjr2")
```

这个结果告诉我们: 若要求模型中只包含1个自变量, 则应选择$\{x_4\}$; 若要求模型中只包含2个自变量, 则应选择$\{x_1, x_2\}$; 若要求模型中只包含3个自变量, 则应选择$\{x_1, x_2, x_4\}$; 若要求模型中包含4个自变量, 则应选择$\{x_1, x_2, x_3, x_4\}$. plot(subsets)默认采用BIC画图, 见图5.2.2, 根据BIC越小越好的准则, 应选择自变量子集$\{x_1, x_2\}$. plot(subsets,scale="Cp") 采用$C_p$画图, 见图5.2.3, 根据$C_p$统计量越小越好的准则, 应选择自变量子集$\{x_1, x_2\}$. plot(subsets,scale="adjr2") 采用调整后的$R^2$画图, 见图5.2.4, 根据调整后的$R^2$越大越好的准则, 应选择自变量子集$\{x_1, x_2, x_4\}$.

**注** plot(subsets, scale="xx")可显示变量选择示意图, 其中"xx" 可以是"Cp", "adjr2", "r2"或"bic", 默认是"bic".

## 5.3 基于检验的自变量选择

多元线性回归分析中, $p$个自变量的所有可能子集构成$2^p - 1$个线性回归模型. 当

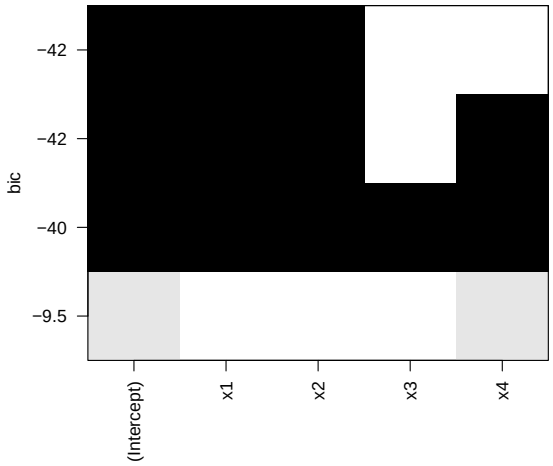


图5.2.2　BIC

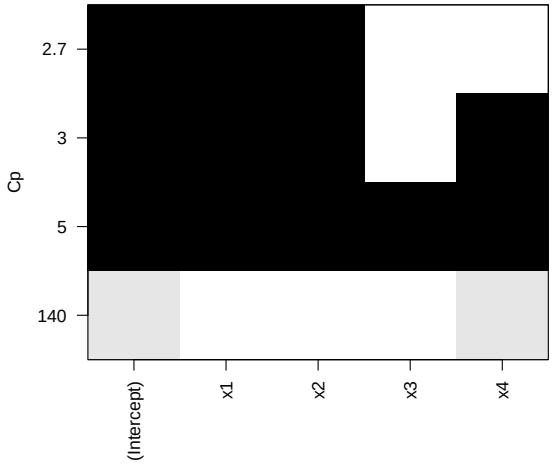


图5.2.3　$C_p$

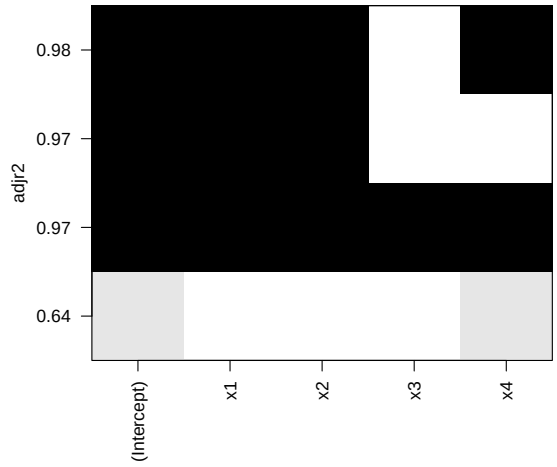


图5.2.4　调整后的$R^2$

可供选择的自变量个数不太多时, 用前面介绍的自变量选择准则可挑选出“最优”的线性回归模型. 但是当自变量的个数较多时, 以上这些方法就不大实用了. 为此, 人们提出了一些较为简便、实用、快捷的选择“最优”的线性回归模型的方法. 这些方法各有优缺点, 没有绝对最优的方法, 目前常用的方法有向前法、向后法和逐步回归法.

(1) 向前法(forward)

向前法的思想是引入回归模型中的自变量个数由少到多, 每次增加一个, 直到没有可引入的自变量为止. 具体做法是:

第一步: 因变量$y$关于每个自变量$x_1, \cdots, x_p$分别建立一元线性回归模型, 因此得到$p$个回归模型. 分别计算这$p$个一元线性回归模型的回归系数的$F$检验值, 记为$F_1^{(1)}, \cdots, F_p^{(1)}$. 选其最大者, 记为

$$F_j^{(1)} = \max\{F_1^{(1)}, \cdots, F_p^{(1)}\}.$$

给定显著性水平$\alpha$, 若$F_j^{(1)} > F_\alpha(1, n-2)$, 则首先将$x_j$引入回归模型. 不妨假设引入的$x_j$就是$x_1$.

第二步: 将因变量$y$分别与$\{x_1, x_2\}, \{x_1, x_3\}, \cdots, \{x_1, x_p\}$建立$p-1$个二元线性回归模型, 对这$p-1$个二元线性回归模型中$x_2, \cdots, x_p$的回归系数进行$F$检验, 得到$F$检验值, 记为$F_2^{(2)}, \cdots, F_p^{(2)}$. 选其最大者, 记为

$$F_j^{(2)} = \max\{F_1^{(2)}, \cdots, F_p^{(2)}\}.$$

若$F_j^{(2)} > F_\alpha(1, n-3)$, 则将$x_j$引入回归模型. 不妨假设引入的$x_j$就是$x_2$.

第三步: 继续以上做法, 假设已确定引入$q$个自变量: $x_1, \cdots, x_q$. 在建立$q+1$元线性回归模型时, 若所有的$x_{q+1}, \cdots, x_p$的$F$检验值均不大于$F_\alpha(1, n-q-2)$, 则变量选择结束. 这时得到的$q$元线性回归模型就是最终确定的回归模型.

向前法的缺点是: 不能反映引入新变量后的变化情况. 因为某个自变量可能刚开始时是显著的, 当引入其它自变量后它就变得不显著了, 但是没有机会将其剔除. 即一旦引入, 就是“终身制”的.

(2) 向后法(backward)

向后法与向前法恰恰相反, 向后法的思想是引入模型的自变量个数由多到少, 每次剔除一个, 直到没有可剔除的自变量为止. 具体做法是:

第一步: 因变量$y$关于所有的自变量$x_1, \cdots, x_p$建立一个$p$元线性回归模型, 分别计算$p$个回归系数的$F$检验值, 记为$F_1^{(p)}, \cdots, F_p^{(p)}$. 选其最小者, 记为

$$F_j^{(p)} = \min\{F_1^{(p)}, \cdots, F_p^{(p)}\}.$$

给定显著性水平$\beta$, 若$F_j^{(p)} \le F_\beta(1, n-p-1)$, 则首先将$x_j$从回归模型中剔除. 不妨假设$x_j$就是$x_p$.

第二步: 因变量$y$关于自变量$x_1, \cdots, x_{p-1}$建立一个$p-1$元的线性回归模型, 分别计算$p-1$个回归系数的$F$检验值, 记为$F_1^{(p-1)}, \cdots, F_{p-1}^{(p-1)}$. 选其最小者, 记为

$$F_j^{(p-1)} = \min\{F_1^{(p-1)}, \cdots, F_{p-1}^{(p-1)}\}.$$

若$F_j^{(p-1)} \le F_\beta(1, n-p)$, 则将$x_j$从回归模型中剔除. 不妨假设$x_j$就是$x_{p-1}$.

第三步: 继续以上做法, 假设已确定剔除了$p-q$个自变量: $x_{q+1}, \cdots, x_p$. 在$y$关于$x_1, \cdots, x_q$的$q$元线性回归模型中, 若所有的$x_1, \cdots, x_q$的$F$检验值均大于$F_\beta(1, n-q-1)$, 则变量选择结束. 这时得到的$q$元线性回归模型就是最终确定的回归模型.

向后法的缺点是: 一开始就把所有自变量引入回归模型, 这样的计算量很大. 另外, 自变量一旦被剔除, 就永远没有机会再重新进入回归模型了, 即是“一棒子打死”的.

(3) 逐步回归法(stepwise)

逐步回归法的基本思想是模型中的自变量可进可出. 具体做法是: 将自变量一个一个地引入回归模型, 每引入一个自变量后, 都要对已引入的自变量逐个进行检验, 当先引入的自变量由于后引入的自变量而变得不再显著时, 就要将其剔除. 将这个过程反复进行下去, 直到既无显著的自变量可引入回归模型, 也无不显著的自变量可从回归模型中剔除为止. 这样就避免了向前法和向后法各自的缺点, 保证最后得到的自变量子集是“最优”的.

应用逐步回归时需注意: 引入自变量与剔除自变量时所选用的显著性水平应是不同的, 要求引入自变量时所使用的显著性水平$\alpha$小于剔除自变量时所使用的显著性水平$\beta$, 否则就可能产生死循环. 也就是说, 当$\alpha \ge \beta$时, 如果某个自变量的$p$值在$\alpha$与$\beta$之间, 那么这个自变量将被引入、剔除、再引入、再剔除, $\cdots$, 循环往复以至无穷.

下面用向后法对例5.2.1的数据进行变量选择, 取$\beta = 0.1$. R代码及分析结果如下:

```
> yx=read.table("cement.txt")
> x1=yx[,1]
```

```
> x2=yx[,2]
> x3=yx[,3]
> x4=yx[,4]
> y=yx[,5]
> cement=data.frame(x1,x2,x3,x4,y)
> lm.sol=lm(y~.,data=cement)
> summary(lm.sol)
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)
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
```

可以发现$x_3$的$p$值大于$\beta = 0.1$, 且它是最不显著的自变量(因为它的$p$值最大), 所以删除$x_3$, 让$y$关于$\{x_1, x_2, x_4\}$进行回归建模:

```
> lm.sol=update(lm.sol,.~.-x3)
> summary(lm.sol)
Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)
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
```

可以发现只有$x_4$的$p$值大于$\beta = 0.1$, 所以删除$x_4$, 让$y$关于$\{x_1, x_2\}$进行回归建模:

```
> lm.sol=update(lm.sol,.~.-x4)
> summary(lm.sol)
Call:
lm(formula = y ~ x1 + x2, data = cement)
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 ***
```

此时发现剩下的$x_1$和$x_2$的$p$值都小于$\beta = 0.1$, 所以向后法选择出来的自变量子集是$\{x_1, x_2\}$.

在R软件中, 基于$p$值或$F$值的向前法、向后法和逐步回归法需要不断地进行人工判断和操作, 非常不方便. step命令可以自动进行向前法、向后法和逐步回

归法的变量选择, 但它是基于AIC准则的. 下面用step命令重新对例5.2.1进行变量选择. 应用向前法的代码及分析结果如下:

```
> min.model=lm(y~1,data=cement)
> fwd.model=step(min.model,direction="forward",scope=(~x1+x2+x3+x4))
> summary(fwd.model)
Start:  AIC=71.44
y ~ 1
       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4
       Df Sum of Sq    RSS    AIC
+ x1    1    809.10  74.76 28.742
+ x3    1    708.13 175.74 39.853
<none>              883.87 58.852
+ x2    1     14.99 868.88 60.629

Step:  AIC=28.74
y ~ x4 + x1
       Df Sum of Sq    RSS    AIC
+ x2    1    26.789 47.973 24.974
+ x3    1    23.926 50.836 25.728
<none>              74.762 28.742

Step:  AIC=24.97
y ~ x4 + x1 + x2
       Df Sum of Sq    RSS    AIC
<none>              47.973 24.974
+ x3    1   0.10909 47.864 26.944

> summary(fwd.model)
Call:
lm(formula = y ~ x4 + x1 + x2, data = cement)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x4           -0.2365     0.1733  -1.365 0.205395
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
```

可以看出, 向前法的变量选择结果是选入$\{x_1, x_2, x_4\}$. 应用向后法的代码及分析结果如下:

```
> max.model=lm(y~.,data=cement)
> bwd.model=step(max.model,direction="backward")
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

> summary(bwd.model)
Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)
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
```

可以看出, 向后法的变量选择结果也是选入$\{x_1, x_2, x_4\}$. 应用逐步回归法的代码及分析结果如下:

```
> min.model=lm(y~1,data=cement)
> step.model=step(min.model,direction="both",scope=(~x1+x2+x3+x4))
Start:  AIC=71.44
y ~ 1
       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4
       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
<none>               883.87 58.852
```

```
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444

Step:  AIC=28.74
y ~ x4 + x1
       Df Sum of Sq     RSS    AIC
+ x2    1     26.79   47.97 24.974
+ x3    1     23.93   50.84 25.728
<none>                74.76 28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190.92 1265.69 63.519

Step:  AIC=24.97
y ~ x4 + x1 + x2
       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

> summary(step.model)
Call:
lm(formula = y ~ x4 + x1 + x2, data = cement)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x4           -0.2365     0.1733  -1.365 0.205395
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
```

可以看出, 逐步回归法的变量选择结果仍是选入$\{x_1, x_2, x_4\}$.

## 5.4 基于惩罚的自变量选择

假设自变量与因变量均已标准化, 线性回归模型的设计矩阵$\boldsymbol{X}$是$n \times p$矩阵. 前面介绍的变量选择方法都要用到最小二乘法, 它要求自变量个数$p$小于样本容量$n$. 若$p \geq n$, 则可以应用基于惩罚的变量选择方法. 注意当$p = n$时, 普通最小二乘法可以估计出回归系数, 但没有多余的自由度可以估计其它参数(例如$\sigma^2$)或做假设检验等统计推断.

第三章的岭估计是把$\boldsymbol{\beta}$的$\ell_2$范数作为惩罚(惩罚项为$\|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^p \beta_i^2$), 这种方法可以将回归系数往原点的方向进行压缩, 但不会把任何一个回归系数压缩到0. 因此, 岭回归给出的模型无法进行自变量的选择. 若把$\boldsymbol{\beta}$的$\ell_1$范数作为惩罚(惩罚项为$\|\boldsymbol{\beta}\|_1^1 = \sum_{i=1}^p |\beta_i|$), 则得到$\boldsymbol{\beta}$的LASSO(least absolute shrinkage and selection operator)估计, 此时的优化问题为

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^p |\beta_i| \right\}, \quad \lambda \geq 0. \tag{5.4.1}$$

(5.4.1)等价于

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \ \text{ s.t. } \ \sum_{i=1}^{p}|\beta_i| \le t, \ \ t\ge 0. \tag{5.4.2}$$

该方法由Tibshirani(1996)提出. 与岭估计不同, 在一般情形下LASSO估计没有解析表达式. 但LASSO方法的优点是它能把某些$\beta_i$的估计取为0, 这是岭估计无法做到的. 所以LASSO方法能用来估计稀疏模型(即绝大部分的自变量对因变量的影响为0或近似为0的模型)的回归系数并达到变量选择的目的.

由于没有高效的算法, LASSO方法在问世后并没有被推广开来. 直到2004年, Efron, Hastie, Johnstone和Tibshirani给出了基于最小角度回归(least angle regression, LAR)的LASSO快速求解算法, 该方法可以非常有效地找到LASSO的解. 2010年, Friedman提出了基于坐标下降的快速求解算法, 更进一步提高了LASSO的算法效率.

此外, 统计学家和数学家把LASSO的思想应用到了很多领域. 其中最引人注目的就是Candes和陶哲轩将LASSO的思想应用到信号处理领域, 开创了一个新的研究方向: 压缩感知(compressive sensing). 这些研究工作的成功, 使得人们在大规模数据和高维数据问题中的回归分析取得了重要的进展.

LASSO方法为何能把某些$\beta_i$的估计取为0呢? 以二维为例, 来了解一下其中的原理. 因为

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

所以优化问题(5.4.2)中的目标函数$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$可替换为$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$, 而后者所表示的图像(即等高线)是一个中心在$\hat{\boldsymbol{\beta}}$的椭圆. 此外, 优化问题(5.4.2)中的约束条件$\sum_{i=1}^{p}|\beta_i|\le t$所表示的图像是一个正方形($p>2$时, 该约束条件对应一个多面体, 这个多面体的顶点落在坐标轴上), 见图5.4.1. 当$t$很小时, 正方形与$\hat{\boldsymbol{\beta}}$的等高线不相交; 当$t$变大时, 它终将与$\hat{\boldsymbol{\beta}}$的等高线相交, 该交点是正方形的某一顶点, 它就是LASSO估计.

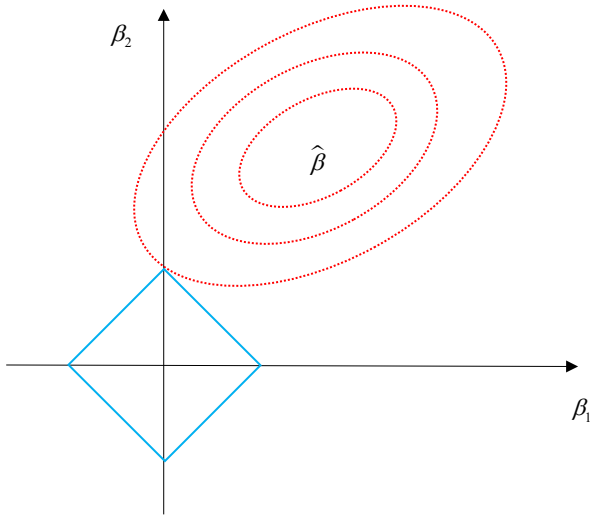


图5.4.1 LASSO估计的几何意义

需要注意的是, $t = 0$时, 所有的自变量都不会进入模型; 随着$t$的变大, 逐渐有自变量进入模型, $t$越大, 进入模型的自变量就会越多, 且回归系数估计值的绝对值也会越大; 当$t = \infty$时, 约束条件$\sum_{i=1}^{p} |\beta_i| \leq t$就是多余的了, 这时LASSO估计就是最小二乘估计. 为了达到自变量选择的目的, 通常选择较小的$t$值(或(5.4.1)中的$\lambda$值). 在实际问题中, 为了更客观地选择$t$值或$\lambda$值, 可以应用交叉验证(cross-validation)的方法.

**例**5.4.1 考虑R中的state.x77数据集, 该数据集收集了上个世纪六七十年代美国50个州的预期寿命以及与此可能相关的其它7个变量的数据. 表5.4.1给出了该数据集的前6组数据.

| | Population | Income | Illiteracy | Life.Exp | Murder | HS.Grad | Frost | Area |
|---|---|---|---|---|---|---|---|---|
| Alabama | 3615 | 3624 | 2.1 | 69.05 | 15.1 | 41.3 | 20 | 50708 |
| Alaska | 365 | 6315 | 1.5 | 69.31 | 11.3 | 66.7 | 152 | 566432 |
| Arizona | 2212 | 4530 | 1.8 | 70.55 | 7.8 | 58.1 | 15 | 113417 |
| Arkansas | 2110 | 3378 | 1.9 | 70.66 | 10.1 | 39.9 | 65 | 51945 |
| California | 21198 | 5114 | 1.1 | 71.71 | 10.3 | 62.6 | 20 | 156361 |
| Colorado | 2541 | 4884 | 0.7 | 72.06 | 6.8 | 63.9 | 166 | 103766 |

表5.4.1 state.x77数据集的前6组数据

来解释一下该数据集中的这几个变量的含义:

Population: 人口估计数(截至1975年7月1日);

Income: 人均收入(1974);

Illiteracy: 文盲人口的百分比(1970);

Life.Exp: 预期寿命(1969-1971);
Murder: 谋杀和非过失杀人率(1976);
HS.Grad: 高中毕业生的百分比(1970);
Frost: 首府或大城市的最低气温低于零度的平均天数(1931-1960);
Area: 土地面积(平方英里).

下面用程序包lars中的lars命令进行变量选择和LASSO估计. R代码及分析结果如下:

```
> library(lars)
> statedata=data.frame(state.x77,row.names=state.abb)  #行名为州的缩写
> head(statedata)
   Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
AK        365   6315        1.5    69.31   11.3    66.7   152 566432
AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
CO       2541   4884        0.7    72.06    6.8    63.9   166 103766
> lasso.sol=lars(as.matrix(statedata[,-4]),statedata$Life.Exp)
> plot(lasso.sol)
```

plot(lasso.sol)得到的图像见图5.4.2. 此图的横坐标$s = |\text{beta}|/\max|\text{beta}|$表示LASSO估计的$\ell_1$范数与最小二乘估计的$\ell_1$范数的比值. $s$其实与$t$有关, 它是$t$的单调增函数. 当$t = 0$时, $s = |\text{beta}|/\max|\text{beta}| = 0$; 当$t = \infty$时, $s = |\text{beta}|/\max|\text{beta}| = 1$. 从图中可以看出, $s = 0$(或$t = 0$)时, 所有自变量都没有进入模型; 随着$s$(或$t$)离开0, 第4个自变量Murder开始进入模型; 随着$s$(或$t$)继续变大, 第5个自变量HS.Grad进入模型; $s$(或$t$)越大, 进入模型的自变量越多, 且回归系数估计值的绝对值也变得越大.

接下来, 用交叉验证的方法来选择$s$值. 在程序包lars中, 默认使用10折交叉验证(即把样本随机等分成10份, 预留1份作为验证集, 剩下的9份用来回归建模, 然后计算回归方程在验证集上的预测均方误差. 遍历所有的验证集后, 再计算10个预测均方误差的平均值). 图5.4.3描述了10折交叉验证的整个过程.

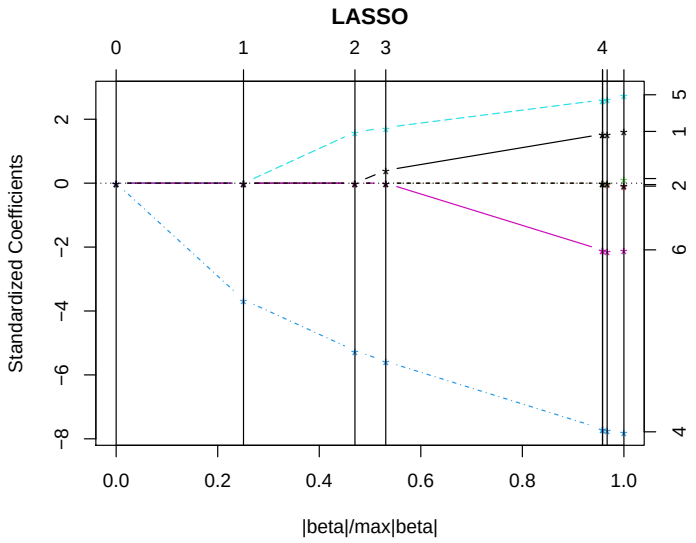


图5.4.2　LASSO估计结果

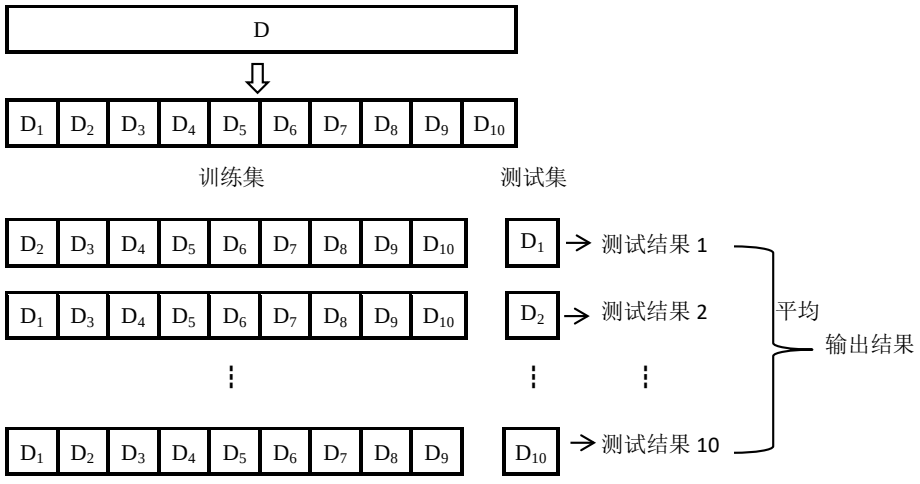


图5.4.3　10折交叉验证

R代码及分析结果如下:

```
> set.seed(123)
> cv.sol=cv.lars(as.matrix(statedata[,-4]),statedata$Life.Exp)  #
```

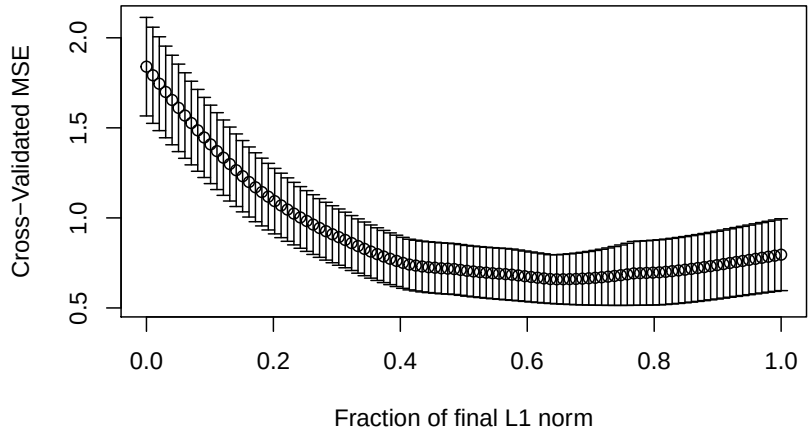


图5.4.4　交叉验证的均方误差

```
  画CV-MSE图
> cv.sol$index[which.min(cv.sol$cv)]
[1] 0.6464646
```

代码框中的cv.sol$index表示$s$, cv.sol$cv表示平均预测均方误差. 由命令cv.lars得到的交叉验证均方误差图见图5.4.4, 当$s = 0.6464646$时预测均方误差达到最小. 来了解一下当$s = 0.6464646$时, 是哪些自变量进入了模型呢?

```
> predict(lasso.sol,s=0.6464646,type="coef",mode="fraction")$coef
   Population        Income    Illiteracy        Murder       HS.Grad
 2.259631e-05  0.000000e+00  0.000000e+00 -2.379447e-01  3.492634e-02
        Frost          Area
-1.558616e-03  0.000000e+00
```

可以发现, 此时LASSO选择了Population, Murder, HS.Grad和Frost这四个自变量.

LASSO能完成变量选择的任务, 所以与岭回归相比, LASSO所建立的模型更具可解释性. 但岭估计有解析表达式, 而LASSO在一般情形下是没有解析表达式的. 下面说明: 在特殊情形下, LASSO也拥有解析表达式.

考虑正交设计情形: $\boldsymbol{X}$是$p \times p$矩阵, 且$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$. 此时$\boldsymbol{\beta}$的LSE为

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

记$\boldsymbol{\beta}$的LASSO估计为$\hat{\boldsymbol{\beta}}^{lasso} = (\hat{\beta}_1^{lasso}, \cdots, \hat{\beta}_p^{lasso})'$. 注意到LASSO的优化函数不

是处处可导的, 所以在给出$\hat{\boldsymbol{\beta}}^{lasso}$的解析表达式之前, 先介绍与次梯度有关的定义和一个引理.

**定义**5.4.1　考虑凸函数$f:\mathbb{R}^p\mapsto\mathbb{R}$. 对$\boldsymbol{x}\in\mathbb{R}^p$, 若向量$\boldsymbol{d}\in\mathbb{R}^p$满足

$$f(\boldsymbol{y})-f(\boldsymbol{x})\geq\boldsymbol{d}'(\boldsymbol{y}-\boldsymbol{x}),\quad\forall\boldsymbol{y}\in\mathbb{R}^p,$$

则称$\boldsymbol{d}$是$f$在$\boldsymbol{x}$处的一个次梯度(sub-gradient). 在$\boldsymbol{x}$处的所有次梯度所组成的集合称为$f$在$\boldsymbol{x}$处的次微分, 记为$\partial f(\boldsymbol{x})$.

考虑凸函数$f:\mathbb{R}\mapsto\mathbb{R}$, 记$f$在$x_0$处的左导数为

$$a=\lim_{x\to x_0-0}\frac{f(x)-f(x_0)}{x-x_0},$$

$f$在$x_0$处的右导数为

$$b=\lim_{x\to x_0+0}\frac{f(x)-f(x_0)}{x-x_0},$$

则$f$在$x_0$处的次微分为闭区间$[a,b]$. 特别地, 若$f$在$x_0$处可导, 则$f$在$x_0$处的次梯度是唯一的, 它就是$f$在$x_0$处的梯度$\nabla f(x_0)$.

易知, 若$f(x)=|x|$, 那么$f$在0点的次微分为闭区间$[-1,1]$.

**引理**5.4.1　$\boldsymbol{x}$是凸函数$f$的全局极小值点当且仅当$\mathbf{0}\in\partial f(\boldsymbol{x})$.

这是凸优化里的一个结论, 见Bertsekas(2016).

**定理**5.4.1　在正交设计情形下,

$$\hat{\beta}_i^{lasso}=\text{sign}(\hat{\beta}_i)(|\hat{\beta}_i|-\frac{\lambda}{2})^+,\quad i=1,\cdots,p,$$

其中, $(|\hat{\beta}_i|-\frac{\lambda}{2})^+$表示$(|\hat{\beta}_i|-\frac{\lambda}{2})$的正部.

**证明**　回忆LASSO的优化函数为

$$Q(\boldsymbol{\beta})=\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda\sum_{i=1}^p|\beta_i|.$$

此外,

$$\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_p,\quad\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{Y}.$$

因此,

$$\begin{aligned}Q(\boldsymbol{\beta})=&(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\sum_{i=1}^p|\beta_i|\\=&\boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+\lambda\sum_{i=1}^p|\beta_i|\\=&\boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{\beta}'\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}'\boldsymbol{\beta}+\lambda\sum_{i=1}^p|\beta_i|\\=&\boldsymbol{Y}'\boldsymbol{Y}-2\sum_{i=1}^p\hat{\beta}_i\beta_i+\sum_{i=1}^p\beta_i^2+\lambda\sum_{i=1}^p|\beta_i|\end{aligned}$$

$$=\boldsymbol{Y}'\boldsymbol{Y}+\sum_{i=1}^{p}L(\beta_i;\hat{\beta}_i,\lambda),$$

其中,

$$L(\beta_i;\hat{\beta}_i,\lambda)=-2\hat{\beta}_i\beta_i+\beta_i^2+\lambda|\beta_i|$$

是关于$\beta_i$的凸函数. 显然, $\hat{\beta}_i^{lasso}$是$L(\beta_i;\hat{\beta}_i,\lambda)$的极小值点, $i=1,\cdots,p$.

下面分情况讨论LASSO解.

(1) 若$\hat{\beta}_i^{lasso}\neq 0$, 则$L(\beta_i;\hat{\beta}_i,\lambda)$在$\hat{\beta}_i^{lasso}$处的次梯度存在且唯一. 根据引理5.4.1可知

$$\frac{\partial L(\beta_i;\hat{\beta}_i,\lambda)}{\partial\beta_i}\Big|_{\beta_i=\hat{\beta}_i^{lasso}}=0,$$

即

$$\hat{\beta}_i-\hat{\beta}_i^{lasso}-\frac{\lambda}{2}\mathrm{sign}(\hat{\beta}_i^{lasso})=0. \tag{5.4.3}$$

由(5.4.3)可看出

$$\mathrm{sign}(\hat{\beta}_i^{lasso})=\mathrm{sign}(\hat{\beta}_i).$$

因此,

$$\begin{aligned}\hat{\beta}_i^{lasso}=&\hat{\beta}_i-\frac{\lambda}{2}\mathrm{sign}(\hat{\beta}_i^{lasso})\\ =&\hat{\beta}_i-\frac{\lambda}{2}\mathrm{sign}(\hat{\beta}_i)\\ =&|\hat{\beta}_i|\mathrm{sign}(\hat{\beta}_i)-\frac{\lambda}{2}\mathrm{sign}(\hat{\beta}_i)\\ =&(|\hat{\beta}_i|-\frac{\lambda}{2})\mathrm{sign}(\hat{\beta}_i).\end{aligned}$$

两边同时乘以$\mathrm{sign}(\hat{\beta}_i^{lasso})$(注意它不等于0), 可知

$$|\hat{\beta}_i|-\frac{\lambda}{2}=|\hat{\beta}_i^{lasso}|>0.$$

因此, 可写

$$\hat{\beta}_i^{lasso}=\mathrm{sign}(\hat{\beta}_i)(|\hat{\beta}_i|-\frac{\lambda}{2})^{+}.$$

(2) 若$\hat{\beta}_i^{lasso}=0$, 则$L(\beta_i;\hat{\beta}_i,\lambda)$在$\hat{\beta}_i^{lasso}$处不可微(因为$|\beta_i|$在0点不可微). 但可知它在$\hat{\beta}_i^{lasso}=0$处的次微分为

$$-2\hat{\beta}_i+2\hat{\beta}_i^{lasso}+\lambda c,\quad c\in[-1,1].$$

由引理5.4.1知, 存在某个$c\in[-1,1]$使得

$$-2\hat{\beta}_i+2\hat{\beta}_i^{lasso}+\lambda c=0.$$

所以,

$$2|\hat{\beta}_i - \hat{\beta}_i^{lasso}| \le \lambda,$$

即

$$|\hat{\beta}_i| \le \frac{\lambda}{2}.$$

也就是说, 当$\hat{\beta}_i^{lasso} = 0$时, 仍可写

$$\hat{\beta}_i^{lasso} = \text{sign}(\hat{\beta}_i)(|\hat{\beta}_i| - \frac{\lambda}{2})^+.$$

综上所述, 在正交设计情形下,

$$\hat{\beta}_i^{lasso} = \text{sign}(\hat{\beta}_i)(|\hat{\beta}_i| - \frac{\lambda}{2})^+, \quad i = 1, \cdots, p.$$

证毕. □

由$\hat{\beta}_i^{lasso}$的表达式可知: 若$|\hat{\beta}_i| \le \frac{\lambda}{2}$, 则LASSO方法把$\hat{\beta}_i$直接收缩到0; 否则, LASSO方法把$|\hat{\beta}_i|$的大小收缩$\lambda/2$, 同时保持$\hat{\beta}_i$的符号不变. 因此, 使用LASSO方法可得到一个稀疏的统计模型.

通常称$\hat{\beta}_i^{lasso} = \text{sign}(\hat{\beta}_i)(|\hat{\beta}_i| - \frac{\lambda}{2})^+$为$\beta_i$的软门槛(soft-threshold)估计. 事实上, $\beta_i$还存在一个硬门槛(hard-threshold)估计:

$$\hat{\beta}_i^{hard} = \hat{\beta}_i I\{|\hat{\beta}_i| > \lambda\},$$

这里$I\{\cdot\}$表示示性函数. $(\hat{\beta}_1^{hard}, \cdots, \hat{\beta}_p^{hard})'$其实是下列优化问题的解:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_0^0 \right\}, \quad \lambda \ge 0,$$

其中的$\|\boldsymbol{\beta}\|_0^0$表示$\boldsymbol{\beta}$的$\ell_0$范数, 即$\|\boldsymbol{\beta}\|_0^0 = \sum_{j=1}^p I\{\beta_i \ne 0\}$.

## 作业

1. 假设样本来自全模型:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = (\boldsymbol{X}_q, \boldsymbol{X}_t) \begin{pmatrix} \boldsymbol{\beta}_q \\ \boldsymbol{\beta}_t \end{pmatrix} + \boldsymbol{e},$$

其中$\boldsymbol{e}$满足Gauss-Markov假设. 若采用选模型

$$\boldsymbol{Y} = \boldsymbol{X}_q \boldsymbol{\beta}_q + \boldsymbol{e}$$

进行回归建模, 得到$\boldsymbol{\beta}_q$的最小二乘估计$\widetilde{\boldsymbol{\beta}}_q$. 问: 当$\boldsymbol{X}$满足什么条件时, $\widetilde{\boldsymbol{\beta}}_q$仍是$\boldsymbol{\beta}_q$的无偏估计? 并解释其条件的意义.

2. 在上题中, 记$\hat{\boldsymbol{\beta}}_q$为全模型中$\boldsymbol{\beta}_q$的最小二乘估计. 问: $\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}_q) = \mathsf{Cov}(\hat{\boldsymbol{\beta}}_q)$的充分必要条件是什么?

3. 假设全模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad i = 1, 2, \cdots, n,$$

选模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, \quad i = 1, 2, \cdots, n,$$

模型误差满足

$$\mathsf{E}(e_i) = 0, \quad \mathsf{Var}(e_i) = \sigma^2, \quad \mathsf{Cov}(e_i, e_j) = 0, \quad i \neq j.$$

记$\tilde{\sigma}^2$表示选模型中$\sigma^2$的最小二乘估计. 当全模型为真时, 求$\mathsf{E}(\tilde{\sigma}^2)$.

4. 在应用逐步回归法时, 引入自变量时的显著性水平$\alpha$与剔除自变量时的显著性水平$\beta$的赋值原则是什么? 如果希望最后的回归模型中多保留一些自变量, $\alpha$ 应该如何赋值?

5. 下表给出了10名中学生的体重($y$)、胸围($x_1$)、腰围之呼吸差($x_2$)和肺活量($x_3$)的数据.
(1) 用$\mathrm{RMS}_q$, $C_p$和AIC进行自变量选择;
(2) 用向前法、向后法以及逐步回归法进行自变量选择.

| $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1600 | 35 | 69 | 0.7 |
| 2600 | 40 | 74 | 2.5 |
| 2100 | 40 | 64 | 2.0 |
| 2650 | 42 | 74 | 3.0 |
| 2400 | 37 | 72 | 1.1 |
| 2200 | 45 | 68 | 1.5 |
| 2750 | 43 | 78 | 4.3 |
| 1600 | 37 | 66 | 2.0 |
| 2750 | 44 | 70 | 3.2 |
| 2500 | 42 | 65 | 3.0 |

题5的数据

6. 中国旅游业的现状分析. 国内旅游市场收入$y$(亿元)受到许多因素的影响, 选取如下的5个因素进行研究:
$x_1$: 国内旅游人数(万人次);
$x_2$: 城镇居民平均旅游支出(元);
$x_3$: 农村居民人均旅游支出(元);
$x_4$: 公路里程(万公里);
$x_5$: 铁路里程(万公里).

| 年份 | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|
| 1994 | 1023.5 | 52400 | 414.7 | 54.9 | 111.78 | 5.9 |
| 1995 | 1375.7 | 62900 | 464 | 61.5 | 115.7 | 5.97 |
| 1996 | 1638.4 | 63900 | 534.1 | 70.5 | 118.58 | 6.49 |
| 1997 | 2112.7 | 64400 | 599.8 | 145.7 | 122.64 | 6.6 |
| 1998 | 2391.2 | 69450 | 607 | 197 | 127.85 | 6.64 |
| 1999 | 2831.9 | 71900 | 614.8 | 249.5 | 135.17 | 6.74 |
| 2000 | 3175.5 | 74400 | 678.6 | 226.6 | 140.27 | 6.87 |
| 2001 | 3522.4 | 78400 | 708.3 | 212.7 | 169.8 | 7.01 |
| 2002 | 3878.4 | 87800 | 739.7 | 209.1 | 176.52 | 7.19 |
| 2003 | 3442.3 | 87000 | 684.9 | 200 | 180.98 | 7.3 |
| 2004 | 4710.7 | 110200 | 731.8 | 210.2 | 187.07 | 7.44 |
| 2005 | 5285.86 | 121200 | 737.1 | 227.6 | 334.52 | 7.54 |
| 2006 | 6229.74 | 139400 | 766.4 | 221.9 | 345.7 | 7.71 |
| 2007 | 7770.62 | 161000 | 906.9 | 222.5 | 358.37 | 7.8 |
| 2008 | 8749.3 | 171200 | 849.4 | 275.3 | 373.02 | 7.97 |
| 2009 | 10183.69 | 190200 | 801.1 | 295.3 | 386.08 | 8.55 |
| 2010 | 12579.77 | 210300 | 883 | 306 | 400.83 | 9.12 |

题6的数据

根据《中国统计年鉴》, 收集了1994-2010年度的数据, 见下表. 试用前进法、后退法和逐步回归法进行自变量的选择.

7. 下表给出了我国1991-2006年猪肉价格及其影响因素的数据. 在这个数据集中, $y$表示猪肉价格(元/公斤), $x_1$表示CPI, $x_2$表示人口数(亿), $x_3$表示年末存栏量(万头), $x_4$表示城镇居民可支配收入(元), $x_5$表示玉米价格(元/吨), $x_6$表示猪肉生成量(万吨). 试用前进法、后退法和逐步回归法进行自变量的选择.

| 年份 | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|---|
| 1990 | 9.84 | 103.1 | 5.28 | 36241 | 1510.2 | 686.7 | 2281 |
| 1991 | 10.32 | 103.4 | 5.89 | 36965 | 1700.6 | 590 | 2452 |
| 1992 | 10.65 | 106.4 | 5.87 | 38421 | 2026.6 | 625 | 2635 |
| 1993 | 10.49 | 114.7 | 6.01 | 39300 | 2577.4 | 726.7 | 2854 |
| 1994 | 9.16 | 124.1 | 6.45 | 41462 | 3496.2 | 1004.2 | 3205 |
| 1995 | 10.18 | 117.1 | 6.95 | 44169 | 4283 | 1576.7 | 3648 |
| 1996 | 14.96 | 107.9 | 7.58 | 36284 | 4838.9 | 1481.7 | 3158 |
| 1997 | 11.81 | 102.8 | 8.18 | 40035 | 5160.3 | 1150.8 | 3596 |
| 1998 | 10.77 | 99.2 | 9.14 | 42256 | 5425.1 | 1269.2 | 3884 |
| 1999 | 8.38 | 98.6 | 10.06 | 43020 | 5854 | 1092.5 | 3891 |
| 2000 | 8.74 | 100.4 | 10.42 | 44682 | 6280 | 887.5 | 4031 |
| 2001 | 10.18 | 100.7 | 10.55 | 45743 | 6859.6 | 1060 | 4184 |
| 2002 | 9.85 | 99.2 | 11.21 | 46292 | 7702.8 | 1033.3 | 4327 |
| 2003 | 10.7 | 101.2 | 11.45 | 46602 | 8472.2 | 1087.5 | 4519 |
| 2004 | 13.97 | 103.9 | 11.60 | 48189 | 9421.6 | 1288.3 | 4702 |
| 2005 | 13.39 | 101.8 | 12.98 | 50335 | 10493 | 1229.2 | 5011 |
| 2006 | 14.03 | 101.5 | 14.39 | 49441 | 13172 | 1280 | 5197 |

题7的数据

8. 为了科学地评估房产售价, 下表收集了美国宾夕法尼亚州伊利县第12区的建筑物特征和房价数据, 各变量的含义如下:

$y$: 房屋售价(千美元);

$x_1$: 税款(地方税/教育税/县税, 千美元);

$x_2$: 盥洗室间数;

$x_3$: 总面积(千平方英尺);

$x_4$: 起居室面积(千平方英尺);

$x_5$: 车库数;

$x_6$: 房间数;

$x_7$: 卧室数;

$x_8$: 房龄(年);

$x_9$: 壁炉数.

回答以下问题, 并附上相应的分析:

(1) 若要拟合一个刻画房产售价关于诸税款和建筑特性的相依关系的回归模型, 是否需要包含所有的变量?

(2) 一个富有经验的房产代理商建议, 只需把税款、房间数和房龄放入回归模型, 就足以刻画房价了, 你同意吗?

(3) 该项目聘请了一位房产评估师, 他认为房屋价格取决于住房的功能, 它显然是建筑物的物理特性的函数. 但是, 这种对功能特性的评价已反映在原屋主所付的地方税中了. 因此, 房价的最佳预测变量(即自变量)是地方税. 在回归模型中, 若已经包含了地方税, 建筑物的物理特性就成为冗余的了. 即, 在回归模型中只需放入税款这个预测变量. 你同意吗? 找出你认为最佳的回归模型.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $y$ |
|---|---|---|---|---|---|---|---|---|---|
| 4.918 | 1.0 | 3.472 | 0.998 | 1.0 | 7 | 4 | 42 | 0 | 25.90 |
| 5.021 | 1.0 | 3.531 | 1.500 | 2.0 | 7 | 4 | 62 | 0 | 29.50 |
| 4.543 | 1.0 | 2.275 | 1.175 | 1.0 | 6 | 3 | 40 | 0 | 27.90 |
| 4.557 | 1.0 | 4.050 | 1.232 | 1.0 | 6 | 3 | 54 | 0 | 25.90 |
| 5.060 | 1.0 | 4.455 | 1.121 | 1.0 | 6 | 3 | 42 | 0 | 29.90 |
| 3.891 | 1.0 | 4.455 | 0.988 | 1.0 | 6 | 3 | 56 | 0 | 29.90 |
| 5.898 | 1.0 | 5.850 | 1.240 | 1.0 | 7 | 3 | 51 | 1 | 30.90 |
| 5.604 | 1.0 | 9.520 | 1.501 | 0.0 | 6 | 3 | 32 | 0 | 28.90 |
| 5.828 | 1.0 | 6.435 | 1.225 | 2.0 | 6 | 3 | 32 | 0 | 35.90 |
| 5.300 | 1.0 | 4.988 | 1.552 | 1.0 | 6 | 3 | 30 | 0 | 31.50 |
| 6.271 | 1.0 | 5.520 | 0.975 | 1.0 | 5 | 2 | 30 | 0 | 31.00 |
| 5.959 | 1.0 | 6.666 | 1.121 | 2.0 | 6 | 3 | 32 | 0 | 30.90 |
| 5.050 | 1.0 | 5.000 | 1.020 | 0.0 | 5 | 2 | 46 | 1 | 30.00 |
| 8.246 | 1.5 | 5.150 | 1.664 | 2.0 | 8 | 4 | 50 | 0 | 36.90 |
| 6.697 | 1.5 | 6.902 | 1.488 | 1.5 | 7 | 3 | 22 | 1 | 41.90 |
| 7.784 | 1.5 | 7.102 | 1.376 | 1.0 | 6 | 3 | 17 | 0 | 40.50 |
| 9.038 | 1.0 | 7.800 | 1.500 | 1.5 | 7 | 3 | 23 | 0 | 43.90 |
| 5.989 | 1.0 | 5.520 | 1.256 | 2.0 | 6 | 3 | 40 | 1 | 37.90 |
| 7.542 | 1.5 | 5.000 | 1.690 | 1.0 | 6 | 3 | 22 | 0 | 37.90 |
| 8.795 | 1.5 | 9.890 | 1.820 | 2.0 | 8 | 4 | 50 | 1 | 44.50 |
| 6.083 | 1.5 | 6.727 | 1.652 | 1.0 | 6 | 3 | 44 | 0 | 37.90 |
| 8.361 | 1.5 | 9.150 | 1.777 | 2.0 | 8 | 4 | 48 | 1 | 38.90 |
| 8.140 | 1.0 | 8.000 | 1.504 | 2.0 | 7 | 3 | 3 | 0 | 36.90 |
| 9.142 | 1.5 | 7.326 | 1.831 | 1.5 | 8 | 4 | 31 | 0 | 45.80 |

题8的数据

思考题: 对于高维线性回归模型(即$p > n$的情形), 除了LASSO, 还有哪些变量选择方法? (完成一篇综述报告, 可作为大作业)

## 参考文献

1. 王松桂, 陈敏, 陈立萍. 线性统计模型. 北京: 高等教育出版社, 1999.

2. 林建忠. 回归分析与线性统计模型. 上海: 上海交通大学出版社, 2018.

3. 唐年胜, 李会琼. 应用回归分析. 北京: 科学出版社, 2014.

4. Bertsekas, D. P. (2016). Nonlinear Programming (3nd Edition). Athena Scientific, Boston.

4. Chatterjee, S. and Hadi A. S. (2012). Regression Analysis by Example (5th Edition). Wiley, New York.

5. Hald, A. (1952). Statistical Theory with Engineering Applications. Wiley, New York.