第五章　自变量的选择

Tianxiao Pang

Zhejiang University

November 30, 2023

内容

内容

内容

内容

第三章和第四章分别讨论了线性回归模型的估计方法和假设检验问题, 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.

第三章和第四章分别讨论了线性回归模型的估计方法和假设检验问题, 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, 即判断是用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.

第三章和第四章分别讨论了线性回归模型的估计方法和假设检验问题, 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, 即判断是用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(见定理5.1.1和定理5.1.2).

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

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

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$p-q$个自变量$x_{q+1},\cdots,x_p$. 记

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

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$p-q$个自变量$x_{q+1},\cdots,x_p$. 记

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

则得到一个新模型

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

这里$\boldsymbol{X}_q$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β和σ^2的最小二乘估计为

在全模型中, 回归系数$\boldsymbol{\beta}$和σ^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$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 =$

在全模型中, 回归系数$\boldsymbol{\beta}$和σ^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$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

在全模型中, 回归系数$\boldsymbol{\beta}$和σ^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$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

在选模型中, 回归系数$\boldsymbol{\beta}_q$和σ^2的最小二乘估计为

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

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

在$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

在选模型中, 回归系数$\boldsymbol{\beta}_q$和σ^2的最小二乘估计为

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

在$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\widetilde{y}_{0q} =$

在全模型中, 回归系数$\boldsymbol{\beta}$和σ^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$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

在选模型中, 回归系数$\boldsymbol{\beta}_q$和σ^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$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\widetilde{y}_{0q} = \boldsymbol{x}_{0q}'\widetilde{\boldsymbol{\beta}}_q$.

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

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

在$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

在选模型中, 回归系数$\boldsymbol{\beta}_q$和σ^2的最小二乘估计为

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

在$\boldsymbol{x}_0' = (\boldsymbol{x}_{0q}', \boldsymbol{x}_{0t}')$点上的预测为$\widetilde{y}_{0q} = \boldsymbol{x}_{0q}'\widetilde{\boldsymbol{\beta}}_q$.

对$\hat{\boldsymbol{\beta}}$作相应的分块:

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

若$\widetilde{\boldsymbol{\theta}}$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, 更合理的度量标准为均方误差矩阵(mean square error matrix, MSEM).

若$\widetilde{\boldsymbol{\theta}}$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, 更合理的度量标准为均方误差矩阵(mean square error matrix, MSEM).

定义

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

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

若$\widetilde{\boldsymbol{\theta}}$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, 更合理的度量标准为均方误差矩阵(mean square error matrix, MSEM).

定义

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

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

不难推得:

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

回忆分块矩阵求逆公式:

回忆分块矩阵求逆公式:

引理 (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

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

定理 (5.1.1, 对估计的影响)

假设全模型(5.1.1)正确, 则

(1) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$; $E(\widetilde{\boldsymbol{\beta}}_q) = \boldsymbol{\beta}_q + \boldsymbol{G}\boldsymbol{\beta}_t$, 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, $E(\widetilde{\boldsymbol{\beta}}_q) \neq \boldsymbol{\beta}_q$;

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

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

(4) $E(\widetilde{\sigma}_q^2) \geq E(\hat{\sigma}^2) = \sigma^2$, 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.

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

(2) 记

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

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

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

为非负定矩阵.

(3) 由公式(5.1.5)以及结论(1)可知

(3) 由公式(5.1.5)以及结论(1)可知

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

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,

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

为非负定矩阵.

(4) $\mathsf{E}(\hat{\sigma}^2) = \sigma^2$是已知的. 而

(4) $\mathsf{E}(\hat{\sigma}^2) = \sigma^2$是已知的. 而

$$
\begin{aligned}
\mathsf{E}(\widetilde{\sigma}_q^2) &= \frac{1}{n-q-1}\mathsf{E}\Big\{\boldsymbol{Y}'[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{Y}\Big\} \\
&= \frac{1}{n-q-1}\mathsf{tr}\Big\{[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}')\Big\} \\
&= \frac{1}{n-q-1}\mathsf{tr}\Big\{[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q'](\sigma^2\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{X}')\Big\} \\
&= \frac{1}{n-q-1}\Big\{(n-q-1)\sigma^2 + \boldsymbol{\beta}'\boldsymbol{X}'[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{X}\boldsymbol{\beta}\Big\} \\
&= \sigma^2 + \frac{1}{n-q-1}\boldsymbol{\beta}_t'\boldsymbol{X}_t'[\boldsymbol{I}_n - \boldsymbol{X}_q(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q']\boldsymbol{X}_t\boldsymbol{\beta}_t \\
&= \sigma^2 + \frac{1}{n-q-1}\boldsymbol{\beta}_t'(\boldsymbol{D} - \boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{C})\boldsymbol{\beta}_t \\
&= \sigma^2 + \frac{1}{n-q-1}\boldsymbol{\beta}_t'\boldsymbol{D}_1^{-1}\boldsymbol{\beta}_t \\
&\geq \sigma^2 = \mathsf{E}(\hat{\sigma}^2),
\end{aligned}
$$

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

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

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

定理 (5.1.2, 对预测的影响)

假设全模型*(5.1.1)*正确, 则

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

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

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

证明 (1) $\mathsf{E}(z_0) = 0$是显然的. 现考察$\mathsf{E}(z_{0q})$.

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

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

所以一般情形下$\widetilde{y}_{0q}$是有偏预测.

(2) 首先, 容易看出

(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

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

马上推得

马上推得

$$
\begin{aligned}
& \mathsf{Var}(z_0) - \mathsf{Var}(z_{0q}) \\
= \; & \sigma^2[\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 - \boldsymbol{x}_{0q}'(\boldsymbol{X}_q'\boldsymbol{X}_q)^{-1}\boldsymbol{x}_{0q}] \\
= \; & \sigma^2[\boldsymbol{x}_{0q}'(\boldsymbol{B}_1 - \boldsymbol{B}^{-1})\boldsymbol{x}_{0q} + \boldsymbol{x}_{0q}'\boldsymbol{C}_1\boldsymbol{x}_{0t} + \boldsymbol{x}_{0t}'\boldsymbol{C}_1'\boldsymbol{x}_{0q} + \boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{x}_{0t}] \\
= \; & \sigma^2[\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} \\
& \quad -\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1\boldsymbol{x}_{0t} - \boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} + \boldsymbol{x}_{0t}'\boldsymbol{D}_1\boldsymbol{x}_{0t}] \\
= \; & \sigma^2[\boldsymbol{x}_{0q}'\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t}) - \boldsymbol{x}_{0t}'\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t})] \\
= \; & \sigma^2(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t})'\boldsymbol{D}_1(\boldsymbol{C}'\boldsymbol{B}^{-1}\boldsymbol{x}_{0q} - \boldsymbol{x}_{0t}) \geq 0.
\end{aligned}
$$

(3) 首先, 容易看出

(3) 首先, 容易看出

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

(3) 首先, 容易看出

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

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

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

总结:

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$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$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.

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$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$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$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$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.

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$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$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$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_t) - \boldsymbol{\beta}_t\boldsymbol{\beta}_t'$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.

(3) 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.

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:

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

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

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

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

改写RSS_q如下:

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

则可知$\mathsf{RSS}_{q+1} \leq \mathsf{RSS}_q$.

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.

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?

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?

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. 但另一方面, 在样本容量给定的条件下, 自变量越多, 回归参数也就越多, 参数估计的准确性会变差.

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?

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. 但另一方面, 在样本容量给定的条件下, 自变量越多, 回归参数也就越多, 参数估计的准确性会变差.

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.

自变量选择的几个常见准则:

自变量选择的几个常见准则:

(1) 平均残差平方和准则(RMS_q)

自变量选择的几个常见准则:

(1) 平均残差平方和准则(RMS_q)

由于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

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

我们按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RMS_q准则.

(2) 调整后的R^2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$R_q^2 = 1 - \mathrm{RSS}_q/\mathrm{TSS}$, 不能直接把$R_q^2$作为选择自变量的准则, 否则将把所有自变量选入模型. 为了克服以上缺点, 引入调整后的判定系数

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

(2) 调整后的R^2准则

判定系数$R_q^2 = \mathsf{ESS}_q/\mathsf{TSS}$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$R_q^2 = 1 - \mathsf{RSS}_q/\mathsf{TSS}$, 不能直接把$R_q^2$作为选择自变量的准则, 否则将把所有自变量选入模型. 为了克服以上缺点, 引入调整后的判定系数

$$\begin{aligned}\bar{R}_q^2 &= 1 - \frac{\mathsf{RSS}_q/(n-q-1)}{\mathsf{TSS}/(n-1)} = 1 - \frac{n-1}{n-q-1}\frac{\mathsf{RSS}_q}{\mathsf{TSS}} \\ &= 1 - \frac{n-1}{n-q-1}(1-R_q^2).\end{aligned}$$

易见$\bar{R}_q^2 \le R_q^2$, 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的增大而增大.

(2) 调整后的R^2准则

判定系数$R_q^2 = \mathsf{ESS}_q/\mathsf{TSS}$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$R_q^2 = 1 - \mathsf{RSS}_q/\mathsf{TSS}$, 不能直接把$R_q^2$作为选择自变量的准则, 否则将把所有自变量选入模型. 为了克服以上缺点, 引入调整后的判定系数

$$\begin{aligned}\bar{R}_q^2 &= 1 - \frac{\mathsf{RSS}_q/(n-q-1)}{\mathsf{TSS}/(n-1)} = 1 - \frac{n-1}{n-q-1}\frac{\mathsf{RSS}_q}{\mathsf{TSS}} \\ &= 1 - \frac{n-1}{n-q-1}(1-R_q^2).\end{aligned}$$

易见$\bar{R}_q^2 \le R_q^2$, 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的增大而增大. 我们选择使$\bar{R}_q^2$达到最大的自变量子集.

(3) C_p准则

(3) C_p准则

C_p准则是C.L.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σ^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.

(3) C_p准则

C_p准则是C.L.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σ^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

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

达到最小.

写

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

$$=$$

写

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

写

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

易见

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

利用定理5.1.1中的结论(1)以及结论(4)的证明过程, 得

$$I_2 \quad =$$

利用定理5.1.1中的结论(1)以及结论(4)的证明过程, 得

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

所以

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

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

所以

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

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

此外, 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.

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

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

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$\mathsf{E}(\mathsf{RSS}_q/\hat{\sigma}^2)$.

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

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

选模型(5.1.2)中的残差平方和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, $\eta \overset{\Delta}{=} \mathsf{RSS}_q - \mathsf{RSS}$与RSS相互独立(根据最小二乘法基本定理), 且

$$\mathsf{E}\left(\frac{\mathsf{RSS}_q}{\hat{\sigma}^2}\right) \quad =$$

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

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

选模型(5.1.2)中的残差平方和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, $\eta \overset{\Delta}{=} \mathsf{RSS}_q - \mathsf{RSS}$与RSS相互独立(根据最小二乘法基本定理), 且

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

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

$$\mathsf{E}\Big(\frac{\sigma^2}{\mathsf{RSS}}\Big) = 2^{-\frac{k}{2}}\Big[\Gamma\Big(\frac{k}{2}\Big)\Big]^{-1}\int_0^{\infty} x^{-1}\cdot e^{-\frac{x}{2}}x^{\frac{k}{2}-1}dx$$

$$= $$

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

因此

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

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

因此

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

假设$H:\boldsymbol{\beta}_t=\mathbf{0}$可以等价地写成线性假设$H:\boldsymbol{A}\boldsymbol{\beta}=\mathbf{0}$, 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$\mathsf{rk}(\boldsymbol{A})=t$. 同时注意到

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

所以由第四章的公式(4.1.8)及非中心卡方分布的定义2.4.1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

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

所以由第四章的公式(4.1.8)及非中心卡方分布的定义2.4.1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

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

因此

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

现在, 已可推知(注意$p = q + t$)

$$\mathsf{E}(C_p) \quad =$$

现在, 已可推知(注意$p = q + t$)

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

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, Γ_q越小越好, 所以我们应该选择具有最小C_p值的自变量子集.

推论

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

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

等价地,

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

其中$u \sim F(t, n - p - 1)$.

证明 记

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

易见u为假设$H : \boldsymbol{\beta}_t = \mathbf{0}$的$F$检验统计量.

证明 记

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

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$u \sim F(t, n-p-1)$. 借助于$u$, C_p可表示为

$$C_p \quad =$$

证明 记

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

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$u \sim F(t, n-p-1)$. 借助于$u$, C_p可表示为

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

来解释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, 我们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.

最后, 关于C_p统计量, 我们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

(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).

(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. 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

$$L(\boldsymbol{\beta}_q, \sigma^2 | \boldsymbol{Y}) =$$

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

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

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

$$\tilde{\boldsymbol{\beta}}_q =$$

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

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

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

$$\tilde{\boldsymbol{\beta}}_q = (\boldsymbol{X}_q' \boldsymbol{X}_q)^{-1} \boldsymbol{X}_q' \boldsymbol{Y},$$

$$\tilde{\sigma}_q^2 =$$

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

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

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

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

代入(5.2.4)得对数似然函数的最大值为

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

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

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

我们选择使上式达到最小的自变量子集.

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

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

我们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.

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.

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).

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

(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

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

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

$$\sum_{i=1}^{n} \mathsf{Var}(\tilde{y}_i - \boldsymbol{x}'_{iq}\widetilde{\boldsymbol{\beta}}_q) \quad =$$

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

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

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

$$\begin{aligned} J_p &= (n+q+1)\widetilde{\sigma}_q^2 \\ &= \frac{n+q+1}{n-q-1}\mathsf{RSS}_q. \end{aligned}$$

这里$(n+q+1)/(n-q-1)$起着惩罚的作用.

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

$$\begin{aligned} J_p &= (n+q+1)\widetilde{\sigma}_q^2 \\ &= \frac{n+q+1}{n-q-1}\mathsf{RSS}_q. \end{aligned}$$

这里$(n+q+1)/(n-q-1)$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PRESS_q(predicted residual sum of squares)准则

(6) 预测残差平方和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

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

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

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

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

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

在第三章中我们已证明

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

所以

$$\hat{e}_{(i)} =$$

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

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

在第三章中我们已证明

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

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. 因此

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

若要计算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$i$个对角元即可.

这里的$\hat{e}_i$是全数据情形下的第i个残差. 因此

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

若要计算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$i$个对角元即可.

我们选择使得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

```
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)
cement
X=matrix(c(x1,x2,x3,x4),nrow=13,byrow=F)
library(leaps)
adjr=leaps(X,y,int=T,method="adjr2")
adjr
adjr$which[which.max(adjr$adjr2),]
```

```
> X=matrix(c(x1,x2,x3,x4),nrow=13,byrow=F)
> 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
```

Figure: $2^4 - 1 = 15$个自变量子集

```
$adjr2
 [1] 0.6449549 0.6359290 0.4915797 0.2209521 0.9744140 0.9669653 0.9223476
 [8] 0.8164305 0.6160725 0.4578001 0.9764473 0.9763796 0.9750415 0.9637599
[15] 0.9735634

> adjr$which[which.max(adjr$adjr2),]
    1     2     3     4
 TRUE  TRUE FALSE  TRUE
>
```

Figure: 调整后的R^2达到最大的自变量子集

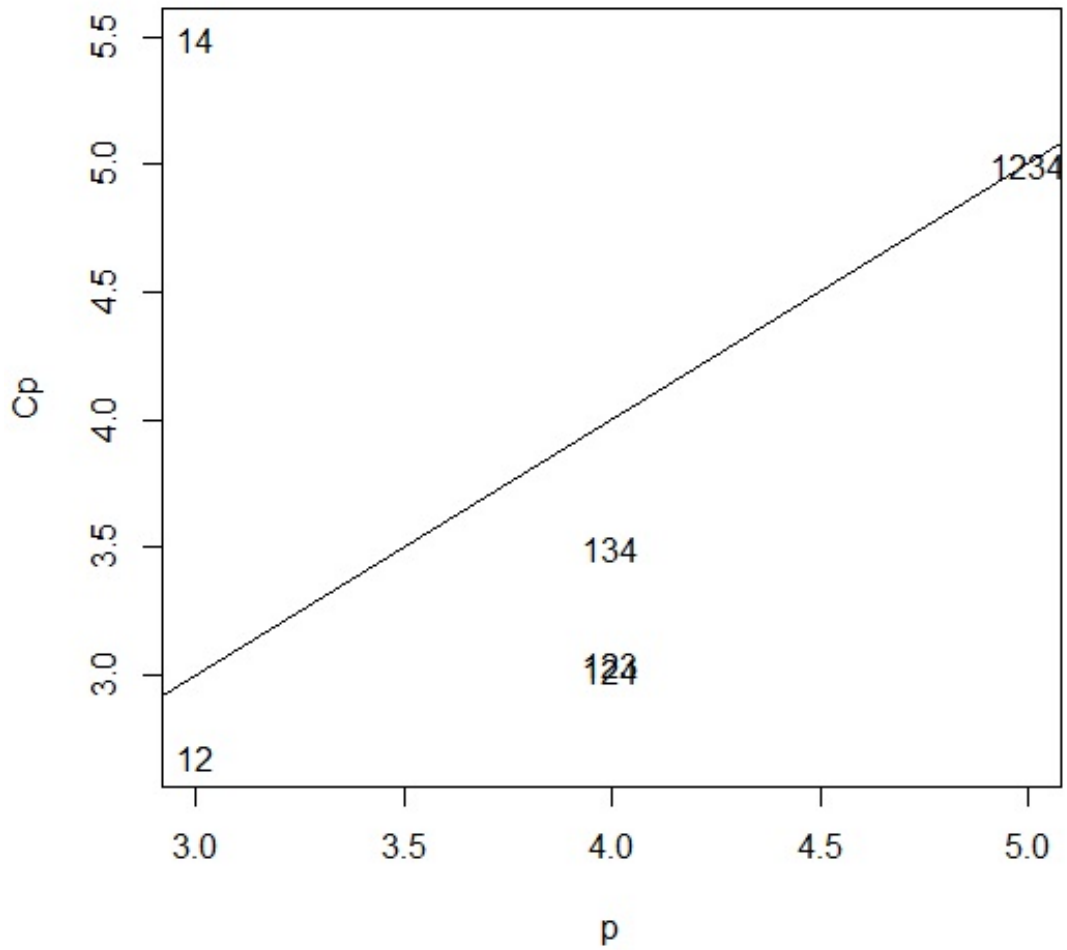
利用调整后的R^2, 最终选择自变量子集: x_1, x_2, x_4.

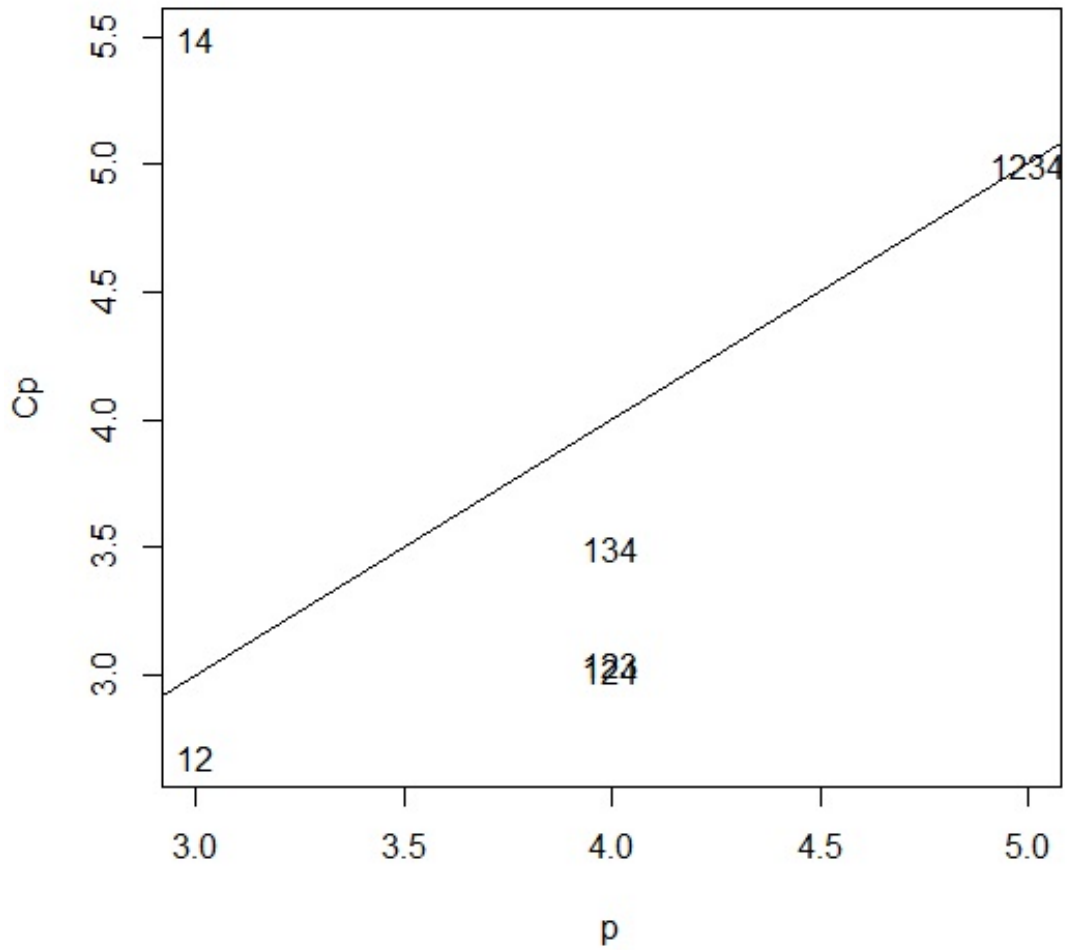
```
library(faraway)
library(leaps)
cp=leaps(X,y,int=T,method="Cp")
cp
cp$which[which.min(cp$Cp),]
Cpplot(cp)
```

```
library(faraway)
library(leaps)
cp=leaps(X,y,int=T,method="Cp")
cp
cp$which[which.min(cp$Cp),]
Cpplot(cp)
```

```
$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)
>
```


14
1234
134
124
12
Cp
3.0
3.5
4.0
4.5
5.0
5.5
p



有些模型的C_p没有显示在图中, 因为它们的C_p值太大. 根据C_p准则, 最终选择自变量子集: x_1, x_2.

leaps中的参数method只有三个选项: method=c(“Cp”, “adjr2”, “r2”). 若需通过其它准则选择自变量, 需自己编写代码, 例如:

leaps中的参数method只有三个选项: method=c(“Cp”, “adjr2”, “r2”). 若需通过其它准则选择自变量, 需自己编写代码, 例如:

```
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)
cement
library(leaps)
search.results=regsubsets(y~x1+x2+x3+x4,data=cement,
    method=“exhaustive”,nbest=15)
selection.criteria=summary(search.results)
selection.criteria
```

(未完, 待续, nbest指“number of subsets of each size to record”)

```
names(selection.criteria)
selection.criteria$which
n=length(cement[, 1])
q=as.integer(row.names(selection.criteria$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)
```

```
> 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
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 ) "*" "*" "*" "*"
```

```
> 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
>
```

	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

DAAG package中的press(model)可以返回model的PRESS值.

DAAG package中的press(model)可以返回model的PRESS值.

```
> lm.reg=lm(y~x1+x2,data=cement)
> AIC(lm.reg)
[1] 64.31239
> library(DAAG)
> press(lm.reg)
[1] 93.88255
>
```

DAAG package中的press(model)可以返回model的PRESS值.

```
> lm.reg=lm(y~x1+x2,data=cement)
> AIC(lm.reg)
[1] 64.31239
> library(DAAG)
> press(lm.reg)
[1] 93.88255
>
```

显然, 若要利用press(model)一一计算选模型的PRESS值, 那就太麻烦了.

汪利军(3140105707)的multipress函数:

```
> 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))
+  }
```

```
yx=read.table("***.txt")
x1=yx[,1]
x2=yx[,2]
x3=yx[,3]
x4=yx[,4]
y=yx[,5]
X=data.frame(x1,x2,x3,x4)
library(DAAG)
multipress(X,y)
```

```
> 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
```

leaps package中的“regsubsets”也有图示法的变量选择功能.

leaps package中的“regsubsets”也有图示法的变量选择功能.

```
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)
cement
library(leaps)
subsets=regsubsets(y~x1+x2+x3+x4,data=cement)
summary(subsets)
plot(subsets)
plot(subsets,scale="Cp")
plot(subsets,scale="adjr2")
```

注: scale=“xx” where “xx” is either “C_p”, “adjr2”, “r2” or “bic”.

```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4,data=cement)
> summary(subsets)
Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
2  ( 1 ) "*" "*" " " " "
3  ( 1 ) "*" "*" " " "*"
4  ( 1 ) "*" "*" "*" "*"
>
```

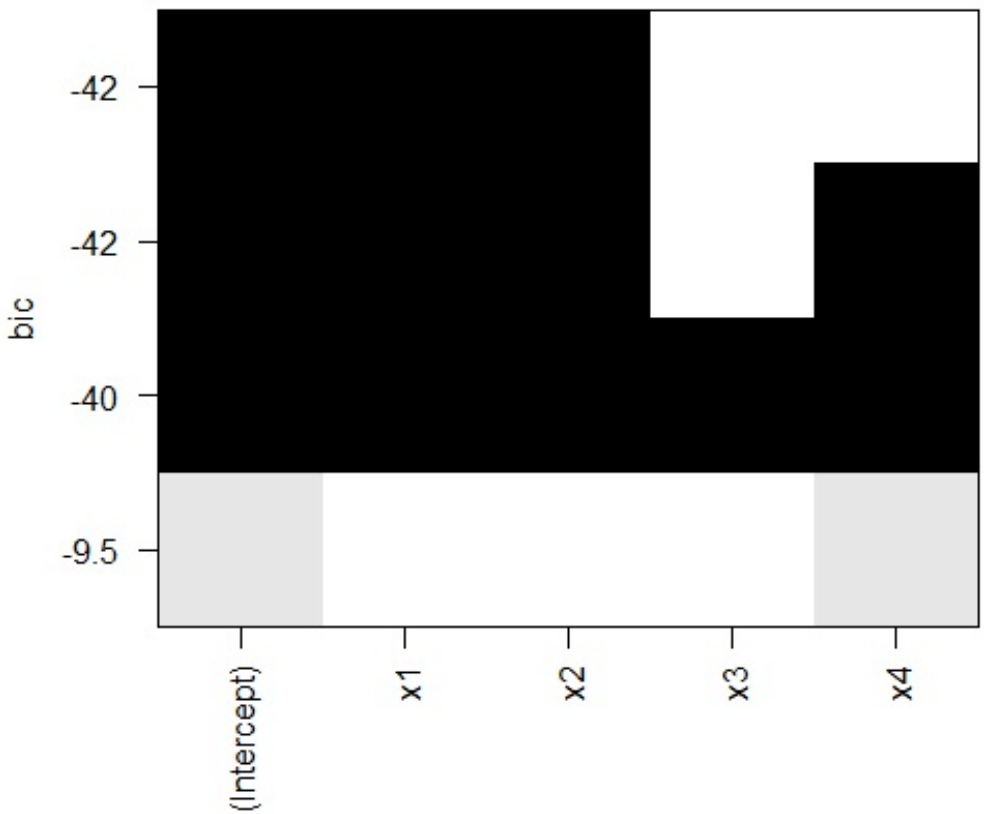
```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4,data=cement)
> summary(subsets)
Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = cement)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
2  ( 1 ) "*" "*" " " " "
3  ( 1 ) "*" "*" " " "*"
4  ( 1 ) "*" "*" "*" "*"
>
```

这结果告诉我们: 若只选择一个自变量, 应选入x_4; 若只选入两个自变量, 应选入x_1和x_2; 若选入三个自变量, 应选入x_1, x_2和x_4; 若选入四个自变量, 则全部选入.

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".

plot(subsets)产生的图形是

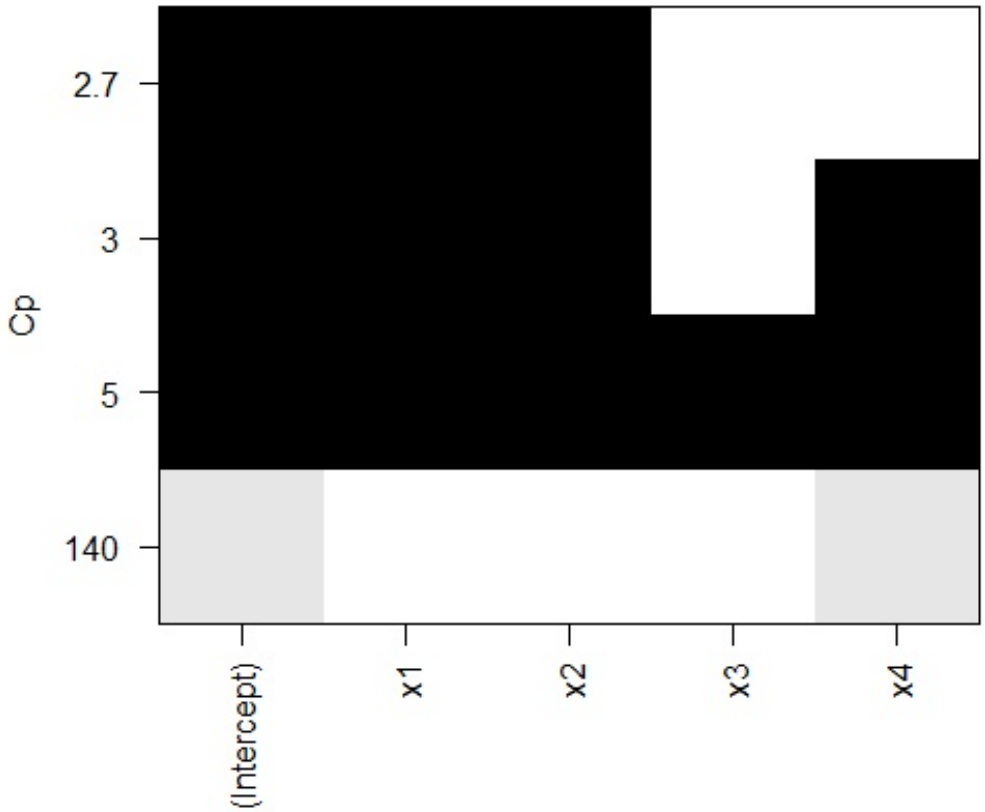
plot(subsets)产生的图形是



因此, 采用BIC准则, 应选入自变量x_1和x_2.

plot(subsets,scale=“Cp”)产生的图形是

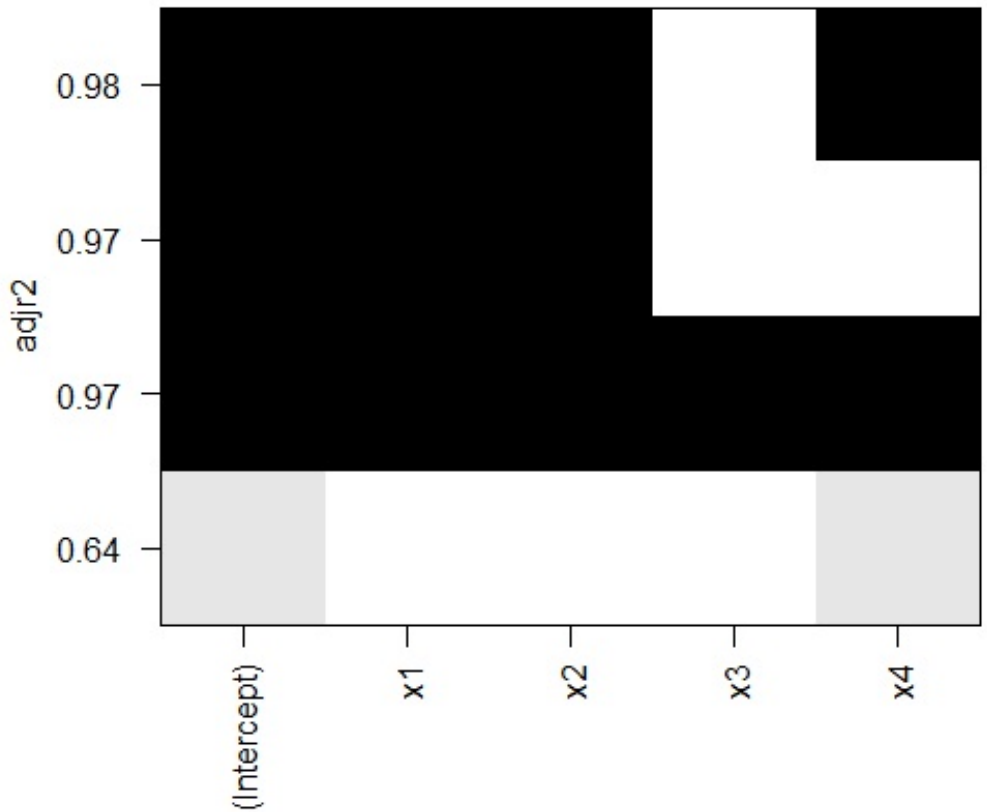
plot(subsets,scale="Cp")产生的图形是



因此, 采用C_p准则, 应选入自变量x_1和x_2.

plot(subsets,scale=“adjr2”)产生的图形是

plot(subsets,scale="adjr2")产生的图形是



因此, 采用调整后的R^2准则, 应选入自变量x_1, x_2和x_4.

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.

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)

(1) 向前法(forward)

向前法的思想是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:

(1) 向前法(forward)

向前法的思想是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:

步骤1: 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p个回归系数的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α, 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$x_j$就是$x_1$;

步骤2: 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$x_j$就是$x_2$;

步骤2: 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$x_j$就是$x_2$;

步骤3: 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$x_{q+1}, \cdots, x_p$的F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)

(2) 向后法(backward)

向后法与向前法恰恰相反, 向后法的思想是选入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:

(2) 向后法(backward)

向后法与向前法恰恰相反, 向后法的思想是选入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:

步骤1: 因变量y关于所有自变量$x_1, \cdots, x_p$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β, 若$F_j^{(p)} \leq F_\beta(1, n-p-1)$, 则首先将$x_j$从回归模型中剔除. 不妨假设$x_j$就是$x_p$;

步骤2: 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$F_j^{(p-1)} \leq F_\beta(1, n-p)$, 则将$x_j$从回归模型中剔除. 不妨假设$x_j$就是$x_{p-1}$;

步骤2: 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$F_j^{(p-1)} \leq F_\beta(1, n-p)$, 则将$x_j$从回归模型中剔除. 不妨假设$x_j$就是$x_{p-1}$;

步骤3: 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$x_1, \cdots, x_q$的F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, 将永远没有机会再重新进入回归模型, 即是“一棒子打死”的.

(3) 逐步回归法(stepwise)

(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, 都要对已选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, 以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使用的显著性水平是不同的, 要求引入自变量时所使用的显著性水平α小于剔除自变量时所使用的显著性水平β, 否则就可能产生死循环.

应用逐步回归时需注意: 引入自变量与剔除自变量时所使用的显著性水平是不同的, 要求引入自变量时所使用的显著性水平α小于剔除自变量时所使用的显著性水平β, 否则就可能产生死循环.

也就是说, 当$\alpha \geq \beta$时, 如果某个自变量的p值在α与β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.

用向后法对cement数据进行变量选择, 设置$\beta = 0.1$.

```
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)
cement
lm.sol=lm(y~x1+x2+x3+x4,data=cement)
summary(lm.sol)
```

```
> summary(lm.sol)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

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
```

可以发现x_3是最不显著的自变量, 所以删除x_3.

lm.sol=update(lm.sol,.~.-x3)
summary(lm.sol)

```
> summary(lm.sol)

Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)

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
```

可以发现x_4是最不显著的自变量, 所以删除x_4.

lm.sol=update(lm.sol,.~.-x4)
summary(lm.sol)

```
> summary(lm.sol)

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

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
```

可以发现剩下的x_1和x_2的p-value都小于$\beta = 0.1$, 所以向后法选择出来的自变量是x_1和x_2.

在R软件中, 基于p-value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.

向前法:

```
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)
cement
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)

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 ***
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 .
---
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
```

向前法的结果: 选入x_1, x_2和x_4.

向后法:

```
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)
cement
max.model=lm(y~x1+x2+x3+x4,data=cement)
bwd.model=step(max.model,direction="backward")
summary(bwd.model)
```

```
> 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)

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
```

向后法的结果: 选入x_1, x_2和x_4.

逐步回归法:

```
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)
cement
min.model=lm(y~1,data=cement)
step.model=step(min.model,direction="both",
                scope=(~x1+x2+x3+x4))
summary(step.model)
```

```
> 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
```

未完,待续.

```
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)

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 ***
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 .
---
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
```

逐步回归法的结果: 选入x_1, x_2和x_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拟合方法, 它要求自变量个数p 小于样本容量n. 若$p \geq n$, 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σ^2)或做假设检验等统计推断.

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拟合方法, 它要求自变量个数p 小于样本容量n. 若$p \geq n$, 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σ^2)或做假设检验等统计推断.

第三章的岭估计是把$\boldsymbol{\beta}$的ℓ_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ℓ_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

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

(5.4.1)等价于

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

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β_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.

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β_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, 更进一步地提高了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).

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β_i的估计取为0? 以二维为例, 来了解一下它的原理.

LASSO方法为何能把某些β_i的估计取为0? 以二维为例, 来了解一下它的原理.

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.

LASSO方法为何能把某些β_i的估计取为0? 以二维为例, 来了解一下它的原理.

因为

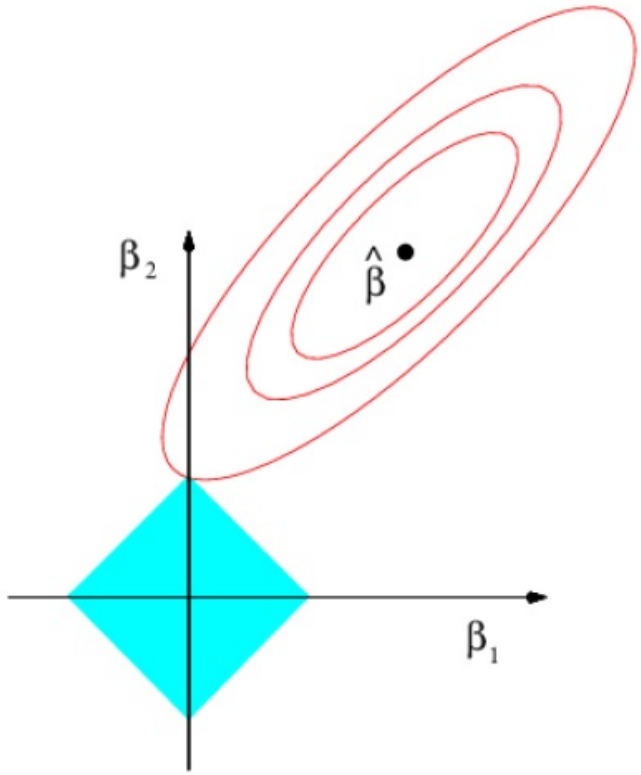
$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 =&\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\\ =&|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsym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

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.

优化问题(5.4.2)中的约束条件$\sum_{i=1}^{p}|\beta_i|\leq t$所表示的图像是一个正方形(在更高维的情形, 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), 见下图.



当t很小时, 正方形与$\hat{\beta}$的等高线不相交; 当t变大时, 它终将与$\hat{\beta}$的等高线相交, 该交点是正方形的某一顶点, 它就是LASSO估计.

需要注意的是, $t = 0$时, 所有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.

需要注意的是, $t = 0$时, 所有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λ值). 在实际问题中, 为了更客观地选择t值或λ值, 可以应用交叉验证(cross-validation)的方法.

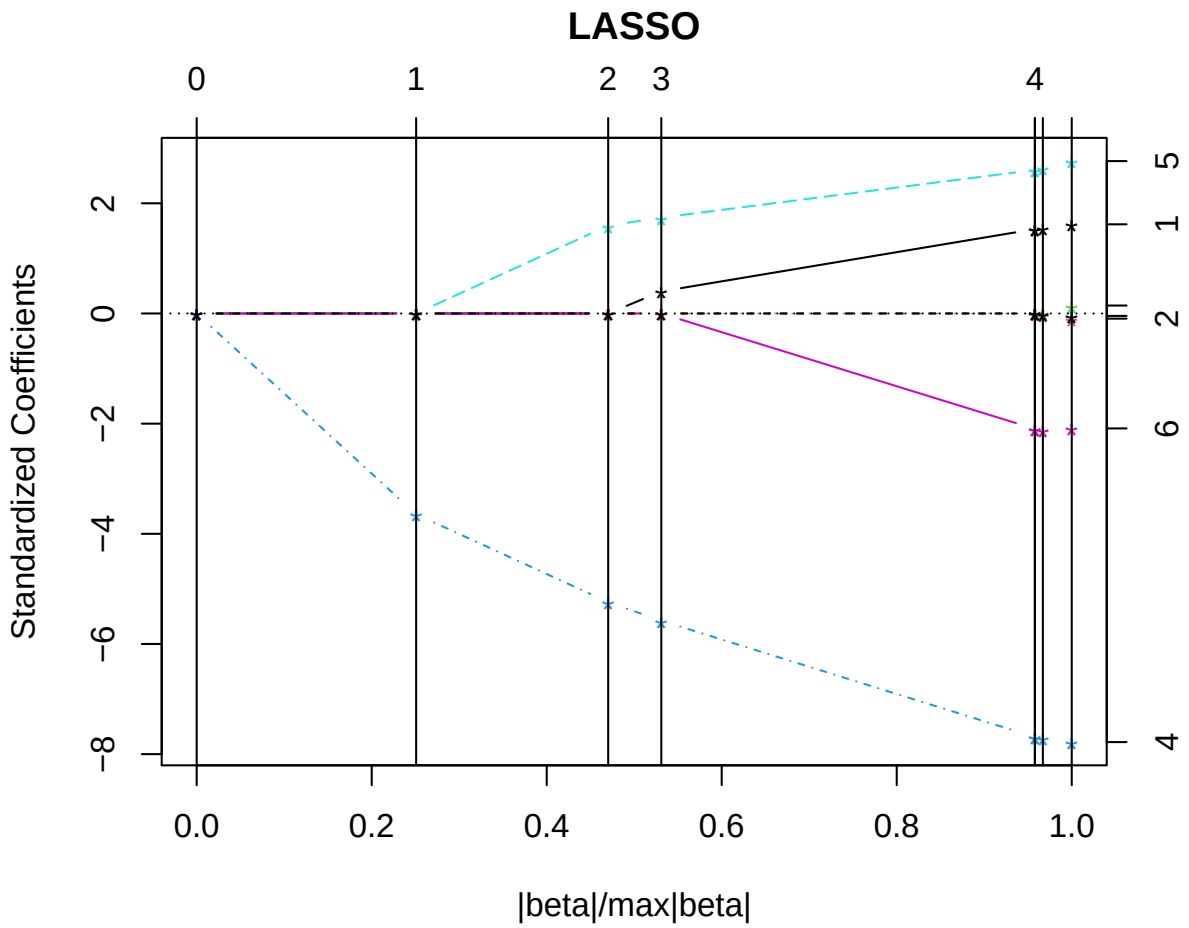
例5.4.1 考虑R中的state.x77数据集, 该数据集收集了上个世纪六七十年代美国50个州的预期寿命以及与此可能相关的其它7个变量的数据. 下表给出该数据集的前6组数据.

	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

来解释一下这几个变量的含义. Population表示人口估计数(截至1975年7月1日); Income表示人均收入(1974); Illiteracy表示文盲人口的百分比(1970); Life.Exp表示预期寿命(1969-1971); Murder表示谋杀和非过失杀人率(1976); HS.Grad表示高中毕业生的百分比(1970); Frost表示首府或大城市的最低气温低于零度的平均天数(1931-1960); Area表示土地面积(平方英里).

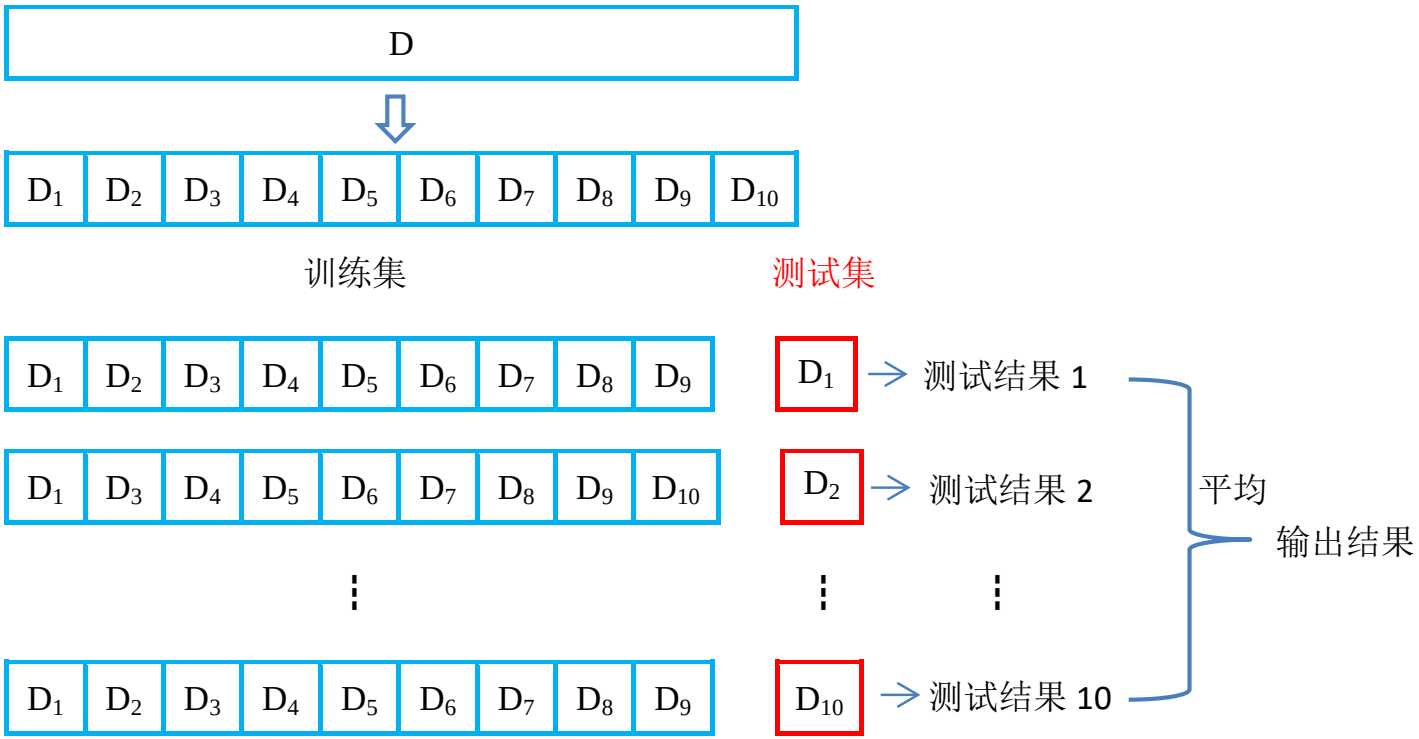
```
library(lars)
statedata=data.frame(state.x77,row.names=state.abb)
lasso.sol=lars(as.matrix(statedata[,-4]),statedata$Life.Exp)
plot(lasso.sol)
```



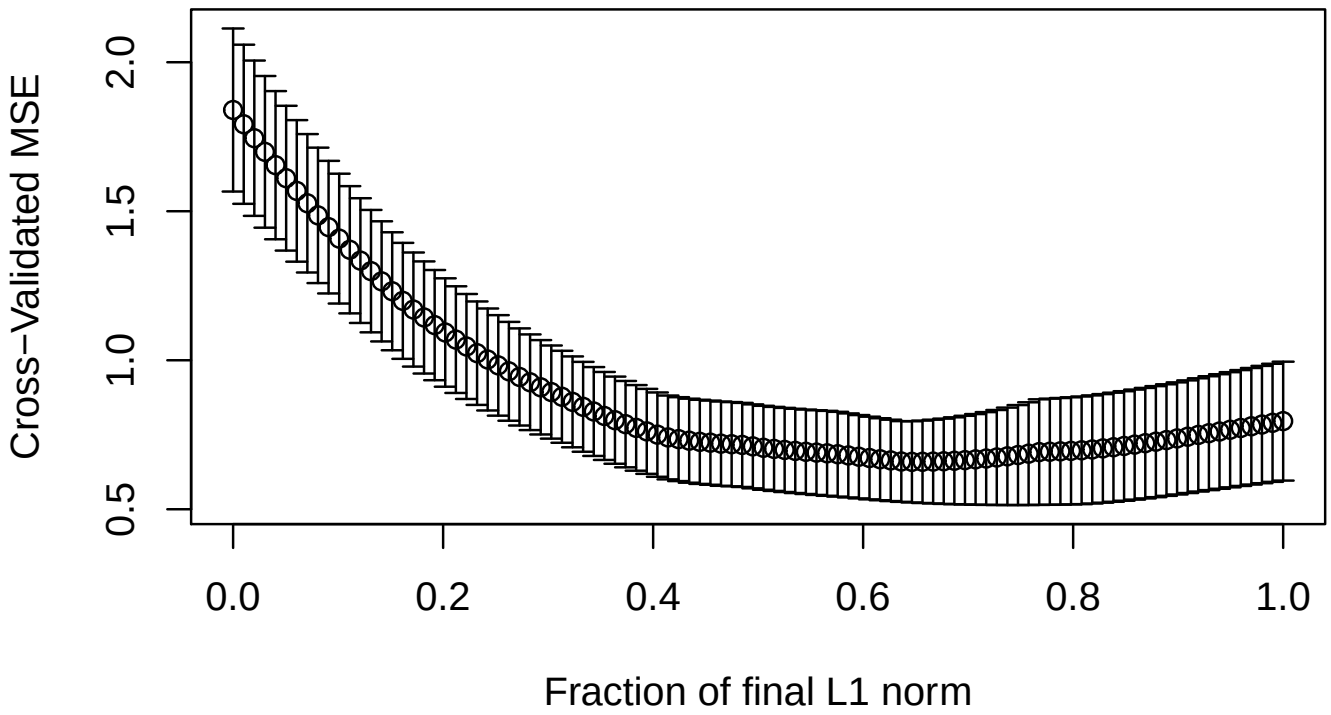
此图的横坐标$s = |\text{beta}|/\max|\text{beta}|$表示LASSO估计的$\ell_1$范数与最小二乘估计的$\ell_1$范数的比值. s其实与t有关(因为LASSO估计与t有关), 它是t的单调增函数. 当$t = 0$时, $s = |\text{beta}|/\max|\text{beta}| = 0$; 当$t = \infty$时, $s = |\text{beta}|/\max|\text{beta}| = 1$.

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R的程序包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).



set.seed(123)
cv.sol=cv.lars(as.matrix(statedata[,-4]),statedata$Life.Exp)
(画CV-MSE图)
cv.sol$index[which.min(cv.sol$cv)] (cv.sol$index表示$s$; cv.sol$cv表示平均预测均方误差)




```
> set.seed(123)
> cv.sol=cv.lars(as.matrix(statedata[,-4]),statedata$Life)
> cv.sol$index[which.min(cv.sol$cv)]
[1] 0.6464646
```

来看一下当$s = 0.6464646$时, 有哪些自变量进入了模型.

predict(lasso.sol,s=0.6464646,type="coef",mode="fraction")$coef

```
> 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.

可以看出, 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.

可以看出, 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$\hat{\boldsymbol{\beta}}^{lasso} = (\hat{\beta}_1^{lasso}, \cdots, \hat{\beta}_p^{lasso})'$.

可以看出, 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, 如何求极小值点?

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. 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$[a,b]$.

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$\nabla f(x_0)$.

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)的经典著作Nonlinear Programming (3nd Edition).

定理 (5.4.1)

在正交设计情形下,

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

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

证明 回忆

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

因此,

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

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.

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,

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

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

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

因此, 可写

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

(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

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

(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,

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

即

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

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

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

综上所述, 在正交设计情形下,

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

证毕.

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

统计里, 把$\hat{\beta}_i^{lasso}$称为β_i的软门槛(soft-threshold)估计.

统计里, 把$\hat{\beta}_i^{lasso}$称为β_i的软门槛(soft-threshold)估计. 事实上, β_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:

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

这里, $\|\boldsymbol{\beta}\|_0^0$表示$\boldsymbol{\beta}$的$\ell_0$范数, 即

$$\|\boldsymbol{\beta}\|_0^0 = \sum_{j=1}^{p} I\{\beta_i \neq 0\}.$$