应用回归分析

上海财经大学　统计与管理学院

第十章变量选择

章节概括：

- 变量选择
- 多重共线性
- 选择标准
- 逐步回归
- 岭回归

变量选择

- X 自变量组， Y 因变量
- 将X划分为两部分， $X = (X_{\mathcal{A}}, X_{\mathcal{I}})$

  $X_{\mathcal{A}}$与回归Y 有关， $X_{\mathcal{I}}$ 与回归Y无关
- $\mathrm{E}(Y|X_{\mathcal{A}}, X_{\mathcal{I}})$ 与 $\mathrm{E}(Y|X_{\mathcal{A}})$ 相似
- 在大样本下，

$$\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{X} = \boldsymbol{\beta}'_{\mathcal{A}}\mathbf{x}_{\mathcal{A}} + \boldsymbol{\beta}'_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}$$

估计检验 $\boldsymbol{\beta}_{\mathcal{I}} = \mathbf{0}$

模拟 I

- 给定线性模型

$$y = 1 + x_1 + x_2 + 0x_3 + 0x_4 + \text{error}$$

error 为标准正态分布

- $X_1 = (x_1, x_2, x_3, x_4)$ 服从正态分布
- Case I， 均值为0

$$\text{Var}(X_1) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- 随机产生n=100的一组样本

模拟 I

TABLE 10.1 Regression Summary for the Simulated Data with No Correlation between the Predictors

	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
(Intercept)	0.8022	0.0919	8.73	0.0000
x_1	0.9141	0.0901	10.14	0.0000
x_2	0.9509	0.0861	11.04	0.0000
x_3	−0.0842	0.1091	−0.77	0.4423
x_4	−0.2453	0.1109	−2.21	0.0294

$$\hat{\sigma} = 0.911, \text{ df} = 95, R^2 = 0.714$$

$$\text{Var}(\hat{\beta}) = \frac{1}{100}\begin{pmatrix} 0.84 & 0.09 & 0.01 & -0.05 & 0.02 \\ 0.09 & 0.81 & -0.03 & -0.04 & -0.06 \\ 0.01 & -0.03 & 0.74 & -0.16 & -0.07 \\ -0.05 & -0.04 & -0.16 & 1.19 & 0.02 \\ 0.02 & -0.06 & -0.07 & 0.02 & 1.23 \end{pmatrix}$$

模拟 II

- Case II， 均值为0

$$\mathrm{Var}(X_2) = \begin{pmatrix} 1 & 0 & .95 & 0 \\ 0 & 1 & 0 & -.95 \\ .95 & 0 & 1 & 0 \\ 0 & -.95 & 0 & 1 \end{pmatrix}$$

- (1, 3)， (2, 4) 自变量强相关

模拟 II

TABLE 10.2 Regression Summary for the Simulated Data with High Correlation between the Predictors

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	0.8022	0.0919	8.73	0.0000
x_1	1.1702	0.3476	3.37	0.0011
x_2	0.2045	0.3426	0.60	0.5519
x_3	−0.2696	0.3494	−0.77	0.4423
x_4	−0.7856	0.3553	−2.21	0.0294

$$\hat{\sigma} = 0.911,\ \mathrm{df} = 95,\ R^2 = 0.702$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{100}\begin{pmatrix} 0.84 & 0.25 & 0.08 & -0.17 & 0.07 \\ 0.25 & 12.08 & 0.14 & -11.73 & -0.34 \\ 0.08 & 0.14 & 11.73 & -0.36 & 11.78 \\ -0.17 & -11.73 & -0.36 & 12.21 & 0.17 \\ 0.07 & -0.34 & 11.78 & 0.17 & 12.63 \end{pmatrix}$$

模拟 II

TABLE 10.3 Regression Summary for the Simulated Data, Correlated Case But with $n = 1100$

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.0354	0.0305	33.92	0.0000
x_1	1.0541	0.0974	10.83	0.0000
x_2	1.1262	0.0989	11.39	0.0000
x_3	−0.0106	0.0978	−0.11	0.9136
x_4	0.1446	0.1006	1.44	0.1511

$$\hat{\sigma} = 1.01,\ \text{df} = 1095,\ R^2 = 0.68$$

$$\text{Var}(\hat{\beta}) = \frac{1}{1100}\begin{pmatrix} 1.02 & -0.10 & 0.00 & 0.09 & -0.05 \\ -0.10 & 10.43 & -0.07 & -9.97 & -0.05 \\ 0.00 & -0.07 & 10.75 & 0.06 & 10.41 \\ 0.09 & -9.97 & 0.06 & 10.52 & 0.06 \\ -0.05 & -0.05 & 10.41 & 0.06 & 11.14 \end{pmatrix}$$

模拟 II

TABLE 10.4 Regression Summary for Two Candidate Subsets in the Simulated Data, Correlated Cases, with $n = 100$

	Estimate	Std. Error	t-value	Pr(>|t|)
(a) Candidate terms are intercept, x_1 and x_4				
(Intercept)	0.7972	0.0912	8.74	0.0000
x_1	0.9146	0.0894	10.23	0.0000
x_4	−0.9796	0.0873	−11.22	0.0000
$\hat{\sigma} = 0.906$, df $= 97$, $R^2 = 0.711$				
	Estimate	Std. Error	t-value	Pr(>|t|)
(b) Candidate terms are intercept, x_1 and x_2				
(Intercept)	0.8028	0.0933	8.60	0.0000
x_1	0.9004	0.0915	9.84	0.0000
x_2	0.9268	0.0861	10.76	0.0000
$\hat{\sigma} = 0.927$, df $= 97$, $R^2 = 0.691$				

多重共线性

- Multi-collinearity,复共线性

如果存在不全为0的$p+1$个数$c_0,c_1,c_2,\ldots,c_p$,使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}=0\ ,\ i=1,2,\ldots,n$$

则称自变量$x_1,x_2,\ldots,x_p$之间存在着完全多重共线性。

- 二元线性相关

$$c_1X_1+c_2X_2=c_0$$

样本相关系数

$$r_{12}^2=1$$

近似多重共线性

- 近似多重共线性

  存在不全为0的$p+1$个数$c_0,c_1,c_2,\ldots,c_p$，使得

  $$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}\approx 0\ ,\ i=1,2,\ldots,n$$

- 多元近似线性相关

  $$c_1X_1+c_2X_2+\cdots+c_pX_p\approx c_0$$

  $$X_j\approx\frac{1}{c_j}\left(c_0-\sum_{\ell\neq j}c_\ell X_\ell\right)=\frac{c_0}{c_j}+\sum_{\ell\neq j}\left(-\frac{c_\ell}{c_j}\right)X_\ell$$

  多元回归 R_j^2

多重共线性

- 多重共线性对回归模型的影响
- 多元回归模型

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

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

- 设计矩阵$\boldsymbol{X}$中的自变量列之间不相关, $\mathbf{X}$是一满秩矩阵
- 若存在完全的多重共线性, 设计矩阵$\mathbf{X}$不是一满秩矩阵，此时$|\mathbf{x}'\mathbf{x}|=0$，$(\mathbf{x}'\mathbf{x})^{-1}$不存在，正规方程组的解不唯一，回归参数的最小二乘估计表达式不成立。

近似（非）完全共线性

- 设计矩阵**X**是一满秩矩阵，但是
- **|x′x|≈0**，
- （**x′x**）$^{-1}$的对角线元素很大
- 最小二乘估计的方差阵对角线元素很大

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- $\beta_0,\beta_1,\ldots,\beta_p$的估计虽然是无偏估计，但精度很低、变差很大。不能正确判断解释变量对被解释变量的影响程度,甚至无法解释。

p=2 例子

- y对两个自变量x_1,x_2的线性回归,假定y与x_1,x_2都已经中心化，此时回归常数项为零，回归方程为

$$\hat{y}=\hat{\beta}_1 x_1+\hat{\beta}_2 x_2$$

记$L_{11}=\sum_{i=1}^{n}x_{i1}^2$，$L_{12}=\sum_{i=1}^{n}x_{i1}x_{i2}$，$L_{22}=\sum_{i=1}^{n}x_{i2}^2$，

则x_1与x_2之间的相关系数为

$$r_{12}=\frac{L_{12}}{\sqrt{L_{11}L_{22}}}$$

例子

$\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2)'$ 的协方差阵为

$$\mathrm{cov}(\hat{\boldsymbol{\beta}})=\sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

$$\mathrm{X}'\mathrm{X}=\begin{pmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{pmatrix}$$

$$(\mathrm{X}'\mathrm{X})^{-1}=\frac{1}{|\mathrm{X}'\mathrm{X}|}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}=\frac{1}{L_{11}L_{22}-L_{12}^2}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

$$=\frac{1}{L_{11}L_{22}(1-r_{12}^2)}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

例子

则

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{(1-r_{12}^2)L_{11}}$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)L_{22}}$$

随着自变量 x_1 与 x_2 的相关性增强,$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差将逐渐增大。

当 x_1 与 x_2 完全相关时,r=1,方差将变为无穷大。

共线性与方差

- p>2

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2} \frac{1}{SX_jX_j}$$

- 方差扩大因子 Variance inflation factor (VIF)

$$1/(1 - R_j^2)$$

- VIF > 1

VIF

- 经验表明,当$VIF_j \geq 10$时,就说明自变量x_j与其余自变量之间有严重的多重共线性,且这种多重共线性可能会过度地影响最小二乘估计值。
- 还可用p个自变量所对应的方差扩大因子的平均数来度量多重共线性。当

$$\overline{VIF} = \frac{1}{p}\sum_{j=1}^{p} VIF_j$$

远远大于1时就表示存在严重的多重共线性问题。

- ……

变量选择

- 有m个可供选择的变量$x_1,x_2,\ldots,x_m$,由于每个自变量都有入选和不入选两种情况，这样y关于这些自变量的所有可能的回归方程就有2^m-1个。从另一个角度看

$$C_m^0 + C_m^1 + \cdots + C_m^m = 2^m$$

- 从数据与模型拟合优劣的直观考虑出发，可用残差平方和RSS最小，或是用复相关系数R平方来衡量回归拟合的好坏。然而这两种方法都有明显的不足，

$$RSS_{p+1} \leq RSS_p \qquad R_{p+1}^2 \geqslant R_p^2$$

自由度调整复相关系数

- 自由度调整复相关系数

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- 显然有 $R_a^2 \leqslant R^2$，R_a^2 随着自变量的增加并不一定增大。

- 从拟合优度的角度追求“最优”，则所有回归子集中 R_a^2 最大者对应的回归方程就是“最优”方程。

- $$R_a^2 = 1 - \frac{n-1}{RSS}\hat{\sigma}^2$$

由于 RSS 是与回归无关的固定值，因而 R_a^2 与 $\hat{\sigma}^2$ 是等价的

信息量准则

- 选取自变量子集 X_C，
- 若 $X_C = X_A$

$$\mathrm{E}(Y|X_C = \mathbf{x}_C) = \beta'_C \mathbf{x}_C$$

$$\mathrm{E}(Y|X = \mathbf{x}) = \beta'\mathbf{X} = \beta'_A \mathbf{x}_A + \beta'_I \mathbf{x}_I$$

- 若 X_C 丢失重要的自变量，则残差平方和会比较大
- 模型拟合好坏 RSS_C
- 模型复杂度 p_C

AIC

- Akaike Information Criterion， AIC
- 赤池信息量 (Akaike, 74)，根据极大似然估计原理提出的一种较为一般的模型选择准则

$$AIC = n\log(RSS_C/n) + p_C$$

- AIC越小越好

对每一个回归子集计算AIC，其中AIC最小者所对应的模型是“最优”回归模型

AIC

- 设回归模型的似然函数为$L(\boldsymbol{\theta},\mathbf{x})$, θ的维数为p，$\mathbf{x}$为样本，在回归分析中样本为$\mathbf{y}=(y_1, y_2, \ldots y_n)'$，则AIC定义为：

$$AIC=-2\ln L(\hat{\boldsymbol{\theta}}_L,\mathbf{x})+2p$$

其中$\hat{\boldsymbol{\theta}}_L$是$\boldsymbol{\theta}$的极大似然估计，p是未知参数的个数。

假定回归模型的随机误差项e遵从正态分布，即

$$e\sim N(0,\sigma^2)$$

对数似然函数为

AIC

$$\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2} RSS$$

将$\hat{\sigma}_L^2 = \frac{RSS}{n}$代入得 $\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\frac{RSS}{n}) - \frac{n}{2}$

带入公式 $AIC=-2\ln L(\hat{\theta}_L, \mathbf{x})+2p$

这里似然函数中的未知参数个数为p+2，略去与p无关的常数，得回归模型的AIC公式为

$$AIC=n\ln(RSS)+2p$$

BIC

- Bayes Information Criterion, Schwarz (78)

$$BIC = n\log(RSS_C/n) + p_C\log(n)$$

- BIC值越小越好

Mallows’ s Cp

- Mallows’ s Cp
- 马勒斯（Mallows 73），从预测的角度提出一个可以用来选择自变量的统计量

$$C_{pc} = \frac{RSS_C}{\hat{\sigma}^2} + 2p_C - n$$

其中 $\hat{\sigma}^2 = \frac{RSS_p}{n - p - 1}$ 是全模型中σ²的无偏估计。

选择使 C_p 最小的自变量子集，这个自变量子集对应的回归方程就是“最优”回归方程。

Cp

- 考虑在n个样本点上，预测值与期望值的相对偏差平方和为：

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{ip} - E(y_i))^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2$$

可以证明，J_p的期望值是

$$E(J_p) = \frac{E(RSS_p)}{\sigma^2} - n + 2(p+1)$$

略去无关的常数2，据此构造出C_p统计量

Highway Accident 数据

TABLE 10.5 Definition of Terms for the Highway Accident Data

Variable	Description
log(*Rate*)	Base-two logarithm of 1973 accident rate per million vehicle miles, the response
log(*Len*)	Base-two logarithm of the length of the segment in miles
log(*ADT*)	Base-two logarithm of average daily traffic count in thousands
log(*Trks*)	Base-two logarithm of truck volume as a percent of the total volume
Slim	1973 speed limit
Lwid	Lane width in feet
Shld	Shoulder width in feet of outer shoulder on the roadway
Itg	Number of freeway-type interchanges per mile in the segment
log(*Sigs1*)	Base-two logarithm of (number of signalized interchanges per mile in the segment + 1)/(length of segment)
Acpt	Number of access points per mile in the segment
Hwy	A factor coded 0 if a federal interstate highway, 1 if a principal arterial highway, 2 if a major arterial, and 3 otherwise

Highway Accident 数据

TABLE 10.6 Regression Summary for the Fit of All Terms in the Highway Accident Data[a]

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	5.7046	2.5471	2.24	0.0342
logLen	−0.2145	0.1000	−2.15	0.0419
logADT	−0.1546	0.1119	−1.38	0.1792
logTrks	−0.1976	0.2398	−0.82	0.4178
logSigs1	0.1923	0.0754	2.55	0.0172
Slim	−0.0393	0.0242	−1.62	0.1172
Shld	0.0043	0.0493	0.09	0.9313
Lane	−0.0161	0.0823	−0.20	0.8468
Acpt	0.0087	0.0117	0.75	0.4622
Itg	0.0515	0.3503	0.15	0.8842
Lwid	0.0608	0.1974	0.31	0.7607
Hwy1	0.3427	0.5768	0.59	0.5578
Hwy2	−0.4123	0.3940	−1.05	0.3053
Hwy3	−0.2074	0.3368	−0.62	0.5437

$\hat{\sigma} = 0.376$ on 25 df, $R^2 = 0.791$

[a] The terms *Hwy1*, *Hwy2*, and *Hwy3* are dummy variables for the highway factor.

Highway Accident 数据

- 全模型

$$AIC = 39\log(3.5377/39) + 2 \times 14 = -65.611$$

$$BIC = 39\log(3.5377/39) + 14\log(39) = -42.322$$

$$C_p = \frac{3.537}{0.1415} + 2 \times 14 - 39 = 14$$

Highway Accident 数据

- 子集 $(\log(Len), Slim, Acpt, \log(Trks), Shld)$

$$AIC = 39\log(5.016/39) + 2 \times 6 = -67.99$$

$$BIC = 39\log(5.016/39) + 6\log(39) = -58.01$$

$$C_p = \frac{5.016}{0.1415} + 2 \times 6 - 39 = 8.453$$

Cross-Validation方法

- 交叉验证方法
- 将数据分为,建模集(construction set)和验证集（validation set)
- 建模集估计模型参数
- 验证集评估模型有效性

- 特例：Press方法

$$PRESS = \sum_{i=1}^{n} (y_i - \mathbf{x}'_{Ci}\hat{\boldsymbol{\beta}}_{C(i)})^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_{Ci}}{1 - h_{Cii}} \right)^2$$

逐步回归

- 自变量的所有可能子集构成2^m-1个回归方程，当可供选择的自变量不太多时，用前边的方法可以求出一切可能的回归方程，然后用几个选择准则去挑出“最好”的方程，但是当自变量的个数较多时，要求出所有可能的回归方程是非常困难的。为此，人们提出了一些较为简便、实用、快速的选择“最优”方程的方法。人们所给出的方法各有优缺点，至今还没有绝对最优的方法，目前常用的方法有“前进法”Forward Selection、“后退法”Backward Selection、“逐步回归法”Stepwise Selection。

逐步回归

- 在接下来的讨论中，无论我们从回归方程中剔除某个自变量，还是给回归方程增加某个自变量都要利用F检验，这个F检验与t检验是等价的

$$F_j = \frac{\Delta SSreg_{(j)}/1}{RSS/(n-p-1)}$$

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$$

Forward Selection

- 前进法
- 前进法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。首先分别对因变量 y 建立 m 个一元线性回归方程，并分别计算这 m 个一元回归方程的 m 个回归系数的 F 检验值，记为$\{F_1^1, F_2^1, \cdots, F_m^1\}$，选其最大者记为：

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著性水平α，若$F_j^1 \geqslant F_\alpha(1, n-2)$，则首先将 x_j 引入回归方程，为方便，设 x_j 就是 x_1。

前进法

- 依上述方法接着做下去。直至所有未被引入方程的自变量的F值均小于$F_{\alpha}(1, n-p-1)$时为止。这时，得到的回归方程就是最终确定的方程。

每步检验中的临界值$F_{\alpha}(1,n-p-1)$与自变量数目p有关。

Backward Selection

- 后退法
- 后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。设对 m 个回归系数进行 F 检验，记求得的 F 值为$\{F_1^m, F_2^m, \cdots, F_m^m\}$，选其最小者记为：

$$F_j^m = \min\{F_1^m, F_2^m, \cdots, F_m^m\}$$

给定显著性水平α，若$F_j^m \leqslant F_\alpha(1, n-m-1)$，则首先将 x_j 从回归方程中剔除，为方便，设 x_j 就是 x_m。

后退法

接着对剩下的 m-1 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_j^{m-1}，如果又有 $F_j^{m-1} \leqslant F_\alpha$（1,n-（m-1）-1），则剔除 x_j，重新建立 y 关于 m-2 个自变量的回归方程，依此下去，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 F_α(1,n-p-1)，没有可剔除的自变量为止。这时，得到的回归方程就是最终确定的方程。

Stepwise Selection

- 逐步回归
- 逐步回归的基本思想是“有进有出”。具体做法是将变量一个一个引入，当每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中剔除为止。这样就避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是“最优”回归子集。

逐步回归

在逐步回归中需要注意的一个问题是引入自变量和剔除自变量的显著性水平α值是不相同的，要求

$$\alpha_{进} < \alpha_{出}$$

否则可能产生“死循环”。也就是当$\alpha_{进} \geq \alpha_{出}$时，如果某个自变量的显著性P值在$\alpha_{进}$与$\alpha_{出}$之间，那末这个自变量将被引入、剔除、再引入、再剔除、…，循环往复，以至无穷。

Highway Accident 数据

TABLE 10.7 Forward Selection for the Highway Accident Data. Subsets within a Step Are Ordered According to the Value of *PRESS*

Step 1: Base terms: (logLen)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: Slim	36	6.11216	3	10.20	-66.28	-61.29	6.93325
Add: Shld	36	7.86104	3	22.56	-56.46	-51.47	9.19173
Add: Acpt	36	7.03982	3	16.76	-60.77	-55.78	9.66532
Add: Hwy	34	8.62481	5	31.96	-48.85	-40.53	10.4634
Add: logSigs1	36	9.41301	3	33.53	-49.44	-44.45	10.8866
Add: logTrks	36	9.89831	3	36.96	-47.48	-42.49	11.5422
Add: logADT	36	10.5218	3	41.37	-45.09	-40.10	12.0428
Add: Itg	36	10.962	3	44.48	-43.50	-38.51	12.5544
Add: Lane	36	10.8671	3	43.81	-43.84	-38.84	12.5791
Add: Lwid	36	11.0287	3	44.95	-43.26	-38.27	15.3326

Highway Accident 数据

Step 2: Base terms: (logLen Slim)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: logTrks	35	5.5644	4	8.33	-67.94	-61.29	6.43729
Add: Hwy	33	5.41187	6	11.25	-65.02	-55.04	6.79799
Add: logSigs1	35	5.80682	4	10.04	-66.28	-59.62	6.94127
Add: Itg	35	6.10666	4	12.16	-64.31	-57.66	7.07987
Add: Lane	35	6.10502	4	12.15	-64.32	-57.67	7.15826
Add: logADT	35	6.05881	4	11.82	-64.62	-57.97	7.18523
Add: Shld	35	6.0442	4	11.72	-64.71	-58.06	7.28524
Add: Acpt	35	5.51181	4	7.96	-68.31	-61.66	7.77756
Add: Lwid	35	6.07752	4	11.96	-64.50	-57.85	8.9025

Highway Accident 数据

Step 3: Base terms: (logLen Slim logTrks)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: Hwy	32	4.82665	7	9.12	-67.49	-55.84	6.28517
Add: Itg	34	5.55929	5	10.29	-65.98	-57.66	6.54584
Add: logADT	34	5.46616	5	9.64	-66.63	-58.32	6.6388
Add: logSigs1	34	5.45673	5	9.57	-66.70	-58.38	6.65431
Add: Lane	34	5.56426	5	10.33	-65.94	-57.62	6.66387
Add: Shld	34	5.41802	5	9.30	-66.98	-58.66	6.71471
Add: Acpt	34	5.15186	5	7.41	-68.94	-60.63	7.341
Add: Lwid	34	5.51339	5	9.97	-66.30	-57.98	8.12161

Step 4: Base terms: (logLen Slim logTrks Hwy)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: logSigs1	31	3.97747	8	5.11	-73.03	-59.73	5.67779
Add: Itg	31	4.8187	8	11.06	-65.55	-52.24	6.49463
Add: Lane	31	4.8047	8	10.96	-65.66	-52.36	6.54448
Add: logADT	31	4.82664	8	11.12	-65.49	-52.18	6.73021
Add: Shld	31	4.82544	8	11.11	-65.50	-52.19	6.88205
Add: Acpt	31	4.61174	8	9.60	-67.26	-53.95	7.72218
Add: Lwid	31	4.80355	8	10.95	-65.67	-52.37	8.05348

Highway Accident 数据

```
Step 5:  Base terms: (logLen Slim logTrks Hwy logSigs1)
Add: logADT     30    3.66683     |    9     4.92 -74.21 -59.23   5.86015
Add: Shld       30    3.97387     |    9     7.09 -71.07 -56.10   6.25166
Add: Acpt       30    3.92837     |    9     6.77 -71.52 -56.55   7.15377
Add: Lwid       30    3.97512     |    9     7.10 -71.06 -56.08   7.68299
Add: Itg        30    3.90937     |    9     6.63 -71.71 -56.74   5.70787
Add: Lane       30    3.91112     |    9     6.64 -71.69 -56.72    5.7465
Add: logADT     30    3.66683     |    9     4.92 -74.21 -59.23   5.86015
Add: Shld       30    3.97387     |    9     7.09 -71.07 -56.10   6.25166
Add: Acpt       30    3.92837     |    9     6.77 -71.52 -56.55   7.15377
Add: Lwid       30    3.97512     |    9     7.10 -71.06 -56.08   7.68299
```

Highway Accident 数据

```
Step 6:  Base terms: (logLen Slim logTrks Hwy logSigs1 Itg)
                df    RSS        |   p    C(p)    AIC    BIC   PRESS
Add: Lane       29   3.86586     |  10    8.32 -70.14 -53.51  5.78305
Add: logADT     29   3.66672     |  10    6.92 -72.21 -55.57   6.1424
Add: Shld       29   3.90652     |  10    8.61 -69.74 -53.10  6.30147
Add: Acpt       29   3.86515     |  10    8.32 -70.15 -53.52  7.17893
Add: Lwid       29   3.90718     |  10    8.62 -69.73 -53.09  7.73347

Step 7:  Base terms: (logLen Slim logTrks Hwy logSigs1 Itg Lane)
                df    RSS        |   p    C(p)    AIC    BIC   PRESS
Add: logADT     28   3.65494     |  11    8.83 -70.33 -52.03  6.31797
Add: Shld       28   3.86395     |  11   10.31 -68.16 -49.86  6.40822
Add: Acpt       28    3.8223     |  11   10.02 -68.59 -50.29  7.32972
Add: Lwid       28   3.86487     |  11   10.32 -68.15 -49.85  7.89833
```

Highway Accident 数据

Step 8: Base terms: (logLen Slim logTrks Hwy logSigs1 Itg Lane logADT)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: Shld	27	3.654	12	10.83	-68.34	-48.38	6.93682
Add: Acpt	27	3.55292	12	10.11	-69.44	-49.47	8.2891
Add: Lwid	27	3.63541	12	10.70	-68.54	-48.58	8.3678

Step 9: Base terms: (logLen Slim logTrks Hwy logSigs1 Itg Lane logADT Shld)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: Lwid	26	3.61585	13	12.56	-66.75	-45.12	8.85795
Add: Acpt	26	3.55037	13	12.09	-67.46	-45.84	9.33926

Step 10: Base terms: (logLen Slim logTrks Hwy logSigs1 Itg Lane logADT Shld Lwid)

	df	RSS	p	C(p)	AIC	BIC	PRESS
Add: Acpt	25	3.53696	14	14.00	-65.61	-42.32	11.2722

模拟数据

- n = 100
- 标准正态分布 Y $X = (X_1, \ldots, X_{50})$

TABLE 10.8 Results of a Simulated Example with 50 Terms and $n = 100$

Method	Number of Terms	R^2	p-value of Overall F	Number $\lvert t\rvert > \sqrt{2}$	Number $\lvert t\rvert > 2$
No selection	50	0.48	0.301	11	7
$\lvert t\rvert > \sqrt{2}$	11	0.28	0.001	7	5
$\lvert t\rvert > 2$	5	0.20	0.003	5	4

Windmills 数据

TABLE 10.9 Description of Data in the Windmill Data in the File `wm4.txt`

Label	Description
Date	Date and time of measurement. "2002/3/4/12" means March 4, 2002 at 12 hours after midnight
Dir1	Wind direction θ at reference site 1 in degrees
Spd1	Wind speed at reference site 1 in meters per second. Site 1 is the closest site to the candidate site
Spd2	Wind speed at reference site 2 in m/s
Spd3	Wind speed at reference site 3 in m/s
Spd4	Wind speed at reference site 4 in m/s
Spd1Lag1	Wind speed at reference site 1 six hours previously
Spd2Lag1	Wind speed at reference site 2 six hours previously
Spd3Lag1	Wind speed at reference site 3 six hours previously
Spd4Lag1	Wind speed at reference site 4 six hours previously
Bin	Bin number
Spd1Sin1	*Spd1* $\times \sin(\theta)$, site 1
Spd1Cos1	*Spd1* $\times \cos(\theta)$, site 1
CSpd	Wind speed in m/s at the candidate site

6个模型

- Model 1: $\mathrm{E}(CSpd|Spd1) = \beta_0 + \beta_1 Spd1.$
- Model 2 : 不同的截距和斜率
- Model 3:

$$\begin{aligned}\mathrm{E}(CSpd|X) &= \beta_0 + \beta_1 Spd1 + \beta_2 \cos(\theta) + \beta_3 \sin(\theta) \\ &= +\beta_4 \cos(\theta) Spd1 + \beta_5 \sin(\theta) Spd1\end{aligned}$$

6个模型

- Model 4: $\mathrm{E}(CSpd|X) = \beta_0 + \beta_1 Spd1 + \beta_2 Spd1Lag1$

- Model 5:

$$\mathrm{E}(CSpd|X) = \beta_0 + \beta_1 Spd1 + \beta_2 Spd2 + \beta_3 Spd3 + \beta_4 Spd4$$

- Model 6:

$$\begin{aligned}\mathrm{E}(CSpd|X) = {} & \beta_0 + \beta_1 Spd1 + \beta_2 Spd2 + \beta_3 Spd3 + \beta_4 Spd4 \\ & + \beta_5 Spd1Lag1 + \beta_6 Spd2Lag1 + \beta_7 Spd3Lag1 \\ & + \beta_8 Spd4Lag1\end{aligned}$$

6个模型

TABLE 10.10 Summary Criteria for the Fit of Six Mean Function to the Windmill Data

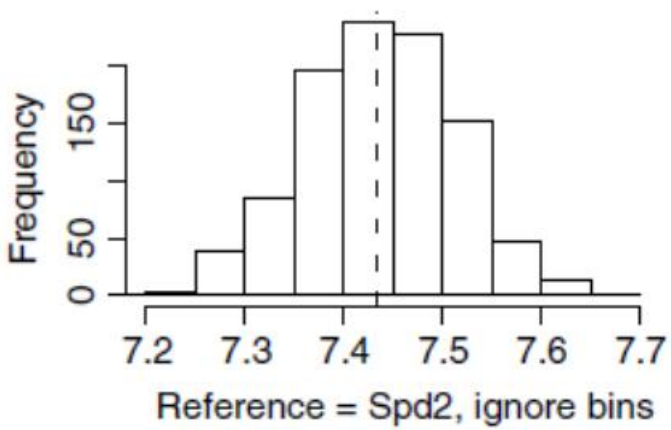
	df	*AIC*	*BIC*	*PRESS*
Model 1	2	2014.9	2024.9	6799.0
Model 2	32	1989.3	2149.8	6660.2
Model 3	6	2020.7	2050.8	6836.3
Model 4	3	1920.6	1935.6	6249.1
Model 5	5	1740.6	1765.7	5320.2
Model 6	9	1711.2	1756.3	5188.5

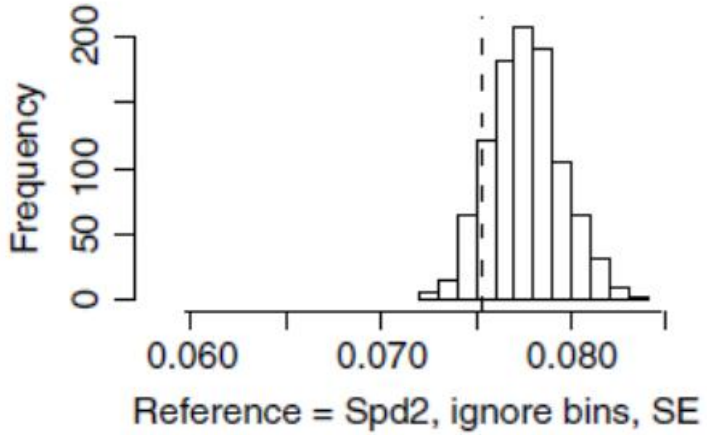
wm5数据

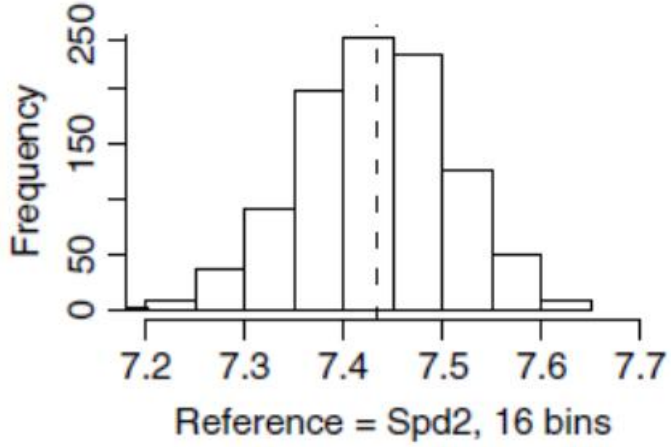
- 4个备选点，55年的数据
- 用3个备选点的数据建立模型
- 用1个备选点的数据检验模型

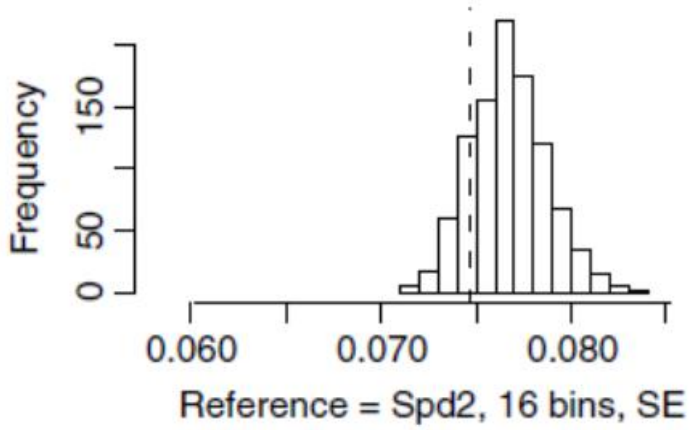
- 1：随机抽样数据
- 2：建立模型，进而预测比较结果
- 3：重复1和2两步1000次，用直方图归纳结果

wm5数据


Frequency
0
50
150
7.2
7.3
7.4
7.5
7.6
7.7
Reference = Spd2, ignore bins


Frequency
0
50
100
200
0.060
0.070
0.080
Reference = Spd2, ignore bins, SE


Frequency
0
50
150
250
7.2
7.3
7.4
7.5
7.6
7.7
Reference = Spd2, 16 bins


Frequency
0
50
150
0.060
0.070
0.080
Reference = Spd2, 16 bins, SE

wm5数据

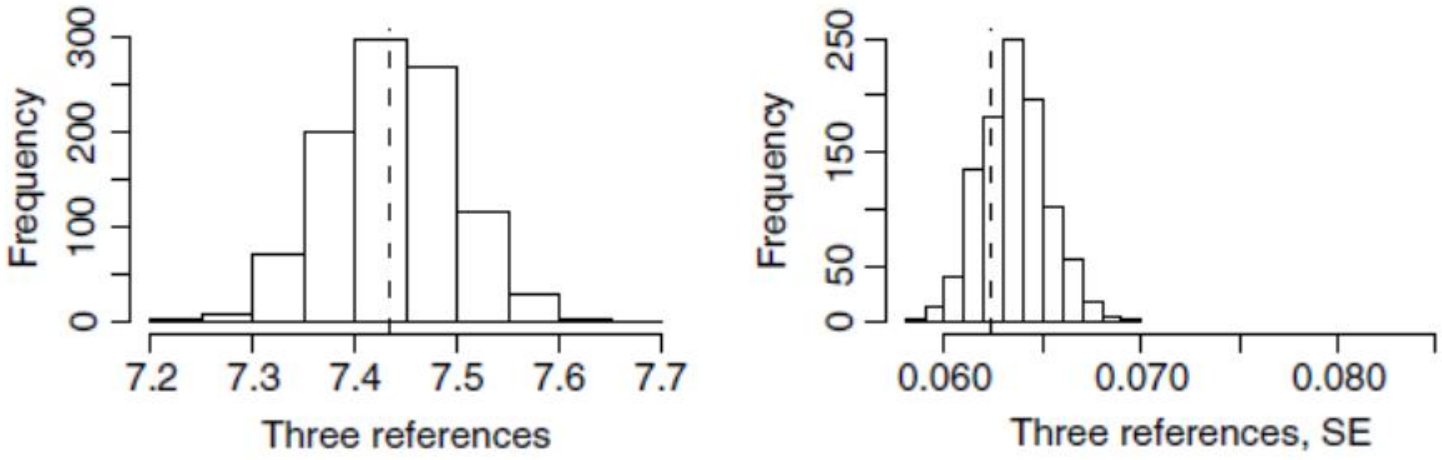


FIG. 10.1 Summary of the simulation for the windmill data. The first column gives a histogram of the estimated mean wind speed at reference site 1 for 1000 simulations using three mean functions. The second column gives a histogram of the 1000 standard errors. The dashed lines give the true values, the average of the wind speed measurements from 1948 to 2003 for the averages, and the standard deviation of the 1000 averages from the simulation for the standard errors.

岭回归 (Ridge Regression)

- 当自变量间存在(近似)复共线性时，回归系数估计的方差就很大， 估计值就很不稳定
- 实例：$y=10+2x_1+3x_2+\varepsilon$

		1	2	3	4	5	6	7	8	9	10
(1)	x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
(2)	x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
(3)	ε_i	0.8	−0.5	0.4	−0.5	0.2	1.9	1.9	0.6	−1.5	−1.5
(4)	y_i	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

最小二乘法估计：$\hat{\beta}_0=11.292$，$\hat{\beta}_1=11.307$，$\hat{\beta}_2=-6.591$

原模型的参数：$\beta_0=10, \beta_1=2, \beta_2=3$

x_1，x_2的样本相关系数得 $r_{12}=0.986$，x_1与 x_2之间高度相关。

岭回归

- 当自变量间存在复共线性时，$|\mathbf{X}'\mathbf{X}|\approx 0$，

  设想给$\mathbf{X}'\mathbf{X}$加上一个正常数矩阵$k\mathbf{I}$，（$k>0$），

  那么$\mathbf{X}'\mathbf{X}+k\mathbf{I}$接近奇异的程度就会比$\mathbf{X}'\mathbf{X}$接近奇异的程度小得多

- 考虑到变量的量纲问题，先对数据做标准化，为了记号方便，标准化后的设计阵仍然用$\mathbf{X}$表示

岭回归估计

- 岭回归估计，其中k称为岭参数

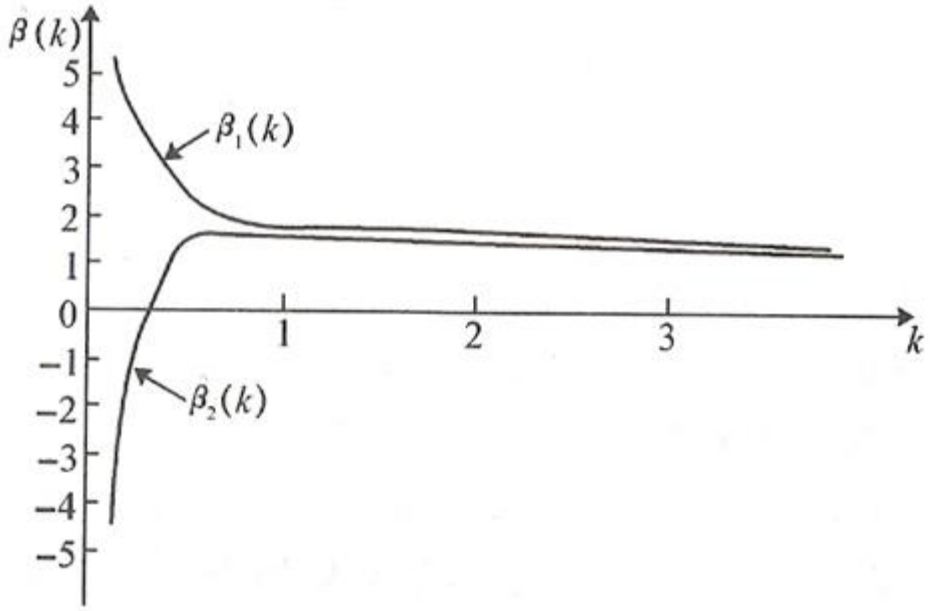
$$\hat{\boldsymbol{\beta}}(\mathrm{k})=(\mathbf{X}'\mathbf{X}+\mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

由于假设$\mathbf{X}$已经标准化，所以$\mathbf{X}'\mathbf{X}$就是自变量样本相关阵。因变量观测向量$\mathbf{y}$可以经过标准化也可以未经标准化。显然，岭回归做为$\boldsymbol{\beta}$的估计应比最小二乘估计稳定，当$\mathrm{k}=0$时的岭回归估计就是普通的最小二乘估计。

岭回归估计

- 因为岭参数 k 不是唯一确定的，所以我们得到的岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 实际是回归参数 $\boldsymbol{\beta}$ 的一个估计族。

k	0	0.1	0.15	0.2	0.3	0.4	0.5	1.0	1.5	2	3
$\hat{\beta}_1(k)$	11.31	3.48	2.99	2.71	2.39	2.20	2.06	1.66	1.43	1.27	1.03
$\hat{\beta}_2(k)$	−6.59	0.63	1.02	1.21	1.39	1.46	1.49	1.41	1.28	1.17	0.98



岭回归估计的性质

- **性质 1**　$\hat{\boldsymbol{\beta}}(k)$ 是回归参数$\boldsymbol{\beta}$ 的有偏估计。

证明：

$$E[\hat{\boldsymbol{\beta}}(k)]=E[(X'X+kI)^{-1}X'y]$$

$$=(X'X+kI)^{-1}X'E(y)$$

$$=(X'X+kI)^{-1}X'X$$

岭回归估计的性质

性质 2 在认为岭参数 k 是与 y 无关的常数时，$\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，也是 y 的线性函数。

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

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

因此，岭估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的一个线性变换，根据定义式 $\hat{\boldsymbol{\beta}}(k)=(X'X+kI)^{-1}X'y$ 知 $\hat{\boldsymbol{\beta}}(k)$ 也是 y 的线性函数。但在实际应用中，由于岭参数 k 总是要通过数据来确定，因而 k 也依赖于 y，因此从本质上说 $\hat{\boldsymbol{\beta}}(k)$ 并非 $\hat{\boldsymbol{\beta}}$ 的线性变换，也不是 y 的线性函数。

岭回归估计的性质

- **性质 3** 对任意 $k>0$，$\|\hat{\boldsymbol{\beta}}\|\neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\|<\|\hat{\boldsymbol{\beta}}\|$$

这里 $\|\bullet\|$ 是向量的模，等于向量各分量的平方和。

这个性质表明 $\hat{\boldsymbol{\beta}}(k)$ 可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种向原点的压缩，从 $\hat{\boldsymbol{\beta}}(k)$ 的表达式可以看到，当 $k\to\infty$ 时，$\hat{\boldsymbol{\beta}}(k)\to 0$，即 $\hat{\boldsymbol{\beta}}(k)$ 化为零向量。用岭回归进行变量选择。

岭回归估计的性质

- **性质 4**　以 MSE 表示估计向量的均方误差，则存在 k＞0，使得

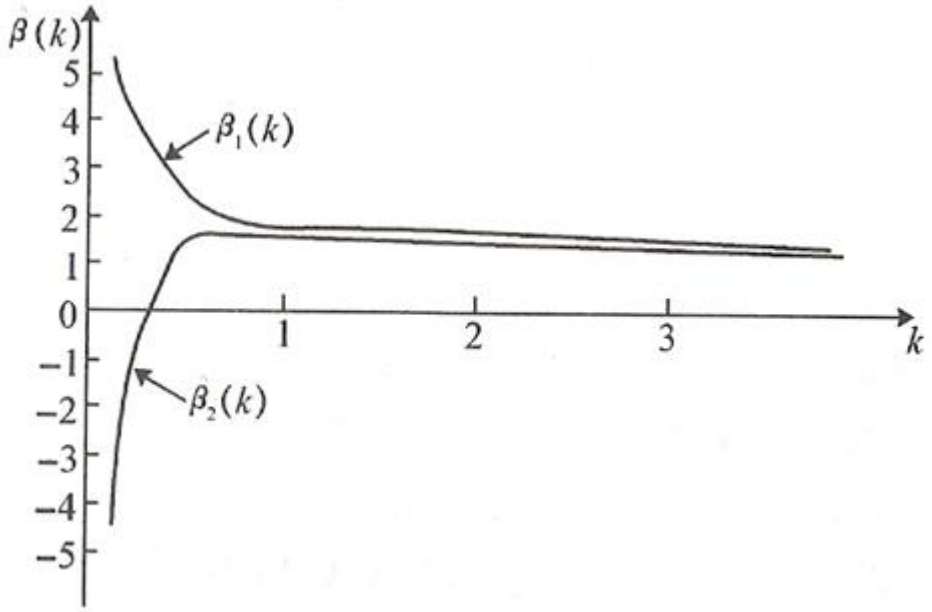
$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) < \mathrm{MSE}(\hat{\boldsymbol{\beta}})$$

即

$$\sum_{j=1}^{p} \mathrm{E}(\hat{\beta}_j(k) - \beta_j)^2 < \sum_{j=1}^{p} Var(\hat{\beta}_j)$$

岭参数选择

- 岭迹法:



（1）各回归系数的岭估计基本稳定；
（2）岭估计的符号、绝对值合理；
（3）残差平方和增大不太多。

岭参数选择

- 方差扩大因子

岭估计 $\hat{\boldsymbol{\beta}}(k)$ 的协方差阵，

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}(k)) &= \operatorname{cov}(\hat{\boldsymbol{\beta}}(k), \hat{\boldsymbol{\beta}}(k)) \\ &= \operatorname{cov}((X'X+kI)^{-1}X'y, (X'X+kI)^{-1}X'y) \\ &= (X'X+kI)^{-1}X'\operatorname{cov}(y, y)X(X'X+kI)^{-1} \\ &= \sigma^2(X'X+kI)^{-1}X'X(X'X+kI)^{-1} \\ &= \sigma^2(c_{ij}(k))\end{aligned}$$

矩阵 $C_{ij}(k)$ 的对角元 $c_{jj}(k)$ 就是岭估计的方差扩大因子。$c_{jj}(k)$ 随着 k 的增大而减少，选择 k 使所有方差扩大因子 $c_{jj}(k) \leqslant 10$。

岭参数选择

- 残差平方和

岭估计在减小均方误差的同时增大了残差平方和，希望岭回归的残差平方和RSS（k）的增加幅度控制在一定的限度以内，可以给定一个大于1的c值，要求：

$$\mathrm{RSS}(k) < c\mathrm{RSS}$$

寻找使上式成立的最大的k值。

Thank You !