第31讲 变量选择

2020.6.10

过犹不及 –《论语·先进》

子贡问：“师与商也孰贤？”子曰：“师也过，商也不及。”曰：“然则师愈与？”子曰：“过犹不及。”

内容较多，除了Cp的推导之外，其它基本都是介绍性的，主要介绍基本思想（高亮）

变量选择则准则：折中RSS与模型复杂度(变量个数k)

线性模型中常用的变量选择准则：

- $C_k = \dfrac{\mathrm{RSS}_k}{\hat{\sigma}^2} - n + 2k$, RSS_k为k个自变量模型的残差平方和；
- $\mathrm{AIC}_k = n\log(\mathrm{RSS}_k) + 2k +$常数项;
- $\mathrm{BIC}_k = n\log(\mathrm{RSS}_k) + (\log n)k +$常数项。

Key:
惩罚模型复杂度 k

其它准则（已不常用）

- 修正的R^2(adjusted R - squared)：

$$\overline{\mathrm{R}}_k^2 = 1 - \frac{n-1}{n-k}(1 - R_k^2)$$

- 平均残差平方和 (Residual mean squares)：

$$\mathrm{RMS}_k = \mathrm{RSS}_k/(n\text{-}k)$$

C_p 准则

- 全模型： $\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1} = \overset{n\times k}{X_1}\ \overset{k\times 1}{\boldsymbol{\beta}_1} + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$， $\boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$,
  残差平方和记为 $\mathrm{RSS}_p = (n-p)\hat{\sigma}^2$
- 子模型： $\mathbf{y} = X_1\boldsymbol{\beta}_{1_{k\times 1}} + \boldsymbol{\delta}$， $\boldsymbol{\delta} \sim (0, \tau^2 I_n)$, 残差平方和 RSS_k。

上节课推论2说明了$\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$的预测误差 (对任何$\mathbf{x}_0 = X^{\mathsf{T}}\mathbf{a}$),

$$e(\tilde{y}_0) = \mathbf{a}^{\mathsf{T}} M(\tilde{\mathbf{y}})\mathbf{a} + \sigma^2$$

因为无法极小化矩阵$M(\tilde{\mathbf{y}})$,为了极小化预测误差，我们极小化

$$m(\tilde{\mathbf{y}}) = trM(\tilde{\mathbf{y}}) = E\,\|\tilde{\mathbf{y}} - X\boldsymbol{\beta}\|^2$$

具有最小$m(\tilde{\mathbf{y}})$的子模型，并不一定具有最小的预测误差，但至少对于完全的in-sample预测，$m(\tilde{\mathbf{y}})$最小 $\Leftrightarrow$ 预测误差最小(新数据也是X,预测误差 $= m(\tilde{\mathbf{y}}) + n\sigma^2$)。

由上节课引理4(2)知对于k变量子模型:

$$m_k = m(\tilde{\mathbf{y}}) = k\sigma^2 + \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2,$$

含有未知参数，将$\sigma^2, \boldsymbol{\beta}_2$在全模型下的LS估计代入得:

$$\hat{m}_k = k\hat{\sigma}^2 + \| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2,$$

我们知道 $E(\hat{\boldsymbol{\beta}}_2) = \boldsymbol{\beta}_2$, 但$E\left(\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2\right) = \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2 + (p-k)\sigma^2$

$E(x) = \theta$, 但$E(x^2) \neq \theta^2$

细节：因为$E(\hat{\boldsymbol{\beta}}_2) = \boldsymbol{\beta}_2, \mathrm{var}(\hat{\boldsymbol{\beta}}_2) = \sigma^2 (X_2^{\perp\mathsf{T}} X_2^{\perp})^{-1}$，所以

$$E(\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2) = E(\hat{\boldsymbol{\beta}}_2{}^{\mathsf{T}} X_2^{\perp\mathsf{T}} X_2^{\perp}\hat{\boldsymbol{\beta}}_2) = \boldsymbol{\beta}_2{}^{\mathsf{T}} X_2^{\perp\mathsf{T}} X_2^{\perp}\boldsymbol{\beta}_2 + \mathrm{tr}\left(X_2^{\perp\mathsf{T}} X_2^{\perp}\mathrm{var}(\hat{\boldsymbol{\beta}}_2)\right)$$

$$= \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2 + \sigma^2 \mathrm{tr}(I_{p-k}) = \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2 + (p-k)\sigma^2$$

若$E(\mathbf{x}) = \boldsymbol{\mu}$，$\mathrm{var}(\mathbf{x}) = \Sigma$，
则$E(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = \boldsymbol{\mu}^{\mathsf{T}} A\boldsymbol{\mu} + \mathrm{tr}(A\Sigma)$,

因此$\hat{m}_k - (p-k)\hat{\sigma}^2$才是$m_k$的无偏估计, 令

$$\tilde{m}_k \doteq \hat{m}_k - (p-k)\hat{\sigma}^2 = \| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 + 2k\hat{\sigma}^2 - p\hat{\sigma}^2$$

为了极小化m_k，我们极小化$\tilde{m}_k$.

另外，注意到$X_2^{\perp}\hat{\boldsymbol{\beta}}_2 = P_{X_2^{\perp}}\mathbf{y} = (P_X - P_{X_1})\mathbf{y} = \hat{\mathbf{y}} - \tilde{\mathbf{y}}$

所以$\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 = \| (\mathbf{y} - \tilde{\mathbf{y}}) - (\mathbf{y} - \hat{\mathbf{y}}) \|^2 = \| \mathbf{y} - \tilde{\mathbf{y}} \|^2 - \| \mathbf{y} - \hat{\mathbf{y}} \|^2$

$$= \mathrm{RSS}_k - \mathrm{RSS}_p = \mathrm{RSS}_k - (n-p)\hat{\sigma}^2$$

所以$\tilde{m}_k = \| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 + 2k\hat{\sigma}^2 - p\hat{\sigma}^2 = \mathrm{RSS}_k - (n-p)\hat{\sigma}^2 + 2k\hat{\sigma}^2 - p\hat{\sigma}^2$

$$= \mathrm{RSS}_k + 2k\hat{\sigma}^2 - n\hat{\sigma}^2$$

两边同除以$\hat{\sigma}^2$(不依赖于子模型)，即得到Mallow's $C_k = \tilde{m}_k / \hat{\sigma}^2 = \dfrac{\mathrm{RSS}_k}{\hat{\sigma}^2} + 2k - n$

Mallow's C_p准则： 在所有子模型中选择使得

$$C_k = \frac{\mathrm{RSS}_k}{\hat{\sigma}^2} - n + 2k,$$

最小的子模型, 其中$\hat{\sigma}^2$为全模型下的误差方差的LS估计。

注：以RSS或R^2作为变量选择标准是不恰当的, 因为

自变量个数k越多，RSS_k越小，R^2越大:

$$RSS_k = \min_{\beta_1} \| \mathbf{y} - X_1\boldsymbol{\beta}_1 \|^2 = \min_{\beta_1,\ \beta_2=0} \| \mathbf{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2 \|^2$$

$$\geq \min_{\beta_1,\ \beta_2} \| \mathbf{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2 \|^2 = RSS_p$$

$C_k = \dfrac{RSS_k}{\hat{\sigma}^2} - n + 2k$中$RSS_k \downarrow k$, 而最后一项是$k$的增函数，两者折中，

C_k可能在某个$0 \leq k < p-1$处达到最小。其它变量选择准则基本与此类似。

AIC、BIC准则

假设训练数据$y_1,...,y_n$服从某个概率模型$f(y|\boldsymbol{\theta})$，似然函数为

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

参数的极大似然估计为$\hat{\boldsymbol{\theta}}$。对于不同的模型 $f(y|\boldsymbol{\theta})$,选择*AIC/BIC*最小的模型：

AIC准则（Akaike Information Criterion, Hirotugu Akaike,1974)：

$$\text{AIC} = -2\log L(\hat{\theta}) + 2k$$

其中$L(\hat{\theta})$为似然函数极大值，k为参数个数。

BIC (Bayesian Information Criterion, Schwarz 1978)是一个类似的准则，

$$\text{BIC} = -2\log L(\hat{\boldsymbol{\theta}}) + (\log n)k$$

其中n是样本量，k是参数个数.

对于正态回归模型

$$\mathrm{AIC} = n\log(\mathrm{RSS}_k) + 2k + \text{常数项};$$

$$\mathrm{BIC} = n\log(\mathrm{RSS}_k) + (\log n)k + \text{常数项}。$$

- AIC，BIC通常称为模型选择准则，有时未必用来选变量 (比如选择Weibull分布、gamma、log-normal模型之一）。
- 为什么称作“信息”准则？与熵 $\mathrm{E}_f \log(f)$ 有关
- AIC，BIC是一般的模型选择(包括变量选择)准则, 适用于一般概率模型。
- 在正态回归情况下，AIC与C_p类似。
- AIC的推导与C_p类似，但使用Kullaback - Leibler距离度量预测误差。
- BIC是在Bayesian框架下得到的准则，相对于AIC或Cp，它倾向于选择参数个数更少的模型。

附录：推导AIC准则

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

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

我们希望预测$y^*\sim g(y^*)$, 使得$\hat{f}=f(y^*;\hat{\boldsymbol{\theta}})$与$g(y^*)$的距离(误差)尽量小。

两个概率密度函数f,g的Kullaback-Leibler距离:

$$K(g,f)=\int g\log\left(\frac{g}{f}\right)=\int g\log(g)-\int g\log(f)\geq 0$$

当$g=f$时，$K(g,f)=0$最小。

两个密度函数之间的距离

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

因为g实际上通常未知，我们需要估计$K=K(\hat{\boldsymbol{\theta}})$.

记对数似然函数 $\log L(\boldsymbol{\theta}) = \sum \log(f(y_i;\boldsymbol{\theta}),$ 注意到

$$E\left(\frac{\log L(\boldsymbol{\theta})}{n}\right) = E\left(\frac{1}{n}\sum \log(f(y_i;\boldsymbol{\theta})\right) = E_{y^*\sim g}\log(f(y^*;\boldsymbol{\theta})) = K(\boldsymbol{\theta})$$

所以可以用 $\frac{1}{n}\sum \log(f(y_i;\boldsymbol{\theta})$ 近似 $K(\boldsymbol{\theta})$，但如果代入参数的估计 $\hat{\boldsymbol{\theta}}$，那么将出现偏差：

$$E\left(\log L(\hat{\boldsymbol{\theta}})/n\right) = E\left(K(\hat{\boldsymbol{\theta}})\right) + k/n + o(1),$$

即 $E\left(\log L(\hat{\boldsymbol{\theta}})/n - k/n\right) = E\left(K(\hat{\boldsymbol{\theta}})\right) + o(1),$ 当 n 较大时，近似地有

$$\log L(\hat{\boldsymbol{\theta}})/n - k/n \approx K(\hat{\boldsymbol{\theta}})$$

为了极大化 $K(\hat{\boldsymbol{\theta}}),$ 我们极小化 $\log L(\hat{\boldsymbol{\theta}})/n - k/n,$ 乘以 $-2n$，即得到 AIC

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}) + 2k.$$

变量选择算法

回归分析中，使用某种准则，通常是AIC(或BIC)选择变量，搜索具有最小AIC的部分变量模型。常用算法有：

(1). 最优子集选择方法(best subset selection)

(2). 逐步回归法 (Stepwise regression):

- **向前逐步回归(forward stepwise regression),**
- **向后逐步回归(backward stepwise regression)**
- **向前-向后逐步回归(forward-backward)**

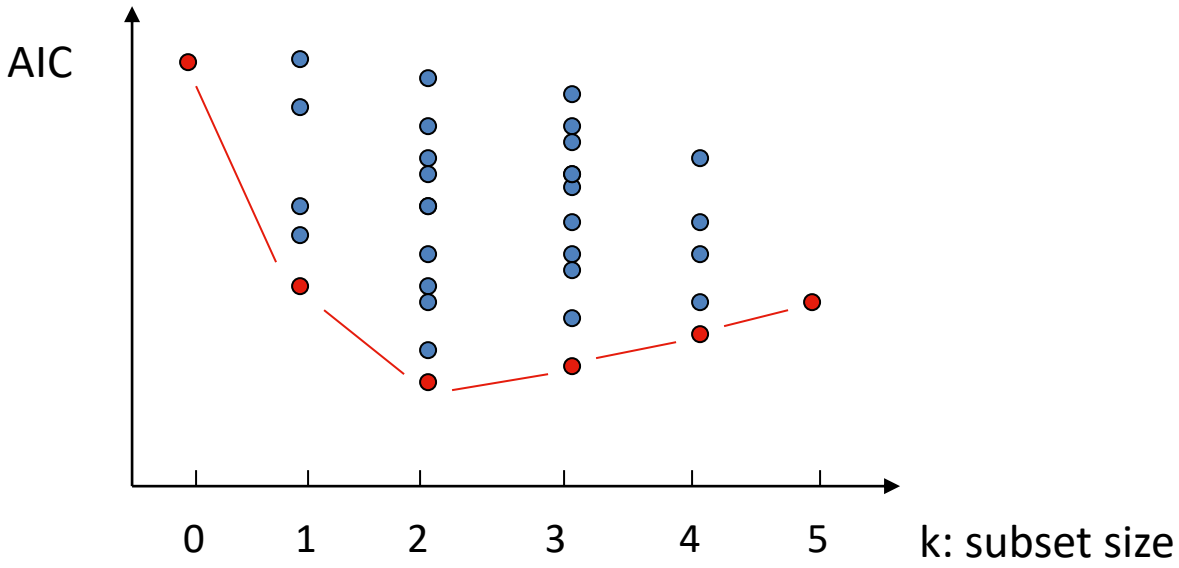
(3). 向前阶段回归（Forward stagewise regression）

1. 最优子集选择方法

$p-1$个自变量，共2^{p-1}个子集。以AIC为例

(1) 对$k=0,1,2,\ldots,p-1$,在所有C_{p-1}^{k}个k-自变量模型中找出RSS最小(等价地R^2最大或AIC最小)的模型.

(2) 比较这p个最优子集的AIC，选出AIC最小的模型。



最优子集选择方法的计算复杂度为2^{p-1},需要搜索所有可能的2^{p-1}个子模型，计算复杂度为$O(2^{p-1})$,这在p较大时不可行,如何降低计算复杂度？

(1) 近似解法：逐步回归方法是一种高效算法（计算复杂度为$O(p^2)$），但得到的解不一定是最优子集。

(2) *Leaps and bound* 算法部分地加快了最优子集选择方法，但复杂度仍是指数阶，不能处理p很大的情况。

(3) 列正交情形最优子集方法计算复杂度可降低到$O(p^2)$:

按投影长度$\|P_{\mathbf{x}_i}\mathbf{y}\|$从大到小逐次添加变量，最多$p$次即可完成。

$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+\mathbf{x}_1\beta_1+\ldots+\mathbf{x}_{p-1}\beta_{p-1}+\boldsymbol{\varepsilon}$, $\mathbf{1},\mathbf{x}_1,\ldots,\mathbf{x}_{p-1}$相互正交，则

$\hat{\mathbf{y}}=P_X\mathbf{y}=\mathbf{1}\bar{y}+P_{\mathbf{x}_1}\mathbf{y}+\ldots+P_{\mathbf{x}_{p-1}}\mathbf{y}, RSS=s_{yy}-a_1-\ldots.-a_{p-1},$

排列投影长度(计算复杂度$O(p^2)$):

$$a_1=\|P_{\mathbf{x}_1}\mathbf{y}\|^2\geq a_2=\|P_{\mathbf{x}_1}\mathbf{y}\|^2\geq\ldots.\geq a_{p-1}=\|P_{\mathbf{x}_{p-1}}\mathbf{y}\|^2,$$

显然所有$\binom{p-1}{k}$个k-自变量模型的最小$\mathrm{RSS}(k)=s_{yy}-a_1-\ldots.-a_k$。

所以只需要对$k=1,2,\ldots,p-1$,求$\mathrm{AIC}_k=n\log(\mathrm{RSS}(k))+2k$的最小值即可。

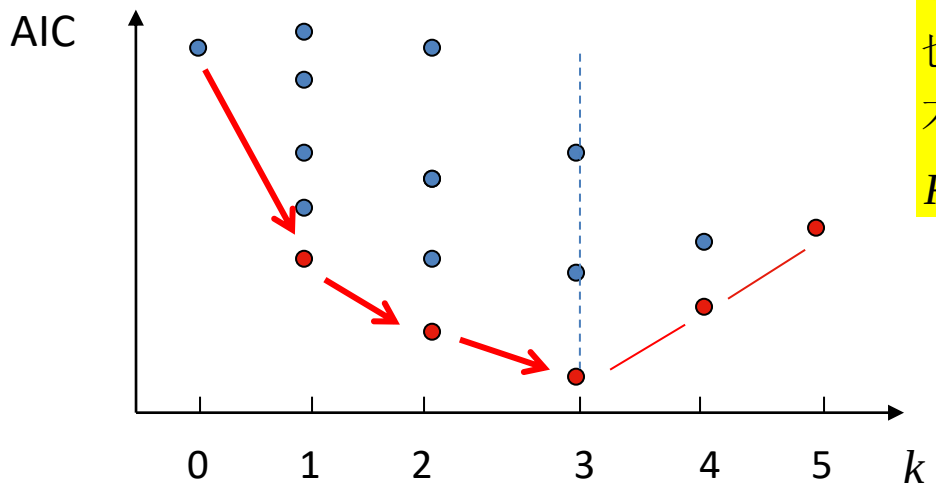
例子：主成分选择

2. 逐步回归方法

逐步回归方法是一种贪心算法(greedy algorithm),不需要搜索所有2^{p-1}个子模型,而是沿*AIC*跳跃最大的路径(故称为greedy)搜索若干子模型。考察的子模型个数最多为$O(p^2)$。逐步回归得到的解可能是最优子集，也可能不是。

向前法(Forward selection)

从0个自变量的回归模型开始,逐步添加变量：每次添加最能改进拟合程度的变量(即使得RSS减少最多的哪个变量),加入后如果*AIC*(或其它准则)变小，则选择该变量进入模型；如果AIC增大，停止.



和列正交情形下的最优子集类似，也是逐步添加变量，但变量之间不正交，可能会互相抵消。*RSS*没有单调性。

变量个数$p-1=5$

向后法(backward elimination)：

从全模型模型开始, 逐步剔除变量：

每次剔除对拟合影响最小的那个变量（即剔除后*RSS*增加最少）。如果剔除该变量使得模型的*AIC*(或其它准则)变小, 则重复上述步骤； 否则，停止.

注：向前或向后法的逐步选取的变量子集是嵌套的(*nested*，递增或递减),

向前法：一个变量一旦被选入就不会再被剔除。

向后法：一个变量一旦被剔除就不会再被选入。

向前–向后法：

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

```
>  step(full.model, method="both"，  k, scale=0，  scope=..)
#method:  both, backward, forward
# AIC: k=2：   BIC:  k=log(n);   Cp:  scale=sigma
```

附：向前阶段回归

Forward stagewise regression：
类似于向前逐步回归，依次添加与当前残差相关系数绝对值最大的自变量，直到相关系数小于某个给定的阈值。

其基本想法来源于简单回归。简单线性回归$y \sim x$的残差平方和$RSS = (1-r_{xy}^2)s_{yy}$
因此，选择使得RSS最小的变量等价于选择与y相关系数绝对值最大的变量。

Stagewise Regression algorithm算法细节：

假设响应 $\mathbf{y}$,自变量$\mathbf{x}_1,...,\mathbf{x}_p$都已经标准化。残差初值：$\mathbf{e}_0 = \mathbf{y}$

- 求与$\mathbf{e}_0$相关系数最大的自变量：$j_1 = \arg\max < \mathbf{e}_0, \mathbf{x}_j >$，回归：$\mathbf{e}_0 \sim \mathbf{x}_{j_1} \Rightarrow$ 残差$\mathbf{e}_1 = \mathbf{e}_0 - \mathbf{x}_{j_1}\hat{\beta}_{j_1}$
- 求与$\mathbf{e}_1$相关系数最大的自变量：$j_2 = \arg\max_{j \neq j_1} < \mathbf{e}_1, \mathbf{x}_j >$，回归：$\mathbf{e}_1 \sim \mathbf{x}_{j_2}$残差$\mathbf{e}_2 = \mathbf{e}_1 - \mathbf{x}_{j_2}\hat{\beta}_{j_2}$

…

若$\| \mathbf{e}_k \| < C$(事先指定的阈值), STOP.　　最终模型：$y = x_{j_1}\hat{\beta}_{j_1} + x_{j_2}\hat{\beta}_{j_2} + ... + x_{j_k}\hat{\beta}_{j_k}$

注：Stagewise regression 是匹配追踪(matching pursuit)的一种，匹配追踪在信号处理领域应用广泛。它不要求逆，每一步都是简单回归，可处理任意多自变量($p > n$)。

主成分回归

最优子集变量选择通常计算量很大，但设计阵列正交时容易实现。因此我们可以将列向量做Schmidt正交化，然后选择变换后的正交变量。困难在于如何正交化？一种策略是使用主成分，

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{假设}\mathbf{y}, X\text{已中心化}$$

假设X有奇异值分解$X_{n\times p} = U_{n\times p}D_{p\times p}V^{\mathsf{T}}{}_{p\times p}, U^{\mathsf{T}}U = V^{\mathsf{T}}V = VV^{\mathsf{T}} = I_p,$

$D = diag(\sqrt{\lambda_1},...,\sqrt{\lambda_p}),\quad UD = XV$各列称为主成分。

记$(n-1)S = X^{\mathsf{T}}X$的标准正交特征向量按列排列成正交矩阵V,

$$X^{\mathsf{T}}X = VD^2V^{\mathsf{T}},\ D^2 = diag(\lambda_1,...,\lambda_{p-1}),\ V^{\mathsf{T}}V = VV^{\mathsf{T}} = I_{p-1}$$

$B = XV$称为主成分，其各列正交，这是因为$B^{\mathsf{T}}B = V^{\mathsf{T}}Z^{\mathsf{T}}ZV = D^2,$

令$BD^{-1} = U,$则$U^{\mathsf{T}}U = I_{p-1},$因此$B = XV = UD \Rightarrow X = UDV^{\mathsf{T}}$

U重新表示了X的列,
V重新表示了X的行。
它们构成了X行,列的"特征向量":
$X\mathbf{v}_k = \mathbf{u}_k\sqrt{\lambda_k}, X^{\mathsf{T}}\mathbf{u}_k = \mathbf{v}_k\sqrt{\lambda_k}$

将 $X = UDV^{\mathsf{T}}$代入模型 (假设$\lambda_1 \geq ... \geq \lambda_p > 0$),模型等价表示为：

$$\mathbf{y} = UDV^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \overset{\text{记为}}{=} U\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

其中新的设计阵为U： 各列正交，$L(X)=L(U)$。

新的回归系数$\boldsymbol{\gamma}=DV^{\top}\boldsymbol{\beta}$： $\boldsymbol{\gamma}$与$\boldsymbol{\beta}$一一对应（因为$DV^{\top}$可逆）

也可表示为：$\mathbf{y}=UDV^{\top}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\overset{\text{记为}}{=}Z\boldsymbol{\eta}+\boldsymbol{\varepsilon}$，其中$Z=UD=XV$为主成分

两个改写的模型都可以叫做主成分回归，它们等价。

但第一个模型的设计阵U是单位化的，较易处理。

在模型表示$\mathbf{y}=U\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$下,设计阵$U$列正交，$LS$拟合结果有简单表示：

- 设$U=(\mathbf{u}_1,...,\mathbf{u}_p)$，则 $P_X=P_U=U(U^{\top}U)^{-1}U^{\top}=UU^{\top}=\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j^{\top}$
- $\boldsymbol{\gamma}$的LS估计$\hat{\boldsymbol{\gamma}}=(U^{\top}U)^{-1}U^{\top}\mathbf{y}=U^{\top}\mathbf{y}$，$\hat{\gamma}_j=\mathbf{u}_j^{\top}\mathbf{y}$
- $\mathbf{y}$在X上的投影可以表示为(其中$\hat{\gamma}_j=\mathbf{u}_j^{\top}\mathbf{y}$为投影坐标,$\gamma_j$的LS估计)：

$$\hat{\mathbf{y}}=P_X\mathbf{y}=UU^{\top}\mathbf{y}=\sum_{j=1}^{p}\mathbf{u}_j\left\{\mathbf{u}_j^{\top}\mathbf{y}\right\}=\sum_{j=1}^{p}\mathbf{u}_j\hat{\gamma}_j$$

- $\|\hat{\mathbf{y}}\|^2=\sum_{j=1}^{p}\hat{\gamma}_j^{\,2}$, $RSS=\|\mathbf{y}\|^2-\|\hat{\mathbf{y}}\|^2=\|\mathbf{y}\|^2-\sum_{j=1}^{p}\hat{\gamma}_j^{\,2}$
- $DV^{\top}\boldsymbol{\beta}=\boldsymbol{\gamma}\Rightarrow\hat{\boldsymbol{\beta}}=VD^{-1}\hat{\boldsymbol{\gamma}}=VD^{-1}U^{\top}\mathbf{y}=(X^{\top}X)^{-1}X^{\top}\mathbf{y}$

传统的主成分回归

传统上，主成分回归选取U的前k列用来预测(前k个方差最大的主成分,方差分别为$\lambda_1 \geq \ldots \geq \lambda_k$)，$k$的取值由累计解释的方差比例 $(\lambda_1 + \ldots + \lambda_k)/(\lambda_1 + \ldots + \lambda_p) > 0.8$决定。

传统的主成分回归使用最大的k个主成分用于预测：

$$\tilde{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k, 0, \ldots 0)^{\mathsf{T}}, \quad \hat{\gamma}_j = \mathbf{u}_j{}^{\mathsf{T}}\mathbf{y}$$

$$\tilde{\mathbf{y}}^{(\mathrm{pc})} = U\tilde{\boldsymbol{\gamma}} = \sum_{j=1}^{k} \mathbf{u}_j \hat{\gamma}_j = \sum_{j=1}^{k} \mathbf{u}_j \left\{ \mathbf{u}_j{}^{\mathsf{T}}\mathbf{y} \right\} = \mathrm{P}_{(\mathbf{u}_1, \ldots, \mathbf{u}_k)}\mathbf{y}$$

即截取 $\hat{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{u}_j \hat{\gamma}_j$ 的前k项，以$L(\mathbf{u}_1, \ldots, \mathbf{u}_k)$逼近$L(\mathbf{u}_1, \ldots, \mathbf{u}_p) = L(X)$。

由引理4MSE公式 $m(\tilde{\mathbf{y}}^{(\mathrm{pc})}) = k\sigma^2 + \sum_{j=k+1}^{p} \gamma_j^2$看出，MSE其实与被选择的$\lambda_1 \geq \ldots \geq \lambda_k$

或被舍弃的$\lambda_{k+1} \geq \ldots \geq \lambda_p$无关，而是与被舍弃的那些$U$的列对应的效应$\gamma_j$有关。

只有当$\gamma_{k+1}^2, \ldots, \gamma_p^2$是$\{\gamma_j^2, 1 \leq j \leq p\}$中最小的$k-p$个时，$\tilde{\mathbf{y}}^{(\mathrm{pc})}$的MSE才最小。

所以，传统的主成分回归除了选取的主成分容易解释之外，在预测方面并没有优势.

主成分选择(最优子集)

假设$S \subseteq \{1,2,\ldots,p\}, |S| = k$, 取$\boldsymbol{\gamma}$的估计 $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1,\ldots,\tilde{\gamma}_p)^{\mathsf{T}}$, 其中

$$\tilde{\gamma}_j = \begin{cases} \hat{\gamma}_j = \mathbf{u}_j^{\mathsf{T}}\mathbf{y}, & \text{若} j \in S \\ 0 & \text{若} j \notin S \end{cases}$$

令 $\tilde{\mathbf{y}}^{(S)} = U\tilde{\boldsymbol{\gamma}} = \sum_{j\in S}\mathbf{u}_j\hat{\gamma}_j = \sum_{j\in S}\mathbf{u}_j\{\mathbf{u}_j^{\mathsf{T}}\mathbf{y}\}$, 由引理4, 可得

命题1. $S \subseteq \{1,2,\ldots,p\}, |S| = k$

(1) $\tilde{\mathbf{y}}^{(S)}$的均方误差$m(\tilde{\mathbf{y}}^{(S)}) = k\sigma^2 + \sum_{j\notin S}\gamma_j^2$

(2) 当$\sum_{j\notin S}\gamma_j^2 \le (p-k)\sigma^2$时, $m(\tilde{\mathbf{y}}^{(S)}) \le m(\hat{\mathbf{y}}) = p\sigma^2$, $\hat{\mathbf{y}}$为OLS拟合。

(3) $\hat{\sigma}^2 C_p = \mathrm{RSS}^{(S)} + 2k\hat{\sigma}^2 - n\hat{\sigma}^2$是$m(\tilde{\mathbf{y}}^{(S)})$的无偏估计.

C_p准则$= \mathrm{RSS}^{(S)} / \hat{\sigma}^2 + 2k - n = \sum_{j \in S} \hat{\gamma}_j^2 / \hat{\sigma}^2 + 2k - n.$

如前所述(p15)，$\mathrm{RSS}^{(S)}$可分拆为$\hat{\gamma}_j^2$的和，最优子集算法可快速完成：

最优主成分选择算法：

1. 将$|\hat{\gamma}_j| = |\mathbf{u}_j^{\mathsf{T}} \mathbf{y}|$排序，不妨设$|\hat{\gamma}_2| \geq |\hat{\gamma}_2| \geq .... \geq |\hat{\gamma}_p|$

2. $k = \arg\min_{k=0,1,...,p} \left(\sum_{j=k+1}^{p} \hat{\gamma}_j^2 / \hat{\sigma}^2 + 2k - n \right)$

计算复杂度为$O(p^2)$。

注1：事实上，将X转换成任意的$Z = XA$（A可逆）：

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = XAA^{-1}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \overset{\text{记为}}{=} Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon}，\ \boldsymbol{\gamma} = A^{-1}\boldsymbol{\beta}$$

注1：部分最小二乘（Partial LS）是另一种试图将主成分的选取与响应变量关联起来的方法。

L2惩罚最小二乘： 岭回归

岭估计的最初提出是考虑解决自变量复共线性时，LS估计的方差过大的现象。

复共线性：条件数$\lambda_{\max}(X^{\mathsf{T}}X)/\lambda_{\min}(X^{\mathsf{T}}X)$过大，即$X^{\mathsf{T}}X$不可逆或接近不可逆.

模型：$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim(0,\sigma^2 I_n)$。定义岭估计(ridge estimator)

$$\tilde{\boldsymbol{\beta}}^{(\text{Ridge})}=(X^{\mathsf{T}}X+\lambda I_p)^{-1}X^{\mathsf{T}}\mathbf{y}\ ,$$

其中$\lambda>0$为常数（Hoerl and Kennard，1970）

为什么称为岭(ridge)？当存在严重复共线性时，误差平方和的二阶导数$-X^{\mathsf{T}}X$接近不可逆，$\text{var}(\hat{\boldsymbol{\beta}}_{\text{LS}})=\sigma^2(X^{\mathsf{T}}X)^{-1}$过大，似然函数或负的误差平方和在最高点$\hat{\boldsymbol{\beta}}_{\text{LS}}$附近不是一个山顶而像一个等高的山脊(*ridge*)，故最优点不唯一或不稳定.

后来人们常从压缩估计或惩罚最小二乘的角度看待岭估计，

容易看出$\tilde{\boldsymbol{\beta}}^{(\text{Ridge})}=(X^{\mathsf{T}}X+\lambda I_p)^{-1}X^{\mathsf{T}}\mathbf{y}$是如下约束问题的解：

$$\tilde{\boldsymbol{\beta}}^{(\text{Ridge})}=\text{argmin}\left\{\|\mathbf{y}-X\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|^2\right\}$$
$$\Leftrightarrow \tilde{\boldsymbol{\beta}}^{(\text{Ridge})}=\text{argmin}\|\mathbf{y}-X\boldsymbol{\beta}\|^2,\quad \text{约束}\|\boldsymbol{\beta}\|<\text{t}$$

所以岭估计是惩罚（penalized）最小二乘估计或规则化（regularized）LS估计。因为对$\boldsymbol{\beta}$的模长进行了约束，故岭估计必是压缩估计。

命题2.

(a) 岭估计是压缩估计：$\|\tilde{\boldsymbol{\beta}}^{(\text{Ridge})}\|<\|\hat{\boldsymbol{\beta}}\|$，$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$

(b) 存在$\lambda>0$，使得$m(\tilde{\boldsymbol{\beta}}^{(\text{Ridge})})\le m(\hat{\boldsymbol{\beta}})$，

(c) 存在$\lambda>0$，使得$m\left(\tilde{\mathbf{y}}=X\tilde{\boldsymbol{\beta}}^{(\text{Ridge})}\right)\le m\left(\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}}\right)$，

即基于某些岭估计比基于LS估计的预测误差更小。

证明(c)：对X进行奇异值分解：$X = UDV^{\mathsf{T}}$，其中$D = \mathrm{diag}\left(\sqrt{\lambda_1},...,\sqrt{\lambda_p}\right)$,

$U^{\mathsf{T}}U = I_p, V^{\mathsf{T}}V = I_p$，　记$U = (\mathbf{u}_1,...,\mathbf{u}_p)$,

岭估计对应的拟合值：$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}^{(\mathrm{Ridge})} = X(X^{\mathsf{T}}X + \lambda I_p)^{-1}X^{\mathsf{T}}\mathbf{y}$

$$= UD(D^2 + \lambda I_p)^{-1}DU^{\mathsf{T}}\mathbf{y} = \sum_{j=1}^{p}\mathbf{u}_j\left\{\frac{\lambda_j}{\lambda_j + \lambda}\mathbf{u}_j{}^{\mathsf{T}}\mathbf{y}\right\}$$

而LS拟合值：$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = P_U\mathbf{y} = \sum_{j=1}^{p}\mathbf{u}_j\left\{\mathbf{u}_j{}^{\mathsf{T}}\mathbf{y}\right\}$，$m(\hat{\mathbf{y}}) = p\sigma^2$,

$$\mathrm{MSE}: \quad m(\tilde{\mathbf{y}}) = E\,\|\tilde{\mathbf{y}} - X\boldsymbol{\beta}\|^2 = \mathrm{tr}\left(\mathrm{var}(\tilde{\mathbf{y}})\right) + \|\,\mathrm{bias}(\tilde{\mathbf{y}})\,\|^2$$

$$= \sigma^2\sum_{j=1}^{p}\left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2 + \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}\sum_{j=1}^{p}\left(\frac{\lambda}{\lambda_j + \lambda}\right)^2\mathbf{u}_j\mathbf{u}_j{}^{\mathsf{T}}X\boldsymbol{\beta}$$

可以验证，一定存在$\lambda = \lambda(\boldsymbol{\beta}, \sigma^2) > 0$，使得$m(\tilde{\mathbf{y}}) \le m(\hat{\mathbf{y}})$。证毕。

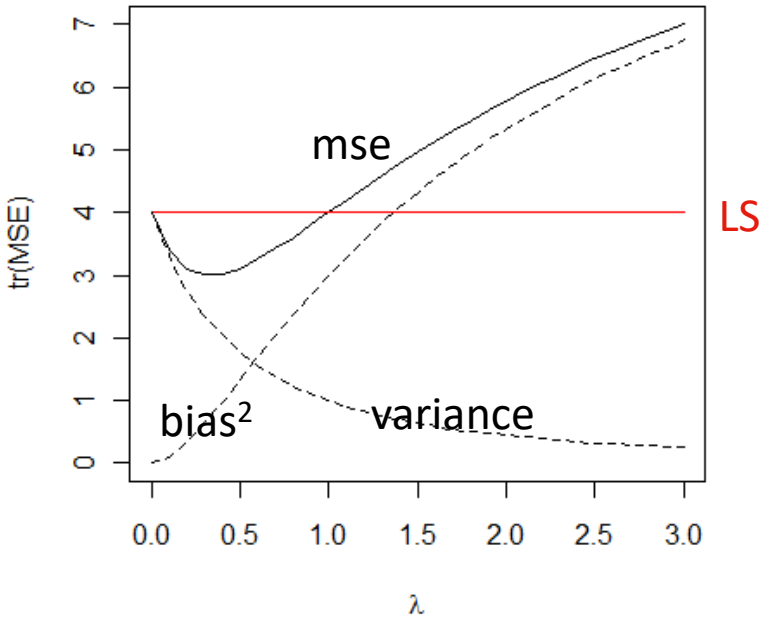
岭回归与James-Stein估计

对于简单情况：$X^{\mathsf{T}}X = I_p$，岭估计具有简单的MSE：

推论1. 假设$X^{\mathsf{T}}X = I_p$,则岭估计：$\tilde{\boldsymbol{\beta}}^{(\text{ridge})} = \hat{\boldsymbol{\beta}}/(1+\lambda)$，是LS估计$\hat{\boldsymbol{\beta}}$的压缩估计！且

$$m(\tilde{\boldsymbol{\beta}}^{(\text{ridge})}) = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} \quad \text{（命题2）}$$

$\lambda\uparrow$时，第一项为variance$\downarrow$,第二项bias$^2\uparrow$，使得上式达到最小的$\lambda_{\text{optimal}} = \frac{p\sigma^2}{\|\boldsymbol{\beta}\|^2}$



LS

plug-in LS 估计，得“最优”岭估计：

$$\tilde{\boldsymbol{\beta}}^{\text{ridge}}(\hat{\lambda}_{\text{optimal}}) = \frac{\hat{\boldsymbol{\beta}}}{1+\frac{p\hat{\sigma}^2}{\|\hat{\boldsymbol{\beta}}\|^2}} \approx \left(1-\frac{(p-2)\hat{\sigma}^2}{\|\hat{\boldsymbol{\beta}}\|^2}\right)\hat{\boldsymbol{\beta}},$$

即James–Stein估计。

统一LS、岭回归、主成分回归

最后，注意到LS，PC回归和岭回归的预测/拟合值可统一如下，它们给予投影坐标以不同的压缩比例 $\lambda_j/(\lambda_j + c_j)$：

$$\tilde{\mathbf{y}}^{(c)} = \sum_{j=1}^{p} \mathbf{u}_j \left\{ \frac{\lambda_j}{\lambda_j + c_j} \mathbf{u}_j^{\mathsf{T}} \mathbf{y} \right\}$$

- 最小二乘：$c_j \equiv 0$
- 岭回归：$c_j \equiv \lambda$
- PC回归：$c_j = \begin{cases} 0, & j \le k \\ \infty, & j > k \end{cases}$

附录：L1惩罚最小二乘 - LASSO

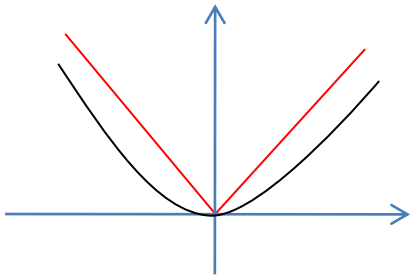
Least Absolute Shrinkage and Selection Operator (Tibshirani,1996)

$$\text{LASSO估计：}\quad \tilde{\boldsymbol{\beta}}^{(\text{lasso})} = \arg\min\left\{ \frac{1}{2}\| \mathbf{y} - X\boldsymbol{\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|_1 \right\}$$

$$\Leftrightarrow \tilde{\boldsymbol{\beta}}^{(\text{lasso})} = \arg\min \| \mathbf{y} - X\boldsymbol{\beta} \|^2, \text{约束} \| \boldsymbol{\beta} \|_1 < t ,$$

其中L_1模$\| \mathrm{u} \|_1 = \sum | u_i |$，$\lambda$与$t$之间1:1对应。

L1（绝对值）与L2（岭估计）的差异



绝对值函数尖锐

LASSO方法把一些回归系数估计为0，被认为是一种变量选择方法。

考虑一维 情况, $\beta \in R^1$, 误差平方和

$$f(\beta) = \| \mathbf{y} - \mathbf{x}\beta \|^2 = \mathbf{x}^{\mathsf{T}}\mathbf{x}\beta^2 - 2\mathbf{x}^{\mathsf{T}}\mathbf{y}\beta + \mathbf{y}^{\mathsf{T}}\mathbf{y} \triangleq a\beta^2 - 2b\beta + c$$

其中$a = \mathbf{x}^{\mathsf{T}}\mathbf{x}, \quad b = \mathbf{x}^{\mathsf{T}}\mathbf{y}.$

	目标函数	一阶导	最优解	二阶导
LS	$f(\beta)/2$	$a\beta - b$	b/a	$a = \| \mathbf{x} \|^2$
$L2$:岭回归	$f(\beta)/2 + \lambda\beta^2$	$(a+\lambda)\beta - b$	$b/(a+\lambda)$	$a + \lambda > a$
$L1$: lasso	$f(\beta)/2 + \lambda \lvert \beta \rvert$	$a\beta - b + \lambda \operatorname{sgn}(\beta)$	$\operatorname{sgn}(b)\left(\left\lvert \frac{b}{a} \right\rvert - \lambda\right)_+$	$a + \delta(0) >> a$

二阶导数(Hessian矩阵，信息、能量)越大，函数越陡峭，优化越容易：

- 岭回归的二阶导数$a + \lambda$大于LS的二阶导数$a = \mathbf{x}^{\mathsf{T}}\mathbf{x}$;
- LASSO回归的二阶导数在$\beta = 0$处无穷大。

Lasso的本质在于极小化二次函数： $a\beta^2 - 2b\beta + \lambda \lvert \beta \rvert + c$
该函数在0点处二阶导无穷。

LS的目标函数 $\|\mathbf{y}-X\boldsymbol{\beta}\|^2=\|\mathbf{y}-X\hat{\boldsymbol{\beta}}\|^2+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathsf{T}}X^{\mathsf{T}}X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$,

$\min\|\mathbf{y}-X\boldsymbol{\beta}\|^2\Leftrightarrow\min(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathsf{T}}X^{\mathsf{T}}X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$,

下图蓝色区域分别是$L1$(左图)和$L2$(右图)约束区域。$\hat{\boldsymbol{\beta}}$是LS估计。

红色椭圆为目标函数$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathsf{T}}X^{\mathsf{T}}X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$的等高线。

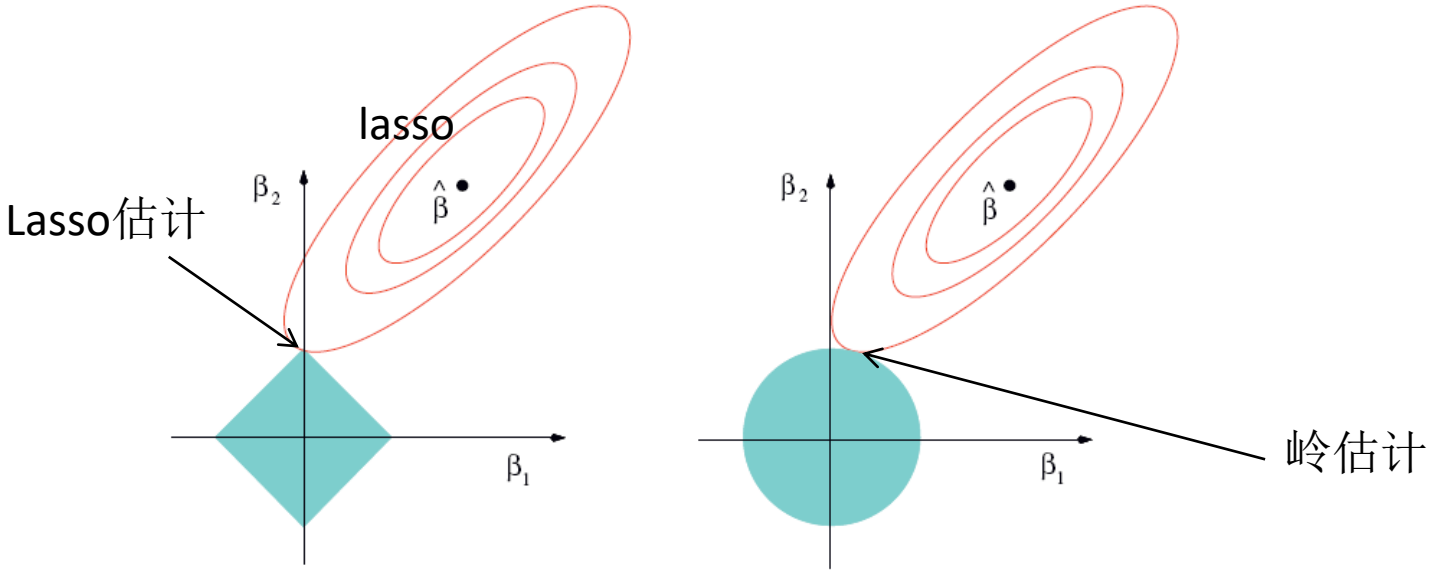


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1|+|\beta_2|\le t$ *and* $\beta_1^2+\beta_2^2\le t^2$, *respectively,*

椭圆与约束区域的切点满足约束，且使得目标函数达到最小。

（左图：尖点更容易与椭球碰到，此时$\beta_1=0$）。

对于X列正交的情形，LASSO估计有显式表达：

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

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

显然，LASSO得到的估计是对LS估计的压缩：

当$|\hat{\beta}_j|$较小时，把它压缩为0

当$|\hat{\beta}_j|$较大时，向0方向平移λ

注1. LASSO对回归系数的L1模进行约束，所以是压缩估计。

注2. 通常只对回归系数（不包括截距）进行约束，故X，y首先需要中心化；

注3. LASSO的常用解法： **least-angle regression (LARS) , 坐标下降法，…**

算法预备知识：坐标下降法（Corrdinate Descent Algorithm）

目标：$\min f(\mathbf{x}) = \min f(x_1,...,x_p)$

坐标下降法：依次给定其它坐标的条件下，对其中一个分量优化。

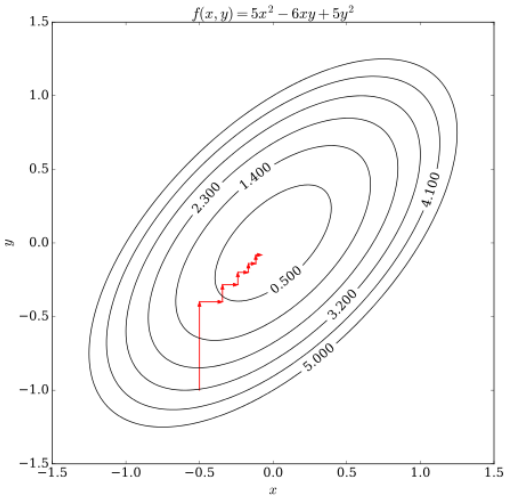
$k=0$, 初始化$\mathbf{x}^{(0)}$

$k = k+1$

For $i = 1,2,...,p$

$$x_i^{(k+1)} = \arg\min_{t \in R^1} f(x_1^{(k+1)},...,x_{i-1}^{(k+1)},\ t,\ x_{i+1}^{(k)},....,x_p^{(k)})$$

至收敛



坐标下降法求LASSO估计

不妨设设计阵X已经中心化且标准化（模为1）。

假设除了β_j之外的其它回归系数$\boldsymbol{\beta}_{(-j)} = (\beta_1^{(k+1)},\ldots,\beta_{j-1}^{(k+1)},\ \beta_{j+1}^{(k)},\ldots\beta_p^{(k)})^T$给定.

记设计阵X的第j列为$\mathbf{x}_j$，其它列为$X_{(-j)}$，则目标函数为β_j的二次函数：

$$R(\beta_j) = \frac{1}{2}\|\mathbf{y} - X_{(-j)}\boldsymbol{\beta}_{(-j)} - \mathbf{x}_j\beta_j\|^2 + \lambda|\boldsymbol{\beta}_{(-j)}|_1 + \lambda|\beta_j| \hat{=} \frac{1}{2}\|\mathbf{y}^* - \mathbf{x}_j\beta_j\|^2 + \lambda|\beta_{(-j)}|_1 + \lambda|\beta_j|$$

$$= \frac{1}{2}\beta_j^2 - \mathbf{x}_j^{\mathsf{T}}\mathbf{y}^*\beta_j + \lambda|\beta_j| + C.$$

其中$\mathbf{y}^* = \mathbf{y} - X_{(-j)}\boldsymbol{\beta}_{(-j)}$"已知",常数$C = \frac{1}{2}\|\mathbf{y}^*\|^2 + \lambda|\boldsymbol{\beta}_{(-j)}|_1$.

记$b = \mathbf{x}_j^{\mathsf{T}}\mathbf{y}^* = (\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j)^{-1}\mathbf{x}_j^{\mathsf{T}}\mathbf{y}^*$，则使得$R(\beta_j)$达到最小的最优解：

$$\beta_j^{(k+1)} = \underset{\beta_j}{\operatorname{argmin}}\ R(\beta_j) = \operatorname{sgn}(b)(|b| - \lambda)_+ = \begin{cases} b - \lambda & b > \lambda, \\ 0 & |b| \le \lambda, \\ b + \lambda & b < -\lambda, \end{cases}$$

其中u_+为u的正部，$\operatorname{sgn}(b)$为符号函数.

对所有$j = 1,\ldots, p$进行上述过程。重复，直到收敛。

$$\operatorname{sgn}(b) = \begin{cases} 1 & b > 0 \\ 0 & b = 0, \\ -1 & b < 0 \end{cases} \qquad u_+ = \begin{cases} u & u > 0 \\ 0 & u \le 0 \end{cases}$$

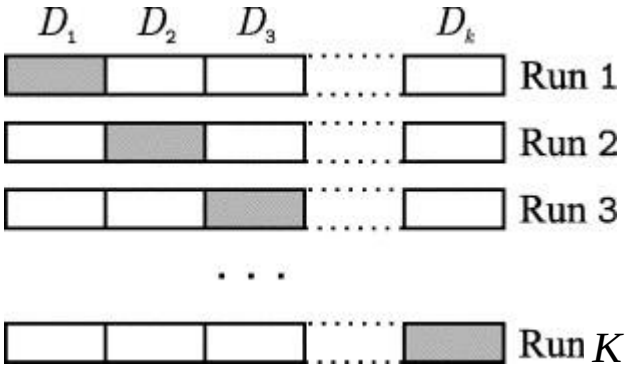
交叉验证(Cross-validation，CV)

C_p/AIC/BIC准则都是基于现有数据对预测误差或MSE的In-sample估计。缺陷：有效性的前提是假设已知或可以估计真模型。

通行的而且更可靠的方法是使用新数据（out-of-sample）测试模型性能、估计预测误差。模仿建模-预测的过程，将数据分为两部分：

- 训练样本(training sample)：用来建立模型和预测方法
- 检验/测试样本(testing sample): 用来评价训练得到的预测或模型 。

训练 (training)	测试/验证 (testing/validation)



K - fold Cross - Validation :

(1) 将数据点$\{1,2,....,n\}$随机划分为K组$D_1,...,D_K$(每组样本量$\approx n/K$),

(2) 对$k=1,2,...,K$,

(a) 使用除了D_k之外的样本拟合模型,

(b) 应用该模型预测D_k样本的响应变量,得到预测误差$\mathrm{CV}(k)$,

其中 $\mathrm{CV}(k)=\sum_{i\in D_k}(y_i-\hat{y}_i)^2/n_k$

y_i 表示D_k中第i个样本点的响应变量； $\hat{y}_i$ 表示其预测。

(3) $\mathrm{CV}=\sum_{k=1}^{K}\mathrm{CV}(k)/K$

总结

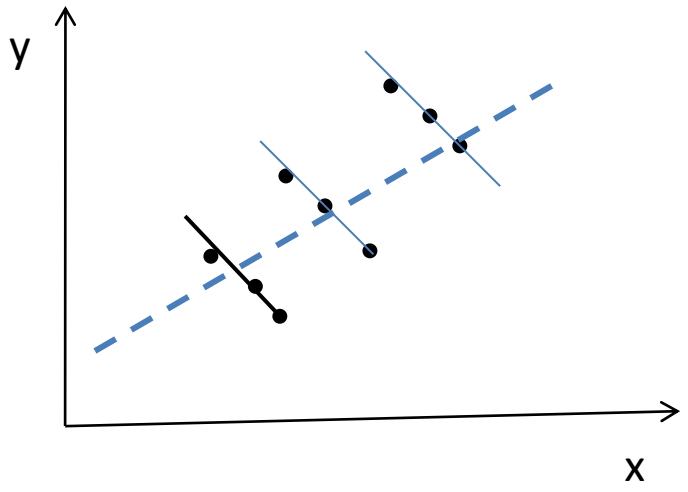
随机化	随机化	分组随机化	观察研究
两正态总体	→ 多正态总体	→ 分层/区组的多个总体	→ 连续控制变量
两样本t检验	→ 单因素方差分析,F	→ 两因素方差分析,F/成对t	→ 回归分析,F / t

- 关联与因果（相关与条件/偏相关）
  - 考虑相关性的时候变量之间对称，但回归分析有响应与自变量之分
  - 相关系数与简单线性模型：回归系数与相关系数成正比；t检验统计量 $t^2 \approx nr^2$
  - 偏相关系数与多重线性模型：控制变量

我们未提到的

Simpson's paradox:
控制干扰因素与不控制干扰因素时相比，x与y的相关性可能不同。

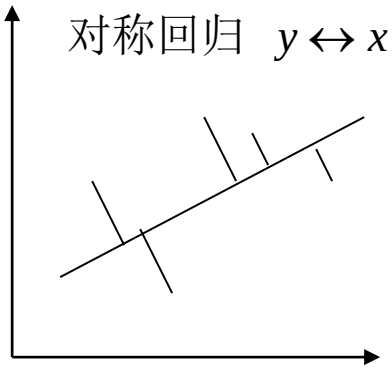
工具变量法试图从观察数据推断因多。



- 回归函数$E(y|x)=\alpha+\beta^T x$
  - 以x的线性函数逼近响应y，从而产生”回归现象(regress to the mean)”。
  - 所有的概率、期望计算都在给定自变量x的条件下进行。

我们未提到的

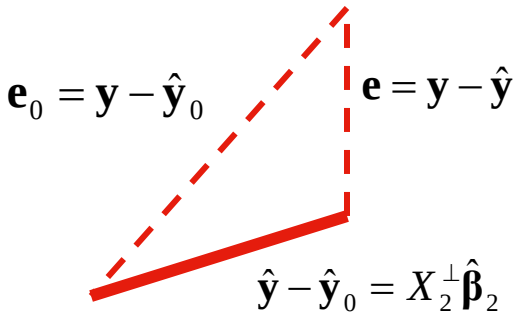
对称回归对称地对待y与x（单个自变量情形极小化图示的距离平方和）；error-in-variable模型考虑自变量的随机性。
优化方法为total least squares。



- 最小二乘(LS)
  - LS即正交投影，控制变量即Gram-Schmidt正交化；
  - LS估计无偏，最优（Gauss-Markov定理）

- 拟合优度 R^2
  - 拟合值/回归函数的方差在总方差中的百分比
  - 响应变量与拟合值(自变量的最优线性组合) 的相关系数

- F检验
  - 正态分布假设
  - F检验比较原假设成立与不成立下的投影/拟合值差别 $\hat{\mathbf{y}}-\hat{\mathbf{y}}_0$



$$F=\frac{n-p}{k}\times\frac{\|\hat{\mathbf{y}}-\hat{\mathbf{y}}_0\|^2}{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}$$

- 方差分析
  - 所有自变量为因子时的模型和检验
  - F检验可从平方和分解简单地得到。

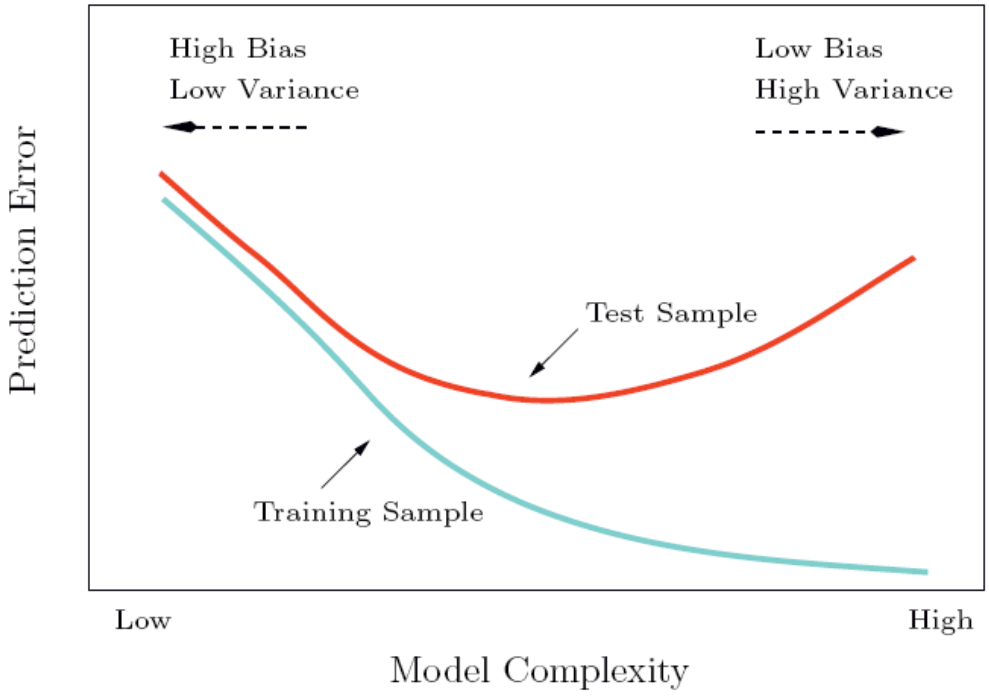
- 回归诊断：

  残差分析和影响分析，残差分析判断模型假设的合理性，影响分析检测对回归影响大的数据点。

我们未提到的

纵向数据/重复测量数据，每个个体重复测量，引入随机效应解释重复测量之间的相关性。由此得到的线性模型是方差不齐的，误差方差通常是分块对角阵$G = \mathrm{diag}(\Sigma, \ldots, \Sigma) = I \otimes \Sigma$ (称作Kronecker乘积）

- 预测： 避免过度复杂的模型/过度拟合（over-fitting）
  - 预测误差 = 均方误差 + 被预测变量的方差
  - 均方误差 = 方差 + bias^2
  - 估计不必无偏，变量之间的关系不必是因果关系，关联即可预测
  - Cp准则选变量，主成分回归选主成分



因为有些课件在课后有改动，建议重新下载所有课件。

考试内容以课件(1-31)、作业(1-9)为准，排除附录内容、虚线框内的内容、以及如下内容：

带交互作用的两因素方差分析模型
Tukey同时置信区间，多重检验
GLS在非线性优化，GLM中的应用
James-Stein估计
AIC推导
LASSO