

第7章 模型选择与正则化

李高荣

北京师范大学统计学院

E-mail: ligaorong@bnu.edu.cn



本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业



- 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

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

■ 考虑如下的多元线性回归模型

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

♠ **问题:** 如何从协变量集 $\{X_1, X_2, \dots, X_p\}$ 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？

■ 考虑如下的多元线性回归模型

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

♠ **问题:** 如何从协变量集 $\{X_1, X_2, \dots, X_p\}$ 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 最优模型选择

最优子集选择(best subset selection)

对 p 个协变量(或预测变量) X_1, X_2, \dots, X_p 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;
- 对含有 p 个协变量的模型, 需拟合 $C_p^p = 1$ 个模型.

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 M_0 表示, 该步只用于估计各观测的样本均值.

步骤2: 对于 $k = 1, 2, \dots, p$:

- 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 M_k .

步骤3: 根据交叉(CV)测试预测误差、 C_p 、AIC、BIC 或者调整的判定系数 R_{adj}^2 从 M_1, \dots, M_p 个模型中选出一个最优子模型.

- 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 M_0 模型，因此，共有 2^p 个模型。例如，
 - ▶ 如果维数 $p = 10$ ，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。

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

■ 向前逐步选择算法:

步骤1: 记不含任何协变量的模型为 M_0 .

步骤2: 对于 $k = 0, 1, \dots, p - 1$:

- ① 从 $p - k$ 个模型中进行选择, 每个模型都在模型 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 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 M_0, \dots, M_p 个模型中选出一个最优模型.

■ 向后逐步选择算法:

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 $k = p, p - 1, \dots, 1$:

- 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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ 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

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

- ▶ 如果维数 $p = 10$, 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.

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

最优模型选择的方法有：

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

调整的判定系数 R_{adj}^2

■ 判定系数 R^2 : 定义为

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

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

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

调整的判定系数 R_{adj}^2

- 当模型中随着显著变量个数的逐渐增加，将使得 $\text{RSS}_d/(n - d - 1)$ 逐渐减小，导致 R_{adj}^2 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 $\text{RSS}_d/(n - d - 1)$ 的增大，从而降低 R_{adj}^2 的值；
- 理论上，拥有最大 R_{adj}^2 的模型只包含了正确的协变量，而没有噪声变量。

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

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

其中 RSS_d 是基于 d 个协变量的残差平方和, $\hat{\sigma}^2$ 是基于全模型 σ^2 的无偏估计.

- Mallow's C_p 有时也被定义为

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

■ 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_d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.

■ 统计学家已经提出了很多信息准则, 可以写成如下一般形式

$$IC(d, \lambda) = n \log \left(\frac{RSS_d}{n} \right) + \lambda d,$$

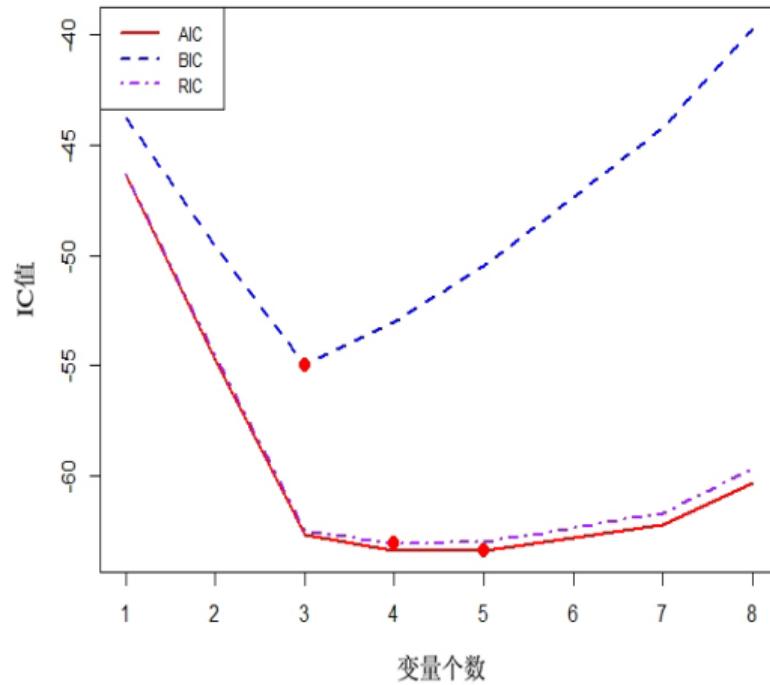
其中 $\lambda > 0$ 是调节参数.

■ 当调节参数 λ 取不同的值, $IC(d, \lambda)$ 将对应不同的信息准则. 经典的信息准则有:

- ① 当 $\lambda = 2$ 时, 则为 Akaike(1974) 提出的 AIC 方法;
- ② 当 $\lambda = \log(n)$ 时, 则为 Schwarz(1978) 提出的 BIC 方法;
- ③ 当 $\lambda = \log(p)$ 时, 则为 Foster 和 George(1994) 提出的 RIC 方法.

信息准则

- 理论上可证明, AIC方法等价于 C_p 准则方法;
- 对任意的 $n > 7$, 有 $\log(n) > 2$, 可知BIC方法通常给包含多个协变量的模型施加较重的惩罚, 与 C_p 和AIC相比, 得到的模型规模会更小;
- 对任意的 $7 < p < n$, 有 $2 < \log(p) < \log(n)$, 可知RIC方法得到的模型规模比AIC方法得到的模型小, 而比BIC方法得到的模型要大.



■ 信息准则 $IC(d, \lambda)$ 等价于 $RSS_d + \lambda^* d$. 当 $RSS_d/n \approx \sigma^2$ 时, 由 Taylor 展式, 有

$$\log\left(\frac{RSS_d}{n}\right) = \log\sigma^2 + \log\left(1 + \frac{RSS_d}{n\sigma^2} - 1\right) \approx \log\sigma^2 + \frac{RSS_d}{n\sigma^2} - 1.$$

- ① AIC 统计量定义为: $AIC = \frac{1}{n}(RSS_d + 2d\hat{\sigma}^2);$
- ② BIC 统计量定义为: $BIC = \frac{1}{n}(RSS_d + \log(n)d\hat{\sigma}^2);$
- ③ RIC 统计量定义为: $RIC = \frac{1}{n}(RSS_d + \log(p)d\hat{\sigma}^2).$

■ 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.

- 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
16	49.156	49	81.42	8.95	44	180	185								

```

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

            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

```

案例与R语言计算—31名中年男性的健康数据

```
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

案例与R语言计算—31名中年男性的健康数据

Step: AIC=55

$Y \sim X_1 + X_2 + X_3 + X_5 + X_6$

	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.

案例与R语言计算—31名中年男性的健康数据

```
> summary(lm.aic)      ## 输出AIC选择模型的估计结果
Residuals:
    Min          1Q      Median          3Q          Max
-5.4714   -0.9249   -0.0131    0.9594    5.4054

                               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
```

案例与R语言计算—31名中年男性的健康数据

```
lm.bic = step(lm.reg, k=log(length(health.data$Y))), trace=FALSE)
> summary(lm.bic)      ## 输出BIC选择模型的估计结果
Residuals:
    Min         1Q     Median        3Q       Max
-4.9590 -1.2603 -0.1512  1.1796  4.8132
                               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

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

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

案例与R语言计算—31名中年男性的健康数据

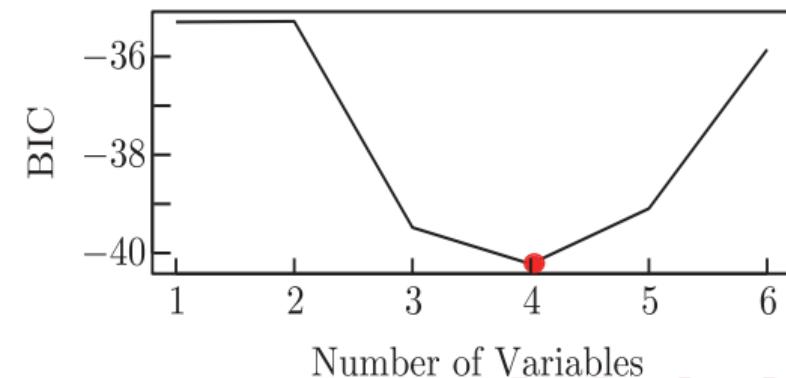
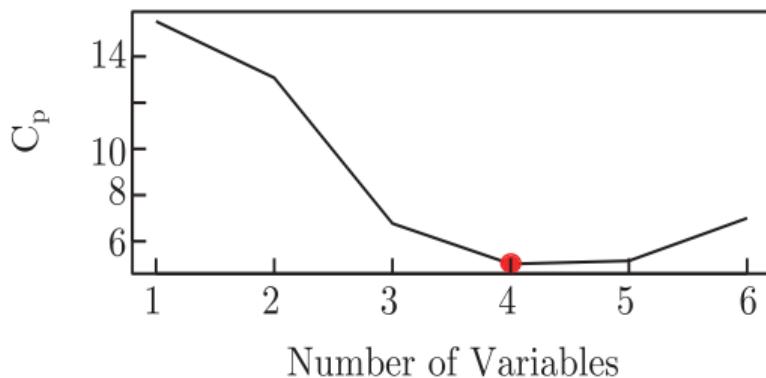
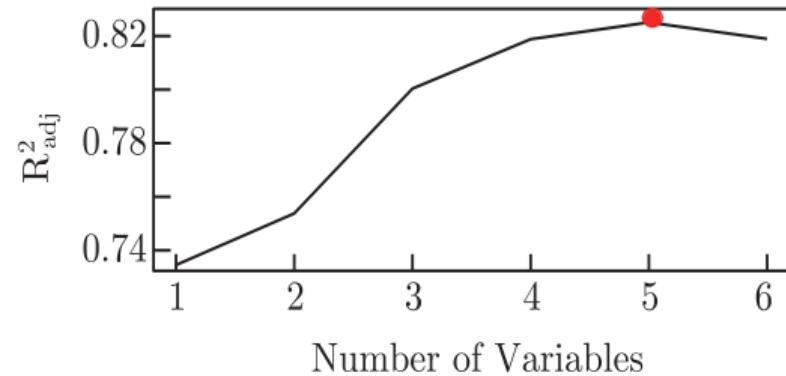
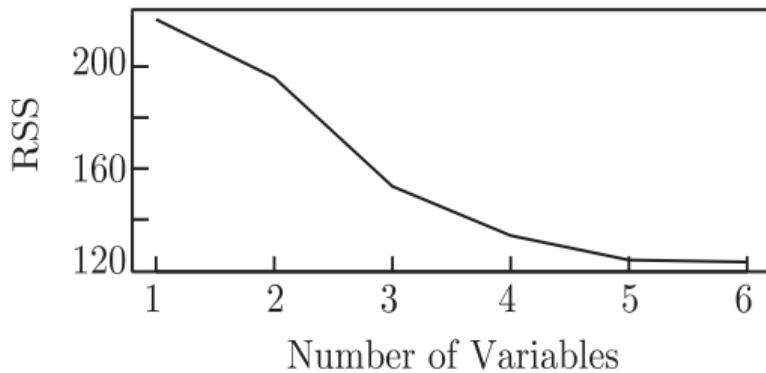
```
library(leaps); health = read.table("health.txt", header=TRUE)
regfit.full = regsubsets(Y~., health)
reg.summary = summary(regfit.full)
> reg.summary      ## 输出结果
Subset selection object
Call: regsubsets.formula(Y ~ ., health)
6 Variables (and intercept)
    Forced in    Forced out
X1        FALSE        FALSE
X2        FALSE        FALSE
X3        FALSE        FALSE
X4        FALSE        FALSE
X5        FALSE        FALSE
X6        FALSE        FALSE
```

案例与R语言计算—31名中年男性的健康数据

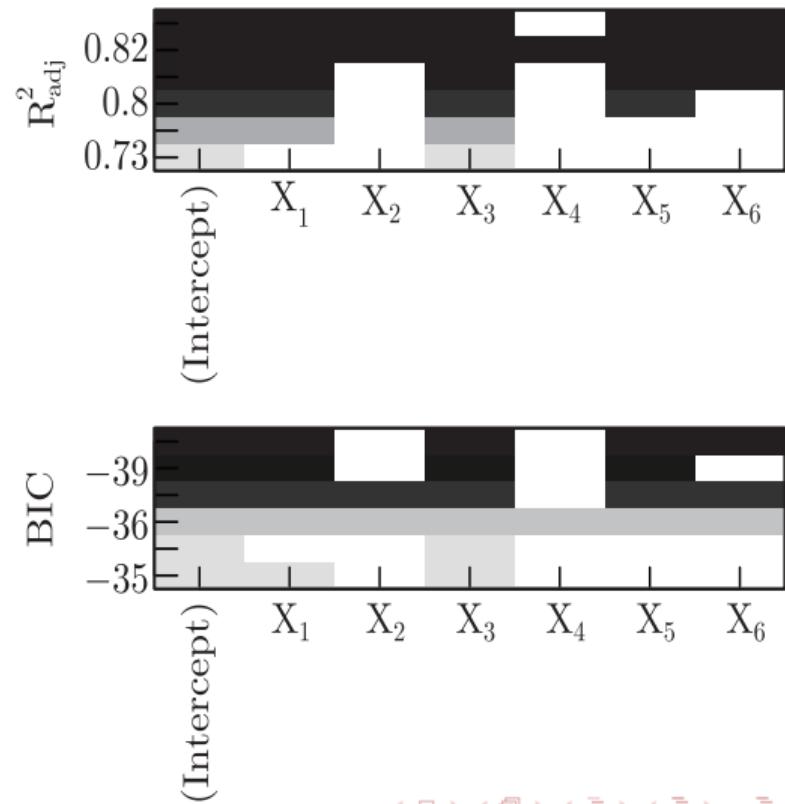
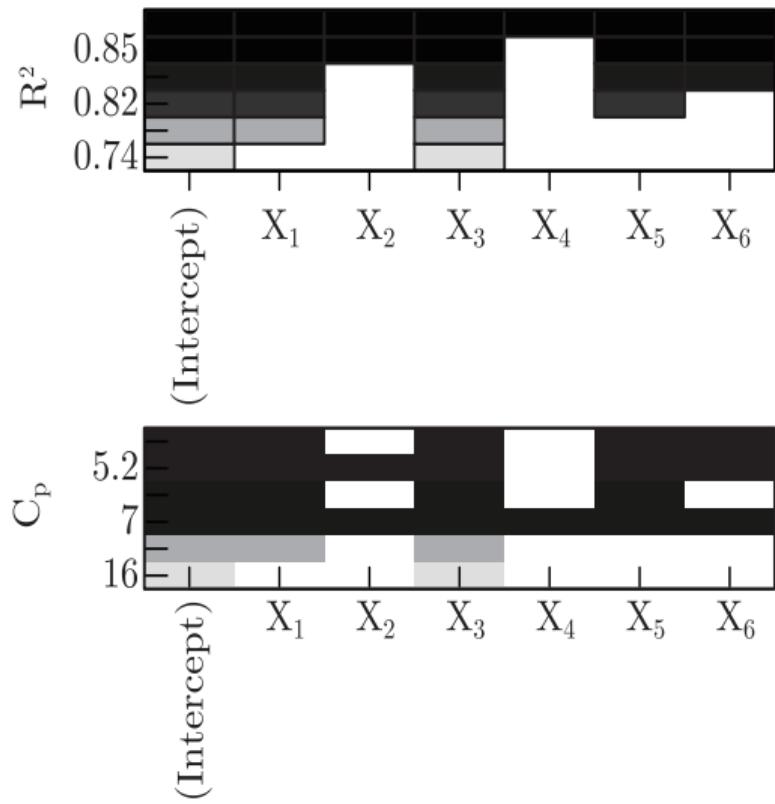
```
1 subsets of each size up to 6
Selection Algorithm: exhaustive
      X1     X2     X3     X4     X5     X6
1 ( 1 ) " "   " "   " * "   " "   " "   " "
2 ( 1 ) " * " " "   " * "   " "   " "   " "
3 ( 1 ) " * " " "   " * "   " "   " * "   " "
4 ( 1 ) " * " " "   " * "   " "   " * "   " * "
5 ( 1 ) " * " " * "   " * "   " "   " * "   " * "
6 ( 1 ) " * " " * "   " * "   " * "   " * "   " * "
```

■ 进一步, 绘制出所有模型的RSS、调整的判定系数 R_{adj}^2 、 C_p 和BIC的图形, 辅助确定最终选择哪一個最优模型.

案例与R语言计算—31名中年男性的健康数据



案例与R语言计算—碎石图



本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

本章介绍几种正则化方法：

- 岭回归(ridge regression)
- 桥回归(bridge regression)
- Lasso (Tibshirani, 1996)
- SCAD (Fan 和Li, 2001)
- 自适应Lasso (Zou, 2006)
- 弹性网(Zou和Hastie, 2005)

岭回归(ridge regression)

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

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

■ 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:

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

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

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 β 的岭回归估计 $\hat{\beta}^R = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$

$$Q(\beta) = \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2,$$

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CV方法进行选取
- ▶ $\lambda \|\beta\|_2^2 = \lambda \sum_{j=1}^p \beta_j^2$ 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

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

- 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 $\lambda \rightarrow \infty$, 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:** 平衡了偏差和方差, 随着 λ 的增加, 岭回归拟合的模型灵活度降低, 尽管方差变小, 但是偏差变大.

岭回归(ridge regression)

■ 由岭回归估计 $\hat{\beta}^R$, 可得 \mathbf{Y} 的岭回归拟合为

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}^R = \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} =: \mathbf{H}(\lambda)\mathbf{Y}.$$

■ 这里, $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T$ 为 **投影矩阵**.

■ 这时, 可得**岭回归估计的自由度**为

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

其中 γ_j 是矩阵 $\mathbf{X}^T\mathbf{X}/n$ 的第 j 个特征值, 且 $j = 1, \dots, p$.

岭回归(ridge regression)

■ 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

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

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

■ 与最优子集选择方法相比, 岭回归方法计算简便.

岭回归(ridge regression)

■ 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

$$\begin{cases} \min_{\beta} \frac{1}{n} \|Y - X\beta\|_2^2, \\ \text{s.t. } \sum_{j=1}^p \beta_j^2 \leq c, \end{cases}$$

其中

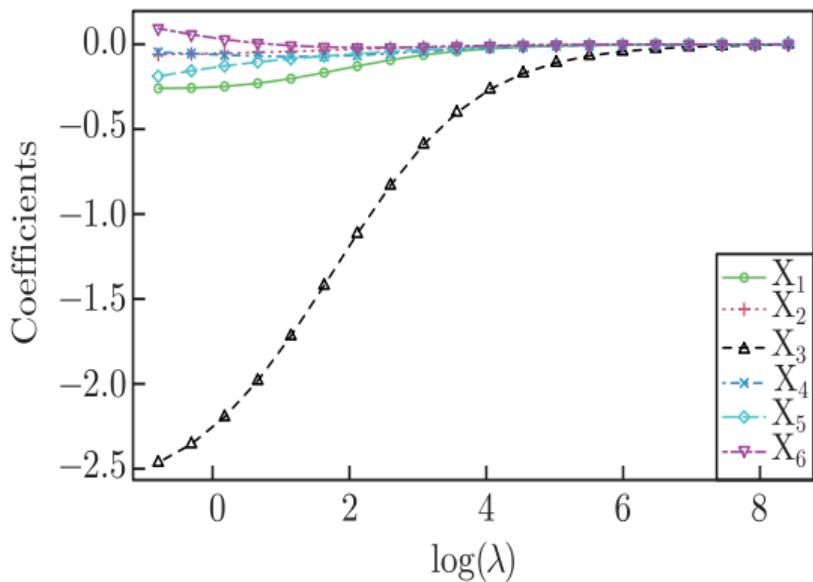
- ▶ c 是一个非负的常数, 作用相当于调节参数 λ
- ▶ 通过 c 的大小控制 $\sum_{j=1}^p \beta_j^2$ 的大小
- ▶ 当 c 的取值为无穷大时, 则约束项不起作用
- ▶ 随着 $c \rightarrow 0$, 约束项的作用增强

■ R 语言中, 岭回归的计算:

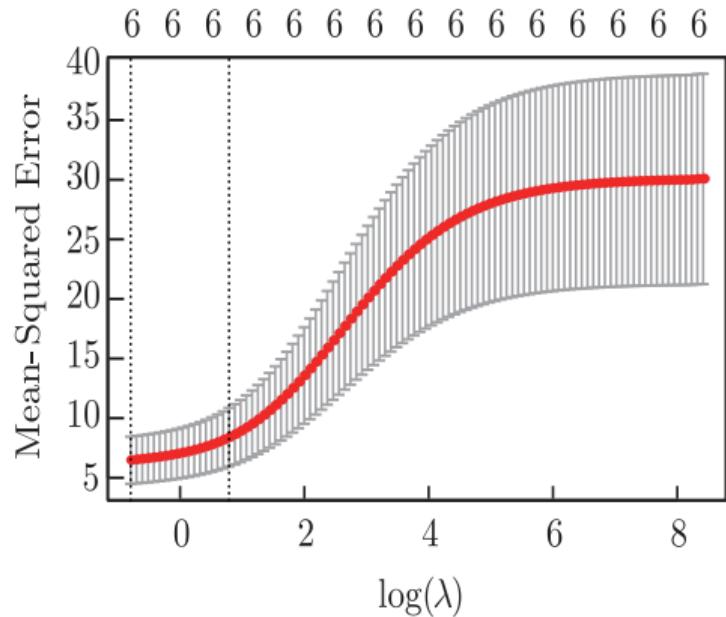
- ▶ 程序包ridge中的函数linearRidge()
- ▶ 程序包MASS中的函数lm.ridge()
- ▶ 程序包glmnet中的函数glmnet(), 其中函数glmnet()中, $\text{alpha}=0$, 拟合岭回归模型; $\text{alpha}=1$, 拟合Lasso 模型; 参数 $0 < \text{alpha} < 1$, 拟合弹性网模型.

■ 例: 31 名中年男性健康数据的岭回归分析.

岭回归—31名中年男性的健康数据



(a)



(b)

(a) 岭回归估计随着 λ 变化的路径图; (b) 岭回归的交叉验证误差图

- 从图(a)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;
- 图(b): 横轴为 $\log(\lambda)$, 而纵轴为交叉验证误差(即 $CV(\lambda) = \overline{MSE}(\lambda)$);
- 图(b)还显示了交叉验证误差的正、负标准差, 即 $\pm sd_{MSE}(\lambda)$;
- 图(b)中左边的垂直虚线表示能使交叉验证误差最小的 $\log(\hat{\lambda})$ 取值, 使用`cv.ridge$lambda.min`获得最优的调节参数为 $\hat{\lambda} \approx 0.452$;
- 图(b)中右边的垂直虚线表示比 $\hat{\lambda}$ 更大, 且与 $CV(\hat{\lambda})$ 相距一个标准差 $sd_{MSE}(\hat{\lambda})$ 的调节参数的取值;
- 使用`cv.ridge$lambda.1se`提取 $\tilde{\lambda}$ 的值为 $\tilde{\lambda} \approx 2.197$.

岭回归—31名中年男性的健康数据

```
library(glmnet); set.seed(2021)

health = read.table("health.txt", header = TRUE)

x=model.matrix(~., 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
```

岭回归—31名中年男性的健康数据

```
> 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
```

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

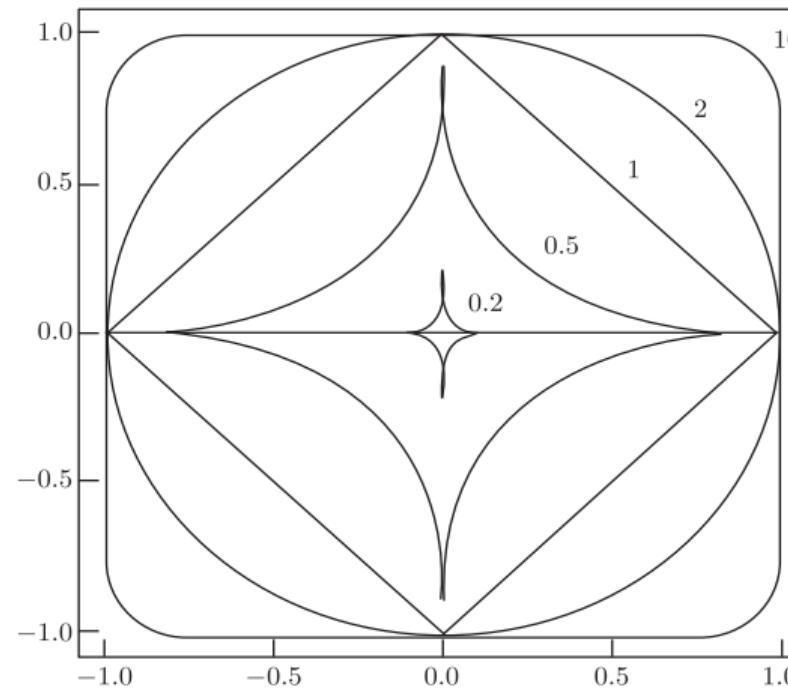
桥回归(bridge regression)

■ 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

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

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

桥回归(bridge regression)



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

$$\begin{cases} \min_{\beta} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2, \\ \text{s.t. } \sum_{j=1}^p |\beta_j|^q \leq c, \end{cases}$$

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>

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

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 (Candès 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

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

其中

- ▶ $p_\lambda(\cdot)$ 是惩罚函数
- ▶ $\lambda \geq 0$ 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。

惩罚变量选择方法

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

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y} / n.$$

- 进一步, 最小二乘目标函数为

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

且

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

惩罚变量选择方法

- $\hat{\beta}_{j,LS} = (\mathbf{X}^T \mathbf{Y}/n)_j$, 即表示第 j 个分量的最小二乘估计, 且 $j = 1, \dots, p$.
- 令 $z_j = \hat{\beta}_{j,LS}$ 表示 β 的第 j 个分量的最小二乘估计, 其中 $j = 1, \dots, p$.
- 对于正交情形, 惩罚最小二乘目标函数变为

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

♠ **问题:** 选择什么样的惩罚函数可以保证变量选择的三个性质?

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

■ 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

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

■ 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

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

- 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**.

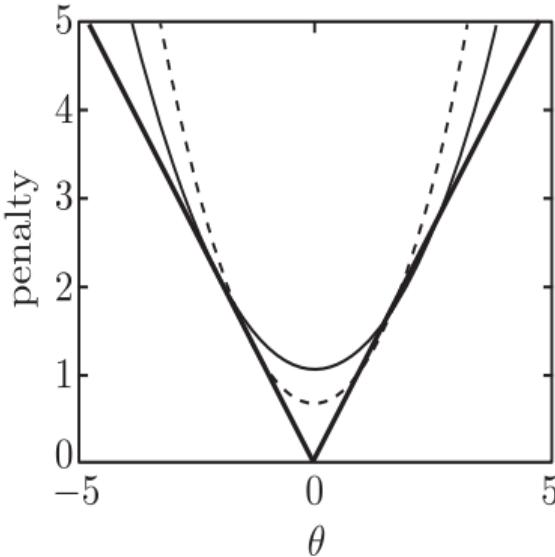
惩罚函数

- 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

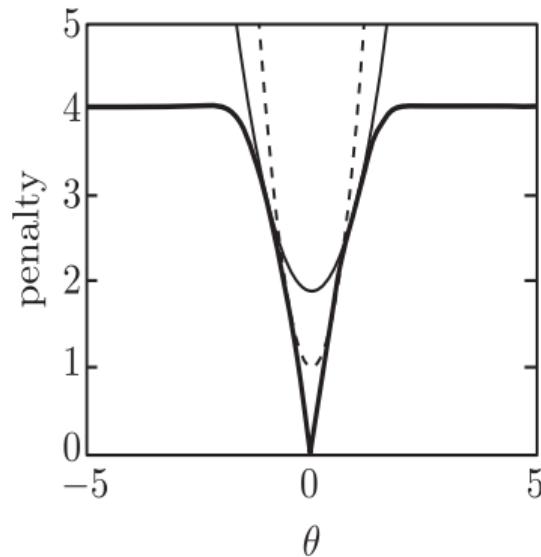
$$\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+$$

- 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.

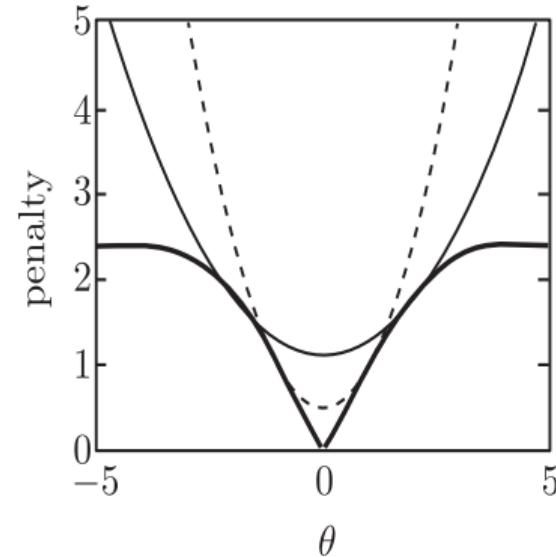
惩罚函数



(a) L_1 惩罚函数



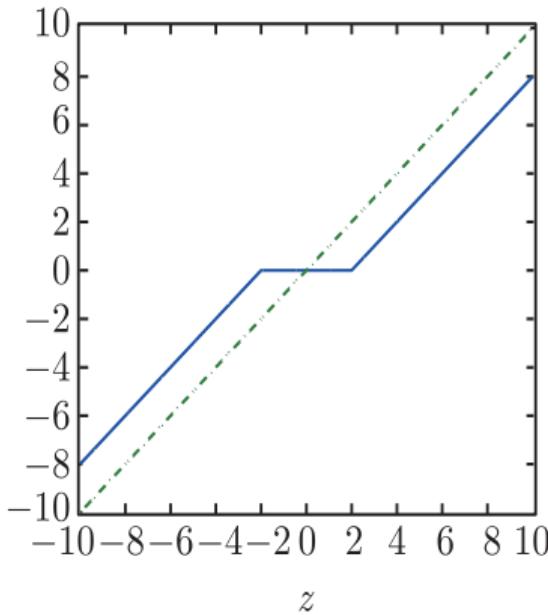
(b) 硬门限惩罚函数



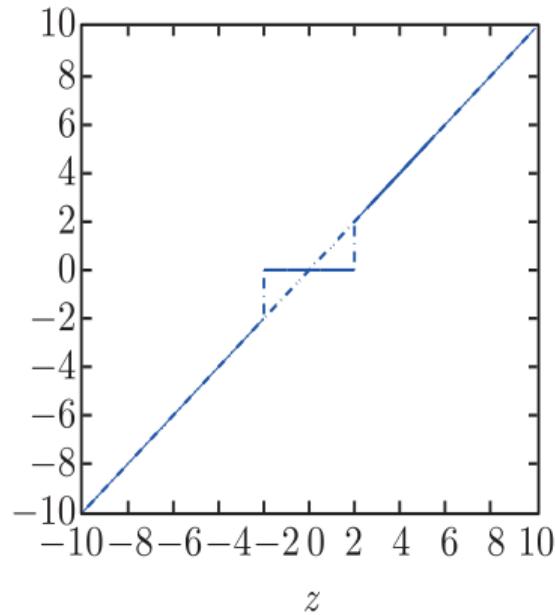
(c) SCAD惩罚函数

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 $a = 3.7$

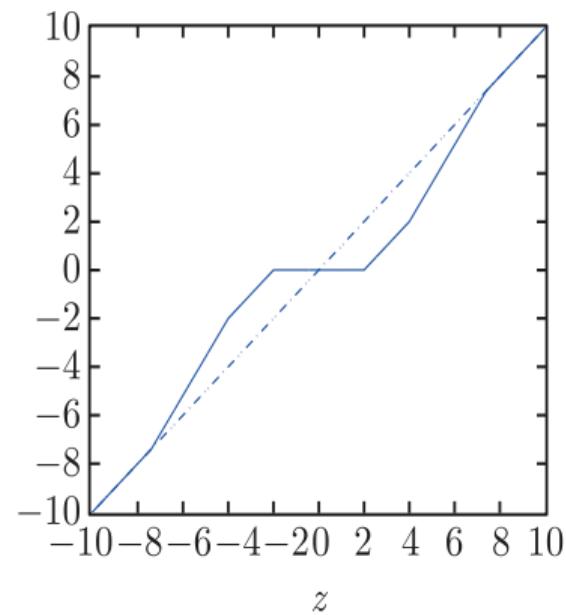
惩罚函数



(a) Lasso门限



(b) 硬门限



(c) SCAD门限

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

惩罚函数

■ Antoniadis (1997) 提出了如下的硬门限惩罚函数

$$p_\lambda(|\theta|) = \frac{1}{2}\lambda^2 - \frac{1}{2}(|\theta| - \lambda)^2 I(|\theta| < \lambda).$$

■ 如果施加硬门限惩罚函数, 可以得到如下的硬门限解

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

■ 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.

惩罚函数

■ 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

$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \quad \text{其中 } a > 2.$$

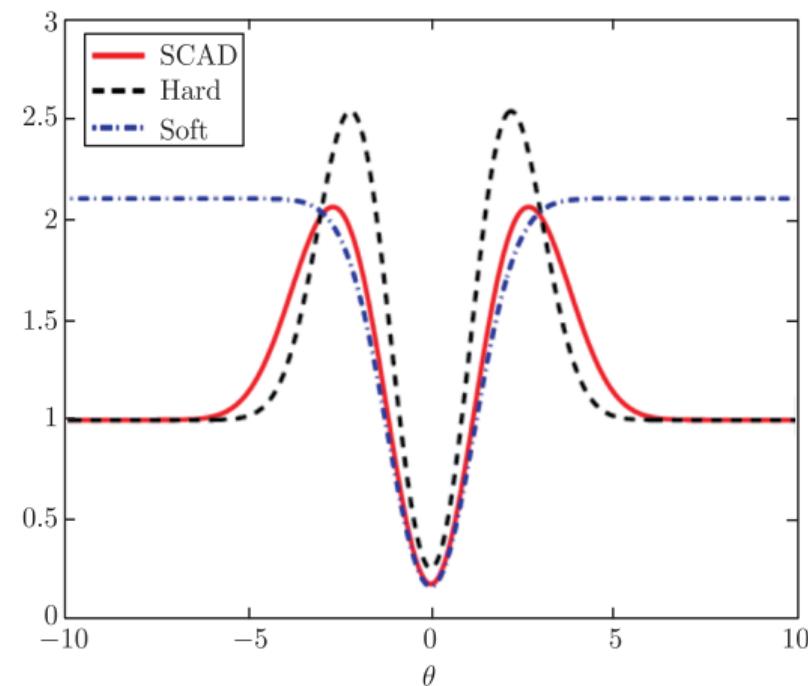
■ SCAD 惩罚最小二乘解为

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

惩罚函数

- 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:

$$R(\hat{\theta}, \theta) = E_{\theta}(\hat{\theta} - \theta)^2.$$



本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

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 $\hat{\beta}^L$

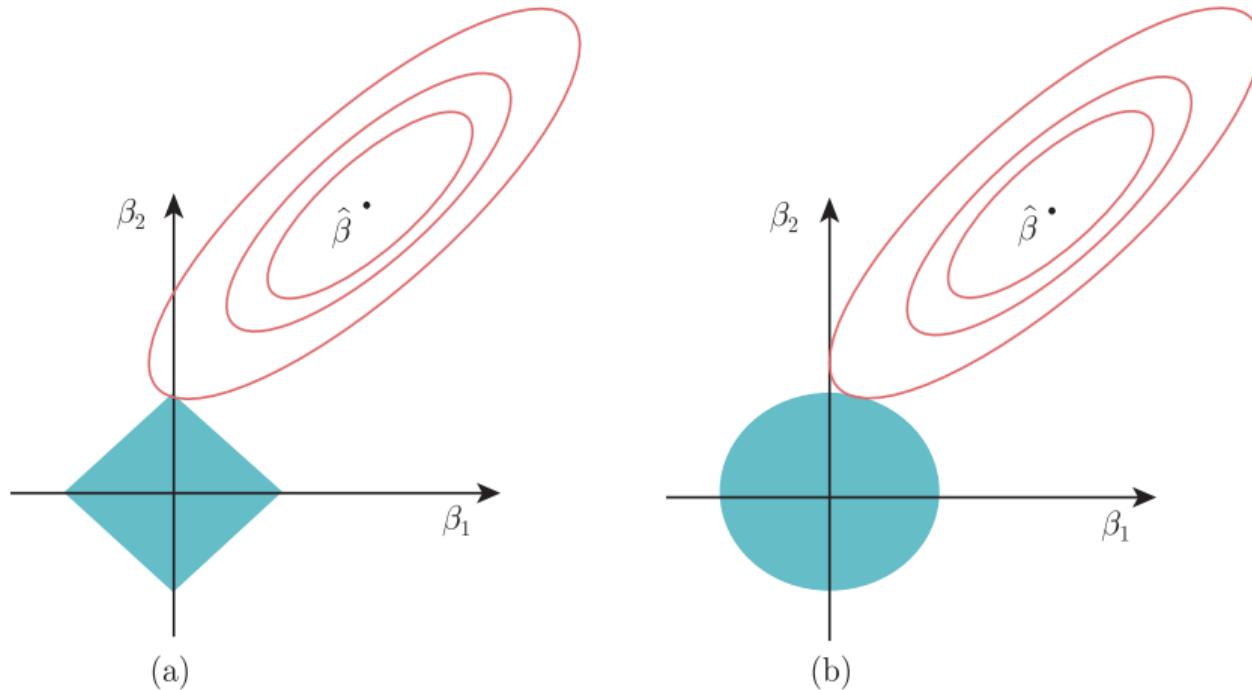
$$\frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

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

$$\begin{cases} \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\}, \\ \text{s.t. } \sum_{j=1}^p |\beta_j| \leq c. \end{cases}$$

- 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



误差等高线和限制条件区域, (a) Lasso; (b) 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 $\beta_1^2 + \beta_2^2 \leq c$; $\hat{\beta}$ 为最小二乘估计

Lasso方法

■ 给定 λ , 令 $\widehat{\beta}(\lambda)$ 表示 β 的Lasso估计, 则Lasso估计的自由度为:

$$\widehat{df}(\lambda) = \#\{j : \widehat{\beta}_j(\lambda) \neq 0\}.$$

■ GCV方法: 可以极小化下面的GCV准则选择调节参数 λ , 即

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

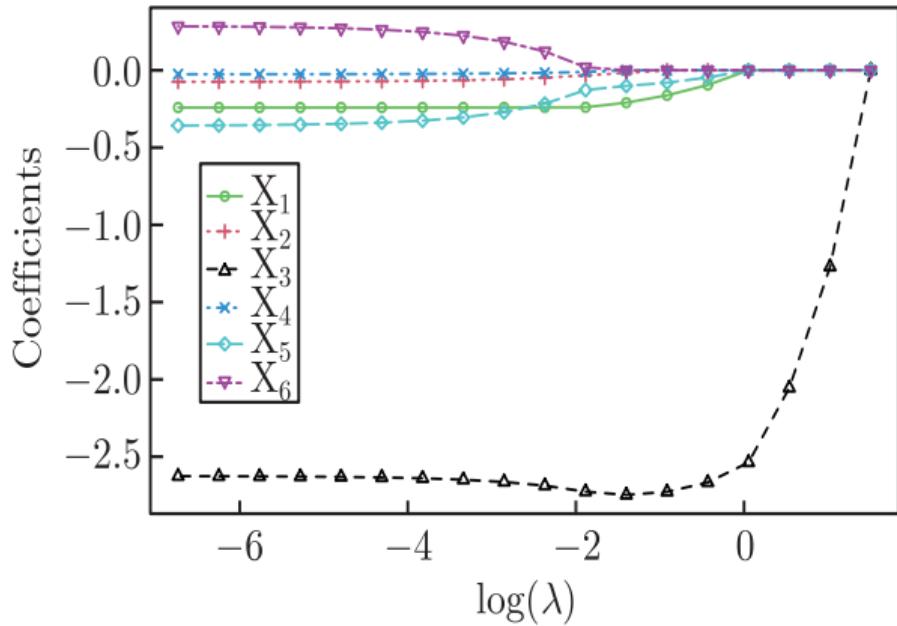
■ BIC方法: 可以通过极小化下面的BIC准则进行选择调节参数 λ , 即

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

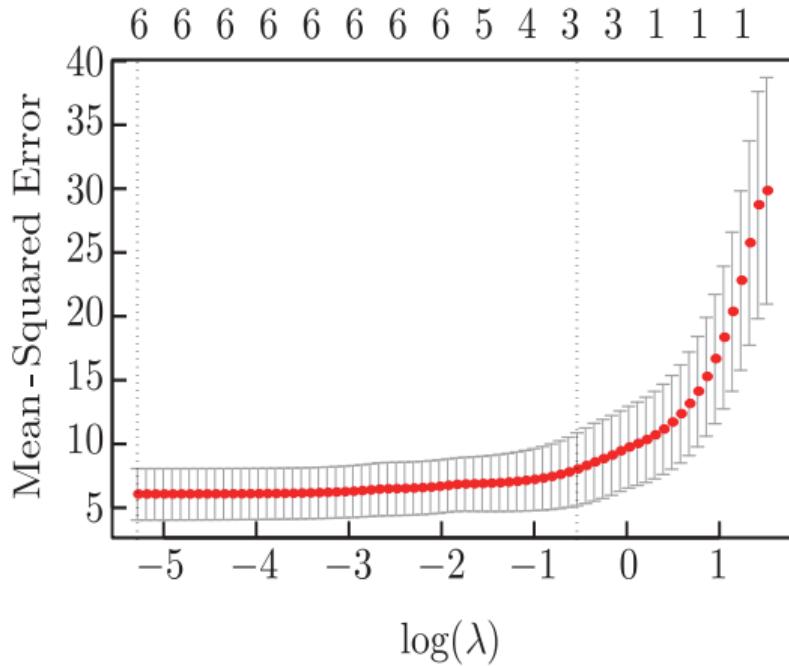
Lasso方法—31名中年男性的健康数据

```
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
```

Lasso方法—31名中年男性的健康数据



(a)



(b)

(a) Lasso估计随着 λ 变化的路径图; (b) Lasso回归的交叉验证误差图

Lasso方法—31名中年男性的健康数据

```
> coef(cv.lasso, s="lambda.min")      > coef(cv.lasso, s="lambda.1se")
7 x 1 sparse Matrix of class "dgCMatrix"

                                         1                               1
(Intercept) 105.01633937          (Intercept) 90.52944150
X1           -0.24071641          X1           -0.11267059
X2           -0.07302177          X2           .
X3           -2.62876371          X3           -2.68260426
X4           -0.02484750          X4           .
X5           -0.35117527          X5           -0.05522621
X6            0.27768283          X6           .
```

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

■ 针对多元线性模型, 考虑下面的SCAD惩罚最小二乘目标函数

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

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 $\beta_0 = (\beta_{01}, \dots, \beta_{0p})^T = (\beta_{0,I}^T, \beta_{0,II}^T)^T$, 其中 $\beta_{0,I}$ 是真参数向量 β_0 的前 s 个分量向量.
- 不失一般性, 假设 $\beta_{0,II} = \mathbf{0}$, 并且 $\beta_{0,I}$ 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 $Q(\beta)$, 可得 β 的一个 **SCAD 估计**, 记为 $\widehat{\beta}$.

定理1

假设 $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ 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 $\Pi = E(\mathbf{X}\mathbf{X}^T)$ 是有限且正定矩阵，模型误差 ε 的均值为 0，方差为 σ^2 . 如果 $a_n \rightarrow 0$ 和 $b_n \rightarrow 0$ ，则依概率趋于 1，存在 $Q(\boldsymbol{\beta})$ 的一个局部最小值 $\hat{\boldsymbol{\beta}}$ ，使得

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

■ 为了讨论下面的oracle性质, 考虑线性回归模型

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

其中 $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ 和 $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

■ $\boldsymbol{\beta}$ 的一个理想估计是

$$\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}, \quad \hat{\boldsymbol{\beta}}_I = (\mathbf{X}_I^T \mathbf{X}_I)^{-1} \mathbf{X}_I^T \mathbf{Y}.$$

■ 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**.

SCAD方法的理论结果

定理2: oracle性质

在定理1的条件下, 假设

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

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

① (稀疏性) $\widehat{\boldsymbol{\beta}}_{II} = \mathbf{0}$;

② (渐近正态性)

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

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

$$\boldsymbol{\Sigma} = \text{diag}\{p''_\lambda(|\beta_{01}|), \dots, p''_\lambda(|\beta_{0s}|)\}, \quad \mathbf{b} = (p'_\lambda(|\beta_{01}|)\text{sgn}(\beta_{01}), \dots, p'_\lambda(|\beta_{0s}|)\text{sgn}(\beta_{0s}))^T.$$

■ 因为SCAD惩罚函数在原点处是奇异的, $Q(\beta)$ 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 β 的最优解.

♠ **问题:** 如何处理非光滑和非凸的惩罚函数?

■ 因为SCAD惩罚函数在原点处是奇异的, $Q(\beta)$ 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 β 的最优解.

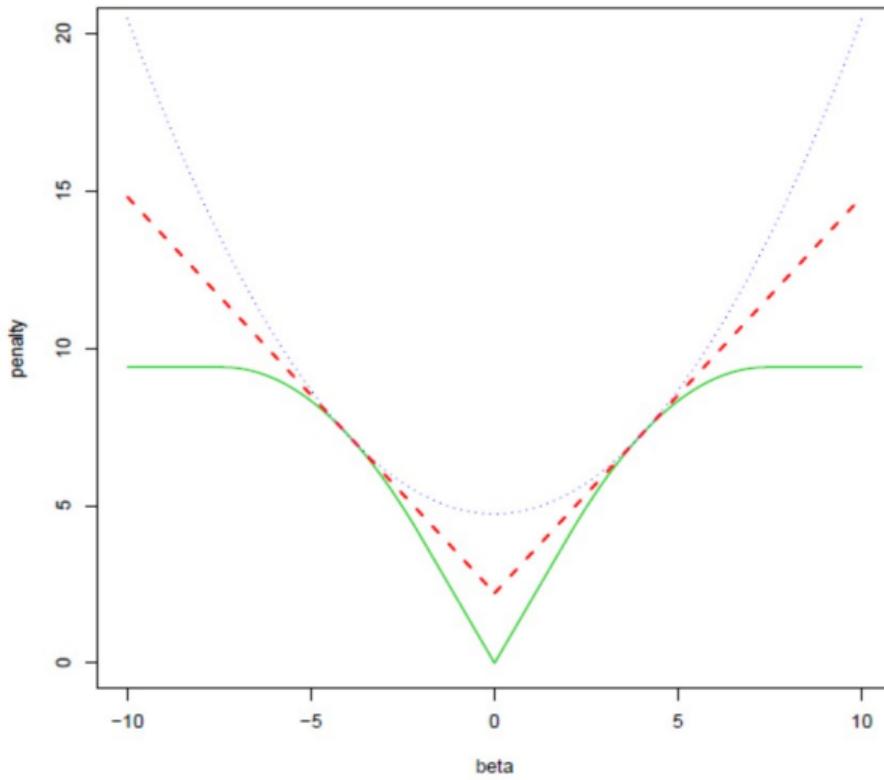
♠ **问题:** 如何处理非光滑和非凸的惩罚函数?

■ 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 $\beta^{(0)} = (\beta_1^{(0)}, \dots, \beta_p^{(0)})^T$. 可用没有惩罚的最小二乘估计作为初始估计;

步骤2 (LQA): 对 $j = 1, \dots, p$, 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

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

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(1)代入到惩罚最小二乘目标函数 $Q(\beta)$ 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:

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

其中

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

- 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

$$-\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + np'_\lambda(|\hat{\beta}_j|)\text{sgn}(\hat{\beta}_j) = \mathbf{0}$$

- 对非零回归系数的估计, 满足

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X} + n \Sigma_\lambda(\hat{\boldsymbol{\beta}}) \right)^{-1} \mathbf{X}^T \mathbf{Y}.$$

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) 提出了局部线性逼近(local linear approximation, LLA) 算法.

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

步骤2: 第 k 步, 令 $\beta^{(k)} = (\beta_1^{(k)}, \dots, \beta_p^{(k)})^T$ 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

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(k)}|) + p'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|); \quad (2)$$

步骤3: 把惩罚函数的局部线性逼近(2)代入到惩罚最小二乘目标函数 $Q(\beta)$ 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

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

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.

- 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 $\hat{\beta}$, 响应变量 \mathbf{Y} 预测的拟合为

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\hat{\beta})\}^{-1}\mathbf{X}^T\mathbf{Y}.$$

- 自由度定义为: $\hat{df}(\lambda) = \text{tr} \left\{ \mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\hat{\beta})\}^{-1}\mathbf{X}^T \right\}.$
- 在一定条件下, Zhang等(2010)证明: $\mathbb{P}\{\hat{df}(\lambda) = df_N(\lambda)\} = 1.$

■ GCV方法: 极小化下面的GCV准则选择调节参数 λ , 即

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

■ BIC准则: 极小化下面的BIC准则选择调节参数 λ , 即

$$\hat{\lambda}_{\text{bic}} = \arg \min_{\lambda} \text{BIC}(\lambda) = \arg \min_{\lambda} \left\{ \log \hat{\sigma}_{\lambda}^2 + \hat{df}(\lambda) \log(n)/n \right\}.$$

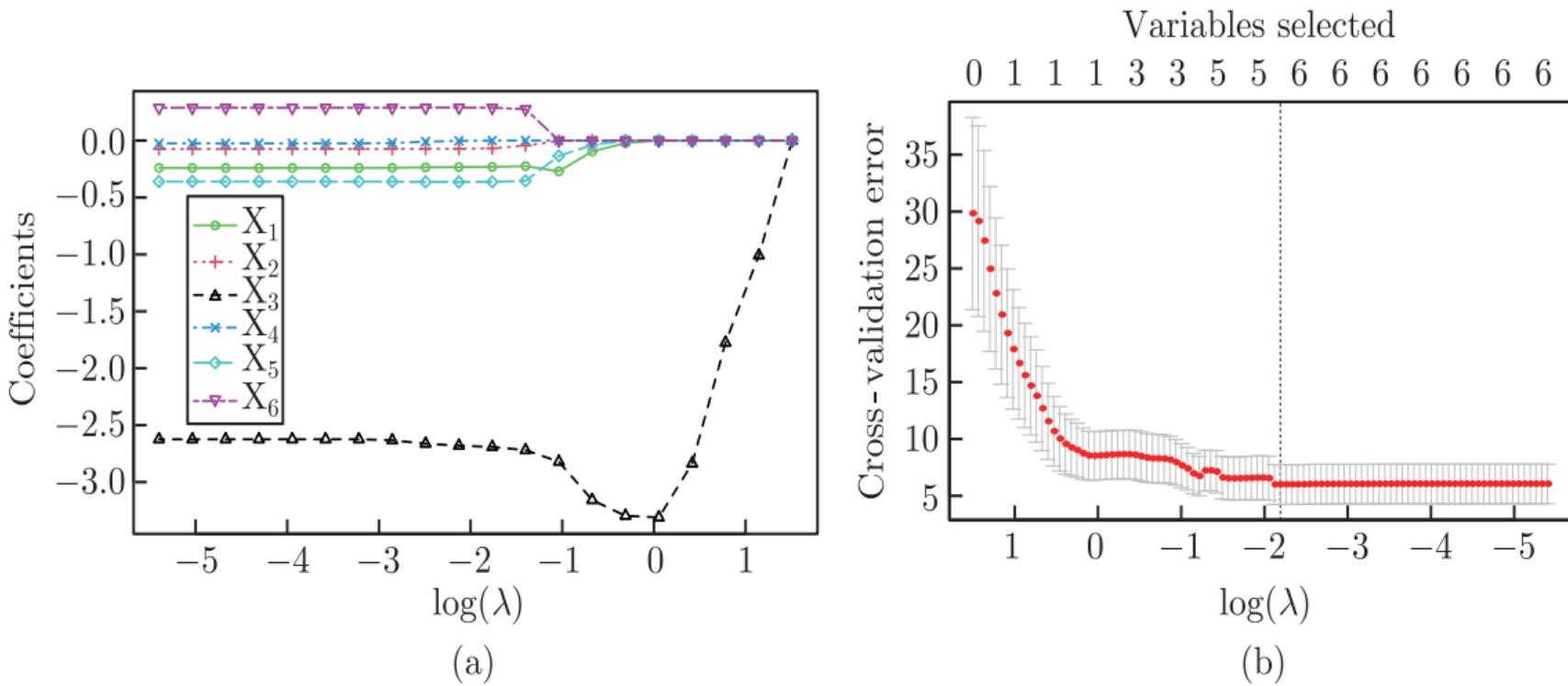
■ 统计性质:

- * 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](#)

SCAD方法—31名中年男性的健康数据



(a) SCAD估计随着 λ 变化的路径图; (b) 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 $CV(\hat{\lambda})$ 最小化的 λ 为 $\hat{\lambda} \approx 0.1119$;
- 对应的回归系数分别为: $\hat{\beta}_0 \approx 104.13, \hat{\beta}_1 \approx -0.23, \hat{\beta}_2 \approx -0.07, \hat{\beta}_3 \approx -2.68, \hat{\beta}_4 \approx -0.00, \hat{\beta}_5 \approx -0.36$ 和 $\hat{\beta}_6 \approx 0.29$;
- 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 $\tilde{\lambda} = 0.3929893$ 时, 可以产生稀疏解, 使得 X_2, X_4 和 X_6 的系数为0.

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

■ 为了解决Lasso估计不具有无偏性的问题, Zou (2006) 提出了**自适应Lasso**(adaptive Lasso, ALasso), 并证明了自适应Lasso具有**oracle性质**.

■ 自适应Lasso估计 $\hat{\beta}^{\text{lasso}}$ 定义为

$$\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \left\{ \frac{1}{2n} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right\},$$

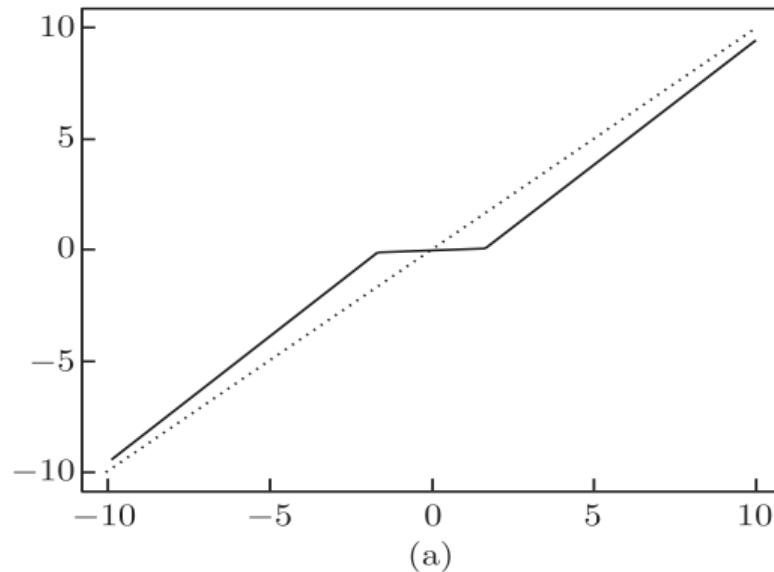
其中 \hat{w}_j 是非负的权重.

- Zou (2006) 建议取自适应权重为 $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$, 其中 $\gamma > 0$, $\hat{\beta}_j$ 是 β_j 的一个 \sqrt{n} -相合估计, 如最小二乘估计.
- 在正交设计下, 取 $\gamma = 1$, β_j 的自适应Lasso估计为

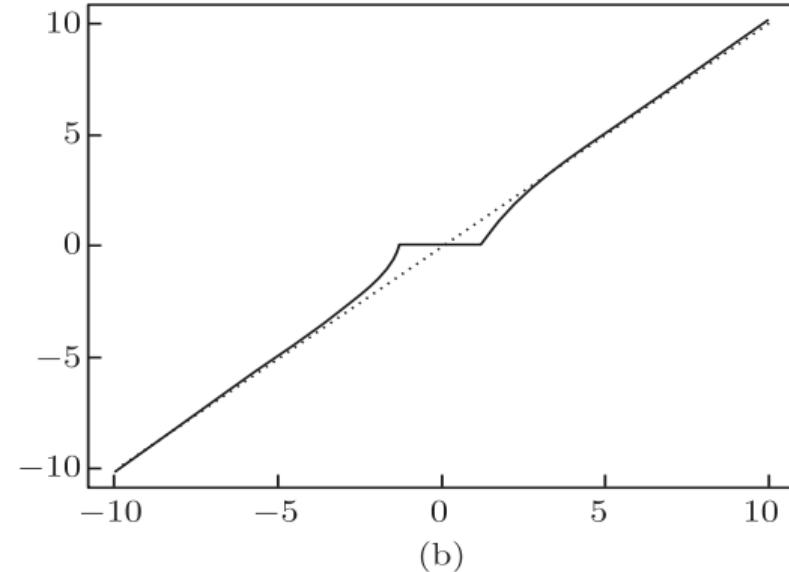
$$\hat{\beta}_j^{\text{lasso}} = \text{sgn}(\hat{\beta}_{j,\text{LS}}) \left(|\hat{\beta}_{j,\text{LS}}| - \frac{\lambda}{|\hat{\beta}_{j,\text{LS}}|} \right)_+, \quad j = 1, \dots, p,$$

其中 $\hat{\beta}_{j,\text{LS}} = (\mathbf{X}^T \mathbf{Y}/n)_j$ 为 β_j 在正交设计下的最小二乘估计.

■ 选取合适的 γ , 自适应Lasso估计将满足无偏性、稀疏性和连续性.



(a)



(b)

自适应Lasso门限函数图. (a) 取 $\gamma = 0.5$ 和 $\lambda = 2$; (b) 取 $\gamma = 2$ 和 $\lambda = 2$

♠ **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, DFTtype="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"lasso"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lasso方法的参数, 默认为1; 其余参数见在线帮助.

自适应Lasso — 31名中年男性的健康数据

```
library(msgps)
alasso_fit = msgps(x, y, penalty = "alasso", gamma = 1, lambda = 0)
plot(alasso_fit, criterion = "gcv", xvar = "t", main = "GCV")
plot(alasso_fit, criterion = "bic", xvar = "t", main = "BIC")
summary(alasso_fit)      ## 用函数summary()汇总结果，并输出结果
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 — 31名中年男性的健康数据

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

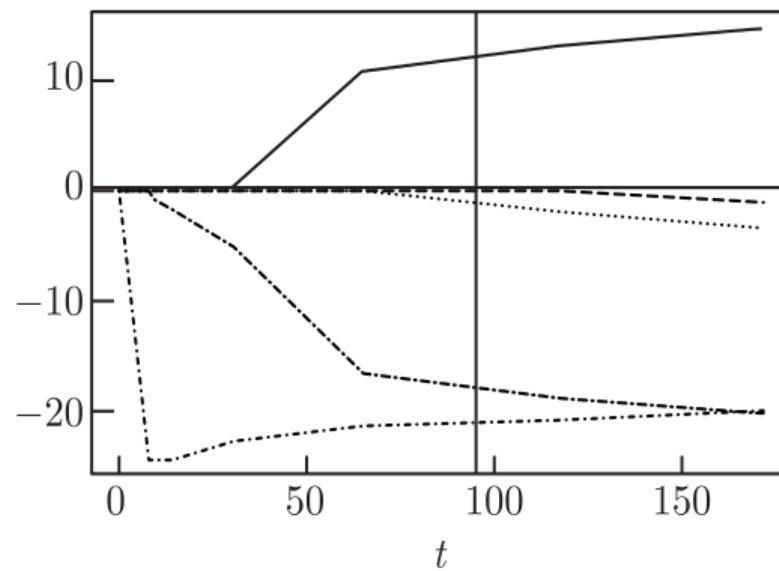
ms.df:

	Cp	AICC	GCV	BIC
[1,]	4.11	3.525	4.154	3.405

自适应Lasso — 31名中年男性的健康数据

standardized coefficients

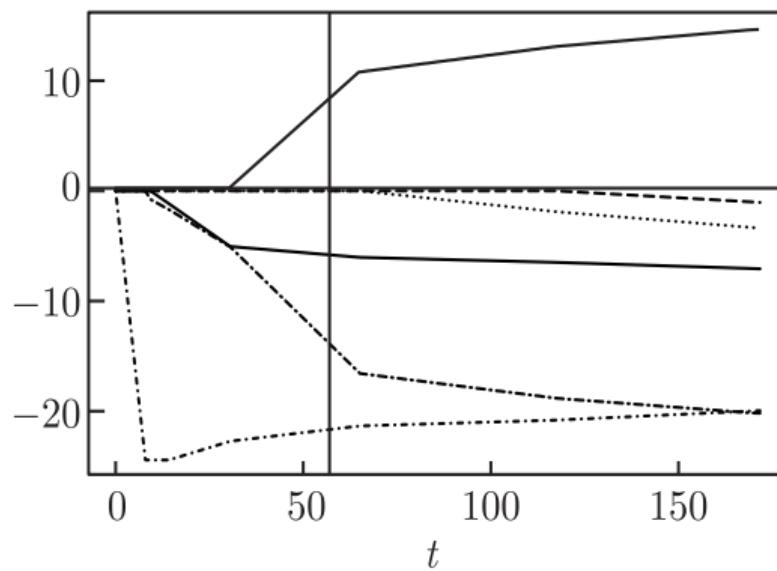
GCV



(a)

standardized coefficients

BIC



(b)

(a) 垂直虚线是用GCV准则选取的调节参数; (b) 垂直虚线是用BIC准则选取的调节参数

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

- 为了解决变量中存在多重共线性问题, Zou和Hastie (2005)将岭回归和Lasso方法进行结合, 提出了**弹性网方法**(elastic net).
- 弹性网方法的核心思想: 在最小二乘目标函数后同时施加 L_1 和 L_2 惩罚函数, 即

$$\frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2,$$

其中 λ_1 和 λ_2 都是非负的调节参数.

- $p_{\lambda_1, \lambda_2}(|t|) = \lambda_1|t| + \lambda_2 t^2$ 称为**弹性网惩罚函数**.

- 由于 λ_1 和 λ_2 的取值范围均为 $[0, \infty)$, 不便于使用数据驱动的CV和GCV等方法同时选择两个最优的调节参数.
- Zou和Hastie (2005)建议考虑如下的弹性网惩罚函数

$$p_{\lambda,\alpha}(|t|) = \lambda [\alpha|t| + (1 - \alpha)t^2],$$

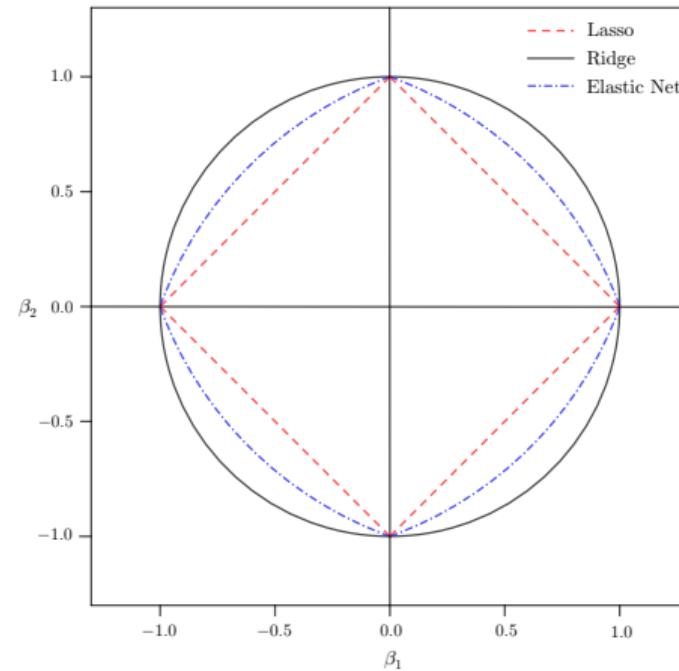
其中 $\lambda = \lambda_1 + \lambda_2 \geq 0$ 和 $\alpha = \lambda_1/\lambda \in [0, 1]$.

- 这时, 回归系数向量 β 的**弹性网估计**为

$$\hat{\beta}^{\text{enet}} = \arg \min_{\beta} \left\{ \frac{1}{n} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \left[\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right] \right\}.$$

- 把调节参数 α 限制在区间 $[0,1]$ 上, 可减少计算量并方便使用数据驱动的CV准则选择最优的调节参数 (λ, α) .
- 例如, 首先取 $\alpha = 0.1k$, 其中 $k = 1, \dots, 10$, 然后利用CV准则选取最优的 (λ, α) .
- 当 $\alpha = 0$ 时, 所得弹性网估计退化为岭回归估计;
- 当 $\alpha = 1$ 时, 所得弹性网估计退化为Lasso估计;
- 如果 $0 < \alpha < 1$ 时, 弹性网估计为岭回归估计和Lasso估计的折中.

弹性网方法



在二维空间中, 岭回归, Lasso和弹性网的约束集, 其中弹性网中 $\alpha = 0.5$

■ 在R语言中，可用三个程序包获得 β 的弹性网估计：

- ① 程序包`glmnet`中的函数`glmnet()`，其中参数`alpha` $\in (0, 1)$ ；
- ② 程序包`elasticnet`中的函数`enet()`，其中参数`lambda`是 L_2 惩罚函数对应的调节参数，当`lambda=0`时，对应Lasso方法。程序包`elasticnet`是基于最小角回归的程序包`lars`开发而成；
- ③ 程序包`msgps`中的函数`msgps()`，当参数`penalty="enet"`和参数`alpha` $\in (0, 1)$ 时，表示拟合弹性网模型。

- 假设响应变量 Y 满足模型: $Y = 1.5Z_1 - 0.5Z_2 + \varepsilon$, 其中 Z_1 和 Z_2 是两个独立的随机变量, 都来自标准正态分布 $N(0, 1)$, 模型误差 $\varepsilon \sim N(0, 1)$.
- 假设观测的协变量 X_1, \dots, X_6 产生于

$$\begin{cases} X_j = Z_1 + \epsilon_j/4, & j = 1, 2, 3, \\ X_j = Z_2 + \epsilon_j/4, & j = 4, 5, 6, \end{cases}$$

其中 $\epsilon_1, \dots, \epsilon_6$ 独立同分布, 且来自于标准正态分布 $N(0, 1)$.

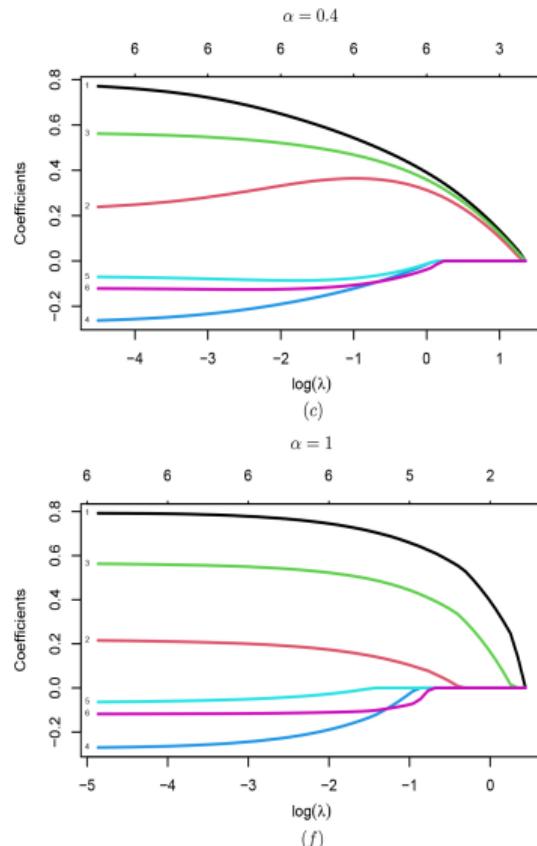
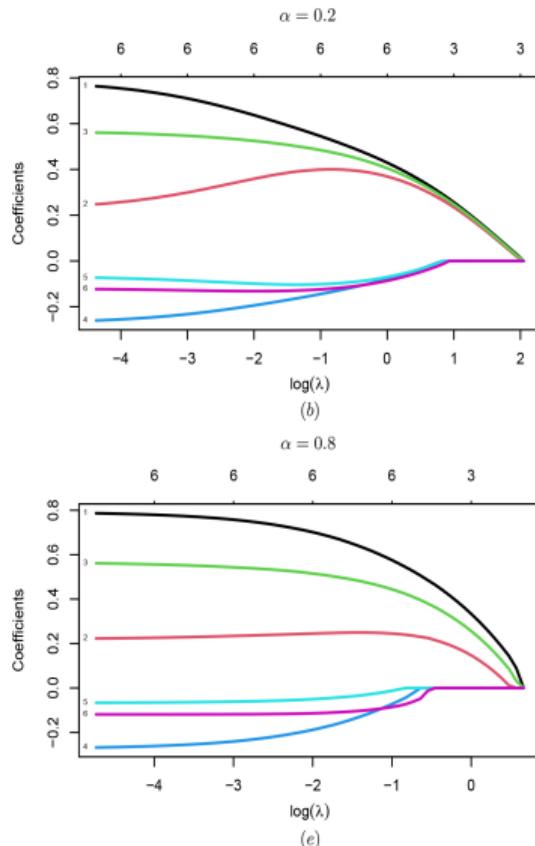
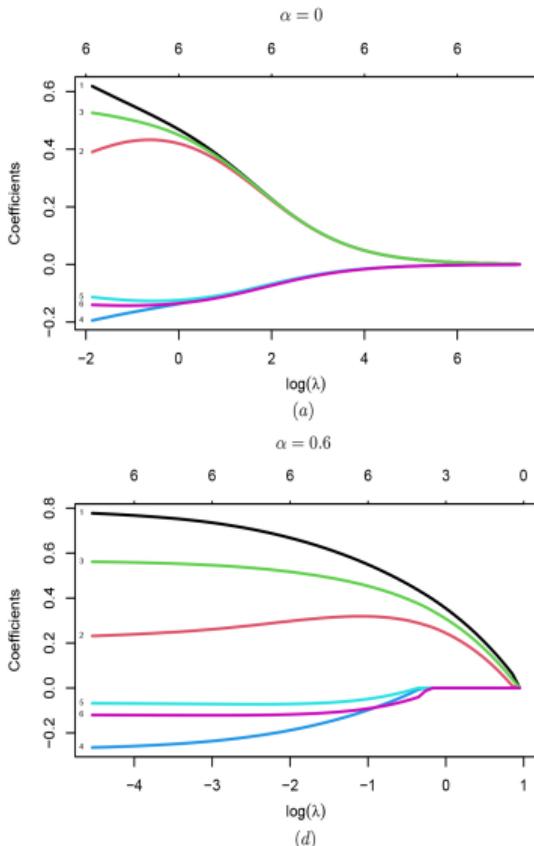
■ 从模型设置可见：

- ① X_1, X_2, X_3 是一个组，依赖于变量 Z_1 ；
- ② X_4, X_5, X_6 是一个组，依赖于变量 Z_2 ；
- ③ X_1, X_2, X_3 之间的相关系数为 1，具有很强的相关性；同样 X_4, X_5, X_6 之间的相关系数也为 1，也具有很强的相关性；而 X_1, X_2, X_3 与 X_4, X_5, X_6 两组之间相互独立。

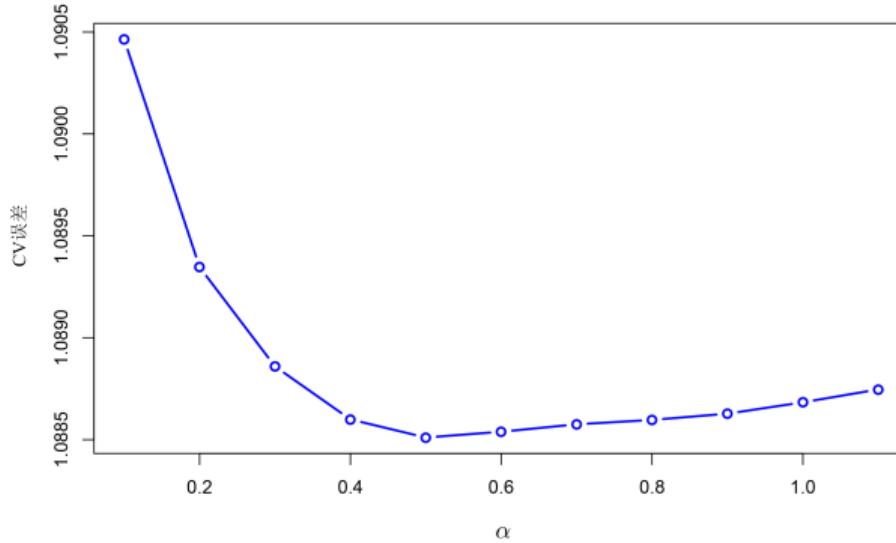
■ 主要目的：针对高度相关的数据能准确识别 Z_1 组的三个变量 X_1, X_2, X_3 。

- 从模型中产生500个独立同分布的随机样本 $\{(y_i, \mathbf{x}_i), i = 1, \dots, 500\}$, 其中 $\mathbf{x}_i = (x_{i1}, \dots, x_{i6})^T$.
- 为了比较, 取 $\alpha = 0, 0.2, 0.4, 0.6, 0.8, 1$, 其中
 - ① $\alpha = 0$ 表示岭回归;
 - ② $\alpha = 1$ 表示Lasso;
 - ③ 其他情况表示弹性网.

弹性网方法



■ CV误差图显示, 当选取 $\alpha = 0.5$ 进行弹性网拟合数据时, 可使得CV误差达到最小.



本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

- 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

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

- ▶ $\beta = (-1.5, 1, 0.8, -0.8, 0.4, 0, \dots, 0)^T$ 为 p 维回归系数向量
- ▶ $x_i = (x_{i1}, \dots, x_{ip})^T$ 从 p 元正态分布 $N_p(\mathbf{0}, \Sigma)$ 中随机产生随机数, 这里, $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$, 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 x_i
- ▶ 响应变量 y_i 可以通过上面多元线性回归模型产生

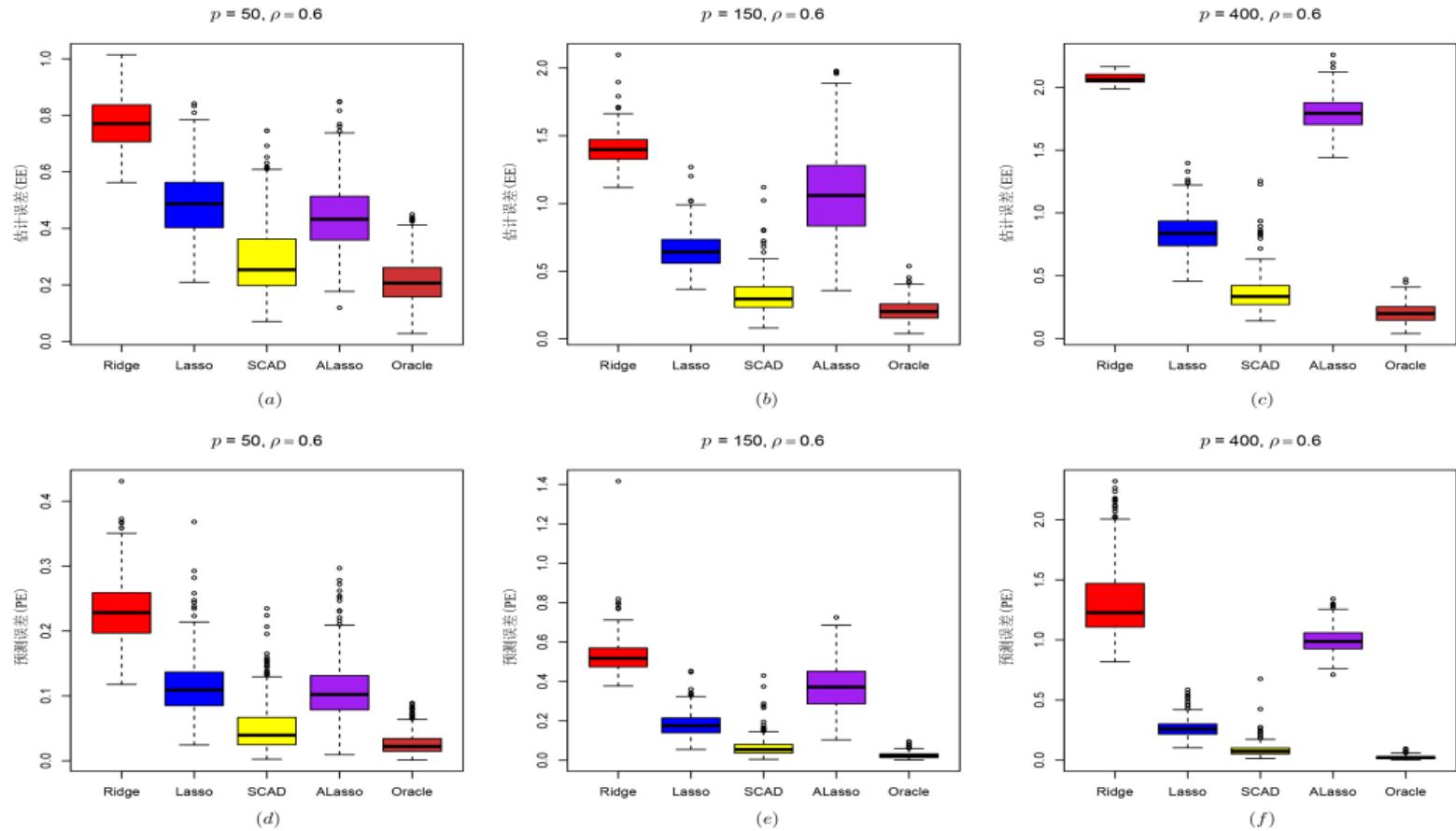
- 前5个协变量对响应变量有显著性影响, 而剩余的 $p - 5$ 个协变量对响应变量不显著, 故设回归系数为零;
- 取样本量 $n = 200$, 维数 $p = 50, 150, 400$, 其中当 $p = 400$ 时, 维数 p 大于样本量 n ;
- 取 $\rho = 0.3, 0.6$ 两种情况;
- 为了对岭回归、Lasso、SCAD和自适应Lasso四种方法进行比较, 重复500次试验;
- 采用10折CV方法选取调节参数 λ .

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 $\|\hat{\beta} - \beta\|_2$ 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 $(\hat{\beta} - \beta)^T E(XX^T)(\hat{\beta} - \beta)$ 来计算.
- **注:** 对自适应Lasso, 当维数 $p > n$ 时, 初始估计取岭回归估计, 其中岭回归的调节参数取0.00001.

模拟研究

		$\rho = 0.3$				$\rho = 0.6$					
		p	指标	Ridge	Lasso	SCAD	ALasso	Ridge	Lasso	SCAD	ALasso
$p = 50$	EE	0.6103	0.3674	0.2183	0.3631	0.7715	0.4887	0.2859	0.4424		
	PE	0.2394	0.1035	0.0446	0.1079	0.2301	0.1144	0.0505	0.1075		
	C	5.0000	5.0000	5.0000	5.0000	5.0000	4.9920	4.9800	4.9940		
	IC	45.000	12.786	3.8060	14.942	45.000	15.242	4.2680	14.694		
$p = 150$	EE	1.2188	0.4497	0.2376	0.8692	1.4046	0.6515	0.3184	1.0750		
	PE	0.5643	0.1455	0.0537	0.3700	0.5271	0.1811	0.0638	0.3727		
	C	5.0000	5.0000	5.0000	4.9900	5.0000	4.9200	4.9540	4.9640		
	IC	145.00	21.464	6.7140	66.638	145.00	27.986	8.7480	67.092		
$p = 400$	EE	1.9926	0.5414	0.2701	1.5421	2.0757	0.8385	0.3588	1.7941		
	PE	1.7661	0.1953	0.0683	0.9959	1.3181	0.2618	0.0822	0.9929		
	C	5.0000	4.9820	4.9840	4.9780	5.0000	4.7040	4.9120	4.9240		
	IC	395.00	30.0980	11.122	291.834	395.00	38.856	16.370	291.922		



总结

考虑下面的正交设计的线性模型: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$.

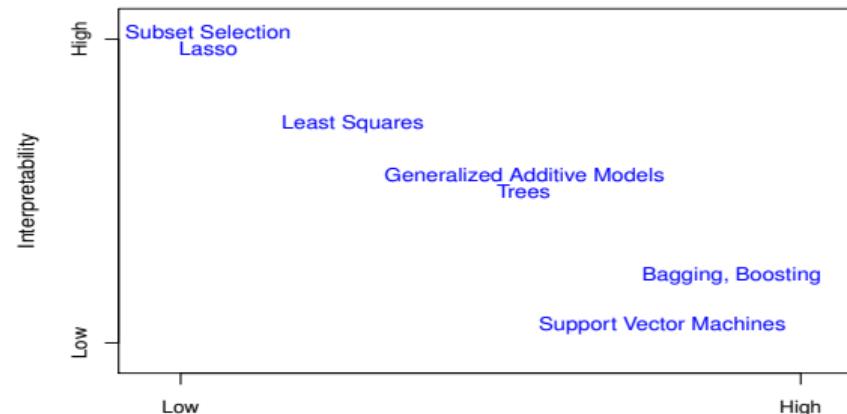
- ▶ Lasso(=软门限, soft-thresholding)估计: $\widehat{\beta}_{\text{lasso},j}(\lambda) = \text{sgn}(Z_j)(|Z_j| - \lambda)_+$, $\underbrace{Z_j}_{=\text{OLS}} = (n^{-1}\mathbf{X}^T\mathbf{Y})_j$.
- ▶ 硬门限(hard-thresholding)估计: $\widehat{\beta}_{\text{hard},j}(\lambda) = Z_j I(|Z_j| > \lambda)$.
- ▶ Adaptive Lasso估计: $\widehat{\beta}_{\text{lasso},j}(\lambda) = \text{sgn}(Z_j)(|Z_j| - \lambda/|Z_j|)_+$.
- ▶ SCAD估计:

$$\widehat{\beta}_{\text{SCAD},j}(\lambda) = \begin{cases} \text{sgn}(Z_j)(|Z_j| - \lambda)_+, & |Z_j| < 2\lambda, \\ \{(a-1)Z_j - \text{sgn}(Z_j)a\lambda\}/(a-2), & 2\lambda \leq |Z_j| \leq a\lambda, \\ Z_j, & |Z_j| > a\lambda. \end{cases}$$

- ▶ 岭回归估计: $\widehat{\beta}_{\text{ridge},j}(\lambda) = Z_j/(1 + \lambda)$.

总结

估计	无偏性	稀疏性	连续性	oracle性质
Lasso估计	✗	✓	✓	✗
硬门限估计	✓	✓	✗	✓
Adaptive Lasso估计	✓	✓	✓	✓
SCAD估计	✓	✓	✓	✓
岭回归估计	✗	✗	✓	✗



本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

- Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Society*, 6: 97–144.
- Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35: 2313–2351.
- Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360.
- Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4): 1947–1975.
- Frank, I. E. and Friedman, J. H. (1993). An statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35: 109–135.
- Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. London: Chapman and Hall.
- Hunter, D. and Li, R. Z. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33: 1617–1642.

- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, 15: 661–675.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58: 267–288.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942.
- Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101: 1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67: 301–320.
- Zou, H. and Li, R. Z. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, 36: 1509–1566.

本章纲要

1 子集选择

- 最优子集选择
- 逐步选择方法
- 最优模型选择

2 岭回归

3 桥回归

4 惩罚变量选择方法

- 惩罚函数
- Lasso方法
- SCAD方法
- 自适应Lasso
- 弹性网方法
- 模拟研究

5 参考文献

6 作业

作业

[习题见教材: 统计学习(R语言版) — 习题7]

- **课后思考题:** 第2题、第5题、第8题、第10题
- **需要完成的课后作业:** 第1题、第7题、第9题、第12题
- **应用:** 第14题、第16题. 具体要求:
 - ① 能使用R语言把数据读入, 并对数据中的每个变量进行了解;
 - ② 能用学过的一些统计方法, 按照题目要求, 利用R语言对数据进行一些简单的分析, 并思考数据分析的结果.



谢谢，请多提宝贵意见！