机器学习 第3周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

多元线性回归的最小二乘解（无偏估计）

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{2n} & \cdots & x_{np} \end{bmatrix},\ \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

则多元线性模型 (6.19) 可表示为

$$Y=X\beta+\varepsilon, \tag{6.20}$$

类似于一元线性回归，求参数 β 的估计值 $\hat{\beta}$, 就是求最小二乘函数

$$Q(\beta)=(y-X\beta)^T(y-X\beta), \tag{6.21}$$

达到最小的 β 值.

可以证明 β 的最小二乘估计

$$\hat{\beta}=\left(X^TX\right)^{-1}X^Ty. \tag{6.22}$$

广义逆的奇异性

$$B = X^{+}Y = (X^{T}X)^{-1}X^{T}Y$$

X^{+}表示X的广义逆（或叫伪逆）。

- 当变量比样本多时，出现奇异性
- 当出现多重共线性时，出现奇异性

例子

- 假设已知x_1，x_2与y的关系服从线性回归型 $y=10+2x_1+3x_2+\varepsilon$

给定 x_1，x_2 的 10 个值，如下表 7.1 的第（2）、（3）两行：

表 7.1

	序号	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

例子

现在我们假设回归系数与误差项是未知的，用普通最小二乘法求回归系数的估计值得：

$$\hat{\beta}_0 = 11.292,\quad \hat{\beta}_1 = 11.307,\quad \hat{\beta}_2 = -6.591$$

而原模型的参数潼

$$\beta_0 = 10, \beta_1 = 2, \beta_2 = 3$$

看来相差太大。计算 x_1，x_2 的样本相关系数得 $r_{12} = 0.986$，表明 x_1 与 x_2 之间高度相关。

岭回归(Ridge Regression，RR)

- 1962年由Heer首先提出，1970年后他与肯纳德合作进一步发展了该方法
- 先对数据做标准化，为了记号方便，标准化后的学习集仍然用X表示
- 我们称

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

为β的岭回归估计，其中k称为岭参数。

性质

- 当自变量间存在复共线性时，$|X'X| \approx 0$，我们设想给$X'X$加上一个正常数矩阵kI，（$k > 0$），那么$X'X+kI$接近奇异的程度就会比$X'X$接近奇异的程度小得多。
- 岭回归做为β的估计应比最小二乘估计稳定，当$k=0$时的岭回归估计就是普通的最小二乘估计。

等价模型：惩罚函数

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3.41}$$

$$\begin{aligned} \hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} & \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ \text{subject to } & \sum_{j=1}^{p} \beta_j^2 \le t, \end{aligned} \tag{3.42}$$

估计族

- 当岭参数为0，得到最小二乘解
- 当岭参数趋向更大时，岭回归系数估计趋向于0

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 β 的一个估计族。

例如对例 7.1 可以算得不同 k 值时的 $\hat{\beta}_1(k)$， $\hat{\beta}_2(k)$，见表 7.2

表7.2

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

岭迹图

- 当不存在奇异性时，岭迹应是稳定地逐渐趋向于0
- 通过岭迹图观察岭估计的情况，可以判断出应该剔除哪些变量

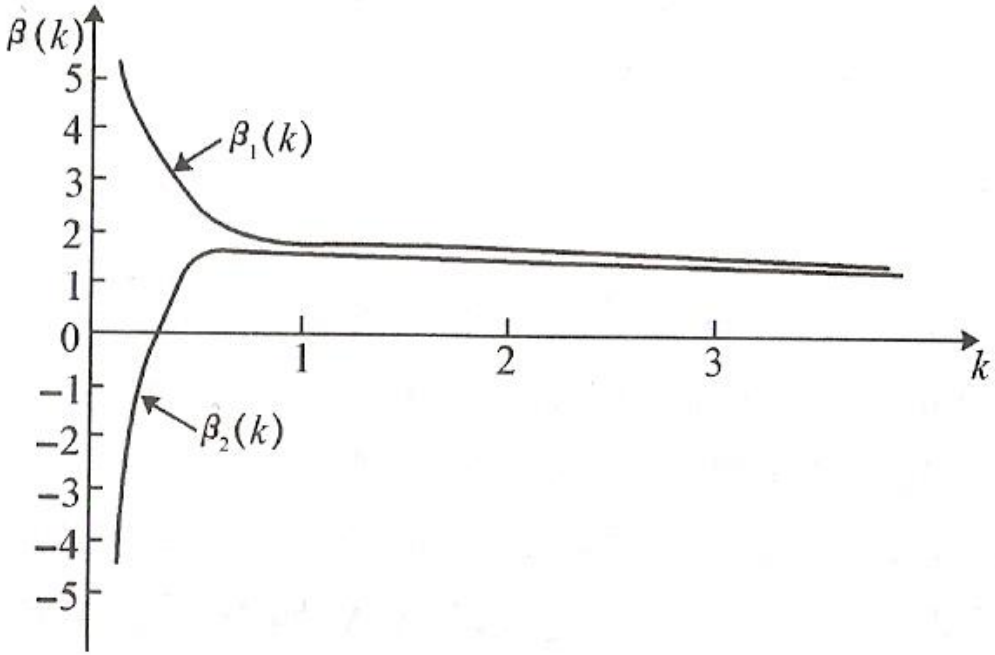


图 7.1

岭回归估计的性质

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

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

显然只有当 $k=0$ 时，$E[\hat{\boldsymbol{\beta}}(0)]=\beta$；当 $k\neq 0$ 时，$\hat{\boldsymbol{\beta}}(k)$ 是 β 的有偏估计。

要特别强调的是 $\hat{\boldsymbol{\beta}}(k)$ 不再是 β 的无偏估计了，

有偏性是岭回归估计的一个重要特性。

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。

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

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

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。

这里需要注意的是，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。

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

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

即

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

岭迹分析

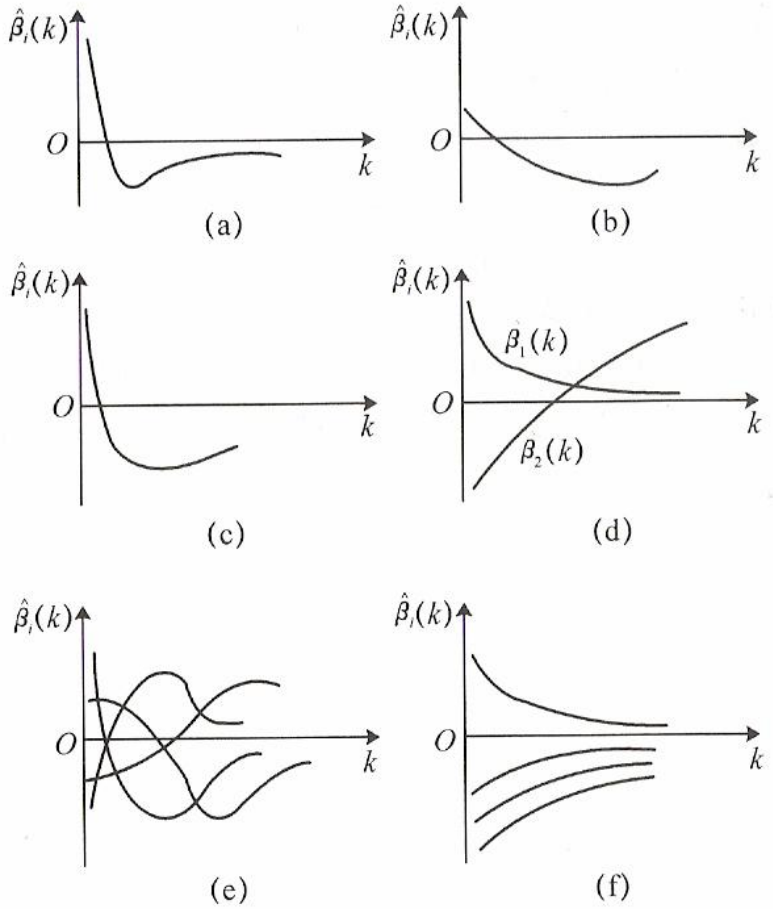


图 7.2

岭参数的一般选择原则

- 选择k（或 lambda）值，使到

（1）各回归系数的岭估计基本稳定；

（2）用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理；

（3）回归系数没有不合乎实际意义的绝对值；

（4）残差平方和增大不太多。

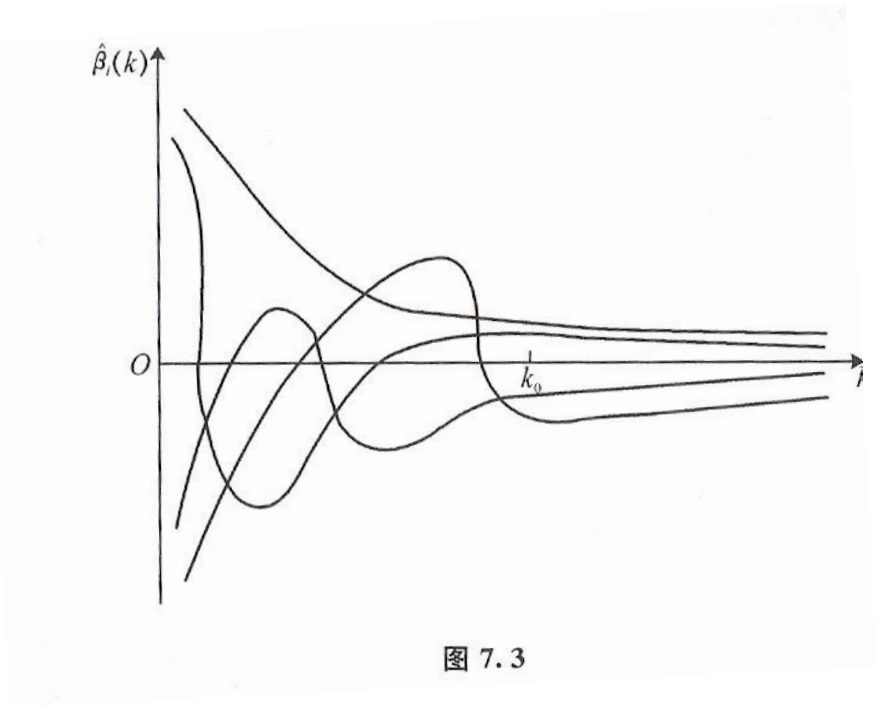


图 7.3

方差扩大因子法

方差扩大因子c_{jj}度量了多重共线性的严重程度，计算岭估计$\hat{\boldsymbol{\beta}}(k)$的协方差阵，得

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

式中矩阵$C_{jj}(k)$的对角元$c_{jj}(k)$就是岭估计的方差扩大因子。不难看出，$c_{jj}(k)$随着k的增大而减少。

选择k使所有方差扩大因子$c_{jj}(k) \leqslant 10$。

用岭回归选择变量

- 岭回归选择变量的原则：

（1）在岭回归中设计矩阵X已经中心化和标准化了，这样可以直接比较标准化岭回归系数的大小。可以剔除掉标准化岭回归系数比较稳定且绝对值很小的自变量。

（2）随着k的增加，回归系数不稳定，震动趋于零的自变量也可以剔除。

（3）如果依照上述去掉变量的原则，有若干个回归系数不稳定，究竟去掉几个，去掉哪几个，这并无一般原则可循，这需根据去掉某个变量后重新进行岭回归分析的效果来确定。

例子

空气污染问题。Mcdonald和Schwing曾研究死亡率与空气污染、气候以及社会经济状况等因素的关系。考虑了15个解释变量，收集了60组样本数据。

x1—Average annual precipitation in inches 平均年降雨量

x2—Average January temperature in degrees F 1月份平均气温

x3—Same for July 7月份平均气温

x4—Percent of 1960 SMSA population aged 65 or older

年龄65岁以上的人口占总人口的百分比

x5—Average household size 每家人口数

x6—Median school years completed by those over 22

年龄在22岁以上的人受教育年限的中位数

x7—Percent of housing units which are sound & with all facilities

住房符合标准的家庭比例数

例子

x8—Population per sq. mile in urbanized areas, 1960 每平方公里人口数

x9—Percent non-white population in urbanized areas,

1960 非白种人占总人口的比例

x10—Percent employed in white collar occupations 白领阶层人口比例

x11—Percent of families with income < $3000

收入在3000美元以下的家庭比例

x12—Relative hydrocarbon pollution potential 碳氢化合物的相对污染势

x13— Same for nitric oxides 氮氧化合物的相对污染势

x14—Same for sulphur dioxide 二氧化硫的相对污染势

x15—Annual average % relative humidity at 1pm 年平均相对湿度

y—Total age-adjusted mortality rate per 100,000

每十万人中的死亡人数

岭迹图

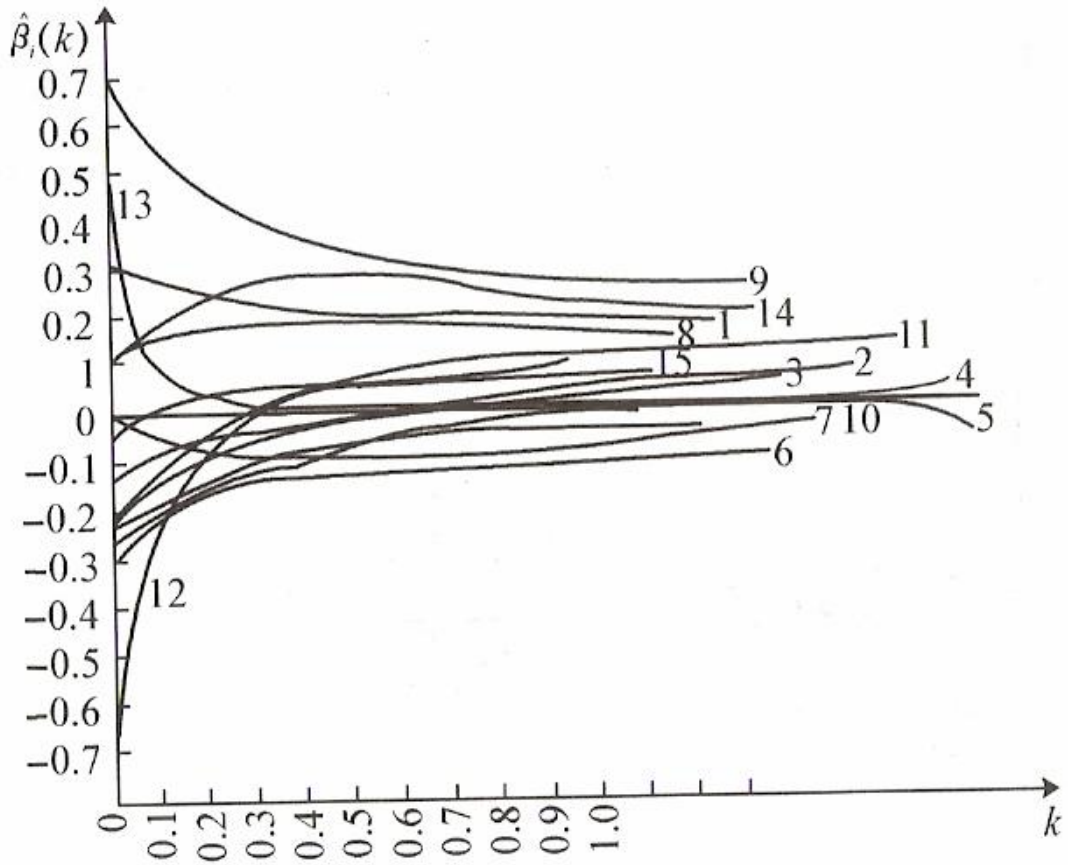

$\hat{\beta}_i(k)$
0.7
0.6
0.5
0.4
0.3
0.2
1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
13
9
14
8
1
11
15
3
2
4
7
10
5
6
12
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
k

图 7.4

岭迹分析

- 把15个回归系数的岭迹画到图中，我们可看到，当k=0.20时岭迹大体上达到稳定。按照岭迹法，应取k=0.2。
- 若用方差扩大因子法，因k=0.18时，方差扩大因子接近于1，当k在0.02～0.08时，方差扩大因子小于10，故应建议在此范围选取k。 由此也看到不同的方法选取k值是不同的。

选择变量

- 在用岭回归进行变量选择时，因为从岭迹看到自变量x4,x7,x10,x11和x15有较稳定且绝对值比较小的岭回归系数，根据变量选择的第一条原则，这些自变量可以去掉。
- 又因为自变量x12和x13的岭回归系数很不稳定，且随着k的增加很快趋于零，根据上面的第二条原则这些自变量也应该去掉。
- 再根据第三条原则去掉变量x3和x5。
- 这个问题最后剩的变量是x1，x2，x6，x8，x9，x14。

用R语言进行岭回归

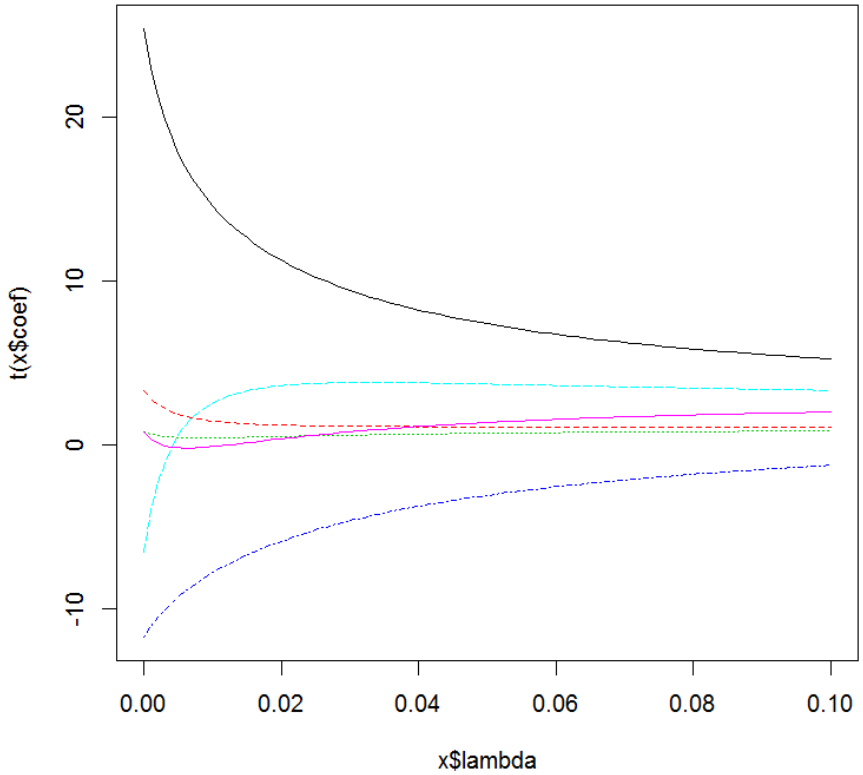
```
> library(MASS)
> longley
          y     GNP Unemployed Armed.Forces Population Year Employed
1947   83.0 234.289      235.6        159.0    107.608 1947   60.323
1948   88.5 259.426      232.5        145.6    108.632 1948   61.122
1949   88.2 258.054      368.2        161.6    109.773 1949   60.171
1950   89.5 284.599      335.1        165.0    110.929 1950   61.187
1951   96.2 328.975      209.9        309.9    112.075 1951   63.221
1952   98.1 346.999      193.2        359.4    113.270 1952   63.639
1953   99.0 365.385      187.0        354.7    115.094 1953   64.989
1954  100.0 363.112      357.8        335.0    116.219 1954   63.761
1955  101.2 397.469      290.4        304.8    117.388 1955   66.019
1956  104.6 419.180      282.2        285.7    118.734 1956   67.857
1957  108.4 442.769      293.6        279.8    120.445 1957   68.169
1958  110.8 444.546      468.1        263.7    121.950 1958   66.513
1959  112.6 482.704      381.3        255.2    123.366 1959   68.655
1960  114.2 502.601      393.1        251.4    125.368 1960   69.564
1961  115.7 518.173      480.6        257.2    127.852 1961   69.331
1962  116.9 554.894      400.7        282.7    130.081 1962   70.551
> 
```

多元线性回归的最小二乘估计

```
> summary(fm1 <- lm(Employed ~ ., data = longley))

Call:
lm(formula = Employed ~ ., data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41011 -0.15767 -0.02816  0.10155  0.45539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.482e+03  8.904e+02  -3.911 0.003560 **
y            1.506e-02  8.492e-02   0.177 0.863141
GNP         -3.582e-02  3.349e-02  -1.070 0.312681
Unemployed  -2.020e-02  4.884e-03  -4.136 0.002535 **
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population  -5.110e-02  2.261e-01  -0.226 0.826212
Year         1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

岭回归

```
names(longley)[1] <- "y "
lm.ridge(y ~ ., longley)
plot(lm.ridge(y ~ ., longley,
     lambda = seq(0,0.1,0.001)))
```




```
>
> names(longley)[1] <- "y"
> lm.ridge(y ~ ., longley)
                      GNP    Unemployed  Armed.Forces    Population          Year      Employed
2946.85636017    0.26352725    0.03648291    0.01116105   -1.73702984   -1.41879853    0.23128785
> plot(lm.ridge(y ~ ., longley,
+              lambda = seq(0,0.1,0.001)))
>
```

岭回归

```
> lm.ridge(y ~ ., longley,lambda = seq(0,0.1,0.001))
                       GNP Unemployed Armed.Forces Population        Year     Employed
0.000  2946.85636 0.26352725 0.03648291  0.011161050 -1.7370298 -1.41879853  0.231287851
0.001  1895.97527 0.23923480 0.03100610  0.009372158 -1.6438029 -0.87657471  0.105607249
0.002  1166.33337 0.22099519 0.02719073  0.008243201 -1.5650260 -0.50108472  0.030290543
0.003   635.78843 0.20661106 0.02440554  0.007514565 -1.4962459 -0.22885815 -0.014755698
0.004   236.65772 0.19485388 0.02230066  0.007043302 -1.4348862 -0.02473192 -0.040566288
0.005   -71.53274 0.18498058 0.02066688  0.006744636 -1.3793225  0.13231532 -0.053663187
0.006  -314.43247 0.17651367 0.01937157  0.006565392 -1.3284596  0.25560068 -0.058119371
0.007  -509.05648 0.16913115 0.01832674  0.006470736 -1.2815187  0.35395451 -0.056588923
0.008  -667.11647 0.16260718 0.01747181  0.006437042 -1.2379217  0.43345188 -0.050860281
0.009  -796.92303 0.15677808 0.01676376  0.006447832 -1.1972245  0.49840118 -0.042171311
0.010  -904.52578 0.15152189 0.01617130  0.006491346 -1.1590755  0.55193667 -0.031397510
0.011  -994.42507 0.14674556 0.01567111  0.006559030 -1.1231903  0.59638825 -0.019168982
0.012 -1070.03184 0.14237663 0.01524553  0.006644564 -1.0893337  0.63352047 -0.005945632
0.013 -1133.97358 0.13835766 0.01488094  0.006743215 -1.0573084  0.66469138  0.007933122
0.014 -1188.30330 0.13464236 0.01456670  0.006851400 -1.0269464  0.69096113  0.022214639
0.015 -1234.64543 0.13119296 0.01429437  0.006966383 -0.9981032  0.71316772  0.036709726
0.016 -1274.29970 0.12797821 0.01405722  0.007086059 -0.9706528  0.73198092  0.051276169
0.017 -1308.31654 0.12497200 0.01384979  0.007208799 -0.9444851  0.74794140  0.065806938
0.018 -1337.55256 0.12215228 0.01366762  0.007333338 -0.9195024  0.76148954  0.080221587
0.019 -1362.71206 0.11950026 0.01350706  0.007458691 -0.8956181  0.77298695  0.094459929
0.020 -1384.37837 0.11699982 0.01336506  0.007584089 -0.8727546  0.78273271  0.108477319
0.021 -1403.03795 0.11463698 0.01323909  0.007708933 -0.8508422  0.79097588  0.122241095
0.022 -1419.09894 0.11239961 0.01312702  0.007832759 -0.8298181  0.79792511  0.135727883
0.023 -1432.90579 0.11027705 0.01302706  0.007955206 -0.8096251  0.80375619  0.148921524
0.024 -1444.75065 0.10825992 0.01293768  0.008075999 -0.7902114  0.80861799  0.161811479
0.025 -1454.88257 0.10633991 0.01285758  0.008194928 -0.7715296  0.81263718  0.174391594
0.026 -1463.51479 0.10450963 0.01278563  0.008311837 -0.7535365  0.81592202  0.186659124
0.027 -1470.83064 0.10276246 0.01272088  0.008426614 -0.7361922  0.81856540  0.198613980
```

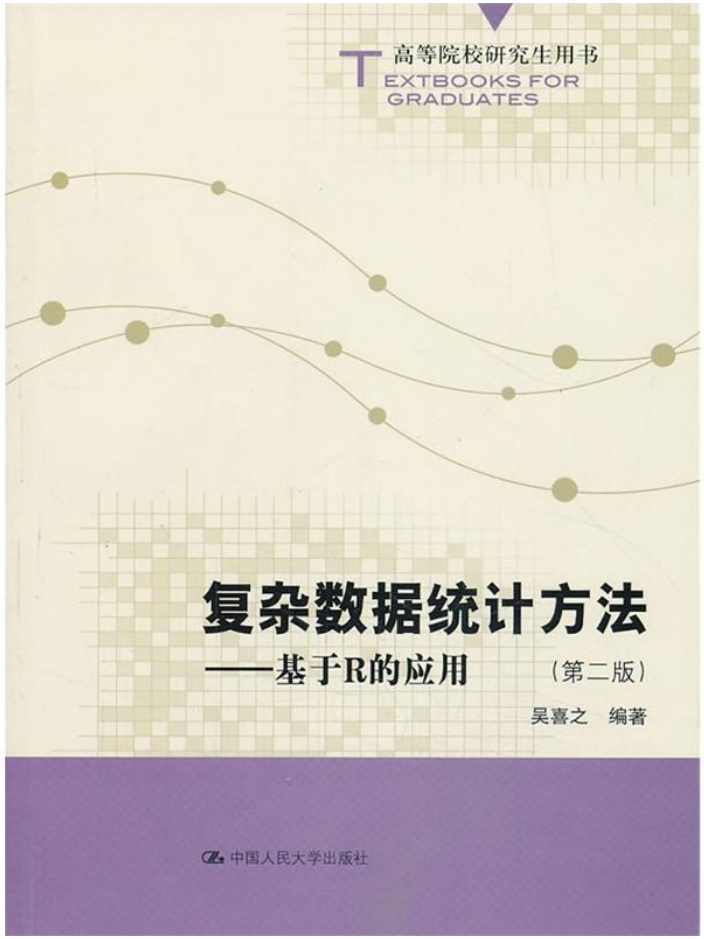
岭回归

```
> select(lm.ridge(y ~ ., longley,lambda = seq(0,0.1,0.001)))
modified HKB estimator is 0.006836982
modified L-W estimator is 0.05267247
smallest value of GCV  at 0.006
```

R的ridge包

```
>
> library(ridge)
> a=linearRidge(GNP.deflator~.,data=longley)
> summary(a)

Call:
linearRidge(formula = GNP.deflator ~ ., data = longley)


Coefficients:
              Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept) -1.247e+03              NA                  NA               NA       NA
GNP          4.338e-02       1.670e+01           3.689e+00            4.526  6.0e-06 ***
Unemployed   1.184e-02       4.286e+00           2.507e+00            1.710   0.0873 .
Armed.Forces 1.381e-02       3.721e+00           1.905e+00            1.953   0.0508 .
Population  -2.831e-02      -7.627e-01           5.285e+00            0.144   0.8853
Year         6.566e-01       1.211e+01           2.691e+00            4.500  6.8e-06 ***
Employed     6.745e-01       9.175e+00           4.996e+00            1.836   0.0663 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.01046912, chosen automatically, computed using 2 PCs

Degrees of freedom: model 3.67 , variance 3.218 , residual 4.123

> |
```

参考书


高等院校研究生用书
EXTBOOKS FOR GRADUATES
复杂数据统计方法
——基于R的应用
（第二版）
吴喜之 编著
中国人民大学出版社

岭回归的问题

- 岭参数计算方法太多，差异太大
- 根据岭迹图进行变量筛选，随意性太大
- 岭回归返回的模型（如果没有经过变量筛选）包含所有的变量

LASSO

- Tibshirani(1996)提出了Lasso(The Least Absolute Shrinkage and Selectionator operator)算法
- 通过构造一个一阶惩罚函数获得一个精炼的模型；通过最终确定一些指标（变量）的系数为零（岭回归估计系数等于0的机会微乎其微，造成筛选变量困难），解释力很强
- 擅长处理具有多重共线性的数据，与岭回归一样是有偏估计

LASSO vs 岭回归

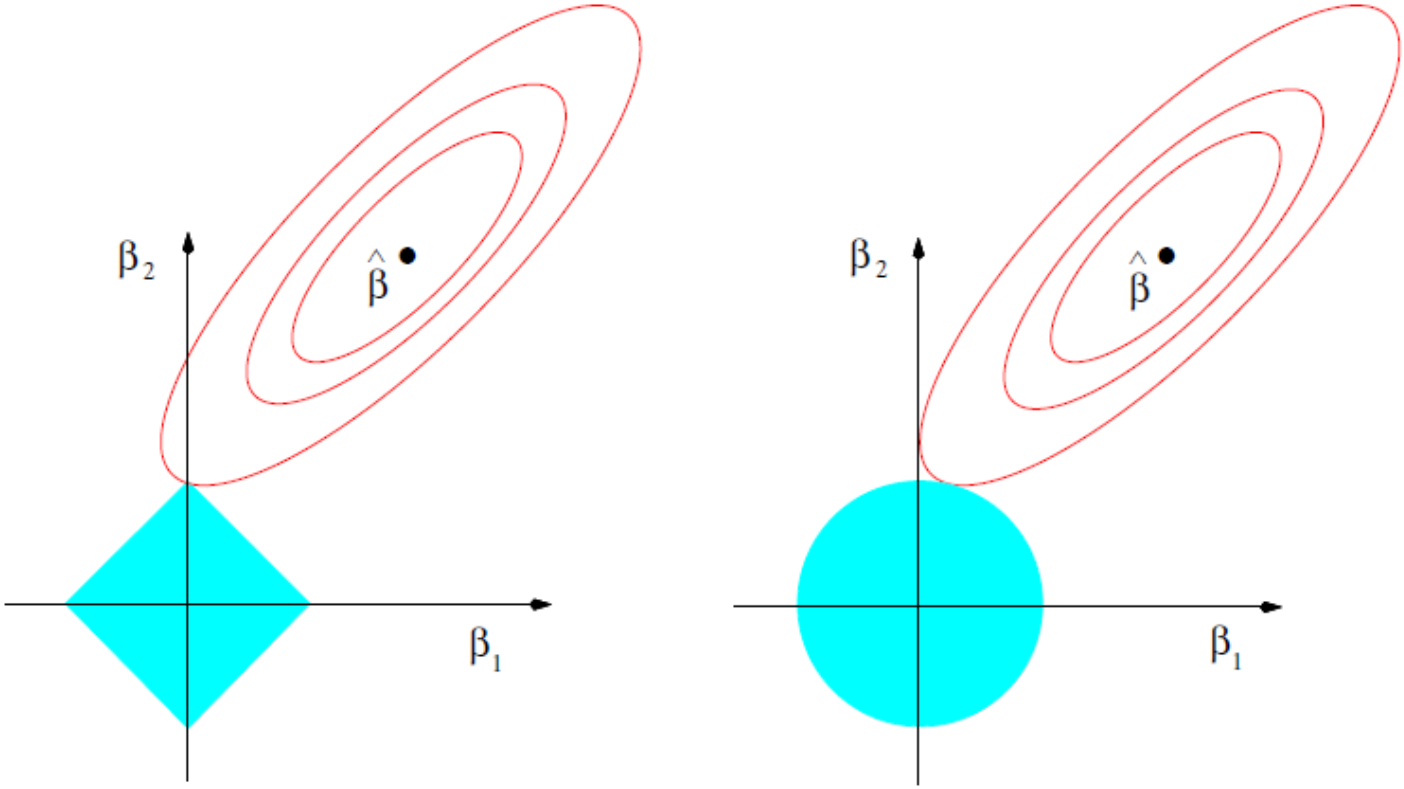
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{i=1}^{p} \beta_j^2 \right\}. \tag{3.41}$$

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ &\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t, \end{aligned} \tag{3.42}$$

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.52}$$

$$\begin{aligned} \hat{\beta}^{\text{lasso}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t. \end{aligned} \tag{3.51}$$

为什么LASSO能直接筛选变量



LASSO vs 岭回归

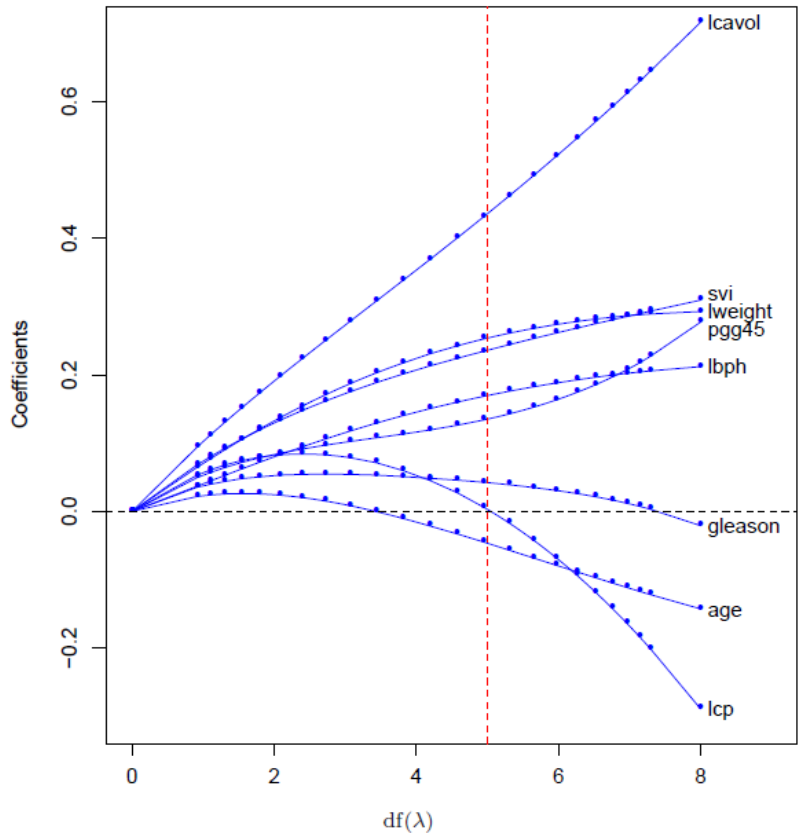


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

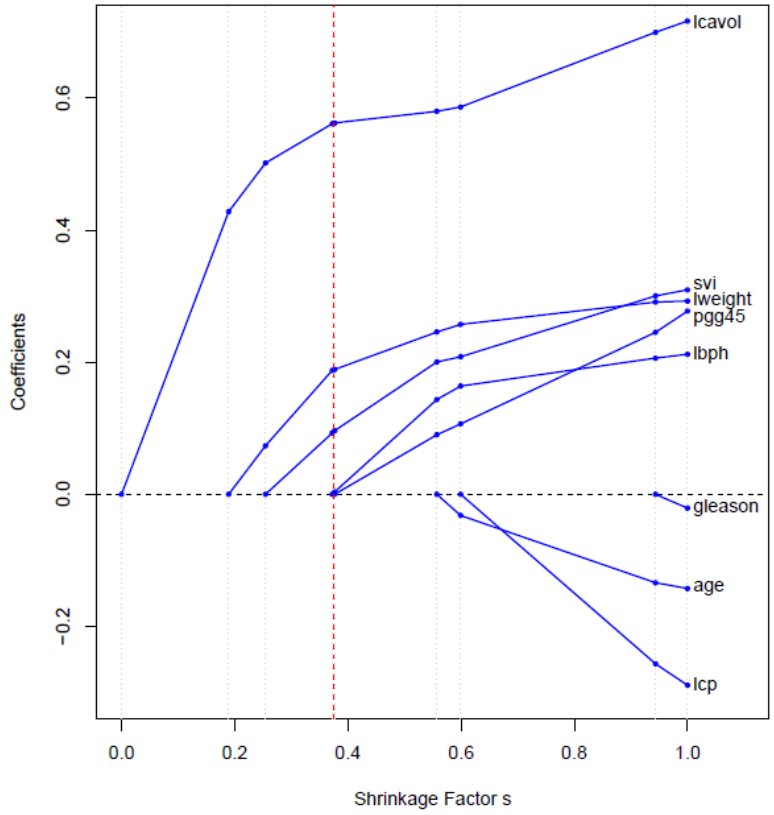


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

更一般化的模型

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \qquad (3.53)$$

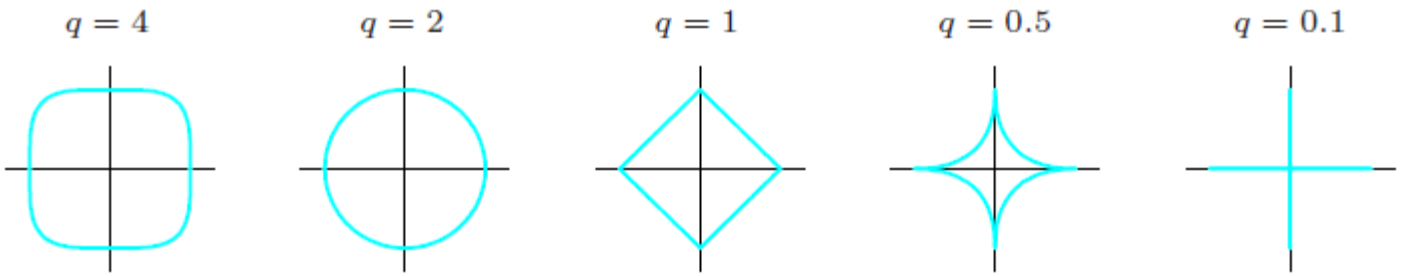
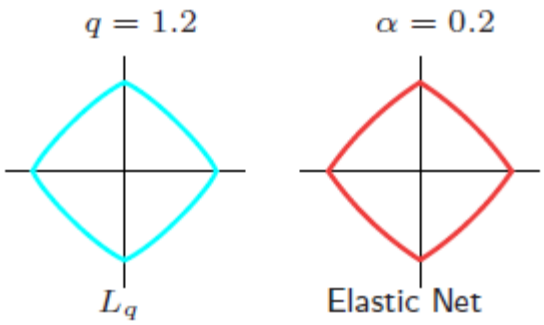


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

弹性网

- Zou and Hastie (2005)提出*elasticnet*

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right), \qquad (3.54)$$



LAR（最小角回归）

- Least Angel Regression
- Efron于2004年提出的一种变量选择的方法，类似于向前逐步回归(Forward Stepwise)的形式。
- 是lasso regression的一种高效解法。
- 向前逐步回归(Forward Stepwise)不同点在于，Forward Stepwise每次都是根据选择的变量子集，完全拟合出线性模型，计算出RSS，再设计统计量（如AIC）对较高的模型复杂度作出惩罚，而LAR是每次先找出和因变量相关度最高的那个变量，再沿着LSE的方向一点点调整这个predictor的系数，在这个过程中，这个变量和残差的相关系数会逐渐减小，等到这个相关性没那么显著的时候，就要选进新的相关性最高的变量，然后重新沿着LSE的方向进行变动。而到最后，所有变量都被选中，就和LSE相同了。

算法步骤

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

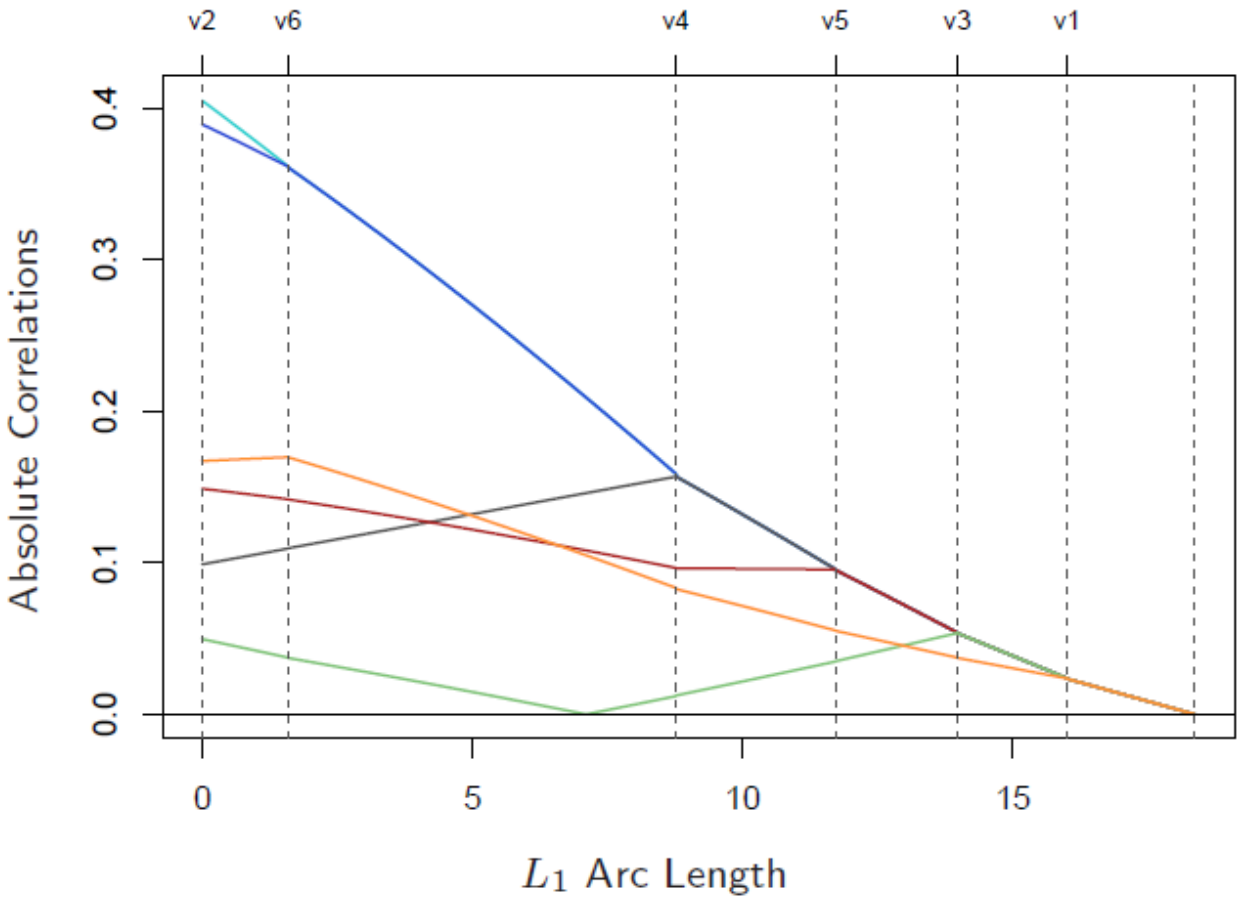
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.

3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.

LAR过程图解


v2
v6
v4
v5
v3
v1
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length

LAR vs LASSO

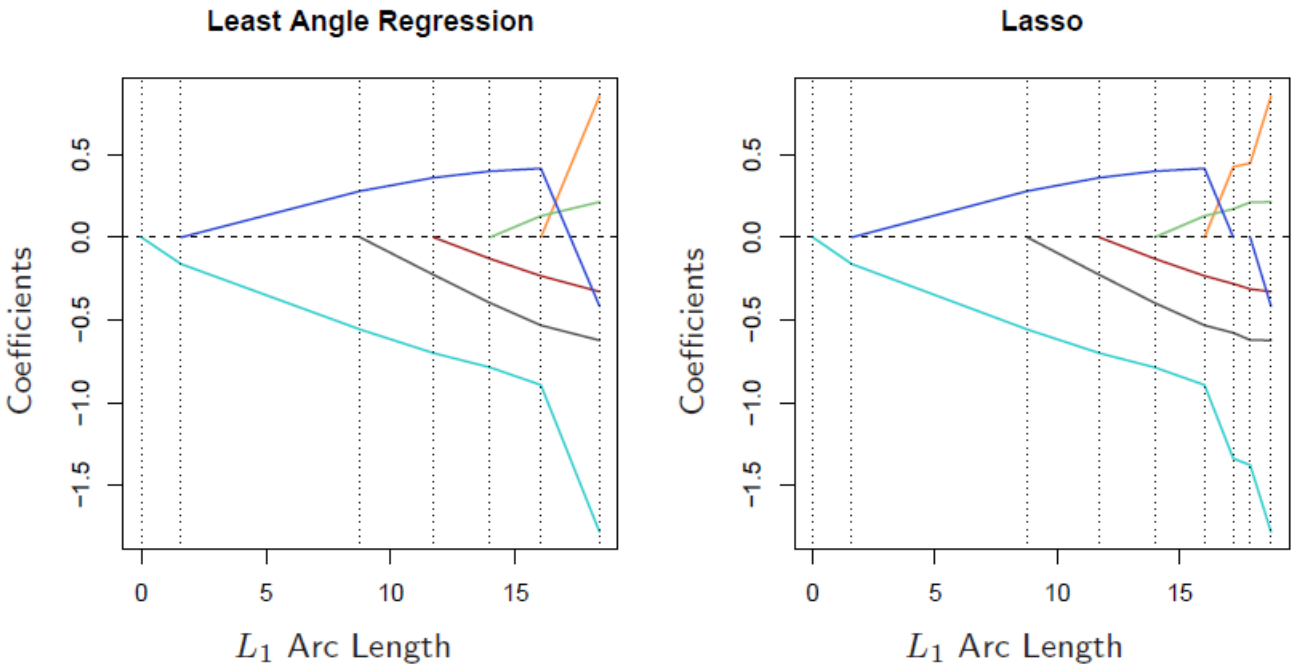


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

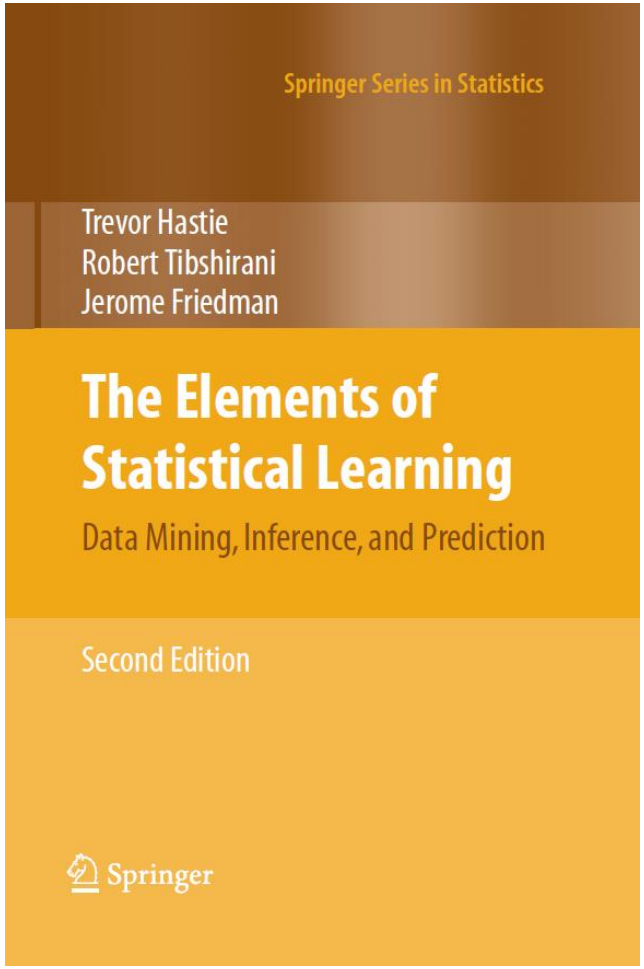
修正的LAR

Algorithm 3.2a *Least Angle Regression: Lasso Modification.*

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

扩展阅读

- LAR的几何意义（第74页）
- 为什么LAR的过程与LASSO过程高度相似（第76页）



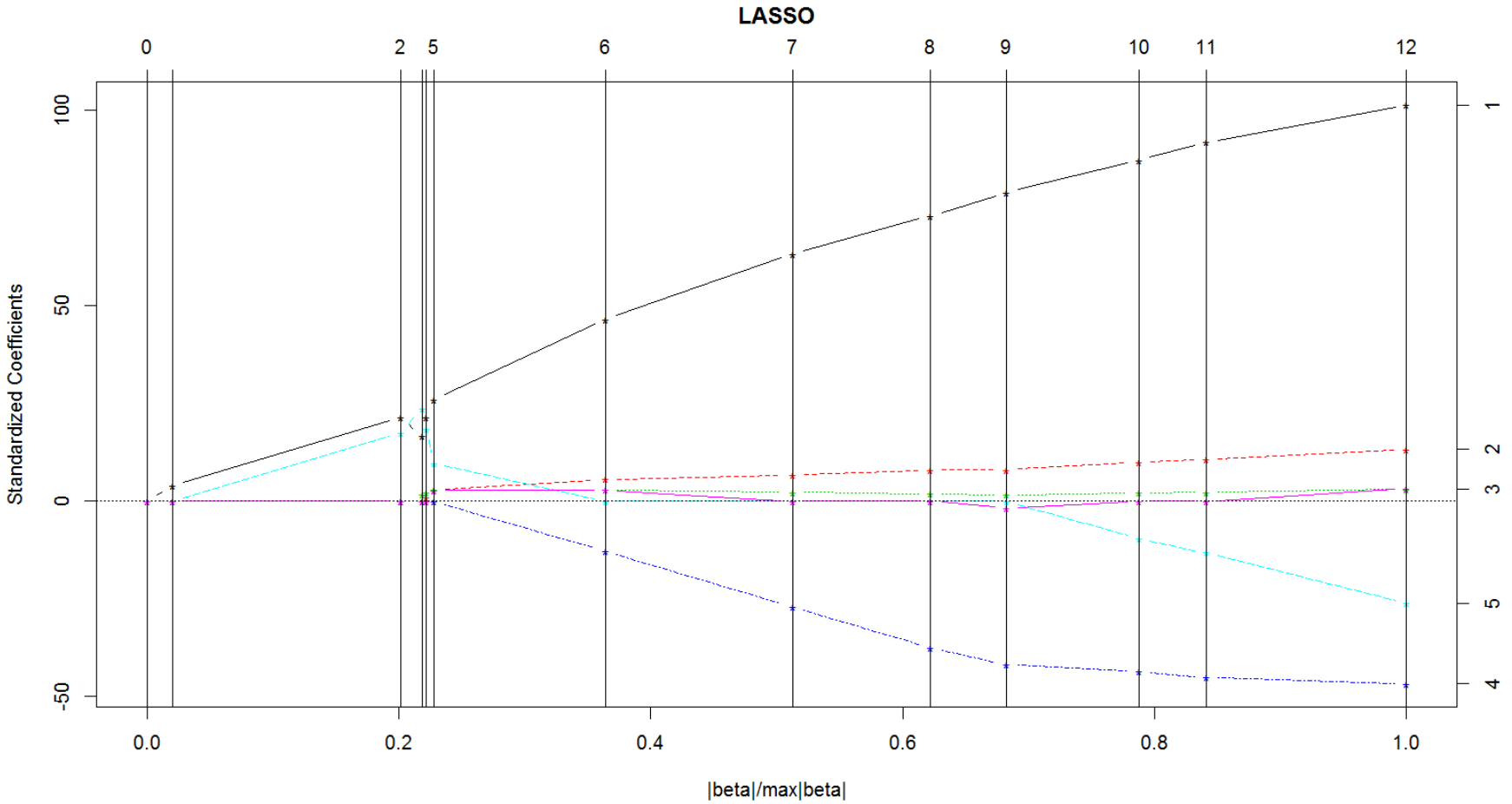
R的lars包

```
> longley
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
1951         96.2 328.975      209.9        309.9    112.075 1951   63.221
1952         98.1 346.999      193.2        359.4    113.270 1952   63.639
1953         99.0 365.385      187.0        354.7    115.094 1953   64.989
1954        100.0 363.112      357.8        335.0    116.219 1954   63.761
1955        101.2 397.469      290.4        304.8    117.388 1955   66.019
1956        104.6 419.180      282.2        285.7    118.734 1956   67.857
1957        108.4 442.769      293.6        279.8    120.445 1957   68.169
1958        110.8 444.546      468.1        263.7    121.950 1958   66.513
1959        112.6 482.704      381.3        255.2    123.366 1959   68.655
1960        114.2 502.601      393.1        251.4    125.368 1960   69.564
1961        115.7 518.173      480.6        257.2    127.852 1961   69.331
1962        116.9 554.894      400.7        282.7    130.081 1962   70.551
> w=as.matrix(longley)
```

R的lars包

```
> laa=lars(w[,2:7],w[,1])
> laa

Call:
lars(x = w[, 2:7], y = w[, 1])
R-squared: 0.993
Sequence of LASSO moves:
     GNP Year Armed.Forces Unemployed Employed Population Year Employed Employed Year Employed Employed
Var    1    5            3          2        6          4   -5       -6        6    5       -6        6
Step   1    2            3          4        5          6    7        8        9   10       11       12
> plot(laa)
```

R的lars包


LASSO
0
2
5
6
7
8
9
10
11
12
100
50
0
-50
Standardized Coefficients
1
2
3
5
4
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|

R的lars包

```
> summary(laa)
LARS/LASSO
Call: lars(x = w[, 2:7], y = w[, 1])
   Df     Rss         Cp
0   1 1746.86 1210.0561
1   2 1439.51  996.6871
2   3   32.31   12.6400
3   4   23.18    8.2425
4   5   22.91   10.0505
5   6   22.63   11.8595
6   7   18.04   10.6409
7   6   14.74    6.3262
8   5   13.54    3.4848
9   6   13.27    5.2974
10  7   13.01    7.1189
11  6   12.93    5.0624
12  7   12.84    7.0000
```

Mallows's C_p

- http://en.wikipedia.org/wiki/Mallows%27_Cp

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

where

- $SSE_p = \sum_{i=1}^{N}(Y_i - Y_{pi})^2$ is the error sum of squares[*disambiguation needed*] for the model with *P* regressors,
- Y_{pi} is the predicted value of the *i*th observation of *Y* from the *P* regressors,
- S^2 is the residual mean square after regression on the complete set of *K* regressors and can be estimated by mean square error *MSE*,
- and *N* is the sample size.

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间