第7章
线性回归模型

李高荣

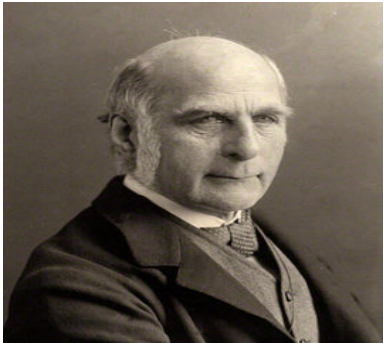
北京师范大学统计学院

E-mail: *ligaorong@bnu.edu.cn*

1. 多元线性回归分析
2. 回归诊断
3. 子集选择
4. 压缩估计方法
   - 岭回归(ridge regression)
   - 桥回归(bridge regression)
   - 惩罚变量选择方法
5. Lasso: 线性回归模型应用
6. SCAD：线性回归模型应用
7. 自适应Lasso
8. 高维回归模型：Lasso应用

- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

- “回归"的概念是1886年由英国统计学家Galton在研究父代身高与子代身高之间的关系时提出的。
- 回归分析已经成为现代统计学中应用最为广泛的方法之一，主要用于探索和检验协变量X与响应变量Y之间的相关关系，也可以通过协变量X的取值变化来预测响应变量Y的取值，进一步可以描述协变量X和响应变量Y之间的相互关系。



模型介绍

- 客观世界中变量之间的关系包括：
  - 确定性关系： 变量之间的关系能用函数来表达
  - 非确定性关系： 相关关系
- 回归分析： 研究相关关系的数学工具。可帮助人们从一个变量的取值去估计另一个变量的值。

- 当给定X的取值时，Y的取值不能确定，只能通过一定的概率分布来描述。
- 因此, 把给定X 的取值时, Y的条件数学期望

$$\mu(x) = E(Y|X = x) \tag{1.1}$$

为随机变量Y对$X = x$的回归函数，或称为随机变量Y 对$X = x$的均值回归函数。

- 式(1.1) 从平均意义下刻画了变量X和Y之间的统计规律。

- **回归分析的任务：** 根据试验观测数据去估计回归函数，讨论模型中未知参数的点估计、区间估计和假设检验，回归模型的拟合优度检验，以及对随机变量Y的观测值作出点预测和区间预测等问题。
  - **预测问题：** 在给定的置信度下，估计出x取某一定值x_0时，随机变量Y的取值情况；
  - **控制问题：** 在给定的置信度下，控制自变量x的取值范围，使Y在给定的范围内取值。

当进行n次试验, 则可得到n组观测样本值，记为

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

问题： 如何根据样本观测值确定Y关于X的回归函数$\mu(x)$?

- 根据专业知识或者经验知识来判断回归函数$\mu(x)$的形式
- 使用函数plot()把每对观测值(x_i, y_i)在直角坐标系中的散点图画出来

- 例：假定一保险公司希望确定居民住宅火灾造成的损失数额与该住户到最近的消防站的距离之间的相关关系，以便准确地定出保险金额。表列出了8起火灾事故的损失及火灾发生地与最近的消防站的距离。

Table: 火灾事故的损失及火灾发生地与最近的消防站距离的数据

距离x(单位：千米)	3.4	1.8	2.1	2.6	4.6	2.3	3.1	5.5
火灾损失y(单位：千元)	26.2	17.8	24.0	19.6	31.3	23.1	27.5	36.0

模型介绍

```
x = c(3.4, 1.8, 2.1, 2.6, 4.6, 2.3, 3.1, 5.5)
y = c(26.2,17.8,24.0,19.6,31.3,23.1,27.5,36.0)
plot(x, y)
```

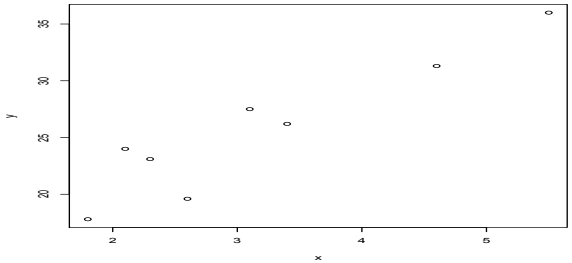


Figure: 火灾事故的损失与距消防站距离的散点图。

- 特别，当$\mu(x)$是线性函数，即$\mu(x) = \beta_0 + \beta_1 x$时，把模型称为一元线性回归模型。
- 在实际问题中，影响响应变量Y的因素往往有很多，这就需要考虑含有多个协变量的回归问题。
- 假设Y为响应变量，$X_1, \cdots, X_p$为p个协变量(或预测变量)，这时多元线性回归模型为：

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

其中

  - β_0为截距项，$\beta_1, \cdots, \beta_p$为回归系数
  - ε为随机模型误差

- 假设对$Y, X_1, \cdots, X_p$进行了n次独立的试验，得到n组观测值，即

$$y_i, x_{i1}, \cdots, x_{ip}, \qquad i = 1, \cdots, n$$

- 它们满足：

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

- 引进矩阵记号：

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- 和

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- 模型写成如下矩阵形式：

$$\boldsymbol{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}.$$

- 模型误差向量ε有如下的Gauss–Markov假设：
  1. $E(\varepsilon_i) = 0,\ \ i = 1, \cdots, n$;
  2. $\mathrm{Var}(\varepsilon_i) = \sigma^2,\ \ i = 1, \cdots, n$;
  3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0,\ \ i \neq j,\ \ i, j = 1, \cdots, n$。
- 进一步：$E(\boldsymbol{\varepsilon}) = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$。

■ 例1：考虑非线性模型：$Y = \alpha \exp(\beta X) \cdot \varepsilon,\ \ln(\varepsilon) \sim N(0, \sigma^2)$，其中$\alpha, \beta, \sigma^2$是与$X$无关的未知参数。

- 对非线性模型两边取对数，得

$$\ln(Y) = \ln(\alpha) + \beta X + \ln(\varepsilon).$$

- 若令$\widetilde{Y} = \ln(Y),\ \beta_0 = \ln(\alpha),\ \beta_1 = \beta,\ \widetilde{X} = X,\ \widetilde{\varepsilon} = \ln(\varepsilon)$，则可转化为一元线性回归模型：

$$\widetilde{Y} = \beta_0 + \beta_1 \widetilde{X} + \widetilde{\varepsilon}.$$

- **例2：** 考虑经济学中著名的Cobb–Douglas生产函数为：$Q_t = aL_t^b K_t^c$，其中Q_t, L_t, K_t分别为t年的产值、劳动力投入量和资金投入量，a, b和c为未知参数。
  - 对生产函数模型两边取对数，有

$$\ln(Q_t) = \ln(a) + b\ln(L_t) + c\ln(K_t).$$

  - 令$Y_t = \ln(Q_t)$, $X_{t1} = \ln(L_t)$, $X_{t2} = \ln(K_t)$，和$\beta_0 = \ln(a)$, $\beta_1 = b$, $\beta_2 = c$，上式右边加一项误差项，则转化为线性回归模型：

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t, \quad t = 1, \cdots, T.$$

■ 例3：考虑多项式回归模型：

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

- 令$X_1 = X,\ X_2 = X^2,\ X_3 = X^3$，则可转化成下面的线性回归模型：

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

■ 例4(多个自变量的多项式)：总所周知，任何光滑函数都可以用足够高阶的多项式来逼近，如

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon.$$

- **例5(单向方差分析回归)：** 比较三种药治疗某种疾病的效果，药效度量指标为Y。
  - 假设采用双盲试验法，即患者和医生都不知道服用三种药中的哪一种，只有试验设计和分析者掌握真实情况。
  - 假设现在对每种药各有n个患者服用，记y_{ij}为服用第i种药的第j个患者的药效测量值，则y_{ij}表示为：

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \cdots, n,$$

其中

- μ称为总平均，α_i表示第i种药的效应
- ε_{ij}表示随机误差，其均值为0，方差都相等，彼此互不相关

- 在这个问题中，感兴趣的因素只有一个，即药品，它有三个不同的品种，称这三个品种为因素的水平或处理，该模型称为单向因素方差分析回归模型。
- 用哑变量处理三种药品的均值：$\mu_1 = \mu + \alpha_1, \mu_2 = \mu + \alpha_2$和$\mu_3 = \mu + \alpha_3$。
- 令

$$x_{1j} = \begin{cases} 1, & \text{第}j\text{个患者服用第1种药;} \\ 0, & \text{其他,} \end{cases}$$

$$x_{2j} = \begin{cases} 1, & \text{第}j\text{个患者服用第2种药;} \\ 0, & \text{其他,} \end{cases}$$

$$x_{3j} = \begin{cases} 1, & \text{第}j\text{个患者服用第3种药;} \\ 0, & \text{其他,} \end{cases} \qquad j = 1, \cdots, n.$$

- 令

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \\ y_{31} \\ \vdots \\ y_{3n} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{3n} \end{pmatrix}.$$

- 单向因素方差分析回归模型可以表示成下面的多元线性回归模型：

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- 为了估计$\boldsymbol{\beta}$，记最小二乘目标函数为：

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

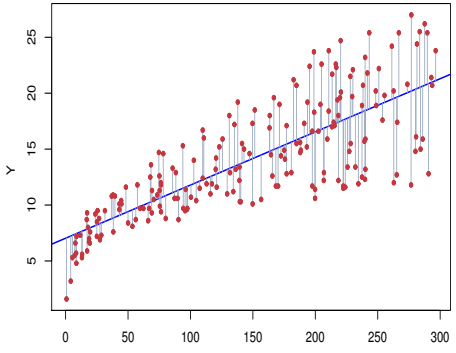
- 对$\boldsymbol{\beta}$求偏导数，并令其为零，则可以得到关于$\boldsymbol{\beta}$ 的正规方程：

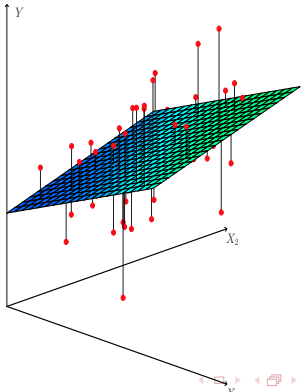
$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\boldsymbol{Y}.$$

- 正规方程有唯一解的充要条件是$\mathbf{X}'\mathbf{X}$ 的秩为$p+1$，或者矩阵$\mathbf{X}'\mathbf{X}$ 的逆存在。

- 解正规方程，可得$\boldsymbol{\beta}$的最小二乘估计为：

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





- 问题：$\widehat{\boldsymbol{\beta}}$是最小二乘目标函数$Q(\boldsymbol{\beta})$的最小值吗？
- 事实上，对任意一个$\boldsymbol{\beta}$，有

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

- 因为$\widehat{\boldsymbol{\beta}}$是正规方程的解，则有$\mathbf{X}'(\boldsymbol{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})=0$。

- 因此，对于任意的$\boldsymbol{\beta}$，有

$$\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|^2=\|\boldsymbol{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}).$$

- 因为$\mathbf{X}'\mathbf{X}$是一个正定矩阵，故上式第二项总是非负的，则有

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

- 等号成立当且仅当

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=0.$$

定理7.1.1

对于多元线性回归模型，最小二乘估计$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}$具有下列性质：

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

- 有了最小二乘估计$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)'$，可得下面的经验线性回归方程：

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p.$$

- 由于模型误差向量$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}$是不可观测的随机向量，可以考虑下面的残差向量，定义为

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

- 可以证明(课堂练习)：

$$\mathbf{X}'\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}, \qquad \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{\varepsilon}} = 0.$$

- 这时，可以定义残差平方和(RSS)为：

$$\mathrm{RSS} = \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2.$$

σ^2的估计

- 残差平方和(RSS)的大小反映了实际数据与理论模型的偏离程度或者拟合程度。
- RSS越小，说明数据与模型拟合得越好。

定理7.1.2

由残差平方和RSS的定义，可以得到σ^2 的无偏估计为：

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1}.$$

证明：由RSS和最小二乘估计$\widehat{\boldsymbol{\beta}}$的定义有

$$\begin{aligned}\mathrm{RSS} &= \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= [(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{Y}]'[(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{Y}] \\ &= \boldsymbol{Y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{Y},\end{aligned}$$

其中$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$称为帽子矩阵。

因为$E(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta}$和$\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2\mathbf{I}_n$，则有

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

利用$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$，可得

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

综上可得

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

定理7.1.3

对于线性回归模型，进一步假设随机模型误差向量$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$，则

1. $\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$；
2. $\dfrac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-p-1}$；
3. $\widehat{\boldsymbol{\beta}}$和RSS相互独立。

推论7.1.1

对于线性回归模型，若$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$，则

1. $\widehat{\beta}_k \sim N(\beta_k, \sigma^2 c_{k+1,k+1})$，其中$c_{k+1,k+1}$表示矩阵$(\mathbf{X}'\mathbf{X})^{-1}$的第$(k+1, k+1)$个元素；
2. 在β_k的一切线性无偏估计中，$\widehat{\beta}_k$是唯一方差最小者，其中$k = 0, 1, \cdots, p$。

- 由推论7.1.1，可得β_k的$100(1-\alpha)\%$的置信区间如下：

$$\widehat{\beta}_k \pm t_{n-p-1}\left(\frac{\alpha}{2}\right)\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}, \quad k=0,1,\cdots,p,$$

其中

- $\widehat{\mathrm{Var}}(\widehat{\beta}_k)=\widehat{\sigma}^2 c_{k+1,k+1}$
- $c_{k+1,k+1}$是矩阵$(\mathbf{X}'\mathbf{X})^{-1}$的第$(k+1,k+1)$个元素

定理7.1.4

对于多元线性回归模型，假设$\mathbf{X}$是$p+1$ 满秩矩阵，$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$，则$\boldsymbol{\beta}$ 的$100(1-\alpha)\%$的置信域为：

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1)\widehat{\sigma}^2 F_{p+1,n-p-1}(\alpha),$$

其中$F_{p+1,n-p-1}(\alpha)$是自由度为$p+1$和$n-p-1$的F分布的上侧α分位数。

- 由定理7.1.4，可构造β_k的$100(1-\alpha)\%$ 的同时置信区间如下：

$$\widehat{\beta}_k \pm \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}\sqrt{(p+1)F_{p+1,n-p-1}(\alpha)}, \quad k = 0, 1, \cdots, p.$$

- 考虑下面的假设检验：

$$H_{k0}: \ \beta_k = 0 \longleftrightarrow H_{k1}: \ \beta_k \neq 0, \quad k = 0, 1, \cdots, p.$$

- 由推论7.1.1，定理7.1.2和定理7.1.3，可以证明，当原假设H_{k0}成立时，统计量

$$T_k = \frac{\widehat{\beta}_k}{\widehat{\sigma}\sqrt{c_{k+1,k+1}}} \sim t_{n-p-1}, \quad k = 0, 1, \cdots, p.$$

- 如果$|T_k| \geq t_{n-p-1}(\alpha/2)$或$p_k = \Pr(t_{n-p-1} \geq |T_k|) < \alpha/2$时，则拒绝原假设$H_{k0}$，认为$\beta_k \neq 0$。

- 问题：如何鉴别用回归方程对观测值$\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$的拟合程度呢？
- 为了解决这个问题，需要考虑回归方程或回归模型的拟合优度(goodness of fit) 方法。
- 考虑下面的假设检验问题：

$$H_0: \quad \beta_0 = \beta_1 = \cdots = \beta_p = 0$$

$$H_1: \quad \beta_0, \beta_1, \cdots, \beta_p \text{不全为} 0.$$

- 考虑总平方和(SST)的分解：

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

其中$\bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$和$\widehat{y}_i=\widehat{\beta}_0+\widehat{\beta}_1x_{i1}+\cdots+\widehat{\beta}_px_{ip}$

- 由$\mathbf{X}'\widehat{\varepsilon}=\mathbf{0}$和$\widehat{\boldsymbol{Y}}'\widehat{\varepsilon}=0$，可证得上式中交叉项等于0。

- 总平方和SST可以分解为：

$$\text{SST} = \text{SSReg} + \text{RSS},$$

其中

$$\text{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad \text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- SSReg称为回归平方和，表示总平方和中被回归方程解释的那部分变异或离差，反映了协变量$\boldsymbol{X}$对响应变量Y 变动平方和的贡献。
- RSS 反映的是随机误差的变动对总平方和的贡献。

- 当原假设H_0成立时，可证得统计量

$$F = \frac{\mathrm{SSReg}/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

- 当$F > F_{p,n-p-1}(\alpha)$时，则拒绝原假设H_0，否则就接受原假设H_0。
- 检验统计量是把SSReg和RSS进行比较，当SSReg 相对RSS比较大时，就拒绝原假设，认为回归直线与样本观测值的拟合效果是显著的。

回归方程的拟合优度

- 对线性回归的拟合优度检验可使用下面的方差分析表进行解释。
- 对给定的显著性水平α，当$F \geq F_{p,n-p-1}(\alpha)$或$p_v = \Pr(F_{p,n-p-1} \geq F) < \alpha$时，拒绝原假设$H_0$，否则就接受原假设$H_0$。

方差来源	平方和	自由度	均方	F比
回归	SSReg	p	$\overline{\text{SSReg}} = \text{SSReg}/p$	$F = \overline{\text{SSReg}}/\overline{\text{RSS}}$
误差	RSS	$n-p-1$	$\overline{\text{RSS}} = \text{RSS}/(n-p-1)$	
总和	SST	$n-1$		

判定系数R^2

- 判定系数R^2定义为SSReg 占SST 的比例，即

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

- 判定系数R^2的取值在[0,1]之间
- 如果在SST 中SSReg所占的比重越大，这时判定系数R^2 越接近于1，则线性回归效果就越好，说明回归方程与样本观测值的拟合效果就越好

- 给定$\boldsymbol{x}_0 = (x_{01}, x_{02}, \cdots, x_{0p})'$，回归方程的真实值为：

$$y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon_0.$$

- 既然ε_0是未知的，只能忽略随机误差ε_0，则可得响应变量的点估计为：

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p}.$$

- 给定置信水平为$1-\alpha$，则可得y_0的预测区间为：

$$\left[\widehat{y}_0 \pm t_{n-p-1}(\alpha/2)\widehat{\sigma}\sqrt{1 + \widetilde{\boldsymbol{x}}_0'(\mathbf{X}'\mathbf{X})^{-1}\widetilde{\boldsymbol{x}}_0}\right],$$

其中$\widetilde{\boldsymbol{x}}_0 = (1, x_{01}, x_{02}, \cdots, x_{0p})'$。

- 当$p=1$时，则退化为一元线性回归模型：$Y=\beta_0+\beta_1 x+\varepsilon$。
- 给定$x=x_0$时，则y_0的置信水平为$1-\alpha$ 预测区间为：

$$\left[\widehat{y}_0 \pm t_{\alpha/2}(n-2)\widehat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right],$$

其中$\widehat{y}_0=\widehat{\beta}_0+\widehat{\beta}_1 x_0$ 和$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$

- 容易看到，该预测区间的长度是x_0 的函数，它随着$|x_0-\bar{x}|$的增加而增加。
- 当$x_0=\bar{x}$ 时，预测区间长度达到最短。

回归函数$\mu(\boldsymbol{x})$的点估计和置信区间

- 回归函数$\mu(\boldsymbol{x})$在$\boldsymbol{x}_0$的点估计为经验回归函数：

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p}.$$

- 可计算得$E(\widehat{y}_0) = \mu(\boldsymbol{x}_0)$，则可得$\mu(\boldsymbol{x}_0)$ 置信水平为$1-\alpha$的置信区间为：

$$\left[\widehat{y}_0 \pm t_{n-p-1}(\alpha/2)\widehat{\sigma}\sqrt{\widetilde{\boldsymbol{x}}_0'(\mathbf{X}'\mathbf{X})^{-1}\widetilde{\boldsymbol{x}}_0}\right].$$

- 当$p=1$时，可得$\mu(x_0) = \beta_0 + \beta_1 x_0$置信水平为$1-\alpha$的置信区间为：

$$\left[\widehat{y}_0 \pm t_{\alpha/2}(n-2)\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\right].$$

- 对于多元线性回归模型的应用，在R语言中，可用函数lm()进行计算，其调用格式为：

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE,
   y = FALSE, qr = TRUE, singular.ok = TRUE,
   contrasts = NULL, offset, ...)
```

其中formula为模型公式；data为数据框数据；subset为可选择向量，表示观测值的子集；weights 为可选择向量，表示用于数据拟合的权重；其余参数见在线帮助。

- 下面再介绍几个常用的函数：

```
anova(object, ...)
```

其中object为函数lm()和glm()得到的对象，其返回值为模型的方差分析表。

```
predict(object, newdata, se.fit=F, scale=NULL, df=Inf,
        interval=c("none", "confidence", "prediction"),
        level = 0.95, type = c("response", "terms"),
        terms = NULL, na.action = na.pass,
        pred.var = res.var/weights, weights = 1, ...)
```

其中object是由函数lm()得到的对象；newdata是预测点的数据框数据；选interval为"confidence"，返回值为回归函数的置信区间；选interval为"prediction"，返回值为Y的预测区间；其余参数见在线帮助。

```
plot(object, ...)
```
其中object是由函数lm()得到的对象，绘制模型诊断的几种图形，显示残差、拟合值和一些诊断情况。

```
confint(object, ...)
```
其中object是由函数lm()得到的对象，返回值为截距和回归系数的置信区间。

```
residuals(object, type =
          c("working", "response", "deviance",
            "pearson", "partial"))
```
其中object是由lm或aov构成的对象，type是返回值的类型，返回值为模型的残差。

```
summary(object, ...)
```
其中object是由lm构成的对象，返回值是显示较为详细的模型拟合结果。

例：31 名中年男性的健康数据

- 为了了解和预测人体吸入氧气的效率，收集了31名中年男性的健康状况调查资料。
- 共调查了7 项指标：吸氧效率(Y)、年龄(X_1，单位：岁)、体重(X_2，单位：千克)、跑1.5千米所需时间(X_3，单位：分钟)、休息时的心率(X_4，次/分钟)、跑步时的心率(X_5，次/分钟) 和最高心率(X_6，次/分钟)，数据见表。
- 在该资料中吸氧效率Y 作为响应变量，其他6 个变量作为协变量，建立多元线性回归模型，并进行统计分析。

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

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

```
library(GGally)
ggcoef(lm.reg, exclude_intercept = T, vline_color = "red",
       errorbar_color = "blue", errorbar_height = 0.1) + theme_bw()
```

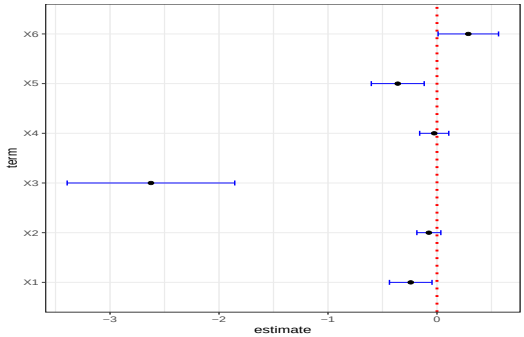


Figure: 回归系数置信区间的可视化。

1. 经验回归方程为：

$$Y = 104.86 - 0.24X_1 - 0.07X_2 - 2.62X_3 - 0.03X_4 - 0.36X_5 + 0.29X_6;$$

2. 变量X_1, X_3, X_5和X_6的p值小于$\alpha = 0.05$，可以认为它们是线性回归显著的，而变量X_2和X_4的p值大于$\alpha = 0.05$，认为它们是不显著的变量。此外，变量X_2和X_4回归系数的置信区间包含了0，也可以判定它们是不显著的；
3. 判定系数$R^2 = 0.8552$，说明经验回归方程对数据的拟合效果显著；
4. 回归方程的检验，F分布的p值为5.823×10^{-9}，远远小于显著性水平$\alpha = 0.05$，说明经验回归方程是显著的。

- 给定$\boldsymbol{x}_0 = (55, 86.26, 10.31, 66, 179, 176)'$，使用函数predict()，求$y_0$的估计值、$y_0$的置信水平为95%的预测区间和回归函数$\mu(\boldsymbol{x}_0)$的置信水平为95%的置信区间。

```
x.0 = data.frame(X1=55, X2=86.26, X3=10.31, X4=66, X5=179, X6=176)
Y.pred = predict(lm.reg, x.0, interval = "prediction", level=0.95)
> Y.pred
       fit      lwr      upr
  42.66741 36.8878 48.44703
mu.conf = predict(lm.reg, x.0, interval = "confidence", level=0.95)
> mu.conf
       fit      lwr      upr
 42.66741 39.27374 46.06109
```

什么是回归诊断？

回归诊断是对回归分析中的假设以及数据的检验与分析。

1. 误差项是否满足独立性、等方差性、正态性；
2. 选择多元线性模型是否合适；
3. 样本数据中是否存在异常值；
4. 回归分析的结果是否对某些样本的依赖性过重，即回归模型是否具备稳健性；
5. 自变量之间是否存在高度相关，即是否存在多重共线性问题。

什么是回归诊断？

- 在R语言中，下面函数与回归诊断有关。

`influence.measures`	`rstandard`
`rstudent`	`dffits`
`cooks.distance`	`dfbeta`
`dfbetas`	`covratio`
`hatvalues`	`hat`

普通残差

- 针对多元线性回归模型，可知残差为：

$$\widehat{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{Y}$$

- 帽子矩阵： $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- 在R语言中，函数residuals()提供了模型残差的计算
- 因此，得到残差后，可对残差进行检验，如正态性检验等

```
y.res = residuals(lm.reg)
shapiro.test(y.res)
#### 输出结果:
        Shapiro-Wilk normality test
data:  y.res
W = 0.9697, p-value = 0.5107
```

- 由模型误差$\boldsymbol{\varepsilon}$的性质，可知

$$E(\widehat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \qquad \mathrm{Cov}(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

- 因此，对每个残差$\widehat{\varepsilon}_i$，有

$$\frac{\widehat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1), \quad i = 1, \cdots, n$$

- 杠杆统计量：帽子矩阵$\mathbf{H}$对角线上的元素h_{ii}，称为杠杆统计量
- 作业：证明：杠杆值h_{ii}满足：$0 < h_{ii} < 1,\ i = 1, \cdots, n$，且 $\sum_{i=1}^{n} h_{ii} = p + 1$，和$1/n \le h_{ii} < 1$。

标准化(内学生化)残差

- 用$\widehat{\sigma}^2$ 作为σ^2的估计值，称

$$r_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$$

为标准化残差(standardized residual)，或称为内学生化残差(internally standardized residual)。

```
rstandard(model,infl=lm.influence(model,do.coef=FALSE),
        sd=sqrt(deviance(model)/df.residual(model)),...)
```

其中model是由lm或glm生成的对象，infl是由lm.influence返回值得到的影响结构，sd是模型的标准差。

外学生化残差

- 首先定义σ^2的估计为：

$$\widehat{\sigma}^2_{(-i)} = \frac{1}{n-p-2}\sum_{j\neq i}(y_j - \boldsymbol{x}_j\widehat{\boldsymbol{\beta}}_{(-i)})^2,$$

其中$\widehat{\boldsymbol{\beta}}_{(-i)}$为删除第$i$个样本后的最小二乘估计。

- 对$i = 1, \cdots, n$，称

$$\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$$

为学生化残差，或者称为外学生化残差。

- 函数rstudent()用来计算回归模型的(外)学生化残差

残差图

- 以残差$\widehat{\varepsilon}_i$为纵坐标，以拟合值$\hat{y}_i$或对应的数据观测序号i，或数据观测时间为横坐标的散点图统称为残差图。

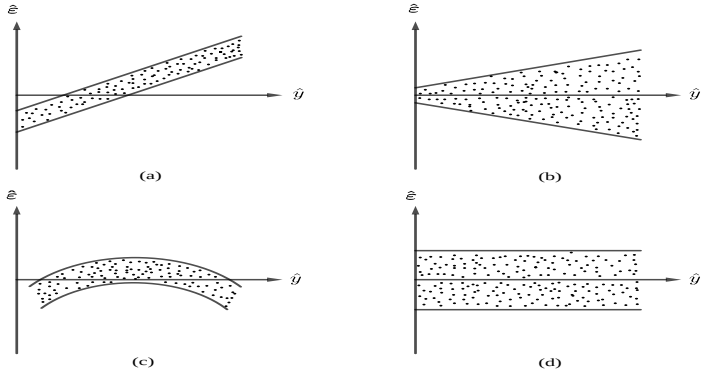


Figure: 回归值$\hat{y}$与残差的散点图。(a) 模型错误的情形；(b) 异方差情形；(c) 非线性情形；(d) 正常情形。

残差的QQ图

- 可用残差的QQ图检验残差的正态性，设$\widehat{\varepsilon}_{(i)}$表示残差$\widehat{\varepsilon}_i$的次序统计量，其中$i = 1, \cdots, n$。
- 令

$$q_{(i)} = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right), \qquad i = 1, \cdots, n,$$

其中

  - $\Phi(x)$为标准正态分布$N(0,1)$的分布函数，$\Phi^{-1}(x)$为其反函数
  - $q_{(i)}$为$\widehat{\varepsilon}_{(i)}$的期望值

- R语言中，可用函数plot(model, 2)绘制残差的QQ图，其中model是由lm生成的对象

杠杆统计量

- 如果某个样本不遵从回归模型，但是其余数据都遵从这个回归模型，则称该样本点为强影响点，也称为异常点。
- 响应变量的拟合值为：$\widehat{\boldsymbol{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = \mathbf{H}\boldsymbol{Y}$。
- 从几何上看，$\widehat{\boldsymbol{Y}}$是$\boldsymbol{Y}$在$\mathbf{X}$的列向量张成子空间内的投影，且满足：

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}.$$

- h_{ii}称为杠杆统计量，h_{ii}的大小可表示第i个样本值对$\hat{y}_i$影响的大小。

杠杆统计量

- $\widehat{y}_i$的方差为：$\mathrm{Var}(\widehat{y}_i) = \sigma^2 h_{ii}$。
- Hoaglin和Welsch (1978)给出了一种判断异常点的方法，当

$$h_{i_0 i_0} \geq \frac{2(p+1)}{n},$$

则可认为第i_0个样本点影响较大，可结合其他准则，考虑是否将其剔除。

```
hatvalues(model, infl = lm.influence(model, do.coef = FALSE), ...)
hat(x, intercept = TRUE)
```

其中model是由lm或glm生成的对象，x是设计矩阵X。

DFFITS准则

- Belsley等(1980)提供了另一种准则，考虑统计量

$$D_i(\widehat{\sigma}_{(-i)}) = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}.$$

- 对于第i个样本，如果有

$$|D_i(\widehat{\sigma}_{(-i)})| > 2\sqrt{\frac{p+1}{n}},$$

则认为第i个样本的影响比较大。

- 计算DFFITS准则的函数dffits()

Cook距离统计量

- Cook (1977)提出了Cook距离统计量，定义为：

$$C_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})}{(p+1)\widehat{\sigma}^2}, \quad i = 1, \cdots, n.$$

- Cook统计量可以改写为：

$$C_i = \frac{1}{p+1}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2, \quad i = 1, \cdots, n,$$

其中r_i是标准化残差。

- Cook距离统计量C_i越大的样本点，越可能是高影响点或异常点
- 计算Cook距离统计量的函数cooks.distance()

COVRATIO准则

- 分别计算$\widehat{\boldsymbol{\beta}}$和$\widehat{\boldsymbol{\beta}}_{(-i)}$的协方差矩阵：

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

- 对$i = 1, \cdots, n$，计算

$$C_i = \frac{\det\left(\widehat{\sigma}^2_{(-i)}(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\right)}{\det\left(\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\right)} = \frac{(\widehat{\sigma}^2_{(-i)})^{p+1}}{(\widehat{\sigma}^2)^{p+1}} \frac{1}{1-h_{ii}}.$$

- 如果第i个样本所对应的C_i值离1越远，则认为该样本影响越大。
- 计算COVRATIO统计量的函数covratio()

回归诊断和影响分析例子

```
par(mfrow = c(3, 2));  plot(lm.reg, which = 1:6)
n = nrow(health.data); p = ncol(health.data)-1
> health.data[hatvalues(lm.reg)>2*(p+1)/n,]
        Y X1    X2   X3 X4  X5  X6
10 60.055 38 81.87 8.63 48 170 186
> health.data[dffits(lm.reg)>2*sqrt((p+1)/n),]
        Y X1    X2    X3 X4  X5  X6
10 60.055 38 81.87  8.63 48 170 186
15 51.855 54 83.12 10.33 50 166 170
20 50.388 49 73.37 10.08 67 168 168
> health.data[cooks.distance(lm.reg)>0.1,]
        Y X1    X2    X3 X4  X5  X6
10 60.055 38 81.87  8.63 48 170 186
15 51.855 54 83.12 10.33 50 166 170
20 50.388 49 73.37 10.08 67 168 168
s = rep("", n);  co = covratio(lm.reg)
abs.co = abs(co-1); s[abs.co==max(abs.co)] = "*"
> data.frame(COVRATIO = co, s)
    COVRATIO s
12 2.3093822 *
```

回归诊断和影响分析例子

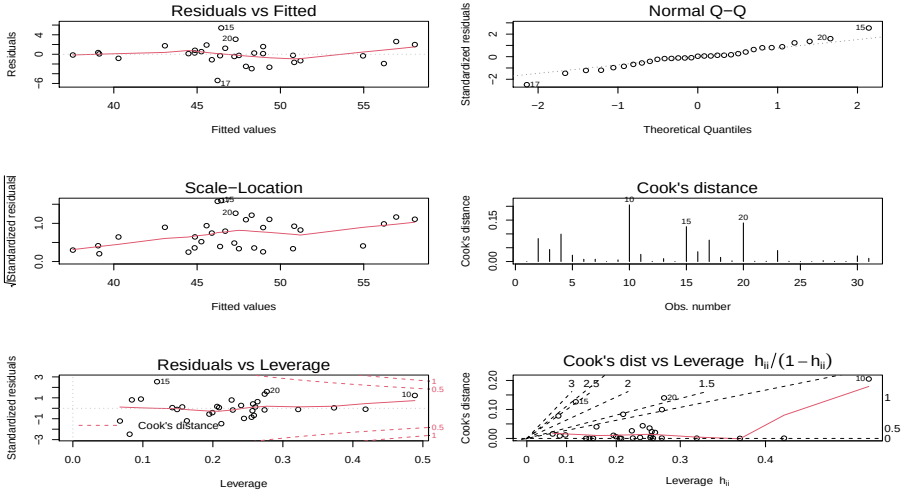


Figure: 31 名中年男性健康数据的模型诊断图。

多重共线性

- 多重共线性是指线性回归模型中的协变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。
- 如果存在不全为0的常数$a_0, a_1, a_2, \cdots, a_p$，使得

$$a_1X_1 + a_2X_2 + \cdots + a_pX_p = a_0.$$

- 如果数据中所有样本都满足上式成立，则称协变量$X_1, X_2, \cdots, X_p$存在精确共线性。
- 如果上式对观测数据近似成立，则有近似共线性，也就表示这p个协变量存在多重共线性。

多重共线性

- 假设$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$是协变量$X_1, X_2, \cdots, X_p$经过中心化和标准化得到的观测向量，记$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$为$n \times p$的设计矩阵
- 考虑多元线性回归模型：$\boldsymbol{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$，其中$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$
- 设λ为$\mathbf{X}'\mathbf{X}$的一个特征值，$\boldsymbol{\phi}$为其对应的特征向量，其长度为1，即$\boldsymbol{\phi}'\boldsymbol{\phi} = 1$
- 若$\lambda \approx 0$，则

$$\mathbf{X}'\mathbf{X}\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \approx \mathbf{0}.$$

- 用$\boldsymbol{\phi}'$左乘上式，得到

$$\boldsymbol{\phi}'\mathbf{X}'\mathbf{X}\boldsymbol{\phi} = \lambda\boldsymbol{\phi}'\boldsymbol{\phi} = \lambda \approx 0.$$

- 则有，$\mathbf{X}\boldsymbol{\phi} \approx \mathbf{0}$，其中$\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_p)'$，即

$$\phi_1\mathbf{x}_1 + \phi_2\mathbf{x}_2 + \cdots + \phi_p\mathbf{x}_p \approx \mathbf{0}.$$

- 若矩阵的某个特征值接近零，就意味着矩阵$\mathbf{X}$的列向量$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$之间存在近似线性关系。

- 下面讨论多重共线性问题对回归系数最小二乘估计的影响。
- 计算$\widehat{\boldsymbol{\beta}}$的均方误差为：

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i},$$

其中$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$是矩阵$\mathbf{X}'\mathbf{X}$的特征值

- 可见，如果矩阵$\mathbf{X}'\mathbf{X}$至少有一个特征值非常接近于零，则$\mathrm{MSE}(\widehat{\boldsymbol{\beta}})$就会很大，则最小二乘估计$\widehat{\boldsymbol{\beta}}$也就不再是回归系数$\boldsymbol{\beta}$的一个好的估计。

- 判断多重共线性及其严重程度的方法有：
  1. 条件数方法
  2. 方差膨胀因子(VIF)方法
- 条件数方法：度量多重共线性严重程度的一个重要指标是矩阵$\mathbf{X}'\mathbf{X}$的条件数，定义为：

$$\kappa(\mathbf{X}'\mathbf{X}) = \|\mathbf{X}'\mathbf{X}\| \cdot \|(\mathbf{X}'\mathbf{X})^{-1}\| = \frac{\lambda_{\max}(\mathbf{X}'\mathbf{X})}{\lambda_{\min}(\mathbf{X}'\mathbf{X})}$$

- $\lambda_{\max}(\mathbf{X}'\mathbf{X})$和$\lambda_{\min}(\mathbf{X}'\mathbf{X})$分别是矩阵$\mathbf{X}'\mathbf{X}$的最大和最小特征值

多重共线性

- 条件数刻画了矩阵$\mathbf{X}'\mathbf{X}$特征值差异性的大小
  - 若$\kappa < 100$，则认为多重共线性的程度很小
  - 若$100 \leq \kappa \leq 1000$，则认为存在中等程度或较强的多重共线性
  - 若$\kappa > 1000$，则认为存在严重的多重共线性
- R语言提供了计算条件数的函数kappa()

- 方差膨胀因子(VIF)是衡量多元线性回归模型中多重共线性严重程度的又一种度量。
- 假设模型中数据已进行中心和标准化，则回归系数最小二乘估计的协差矩阵为$\sigma^2\mathbf{R}_{\boldsymbol{X}}^{-1}$，其中$\mathbf{R}_{\boldsymbol{X}}$是协变量$\boldsymbol{X}$的样本相关矩阵。
- 因此，第k个回归系数估计$\widehat{\beta}_k$的方差为σ^2和矩阵$\mathbf{R}_{\boldsymbol{X}}^{-1}$中第$k$个对角线元素的乘积，则把矩阵$\mathbf{R}_{\boldsymbol{X}}^{-1}$中第$k$个对角线元素称为方差膨胀因子，记为$\mathrm{VIF}(\widehat{\beta}_k)$，其中$k=1,\cdots,p$。

- 可以证明：

$$\mathrm{VIF}(\widehat{\beta}_k) = \frac{1}{1 - R^2_{X_k|\boldsymbol{X}_{(-k)}}}, \quad k = 1, \cdots, p,$$

其中$R^2_{X_k|\boldsymbol{X}_{(-k)}}$是第$k$个协变量$X_k$与其余$p-1$个协变量$\boldsymbol{X}_{(-k)}$之间的判定系数

- VIF的取值为1，表示完全不存在多重共线性。
- 实际应用中，一个经验法则是当方差膨胀因子VIF的值超过5或10，就表示存在多重共线性问题。
- 可用程序包car中的函数vif()计算方差膨胀因子

例：薛毅和陈立萍(2007)的例6.18

- 考虑一个有6个协变量的线性回归问题，数据见表。
- 这里共有12 组数据，除第一组外，协变量$X_1, X_2, \cdots, X_6$的其余11组数据满足线性关系：

$$X_1 + X_2 + X_3 + X_4 = 10.$$

- 试用求矩阵条件数和方差膨胀因子的方法，分析出自变量间存在多重共线性。

序号	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	10.006	8	1	1	1	0.541	−0.099
2	9.737	8	1	1	0	0.130	0.070
3	15.087	8	1	1	0	2.116	0.115
4	8.422	0	0	9	1	−2.397	0.252
5	8.625	0	0	9	1	−0.046	0.017
6	16.289	0	0	9	1	0.365	1.504
7	5.958	2	7	0	1	1.996	−0.865
8	9.313	2	7	0	1	0.228	−0.055
9	12.960	2	7	0	1	1.380	0.502
10	5.541	0	0	0	10	−0.798	−0.399
11	8.756	0	0	0	10	0.257	0.101
12	10.937	0	0	0	10	0.440	0.432

多重共线性

```
library(car)
collinear = read.table("collinear.txt", header=TRUE)
XX = cor(collinear[2:7]); kappa(XX, exact=TRUE)
lm.fit = lm(Y~., data=collinear)
> round(vif(lm.fit), 3)
     X1       X2       X3       X4       X5       X6
182.052 161.362 266.264 297.715   1.920   1.455
```

- 得到的条件数是$\kappa = 2195.908 > 1000$，认为有严重的多重共线性。
- 矩阵$\mathbf{X}'\mathbf{X}$的最小特征值和相应的特征向量为：

$$\lambda_{\min} = 0.001106,$$

$$\phi = (0.4477, 0.4211, 0.5417, 0.5734, 0.0061, 0.0022)'.$$

多重共线性

- 可得：

$$0.4477\mathbf{x}_1 + 0.4211\mathbf{x}_2 + 0.5417\mathbf{x}_3 + 0.5734\mathbf{x}_4 + 0.0061\mathbf{x}_5 + 0.0022\mathbf{x}_6 \approx \mathbf{0}.$$

- 由于$\mathbf{x}_5$和$\mathbf{x}_6$的系数近似为0，因此有

$$0.4477\mathbf{x}_1 + 0.4211\mathbf{x}_2 + 0.5417\mathbf{x}_3 + 0.5734\mathbf{x}_4 \approx \mathbf{0}.$$

- 所以，存在不全为0的常数a_0, a_1, a_2, a_3, a_4，使得

$$a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 \approx a_0.$$

- 说明变量X_1, X_2, X_3, X_4存在多重共线性。

最优子集选择

- 问题：如何从协变量集$\{X_1, X_2, \cdots, X_p\}$中选出一个“最优子集"，使得这个子集中的变量对响应变量Y有显著性的影响？
- 针对多元线性回归模型，介绍几种子集选择的方法：
  1. 最优子集选择方法
  2. 逐步选择方法
  3. L_0惩罚模型选择方法

最优子集选择(best subset selection)

对p个协变量(或预测变量)的所有可能组合分别使用最小二乘回归方法进行拟合。对含有一个协变量的模型，需拟合p个模型；对含有两个协变量的模型，需拟合$C_p^2 = p(p-1)/2$个模型，依次类推。最后在所有可能模型中选取一个最优子模型。

最优子集选择的算法：

步骤1：记不含任何协变量的模型为零模型，用$\mathcal{M}_0$表示，该步只用于估计各观测的样本均值。

步骤2：对于$k = 1, 2, \cdots, p$：

1. 拟合C_p^k个包含k个协变量的模型；
2. 在C_p^k个模型中选择使RSS最小或判定系数R^2最大的模型作为最优子模型，记为$\mathcal{M}_k$。

步骤3：根据交叉(CV)测试预测误差、C_p、AIC、BIC 或者判定系数R^2 从$\mathcal{M}_1, \cdots, \mathcal{M}_p$个模型中选出一个最优子模型。

最优子集选择

- 最优子集选择方法简单直观，但是计算效率不高
- 步骤2在不同子集规模下进行模型选择，一共要拟合$C_p^1 + \cdots + C_p^p = 2^p - 1$个模型
- 包含$\mathcal{M}_0$模型，因此，共有2^p个模型，例如
  - 如果协变量的维数$p = 10$，则需要拟合1000 多个模型
  - 如果维数$p = 20$，则需要拟合超过100 万个模型
  - 如果维数$p = 30$，则需要拟合超过10亿个模型

向前逐步选择(forward stepwise selection)

向前逐步选择(forward stepwise selection)算法：

步骤1： 记不含任何协变量的模型为$\mathcal{M}_0$。

步骤2： 对于$k = 0, 1, \cdots, p-1$：

1. 从$p-k$个模型中进行选择，每个模型都在模型$\mathcal{M}_k$的基础上增加一个协变量；
2. 在$p-k$个模型中选择RSS最小或判定系数R^2最大的最优模型，记为$\mathcal{M}_{k+1}$。

步骤3： 根据交叉验证误差、C_p、AIC、BIC或者调整的判定系数R^2从$\mathcal{M}_0, \cdots, \mathcal{M}_p$个模型中选出一个最优模型。

向后逐步选择(backword stepwise selection)

向后逐步选择(backword stepwise selection)算法：

步骤1： 记包含全部p个协变量的模型为$\mathcal{M}_p$。

步骤2： 对于$k = p, p-1, \cdots, 1$：

1. 在k个模型中进行选择，在模型$\mathcal{M}_k$的基础上减少一个协变量，则模型只含$k-1$个协变量；
2. 在k个模型中选择RSS最小或判定系数R^2最大的最优模型，记为$\mathcal{M}_{k-1}$。

步骤3： 根据交叉验证误差、C_p、AIC、BIC或者调整的判定系数R^2从$\mathcal{M}_0, \cdots, \mathcal{M}_p$个模型中选出一个最优模型。

- 向前和向后逐步选择方法的计算复杂度为：

$$1+\sum_{k=0}^{p-1}(p-k)=1+p(p+1)/2$$

  - 当协变量的维数$p=10$，只需要拟合56个模型
  - 如果维数$p=20$，则仅需要拟合211个模型
  - 如果维数$p=30$，则仅需要拟合466个模型

- 步骤3，根据交叉测试预测误差、C_p、AIC、BIC 或者调整的判定系数R^2，从p个候选模型中选择一个最优模型。

最优模型选择

最优模型选择的方法有：

- C_p准则
- Akaike 信息准则(Akaike information criterion，简写为AIC)
- Bayes信息准则(Bayesian information criterian，简写为BIC)
- 风险膨胀准则(risk inflation criterion，简写为RIC)
- 调整的判定系数R^2 (adjusted R^2)

C_p准则

- Mallows (1973)提出C_p准则，通过采用最小二乘拟合一个包含d个协变量的模型，极小化下面的C_p准则，获得一个最优模型。
- C_p值的公式如下：

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2),$$

其中$\hat{\sigma}^2$是基于d个协变量得到的模型误差ε方差的估计

- Mallow's C_p 有时也被定义为：

$$C_p' = \frac{\text{RSS}}{\hat{\sigma}^2} - (n - 2d).$$

- 可以发现：

$$C_p = \frac{1}{n}\hat{\sigma}^2(C_p' + n).$$

- C_p统计量在训练集RSS 的基础上增加了惩罚项$2d\hat{\sigma}^2$，目的是调整训练误差倾向于低估测试误差。
- 显然，惩罚项$2d\hat{\sigma}^2$的大小随着d的增大而增大，但是可以调节由于变量个数d增加而不断降低训练集的RSS。
- 这时，可以选择使统计量C_p 达到最低的模型作为最优模型。

- 考虑下面的L_0惩罚最小二乘目标函数：

$$\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+n\lambda^2\sum_{j=1}^{p}I(|\beta_j|\neq 0),$$

其中

  - λ是调节参数(tuning parameter)
  - $I(\cdot)$是示性函数

- 当调节参数λ取不同的值，L_0惩罚最小二乘目标函数将对应不同的信息准则。

- 令$\lambda = \sigma\sqrt{2/n}$，则惩罚最小二乘目标函数变为AIC 准则，定义为：

$$\begin{aligned}\text{AIC} &= \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + n(\sigma\sqrt{2/n})^2 \sum_{j=1}^{p} I(|\beta_j| \neq 0) \\ &= \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2\sigma^2 \sum_{j=1}^{p} I(|\beta_j| \neq 0).\end{aligned}$$

- 用最小二乘方法拟合一个包含d个协变量的模型，ε方差估计记为$\widehat{\sigma}^2$，则AIC 统计量变为：

$$\text{AIC} = \frac{1}{n\widehat{\sigma}^2}(\text{RSS} + 2d\widehat{\sigma}^2).$$

- 令$\lambda = \sigma\sqrt{\log(n)/n}$，则$L_0$惩罚最小二乘目标函数退化为BIC 准则，定义为：

$$\text{BIC} = \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sigma^2 \log(n) \sum_{j=1}^{p} I(|\beta_j| \neq 0).$$

- 对于包含d 个预测变量的最小二乘模型，则BIC统计量变为：

$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n) d \widehat{\sigma}^2).$$

- 对任意的$n > 7$，有$\log(n) > 2$，可知BIC统计量通常给包含多个协变量的模型施加较重的惩罚。

- 令$\lambda = \sigma\sqrt{\log(p)/n}$，Foster 和George (1994) 提出下面的RIC模型选择方法，RIC 统计量定义为：

$$\text{RIC} = \frac{1}{n}(\text{RSS} + \log(p)d\widehat{\sigma}^2).$$

- 对任意的$7 < p < n$，有$2 < \log(p) < \log(n)$，可知RIC得到的模型规模比AIC 得到的模型小，而比BIC得到的模型要大。

- 首先判定系数R^2的定义：

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}$$

- 对于包含d个协变量的最小二乘模型，调整的判定系数R^2_{adj}统计量定义为：

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-d-1)}{\text{SST}/(n-1)}.$$

调整的判定系数R^2_{adj}

- 当模型中随着显著变量个数的逐渐增加，将使得$\text{RSS}/(n-d-1)$逐渐减小，导致R^2_{adj} 逐渐增大；
- 当模型中包含了所有正确的协变量后，再增加其他噪声变量只会导致RSS 小幅度的减小；
- 由于加入这些噪声变量的同时增加了d的值，因此这些协变量的加入会导致$\text{RSS}/(n-d-1)$的增大，从而降低R^2_{adj}的值；
- 理论上，拥有最大R^2_{adj}的模型只包含了正确的协变量，而没有噪声变量。

- 在R语言中，使用函数step()进行变量选择和选取“最优子集"

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

其中object是函数lm()或glm()分析的结果，scope是确定逐步搜索的区域，direction确定逐步搜索的方向："both"是"一切子集回归法"，"backward"是后退法(只减少变量)，"forward" 是前进法(只增加变量)，默认值为"both"。k为正数，表示自由度数目的倍数，只有当k=2时，才能给出真正的AIC。当k=log(n)时，是BIC 准则，其中n表示样本量大小。其他参数见在线帮助，使用命令?step。

```
lm.aic = step(lm.reg)
#### 输出结果:
Start:  AIC=56.8
Y ~ X1 + X2 + X3 + X4 + X5 + X6
       Df Sum of Sq    RSS    AIC
- X4    1     0.787 124.10 55.001
<none>              123.32 56.804
- X2    1    10.053 133.37 57.233
- X6    1    23.545 146.86 60.221
- X1    1    33.272 156.59 62.209
- X5    1    48.153 171.47 65.023
- X3    1   255.036 378.35 89.557
Step:  AIC=55
Y ~ X1 + X2 + X3 + X5 + X6
       Df Sum of Sq    RSS    AIC
<none>              124.10 55.001
- X2    1      9.58 133.69 55.307
- X6    1     23.97 148.07 58.475
- X1    1     32.68 156.79 60.248
- X5    1     49.94 174.04 63.484
- X3    1    327.46 451.56 93.041
```

1. 如果用全部变量作回归方程时，AIC统计量的值为56.8，如果去掉变量X_4时，AIC 统计量的值为55.001；
2. 如果去掉变量X_2时，AIC统计量的值为57.233，依次类推；
3. 由于去掉变量X_4使AIC统计量的取值达到最小，因此step() 函数会自动去掉变量X_4，进入下一轮计算；
4. 在下一轮中，无论去掉哪一个变量，AIC统计量的取值都会升高，这时自动终止计算，得到最优回归方程。

```
summary(lm.aic)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X5 + X6, data = health.data)
Residuals:
    Min      1Q  Median      3Q     Max
-5.4714 -0.9249 -0.0131  0.9594  5.4054
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.99470   11.71953   8.874 3.38e-09 ***
X1           -0.23223    0.09051  -2.566  0.01667 *
X2           -0.07237    0.05209  -1.389  0.17697
X3           -2.68692    0.33082  -8.122 1.78e-08 ***
X5           -0.36462    0.11496  -3.172  0.00398 **
X6            0.28996    0.13196   2.197  0.03747 *
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.228 on 25 degrees of freedom
Multiple R-squared:  0.8542,    Adjusted R-squared:  0.8251
F-statistic:  29.3 on 5 and 25 DF,  p-value: 1.084e-09
```

```
lm.bic = step(lm.reg, k=log(length(health.data$Y)), trace=FALSE)
summary(lm.bic)
Call:
lm(formula = Y ~ X1 + X3 + X5 + X6, data = health.data)
Residuals:
    Min      1Q  Median      3Q     Max
-4.9590 -1.2603 -0.1512  1.1796  4.8132
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.07910   11.57739   8.644 4.02e-09 ***
X1           -0.21266    0.09099  -2.337  0.02740 *
X3           -2.76824    0.33138  -8.354 7.79e-09 ***
X5           -0.33957    0.11555  -2.939  0.00683 **
X6            0.25535    0.13188   1.936  0.06378 .
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.268 on 26 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.8188
F-statistic:  34.9 on 4 and 26 DF,  p-value: 4.219e-10
```

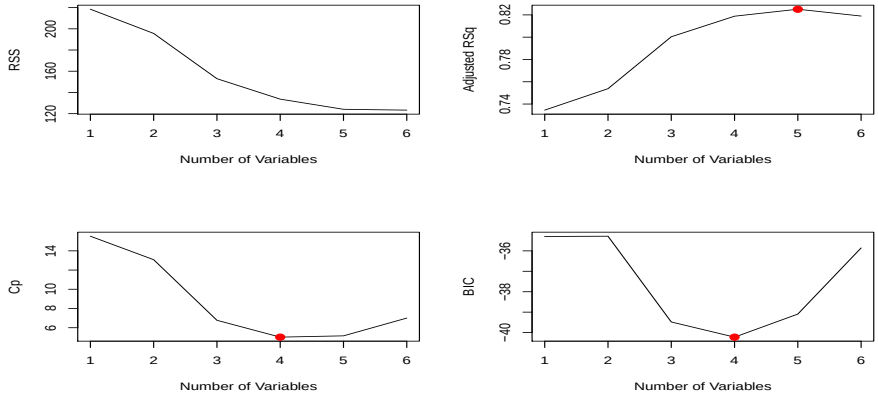
- 使用BIC准则，从模型中进一步去掉了变量X_2，但是判定系数R^2变化比较小，可以看出AIC 准则比较保守。
- 因此，最后得到经验回归方程为：

$$\begin{aligned} Y &= 100.07910 - 0.21266X_1 - 2.76824X_3 \\ &\quad -0.33957X_5 + 0.25535X_6. \end{aligned}$$

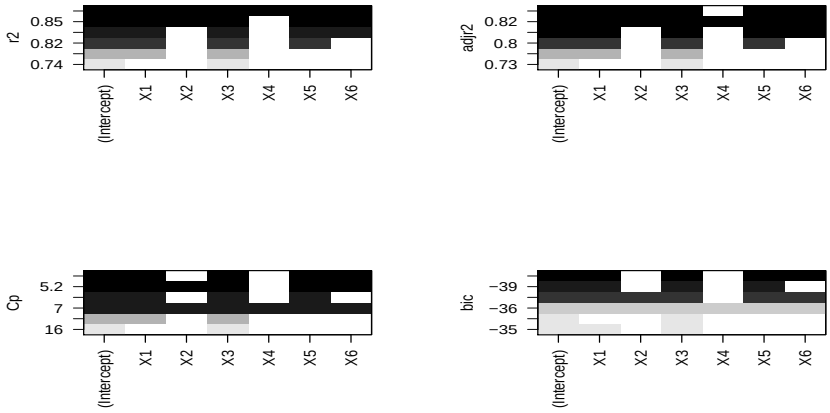
- 下面用程序包leaps中的函数regsubsets()来实现最优变量子集的筛选

```
library(leaps); health = read.table("health.txt", header=TRUE)
regfit.full = regsubsets(Y~., health)
reg.summary = summary(regfit.full); reg.summary
#### 输出结果:
Subset selection object
Call: regsubsets.formula(Y ~ ., health)
6 Variables  (and intercept)
   Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X5  X6
1  ( 1 ) " " " " "*" " " " " " "
2  ( 1 ) "*" " " "*" " " " " " "
3  ( 1 ) "*" " " "*" " " "*" " "
4  ( 1 ) "*" " " "*" " " "*" "*"
5  ( 1 ) "*" "*" "*" " " "*" "*"
6  ( 1 ) "*" "*" "*" "*" "*" "*"
```

- 绘制出所有模型的RSS、调整的判定系数R^2_{adj}、C_p和BIC的图形，辅助确定最终选择哪一个最优模型



- 进一步，通过绘制碎石图来展示最优子集的结果



- 考虑多元线性回归模型，假设对$Y, X_1, \cdots, X_p$进行了n次独立的试验，得到n组独立的观测值，即

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

- 为了简单，响应变量作中心化，对协变量数据进行标准化处理，使得$\sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1$。
- 标准化后的数据满足：

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \cdots, n.$$

- 线性模型的矩阵形式为：$\boldsymbol{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$.

压缩估计方法

本章介绍几种压缩估计方法：

- 岭回归(ridge regression)
- 桥回归(bridge regression)
- Lasso (Tibshirani, 1996)
- SCAD (Fan 和Li, 2001)
- 自适应Lasso (Zou, 2006)

岭回归(ridge regression)

- 极小化下面的惩罚最小二乘目标函数，可得到未知回归系数$\boldsymbol{\beta}$的岭回归估计$\widehat{\boldsymbol{\beta}}^R = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)'$：

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2,$$

其中

  - $\lambda \geq 0$是一个调节参数(tuning parameter)
  - λ的作用是控制回归系数估计的相对影响程度，可以通过数据驱动的方法进行选取
  - $\lambda\sum_{j=1}^{p}\beta_j^2$ 是压缩惩罚

- 极小化惩罚最小二乘目标函数，求得岭回归估计为：

$$\widehat{\boldsymbol{\beta}}^R = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)' = (\mathbf{X}'\mathbf{X} + n\lambda \mathbf{I}_p)^{-1}\mathbf{X}'\boldsymbol{Y}.$$

- 当$\lambda = 0$时，岭回归估计就是最小二乘估计，即惩罚项不起任何作用
- 随着$\lambda \to \infty$，惩罚项的作用增强，岭回归估计也会随着λ增大越来越接近于0
- 岭回归的优势是平衡了偏差和方差，随着λ的增加，岭回归拟合的光滑度降低，尽管方差变小，但是偏差变大

- 广义交叉验证(GCV)方法：极小化下面的GCV目标函数，获得最优的调节参数λ，即

$$\widehat{\lambda}_{\text{gcv}} = \arg\min_{\lambda} \text{GCV}(\lambda) = \arg\min_{\lambda} \frac{\frac{1}{n}\|(\mathbf{I}_n - \mathbf{H}(\lambda))\boldsymbol{Y}\|_2^2}{[n^{-1}\text{tr}(\mathbf{I}_n - \mathbf{H}(\lambda))]^2},$$

其中$\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I}_p)^{-1}\mathbf{X}'$。

- 与最优子集选择方法相比，岭回归方法计算简便。

- 极小化惩罚最小二乘目标函数，等价于求解约束的最小二乘问题：

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} \dfrac{1}{n}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2, \\ \text{s.t.} \quad \sum\limits_{j=1}^{p}\beta_j^2 \le c, \end{cases}$$

其中

- c是一个非负的常数，作用相同于调节参数λ
- 通过c 的大小控制$\sum\limits_{j=1}^{p}\beta_j^2$的大小
- 当c的取值为无穷大时，则约束项不起作用
- 随着$c \to 0$，约束项的作用增强

- R语言中，岭回归的计算：
  - 程序包ridge中的函数linearRidge()
  - 程序包MASS中的函数lm.ridge()
  - 程序包glmnet中的函数glmnet()，其中函数glmnet()中，alpha=0，拟合岭回归模型；alpha=1，拟合Lasso 模型
- 例： 31 名中年男性健康数据的岭回归分析

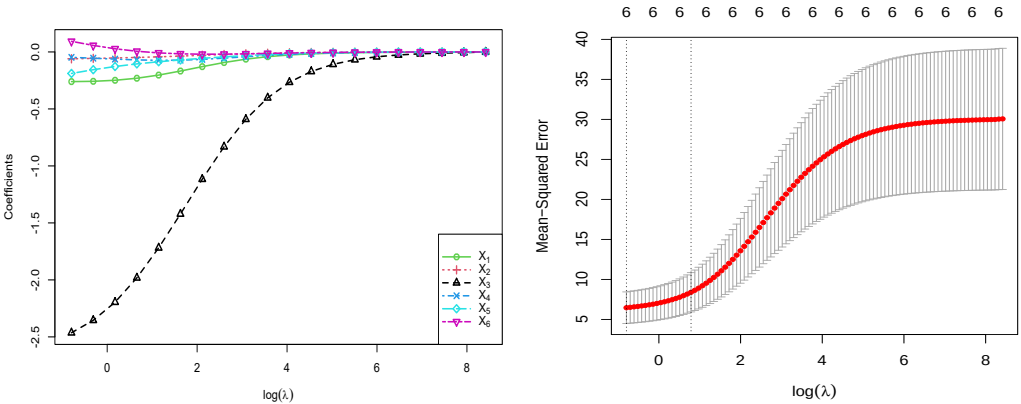


Figure: 左图：岭回归估计随着λ变化的路径图；右图：岭回归的交叉验证误差图。

- 从左图可看出，随着调节参数λ的增大，岭回归估计向原点收缩，但并不会使任何回归系数严格等于零。
- 右图：横轴为$\log(\lambda)$，而纵轴为交叉验证误差(即$\mathrm{CV}(\lambda) = \overline{\mathrm{MSE}}(\lambda)$)
- 右图还显示了交叉验证误差的正、负标准差，即$\pm \mathrm{sd}_{\mathrm{MSE}}(\lambda)$
- 图中左边的垂直虚线表示能使交叉验证误差最小的$\log(\widehat{\lambda})$取值，使用cv.ridge$lambda.min获得最优的调节参数为$\widehat{\lambda} \approx 0.452$
- 右边的垂直虚线表示比$\widehat{\lambda}$更大，且与$\mathrm{CV}(\widehat{\lambda})$相距一个标准差$\mathrm{sd}_{\mathrm{MSE}}(\widehat{\lambda})$的调节参数的取值
- 使用cv.ridge$lambda.1se提取$\widetilde{\lambda}$的值为$\widetilde{\lambda} \approx 2.197$

岭回归

```
library(glmnet); set.seed(2021)
health = read.table("health.txt", header=TRUE)
x=model.matrix(Y~., health)[,-1];  y=health$Y
fit_ridge = glmnet(x, y, alpha = 0, nlambda = 20)
lam = fit_ridge$lambda
cv.ridge = cv.glmnet(x, y, alpha = 0)
plot(cv.ridge)                  ## 绘制交叉验证误差图
> cv.ridge$lambda.min           > cv.ridge$lambda.1se
  [1] 0.4518421                   [1] 2.197128
> coef(cv.ridge, s="lambda.min")    > coef(cv.ridge, s="lambda.1se")
7 x 1 sparse Matrix of class "dgCMatrix"
                      1                                 1
(Intercept) 108.88840742          (Intercept) 102.45932307
X1           -0.26026890          X1           -0.22433706
X2           -0.06010871          X2           -0.04709489
X3           -2.46384363          X3           -1.91745239
X4           -0.04742507          X4           -0.07122858
X5           -0.18868679          X5           -0.09953895
X6            0.09329519          X6            0.00161379
```

桥回归(bridge regression)

- Frank 和Friedman (1993)提出桥回归，即考虑下面的L_q惩罚最小二乘目标函数：

$$\frac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda^*\sum_{j=1}^{p}|\beta_j|^q,$$

其中$\lambda^*=\lambda/q$，且$0\leq q\leq 2$。

桥回归(bridge regression)

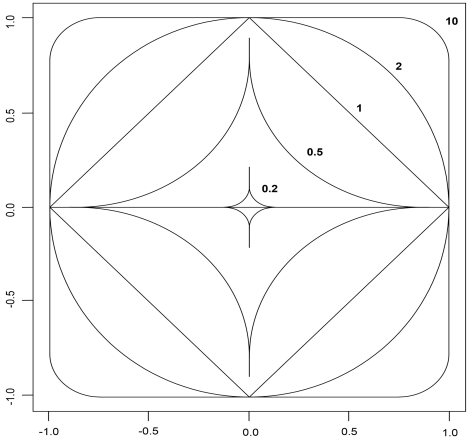


Figure: L_q惩罚函数的等值线图，$q = 0.2, 0.5, 1, 2$和10。

桥回归(bridge regression)

- 极小化L_q惩罚最小二乘目标函数，等价于求解下面约束的最小二乘问题：

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} \dfrac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2, \\ \text{s.t.} \quad \sum\limits_{j=1}^{p}|\beta_j|^q \le c, \end{cases}$$

其中c是一个非负的常数。

- 桥回归的R语言应用，可见

http://statweb.stanford.edu/~ jhf/R-GPS.html

惩罚变量选择方法

惩罚变量选择方法：

- Lasso (Tibshirani, 1996, JRSSB)
- SCAD (Fan and Li, 2001, JASA)
- Adaptive Lasso (Zou, 2006, JASA)
- 桥回归(Frank and Friedman, 1993, Technometrics)
- Elastic Net (Zou and Hastie, 2005, JRSSB)
- MCP (Zhang, 2010, AOS)
- Dantzig (Candes and Tao, 2007, AOS)

优点：

- 计算量小, 可以同时进行变量选择和参数估计
- 统计性质很容易证明

- 为了选择对响应变量Y有显著影响的协变量，即进行变量选择，考虑下面的惩罚最小二乘目标函数：

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

其中

- $p_\lambda(\cdot)$是惩罚函数
- λ是调节参数或截断参数，是用来控制模型的复杂度
- 采用交叉验证(CV)方法、广义交叉验证(GCV) 方法或BIC等数据驱动的准则进行选取λ

惩罚变量选择方法

- 惩罚变量选择方法的优点是计算量小，可以同时进行变量选择和参数估计，而且统计性质很容易证明
- Fan 和Li (2001) 建议一个好的惩罚函数将导致具有三个性质的估计量：
  1. 无偏性：当真参数很大时，得到的估计量是渐近无偏的，以避免不必要的建模偏差；
  2. 稀疏性：所得到的估计量是一个门限值，自动把小的参数分量估计成0，以便减少模型的复杂性；
  3. 连续性：所得估计量在数据点处是连续的，避免模型预测的不稳定性。

- 为更好理解惩罚最小二乘惩变量选择方法，考虑简单的正交情形。
- 假设$\mathbf{X}'\mathbf{X}/n = \mathbf{I}_p$，这时最小二乘估计为：

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = \mathbf{X}'\boldsymbol{Y}/n$$

- 进一步，最小二乘目标函数为：

$$\frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|_2^2 + \frac{1}{2n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

且

$$\frac{1}{2n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}\|_2^2 = \frac{1}{2}\sum_{j=1}^{p}(\widehat{\beta}_{j,\mathrm{LS}} - \beta_j)^2$$

- $\widehat{\beta}_{j,\mathrm{LS}} = (\mathbf{X}'\boldsymbol{Y}/n)_j$，即表示第$j$个分量的最小二乘估计，且$j = 1, \cdots, p$。
- 令$z_j = \widehat{\beta}_{j,\mathrm{LS}}$表示$\boldsymbol{\beta}$的第$j$个分量的最小二乘估计，其中$j = 1, \cdots, p$。
- 对于正交情形，惩罚最小二乘目标函数变为：

$$\frac{1}{2}\sum_{j=1}^{p}(z_j - \beta_j)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

- 为更好理解惩罚变量选择方法，首先考虑下面一般形式的惩罚最小二乘目标函数：

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|).$$

- 上式的导数为：

$$\theta - z + p'_\lambda(|\theta|)\mathrm{sgn}(\theta) = \mathrm{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z.$$

- Fan 和Li (2001)讨论了满足上面三个性质的惩罚函数所应满足的条件，结论是：

1. 无偏性：取值较大的真参数估计具有无偏性的充要条件是对取值较大的$|\theta|$ 有$p'_\lambda(|\theta|) = 0$;
2. 稀疏性：具有稀疏性的充分条件是

$$\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} > 0;$$

3. 连续性：具有连续性的充要条件是

$$\arg\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} = 0.$$

惩罚变量选择方法

- L_2惩罚函数：$p_\lambda(|\theta|) = \lambda|\theta|^2$
- 极小化惩罚最小二乘函数，得到的是岭回归估计
- 明显，L_2 惩罚函数在原点处不是奇异的，因此L_2惩罚函数不能产生稀疏解
- L_2 惩罚函数的一个推广形式是L_q惩罚函数：$p_\lambda(|\theta|) = \lambda|\theta|^q$，$q > 1$
- 这类惩罚函数只能减小估计的方差，产生的是有偏估计，但是不具有稀疏性

- L_1惩罚函数：$p_\lambda(|\theta|) = \lambda|\theta|$
- 极小化惩罚最小二乘函数，产生一个软门限解：

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

- Tibshirani (1996) 把L_1 惩罚函数施加于回归模型的一般最小二乘和似然函数，提出了Lasso变量选择方法
- Lasso 变量选择方法尽管可以产生稀疏解，并满足连续性，但是所得估计不具有无偏性。

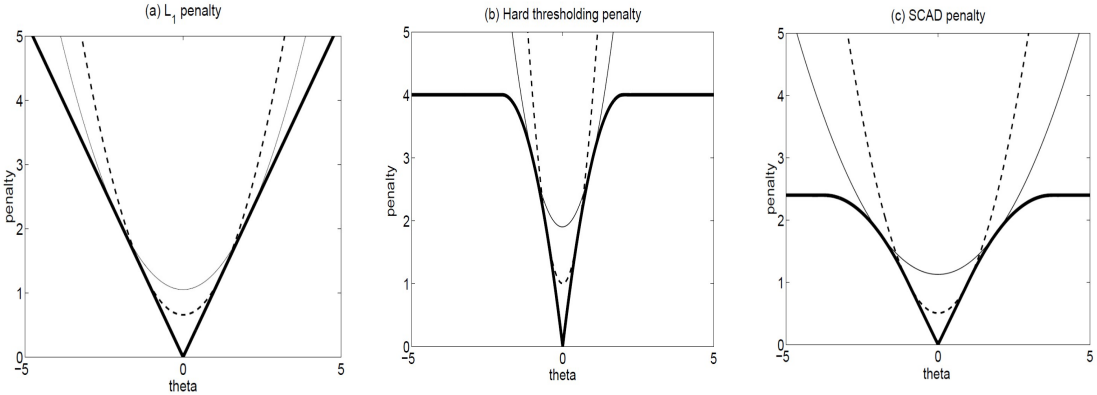


Figure: 三个惩罚函数$p_\lambda(|\theta|)$和它们的二次逼近，$\lambda = 2$，SCAD惩罚函数中取$a = 3.7$。

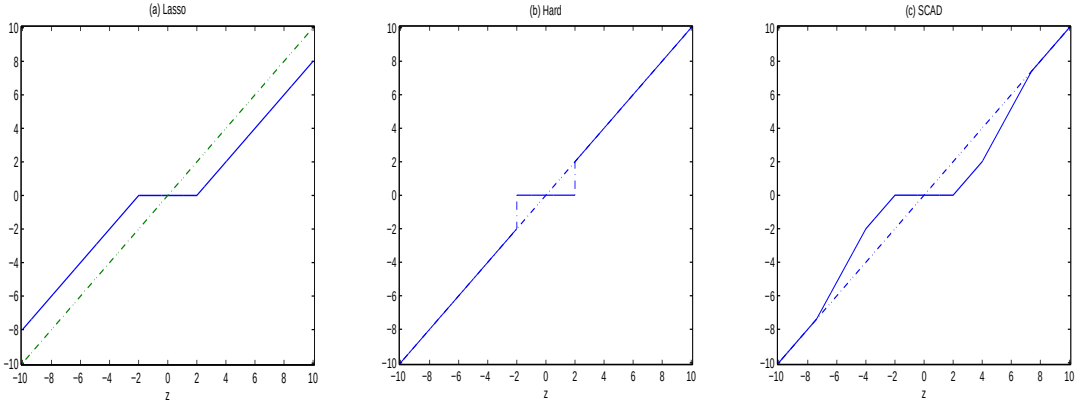


Figure: 门限函数图：(a) Lasso门限；(b) 硬门限；(c) SCAD门限，其中$\lambda = 2$，$a = 3.7$。

- Antoniadis (1997)提出了如下的硬门限惩罚函数：

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

- 如果施加硬门限惩罚函数，可以得到如下的硬门限解：

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

- 硬门限解满足无偏性和稀疏性，但是对数据点z，不满足连续性。

- Fan (1997) 提出了一个连续可微的惩罚函数，称为SCAD惩罚函数，定义为：

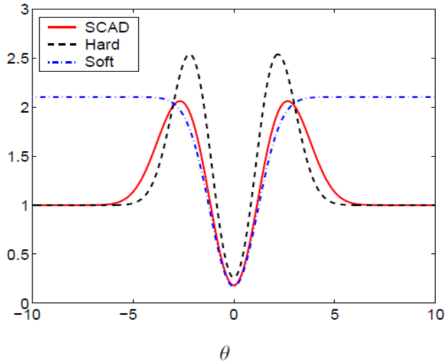
$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \le \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\},$$

其中$a > 2$。

- SCAD 惩罚最小二乘解为：

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

- Fan 和Li (2001)证明SCAD惩罚函数可以同时满足无偏性，稀疏性和连续性，并且具有oracle性质。
- 令$z \sim N(\theta, 1)$，考虑风险函数为：$R(\widehat{\theta}, \theta) = E_\theta(\widehat{\theta} - \theta)^2$。



与L_0惩罚函数进行比较

- 为了方便比较，考虑下面一般形式的L_0惩罚最小二乘目标函数：

$$\frac{1}{2}(z-\theta)^2+\frac{\lambda^2}{2}I(|\theta|\neq 0).$$

- 对C_p和AIC准则，取$\lambda=\sigma\sqrt{2/n}$，可得θ的解为：

$$\widehat{\theta}=zI\left(|z|\geq\sigma\sqrt{2/n}\right).$$

- 考虑多元线性模型，对于正交情形，即$n^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}_p$，假设最小二乘估计为：

$$z =: \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = \mathbf{X}'\boldsymbol{Y}/n.$$

- 则最小二乘估计的协方差矩阵为：

$$\mathrm{Cov}(z|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = (\sigma^2/n)\mathbf{I}_p$$

- 可得$z_j = \widehat{\beta}_{j,\mathrm{LS}}$的标准差为：$\mathrm{SE}(z_j) = \sigma/\sqrt{n}$

- 推广到一般情形的解，则有

$$\widehat{\theta} = zI\left(|z| \geq \sqrt{2}\mathrm{SE}(z)\right) = zI\left(\frac{|z|}{\mathrm{SE}(z)} \geq \sqrt{2}\right).$$

- 意味着：如果t检验统计量的值满足$|t| \leq \sqrt{2}$或p 值≥ 0.1573时，将删掉该变量。
- 对BIC准则，取$\lambda = \sigma\sqrt{\log(n)/n}$，可得$\theta$ 的解为：

$$\widehat{\theta} = zI\left(\frac{|z|}{\mathrm{SE}(z)} \geq \sqrt{\log(n)}\right).$$

- 意味着：如果t检验统计量的值满足：$|t| \leq \sqrt{\log(n)}$ 时将删掉该变量。

Table: BIC 准则的p 值

n	50	100	150	200	400	800	1600
$\sqrt{\log(n)}$	1.9779	2.1460	2.2384	2.3018	2.4477	2.5855	2.7162
p值	0.0479	0.0319	0.0252	0.0213	0.0144	0.0097	0.0066

- 对RIC准则，$\lambda = \sigma\sqrt{\log(p)/n}$，可得$\theta$ 的解为：

$$\widehat{\theta} = zI\left(\frac{|z|}{\mathrm{SE}(z)} \geq \sqrt{\log(p)}\right).$$

- 意味着：如果t检验统计量的值满足：$|t| \leq \sqrt{\log(p)}$ 时将删掉该变量

Table: RIC 准则的p 值

p	10	20	30	40	50	100	200
$\sqrt{\log(p)}$	1.5174	1.7308	1.8442	1.9206	1.9779	2.1460	2.3018
p值	0.1292	0.0835	0.0651	0.0548	0.0479	0.0319	0.0213

Lasso方法

- 针对多元线性回归模型，极小化下面的L_1 惩罚最小二乘目标函数，可得回归系数β的Lasso 估计$\widehat{\beta}^L$：

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

- Lasso估计$\widehat{\beta}^L$也等价于求解下面的约束优化问题：

$$\begin{cases}\min\limits_{\beta}\left\{\dfrac{1}{2n}\sum\limits_{i=1}^{n}\left(y_i-\sum\limits_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}, \\ \text{s.t.} \quad \sum\limits_{j=1}^{p}|\beta_j|\le c.\end{cases}$$

Lasso方法

- 寻找Lasso 估计，就是寻找最优的调节参数λ或控制最优的c，找使得RSS最小的回归系数的估计
- 通过一些数据驱动的方法选取λ，如CV，GCV 或BIC 准则
- Lasso解的问题，可使用R 程序包：
  - glmnet
  - gcdnet
  - lars
- 下面对维数$p = 2$ 时，说明为什么Lasso 可以产生稀疏模型，而岭回归不可以？

Lasso方法

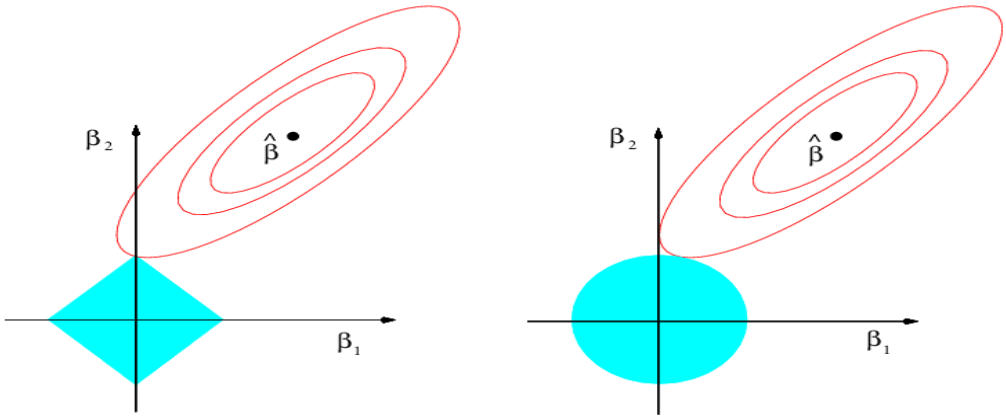


Figure: 误差等高线和限制条件区域，左：Lasso；右：岭回归。椭圆是RSS 等高线，实心区域是限制条件：$|\beta_1| + |\beta_2| \leq c$ 和$\beta_1^2 + \beta_2^2 \leq c$；$\widehat{\boldsymbol{\beta}}$为最小二乘估计。

自由度

- Hastie 和Tibshirani (1990)定义了有效的自由度，它表示模型的复杂度
- 例如，对于估计$\widehat{\boldsymbol{\mu}} = \mathbf{S}\boldsymbol{Y}$，则自由度定义为：$df(\widehat{\boldsymbol{\mu}}) = \text{tr}(\mathbf{S})$。

引理7.5.1：Stein引理

假设$\widehat{\boldsymbol{\mu}} : \mathbb{R}^n \to \mathbb{R}^n$是几乎处处可微的，且令$\nabla \cdot \widehat{\boldsymbol{\mu}} = \sum_{i=1}^{n} \partial \widehat{\mu}_i / \partial y_i$。如果$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$，则

$$\sum_{i=1}^{n} \text{Cov}(\widehat{\mu}_i, y_i)/\sigma^2 = E[\nabla \cdot \widehat{\boldsymbol{\mu}}].$$

自由度

- 对$\boldsymbol{Y}$的一个线性算子拟合：$\widehat{\boldsymbol{\mu}} = \mathbf{S}\boldsymbol{Y}$，有$\nabla \cdot \widehat{\boldsymbol{\mu}} = \sum_{i=1}^{n} s_{ii} = \mathrm{tr}(\mathbf{S})$。
- 对线性模型的最小二乘拟合：

$$\widehat{\boldsymbol{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}.$$

- 则最小二乘估计的自由度可以定义为：

$$\begin{aligned} df(ols) &= \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= \mathrm{tr}(\mathbf{I}_p) = p. \end{aligned}$$

自由度

- 对岭回归拟合：$\widehat{\boldsymbol{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X} + n\lambda \mathbf{I}_p)^{-1}\mathbf{X}'\boldsymbol{Y}$。
- 则岭回归估计的自由度为：

$$df(ridge) = \mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + n\lambda \mathbf{I}_p)^{-1}) = \sum_{j=1}^{p} \frac{\gamma_j}{\gamma_j + \lambda}$$

  - γ_j是矩阵$\mathbf{X}'\mathbf{X}/n$ 的第j个特征值，且$j = 1, \cdots, p$

- 对给定的调节参数λ，令$\widehat{\boldsymbol{\beta}}(\lambda)$表示回归系数向量$\boldsymbol{\beta}$的Lasso 估计
- 则Lasso的自由度定义为：$\widehat{df}(\lambda) = \#\{j : \widehat{\beta}_j(\lambda) \neq 0\}$

- GCV方法：可以极小化下面的GCV 准则选择调节参数λ，即

$$\widehat{\lambda}_{\text{gcv}} = \arg\min_{\lambda} \text{GCV}(\lambda) = \arg\min_{\lambda} \frac{\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2}{n(1 - \widehat{df}(\lambda)/n)^2}$$

- BIC方法： 可以通过极小化下面的BIC准则进行选择调节参数λ，即

$$\text{BIC}(\lambda) = \log \widehat{\sigma}_{\lambda}^2 + \widehat{df}(\lambda) \log(n)/n$$

- 极小化上面的BIC准则目标函数$\text{BIC}(\lambda)$，可以选择最优的调节参数λ

```
library(glmnet); library(latex2exp)
fit_lasso = glmnet(x, y, alpha = 1, nlambda = 20)
lam = fit_lasso$lambda
beta.hat = as.matrix(fit_lasso$beta)
## 绘制Lasso估计的路径图
path.plot(lam, beta.hat)  ## 函数path.plot()见教材
## 用函数cv.glmnet()选择最优的lambda
set.seed(2021)
cv.lasso = cv.glmnet(x, y, alpha = 1)
plot(cv.lasso)                    ## 绘制交叉验证误差图
> cv.lasso$lambda.min          > cv.lasso$lambda.1se
  [1] 0.005075651                [1] 0.5835766
```

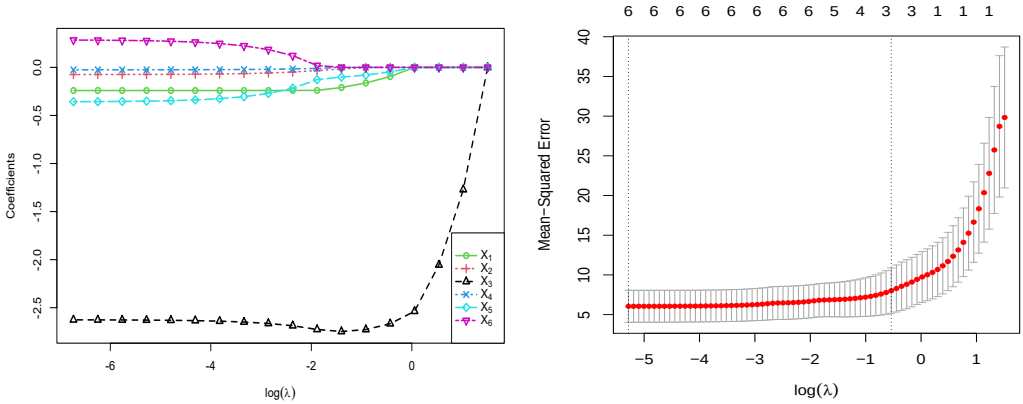


Figure: 31名中年男性健康数据的Lasso回归分析，其中，左图：Lasso估计随着λ变化的路径图；右图：Lasso回归的交叉验证误差图。

```
> coef(cv.lasso, s="lambda.min")
7 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) 105.01633937
X1           -0.24071641
X2           -0.07302177
X3           -2.62876371
X4           -0.02484750
X5           -0.35117527
X6            0.27768283
```

```
> coef(cv.lasso, s="lambda.1se")
                       1
(Intercept) 90.52944150
X1          -0.11267059
X2           .
X3          -2.68260426
X4           .
X5          -0.05522621
X6           .
```

SCAD方法的理论结果

- 针对多元线性模型，考虑下面的SCAD惩罚最小二乘目标函数：

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

其中$p_\lambda(\cdot)$是SCAD惩罚函数。

- 惩罚函数满足假设：
  1. $p_\lambda(\cdot)$是一个非负、非降函数，且$p_\lambda(0) = 0$；
  2. $p_\lambda(\cdot)$在$\boldsymbol{\beta}_0$的非零分量处存在二阶连续偏导数，其中假设$\boldsymbol{\beta}_0$是$\boldsymbol{\beta}$的真值。

SCAD方法的理论结果

- 令$\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})' = (\boldsymbol{\beta}'_{0,\mathrm{I}}, \boldsymbol{\beta}'_{0,\mathrm{II}})'$，其中$\boldsymbol{\beta}_{0,\mathrm{I}}$是真参数向量$\boldsymbol{\beta}_0$的前$s$个分量向量。
- 不失一般性，假设$\boldsymbol{\beta}_{0,\mathrm{II}} = \mathbf{0}$，并且$\boldsymbol{\beta}_{0,\mathrm{I}}$的所有分量都不等于0。
- 令

$$a_n = \max\{|p'_\lambda(|\beta_{0j}|)| : \beta_{0j} \neq 0\}$$

和

$$b_n = \max\{|p''_\lambda(|\beta_{0j}|)| : \beta_{0j} \neq 0\}.$$

- 极小化$Q(\boldsymbol{\beta})$，可得$\boldsymbol{\beta}$的一个SCAD估计，记为$\widehat{\boldsymbol{\beta}}$。

定理7.6.1

假设$(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$来自多元线性回归模型的独立同分布(i.i.d.)的观测样本，并令$\boldsymbol{\Pi} = E(\boldsymbol{X}\boldsymbol{X}')$是有限且正定矩阵，模型误差$\varepsilon$的均值为0，方差为$\sigma^2$。如果$a_n \to 0$和$b_n \to 0$，则依概率趋于1，存在$Q(\boldsymbol{\beta})$的一个局部最小值$\widehat{\boldsymbol{\beta}}$，使得

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_P(n^{-1/2} + a_n).$$

SCAD方法的理论结果

- 为了获得$\widehat{\boldsymbol{\beta}}$的$\sqrt{n}$-相合性，要求$a_n = O(1/\sqrt{n})$和$b_n \to 0$。
- 对其他惩罚变量选择方法：
  1. 对硬门限惩罚和SCAD惩罚函数，要求$\lambda \to 0$;
  2. 对L_1惩罚函数，要求$\lambda = O_P(n^{-1/2})$;
  3. 对L_q和对数变换的L_q惩罚函数$(0 < q < 1)$，要求$\lambda = O_P(n^{-1/2})$。

- 为了讨论下面的oracle性质，考虑线性回归模型：

$$\boldsymbol{Y} = \mathbf{X}_{\mathrm{I}}\boldsymbol{\beta}_{\mathrm{I}} + \mathbf{X}_{\mathrm{II}}\boldsymbol{\beta}_{\mathrm{II}} + \boldsymbol{\varepsilon},$$

其中

  - $E(\varepsilon|\mathbf{X}) = \mathbf{0}$
  - $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}_n$

- $\boldsymbol{\beta}$的一个理想估计是：

$$\widehat{\boldsymbol{\beta}}_{\mathrm{II}} = \mathbf{0}, \qquad \widehat{\boldsymbol{\beta}}_{\mathrm{I}} = (\mathbf{X}_{\mathrm{I}}'\mathbf{X}_{\mathrm{I}})^{-1}\mathbf{X}_{\mathrm{I}}'\boldsymbol{Y}.$$

- 上面估计正确识别了正确的模型，好像提前知道了正确的模型，这就是oracle 估计。

定理7.6.2：oracle性质

在定理7.6.1的条件下，假设

$$\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\sqrt{n}p'_\lambda(\theta)=+\infty,$$

则依概率趋于1，$\sqrt{n}$-相合局部最小估计$\widehat{\boldsymbol{\beta}}=(\widehat{\boldsymbol{\beta}}_{\mathrm{I}}',\widehat{\boldsymbol{\beta}}_{\mathrm{II}}')'$满足：

1. (稀疏性) $\widehat{\boldsymbol{\beta}}_{\mathrm{II}}=\mathbf{0}$;
2. (渐近正态性)

$$\sqrt{n}(\boldsymbol{\Pi}_{\mathrm{I}}+\boldsymbol{\Sigma})\left\{\widehat{\boldsymbol{\beta}}_{\mathrm{I}}-\boldsymbol{\beta}_{0,\mathrm{I}}+(\boldsymbol{\Pi}_{\mathrm{I}}+\boldsymbol{\Sigma})^{-1}\boldsymbol{b}\right\}\xrightarrow{d}N(\mathbf{0},\sigma^2\boldsymbol{\Pi}_{\mathrm{I}}),$$

其中$\boldsymbol{\Pi}_{\mathrm{I}}$是矩阵$\boldsymbol{\Pi}$的前$s$行和列构成的矩阵，且

$$\boldsymbol{\Sigma}=\mathrm{diag}\{p''_\lambda(|\beta_{01}|),\cdots,p''_\lambda(|\beta_{0s}|)\},\quad \boldsymbol{b}=(p'_\lambda(|\beta_{01}|)\mathrm{sgn}(\beta_{01}),\cdots,p'_\lambda(|\beta_{0s}|)\mathrm{sgn}(\beta_{0s}))'.$$

SCAD方法的理论结果

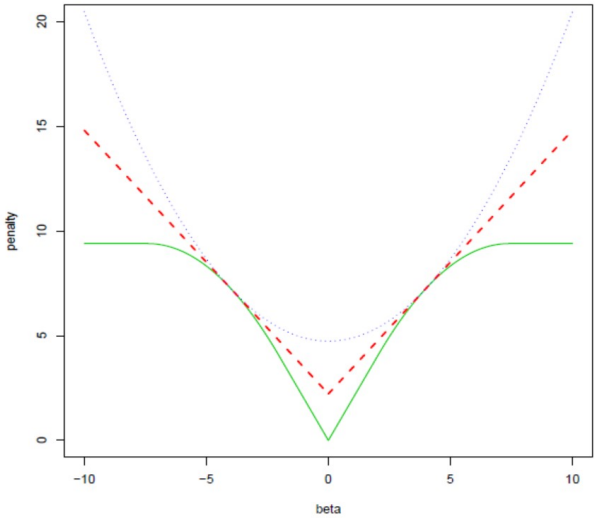
- 对硬门限和SCAD惩罚函数：
  - 当$n \to \infty$时，如果$\lambda \to 0$, $\sqrt{n}\lambda \to \infty$，则$a_n = 0$。
  - 由定理7.6.1可知，存在$\sqrt{n}$-相合的局部最小值
  - 由定理7.6.2可知，$\sqrt{n}$-相合的局部最小值满足：$\widehat{\boldsymbol{\beta}}_{\mathrm{II}} = \mathbf{0}$，且当$b_n = 0$时，$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{I}} - \boldsymbol{\beta}_{0,\mathrm{I}})$ 具有均值为$\mathbf{0}$，协方差矩阵为$\boldsymbol{\Pi}_{\mathrm{I}}^{-1}$的渐近正态分布
- 对Lasso惩罚函数：
  - 定理7.6.1的$\sqrt{n}$-相合性要求$\lambda = O(1/\sqrt{n})$
  - 定理7.6.2的oracle 性质要求$\sqrt{n}\lambda \to \infty$，但是要求$\lambda = O(1/\sqrt{n})$ 和$\sqrt{n}\lambda \to \infty$同时成立，显然是不可能的
  - 因此，Fan 和Li (2001) 认为Lasso估计不具有oracle性质

LQA算法

- 因为SCAD惩罚函数在原点处是奇异的，$Q(\boldsymbol{\beta})$是非光滑、非凸和高维的函数，这时Newton-Raphson算法不能直接被用于求解$\boldsymbol{\beta}$的最优解。
- 问题：如何处理非光滑和非凸的惩罚函数？
- Fan 和Li (2001)提出对惩罚函数$p_\lambda(\cdot)$进行局部二次逼近，提出了一个局部二次逼近(LQA) 的迭代算法。
- 对任给非零θ_0 的某个小邻域内，$p_\lambda(\cdot)$ 在θ_0的局部渐近表示为：

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta_0|) + \frac{1}{2}\frac{p'_\lambda(|\theta_0|)}{|\theta_0|}(\theta^2 - \theta_0^2).$$

LLA and LQA for SCAD penalty



步骤1： 初始估计，记为$\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \cdots, \beta_p^{(0)})'$。可用没有惩罚的最小二乘估计作为初始估计；

步骤2 (LQA)： 对$j = 1, \cdots, p$，对给定非零$\beta_j^{(0)}$的某个小邻域内，惩罚函数$p_\lambda(|\beta_j|)$在$\beta_j^{(0)}$的渐近表示为：

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}). \tag{6.1}$$

或者惩罚函数的导数在$\beta_j^{(0)}$处表示为：

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}(\beta_j) \approx \frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}\beta_j;$$

步骤3： 把局部二次逼近的惩罚函数(6.1)代入到惩罚最小二乘目标函数$Q(\boldsymbol{\beta})$中，应用调整的Newton-Raphson 算法进行求解，如果$|\beta_j| < \eta$，则删掉该变量；

步骤4： 在步骤2和步骤3之间进行迭代，直到收敛。

- LQA算法可直接提供一个估计量标准误差的直接估计量
- 对于线性回归模型，LQA算法变成了下面的迭代岭回归算法：

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{X}'\boldsymbol{Y},$$

其中

- $\boldsymbol{\beta}^{(k)}$表示第k步的估计值
- $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)}) = \text{diag}\left\{\dfrac{p'_\lambda(|\beta_1^{(k)}|)}{|\beta_1^{(k)}|}, \cdots, \dfrac{p'_\lambda(|\beta_p^{(k)}|)}{|\beta_p^{(k)}|}\right\}$

- LQA算法将删掉回归系数小的变量，对于第$k+1$步，如果$|\beta_j^{(k+1)}| < \eta$时，则删掉第j个变量
- 当算法收敛时，估计满足下面的惩罚最小二乘估计方程：

$$-\mathbf{x}_j'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + np_\lambda'(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) = \mathbf{0}$$

- 对非零回归系数的估计，满足：

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\boldsymbol{Y}.$$

- 这时，可计算$\widehat{\boldsymbol{\beta}}$的协方差矩阵为

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) \approx \{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\mathrm{Cov}(\boldsymbol{Y}|\mathbf{X})\mathbf{X}\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}.$$

- 对同方差情形，有

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) \approx \widehat{\sigma}^2\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\mathbf{X}\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}.$$

- 对异方差情形，有

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) \approx \{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\mathbf{E}\mathbf{X}\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1},$$

其中$\mathbf{E} = \mathrm{diag}\{\widehat{\varepsilon}_1^2, \cdots, \widehat{\varepsilon}_n^2\}$。

LQA算法的主要问题：

LQA算法的主要问题：

- LQA算法在实际应用中需要给定阈值η
- LQA算法在迭代过程中把回归系数$|\beta_j| < \eta$的变量删掉，在后面的迭代过程删掉的变量不再回到计算过程中

- 为了解决这个问题，Hunter 和Li (2005)在LQA算法的步骤2中，对惩罚函数提出了下面扰动版本的局部二次逼近，即

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}| + \tau_0}(\beta_j^2 - \beta_j^{(0)2}),$$

其中τ_0是一个非负的扰动参数。

- Hunter 和Li (2005) 把修正以后的算法称为MM算法。

Zou 和Li (2008)提出了下面的局部线性逼近(local linear approximation，简写为LLA) 算法。

步骤1： 给定初始估计$\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \cdots, \beta_p^{(0)})'$，可用没有惩罚的最小二乘估计作为初始估计；

步骤2： 在第k步，令$\boldsymbol{\beta}^{(k)} = (\beta_1^{(k)}, \cdots, \beta_p^{(k)})'$ 为第k 步的估计值。对给定非零$\beta_j^{(k)}$的某个小邻域内，惩罚函数$p_\lambda(|\beta_j|)$ 在$\beta_j^{(k)}$ 的局部线性表示为：

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(k)}|) + p'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|); \tag{6.2}$$

步骤3： 把惩罚函数的局部线性逼近(6.2) 代入到惩罚最小二乘目标函数$Q(\boldsymbol{\beta})$中，并去掉常数项，在LLA 的帮助下，极小化下面的目标函数，可得$k+1$ 步估计为：

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right\}.$$

步骤4： 在步骤2和步骤3之间进行迭代，直到收敛。

LLA 算法

- LLA算法成功避免选取LQA算法中的阈值η 和MM 算法中的τ_0
- 从步骤3也可以看出，采用求Lasso 解的算法可以得到回归系数的SCAD估计
- LLA算法把极小化非凸目标函数的问题转化成了一个极小化凸函数的问题
- 因此，LLA 算法能够找到一个理想的局部最小值，且具有oracle 性质

Zou 和Li (2008) 基于LARS算法，提出了下面的一步估计(one-step estimates，简写为ose) 算法：

步骤1： 对$i=1,\cdots,n;\ j=1,\cdots,p$，作变换：$x_{ij}^{*}=x_{ij}/p'(|\beta_j^{(0)}|)$.

步骤2： 应用LARS算法，求解下面的Lasso解：

$$\widehat{\boldsymbol{\beta}}^{*}=\arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\|\boldsymbol{Y}-\mathbf{X}^{*}\boldsymbol{\beta}\|_2^2+n\lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

步骤3： 对$j=1,\cdots,p$，最终一步估计表示为：

$$\widehat{\beta}_j^{ose}=\widehat{\beta}_j^{*}/p'(|\beta_j^{(0)}|).$$

一步估计(ose) 算法

- 从一步估计算法中可以看到，如果$\widehat{\beta}_j^* \neq 0$，则相应的预测变量被选入模型
- 问题：
  - ※ 对于一些惩罚函数，可能导数为0
  - ※ 调节参数λ不能从惩罚函数中分离出来，如SCAD惩罚函数
- 解决问题的方法：令

$$U = \{j : p'_\lambda(|\beta_j^{(0)}|) = 0\}, \quad V = \{j : p'_\lambda(|\beta_j^{(0)}|) > 0\}.$$

- 根据子集U和V，可以把设计矩阵$\mathbf{X}$和回归系数向量$\boldsymbol{\beta}$进行如下划分：

$$\mathbf{X} = [\mathbf{X}_U, \mathbf{X}_V], \quad \boldsymbol{\beta} = (\boldsymbol{\beta}'_U, \boldsymbol{\beta}'_V)'.$$

修正的一步估计算法

步骤1： 对子集U和V，对数据作变换：

1. 对$j \in V$，令$x_{ij}^* = x_{ij} \dfrac{\lambda}{p'_\lambda(|\beta_j^{(0)}|)}$；
2. 令$\mathbf{H}_U$是空间$\{\boldsymbol{X}_j, j \in U\}$的一个投影矩阵，计算

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \mathbf{H}_U \boldsymbol{Y}, \quad \mathbf{X}_V^{**} = \mathbf{X}_V^* - \mathbf{H}_U \mathbf{X}_V^*;$$

步骤2： 应用LARS算法，求解下面的Lasso解：

$$\widehat{\boldsymbol{\beta}}_V^* = \arg\min_{\boldsymbol{\beta}_V} \left\{ \frac{1}{2} \|\boldsymbol{Y}^* - \mathbf{X}_V^{**} \boldsymbol{\beta}_V\|_2^2 + n\lambda \|\boldsymbol{\beta}_V\|_1 \right\};$$

步骤3： 计算$\widehat{\boldsymbol{\beta}}_U^* = (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{X}_U' (\boldsymbol{Y} - \mathbf{X}_V^* \widehat{\boldsymbol{\beta}}_V^*)$。这时，最终一步估计为：

$$\widehat{\boldsymbol{\beta}}_U^{ose} = \widehat{\boldsymbol{\beta}}_U^*, \qquad \widehat{\beta}_j^{ose} = \widehat{\beta}_j^* \frac{\lambda}{p'_\lambda(|\beta_j^{(0)}|)}, \quad j \in V.$$

调节参数λ的选择

- 令$df_N(\lambda)$表示真实模型的自由度，即正确模型中非零回归系数的个数
- 对SCAD估计$\widehat{\boldsymbol{\beta}}$，响应变量$\boldsymbol{Y}$预测的拟合为：

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\boldsymbol{Y}.$$

- 自由度定义为：

$$\widehat{df}(\lambda) = \text{tr}\left\{\mathbf{X}\{\mathbf{X}'\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}'\right\}.$$

- 在一定条件下，Zhang等(2010)证明：

$$\Pr\{\widehat{df}(\lambda) = df_N(\lambda)\} = 1$$

- GCV方法：极小化下面的GCV 准则选择调节参数λ，即

$$\widehat{\lambda}_{\text{gcv}} = \arg\min_{\lambda} \text{GCV}(\lambda) = \arg\min_{\lambda} \frac{\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2}{n(1 - \widehat{df}(\lambda)/n)^2}$$

- 问题： 基于GCV准则选择的调节参数$\widehat{\lambda}_{\text{gcv}}$，所得到的SCAD估计是否具有oracle 性质？
- 令$\widehat{\sigma}_{\lambda}^2 = n^{-1}\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2$，则

$$\text{GCV}(\lambda) = \frac{\widehat{\sigma}_{\lambda}^2}{(1 - \widehat{df}(\lambda)/n)^2}.$$

- 这时，有

$$\begin{aligned}\log\{\mathrm{GCV}(\lambda)\} &= \log\widehat{\sigma}_{\lambda}^{2} - 2\log[1-\widehat{df}(\lambda)/n] \\ &\approx \log\widehat{\sigma}_{\lambda}^{2} + 2\widehat{df}(\lambda)/n =: \mathrm{AIC}(\lambda).\end{aligned}$$

- 可见，$\log\{\mathrm{GCV}(\lambda)\}$是类似于传统的AIC准则

定理7.6.3

在一定正则条件下，依概率趋于1，由SCAD-GCV选择的模型包含了所有显著性协变量(或预测变量)，但是也依非零概率包含了一些不显著的协变量(或预测变量)。

- SCAD-GCV选择的模型是过拟合的，尽管包含了正确的模型，但是也包含了很多噪声变量。
- BIC准则： Wang等(2007) 定义下面的BIC 准则：

$$\mathrm{BIC}(\lambda) = \log \widehat{\sigma}_{\lambda}^2 + \widehat{df}(\lambda) \log(n)/n.$$

- $\widehat{\lambda}_{\mathrm{bic}} = \arg\min_{\lambda} \mathrm{BIC}(\lambda)$

定理7.6.4

在一定正则条件下，依概率趋于1，SCAD-BIC估计具有oracle性质。

Lasso和SCAD的比较

- **统计性质：**
  - Lasso $\Rightarrow$ 有偏估计，没有oracle 性质
  - SCAD解决了Lasso对大的系数有偏估计的问题，并具有oracle性质
- **优化方面：**
  - Lasso $\Rightarrow$ 惩罚最小二乘目标函数(PLS)是凸函数

    $\Rightarrow$ 存在唯一的全局最小值

    能够通过求解线性约束的二次规划得到最优解

    R package: 程序包lars和glmnet或算法能够找到解的路径
  - SCAD $\Rightarrow$ 惩罚最小二乘目标函数(PLS)非凸

    $\Rightarrow$ 存在多个局部最小值解

    通过LQA算法进行求解，R package: ncvreg 和SIS

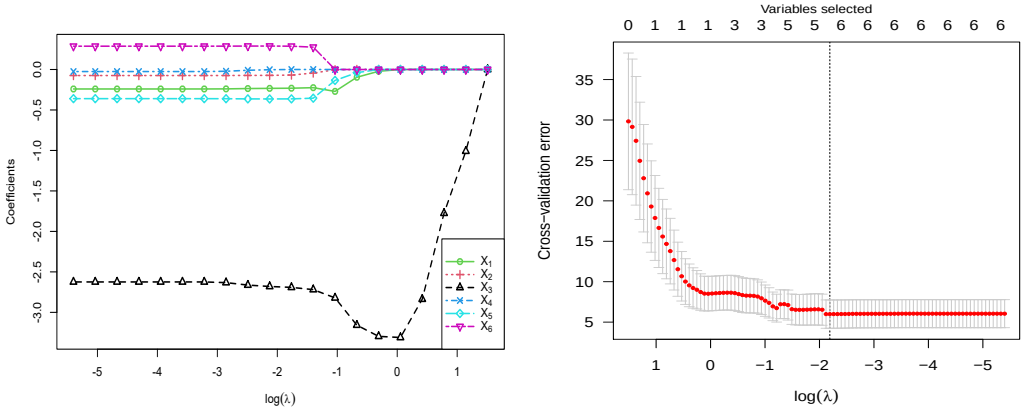


Figure: 31名中年男性健康数据的SCAD回归分析，其中，左图：SCAD估计随着λ变化的路径图；右图：SCAD回归的交叉验证误差图。

- SCAD估计最后被压缩成0的系数变量依次为X_3, X_1, X_5, X_6，而变量X_2和X_4的系数随着λ变大，几乎很快同时被压缩成0
- 使$\mathrm{CV}(\widehat{\lambda})$最小化的$\lambda$为$\widehat{\lambda} \approx 0.1119$
- 对应的回归系数分别为：$\widehat{\beta}_0 \approx 104.13, \widehat{\beta}_1 \approx -0.23, \widehat{\beta}_2 \approx -0.07, \widehat{\beta}_3 \approx -2.68, \widehat{\beta}_4 \approx -0.00, \widehat{\beta}_5 \approx -0.36$和$\widehat{\beta}_6 \approx 0.29$
- 如果采用"一个标准差"准则选取稍微大的调节参数，则同样会产生稀疏解。
- 可见，当取$\widetilde{\lambda} = 0.3929893$时，可以产生稀疏解，使得$X_2, X_4$和$X_6$的系数为0。

自适应Lasso

- 通过极小化L_1惩罚最小二乘目标函数，得到了Lasso 估计，即

$$\widehat{\boldsymbol{\beta}}^{L} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

- 令$\mathcal{A} = \{j : \beta_{0j} \neq 0\}$表示活动模型，$\widehat{\mathcal{A}}_n = \{j : \widehat{\beta}_j^L \neq 0\}$ 表示选择模型
- 变量选择相合性，即：

$$\lim_{n\to\infty} \Pr(\widehat{\mathcal{A}}_n = \mathcal{A}) = 1.$$

自适应Lasso

- 为了证明变量选择相合性，假设$n^{-1}\mathbf{X}'\mathbf{X} \to \mathbf{C}$，其中$\mathbf{C}$是正定矩阵。
- 不失一般性，假设$\mathcal{A} = \{1, \cdots, s\}$。
- 令

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}, \qquad \text{其中}\mathbf{C}_{11}\text{是一个}s \times s\text{矩阵.}$$

Zou (2006)证明Lasso估计有下面的结论：

1. 如果$\sqrt{n}\lambda \to \lambda_0 \geq 0$时，则$\sup_n \Pr(\widehat{\mathcal{A}}_n = \mathcal{A}) \leq c < 1$，其中$c$是依赖于正确模型的常数；
2. 如果$\lambda \to 0$和$\sqrt{n}\lambda \to \infty$时，则

$$\lambda^{-1}(\widehat{\boldsymbol{\beta}}^L - \boldsymbol{\beta}_0) \xrightarrow{P} \operatorname{argmin}(V),$$

其中

$$V(\boldsymbol{u}) = \boldsymbol{u}'\mathbf{C}\boldsymbol{u} + \sum_{j=1}^{p} \{u_j \operatorname{sgn}(\beta_{0j}) I(\beta_{0j} \neq 0) + |u_j| I(\beta_{0j} = 0)\}.$$

自适应Lasso

Zou (2006) 证明：

- Lasso估计$\widehat{\beta}^L$的收敛速度比$n^{-1/2}$慢
- 最优收敛速度要求调节参数满足：$\lambda = O(1/\sqrt{n})$
- Lasso方法导致了变量选择的不相合性，从而证明了Fan 和Li (2001) 的猜想，即Lasso 估计不具有oracle 性质

- Zou (2006) 进一步证明，假设 $\lim_{n\to\infty} \Pr(\widehat{\mathcal{A}}_n = \mathcal{A}) = 1$，则存在一些符号向量 $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_s)'$，其中 $\nu_j = 1$ 或 -1，使得

$$|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\boldsymbol{\nu}| \leq 1.$$

- 如果上面的条件不满足，则Lasso变量选择是不相合的。
- Zhao 和Yu (2006)把上面条件称为不可表示条件(irrepresentable condition)。

自适应Lasso

- 假设$s = 2m + 1 \geq 3$ 和$p = s + 1$，即存在1个不显著预测变量。
- 令$\mathbf{C}_{11} = (1 - \rho_1)\mathbf{I}_s + \rho_1\mathbf{J}_1$，其中$\mathbf{J}_1$ 是所有元素为1 的矩阵。
- $\mathbf{C}_{12} = \rho_2\mathbf{1}$, $\mathbf{C}_{22} = 1$，其中$\mathbf{1}$是所有元素为1 的s 维列向量。
- 如果$-\frac{1}{s-1} < \rho_1 < -\frac{1}{s}$，$1+(s-1)\rho_1 < |\rho_2| < \sqrt{\frac{1+(s-1)\rho_1}{s}}$时，则不可表示条件将不再满足。
- 如，令$s = 3$，则$-1/2 < \rho_1 < -1/3$ 和$1+2\rho_1 < |\rho_2| < \sqrt{(1+2\rho_1)/3}$。

自适应Lasso

- 令$\widehat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的一个$\sqrt{n}$-相合估计，如最小二乘估计。
- 定义权向量：$\widehat{\boldsymbol{w}} = 1/|\widehat{\boldsymbol{\beta}}|^{\gamma},\ \gamma > 0$。
- 自适应Lasso 估计$\widehat{\boldsymbol{\beta}}^{alasso}$定义为：

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

- 自适应Lasso 的基本思想：对于最小二乘估计大的回归系数，不进行惩罚，而对于接近于0 的回归系数给尽量大的惩罚，并压缩到0。

自适应Lasso

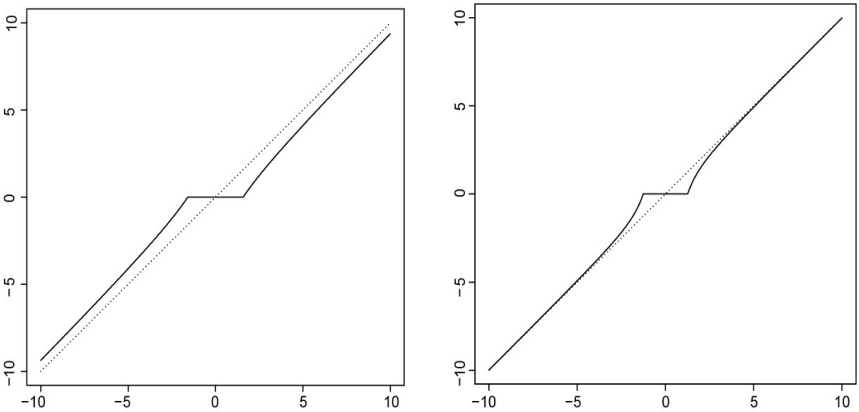


Figure: 自适应Lasso门限函数图，左图取$\gamma = 0.5$和$\lambda = 2$，右图取$\gamma = 2$和$\lambda = 2$。

自适应Lasso

- 令$\widehat{\mathcal{A}}^{alasso} = \{j : \widehat{\beta}_j^{alasso} \neq 0\}$表示由自适应Lasso选择的模型。Zou (2006)证明了下面的定理。

定理7.7.1

假设$\sqrt{n}\lambda \to 0$和$\lambda n^{(\gamma+1)/2} \to \infty$，则自适应Lasso估计满足：

1. 变量选择的相合性：$\lim_{n\to\infty} \Pr(\widehat{\mathcal{A}}^{alasso} = \mathcal{A}) = 1$；
2. 渐近正态性：$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{alasso} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}_{11}^{-1})$。

自适应Lasso

- 由定理7.7.1知，自适应Lasso估计不仅具有变量选择的相合性，而且具有oracle性质。
- 在应用中，权重中的初始估计$\widehat{\boldsymbol{\beta}}^{initial}$不需要满足$\sqrt{n}$-相合性，能够被弱化。
- 假设存在一个序列a_n使得$a_n \to \infty$和$a_n(\widehat{\boldsymbol{\beta}}^{initial} - \boldsymbol{\beta}_0) = O_P(1)$，如果$\lambda = o(n^{-1/2})$和$\sqrt{n}a_n^{\gamma}\lambda \to \infty$时，则定理7.7.1 中的oracle 性质成立。

nn-garotte(nonnegative garotte)方法

- Breiman (1995)提出nn-garotte(nonnegative garotte) 方法，即极小化下面的目标函数去找一个非负的刻度因子$\{c_j\}$：

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\widehat{\beta}_j^{ols}c_j\right)^2+\lambda\sum_{j=1}^{p}c_j$$

$$\text{s.t.} \qquad c_j\geq 0 \quad \forall j$$

- 则Garotte估计记为：$\widehat{\beta}_j^{garotte}=\widehat{c}_j\widehat{\beta}_j^{ols}$。

- 等价地，能够得到nn-garotte估计如下：

$$\widehat{\boldsymbol{\beta}}^{nn-garotte} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{ols}|} \right\},$$

$$\text{s.t.} \qquad \beta_j \widehat{\beta}_j^{ols} \geq 0 \quad \forall j.$$

- 在自适应Lasso中，令$\gamma = 1$，并选择自适应权重为$\widehat{w}_j = 1/|\widehat{\beta}_j^{ols}|$。
- 这时，自适应Lasso 解为：

$$\widehat{\boldsymbol{\beta}}^{alasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{ols}|} \right\}.$$

自适应Lasso

- 明显可以看出，nn-garotte估计能够看出是具有符号约束的自适应Lasso 估计($\gamma = 1$)。
- 如果调节参数λ满足：$\sqrt{n}\lambda \to 0$和$n\lambda \to \infty$，则nn-garotte估计方法是变量选择相合的。
- 问题：在实际应用中，如何求解自适应Lasso估计呢？

自适应Lasso

- 明显可以看出，nn-garotte估计能够看出是具有符号约束的自适应Lasso 估计($\gamma = 1$)。
- 如果调节参数λ满足：$\sqrt{n}\lambda \to 0$和$n\lambda \to \infty$，则nn-garotte估计方法是变量选择相合的。
- 问题：在实际应用中，如何求解自适应Lasso估计呢？
- 可以借助存在的Lasso程序包进行计算，例如lars，glmnet 和gcdnet。

自适应Lasso

- 下面介绍利用LARS算法求解自适应Lasso估计。
- 首先对数据作变换：$\mathbf{x}_j^* = \mathbf{x}_j / \widehat{w}_j$，其中$\mathbf{x}_j$是设计矩阵$\mathbf{X}$的第$j$ 列的n 个样本，且$j = 1, \cdots, p$。
- 然后应用LARS 算法，求解得

$$\widehat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}^* \boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- 这时可得回归系数$\boldsymbol{\beta}$的最终自适应Lasso估计为：

$$\widehat{\beta}_j = \widehat{\beta}_j^* / \widehat{w}_j, \ j = 1, \cdots, p.$$

自适应Lasso

- 问题：如何选择最优的调节参数λ和参数γ?
- 可以通过在二维空间中利用CV方法获得最优的λ和γ。
- 程序包msgps中的函数msgps()进行自适应Lasso分析，调用格式为：

```
msgps(X, y, penalty = "enet", alpha=0, gamma=1,
      lambda=0.001, tau2, STEP=20000, STEP.max=200000,
      DFtype="MODIFIED",  p.max=300,
      intercept=TRUE, stand.coef=FALSE)
```

其中X为协变量数据矩阵，y为响应变量数据；penalty="enet"时，表示elastic net方法，取"genet"表示推广的elastic net方法，取"alasso"时，表示自适应Lasso方法；alpha对应的enet 和genet方法的参数；gamma对应的是alasso方法的参数，默认为1；其余参数见在线帮助。

自适应Lasso

```
library(msgps)
alasso_fit = msgps(x, y, penalty = "alasso", gamma = 1, lambda=0)
par(mfrow=c(1,2))
plot(alasso_fit, criterion = "gcv", xvar = "t", main = "GCV")
plot(alasso_fit, criterion = "bic", xvar = "t", main = "BIC")
#### 用函数summary()汇总结果，并输出结果:
summary(alasso_fit)
Call:msgps(X = x, y = y, penalty = "alasso", gamma = 1, lambda = 0)
Penalty: "alasso"
gamma: 1
lambda: 0
df:
      tuning     df      ##只列出了2行结果
 [1,] 0.0000 0.0000
 [2,] 0.1686 0.1569
tuning.max: 4.567
```

自适应Lasso

```
ms.coef:
                    Cp     AICC       GCV      BIC
(Intercept) 101.59412 100.7412 101.68937 100.9867
X1           -0.21351  -0.2017  -0.21429  -0.1977
X2           -0.02203   0.0000  -0.02421   0.0000
X3           -2.77115  -2.8273  -2.76767  -2.8490
X4            0.00000   0.0000   0.00000   0.0000
X5           -0.31626  -0.2737  -0.31832  -0.2491
X6            0.23411   0.1879   0.23655   0.1626
ms.tuning:
        Cp  AICC   GCV   BIC
[1,] 2.843 2.182 2.901 2.105
ms.df:
       Cp  AICC   GCV   BIC
[1,] 4.11 3.525 4.154 3.405
```

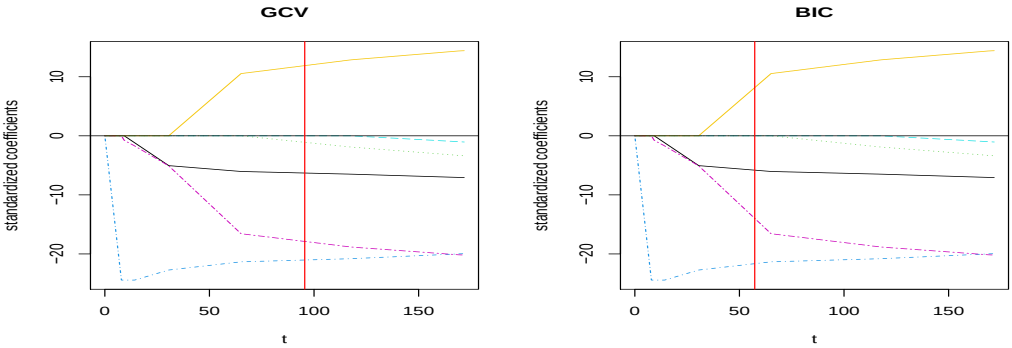


Figure: 31名中年男性健康数据的自适应Lasso估计的路径图，其中，左图：垂直虚线是用GCV 准则选取的调节参数；右图：垂直虚线是用BIC准则选取的调节参数。

- 通过一个模拟例子对岭回归、Lasso、SCAD和自适应Lasso(记为ALasso)方法进行比较。
- 例：考虑下面的多元线性回归模型：

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

其中

- $\boldsymbol{\beta} = (2, -1.5, 0.5, -0.5, 0.3, 0, \cdots, 0)'$ 为p维回归系数向量
- $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$从$p$元正态分布$N_p(\mathbf{0}, \boldsymbol{\Sigma})$中随机产生随机数，这里，$\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \le i,j \le p}$，且$\sigma_{ij} = \rho^{|i-j|}$
- 模型误差$\varepsilon_i \sim N(0,1)$，且独立于协变量向量$\boldsymbol{x}_i$
- 响应变量y_i可以通过上面多元线性回归模型产生

模拟数据分析

- 前5个协变量对响应变量有显著性影响，而剩余的$p-5$个协变量对响应变量不显著
- 取样本量$n=200$，维数$p=20,\ 50,\ 100$三种情况
- $\rho=0.3,\ 0.6$两种情况
- 为了对岭回归、Lasso、SCAD和自适应Lasso四种方法进行比较，重复500次试验
- 采用10折CV方法选取调节参数λ

四个指标进行评价：

1. 非零回归系数被正确估成非零的平均个数，用“C" 表示；
2. 零回归系数被错误估成非零的平均个数，用“IC"表示；
3. 基于500次重复试验的平均估计误差，用“EE"表示，其中估计误差用$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$来计算；
4. 基于500重复试验的平均预测误差，用“PE"表示，其中预测误差用$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'E(\boldsymbol{X}\boldsymbol{X}')(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ 来计算。

Table: 模拟结果，其中$n = 200$

		$\rho = 0.3$				$\rho = 0.6$			
p	指标	Ridge	Lasso	SCAD	ALasso	Ridge	Lasso	SCAD	ALasso
$p = 20$	EE	0.4289	0.3011	0.2205	0.2525	0.6025	0.4056	0.3075	0.3272
	PE	0.1293	0.0713	0.0456	0.0559	0.1375	0.0765	0.0554	0.0594
	C	5.0000	4.9980	4.9900	4.9940	5.0000	4.9560	4.8340	4.9620
	IC	15.0000	7.6840	3.0460	5.0040	15.0000	8.4320	3.1740	4.8080
$p = 50$	EE	0.6458	0.3893	0.2362	0.3700	0.8502	0.5446	0.3545	0.4589
	PE	0.2536	0.1075	0.0521	0.1098	0.2517	0.1228	0.0709	0.1116
	C	5.0000	4.9760	4.9680	4.9920	5.0000	4.8360	4.6680	4.9580
	IC	45.0000	13.9460	5.4940	15.1780	45.0000	16.2520	6.1720	14.8520
$p = 100$	EE	0.9890	0.4523	0.2595	0.5694	1.2330	0.6702	0.4144	0.7007
	PE	0.4503	0.1365	0.0613	0.2180	0.4282	0.1694	0.0912	0.2215
	C	5.0000	4.9720	4.9560	4.9880	5.0000	4.6040	4.4080	4.9300
	IC	95.0000	20.3440	9.0160	35.8340	95.0000	23.5360	9.6640	36.4820

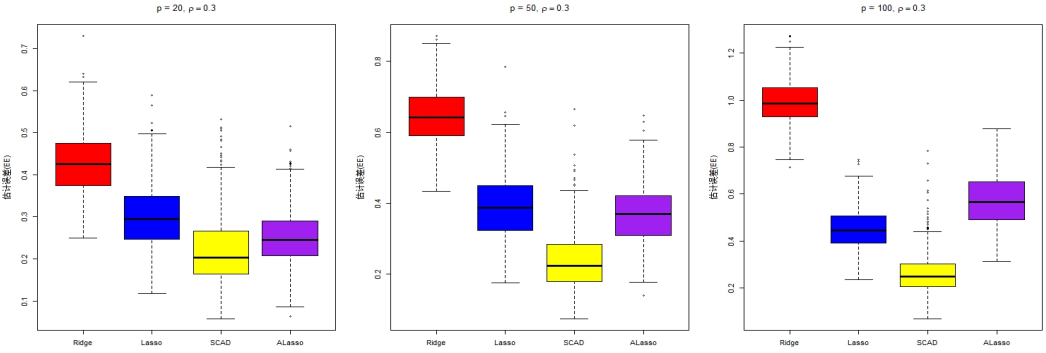


Figure: 当$n = 200$和$\rho = 0.3$时，基于500次重复试验，模拟数据的估计误差(EE)的箱线图，从左到右分别为$p = 20,\ 50,\ 100$的估计误差箱线图。

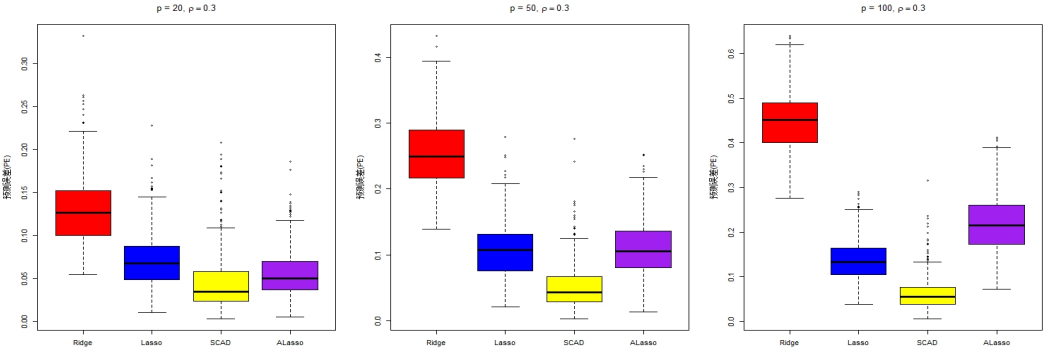
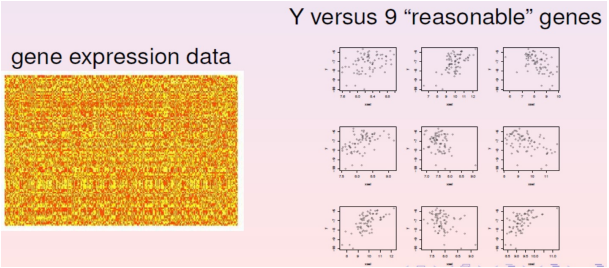
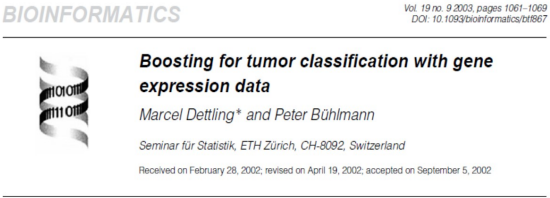


Figure: 当$n = 200$和$\rho = 0.3$时，基于500次重复试验，模拟数据预测误差(PE)的箱线图，从左到右分别为$p = 20,\ 50,\ 100$的预测误差箱线图。

- Riboflavin production with Bacillus Subtilis (枯草芽孢杆菌生产核黄素)(in collaboration with DSM (Switzerland))
  - 目的：利用基因工程提高枯草芽孢杆菌核黄素的产生率
  - 响应变量$Y \in \mathbb{R}$：riboflavin (log-) production rate
  - 协变量$\boldsymbol{X} \in \mathbb{R}^p$：expressions from $p = 4088$ genes
  - 样本量大小：$n = 115$
  - 典型的$p \gg n$



- 基因表达肿瘤(tumor)样本的microarray数据
  - 目的：分类
  - 响应变量：$Y \in \{0, 1, \cdots, J-1\}$
  - 协变量$\boldsymbol{X} \in \mathbb{R}^p$：expressions from $p = 7130$ genes
  - 样本量大小：$n = 49$
  - 典型的$p \gg n$



BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1061–1069
DOI: 10.1093/bioinformatics/btf867

Boosting for tumor classification with gene expression data

Marcel Dettling* and Peter Bühlmann

Seminar für Statistik, ETH Zürich, CH-8092, Switzerland

Received on February 28, 2002; revised on April 19, 2002; accepted on September 5, 2002

高维统计推断方法：$p \geq n$

- 监督学习： 回归模型、分类模型;
- 无监督学习： 聚类、图模型、多重假设检验;
- 非参数统计和机器学习： 高度复杂性、正则化;
- 统计和机器学习的交互: 统计学习和数据挖掘。

机会和挑战

挑战:

- 统计理论、统计方法和应用
- 统计精度、模型的解释能力、计算复杂度和稳定性
- 传统统计方法和理论不再适用，遇到了很大的挑战

机会:

- 关键想法: 稀疏性
- 产生稀疏模型的一般框架:
  Goodness of fit (RSS, -log-likelihood, or empirical risk)+ penalty on model complexity
- Lasso, SCAD, Elastic Net, Adaptive Lasso, MCP $\cdots$

- 考虑下面的高维多元线性回归模型：

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},$$

其中

  - $\boldsymbol{Y} = (y_1, \cdots, y_n)'$为$n \times 1$的响应变量向量
  - $\mathbf{X}$ 为$n \times p$的设计矩阵
  - $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})'$ 为$p \times 1$的正确回归系数向量
  - 假设模型误差$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$；对于模型误差，也可以放松为轻尾分布

- 假设协变量的维数p远远大于样本量n，即$p \gg n$

- 对于给定的调节参数λ，Lasso估计定义为：

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

- 由优化理论，Lasso解存在的充分必要条件是满足下面的Karush-Kuhn-Tucker (KKT) 条件：

$$\mathbf{X}'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})/n = \lambda \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) =: \lambda \boldsymbol{\gamma},$$

其中$\boldsymbol{\gamma} = \mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = (\mathrm{sgn}(\widehat{\beta}_1), \cdots, \mathrm{sgn}(\widehat{\beta}_p))'$。

- 对$j=1,\cdots,p$，有

$$\gamma_j=\operatorname{sgn}(\widehat{\beta}_j)\in\begin{cases}\widehat{\beta}_j, & \text{如果}\widehat{\beta}_j\neq 0,\\ [-1,1], & \text{如果}\widehat{\beta}_j=0.\end{cases}$$

定理7.8.1：基础不等式

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1\leq\boldsymbol{\varepsilon}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/n+\lambda\|\boldsymbol{\beta}_0\|_1.$$

证明：假设$\boldsymbol{\beta}_0$是真实的回归系数向量，并由$\widehat{\boldsymbol{\beta}}$是Lasso 解，则有

$$\|\boldsymbol{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}_0\|_2^2/(2n)+\lambda\|\boldsymbol{\beta}_0\|_1,$$

即，进一步有

$$\begin{aligned}
&\|\boldsymbol{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1\\
=&\|\mathbf{X}\boldsymbol{\beta}_0+\boldsymbol{\varepsilon}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1\\
=&\|\mathbf{X}\boldsymbol{\beta}_0-\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2/(2n)+\|\boldsymbol{\varepsilon}\|_2^2/(2n)-2\boldsymbol{\varepsilon}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1\\
\leq&\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}_0\|_2^2/(2n)+\lambda\|\boldsymbol{\beta}_0\|_1\\
=&\|\boldsymbol{\varepsilon}\|_2^2/(2n)+\lambda\|\boldsymbol{\beta}_0\|_1.
\end{aligned}$$

对上式进行化简，即完成了基础不等式的证明。

- 为了证明Lasso估计的相合性，下面需要给出$\boldsymbol{\varepsilon}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/n$的上界。
- 由

$$\left|\boldsymbol{\varepsilon}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/n\right| \leq \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ij}\right| \cdot \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1.$$

- 定义下面的事件：

$$\mathcal{T}=\mathcal{T}(\lambda_0)=\left\{\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ij}\right| \leq \lambda_0\right\}.$$

引理7.8.1

为了简单，对所有的$j = 1, \cdots, p$，假设$\widehat{\sigma}_{jj} = \widehat{\Sigma}_{j,j} = 1$，其中$\widehat{\Sigma} = n^{-1}\mathbf{X}'\mathbf{X}$。对所有的$t > 0$和$z = \sigma\sqrt{\dfrac{t^2 + 2\log p}{n}}$，则

$$\Pr\left(\left\|\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right\|_\infty \leq z\right) \geq 1 - 2\exp\{-t^2/2\},$$

其中

$$\left\|\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right\|_\infty = \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ij}\right|.$$

证明：因为$\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$，且对所有的$t > 0$，则有

$$
\begin{aligned}
&1 - \Pr\left(\left\|\frac{1}{n}\mathbf{X}'\varepsilon\right\|_\infty \le z\right) = \Pr\left(\left\|\frac{1}{n}\mathbf{X}'\varepsilon\right\|_\infty > z\right) \\
=&\Pr\left(\max_{1\le j\le p} |\varepsilon'\mathbf{x}_j|/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p}\right) \\
\le& p\Pr\left(|\varepsilon'\mathbf{x}_j|/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p}\right) \\
=&2p\Pr\left(\varepsilon'\mathbf{x}_j/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p}\right) \\
\le&2p\exp\left\{-\frac{t^2 + 2\log p}{2}\right\} = 2\exp\{-t^2/2\},
\end{aligned}
$$

其中$\mathbf{x}_j$表示$\mathbf{X}$的第j列。

则有

$$\Pr\left(\max_{1\le j\le p}|\varepsilon'\mathbf{x}_j|/n\le z\right)\ge 1-2\exp\{-t^2/2\}.$$

- 由引理7.8.1，令$\lambda_0=O(\sqrt{\log p/n})$，在事件$\mathcal{T}$上，则有

$$\left|\varepsilon'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/n\right|\le\lambda_0\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1.$$

- 由上式和基础不等式，在事件$\mathcal{T}$上，有

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}\|_1\le\lambda_0\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1+\lambda\|\boldsymbol{\beta}_0\|_1.$$

- 令$S = \{j : \beta_{0j} \neq 0\}$表示活动集，且假设$s = \|S\|_0 = \#\{j : \beta_{0j} \neq 0\}$ 表示非零回归系数的个数。
- 由$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 \geq \|\boldsymbol{\beta}_{0,S}\|_1 - \|\widehat{\boldsymbol{\beta}}_S\|_1$，有

$$\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \geq \|\boldsymbol{\beta}_{0,S}\|_1 - \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1.$$

- 由$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$ 和$\|\boldsymbol{\beta}_0\|_1 = \|\boldsymbol{\beta}_{0,S}\|_1$，则有

$$\begin{aligned}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(2n) &\leq \lambda_0\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 + \lambda\|\boldsymbol{\beta}_0\|_1 - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \\ &\leq (\lambda_0 + \lambda)\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 + (\lambda_0 - \lambda)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1.\end{aligned}$$

Lasso的理论: 预测和估计

- 如果$\lambda_0 \le \lambda/2$，则有

$$\begin{aligned}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(2n) &\le (\lambda_0 + \lambda)\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 + (\lambda_0 - \lambda)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \\ &\le (3\lambda/2)\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 - (\lambda/2)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1,\end{aligned}$$

- 即

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(2n) + (\lambda/2)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \le (3\lambda/2)\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1.$$

- 由$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \ge 0$，在事件$\mathcal{T}$上，如果$\lambda_0 \le \lambda/2$，依概率趋于1，有

$$\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \le 3\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1,$$

- 即

$$\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)_{S^c}\|_1=\|\widehat{\boldsymbol{\beta}}_{S^c}-\boldsymbol{\beta}_{0,S^c}\|_1\leq 3\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1=3\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)_S\|_1.$$

- 可见，误差向量$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$属于一个指定的区域，这时引入相容性条件(compatibility condition)。

定义7.1：相容性条件(compatibility condition)

存在活动集S，和常数$\phi_0 > 0$，相容性条件满足，对所有的$\boldsymbol{\beta} \in \mathcal{E}(S,3) =: \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S^c}\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1\}$，成立

$$\|\boldsymbol{\beta}_S\|_1^2 \leq s(\boldsymbol{\beta}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta})/\phi_0^2,$$

其中

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{X}'\mathbf{X}, \quad \phi^2(\widehat{\boldsymbol{\Sigma}}, S) =: \min_{\boldsymbol{\beta} \in \mathcal{E}(S,3)} \frac{s\boldsymbol{\beta}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}}{\|\boldsymbol{\beta}_S\|_1^2}.$$

- 相容性条件要比不可表示条件要弱。在相容性条件下，只能证明下面的定理7.8.2和oracle不等式。
- 要想证明Lasso估计的L_2相合性，还需要对设计矩阵$\mathbf{X}$施加更强的限制特征值条件。
- 进一步，有

$$
\begin{aligned}
&\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+(\lambda/2)\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1\\
=&\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+(\lambda/2)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1+(\lambda/2)\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1\\
\le&(3\lambda/2)\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1+(\lambda/2)\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1\\
=&2\lambda\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1.
\end{aligned}
$$

- 容易看到，在事件$\mathcal{T}$上，如果$\lambda_0 \leq \lambda/2$时，误差向量$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$满足：$\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_{S^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_S\|_1$。
- 由相容性条件和不等式$2ab \leq a^2/4 + 4b^2$，有

$$\begin{aligned}
&\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(2n) + (\lambda/2)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \\
\leq &2\lambda\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}\|_1 \leq 2\lambda\sqrt{s[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)]/\phi_0^2} \\
= &2\lambda\sqrt{s}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2/(\sqrt{n}\phi_0) \leq \|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(4n) + 4\lambda^2 s/\phi_0^2.
\end{aligned}$$

定理7.8.2

对活动集S，假设相容性条件成立，在事件$\mathcal{T}$上，如果$\lambda_0 \leq \lambda/2$，则依概率趋于1，有

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/(2n) + \lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 8\lambda^2 s/\phi_0^2.$$

- Oracle不等式： 假设$\lambda \asymp \sqrt{\log(p)/n}$，有

$$n^{-1}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq \frac{s}{\phi_0^2} O_P(\log(p)/n),$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \frac{s}{\phi_0^2} O_P(\sqrt{\log(p)/n}).$$

- 为了获得L_2相合性，下面引入一个比相容性条件更强的一个条件：限制特征值条件(restricted eigenvalue conditions，简写为REC)。

定义7.2：限制特征值条件(REC)

令$\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{X}'\mathbf{X}$是$p \times p$的样本协方差矩阵，如果

$$\boldsymbol{\beta}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} = n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|_2^2 \geq \phi_0^2\|\boldsymbol{\beta}\|_2^2, \quad \forall \boldsymbol{\beta} \in \mathcal{E}(S,3),$$

则称$\widehat{\boldsymbol{\Sigma}}$活动集$S$上满足限制特征值条件(REC)。

- 利用下面不等式:

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+(\lambda/2)\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \le (3\lambda/2)\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1.$$

- 对上面不等式，两边乘2，并加$\lambda\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|$，则有

$$n^{-1}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2+\lambda\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le 4\lambda\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1.$$

- 由Cauchy-Schwarz不等式，限制特征值条件和$2ab \le a^2/4+4b^2$，则有

$$\begin{aligned} & n^{-1}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2+\lambda\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le 4\lambda\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_1 \\ \le & 4\lambda\sqrt{s}\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0,S}\|_2 \le 4\lambda\sqrt{s}\sqrt{\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{\phi_0^2}} \\ \le & \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{2}+\frac{8\lambda^2 s}{\phi_0^2}. \end{aligned}$$

- 整理上面结果，有

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2/(2n)+\lambda\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \leq \frac{8\lambda^2 s}{\phi_0^2}.$$

- 由限制特征值条件，和Bühlmann 和van de Geer (2011)，Bickel等(2009)的证明，有

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_q = O_P\left(s^{1/q}\phi_0^{-2}\sqrt{\log(p)/n}\right), \qquad q \in \{1,2\}.$$

例

考虑程序包care中的人脑数据集lu2004，该数据集包含年龄从26到106的30位不同年龄的样本，每个样本有404 个变量，Lu等(2004)，Zuber 和Strimmer (2011)，与Lin和Pang (2014) 分析了该数据集，目的是研究年龄与基因表达之间的关系。在研究中，响应变量取年龄的对数，并进行中心化，把其余403个基因作为协变量，并进行标准化处理。试应用本节的Lasso 方法用于该数据集的分析。

Lasso的理论: 预测和估计

- 从30 个样本中随机选取21 个作为训练集，其余9个样本作为测试集
- 用训练集进行Lasso 回归
- 采用10 折CV 选取调节参数λ
- 通过训练集数据得到回归系数的Lasso估计后，用于测试集进行预测，计算预测误差(PE)
- 重复以上试验500 次，并计算平均预测误差大小和展示箱线图

Lasso的理论: 预测和估计

- 预测误差定义为:

$$\mathrm{PE} = \frac{1}{9}\sum_{i=1}^{9}(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})^2,$$

  - $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, 9\}$为测试样本
  - $\widehat{\boldsymbol{\beta}}$为训练集得到的Lasso估计

- 图形报告了当调节参数λ取使$\mathrm{CV}(\widehat{\lambda})$最小化的$\widehat{\lambda}$和"一个标准差"准则选取的$\widetilde{\lambda}$所对应的预测误差箱线图

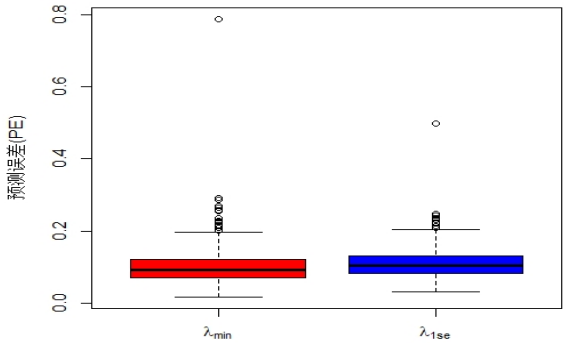


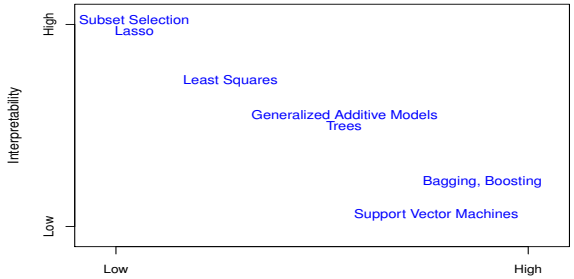
Figure: 基于500次重复试验，预测误差的箱线图。

Lasso的理论: 预测和估计

1. 当调节参数λ取使$\mathrm{CV}(\widehat{\lambda})$最小化的$\widehat{\lambda}$时，基于500次重复试验，平均模型大小约为16.20，平均预测误差约为0.1025；
2. 当采用“一个标准差"准则选取调节参数$\widetilde{\lambda}$时，平均模型大小约为11.19，平均预测误差约为0.1112；
3. $\widetilde{\lambda} > \widehat{\lambda}$，则$\widehat{\lambda}$产生的模型比$\widetilde{\lambda}$产生的模型要大，平均多了四个变量进入模型，但是平均预测误差要稍微小。

总结

估计	无偏性	稀疏性	连续性
Lasso估计	×	√	√
硬门限估计	√	√	×
Adaptive Lasso估计	√	√	√
SCAD估计	√	√	√
岭回归估计	×	×	√



总结

- Lasso估计的计算问题:
  - Use a standard quadratic programming solver (Tibshirani,1996)
  - Shooting algorithm (Fu, 1998)
  - Coordinate descent (Friedman et al., 2008, Wu and Lang, 2008), R package: glmnet, gcdnet
  - Homotopy method (Obsborn et al., 2000)
  - Least Angle Regression and LARS algorithm (Efron et al., 2003), R package: lars
- SCAD估计的计算问题:
  - LQA算法(Fan and Li, 2001)
  - MM算法(Hunter and Li, 2005)
  - LLA算法(Zou and Li, 2008)

- **大维问题：** 当$n \to \infty$时，$p \to \infty$或$p = O(n^{\kappa})$，其中$0 < \kappa < 1$
  - Oracle性质：(1) (稀疏性) $\widehat{\boldsymbol{\beta}}_{\mathrm{II}} = \mathbf{0}$；(2) (渐近正态性) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{I}} - \boldsymbol{\beta}_{0,\mathrm{I}}) \xrightarrow{L} N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{\mathrm{I}}^{*})$.
  - 思考：Lasso估计、硬门限估计、Adaptive Lasso估计、SCAD估计中哪些具有Oracle 性质吗？
  - Lasso估计满足$\sqrt{n}$-相合性要求:$\lambda = O_P(n^{-1/2})$; Lasso估计满足oracle性质要求:$\sqrt{n}\lambda \to \infty$.
  - 但是这两个条件不能同时满足，因此Lasso不具有oracle 性质
- **高维问题：** 当$n \to \infty$时，$p > n$，但是$p = O(n^{\kappa})$，其中$\kappa > 1$
  - Oracle不等式: 需要满足相容性条件和限制特征值条件(REC)。假设$\lambda \asymp \sqrt{\log(p)/n}$，有

$$n^{-1}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \le O_P(s\phi_0^{-2}\log(p)/n),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q = O_P(s^{1/q}\phi_0^{-2}\sqrt{\log(p)/n}), \quad q \in \{1, 2\}.$$

- 超高维问题：当$n \to \infty$时，$p > n$，但是$p = O(\exp(n^{\kappa}))$
  - Fan and Lv (2008)提出的SIS程序，主要定义了边际相关系数进行变量筛选
    - Sure independence screening (SIS, Fan and Lv, 2008); Generalized correlation screening (Hall and Miller, 2009); Forward regression (FR, Wang, 2009); Marginal Likelihood (Fan, et al, 2009; Fan and Song, 2010); Nonparametric screening (Fan, Feng and Song, 2011); Robust rank correlation screening (RRCS, Li et al, 2012); Model-free screening (Zhu et al, 2011); Factor profile sure screening (Wang, 2012); Marginal empirical likelihood sure screening, (Chang, 2013).
  - Lasso方法：在beta-min条件和相容性条件下，证明Lasso 方法能筛选出正确变量，即

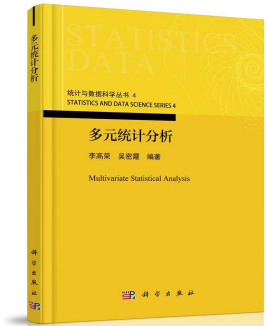
$$\underbrace{\Pr(S \subseteq \widehat{S}(\lambda)) \to 1 \quad (n \to \infty)}_{\text{Lasso is an excellent screening procedure}}$$

总结

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag Berlin Heidelberg, New York.

property	design condition	size of non-zero coeff.
slow converg. rate	no requirement	no requirement
fast converg. rate	restricted eigenvalue	no requirement
variable screening	restricted eigenvalue	beta-min condition
variable selection	neighborhood stability $\Leftrightarrow$ irrepresentable cond.	beta-min condition

- variable screening: $\Pr(S \subseteq \widehat{S}(\lambda)) \to 1 \quad (p > n \to \infty)$
- variable selection: $\Pr(S = \widehat{S}(\lambda)) \to 1 \quad (p > n \to \infty)$


STATISTICS
DATA
统计与数据科学丛书 4
STATISTICS AND DATA SCIENCE SERIES 4
多元统计分析
李高荣 吴密霞 编著
Multivariate Statistical Analysis
科学出版社

谢谢，请多提宝贵意见！