第十七讲 预测

2022.12.23

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。”

1. 预测误差与均方误差

预测、预言出于人类对未知和不确定性的恐惧或好奇。

预测：基于历史数据对未发生的随机变量进行“估计”。

具体到线性模型，预测就是基于响应变量和自变量的历史数据，建立恰当的线性模型，对仅含自变量的新数据预测其对应的响应。

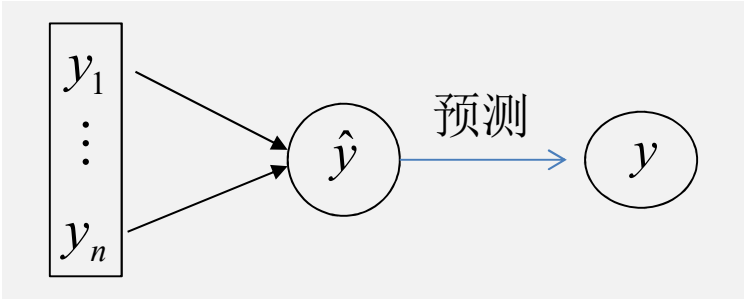
预测的一般原则

- 可泛化（generalization）：可推广，适用于不同场景
- 简约原则（Occam’s Razor，Occam剃刀原则)：若无必要,勿增实体　Entities should not be multiplied unnecessarily
- 模型不必正确，预测量不必无偏。

历史数据/样本： $y_1,\ldots,y_n$

待预测随机变量：y（与y_i's独立）

预测统计量：$\hat{y}=f(y_1,\ldots,y_n)$



预测误差

定义. 预测误差(prediction error, expected generalization error):

$$\mathrm{pe}(\hat{y})=E(\hat{y}-y)^2$$

向量情形：$\mathrm{pe}(\hat{\mathbf{y}})=E\|\hat{\mathbf{y}}-\mathbf{y}\|^2=\mathrm{E}(\hat{\mathbf{y}}-\mathbf{y})^{\mathrm{T}}(\hat{\mathbf{y}}-\mathbf{y})$

均方误差

定义：统计量 $\hat{y}$ 估计参数θ的均方误差（MSE: mean squared error）

$$\mathrm{mse}(\hat{y})=E(\hat{y}-\theta)^2$$

向量情形：$\mathrm{mse}(\hat{\mathbf{y}})=E\|\hat{\mathbf{y}}-\boldsymbol{\theta}\|^2$。

后续内容中，θ是待预测变量的期望，$\theta=E(y)$

记$\theta=E(y)$, 则以$\hat{y}$预测y的预测误差 $\mathrm{pe}(\hat{y})=E(\hat{y}-\theta+\theta-y)^2$

$=E(\hat{y}-\theta)^2+E(y-\theta)^2+2E(\hat{y}-\theta)(y-\theta)=E(\hat{y}-\theta)^2+E(y-\theta)^2$

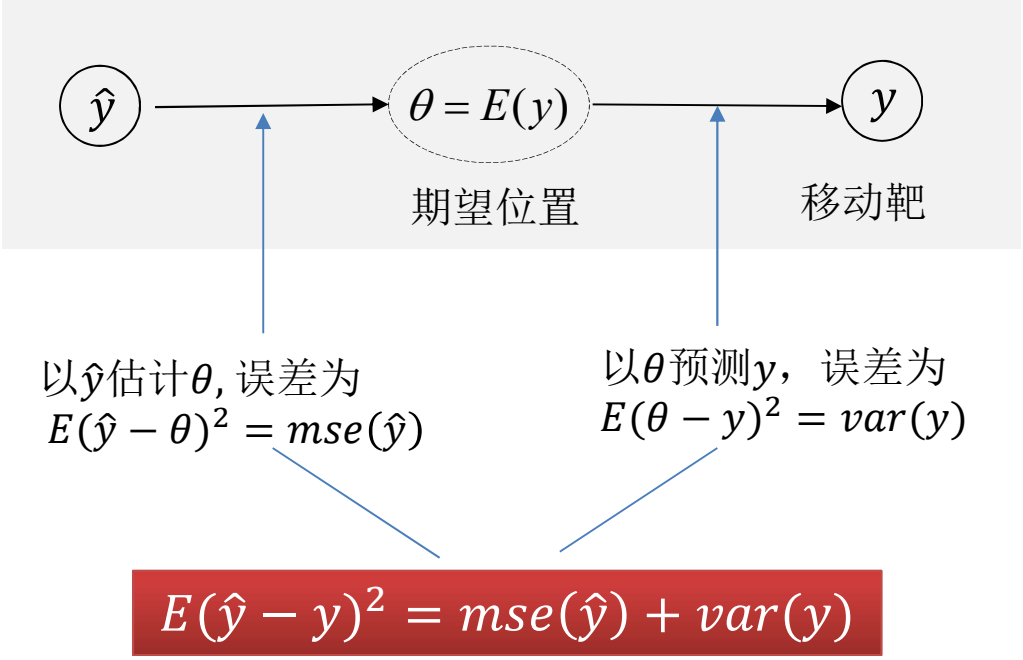
其中因为$\hat{y}$与y独立，所以交叉项为0。因此我们有命题1.

分解预测误差

命题1：假设历史数据$y_1,\ldots,y_n$与待预测随机变量y独立，
假设$E(y)=\theta$，则以统计量 $\hat{y}=f(y_1,\ldots,y_n)$预测$y$的误差可分解为

$$\mathrm{pe}(\hat{y})=\mathrm{mse}(\hat{y})+\mathrm{var}(y)$$

注：预测随机变量y类似于移动靶射击，关键在于判断移动靶的期望位置。



The bias-variance trade-off/dilemma

被预测对象的方差不可控，为了减小预测误差，只能减小MSE。
命题2说明MSE又可分解成方差与偏差平方之和。

MSE
= variance
+ bias^2

命题2. 设 $\hat{y}$ 是参数θ的一个估计， $\hat{y}$的偏差为$\text{bias}(\hat{y}) = E(\hat{y}) - \theta$,
则均方误差可分解为：$\text{mse}(\hat{y}) = \text{var}(\hat{y}) + \text{bias}(\hat{y})^2$.

证明：记$a = E(\hat{y})$, $\text{mse}(\hat{y}) = E(\hat{y} - \theta)^2 = E(\hat{y} - a + a - \theta)^2$

$$= E(\hat{y} - a)^2 + (a - \theta)^2 = \text{var}(\hat{y}) + \text{bias}(\hat{y})^2$$

注： $\hat{y}$ 的期望$a = E(\hat{y})$未必等于待估计参数 θ,
若$\hat{y}$是θ的无偏估计，即 $E(\hat{y}) = \theta$,则$\text{mse}(\hat{y}) = \text{var}(\hat{y})$。

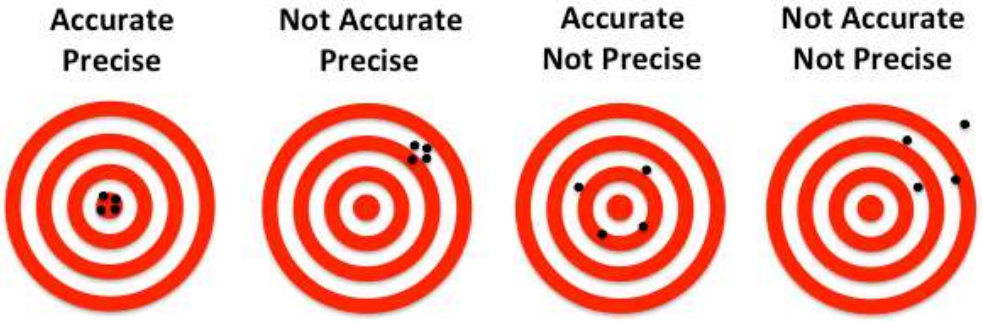
精准：
准确性和精确性的折中

偏差代表准确度(accuracy), 方差代表精确度(precision), MSE是精确度和准确度的折中：

$$MSE = vaiance + bias^2 = \text{精确度} + \text{准确度}$$

极小化其中之一并不意味着MSE最小：

- 无偏估计: 准确度 最高, $bias^2 = 0$，但 方差可能过大（图3）。
- 常数估计: 精确度最高, $vaiance = 0$，但偏差可能过大（图2)。



偏差和方差都小

偏差大 方差小

偏差小 方差大

偏差和方差都大

图1是理想情况，但难以实现

图2能否减小偏差？

图3能否减小方差？

一个常用的策略是，对于普通的统计量（比如样本均值，通常是无偏的），我们设法大幅度地降低其方差但同时允许出现一定的偏差。问题是，给定一个统计量$\hat{\theta}$, 如何减小其方差？

- 压缩，乘以一个小于1的正数或优化求解阶段增加约束/惩罚

$$\hat{\theta} \to \lambda\hat{\theta}，\ 0 \le \lambda \le 1$$

- 截断，缩减其取值范围

$$\hat{\theta} \to \hat{\theta}1_{(|\hat{\theta}| \le c)}$$

有偏统计、统计学习的发展历史：

- James-Stein (1956,1961)：正态分布均值向量的有偏估计（下页）。
- Hoerl and Kennard (1970)：线性模型的岭估计(ridge estimator).
- 规则化/带惩罚的最小二乘：LASSO，贝叶斯方法.
- Vapnik 统计学习/机器学习理论.

James-Stein估计

James & Stein (1956,1961) 发现了一个多元正态分布均值的有偏估计（后被称为James-Stein估计），其MSE表现好于经典的样本均值。这是一个惊人的发现，因为传统上认为样本均值是最好的一个正态均值估计。

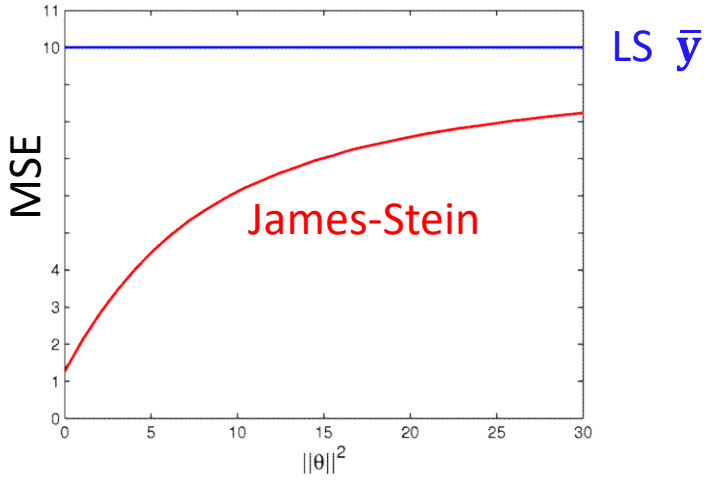
(James-Stein估计). 假设 $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ iid $\sim N(\boldsymbol{\theta}, \sigma^2 I_p)$, $p \geq 3$, 假设σ^2已知。定义$\bar{\mathbf{y}}$的一个压缩估计(有偏):

$$\hat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{n \| \bar{\mathbf{y}} \|^2}\right) \bar{\mathbf{y}}$$

它比LS估计$\hat{\boldsymbol{\theta}} = \bar{\mathbf{y}}$具有更小的MSE:

$$\mathrm{E} \| \hat{\boldsymbol{\theta}}_{\mathrm{JS}} - \boldsymbol{\theta} \|^2 < \mathrm{E} \| \bar{\mathbf{y}} - \boldsymbol{\theta} \|^2$$

θ_k的JS估计不仅仅与样本的第k个分量的平均$\bar{y}_k$有关，也与其它分量有关。考虑到样本的各个分量独立，所以JS估计是反直觉的。



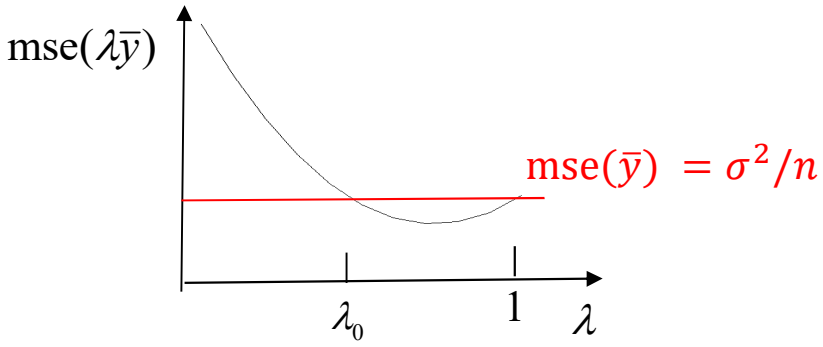
例1：假设$y_1,\ldots,y_n,y$ iid $\sim(\theta,\sigma^2)$，y是待预测随机变量。基于历史样本$y_1,\ldots,y_n$构造y的预测$\hat{y}$

(1) 若预测量$\hat{y}=\bar{y}$，则

$$bias(\bar{y})=0,\ \mathrm{var}(\bar{y})=\sigma^2/n,\ \ \mathrm{mse}(\bar{y})=\sigma^2/n$$

(2) 若 $\hat{y}=\lambda\bar{y}$，则

$$bias(\lambda\bar{y})=(\lambda-1)\theta,\ \ \mathrm{var}(\lambda\bar{y})=\lambda^2\sigma^2/n,\ \ \mathrm{mse}(\lambda\bar{y})=\lambda^2\sigma^2/n+(1-\lambda)^2\theta^2\ 。$$



容易验证，当 $\lambda_0\overset{\Delta}{=}\dfrac{\theta^2-\sigma^2/n}{\theta^2+\sigma^2/n}<\lambda<1$时，$\mathrm{mse}(\lambda\bar{y})<\mathrm{mse}(\bar{y})$

特别地，若 $\lambda_0<0$即 $|\theta|\le\sigma/\sqrt{n}$ (这意味着θ较小或σ较大)，则λ取0时，$\mathrm{mse}(0)<\mathrm{mse}(\bar{y})$,此时常数预测 $\hat{y}=0$优于$\bar{y}$。

例2(惩罚最小二乘：截断).设样本$x_1, x_2, ..., x_n$ iid $\sim (\theta, \sigma^2)$, 假设已知$|\theta| \le c$, 在此约束下，我们极小化误差平方和：

$$\min \sum (x_i - \theta)^2, \quad s.t. |\theta| \le c \quad (\text{约束}, \text{subject to})$$

因为误差平方和 $\sum (x_i - \theta)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$,

约束LS问题转化为：

$$\min (\theta - \bar{x})^2, \quad s.t. \quad |\theta| \le c$$

$$\Rightarrow \text{最优解} \tilde{\theta}_c = \begin{cases} \bar{x} & |\bar{x}| \le c \\ c & \bar{x} > c \\ -c & \bar{x} < -c \end{cases}$$

,它是经典估计$\bar{x}$的截断，有偏，但方差小于$\bar{x}$的方差

例3（贝叶斯估计：压缩）. 设样本$y_1, y_2, ..., y_n$ iid $\sim N(\theta, \sigma^2)$, 假设θ服从先验分布$N(\mu_0, \tau^2)$, 其中μ_0, τ^2已知，则后验分布

$$\theta \mid y\text{'s} \sim N\left(\frac{\tau^2 \bar{y} + \sigma^2 \mu_0}{\tau^2 + \sigma^2}, \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}\right)$$

θ的后验估计 $\tilde{\theta}_{\text{Bayes}} = \dfrac{\bar{y}/\sigma^2 + \mu_0/\tau^2}{1/\tau^2 + 1/\sigma^2}$。

特别地当已知 θ 较小，$\mu_0 = 0$， $\tilde{\theta}_{\text{Bayes}} = \dfrac{\tau^2}{\tau^2 + \sigma^2}\bar{y}$。

假设先验分布$\theta \sim N(\mu_0, \tau^2)$实际上是对$\theta$取值的一种“约束”，即先验上我们已知$\theta \approx \mu_0$

预测误差与均方误差：向量情形

定义（预测误差）. 以随机向量$\hat{\mathbf{y}}$预测随机向量$\mathbf{y}$，记$\boldsymbol{\theta} = E(\mathbf{y})$,

预测误差： $\mathrm{pe}(\hat{\mathbf{y}}) = E\,\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = E(\hat{\mathbf{y}} - \mathbf{y})^{\mathsf{T}}(\hat{\mathbf{y}} - \mathbf{y})$

定义（均方误差）. 若参数$\boldsymbol{\theta}$、统计量$\hat{\mathbf{y}}$是向量，定义

- 均方误差矩阵： $M(\hat{\mathbf{y}}) = \mathrm{MSE}(\hat{\mathbf{y}}) = E\left((\hat{\mathbf{y}} - \boldsymbol{\theta})(\hat{\mathbf{y}} - \boldsymbol{\theta})^{\mathsf{T}}\right) = \mathrm{var}(\hat{\mathbf{y}}) + \mathbf{b}\mathbf{b}^{\mathsf{T}}$,
  其中 $\mathbf{b} = \mathrm{bias}(\hat{\mathbf{y}}) = E\hat{\mathbf{y}} - \boldsymbol{\theta}$.
- 均方误差： $m(\hat{\mathbf{y}}) = \mathrm{mse}(\hat{\mathbf{y}}) = E\,\|\hat{\mathbf{y}} - \boldsymbol{\theta}\|^2 = \mathrm{tr}(\mathrm{MSE}(\hat{\mathbf{y}})) = \mathrm{tr}\left(\mathrm{var}(\hat{\mathbf{y}})\right) + \mathbf{b}^{\mathsf{T}}\mathbf{b}$

所以，预测误差分解为

$$pe(\hat{\mathbf{y}}) = E\,\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \mathrm{trM}(\hat{\mathbf{y}}) + \mathrm{tr}(\mathrm{var}(\mathbf{y})) = \mathrm{tr}(\mathrm{var}(\hat{\mathbf{y}})) + \|\,bias(\hat{\mathbf{y}})\,\|^2 + \mathrm{tr}\left(\mathrm{var}(\mathbf{y})\right)$$

我们将以M代表均方差误差矩阵，m代表均方误差

2. 线性模型中的预测问题

问题框架

- 训练数据： $(\mathbf{x}_i, y_i), i = 1,....,n,$ $p\times 1$自变量$\mathbf{x}_i$的第一个元素为1。
- 模型： $y_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \sim (0,\sigma^2),\quad \mathbf{y} = X_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (0,\sigma^2 I_n),$
- 假设$\tilde{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的某一个估计(未必是LS），$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}$为拟合值向量。

- 待预测数据： 假设$(\mathbf{x}_0, y_0)$满足与训练数据同样的 模型

$$y_0 = \mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_0, \varepsilon_0 \sim (0,\sigma^2)$$

其中自变量$\mathbf{x}_0$已知，但对应的y_0需要预测。假设y_0与$y_1,..., y_n$独立。

- y_0的 预测取为： $\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$, 预测误差 $\mathrm{pe}(\tilde{y}_0) = m(\tilde{y}_0) + \sigma^2$,其中MSE：

$$m(\tilde{y}_0) = E(\tilde{y}_0 - \theta)^2 = \mathbf{x}_0^{\mathsf{T}} E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{x}_0 = \mathbf{x}_0^{\mathsf{T}} M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0$$

引理1. 当$\mathbf{x}_0 \in C(X^{\mathsf{T}})$时, 存在$\mathbf{a} \in R^n$,使得$\mathbf{x}_0 = X^{\mathsf{T}}\mathbf{a}$，此时

$$m(\tilde{y}_0) = \mathbf{x}_0^{\mathsf{T}} M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0 = \mathbf{a}^{\mathsf{T}} M(\tilde{\mathbf{y}})\mathbf{a},\ M(\tilde{\mathbf{y}}) = E(\tilde{\mathbf{y}} - X\boldsymbol{\beta})(\tilde{\mathbf{y}} - X\boldsymbol{\beta})^{\mathsf{T}}$$

即$\tilde{y}_0$的预测误差与训练模型 的拟合$\tilde{\mathbf{y}}$的均方误差矩阵有关。

注:当X列满秩时，$C(X^{\mathsf{T}}) = R^p, \mathbf{x}_0 \in C(X^{\mathsf{T}})$自然成立。

$\mathbf{x}_0 \in C(X^{\mathsf{T}}) \Leftrightarrow \mathbf{x}_0 = \sum \mathbf{x}_i a_i$,新数据$\mathbf{x}_0$与训练数据相似。

引理2：对任何$\mathbf{x} \in R^n$,实数$\lambda > 0$和矩阵$A_{n\times n} > 0$,则$\mathbf{x}^{\mathsf{T}} A\mathbf{x} \le \lambda \Leftrightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda A^{-1}$

证明：令$\mathbf{y} = A^{1/2}\mathbf{x}$, $\mathbf{x}^{\mathsf{T}} A\mathbf{x} \le \lambda \Leftrightarrow \mathbf{y}^{\mathsf{T}}\mathbf{y} \le \lambda \Leftrightarrow \mathbf{y}\mathbf{y}^{\mathsf{T}} \le \lambda I_n$

$\Leftrightarrow A^{1/2}\mathbf{x}\mathbf{x}^{\mathsf{T}} A^{1/2} \le \lambda I_n \Leftrightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda A^{-1}$.

线性模型的预测问题概括为

- 训练得到估计$\tilde{\boldsymbol{\beta}}$及拟合$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}$， 均方误差矩阵为$M(\tilde{\boldsymbol{\beta}})$、$M(\tilde{\mathbf{y}})$。
- 预测统计量$\tilde{\mathbf{y}}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$。
- $\tilde{\mathbf{y}}_0$的预测误差与$M(\tilde{\boldsymbol{\beta}})$、$M(\tilde{\mathbf{y}})$ 有关。后续内容主要关注$M(\tilde{\mathbf{y}})$。

基于子模型的预测

下面考虑基于子模型构建预测统计量，子模型指的是只使用部分自变量的模型（换言之，其它自变量的回归系数压缩为0）

全模型（真模型）：$\mathbf{y} = X\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$，$\boldsymbol{\beta}_1$为$q\times 1$，

LS估计$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$，拟合值$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$。

子模型(工作模型）：我们以$\tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$估计$\boldsymbol{\beta}_2$，以$\tilde{\boldsymbol{\beta}}_1 = (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y}$估计$\boldsymbol{\beta}_1$，

即全模型中$\boldsymbol{\beta}$的估计取为$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \tilde{\boldsymbol{\beta}}_1 \\ 0 \end{pmatrix}$，拟合值：$\tilde{\mathbf{y}} = P_{X_1}\mathbf{y} = X_1\tilde{\boldsymbol{\beta}}_1 = X\tilde{\boldsymbol{\beta}}$.

- 显然$\tilde{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的无偏估计，且$\mathrm{var}(\tilde{\boldsymbol{\beta}}) \le \mathrm{var}(\hat{\boldsymbol{\beta}})$
- 基于$\tilde{\boldsymbol{\beta}}$的预测误差与$M(\tilde{\boldsymbol{\beta}})$或$M(\tilde{\mathbf{y}})$有关，下面我们讨论$M$ 。

命题3：假设$\tilde{\boldsymbol{\beta}}=\begin{pmatrix}\tilde{\boldsymbol{\beta}}_1\\0\end{pmatrix}$，$\tilde{\boldsymbol{\beta}}_1=(X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y}$，$\tilde{\mathbf{y}}=X_1\tilde{\boldsymbol{\beta}}_1=X\tilde{\boldsymbol{\beta}}$，在上述全模型是正确模型的假定下，$\tilde{\mathbf{y}}$的均方误差矩阵$M(\tilde{\mathbf{y}})=E(\tilde{\mathbf{y}}-X\boldsymbol{\beta})(\tilde{\mathbf{y}}-X\boldsymbol{\beta})^{\mathsf{T}}$,我们有

(1) $M(\tilde{\mathbf{y}})=\sigma^2P_{X_1}+X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}$，$m(\tilde{\mathbf{y}})=q\sigma^2+\|X_2^{\perp}\boldsymbol{\beta}_2\|^2$.

$q=p$时，$m(\hat{\mathbf{y}})=p\sigma^2$

$M(\hat{\mathbf{y}})=\mathrm{var}(\hat{\mathbf{y}})=\sigma^2P_X$

(2) $\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma\Leftrightarrow M(\tilde{\mathbf{y}})\le M(\hat{\mathbf{y}})\Leftrightarrow M(\tilde{\boldsymbol{\beta}})\le M(\hat{\boldsymbol{\beta}})$.

注1：对于$\mathbf{x}_0=(\mathbf{x}_{01}^{\mathsf{T}},\mathbf{x}_{02}^{\mathsf{T}})^{\mathsf{T}}$，$y_0=\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta}+\varepsilon_0=\mathbf{x}_{01}^{\mathsf{T}}\boldsymbol{\beta}_1+\mathbf{x}_{02}^{\mathsf{T}}\boldsymbol{\beta}_2+\varepsilon_0$，

仅用$\mathbf{x}_0$的前q个分量$\mathbf{x}_{01}$预测：$y_0=\mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}=\mathbf{x}_{01}^{\mathsf{T}}\tilde{\boldsymbol{\beta}}_1$,

命题3说明当$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma$时，$pe(\tilde{y}_0)=\mathbf{a}^{\mathsf{T}}M(\tilde{\mathbf{y}})\mathbf{a}+\sigma^2\le\mathbf{a}^{\mathsf{T}}M(\hat{\mathbf{y}})\mathbf{a}+\sigma^2=pe(\hat{y}_0)$.

注2：条件“$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma$”意味着$X_2^{\perp}\boldsymbol{\beta}_2$相对于$\sigma$较小。实际操作中需代入$(plug-in)$未知参数的LS估计，即“$\frac{\|X_2^{\perp}\hat{\boldsymbol{\beta}}_2\|^2}{\hat{\sigma}^2}\le 1$”（回忆$H_0:\boldsymbol{\beta}_2=0$的检验$F=\frac{\|X_2^{\perp}\hat{\boldsymbol{\beta}}_2\|^2}{(p-q)\hat{\sigma}^2}$），但代入未知参数的估计会导致偏差（参见后面$C_p$的推导）。

证明：(1) 根据MSE的分解公式，我们下面分别计算$\tilde{\mathbf{y}}$的偏差和方差，注意

$\tilde{\mathbf{y}} = X_1\tilde{\boldsymbol{\beta}}_1 = P_{X_1}\mathbf{y},$

- 偏差 $\mathbf{b} = E(\tilde{\mathbf{y}}) - X\boldsymbol{\beta} = P_{X_1}X\boldsymbol{\beta} - X\boldsymbol{\beta} = P_{X_1}X_2\boldsymbol{\beta}_2 - X_2\boldsymbol{\beta}_2 = -X_2^{\perp}\boldsymbol{\beta}_2,$
- $\mathrm{var}(\tilde{\mathbf{y}}) = \mathrm{var}(P_{X_1}\mathbf{y}) = \sigma^2 P_{X_1},$

$\Rightarrow M(\tilde{\mathbf{y}}) = \mathrm{var}(\tilde{\mathbf{y}}) + \mathbf{b}\mathbf{b}^{\mathsf{T}} = \sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2(X_2^{\perp}\boldsymbol{\beta}_2)^{\mathsf{T}}.$

$\Rightarrow m(\tilde{\mathbf{y}}) = \mathrm{tr}M(\tilde{\mathbf{y}}) = \sigma^2 trP_{X_1} + trX_2^{\perp}\boldsymbol{\beta}_2(X_2^{\perp}\boldsymbol{\beta}_2)^{\mathsf{T}} = q\sigma^2 + \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2 .$

（2）因为 $M(\tilde{\mathbf{y}})=\sigma^2 P_{X_1}+X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}$，$M(\hat{\mathbf{y}})=\sigma^2 P_X$，所以

$$M(\tilde{\mathbf{y}})\le M(\hat{\mathbf{y}}) \Leftrightarrow X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}\le\sigma^2 P_X-\sigma^2 P_{X_1}=\sigma^2 X_2^{\perp}(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}$$

$$\Leftrightarrow \boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}\le\sigma^2(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}$$

“$\Rightarrow$” 左乘$A=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}$,右乘$A^{\mathsf{T}}$

“$\Leftarrow$” 两边同时左乘$X_2^{\perp}$，右乘$X_2^{\perp\mathsf{T}}$

$$\overset{\text{引理2}}{\Leftrightarrow}\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}X_2^{\perp}\boldsymbol{\beta}_2\le\sigma^2 \text{即} \| X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma$$

因为$M(\tilde{\mathbf{y}})=XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}$，$M(\hat{\mathbf{y}})=XM(\hat{\boldsymbol{\beta}})X^{\mathsf{T}}$,

则类似地有$M(\tilde{\boldsymbol{\beta}})\le M(\hat{\boldsymbol{\beta}})\Leftrightarrow M(\tilde{\mathbf{y}})\le M(\hat{\mathbf{y}})$

“$\Rightarrow$” 左乘X,右乘X^{T}

“$\Leftarrow$” 左乘$(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$,右乘$X(X^{\mathsf{T}}X)^{-1}$

变量选择(选择子模型): C_p 准则

过犹不及

过拟合：自变量越多，模型拟合效果越好，但预测效果反而变差。

对于线性模型 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$（全模型），
命题3说明，在条件 $\| X_2^{\perp}\boldsymbol{\beta}_2 \| \le \sigma$ 下，仅含一部分（q个）变量 X_1 的子模型得到的拟合 $\tilde{\mathbf{y}}$，比全模型下的拟合 $\hat{\mathbf{y}}$ 具有更小的MSE，进而相应的预测有更小的预测误差。

Mallow's C_p 准则

Mallow's C_p 准则（1973）：选择变量个数 q 的子模型使得

$$C_q = \frac{\mathrm{RSS}_q}{\hat{\sigma}^2} - n + 2q,$$

达到最小，其中 RSS_q 为子模型下的误差平方和，$\hat{\sigma}^2 = \mathrm{RSS}_p/(n-p)$ 为全模型下的误差方差估计。

C_q 中，主体是 RSS_q，惩罚变量个数 q。

$\mathrm{RSS}_q \downarrow q$，而 $2q \uparrow q$，两者折中，C_q 可能在某个 $0 \le q < p-1$ 处达到最小。

C_p准则的构建过程

对于任何$\mathbf{x}_0$，待预测的y_0满足$y_0 = \mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_0$。我们希望用$\mathbf{x}_0$的部分分量预测$y_0$。

对于$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y} \\ 0 \end{pmatrix}$，由引理1和命题3，预测统计量$\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}} = \mathbf{a}^{\mathsf{T}}\tilde{\mathbf{y}}$的MSE为

$$\mathbf{x}_0^{\mathsf{T}}M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0 = \mathbf{a}^{\mathsf{T}}M(\tilde{\mathbf{y}})\mathbf{a} = \mathbf{a}^{\mathsf{T}}\left(\sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}\right)\mathbf{a}$$

它依赖于具体的$\mathbf{x}_0$或$\mathbf{a}$，不适合作为准则（准 则应与具体的$\mathbf{x}_0$无关）。

我们以$M(\tilde{\mathbf{y}}) = \sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}$的trace，即

$$m(\tilde{\mathbf{y}}) = \mathrm{tr}(\sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}X_2^{\perp\mathsf{T}}) = q\sigma^2 + \| X_2^{\perp}\boldsymbol{\beta}_2 \|^2$$

作为度量。Mallow(1973)基于该MSE构建了C_p变量选择准则。

1.以$m_q = m(\tilde{\mathbf{y}})$代表预测误差

对于q变量子模型$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\delta}$，$\tilde{\boldsymbol{\beta}}_1 = (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y}$, $\tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$,

$\tilde{\mathbf{y}} = X_1\tilde{\boldsymbol{\beta}}_1 = X\tilde{\boldsymbol{\beta}}$的均方误差

$$m_q \overset{\Delta}{=} m(\tilde{\mathbf{y}}) = trM(\tilde{\mathbf{y}}) = E\,\|\,\tilde{\mathbf{y}} - X\boldsymbol{\beta}\,\|^2 = \|\,X_2^{\perp}\boldsymbol{\beta}_2\,\|^2 + q\sigma^2 \qquad (*)$$

m_q中含未知参数,代入LS估计得$q\hat{\sigma}^2 + \|\,X_2^{\perp}\hat{\boldsymbol{\beta}}_2\,\|^2$，$\hat{\sigma}^2$是$\sigma^2$的无偏估计，

但$\|\,X_2^{\perp}\hat{\boldsymbol{\beta}}_2\,\|^2$是$\|\,X_2^{\perp}\boldsymbol{\beta}_2\,\|^2$的有偏估计：

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

所以$E(\|\,X_2^{\perp}\hat{\boldsymbol{\beta}}_2\,\|^2 - (p-q)\hat{\sigma}^2) = \|\,X_2^{\perp}\boldsymbol{\beta}_2\,\|^2$。

2. 构造m_q的无偏估计$\hat{m}_q$

以$\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 -(p-q)\hat{\sigma}^2$作为$\| X_2^{\perp}\boldsymbol{\beta}_2 \|^2$的无偏估计，令

$$\hat{m}_q = \left(\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 -(p-q)\hat{\sigma}^2\right)+q\hat{\sigma}^2 =\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 +2q\hat{\sigma}^2 - p\hat{\sigma}^2$$

$\hat{m}_q$是m_q的无偏估计。

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

$$\| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 = \| (\mathbf{y}-\tilde{\mathbf{y}})-(\mathbf{y}-\hat{\mathbf{y}}) \|^2 = \| \mathbf{y}-\tilde{\mathbf{y}} \|^2 - \| \mathbf{y}-\hat{\mathbf{y}} \|^2 \overset{\Delta}{=} \mathrm{RSS}_q - (n-p)\hat{\sigma}^2,$$

所以$\hat{m}_q = \| X_2^{\perp}\hat{\boldsymbol{\beta}}_2 \|^2 +2q\hat{\sigma}^2 - p\hat{\sigma}^2 = \mathrm{RSS}_q + 2q\hat{\sigma}^2 - n\hat{\sigma}^2.$

3. $C_q = \hat{m}_q/\hat{\sigma}^2$

标准化：$\hat{m}_q$除以$\hat{\sigma}^2$(不依赖于子模型)，即得到Mallow's C_p准则

$$C_q = \hat{m}_q/\hat{\sigma}^2 = \frac{\mathrm{RSS}_q}{\hat{\sigma}^2} + 2q - n$$

变量选择：AIC、BIC准则

AIC和BIC准则是一般统计模型（不局限于线性模型）的变量选择方法。当应用于正态线性模型的变量选择时，AIC与 C_p 类似，但BIC倾向于选择更少的变量。

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 $\hat{\boldsymbol{\theta}}$。

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

  极小化： $\mathrm{AIC} = -2\log L(\hat{\theta}) + 2q$

  其中 $L(\hat{\theta})$ 为似然函数极大值， q 为参数个数。
- BIC (Bayesian Information Criterion, Schwarz 1978)：

  极小化： $\mathrm{BIC} = -2\log L(\hat{\boldsymbol{\theta}}) + (\log n)q$

  其中 n 是样本量， q 是参数个数.

Key:
惩罚模型复杂度 q

对于正态回归模型

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

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

- 为什么称为信息准则？与熵$E_f \log(f) = \int f \log(f)$有关;
- 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C_p，它倾向于选择参数个数更少的模型。

附录：推导AIC准则

密度的Kullaback-Leibler距离：

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

当$g=f$时，$K(g,f)=0$最小。

假设样本$y_1,\ldots,y_n$ iid ~概率密度g，g未知。

待预测随机变量$y_0\sim g$。

设 $f_{\boldsymbol{\theta}}(y)=f(y;\boldsymbol{\theta})$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$\widehat{\boldsymbol{\theta}}$。

我们希望y_0的“概率”$f_{\boldsymbol{\theta}}=f(y_0;\boldsymbol{\theta})$接近真正的$g(y_0)$, 为此极小化KL距离(预测误差)：$K(g,f_{\boldsymbol{\theta}})=\int g\log(g)-\int g\log(f_{\boldsymbol{\theta}})$，这等价于极大化

$$K=K(\boldsymbol{\theta})=\int g\log f_{\boldsymbol{\theta}}$$

因为g未知，K未知。由大数律，对固定的$\boldsymbol{\theta}$

$$\frac{\log L(\boldsymbol{\theta})}{n}=\frac{1}{n}\sum\log(f(y_i;\boldsymbol{\theta})\ \rightarrow\int g(y)\log f(y;\boldsymbol{\theta})=K(\boldsymbol{\theta}),$$

即$\frac{\log L(\boldsymbol{\theta})}{n}\approx K(\boldsymbol{\theta})$。代入$\boldsymbol{\theta}$的估计$\hat{\boldsymbol{\theta}}$，是否还有$\frac{\log L(\hat{\boldsymbol{\theta}})}{n}\approx K(\hat{\boldsymbol{\theta}})$?

$$E(\hat{\boldsymbol{\theta}})=\boldsymbol{\theta} \quad \nRightarrow \quad E\psi(\hat{\boldsymbol{\theta}})=\psi(\boldsymbol{\theta})$$

代入$\boldsymbol{\theta}$的估计$\hat{\boldsymbol{\theta}}$会出现偏差(即使$\hat{\boldsymbol{\theta}}$是无偏估计)，实际上，对数似然函数$\log L(\hat{\boldsymbol{\theta}})$在$\boldsymbol{\theta}$处Taylor展开，可以得到：

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

注意偏差部分$q/n\to 0$，当$n\to\infty$,但我们需保留该项。

即n较大时，近似地有

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

$$-2\log L(\hat{\boldsymbol{\theta}})+2q\approx -2nK(\hat{\boldsymbol{\theta}})$$

注意最大似然函数$L(\hat{\boldsymbol{\theta}})$是一个可计算的量。

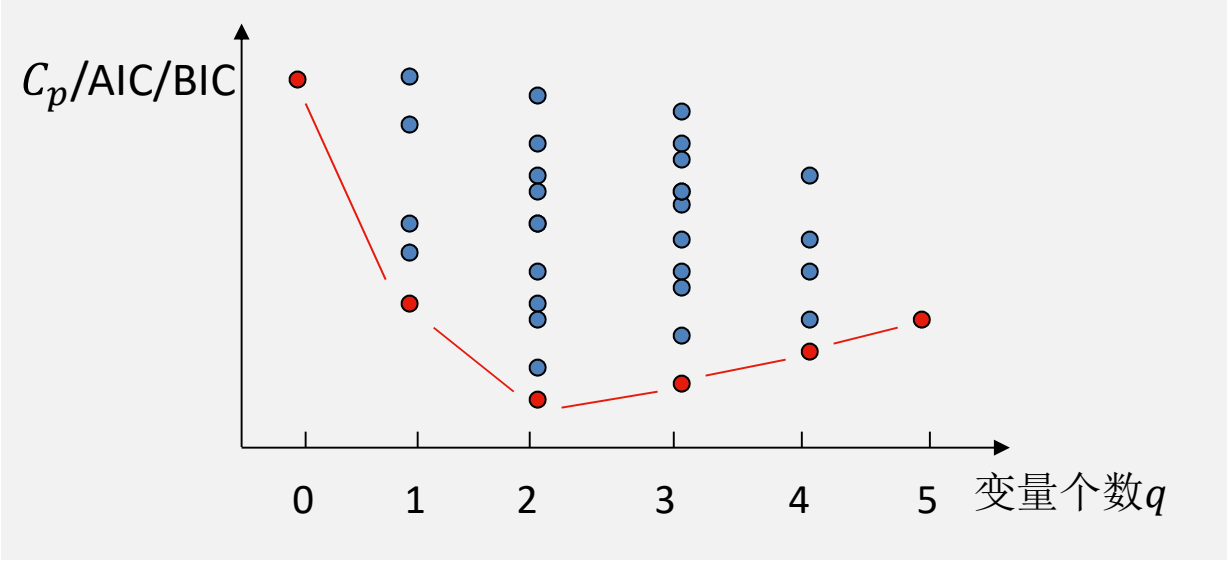
为了极大化$K(\hat{\boldsymbol{\theta}})$,我们极小化

$$AIC=-2\log L(\hat{\boldsymbol{\theta}})+2q,\ q=\boldsymbol{\theta}\text{的长度}$$

3. 最优子集回归

最优子模型

假设设计阵X为$n\times p$矩阵，共有$p-1$个自变量，每个自变量 都有可能入选模型或被舍弃，子模 型总个数为2^{p-1}。最优子集回归对所有 的2^{p-1}个子模型都计算变量选择 准则C_p(或AIC / BIC)。除非p较小，或者特殊的设计阵情形（参见例1），一般情况下，最优 子集算法计算量巨大。



具体执行最优子集搜索时，对不同的变量个数$q = 0,1,2,\ldots,p-1$，需要比较C_{p-1}^{q}个含q个变量的模型的RSS。

例4. 各列正交情形最优子集方法计算复杂度可降低到$O(p^2)$:

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

则$\hat{\mathbf{y}}=P_X\mathbf{y}=\mathbf{1}\bar{y}+P_{\mathbf{x}_1}\mathbf{y}+\ldots+P_{\mathbf{x}_{p-1}}\mathbf{y},$

$RSS=s_{yy}-\|P_{\mathbf{x}_1}\mathbf{y}\|^2-\ldots-\|P_{\mathbf{x}_{p-1}}\mathbf{y}\|^2$

从大到校排列投影长度(计算复杂度$O(p^2)$)，不妨设次序为：

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

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

所以只需要对$q=1,2,\ldots,p-1,$求$C_q=\dfrac{\mathrm{RSS}(q)}{\hat{\sigma}^2}+2q-n$的最小值即可。

主成分回归将一般设计阵变换为列正交（主成分）情形。

主成分回归

为了叙述简便，不妨假设设计阵$X_{n\times p}$和相应变量$\mathbf{y}$都已经中心化（此时回归模型没有截距项）。

如果X的列相互正交，那么投影、选子模型都将非常容易实现。对于一般X,我们可以通过变换使得变换后的矩阵各列正交（比如Gram-Schmidt正交化、奇异值分解）。

奇异值分解

假设X有奇异值分解：

$$X_{n\times p}=U_{n\times p}D_{p\times p}V^{\mathsf{T}}{}_{p\times p},$$

其中$U^{\mathsf{T}}U=V^{\mathsf{T}}V=VV^{\mathsf{T}}=I_p, D=diag(\sqrt{\lambda_1},...,\sqrt{\lambda_p}),\ \sqrt{\lambda_1}\geq...\geq\sqrt{\lambda_p}$。

X的变换$W_{n\times p}=XV=UD$的各列相互正交，称为主成分矩阵，其第k列称为第k个主成分。$U_{n\times p}=XVD^{-1}$也是列正交矩阵，且每列模长为1。

主成分回归

将 $X=UDV^{\mathsf{T}}$ 代入模型（假设 $\lambda_1 \geq \ldots \geq \lambda_p > 0$），模型改写为：

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

U作为设计阵是列正交的，$U^{\mathsf{T}}U=I_p, U=XVD^{-1}$是$X$的变换，其各列视作“新”的变量，是$X$的各列（变量）的线性组合。

（也可以W为设计阵：$\mathbf{y}=W\boldsymbol{\eta}+\boldsymbol{\varepsilon}$，$W=UD$，$\boldsymbol{\eta}=V^{\mathsf{T}}\boldsymbol{\beta}$，此处省略）

在模型表示$\mathbf{y}=U\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$下：

- 设$U=(\mathbf{u}_1,\ldots,\mathbf{u}_p)$，则 $P_X=P_U=U(U^{\mathsf{T}}U)^{-1}U^{\mathsf{T}}=UU^{\mathsf{T}}=\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j^{\mathsf{T}}$
- $\mathbf{y}$在X上的投影可以表示为：

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

- 其中γ_j的LS估计$\boldsymbol{\gamma}$的LS估计$\hat{\gamma}_j=\mathbf{u}_j^{\mathsf{T}}\mathbf{y}$，$\hat{\boldsymbol{\gamma}}=(U^{\mathsf{T}}U)^{-1}U^{\mathsf{T}}\mathbf{y}=U^{\mathsf{T}}\mathbf{y}$.
- $\|\hat{\mathbf{y}}\|^2=\sum_{j=1}^{p}\hat{\gamma}_j^{\,2}$，$RSS=\|\mathbf{y}\|^2-\|\hat{\mathbf{y}}\|^2=\|\mathbf{y}\|^2-\sum_{j=1}^{p}\hat{\gamma}_j^{\,2}$，
- $DV^{\mathsf{T}}\boldsymbol{\beta}=\boldsymbol{\gamma}\Rightarrow\hat{\boldsymbol{\beta}}=VD^{-1}\hat{\boldsymbol{\gamma}}=VD^{-1}U^{\mathsf{T}}\mathbf{y}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$

下标集合$S \subset \{1,\ldots,p\}$对应的子模型为：$\mathbf{y} = U_S \boldsymbol{\gamma}_S + \boldsymbol{\delta} = \sum_{j\in S} \mathbf{u}_j \gamma_j + \boldsymbol{\delta}$

其中$U_S = (\mathbf{u}_j, j \in S), \boldsymbol{\gamma}_S = (\gamma_j, j \in S)^T$，子模型的拟合

$$\tilde{\mathbf{y}} = P_{U_S}\mathbf{y} = \sum_{j\in S} \mathbf{u}_j \hat{\gamma}_j, \quad m(\tilde{\mathbf{y}}) = |S|\sigma^2 + \sum_{j\in S} \gamma_j^2.$$

$$m(\tilde{\mathbf{y}}) = E\|\tilde{\mathbf{y}} - U\boldsymbol{\gamma}\|^2 = E\|\sum_{j\in S}\mathbf{u}_j\hat{\gamma}_j - \sum_{j=1}^{p}\mathbf{u}_j\gamma_j\|^2$$

$$= E\|\sum_{j\in S}\mathbf{u}_j(\hat{\gamma}_j - \gamma_j) - \sum_{j\in S^c}\mathbf{u}_j\gamma_j\|^2 = \sum_{j\in S}E(\hat{\gamma}_j - \gamma_j)^2 + \sum_{j\in S^c}\gamma_j^2$$

最优子集主成分回归

1. 将$|\hat{\gamma}_j| = |\mathbf{u}_j^{\mathsf{T}}\mathbf{y}|$排序，不妨设$|\hat{\gamma}_{i_1}| \geq |\hat{\gamma}_{i_2}| \geq \ldots \geq |\hat{\gamma}_{i_p}|$

2. $q^* = \arg\min_{q=0,1,\ldots,p}\left(\sum_{j=q+1}^{p}\hat{\gamma}_{i_j}^2/\hat{\sigma}^2 + 2q - n\right)$

最优子模型$\mathbf{y} = \sum_{j=1}^{q^*}\mathbf{u}_{i_j}\gamma_{i_j} + \boldsymbol{\delta}.$

传统的主成分回归

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

传统的主成分回归使用最大的q个主成分用于预测：

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

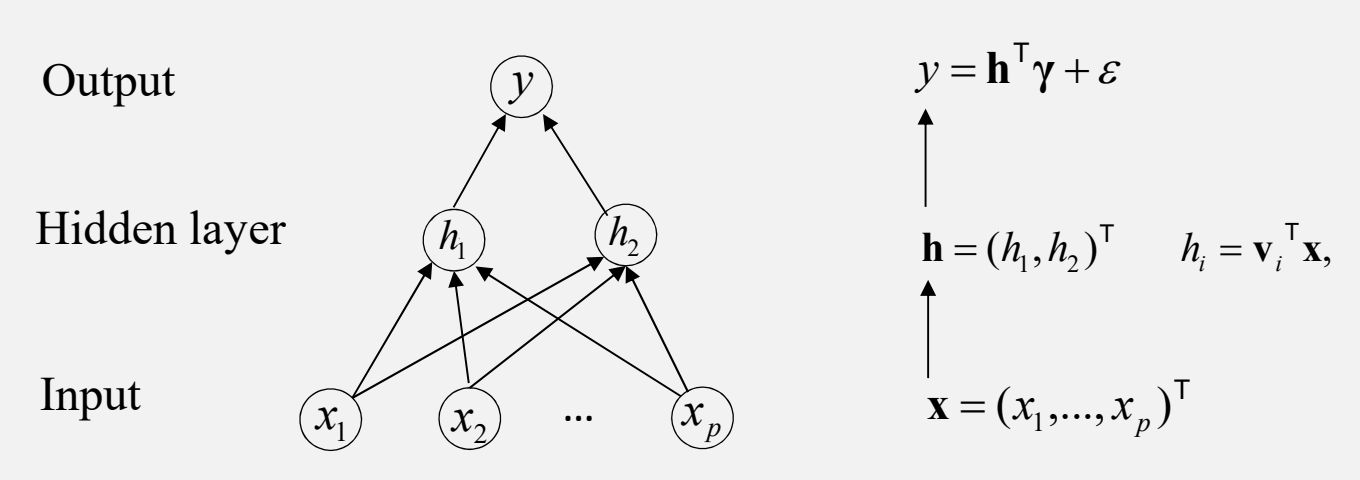
$$\tilde{\mathbf{y}}^{(\mathrm{pc})} = U\tilde{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \mathbf{u}_j \hat{\gamma}_j = \sum_{j=1}^{q} \mathbf{u}_j \left\{ \mathbf{u}_j{}^{\mathsf{T}}\mathbf{y} \right\}$$

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

显然这种方法得到的子模型的自变量有良好的统计解释(前q个主成分)，并不会得到最优的预测子模型。

主成分回归与神经网络

主成分回归把原始p个自变量$\mathbf{x}$变换为q个"新"的或隐藏变量$\mathbf{h}$,然后对响应变量y和隐藏层变量$\mathbf{h}$建立线性回归模型。如下图表示：



一般的神经网络与此类似，它将原始自变量进行多次非线性变换(多个隐藏层)，对最终得到的新变量和响应变量建立通常的回归模型。不同的是

- 隐藏层变量一般是原变量的非线性变换；
- 隐藏层可以不止一个；
- 输出变量/响应变量为连续变量时，建立线性回归模型.

4. 变量选择的贪心算法

最优子集回归选择/搜索Cp或AIC/BIC等准则最小的子模型，计算量太大。贪心算法（greedy）是一种计算高效的近似求解最优子集的方法。常用的贪心算法有：

(1). 逐步回归法 (Stepwise regression):

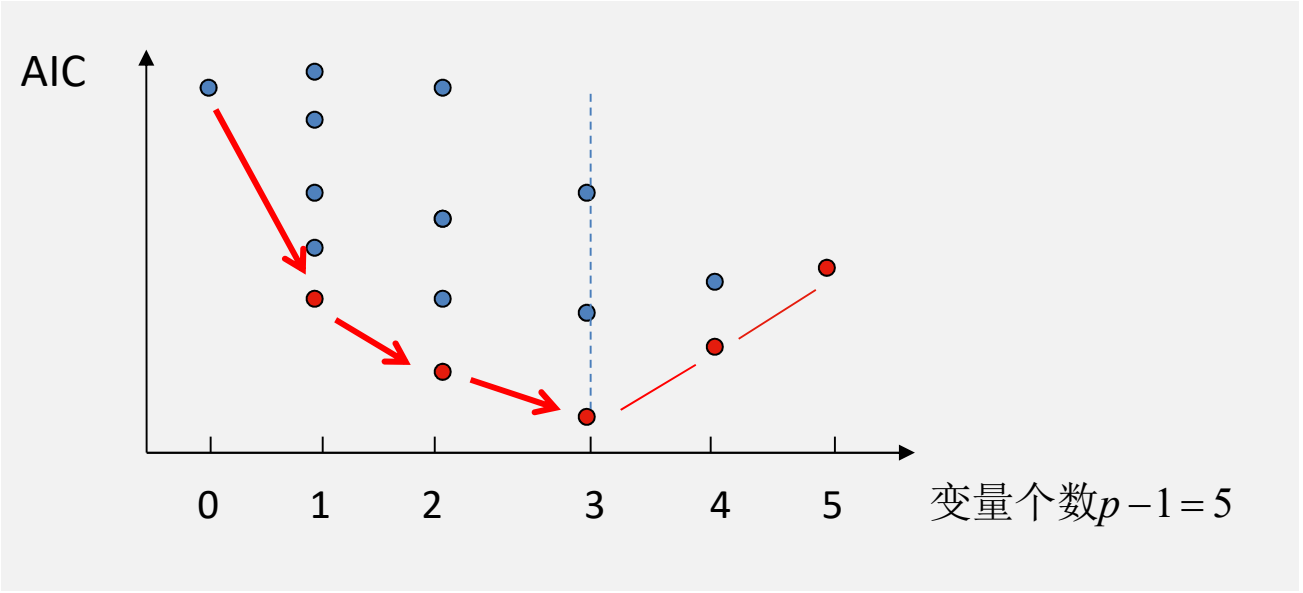
- 向前逐步回归(forward stepwise regression),
- 向后逐步回归(backward stepwise regression)
- 向前-向后逐步回归(forward-backward)

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

逐步回归方法沿*AIC*跳跃最大的路径(故称为greedy)搜索子模型。考察的子模型个数最多为$O(p^2)$。逐步回归得到的解可能不是最优子集。

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减少最多的那个变量),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.



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),
变量一旦被选入(剔除)就不会再被剔除(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
```

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。

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,可处理任意多自变量($p > n$)。

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$\mathbf{x}_1,...,\mathbf{x}_p$都已经中心标准化。残差初值：$\mathbf{e}_0 = \mathbf{y}$

- 求与$\mathbf{e}_0$相关系数最大的自变量：$j_1 = \arg\max \mathbf{e}_0^{\mathsf{T}} \mathbf{x}_j$,

  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：$j_2 = \arg\max_{j \neq j_1} \mathbf{e}_1^{\mathsf{T}} \mathbf{x}_j$,

  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.