第十四讲 广义最小二乘

2023.12.22

$$\boxed{var\left(\frac{\Sigma y_i/\sigma_i^2}{1/\sigma_i^2}\right) \le var(\bar{y}),\ \ \sigma_i^2 = var(y_i)}$$

异方差：一个简单例子

例1. 假设独立样本 $y_1, \dots, y_n$，$\mu = E(y_i)$，方差 $\sigma_i^2 = \mathrm{var}(y_i)$不同，假设$\sigma_i^2$已知。为了估计$\mu$，我们可以应用最小二乘：

$$\min_{\mu} \Sigma (y_i - \mu)^2$$

得OLS估计$\hat{\mu}_{OLS} = \bar{y}$（为了和广义最小二乘或加权最小二乘区分，我们称之为普通最小二乘(OLS, ordinary LS)）。

OLS平等对待各个y_i，直观上不尽合理（因为方差不同），方差大的样本应给与较小权重。令标准化$z_i = (y_i - \mu)/\sigma_i$ ，极小化加权误差平方和

$$\min_{\mu} \Sigma {z_i}^2 = \min_{\mu} \Sigma (y_i - \mu)^2 / \sigma_i^2$$

由此解得加权最小二乘估计（WLS，Weighted LS）

$$\hat{\mu}_{WLS} = \frac{\sum y_i/\sigma_i^2}{\sum 1/\sigma_i^2} = \Sigma w_i y_i / \Sigma w_i, \quad w_i = 1/\sigma_i^2$$

加权最小二乘估计是BLUE，特别地

$$\mathrm{var}(\hat{\mu}_{WLS}) = \frac{1}{\sum 1/\sigma_i^2} \le \Sigma \sigma_i^2 / n^2 = \mathrm{var}(\hat{\mu}_{OLS})$$

广义最小二乘（GLS）

Box-Cox变换有可能能够解决现误差方差不齐(heteroscedasticity)现象。但如果方差结构已知或部分已知，可应用广义最小二乘(GLS, Generalized LS)。

假设异方差线性模型（方差不齐且样本有可能是相依的）

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim (0, \Sigma),\ \ \boldsymbol{\varepsilon} \perp\!\!\!\perp X \qquad (1)$$

其中$\mathrm{var}(\boldsymbol{\varepsilon}) = \Sigma > 0$，通常 $\Sigma = \Sigma(\boldsymbol{\theta})$与参数$\boldsymbol{\theta}$有关。

尽管$\Sigma \neq \sigma^2 I_n$，我们仍可应用最小二乘法(称为OLS，ordinary LS)得

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

容易验证它是无偏估计， 且

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} | X) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}$$

但它不是最优的（除非$\Sigma = \sigma^2 I_n$），最优的估计应该利用方差结构。

Σ 已知情形

假设Σ完全已知,我们可将异方差模型变换为方差齐性模型（虽然这种情况在实际问题中几乎不存在，但对于了解问题有帮助）：

模型$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$两端同时左乘$\Sigma^{-1/2}$,

$$\mathbf{y}^* \triangleq \Sigma^{-1/2}\mathbf{y}=\Sigma^{-1/2}X\boldsymbol{\beta}+\Sigma^{-1/2}\boldsymbol{\varepsilon} \triangleq X^*\boldsymbol{\beta}+\boldsymbol{\varepsilon}^*, \boldsymbol{\varepsilon}^* \sim (0, I_n), \qquad (2)$$

该模型满足GM假设，则模型(2)的误差平方和

$$\|\mathbf{y}^*-X^*\boldsymbol{\beta}\|^2=(\mathbf{y}-X\boldsymbol{\beta})^\top\Sigma^{-1}(\mathbf{y}-X\boldsymbol{\beta})$$

基于模型(2)的$\boldsymbol{\beta}$的LS估计（BLUE最优）

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}=(X^{*\top}X^*)^{-1}X^{*\top}\mathbf{y}=(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\mathbf{y}$$

对模型(1)来说，该估计称为广义最小二乘估计(GLS, generalized LS)

命题1. 假设异方差模型(1)中Σ已知，则最优线性无偏估计$\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}=(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\mathbf{y}$，它使得$(\mathbf{y}-X\boldsymbol{\beta})^\top\Sigma^{-1}(\mathbf{y}-X\boldsymbol{\beta})$极小。

注：由GM定理，$\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}|X)\le\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}|X)$，

$$(X^\top\Sigma^{-1}X)^{-1}\le(X^\top X)^{-1}X^\top\Sigma X(X^\top X)^{-1}$$

抽样调查数据：$\Sigma = \sigma^2\Sigma_0$

假设误差异方差几乎完全已知，具有形式：

$$\Sigma = \sigma^2\Sigma_0\text{，其中 }\Sigma_0\text{已知，}\sigma^2\text{未知，}$$

代入命题1中，σ^2被约掉，所以此时

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (X^\top\Sigma_0^{-1}X)^{-1}X^\top\Sigma_0^{-1}\mathbf{y}$$

而σ^2的GLS估计$\hat{\sigma}^2_{\text{GLS}} = \left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{\text{GLS}}\right)^\top\Sigma_0^{-1}(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{\text{GLS}})/(n-p)$

例2. 在cluster survey中，第i个$cluster$的响应通常是类内所有m_i个个体的汇总，比如平均值，自变量$\mathbf{x}_i$为第i个$cluster$的特征,因为平均值的方差与个数成反比，假设异方差模型

$$y_i = a + \mathbf{b}^\top\mathbf{x}_i + \varepsilon_i = \tilde{\mathbf{x}}_i^\top\boldsymbol{\beta} + \varepsilon_i, E(\varepsilon_i) = 0,\quad \text{var}(\varepsilon_i) = \sigma^2/m_i = \sigma^2/w_i,$$

即$\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 diag(1/m_1,\ldots,1/m_n) = \sigma^2 W^{-1}, W = diag(w_1,\ldots,w_n), w_i = m_i$称为权重.

GLS(此时一般称为WLS)的目标函数是加权误差平方和：

$$\sum_{i=1}^n w_i(y_i - \tilde{\mathbf{x}}_i^\top\boldsymbol{\beta})^2$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{i=1}^n w_i\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\right)^{-1}\left(\sum_{i=1}^n w_i\tilde{\mathbf{x}}_i y_i\right),\quad \hat{\sigma}^2_{\text{WLS}} = \sum_{i=1}^n w_i(y_i - \tilde{\mathbf{x}}_i^\top\hat{\boldsymbol{\beta}}_{WLS})^2/(n-p),$$

其中$e_i = \sqrt{w_i}(y_i - \tilde{\mathbf{x}}_i^\top\hat{\boldsymbol{\beta}}_{WLS})$称为加权残差。OLS是$w_i \equiv 1$的情形。

特别地，

(1) 没有自变量的情形：$y_i = a + \varepsilon_i, E(\varepsilon_i) = 0$， $\mathrm{var}(\varepsilon_i) = \sigma^2/m_i = \sigma^2/w_i$，

目标函数$\sum_{i=1}^{n} w_i (y_i - a)^2$,这就是例1.

（2）简单线性回归：假设$y_i = a + bx_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 / w_i), w_i$已知,目标函数

$$\sum_{i=1}^{n} w_i (y_i - a - bx_i)^2$$

$$\hat{b} = \frac{\sum w_i (x_i - \bar{x}_w) y_i}{\sum w_i (x_i - \bar{x}_w)^2}, \quad \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}; \ \hat{a} = \bar{y}_w - \hat{\beta}_1 \bar{x}_w, \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i},$$

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum w_i (x_i - \bar{x}_w)^2}, \ \mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2}{\sum w_i} + \frac{\bar{x}_w{}^2 \sigma^2}{\sum w_i (x_i - \bar{x}_w)^2}$$

注意：如果将加权最小二乘目标函数写成普通最小二乘的形式(变换y, x):

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i} \beta_0 - \beta_1 \sqrt{w_i} x_i)^2$$

记$\tilde{y}_i = \sqrt{w_i} y_i, \tilde{x}_i = \sqrt{w_i} x_i, \tilde{z}_i = \sqrt{w_i}$,上述右端平方和为$\sum_{i=1}^{n} (\tilde{y}_i - \tilde{z}_i \beta_0 - \tilde{x}_i \beta_1)^2$

对应的方差齐性模型$\tilde{y}_i = \tilde{z}_i \beta_0 + \tilde{x}_i \beta_1 + \tilde{\varepsilon}_i$不含截距项。

$\Sigma = \Sigma(\boldsymbol{\theta})$ 情形

假设异方差模型的方差具有复杂结构$\Sigma = \Sigma(\boldsymbol{\theta}), \boldsymbol{\theta}$是未知参数，需要同时估计回归系数$\boldsymbol{\beta}$和方差中的$\boldsymbol{\theta}$。此时LS或GLS都不能直接应用, 通常我们

- 假设误差服从正态分布并应用极大似然方法估计参数$(\boldsymbol{\beta}, \boldsymbol{\theta})$;
- 另一种方法是迭代加权最小二乘法，计算相对简单。

例3. 纵向数据(longitudinal data). 同一个个体跟踪反复测量若干次，不同的个体独立。数据：$(y_{ij}, \mathbf{x}_{ij}), j=1,\dots,m_i; i=1,\dots,n$。 假设模型

$$y_{ij} = a + a_i + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \text{ iid} \sim N(0,\sigma^2),\ a_i \text{ iid} \sim N(0,\tau^2), \varepsilon_{ij}, a_i\text{与独立}$$

对于固定的i, 各次测量$y_{i1},\dots,y_{im_i}$共享同一个随机变量a_i（随机效应）：

$$\mathrm{cov}(y_{ij}, y_{ik} \mid \mathbf{x}_{ij}, \mathbf{x}_{ik}) = \mathrm{cov}(a_i, a_i) = \tau^2$$

a_i代表个体i的效应(不同的i有不同的a_i)，令$\tilde{\varepsilon}_{ij} = a_i + \varepsilon_{ij}$，则

$$y_{ij} = a + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{ij} + \tilde{\varepsilon}_{ij},\ (\tilde{\varepsilon}_{i1},\dots,\tilde{\varepsilon}_{im_i})^{\mathrm{T}} \sim N(0,\Sigma_i),\ \Sigma_i = \sigma^2 I_{m_i} + \tau^2\mathbf{1}\mathbf{1}^{\mathrm{T}}$$

写成$\mathbf{y} = (y_{11}, y_{12},\dots, y_{21}, y_{22},\dots)^{\mathrm{T}} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$的形式，则$\boldsymbol{\varepsilon}$的方差矩阵是分块对角矩阵$\mathrm{diag}(\Sigma_1, \Sigma_2,\dots,\Sigma_n)$.

极大似然方法

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 假设 $\boldsymbol{\varepsilon} \sim N(0, \Sigma)$, $\Sigma = \Sigma(\boldsymbol{\theta})$, 似然函数:

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{n/2} |\Sigma(\boldsymbol{\theta})|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} \Sigma(\boldsymbol{\theta})^{-1} (\mathbf{y} - X\boldsymbol{\beta}) \right)$$

极大似然函数等价于极小化:

$$Q(\boldsymbol{\theta}, \boldsymbol{\beta}) = -2\log L = \log|\Sigma(\boldsymbol{\theta})| + (\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} \Sigma(\boldsymbol{\theta})^{-1} (\mathbf{y} - X\boldsymbol{\beta}).$$

注1：当$\Sigma = \sigma^2 I_n$时，

$$-2\log L = n\log(\sigma^2) + \|\mathbf{y} - X\boldsymbol{\beta}\|^2 / \sigma^2.$$

$$\min_{\boldsymbol{\beta}}(-2\log L) \Leftrightarrow \min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 / \sigma^2 \Leftrightarrow \min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2$$

所以关于$\boldsymbol{\beta}$，极大似然法与LS完全相同，$\hat{\boldsymbol{\beta}}_{mle} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} \mathbf{y}$，

但$\hat{\sigma}^2_{\mathrm{mle}} = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}_{mle}\|^2 / n = \dfrac{n-p}{n}\hat{\sigma}^2$与LS估计$\hat{\sigma}^2$略有差异。

类似地，当$\Sigma = \sigma^2 \Sigma_0$时($\Sigma_0$已知)，MLE与GLS基本相同。

注2：如果极大似然估计不容易求解，可使用迭代加权最小二乘法。

迭代加权最小二乘方法(IRLS）

复杂的误差方差结构$\Sigma = \Sigma(\boldsymbol{\theta})$下，通用的方法是极大似然法，即极小化

$$Q(\boldsymbol{\theta}, \boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^{\mathrm{T}} \Sigma^{-1}(\theta)(\mathbf{y} - X\boldsymbol{\beta}) + \log|\Sigma(\boldsymbol{\theta})|,$$

这可以认为是带有惩罚项$\log|\Sigma(\boldsymbol{\theta})|$的GLS误差平方和的极小化。该优化问题有时会非常复杂，难以求解最优的$\boldsymbol{\theta}$。

注意到如果$\boldsymbol{\theta}$已知，那么关于$\boldsymbol{\beta}$的最优化是GLS问题，有显式解;
而如果$\boldsymbol{\beta}$已知，那么我们可以求出残差，假如利用残差估计$\boldsymbol{\theta}$比较容易，那么我们可以分步迭代求解$\boldsymbol{\beta}$和$\boldsymbol{\theta}$:

IRLS（Iteratively Reweighted Least Sqaures）:

- 给定当前的$\boldsymbol{\theta}$解，利用GLS求解$\boldsymbol{\beta}$;
- 给定$\boldsymbol{\beta}$，利用残差求解$\boldsymbol{\theta}$

上述两步反复迭代直至收敛。

例4. 假设模型$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \Sigma)$, 其中$\Sigma = \sigma^2 \left(\rho^{|i-j|}\right)_{1\le i,j\le n}, \sigma^2, \rho$未知。

误差满足1阶自回归AR(1)模型：$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i, \delta_i \sim (0, \sigma^2)$与$\varepsilon_i$独立

IRLS:

$k = 0$, $\boldsymbol{\beta}$的初始估计可取为OLS估计$\hat{\boldsymbol{\beta}}^{(0)} = (X^\top X)^{-1} X^\top \mathbf{y}$,

REPEAT

$k = k + 1$

$\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}}^{(k-1)}$,

$$\hat{\sigma}_k^2 = \sum_{i=1}^{n} e_i^2 / (n - p), \quad \hat{\rho}_k = \sum_{i=1}^{n-1} e_i e_{i+1} \Big/ \sum_{i=1}^{n} e_i^2$$

$\leftarrow \varepsilon_i, \varepsilon_{i+1}$的相关系数为$\rho$

$$\hat{\Sigma}^{(k)} = \hat{\sigma}_k^2 \left(\hat{\rho}_k^{|i-j|}\right)_{1\le i,j\le n}$$

$$\hat{\boldsymbol{\beta}}^{(k)} = (X^\top \hat{\Sigma}^{(k)-1} X)^{-1} X^\top \hat{\Sigma}^{(k)-1} \mathbf{y}$$

极小化加权平方和

有很多优化问题可凑成加权平方和的形式：

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \sum w_i(\boldsymbol{\beta},\boldsymbol{\theta})\left(y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2,$$

其中权函数与$\boldsymbol{\beta},\boldsymbol{\theta}$有关，$\boldsymbol{\theta}$一般与误差方差有关。假设分别优化$\boldsymbol{\beta}$和$\boldsymbol{\theta}$比较容易，可以转化为IRLS：

$$\boldsymbol{\beta}^{(k)} = \arg\min_{\boldsymbol{\beta}} \sum w_i(\boldsymbol{\beta}^{(k-1)},\boldsymbol{\theta}^{(k-1)})\left(y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2,$$

$\leftarrow w_i$中的$\boldsymbol{\beta}$给定，
平方项中的$\boldsymbol{\beta}$未知

$$e_i^{(k)} = y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}^{(k)}$$

$$\boldsymbol{\theta}^{(k)} = h(e_i^{(k)}, i=1,\dots,n)$$

具体如下：

- 给定$\boldsymbol{\theta} = \boldsymbol{\theta}^{(k-1)}$，$\boldsymbol{\beta} = \boldsymbol{\beta}^{(k-1)}$，计算$w_i(\boldsymbol{\beta}^{(k-1)},\boldsymbol{\theta}^{(k-1)})$，WLS方法更新$\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$;
- 使用更新的$\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$，更新残差$e_i^{(k)} = y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}^{(k)}$，$i=1,\dots,n$,
- 使用残差$\{e_i^{(k)}, i=1,\dots,n\}$更新$\boldsymbol{\theta}$的估计。

例5. 线性模型的M估计方法（稳健回归）: 模型 $:\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$,

$$\min \sum \rho(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),\quad \rho(\cdot) \geq 0\text{关于}0\text{对称，}\quad \rho(0) = 0.$$

当$\rho(t) = |t|$时称为最小一乘法或LAD (least absolute deviation), 如何优化$\min \sum | y_i - \mathbf{x}_i^\top \boldsymbol{\beta} |$?

对于LAD方法, 改写目标函数:

$$\sum | y_i - \mathbf{x}_i^\top \boldsymbol{\beta} | = \sum w_i(\boldsymbol{\beta}) | y_i - \mathbf{x}_i^\top \boldsymbol{\beta} |^2, \text{ 其中} w_i(\boldsymbol{\beta}) = | y_i - \mathbf{x}_i^\top \boldsymbol{\beta} |^{-1}.$$

$IRLS$:

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\beta} \sum w_i(\boldsymbol{\beta}^{(k)}) | y_i - \mathbf{x}_i^\top \boldsymbol{\beta} |^2.$$

预测

人类对未知和不确定性存在恐惧或好奇 ⇒ 预测、预言、先知。

预测： “估计”随机变量

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。

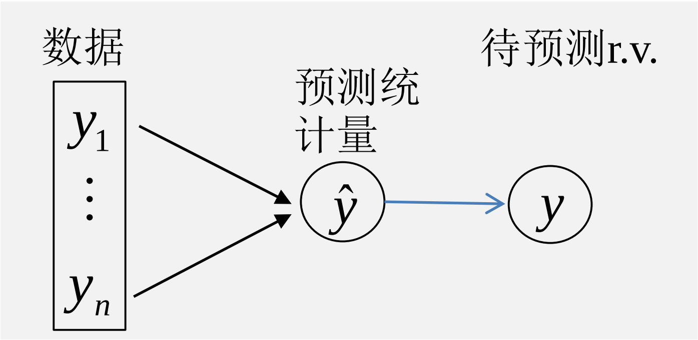
预测的一般原则

- 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： $y_1, \dots, y_n$

待预测随机变量：y（与y_i's独立）

预测统计量：$\hat{y} = h(y_1, \dots, y_n)$



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

回归预测：这里我们假设y_i，y 是连续随机变量，误差为平方误差。每个y_i 可能与自变量$\mathbf{x}_i$有关，待预测量y与$\mathbf{x}$有关，通过线性回归或非线性回归建立回归关系$y_i = f(\mathbf{x}_i) + \varepsilon_i$, 对于给定的$\mathbf{x}$预测$y$为$\hat{y} = f(\mathbf{x})$。

当 y 是0-1变量时，通常误差取为交叉熵。

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

记$\theta = E(y)$，则以$\hat{y}$预测y的预测误差

$$\begin{aligned}\mathrm{pe}(\hat{y}) &= E(\hat{y}-\theta+\theta-y)^2 \\ &= E(\hat{y}-\theta)^2 + E(y-\theta)^2 + 2E(\hat{y}-\theta)(y-\theta) \\ &= E(\hat{y}-\theta)^2 + E(y-\theta)^2\end{aligned}$$

$\hat{y}$与y独立，所以交叉项为0

其中$E(\hat{y}-\theta)^2$称为$\hat{y}$的均方误差，$E(y-\theta)^2 = \mathrm{var}(y)$。

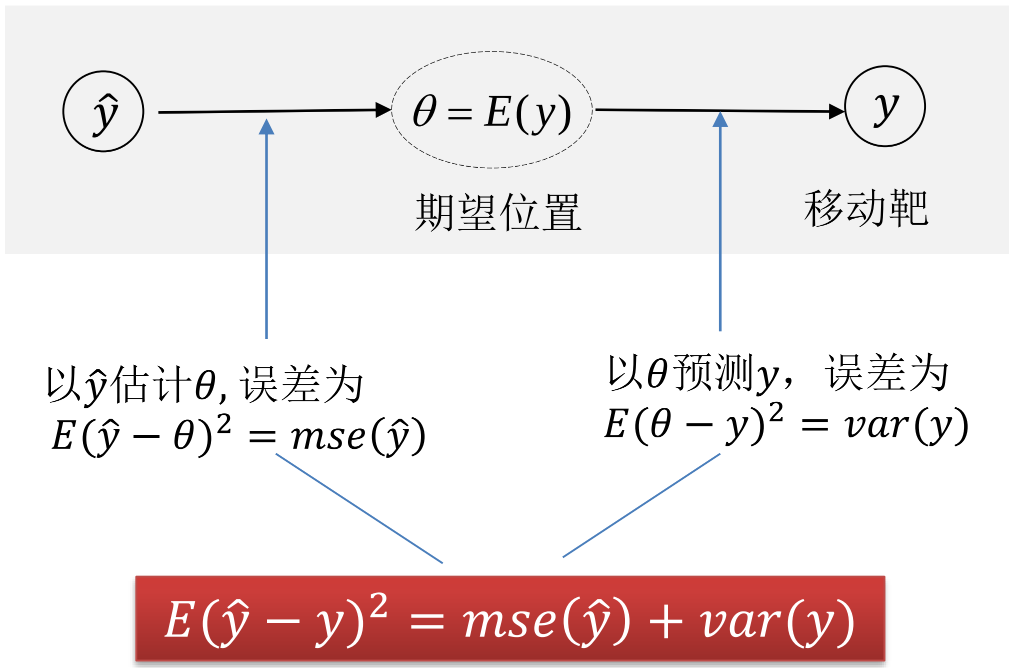
预测误差的分解

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

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

后续内容中，θ是待预测变量的期望，$\theta = E(y)$

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。



注意$\hat{y}$ 未必是θ的无偏估计，故$mse(\hat{y}) = E(\hat{y}-\theta)^2$未必是$\hat{y}$的方差。

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，而MSE又可分解成方差与偏差平方之和（命题2），所以需要折中预测统计量的方差和偏差。

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 $a = E(\hat{y})$, $\text{mse}(\hat{y}) = E(\hat{y} - \theta)^2 = E\left(\hat{y} - a + a - \theta\right)^2$

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

注1：若 $\hat{y}$ 是 θ 的无偏估计，即 $E(\hat{y}) = \theta$，则 $\text{mse}(\hat{y}) = \text{var}(\hat{y})$。

注2：对无偏估计做适当压缩（从而有偏），通常能有效地减小方差。

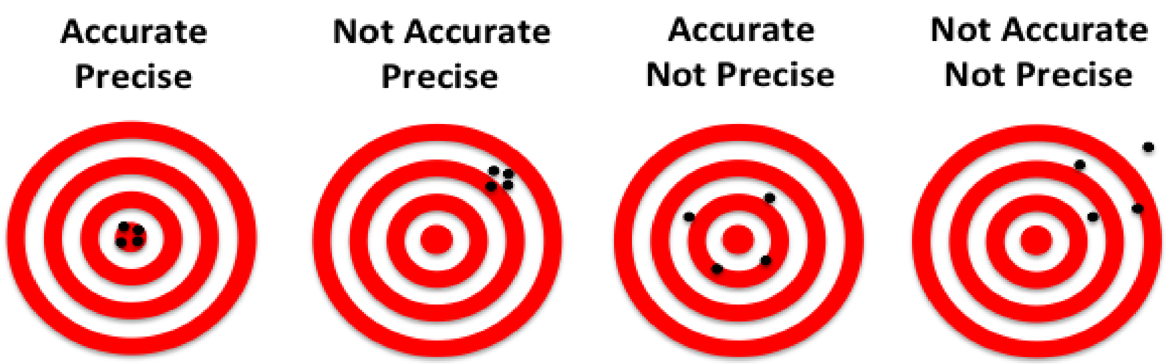
精准：
准确性+
精确性

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(难以实现)　　偏差大、方差小(如何减小偏差?)　　偏差小、方差大(如何减小方差?)　　偏差和方差都大

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

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

- 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。

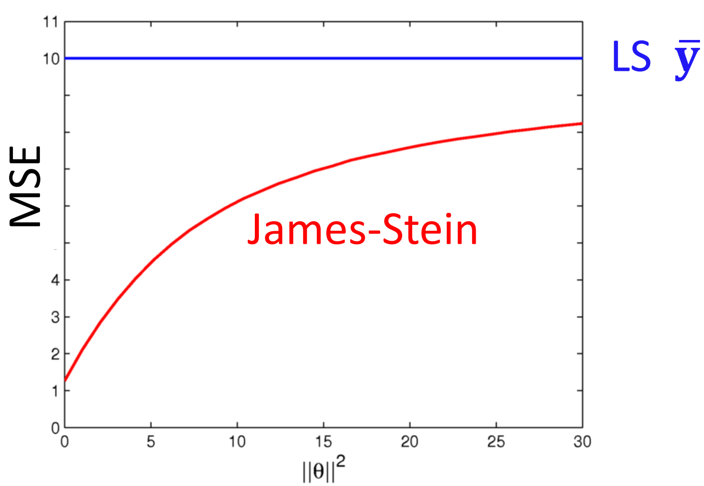
(James-Stein估计). 假设 $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ iid $\sim N(\boldsymbol{\theta}, \sigma^2 I_p)$， $p \geq 3$, 假设σ^2已知。定义$\bar{\mathbf{y}}$的一个压缩估计(有偏):

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

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。



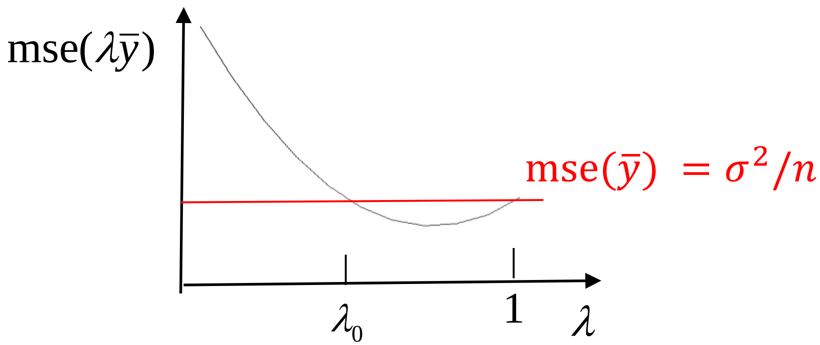
例1：假设$y_1,\ldots,y_n,y$ iid $\sim(\theta,\sigma^2)$，y是待预测随机变量。基于历史样本$y_1,\ldots,y_n$构造y的预测$\hat{y}$

(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：

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

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：

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

$\Rightarrow$ 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$y_1, y_2, \ldots, y_n$ 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