2013.5.22

第二十七讲. 统计学习 (III)：变量选择算法以及惩罚最小二乘

回顾AIC准则

$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2q$，其中$L(\hat{\theta})$为似然函数极大值，$q$为参数个数

- AIC通常称为模型选择准则，有时未必用来选变量 (比如选择Weibull分布、gamma、log-normal模型之一）。
- 为什么称作“信息”准则？与熵 $E_f\log(f)$ 有关

Derivation of AIC：

假设$y_1,\ldots,y_n$来自于真模型 $g(y)$，设 $f(y;\theta)$为候选模型之一.

记号：似然函数$\mathrm{L}(\theta) = \prod f(y_i;\theta)$，极大似然估计$\hat{\theta}$.

我们希望估计得到的模型 $\hat{f} = f(y;\hat{\theta})$与$g$的距离尽量小。

$\hat{f}, g$的Kullback-Leibler距离：$\mathrm{K}(g,\hat{f}) = \int g\log\left(\frac{g}{\hat{f}}\right) = \int g\log(g) - \int g\log(\hat{f})$

极小化 $\mathrm{K}(g,\hat{f}) \Leftrightarrow$ 极大化 $\mathrm{K} = \mathrm{K}(\hat{\theta}) = \int g(y)\log(f(y;\hat{\theta}))\mathrm{d}y$

试图以如下 $\overline{\mathrm{K}}$逼近K

$$\overline{\mathrm{K}} = \frac{1}{n}\sum\log(f(y_i;\hat{\theta})) = \frac{\log L(\hat{\theta})}{n}$$

我们将证明其偏差： $E(\overline{\mathrm{K}}-\mathrm{K}) \approx \mathrm{q/n}$。

(1) 首先证明 $\mathrm{E}(\overline{\mathrm{K}}) \approx \int g(y)\log(f(y;\theta))\mathrm{d}y + q/2n$

利用如下事实：

- $\log L(\theta)$在$\hat{\theta}$处Taylor展开，注意$\left.\frac{\partial \log L}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0$，近似地有

$$\log L(\theta) = \log L(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})'\frac{\partial^2\log L}{\partial\theta^2}(\theta-\hat{\theta}),$$

- $\hat{\theta} \sim \mathrm{N}\left(\theta, \left(-\frac{\partial^2\log L}{\partial\theta^2}\right)^{-1}\right)$

$$\Rightarrow \mathrm{E}\log L(\theta) \approx \mathrm{E}\log L(\hat{\theta}) - q/2$$

$$\Rightarrow \mathrm{E}(\overline{\mathrm{K}}) \approx \mathrm{E}\log L(\theta)/n + q/2n \approx \int g(y)\log(f(y;\theta))\mathrm{d}y + q/2n$$

（2）另一方面，我们将证明 $E(\mathrm{K}) = \int g(y)\{\log f(\mathrm{y},\theta)\}\mathrm{d}y - \frac{q}{2n}$。

我们假设 在真参数 θ处，$\int g(y)\{\log f(\mathrm{y},\theta)\}\mathrm{d}y$达到极大，

则$\int g(y)\left\{\frac{\partial\log f(\mathrm{y},\theta)}{\partial\theta}\right\}\mathrm{d}y = 0$

利用事实

- $\log(f(y;\hat{\theta}))$在θ处展开：

$$\log(f(y;\hat{\theta})) \approx \log f(\mathrm{y},\theta) + \frac{\partial\log f}{\partial\theta}(\hat{\theta}-\theta) + \frac{1}{2}(\hat{\theta}-\theta)'\frac{\partial^2\log f}{\partial\theta^2}(\hat{\theta}-\theta)$$

- $\mathrm{E}\frac{\partial^2\log f}{\partial\theta^2} \approx \frac{1}{n}\frac{\partial^2\log L}{\partial\theta^2}$（大数律）

$$\Rightarrow \mathrm{K} = \int g(y)\log(f(y;\hat{\theta}))\mathrm{d}y$$

$$\approx \int g(y)\{\log f(\mathrm{y},\theta)\}\mathrm{d}y + \frac{1}{2}E\left\{(\hat{\theta}-\theta)'\left(\frac{1}{n}\frac{\partial^2 \log L}{\partial\theta^2}\right)(\hat{\theta}-\theta)\right\}$$

$$\Rightarrow \mathrm{E(K)} = \int g(y)\{\log f(\mathrm{y},\theta)\}\mathrm{d}y - \frac{q}{2n}$$

(3) 结合前面的 $\mathrm{E}\overline{\mathrm{K}} \approx \int g(y)\log(f(y;\theta))\mathrm{d}y + q/2n$

$\Rightarrow \mathrm{E}\log\left(\overline{\mathrm{K}} - q/n\right) = E(\mathrm{K})$

为了极大化K，可近似地极大化 $\overline{\mathrm{K}} - q/n = L(\hat{\theta})/n - q/n$

故取 $\mathrm{AIC} = -2n\left\{L(\hat{\theta})/n - q/n\right\} = -2\mathrm{L}(\hat{\theta}) + 2q$，并极小化之。

6. 变量选择算法

回归分析中，使用某种准则，通常是**AIC(或BIC)**，搜索具有最小**AIC**的模型。搜索方法有：

(1). 最优子集选择方法(best subset selection)

(2). 逐步回归法 (Stepwise selection):

向前法,

向后法,

向前-向后法

(3). Forward stagewise selection (matching pursuit)

(4). Leaps and bound

…

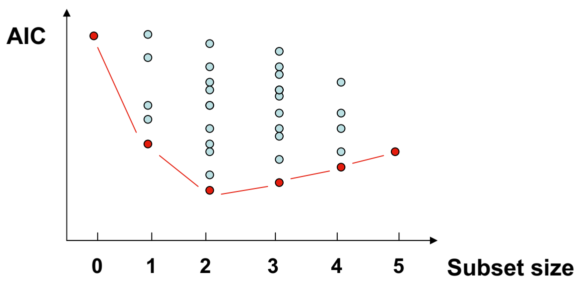
(1). 最优子集选择方法(subset selection)

$p-1$个自变量，共2^{p-1}个子集

(1) 对$k=0,1,2,\ldots,p-1$,在所有$\binom{p-1}{k}$个k-自变量模型中找出

RSS最小（R^2最大，AIC最小）的模型.

(2) 比较这p个最优子集的AIC，选出最后的模型。



搜索所有可能的2^{p-1}个子模型，使

某种准则达到最小。p较大时不可行。$p-1=30: 2^{p-1} \sim 10$亿

(2). 逐步回归(stepwise regression)

逐步回归方法是一种贪心算法(greedy algorithm),不需要搜索所有2^{p-1}个子模型,而是AIC沿跳跃最大的路径(故称为greedy),搜索O(p^2)个子模型.

包括向前法,向后法,以及两者的综合.

- 向前法(Forward selection):

从0个自变量的回归模型开始,逐步添加变量：

从尚未进入模型的变量中,找出最能改进模型RSS的变量,

如果加入该变量使得模型的AIC(或其它准则)变小,

则将该变量加入模型;

否则如果加入任何变量都不能改善AIC,停止.

• 向后法(backward elimination):
从全模型开始，每次删除一个使RSS变化最小的自变量.
每步保证ＡＩＣ减小，否则停止.

• 向前－向后法：基本是向前法，结合向后法，
即在每步添加变量后,考察已入选的自变量是否需要删除.

Remark:
- 向前或向后法的逐步选取的变量子集是嵌套的(*nested*),递增或递减的.
- 逐步回归方法得到的解很多情况下是全局最优的.

```
> step(full.model, method="both")
#method:  both, backward, forward
```


(3). Forward stagewise regression (向前阶段回归)

响应 y,自变量$x_1,\ldots,x_p$

(1°) $y \sim 1$,残差r_0

(2°) $j_1 = \arg\min < y, x_j >, \quad r_0 \sim x_{j_1} \Rightarrow \hat{\beta}_{j_1}$,残差$r_1$

(3°) $j_2 = \arg\min_{j \neq j_1} < y, x_j >, \quad r_1 \sim x_{j_2} \Rightarrow \hat{\beta}_{j_2}$,残差$r_2$

....

若$\| r_k \| < C$, STOP.

拟合模型为$y = \bar{y} + x_{j_1}\hat{\beta}_{j_1} + x_{j_2}\hat{\beta}_{j_2} + \ldots + x_{j_k}\hat{\beta}_{j_k}$

注1: Stagewise regression 可处理 $p >> n$的情况.

注2: Stagewise regression 是 matching pursuit的一种.

注:主成分回归与此类似

在X平面上寻找最具有代表性的方向：主成分(正交特征向量):

奇异值分解：$X_{n\times p} = U_{n\times p}\Lambda_{p\times p}V'_{p\times p}$,其中$U$, V列正交,
记$U = (\mathbf{u}_1,\ldots,\mathbf{u}_p)$
所有主成分：$Z = XV = U\Lambda = (\mathbf{u}_1\lambda_1,\ldots,\mathbf{u}_p\lambda_p)$

用前k个主成分 $\mathbf{z}_1 = \mathbf{u}_1\lambda_1$, ..., $\mathbf{z}_k = \mathbf{u}_k\lambda_k$ 张成的空间逼近 $\mathrm{L(X)} = \mathrm{L(Z)}$
投影：$\tilde{\mathrm{Y}} = \mathrm{P_Z Y}$

7. 惩罚最小二乘（规则化方法）

（1）岭估计

岭估计(ridge estimator): 模型 $\mathrm{Y} = \mathrm{X}\beta + \varepsilon$,
$\tilde{\beta}^{(\mathrm{Ridge})} = (X'X + \lambda I_p)^{-1}X'Y$ 称为岭估计 ($\lambda > 0$为常数).

定理:
(a) 岭估计是压缩估计: $\| \tilde{\beta}^{(\mathrm{Ridge})} \| < \| \hat{\beta} \|$, $\hat{\beta} = (X'X)^{-1}X'Y$
(b) 存在$\lambda > 0$, 使得$\mathrm{MSE}(\tilde{\beta}^{(\mathrm{Ridge})}) \leq \mathrm{MSE}(\hat{\beta})$,
即基于某些岭估计比基于LS估计的预测误差更小。

证：略

注意到：

$$\tilde{\beta}^{(\text{Ridge})} = \text{argmin}\left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|^2 \right\}$$

$$\Leftrightarrow \tilde{\beta}^{(\text{Ridge})} = \text{argmin} \| Y - X\beta \|^2, \text{约束} \| \beta \| < t$$

所以岭估计是规则化估计, 即约束/惩罚 (β 长度) 的LS估计。

一般情况下使得MSE达到最小的λ，可由CV决定。

当如果X列正交标准化($X'X = I$)时,最优的λ显式表达，

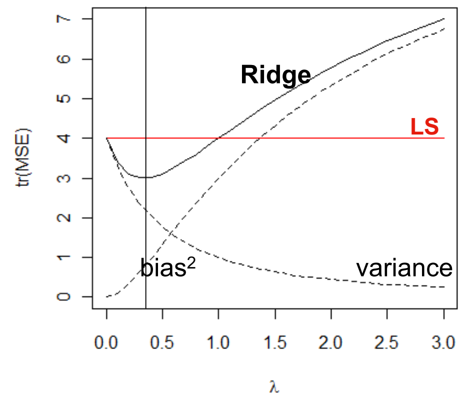
此时岭估计： $\tilde{\beta}^{(\text{ridge})} = \hat{\beta}/(1+\lambda)$,

$$MSE(\tilde{\beta}^{(\text{ridge})}) = \text{variance} + \text{bias}^2 = \frac{\sigma^2}{(1+\lambda)^2} I_p + \frac{\lambda^2}{(1+\lambda)^2}\beta\beta'$$

λ_{optimal}有显示表达：$\lambda_{\text{optimal}} = \dfrac{p\sigma^2}{\| \beta \|^2}$,(效应弱或误差方差大时大幅度压缩)

$$MSE(\tilde{\beta}^{(\text{ridge})}) = \text{variance} + \text{bias}^2 = \frac{\sigma^2}{(1+\lambda)^2} I_p + \frac{\lambda^2}{(1+\lambda)^2}\beta\beta'$$

$\lambda \uparrow$, variance $\downarrow$, bias^2 $\uparrow$



注：岭估计与James-Stein 估计

线性模型中β的James-Stein估计定义为:

$$\tilde{\beta}_{JS} = \left(1 - \frac{(p-2)\hat{\sigma}^2}{\| \hat{\beta} \|^2}\right)\hat{\beta}, \text{ 其中}\hat{\beta}\text{是LS估计},$$

则 $\text{MSE}(\tilde{\beta}_{JS}) \le \text{MSE}(\hat{\beta})$

X列正交情形下，岭估计的$\lambda_{\text{optimal}} = \dfrac{p\sigma^2}{\| \beta \|^2}$,

但是σ^2，β是未知参数， plug - in LS 估计，

$$\text{岭估计 } \tilde{\beta}^{\text{ridge}}(\lambda_{\text{optimal}}) = \frac{\hat{\beta}}{1 + \dfrac{p\hat{\sigma}^2}{\| \hat{\beta} \|^2}} \approx \left(1 - \frac{p\hat{\sigma}^2}{\| \hat{\beta} \|^2}\right)\hat{\beta},$$

以$p-2$替代p,即James – Stein估计。对一般的 X，类似。

(2) LASSO估计

LASSO: least absolute shrinkage and selection operator

LASSO 估计 ：

$$\tilde{\beta}^{(\text{lasso})} = \text{argmin}\left\{ \| Y - X\beta \|^2 + 2\lambda \| \beta \|_1 \right\}$$

$$\Leftrightarrow \tilde{\beta}^{(\text{lasso})} = \text{argmin} \| Y - X\beta \|^2, \text{约束} \| \beta \|_1 < t,$$

其中 L_1模 $\| u \|_1 = \sum | u_i |$.

注1．LASSO对回归系数的**L1**模进行约束，所以是压缩估计。

注2．LASSO方法把一些回归系数估计为**0**，可认为是一种变量选择方法。

LS的目标函数 $\|Y-X\beta\|^2=\|Y-X\hat{\beta}\|^2+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})$,

$\min\|Y-X\beta\|^2 \Leftrightarrow \min(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})$， 下图中的椭圆

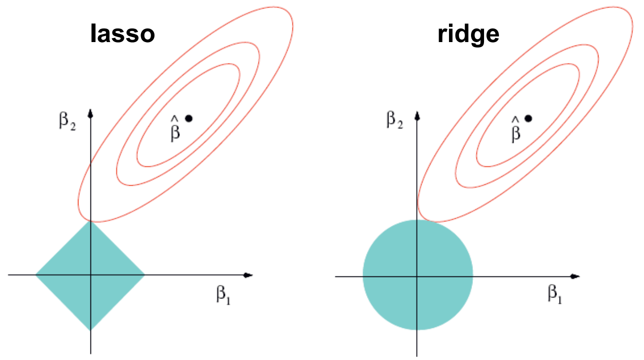


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1|+|\beta_2|\le t$ and $\beta_1^2+\beta_2^2\le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

对于X列正交标准化的情形，LASSO估计有显式表达：

$\tilde{\beta}_j^{(\text{lasso})}=\text{sgn}(\hat{\beta}_j)(|\hat{\beta}_j|-\lambda)_+$，其中 $x_+=x1_{(x>0)}$，

$$即\tilde{\beta}_j^{(\text{lasso})}=\begin{cases}\hat{\beta}_j-\lambda, & 若\hat{\beta}_j>\lambda\\ \hat{\beta}_j+\lambda, & 若\hat{\beta}_j<-\lambda\\ 0, & 若|\hat{\beta}_j|<\lambda\end{cases}$$

所以，LASSO得到的估计是对LS估计的压缩：

当$|\hat{\beta}_j|$较小时，把它压缩为0

当$|\hat{\beta}_j|$较大时，减去λ