变量选择以及R实现

马学俊

xuejunma@suda.edu.cn

苏州大学

内容

Caredit数据

```
data <- read.csv("D://Credit.csv",header=T)#读入数据
kable(head(data)[1:7])
```

X	Income	Limit	Rating	Cards	Age	Education
1	14.891	3606	283	2	34	11
2	106.025	6645	483	3	82	15
3	104.593	7075	514	4	71	11
4	148.924	9504	681	3	36	11
5	55.882	4897	357	2	68	16
6	80.180	8047	569	4	77	10

```
kable(head(data)[8:12])
```

Gender	Student	Married	Ethnicity	Balance
Male	No	Yes	Caucasian	333
Female	Yes	Yes	Asian	903
Male	No	No	Asian	580
Female	No	No	Asian	964
Male	No	Yes	Caucasian	331
Male	No	No	Caucasian	1151

- Caredit数据包含10个变量:X,Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married, Ethnicity, Balance, 分别是标识、收入（单位是千美元）、信用额度、信用级别（连续变量）、信用卡数量、年龄、受教育年限、性别、学生、婚姻状况、种族（白种人、非裔美国人、亚洲人）、个人平均信用卡债务。
- 我们目的是分析哪些因素影响Balance。
- 样本量为400。

变量选择的意义

变量选择是统计学研究的重要领域。在实际中，研究问题包括很多个变量，有些变量是不显著的，应该去掉这些变量，以增加模型的精度。具体来说：

- 预测准确率(Prediction Accuracy)：
  - $n \gg p$ OLS 估计得到的方差比较小，测试集表现比较好；但不满足这个条件时（此时$n > p$），OLS得到的结果可能存在比较大的差异，导致过度拟合。即训练集表现非常好，而测试集表现不理想。
  - $p > n$，OLS得到的结果不唯一。此时需要利用其它技术估计系数，该技术以牺牲估计偏差为代价(偏差比较大)，显著减小估计量的方差。
- 模型的解释能力(Model Interpretablity)：常常存在自变量与因变量无关或者关系比较弱的情况，这样增加的模型的复杂度。

变量选择的方法主要有最优子集方法、逐步回归法、LASSO和SCAD等。

最优子集逐步回归

- 最优子集方法是对自变量的所有可能的组合进行OLS拟合，通过某种规则（R^2和BIC等）选择一个最优的模型
- 缺点：当自变量个数多时，运行速度非常慢。
- 逐步回归：向前逐步回归和向后逐步回归。

常见的规则的讨论：

- $RSS = \sum(y_i - \hat{y}_i)^2$
- $R^2 = 1 - \frac{RSS}{TSS}$，其中$TSS = \sum(y_i - \bar{y})$

这两种基于测试集误差选择最优模型的，没有考虑到模型复杂的指标、测试误差。通常有两种解决方法:

- 根据拟合导致的偏差对训练误差进行调整，间接地估计测试误差，如C_p、AIC、BIC。
  - $C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2)$，其中$\hat{\sigma}$是误差项的标准差。
  - $AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2p\hat{\sigma}^2)$
  - $BIC = \frac{1}{n}(RSS + \log(n)p\hat{\sigma}^2)$，$BIC$得到的模型规模比$C_p$小，因为$n > 7$时$\log n > 2$
- 通过交叉核实(Cross-validation)方法，直接估计测试误差。

Hitters数据

```
library(ISLR)
dim(Hitters)
```

```
## [1] 322  20
```

```
sum(is.na(Hitters$Salary))
```

```
## [1] 59
```

```
Hitters=na.omit(Hitters)
dim(Hitters)
```

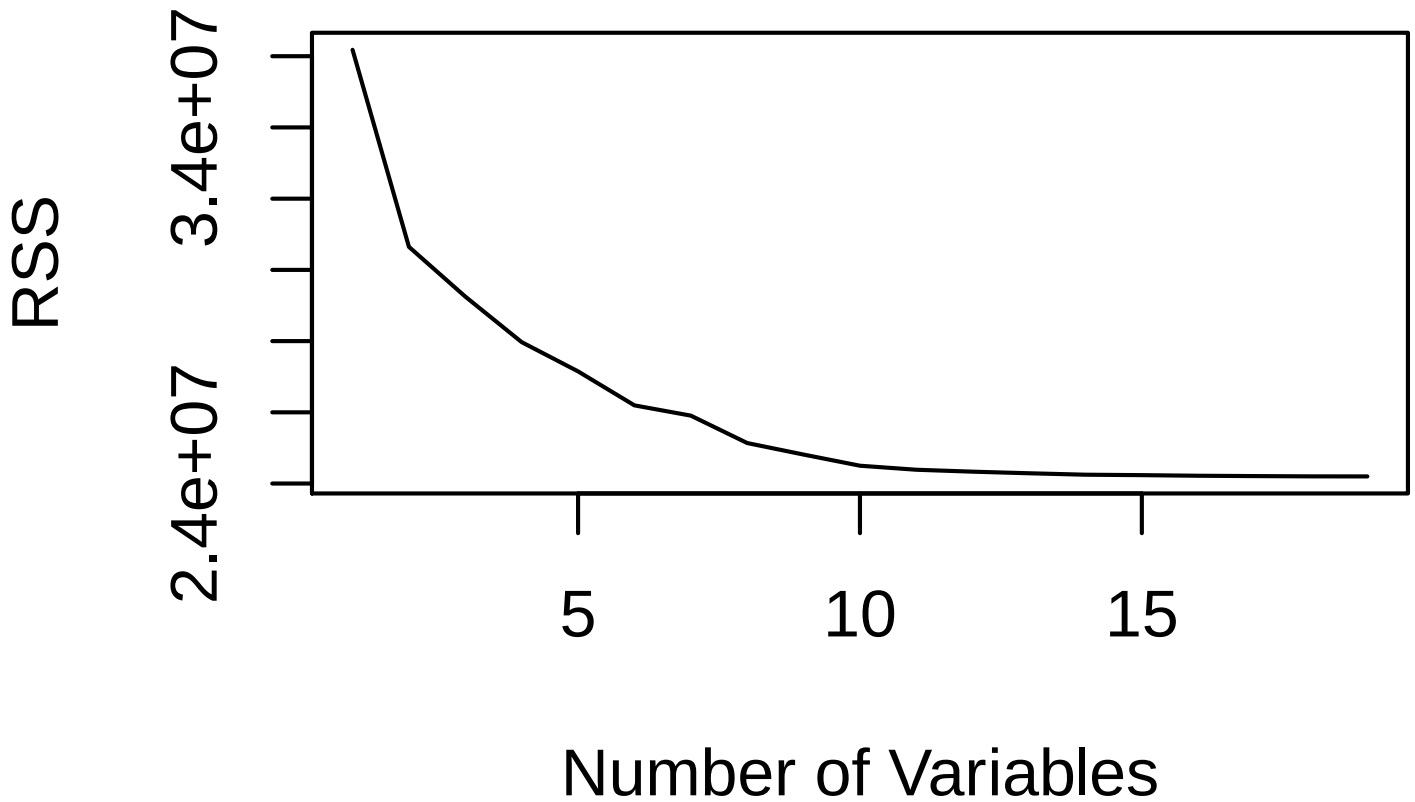
```
## [1] 263  20
```

```
sum(is.na(Hitters))
```

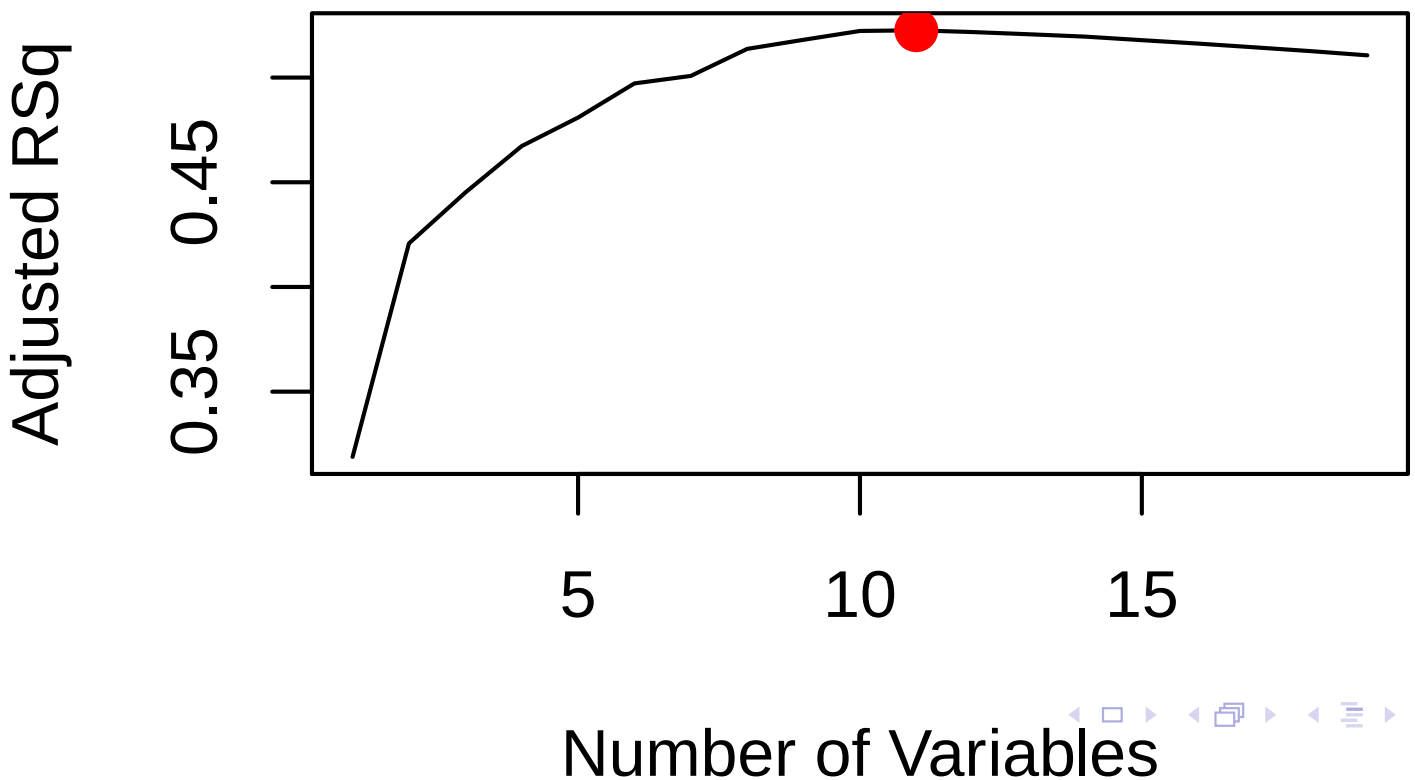
```
## [1] 0
```

```
library(leaps)
regfit.full=regsubsets(Salary~.,data=Hitters,nvmax=19)
reg.summary=summary(regfit.full)
```

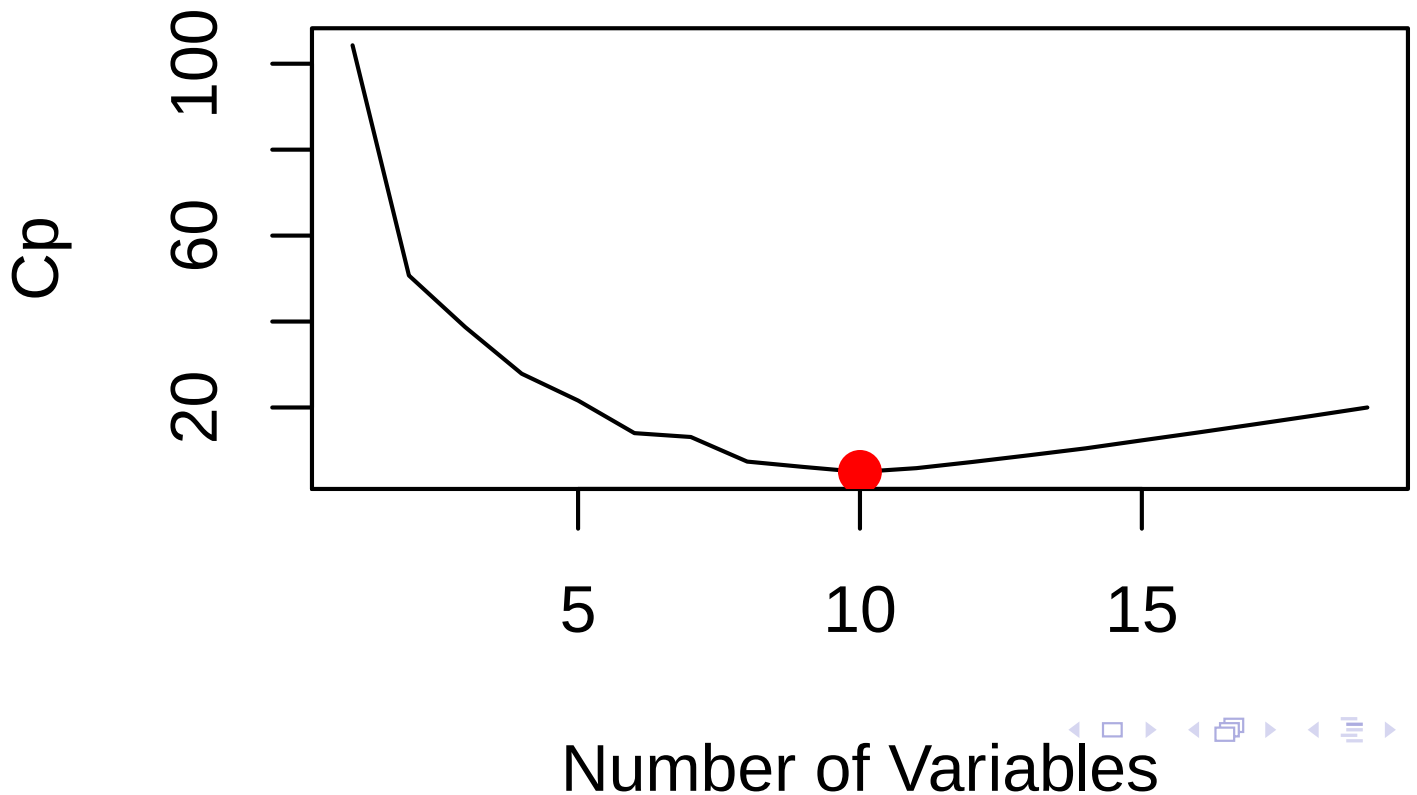
```
plot(reg.summary$rss,xlab="Number of Variables",
     ylab="RSS",type="l")
```



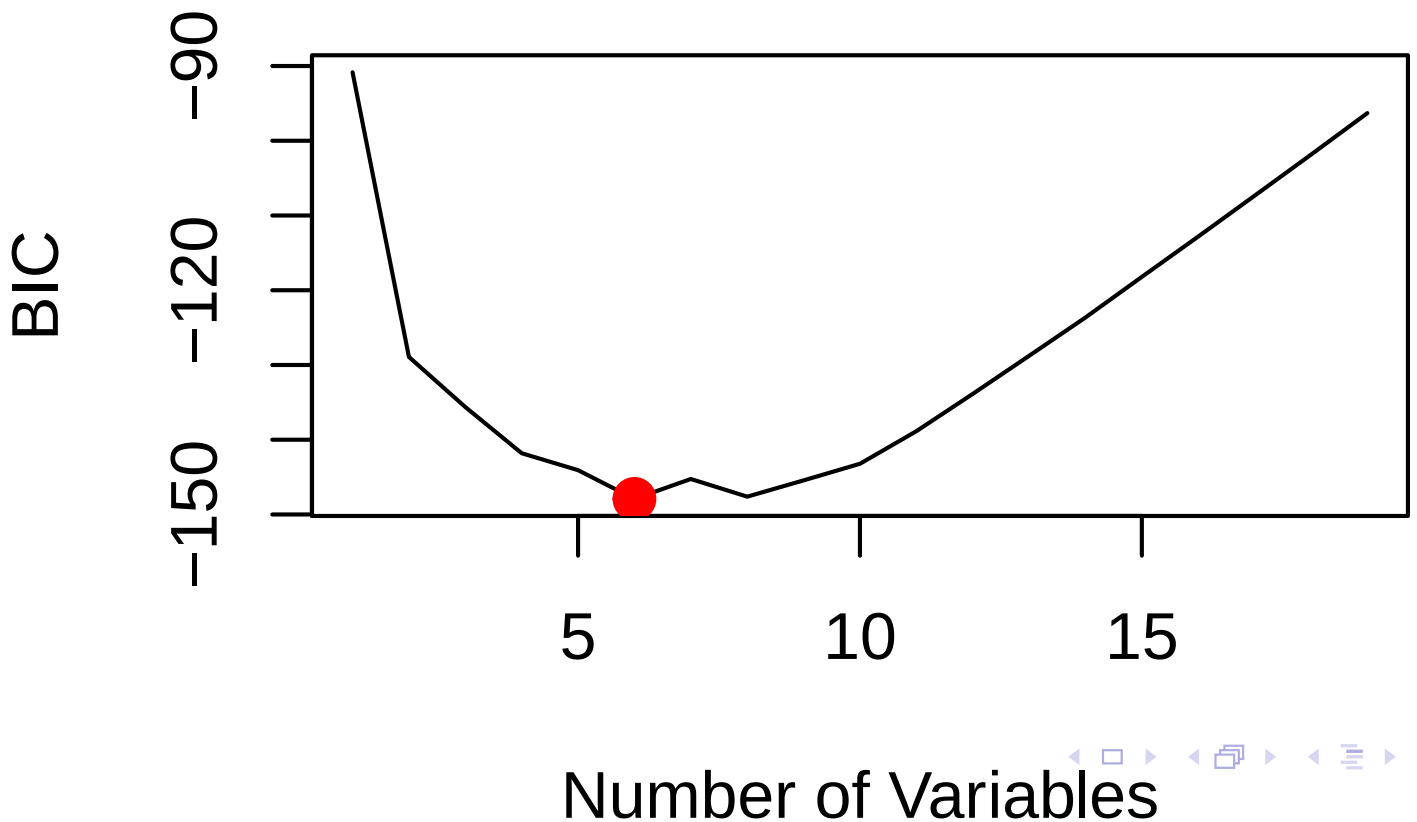
```
plot(reg.summary$adjr2,xlab="Number of Variables",
     ylab="Adjusted RSq",type="l")
## which.max(reg.summary$adjr2)
points(11,reg.summary$adjr2[11], col="red",cex=2,pch=20)
```



```
plot(reg.summary$cp,xlab="Number of Variables",
     ylab="Cp",type='l')
## which.min(reg.summary£cp)
points(10,reg.summary$cp[10],col="red",cex=2,pch=20)
```



```
plot(reg.summary$bic,xlab="Number of Variables",
     ylab="BIC",type='l')
## which.min(reg.summary£bic)
points(6,reg.summary$bic[6],col="red",cex=2,pch=20)
```



岭回归

岭回归(Ridge Regression)利用惩罚的方式选择最优的模型。设线性模型为

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i,$$

岭回归是求下面目标函数的最小值：

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

其中$RSS = \sum_{i=1}^{n}(Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2$。$\lambda$是截断参数(Tuning Parameter)。

- $\lambda \to \infty$时，岭回归的系数估计值趋于0。
- $\lambda \to 0$时，岭回归和OLS相同。

$\lambda \sum_{j=1}^{p} \beta_j^2$是惩罚项，对参数具有压缩作用。需要注意的时，惩罚项没有对常数项进行惩罚，因为当自变量进行中心化时，$\hat{\beta}_0 = \sum_{i=1}^{n} \frac{Y_i}{n}$。

下面以Credit数据说明岭回归的应用。从图1可以看出：

1. 随着λ增加系数向零趋近。
2. 数值比较小的已经压缩接近为零了。
3. 随着λ的增加$\hat{\beta}^R$的l_2在降低，从而$\|\hat{\beta}^R\|_2/\|\hat{\beta}\|_2$也在降低。
4. OLS的估计系数具有尺度不变形(Scale Invariant)，但岭回归没有这个性质，为了消除量纲的影响，通常进行标准化处理。

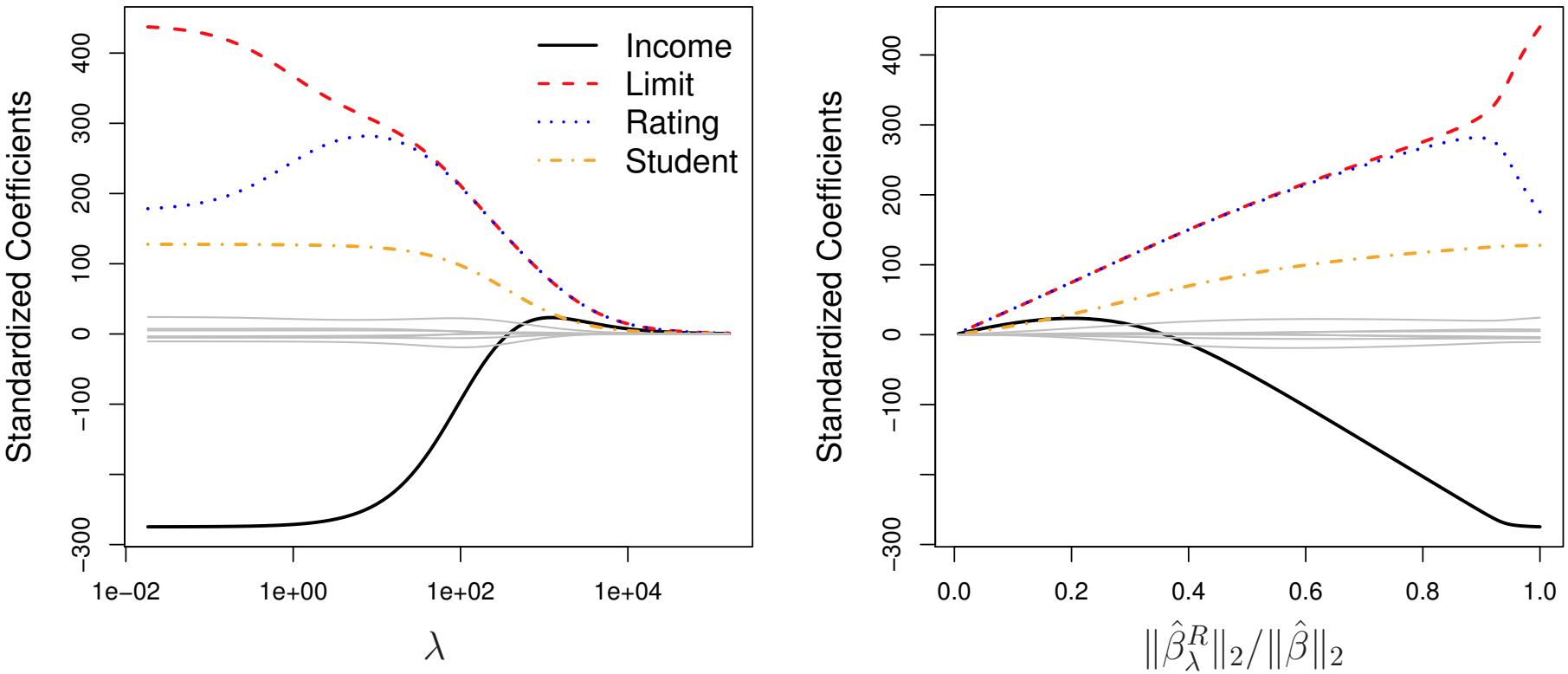


Figure 1: Credit数据标准化后岭回归系数变化情况

比较OLS和岭回归

- 岭回归的优势在于权衡了偏差和方差。
- 随着λ的增加，方差在降低，但偏差在增加。
- 当$\lambda = 0$，OLS的偏差很小，但方差很大。
- 通常来说，当因变量与自变量关系近似线性时，OLS会有较低的偏差，但方差比较大。也就是说训练集一个微小的改变可能导致OLS系数的变化比较大。特别当自变量个数和样本量差不多时，最小二乘的方差很大(图2)。
- 如果$p > n$，OLS没有唯一解。而岭回归通过增加小幅度的变差换取方差大幅度的下降，权衡了偏差和方差获得了比较好的效果。

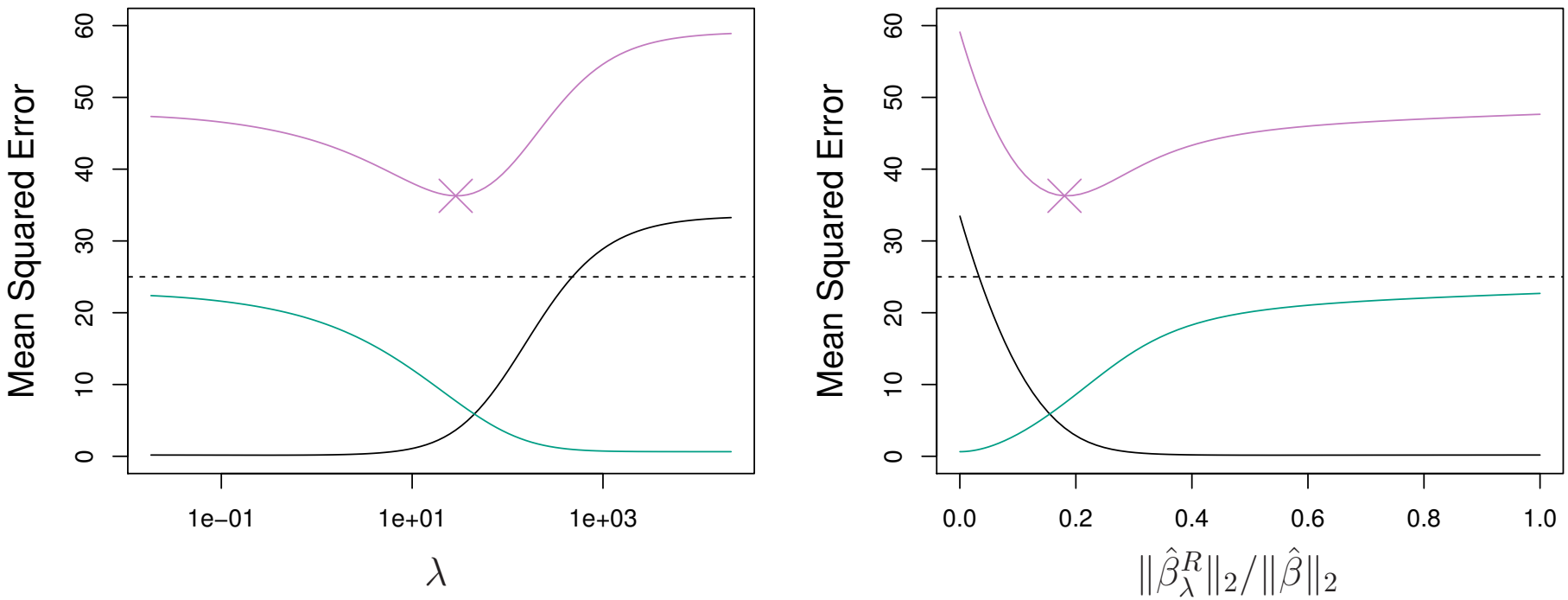


Figure 2: 模拟数据的预测结果：黑线是偏差的平方，绿线是方差，紫线是均方误差；水平线是MSE可能达到的最低水平。$p = 45, n = 50$。

LASSO

岭回归的缺点是不能将系数压缩为0，当自变量比较大时不变解释。Tibshirani(1996)提出LASSO (Least Absolute Shrinkage and Selection Operator)，它利用l_1范数对参数进行惩罚，其目标函数：

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- LASSO得到是稀疏模型(Sparse Model)。
- “稀疏性”假定，即假定只有少数的自变量对于因变量产生影响，也就说自变量系数为零的很多，非零的很少。这种假定在一定程度是合理性的，因为对某一个事物的影响也许有很多因素，但起主要作用也许只有很少的几个。
- 稀疏性假定是处理超高维（高维）问题的基本假定。如果我们的目的不是得到稀疏模型，那么LASSO不是最优选择，降维(Reduction Dimensiona)是不错的选择。

LASOO和岭回归的比较

LASSO和岭回归可以等价于以下问题：

$$\text{minimize} RSS, \sum_{j=1}^{p} |\beta_j| \leq s \tag{5.1}$$

和

$$\text{minimize} RSS, \sum_{j=1}^{p} (\beta_j)^2 \leq s \tag{5.2}$$

以两维为例说明两种方法的区别，也解释LASSO可以选择变量的原因。

- 5.1的限制条件是$|\beta_1 + \beta_2| \leq s$是菱形，
- 5.2的限制条件是$\beta_1^2 + \beta_2^2 \leq s$是个圆。
- 图3中以OLS估计$\hat{\beta}$为中心的每一椭圆代表某一个RSS数值，即误差的等高线，也就是说椭圆边界上的每一个RSS值相同的。
- 椭圆与$\hat{\beta}$远近，RSS越大。
- 当s很大时，LASSSO和岭回归可以得到OLS结果。但通常他们限制s的取值。LASSO和岭回归的估计是有约束区域可椭圆第一次交点决定的。
- 岭回归的约束区域是圆的，与椭圆的交点每一点的概率几乎一样，一般不会出现在坐标轴上，所以得到的估计不是稀疏的，也就数估计系数不会为零。
- LASSO的约束区域是菱形，与椭圆的交点最易出现在坐标轴上，所以LASSO得到的估计是稀疏的。

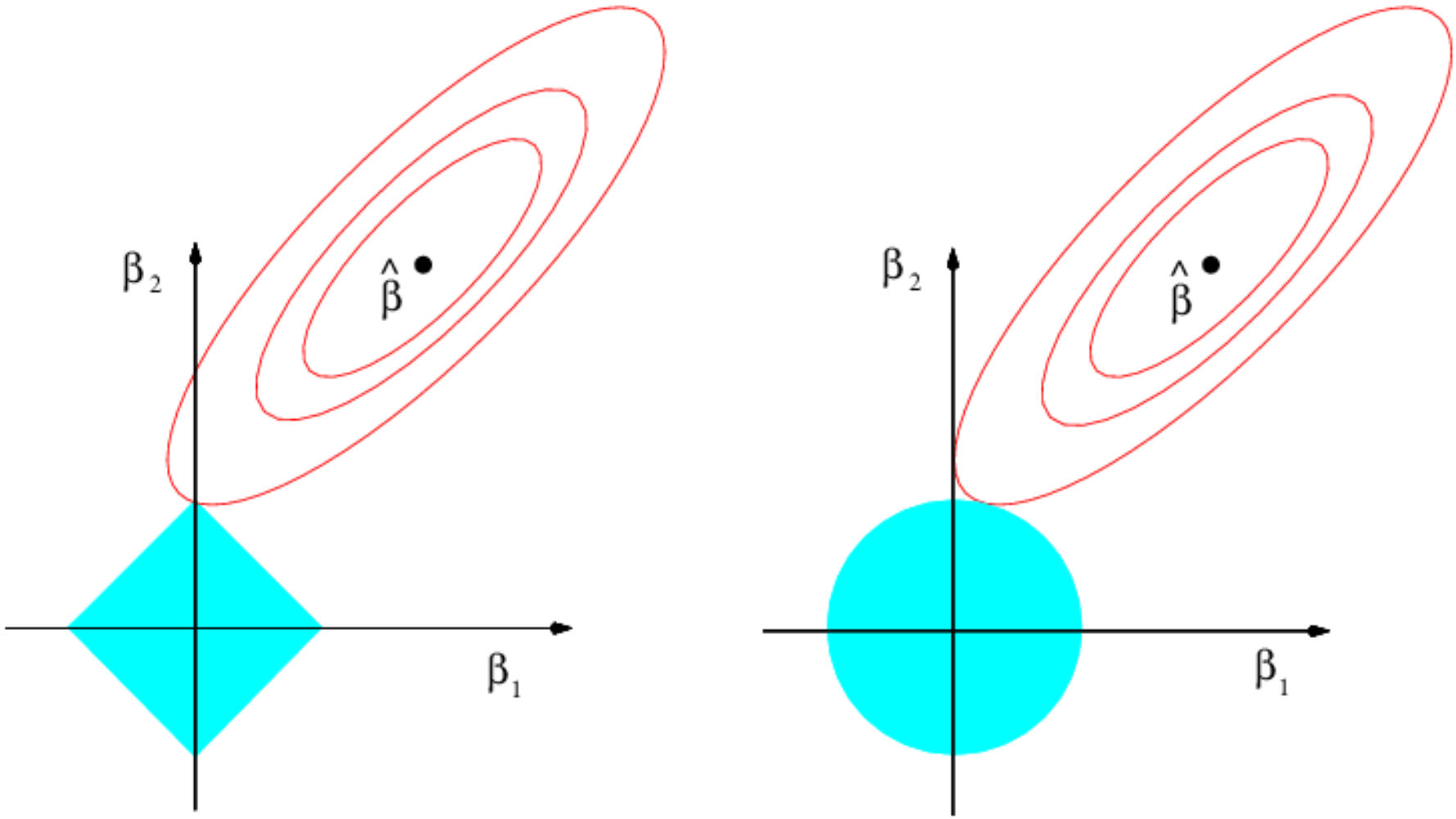


Figure 3: 误差等高线和限制区域，左图是LASSO，右图是岭回归

实例分析

我们仍利用Hiters数据。λ可以利用AIC和CV方法等得到。由于CV是随机抽样，所以不同时刻的运行结果不一样。一般差异不会太大。

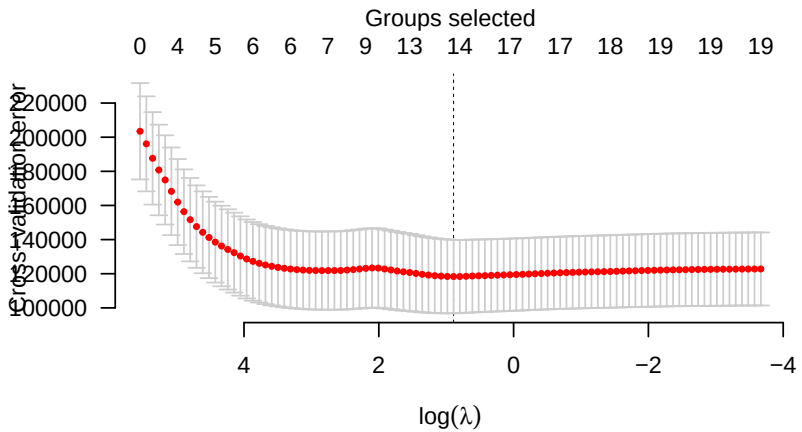
```
rm(list=ls())
library(ISLR)
Hitters=na.omit(Hitters)
x=model.matrix(Salary~.,Hitters)[,-1]
y=Hitters$Salary
library(grpreg)

## Loading required package:  Matrix

fit <- grpreg(x,y,penalty="grLasso")
plot(fit)
```

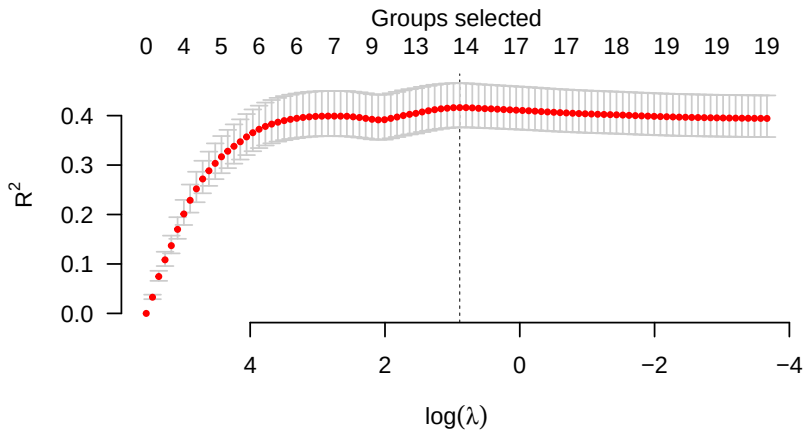
```
cvfit <- cv.grpreg(x, y)
summary(cvfit)

## grLasso-penalized linear regression with n=263, p=19
## At minimum cross-validation error (lambda=2.4368):
## -------------------------------------------------
##   Nonzero coefficients: 13
##   Nonzero groups: 13
##   Cross-validation error of 118353.74
##   Maximum R-squared: 0.42
##   Maximum signal-to-noise ratio: 0.71
##   Scale estimate (sigma) at lambda.min: 344.026
```

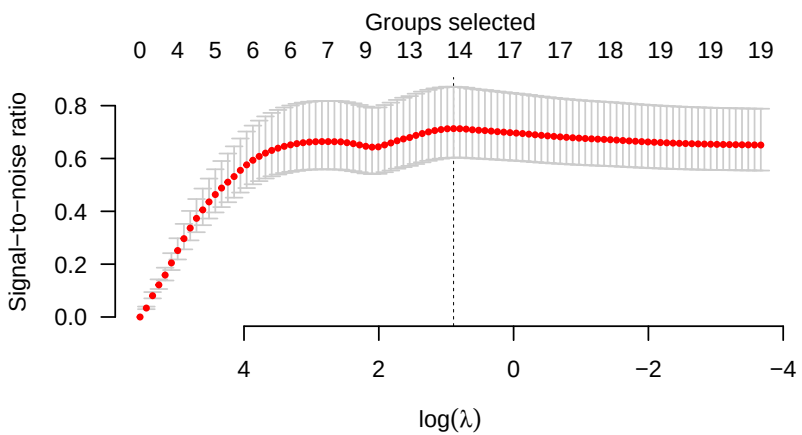
```
coef(cvfit)
```

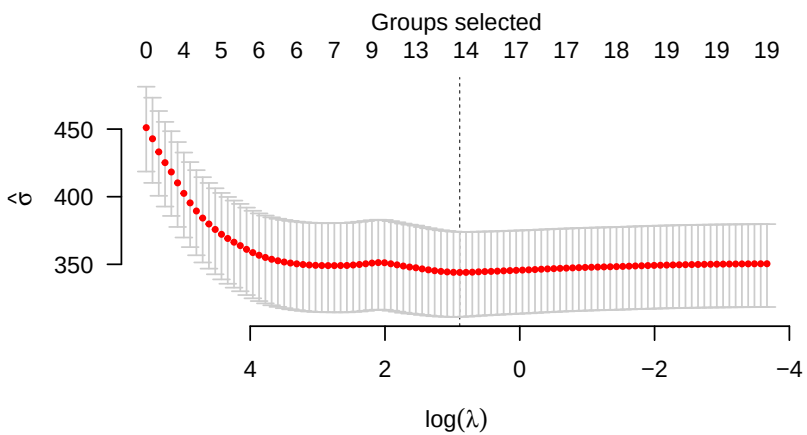
```
##  (Intercept)        AtBat         Hits        HmRun         Runs
##  129.0937321   -1.6277628    5.8466448    0.0000000    0.0000000
##          RBI        Walks        Years       CAtBat        CHits
##    0.0000000    4.8627124   -9.7615500    0.0000000    0.0000000
##       CHmRun        CRuns         CRBI       CWalks      LeagueN
##    0.5667480    0.6876766    0.3777093   -0.5605903   32.6757077
##    DivisionW      PutOuts      Assists       Errors   NewLeagueN
## -119.1496910    0.2747051    0.1868238   -2.1425844    0.0000000
```

```
par(mfrow=c(2,2))
plot(cvfit, type="all")
```









进一步讨论LASSO

下面我们一个简单的例子讨论LASSO、岭回归和OLS的关系。假设$n = p$，并X是单位矩阵。OLS是最小化

$$\sum_{i=1}^{n}(Y_i - \beta_i)^2$$

可以得到$\hat{\beta}_j = Y_j$。那么岭回归和LASSO最小化：

$$\sum_{i=1}^{n}(Y_i - \beta_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

和

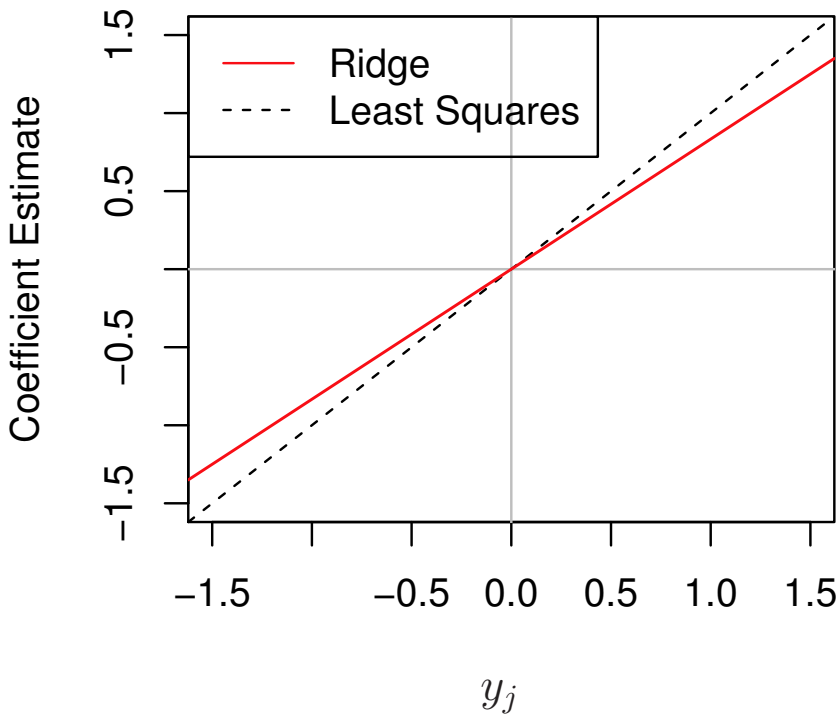
$$\sum_{i=1}^{n}(Y_i - \beta_i)^2 + \lambda \sum_{j=1}^{n} |\beta_j|$$

可以证明它们的估计是

$$\hat{\beta}_j^R = \frac{Y_j}{1+\lambda} \tag{5.3}$$

和

$$\hat{\beta}_j^L = \begin{cases} Y_j - \frac{\lambda}{2} & Y_j > \frac{\lambda}{2} \\ Y_j + \frac{\lambda}{2} & Y_j < -\frac{\lambda}{2} \\ 0 & |Y_j| \leq \frac{\lambda}{2} \end{cases} \tag{5.4}$$



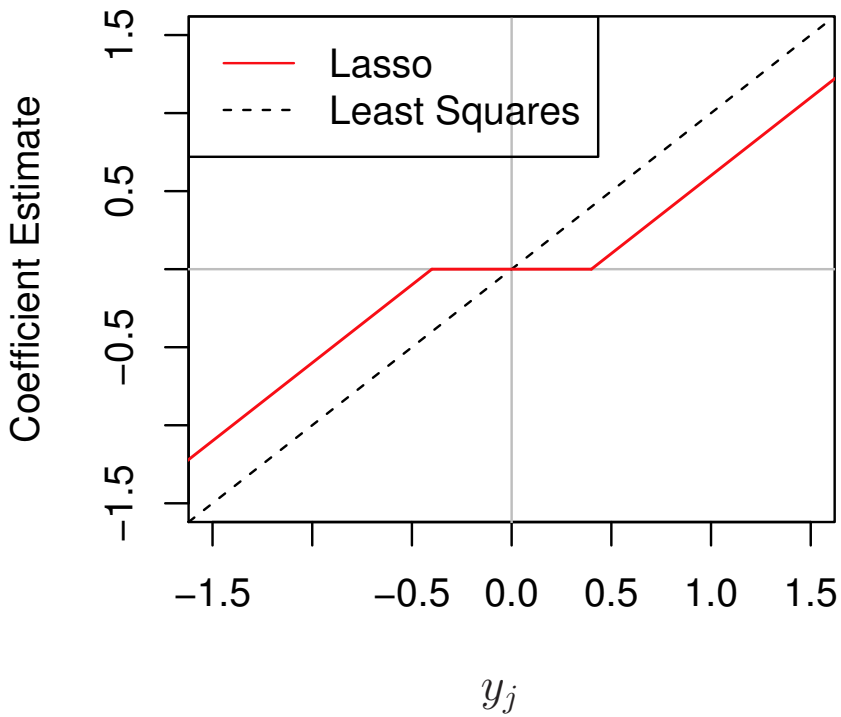


Figure 4: LASSO、岭回归和OLS估计比较

LASSO从提出了，收到统计学界以及其他相关学科学者的广泛关注，并且在此基础上开展了大量的研究工作Adaptive LASSO (Zou, 2006)、Fused LASSO (Tibshirani等, 2005)、SCAD (Fan和Li, 2001), Elastic Net (Zou和Hastie, 2005)、Dantzig (Candes 和Tao, 2007)、MCP(Zhang, 2010)、GLASSO、GSCAD和GMCP (Huang 和Breheny, 2012; Breheny和Huang, 2015) 等

下面我们再进一步讨论LASSO和延伸方法。我们考虑线性模型

$$Y = X^\top \boldsymbol{\beta} + \varepsilon$$

设模型中不含常数项，$Y \in R^n$，$X \in R^{n\times p}$，且$X^\top X = E_n$其中E_n是单位矩阵。则$\hat{\boldsymbol{\beta}}_{OLS} = (X^\top X)^{-1} X^\top Y = X^\top Y$。我们将LASSO的目标函数，写成一般的形式：

$$\hat{\boldsymbol{\beta}} = \text{argmin} \frac{1}{2} \|Y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|) \tag{5.5}$$

由于$\hat{Y}_{OLS} = XX^\top Y, \hat{\boldsymbol{\beta}}_{OLS} = X^\top Y$，且$X^\top(Y - \hat{Y}_{OLS}) = 0$，所以

$$\begin{aligned}
\|Y - X\boldsymbol{\beta}\|^2 &= (Y - X\boldsymbol{\beta})^\top (Y - X\boldsymbol{\beta}) \\
&= (Y - \hat{Y}_{OLS} + \hat{Y}_{OLS} - X\boldsymbol{\beta})^\top (Y - \hat{Y}_{OLS} + \hat{Y}_{OLS} - X\boldsymbol{\beta}) \\
&= \|Y - \hat{Y}_{OLS}\|^2 + (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \\
&\quad + 2(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^\top X^\top (Y - \hat{Y}_{OLS}) \\
&= \|Y - \hat{Y}_{OLS}\|^2 + (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})
\end{aligned}$$

进一步得到5.5等于

$$\hat{\boldsymbol{\beta}} = \text{argmin}\frac{1}{2}\sum_{j=1}^{p}(\hat{\beta}_{OLS,j} - \beta_j) + \lambda\sum_{j=1}^{p}p_j(|\beta_j|) \tag{5.6}$$

为了简单，我们将问题简单转化为

$$\min\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \tag{5.7}$$

其中$p_\lambda(|\theta|)$取不同的函数得到不同的估计量。

1. $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)I(|\theta < \lambda|) \Longrightarrow$ 最优子集估计
2. $p_\lambda(|\theta|) = \lambda|\theta|^2 \Longrightarrow$ 岭回归估计
3. $p_\lambda(|\theta|) = \lambda|\theta| \Longrightarrow$ LASSO估计
4. $p'_\lambda(|\theta|) = \lambda\left\{I(\theta \leq \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta > \lambda)\right\} \Longrightarrow$ SCAD估计

它们的估计量与OLS估计量之间的关系。

- 最优子集：$\hat{\theta} = zI(|z| > \lambda)$ 硬罚(Hard Thresholding Penalty)
- 岭回归：$\hat{\theta} = \frac{z}{1+2\lambda}$
- LASSO ：$\hat{\theta} = sgn(z)(|z| - \lambda)_+$ 软罚(Soft Thresholding Penalty)
- SCAD:

$$\hat{\theta} = \begin{cases} sgn(z)(|z| - \lambda)_+ & |z| < 2\lambda \\ \frac{(a-1)z - sgn(z)a\lambda}{a-2} & 2\lambda \leq |z| \leq a\lambda \\ z & |z| \geq a\lambda \end{cases} \tag{5.8}$$

为了比较不同罚函数，我们重新考虑5.7，并且给出评价的三种准则(Fan 和Li,2001)。

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

对上述式子进行一阶求导得：

$$sgn(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z \tag{5.9}$$

三种准则是：

- 无偏性(Unbiasedness)：对于重要变量的系数估计要尽量无偏。显然对于较大的θ，$p'_\lambda(|\theta|)=0$,可以得到$\hat{\theta}=z$，即得到渐近无偏的。LASSO的导数λ，显然不满足无偏性。无偏的**充分条件**是$p'_\lambda(|\theta|)=0$。
- 稀疏性(Sparsity)：得到的估计应该是稀疏的，从而能够起到变量选择的效果。当z比较小时，$\hat{\theta}$应该为0。如果$|\theta|+p'_\lambda(|\theta|)$关于$\theta$的最小值大于0，即，则$|z|<\min_{\theta\neq 0}\{|\theta|+p'_\lambda(|\theta|)\}$ 时，则5.9在$(0,+\infty)$为正，$(-\infty,0)$为负，即5.7式的目标函数在在$(0,+\infty)$为递增，$(-\infty,0)$为递减，因此最优解$\hat{\theta}=0$。稀疏性的**充分条件**是$\min_\theta\{|\theta|+p'_\lambda(|\theta|)\}>0$。
- 连续性(Continuity)：得到的估计应该是样本数据的连续函数，从而保证变量选择的稳定性。连续的**充要条件**是$\min_\theta\{|\theta|+p'_\lambda(|\theta|)\}$ 在零处取得。LASSO的$\min_\theta\{|\theta|+p'_\lambda(|\theta|)\}=\min_\theta|\theta|+\lambda$ 在$\theta=0$ 处取得，所以LASSSO估计量是连续的。

表1总结了四种惩罚的特点。

Table 1: 不同惩罚函数的特点

	无偏性	稀疏性	连续性
最优子集	无偏	稀疏	不连续
岭回归	有偏	不稀疏	连续
LASSO	有偏	稀疏	连续
SCAD	无偏	稀疏	连续

组惩罚

- 当某些协变量为多分类离散变量时，我们通常引入哑变量(Dummy Variables)转化为多个二分类解释变量
- 另一种情形是对于非线性可加模型，我们通常选择一组基函数(比如B-样条基函数等)，将非参函数在该组基函数下展开，从而将非线性可加模型转化为线性模型.

在这两种或其他类似情况下，我们考虑变量选择时，与某个协变量相对应的是一组(group) 哑变量(基函数变量)必须同时被选中或者同时不被选中.

表达式

Zou和Yuan(2008)提出Group LASSO, Huang和Breheny(2012)进行了详细的研究。

$$\hat{\boldsymbol{\beta}} = \text{argmin}\frac{1}{2}\|Y - \sum_{j=1}^{J} X_j\boldsymbol{\beta}_j\|^2 + \lambda\sum_{j=1}^{p} p_j(||\boldsymbol{\beta}_j||_{R_j}) \tag{5.10}$$

其中$\|v\|_R = (v^\top Rv)$。

- 将自变量进行分块，对每一块的系数进行惩罚。该惩罚可以使LASSO、SCAD和MCP。

R实现

```
grpreg(X, y, group=1:ncol(X),
       penalty=c("grLasso", "grMCP", "grSCAD"),
       family=c("gaussian","binomial","poisson"))
```

- group需要设置，必须从1开始的整数。
  如group=c(1,1,2,3)表示前两个变量是一组变量。
- X和y必须是矩阵。

Caredit数

Caredit

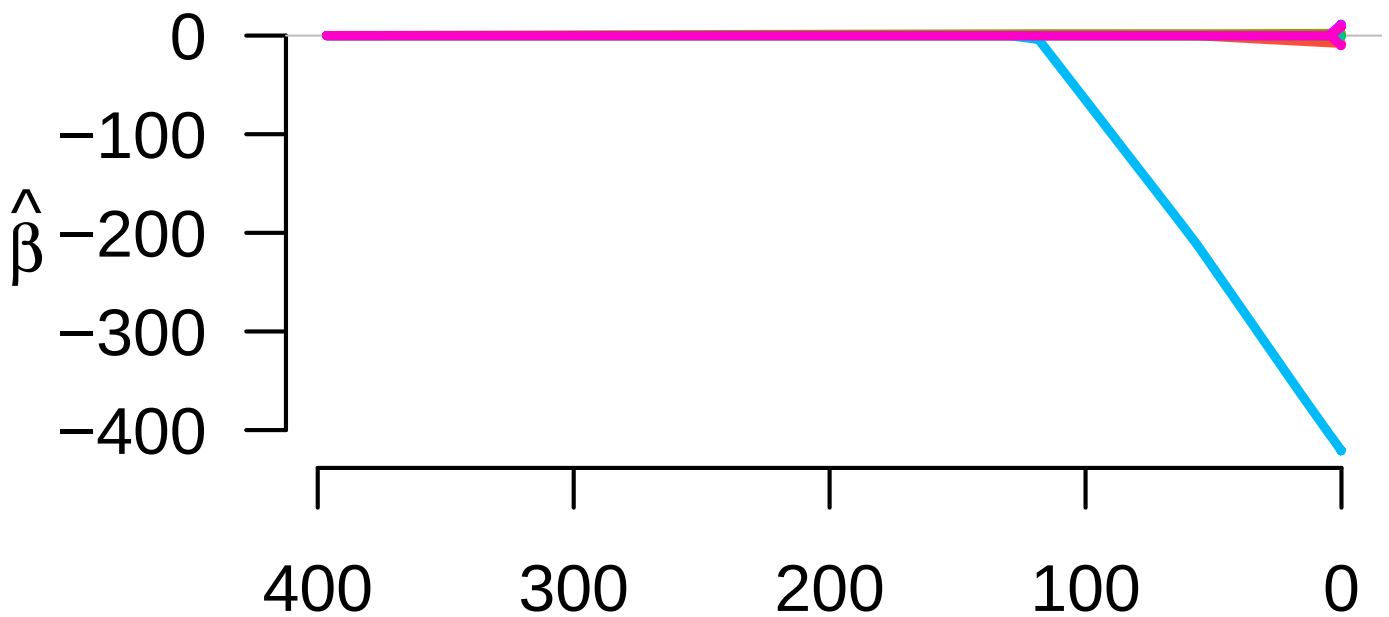
数据包含10个变量:, 分别是标识、收入（单位是千美元）、信用额度、信用级别（连续变量）、信用卡数量、年龄、受教育年限、性别、学生、婚姻状况、种族（白种人、非裔美国人、亚洲人）、个人平均信用卡债务。我们目的是分析哪些因素影响。样本量为。

```
data <- read.csv("D://Credit.csv",header=T)
attach(data)
data <- data[-1]
#转化为虚拟变量
library(nnet)
stu <- class.ind(data$Student)
head(stu,1)

##       No Yes
## [1,]   1   0

Student1 <- stu[,1]
```

```
gen <- class.ind(data$Gender)
head(gen,1)

##        Male Female
## [1,]      1      0

Gender1 <- gen[,1]
mar <- class.ind(data$Married)
head(mar,1)

##       No Yes
## [1,]  0   1

Married1 <- mar[,1]
```

```
eth <- class.ind(data$Ethnicity)
head(eth,1)

##      African American Asian Caucasian
## [1,]                0     0         1

Ethnicity1 <- eth[,1]
Ethnicity2 <- eth[,2]
x <- cbind( Income,Limit,Rating,Age,Education ,
            Student1,Gender1, Married1,
            Ethnicity1,Ethnicity2)
y <- Balance
```

```
library(grpreg)
group<- c(1,2,3,4,5,6,7,8,9,9)
fit <- grpreg(X=x, y=y, penalty="grLasso",group=group)
plot(fit)
```

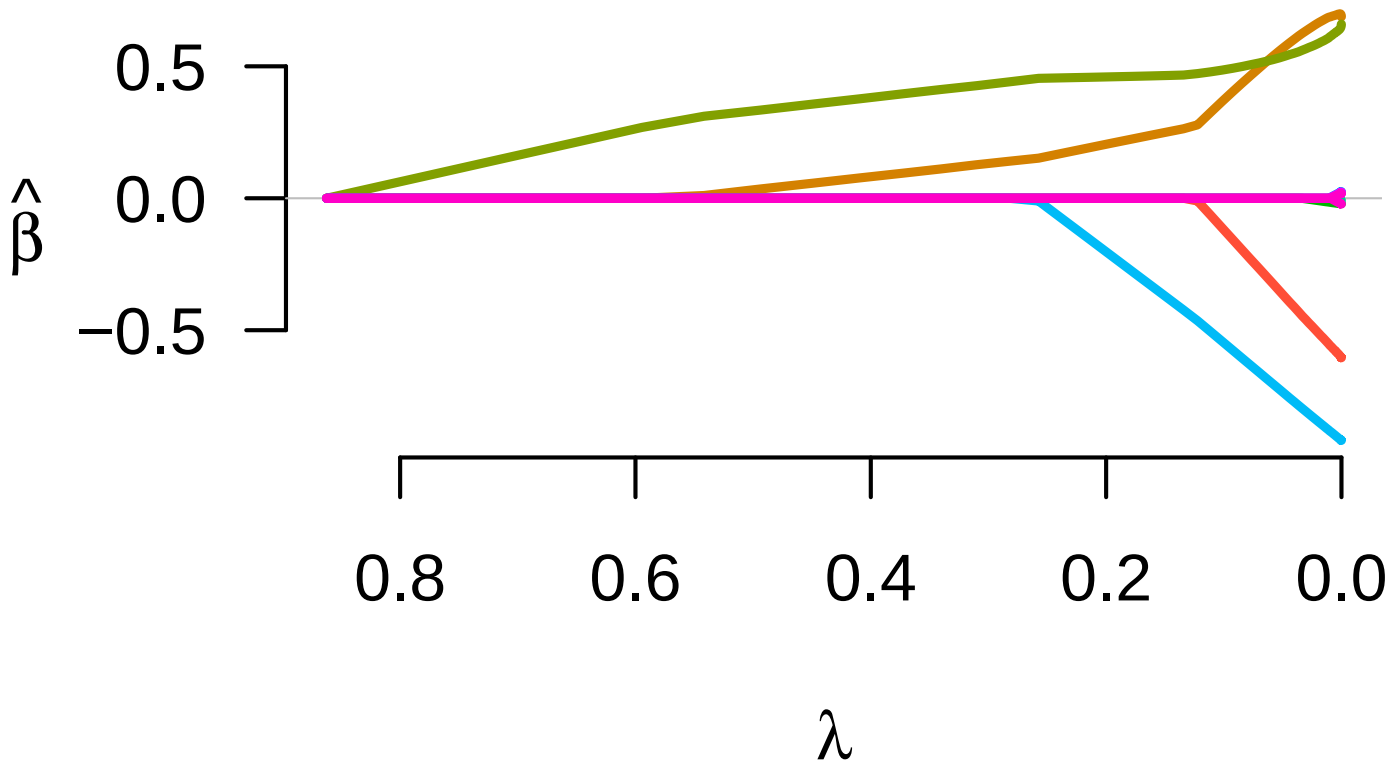


```
select(fit,"AIC")$beta
```

```
##  (Intercept)       Income        Limit       Rating          Age
##  -68.5254909   -7.6335419    0.1157769    2.2183061   -0.5088343
##    Education     Student1      Gender1     Married1   Ethnicity1
##   -0.4923560 -414.5777552    7.3020732    7.3144774   -6.2052974
##   Ethnicity2
##    5.2081229
```

```
Income1 <- (Income-mean(Income))/sd(Income)
Limit1 <- (Limit - mean(Limit))/sd (Limit)
Rating1 <- (Rating-mean(Rating))/sd(Rating)
Age1 <- (Age-mean(Age))/sd(Age)
Education1 <- (Education - mean(Education))/sd(Education)
xx <- cbind( Income1, Limit1, Rating1, Age1, Education1,
             Student1, Gender1,Married1, Ethnicity1, Ethnicity2)
yy <- scale(Balance)
```

```
fit1 <- grpreg(X=xx, y=yy, penalty="grLasso",group=group)
plot(fit1)
```



```
select(fit1,"AIC")$beta
```

```
##  (Intercept)      Income1      Limit1      Rating1         Age1
##  0.800120501 -0.583131459  0.695292789  0.630705030 -0.018604919
##   Education1     Student1     Gender1     Married1   Ethnicity1
## -0.003769765 -0.902332411  0.015167510  0.013461518 -0.012285683
##   Ethnicity2
##  0.009743976
```

```
select(fit,"AIC")$beta
```

```
##  (Intercept)       Income        Limit       Rating          Age
##  -68.5254909   -7.6335419    0.1157769    2.2183061   -0.5088343
##    Education     Student1     Gender1     Married1   Ethnicity1
##   -0.4923560 -414.5777552    7.3020732    7.3144774   -6.2052974
##   Ethnicity2
##    5.2081229
```

参考文献

1. Breheny P., Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing, 2015, 25(2): 173–187.
2. Breiman L. Better subset selection using nonnegative garrote. Techonometrics, 1995, 37(4): 373-384.
3. Candes E., Tao T. The dantzig selector statistical estimation when p is much larger than n. The Annals of Statistics, 2007, 35(6): 2313-2351.
4. Fan J., Li R.. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.
5. Tibshirani R., Saunders M., Rosset S., Zhu J., Knight K. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Ser. B, 2005, 67(1): 91-108.

1. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Ser. B, 1996, 58(1): 267-288.
2. Zhang C. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 2010, 38(2): 894-942.
3. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 2006 ,101(476): 1416-1429.
4. Zou H., Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Ser. B, 2005, 67(2): 301-320.

谢谢大家！