课程主页：http://staff.ustc.edu.cn/~ynyang/2023

第十讲 最小二乘法

2023.11.24

$g_n(\theta)=0$是一个合理的估计方程的必要条件是$Eg_n(\theta)=0$

多变量线性回归模型

多变量回归模型

多变量/多重线性回归模型(multiple linear regression model)总体模型：

$$y=\beta_0+x_1\beta_1+\ldots+x_{p-1}\beta_{p-1}+\varepsilon=\beta_0+\mathbf{x}^{\mathsf{T}}\mathbf{b}+\varepsilon,\quad \varepsilon\sim(0,\sigma^2),\varepsilon与\mathbf{x}独立.$$

假设样本 $(y_i,x_{i1},\ldots,x_{i,p-1})$，$i=1,2\ldots,n$ 独立，来自于上述总体模型，即

$$y_i=\beta_0+x_{i1}\beta_1+\ldots+x_{i,p-1}\beta_{p-1}+\varepsilon_i=\beta_0+\mathbf{x}_i^{\mathsf{T}}\mathbf{b}+\varepsilon_i,\qquad (1)$$

ε_i iid $\sim(0,\sigma^2)$，ε_i与$\mathbf{x}_i$独立.

记所有响应变量 $\mathbf{y}=(y_1,\ldots,y_n)^{\mathsf{T}}$，所有误差$\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^{\mathsf{T}}$，

所有回归系数$\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_{p-1})^{\mathsf{T}}=\begin{pmatrix}\beta_0\\ \mathbf{b}\end{pmatrix}$。所有自变量(包含1)

组成的矩阵称为设计阵(design matrix)：

$$X_{n\times p}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1,p-1}\\ 1 & x_{21} & \cdots & x_{2,p-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{n,p-1}\end{pmatrix}=\begin{pmatrix}1 & \mathbf{x}_1^{\mathsf{T}}\\ 1 & \mathbf{x}_2^{\mathsf{T}}\\ \vdots & \vdots\\ 1 & \mathbf{x}_n^{\mathsf{T}}\end{pmatrix}=\begin{pmatrix}\tilde{\mathbf{x}}_1^{\mathsf{T}}\\ \tilde{\mathbf{x}}_2^{\mathsf{T}}\\ \vdots\\ \tilde{\mathbf{x}}_n^{\mathsf{T}}\end{pmatrix}$$

矩阵-向量形式

模型（1）以矩阵-向量表达为

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 I_n), \ \boldsymbol{\varepsilon} \text{与} X \text{ 独立} \qquad (2)$$

注：模型也可写作：

$$\mathbf{y} = \mathbf{u} + \boldsymbol{\varepsilon}, \ \mathbf{u} \in C(X)$$

记X的各列为$X = (\mathbf{1}, \mathbf{x}_{(1)}, ..., \mathbf{x}_{(p-1)})$, 模型为

$$\mathbf{y} = \mathbf{u} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + \mathbf{x}_{(1)}\beta_1 + ... + \mathbf{x}_{(p-1)}\beta_{p-1} + \boldsymbol{\varepsilon}$$

例1. 假设 (y_i, x_i), $i = 1,2..., n$ 独立，满足简单线性回归模型：

$$y_i = a + bx_i + \varepsilon_i,$$

记$\boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \mathbf{y} = (y_1, ..., y_n)^{\mathsf{T}}$, 设计阵 $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = (\mathbf{1}, \mathbf{x})$

模型为 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}a + \mathbf{x}b + \boldsymbol{\varepsilon}$

最小二乘法 (LS: Least Squares)

最小化误差平方和

为了估计参数$\boldsymbol{\beta}$，最小二乘法最小化误差平方和：

$$\min_{\boldsymbol{\beta}\in R^p}\sum\varepsilon_i^2=\min_{\boldsymbol{\beta}\in R^p}\sum\left(y_i-\beta_0-\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right)^2=\min_{\boldsymbol{\beta}\in R^p}\|\boldsymbol{\varepsilon}\|^2=\min_{\boldsymbol{\beta}\in R^p}\|\mathbf{y}-X\boldsymbol{\beta}\|^2$$

使得平方和达到最小值的$\hat{\boldsymbol{\beta}}$称为最小二乘(LS)估计。

定理1. 假设线性回归模型满足Gauss - Markov(GM)假设：

$$\mathbf{y}_{n\times1}=X_{n\times p}\boldsymbol{\beta}_{p\times1}+\boldsymbol{\varepsilon}_{n\times1},\quad \boldsymbol{\varepsilon}\sim(\mathbf{0},\sigma^2I_n),\ \boldsymbol{\varepsilon}\text{与}X\text{ 独立}$$

最小二乘估计为$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$，满足正则方程（估计方程）

$$X^{\mathsf{T}}\boldsymbol{\varepsilon}=X^{\mathsf{T}}(\mathbf{y}-X\boldsymbol{\beta})=\mathbf{0}$$

若X列满秩，则LS估计$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$唯一，且$\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的无偏估计。

无偏性是LS估计最重要的性质，无偏性说明在GM假设成立的条件下，基于LS方法能得到参数正确的估计。LS估计的另一个重要性质是在所有线性无偏估计中其方差最小（GM定理，后面）

证法1：投影

证：由第9讲定理2(投影的最小二乘性质，P23),

$$\min_{\boldsymbol{\beta}\in R^p} \|\mathbf{y}-X\boldsymbol{\beta}\|^2 = \min_{\mathbf{u}\in C(X)} \|\mathbf{y}-\mathbf{u}\|^2$$

最小值在$\mathbf{u}$等于$\mathbf{y}$在$C(X)$上的正交投影时达到最小，投影

$$\hat{\mathbf{y}} = P_X\mathbf{y} = X\underline{\underline{(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}}} = X\underline{\hat{\boldsymbol{\beta}}}$$

$\Rightarrow LS$估计$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$

$\mathbf{y}-\hat{\mathbf{y}} \perp C(X)$,特别地与$X$的每一列正交，满足正则方程：

$$X^{\mathsf{T}}(\mathbf{y}-\hat{\mathbf{y}}) = X^{\mathsf{T}}(\mathbf{y}-X\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

若X列满秩，即若$X^{\mathsf{T}}X$可逆，则$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$唯一.

因为$\boldsymbol{\varepsilon}$与X独立，所以

$$E(\hat{\boldsymbol{\beta}}\mid X) = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}E(\mathbf{y}\mid X) = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}E(X\boldsymbol{\beta}+\boldsymbol{\varepsilon}\mid X)$$
$$= \boldsymbol{\beta} + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}E(\boldsymbol{\varepsilon}\mid X) = \boldsymbol{\beta} + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}$$

$$\Rightarrow E(\hat{\boldsymbol{\beta}}) = E(E(\hat{\boldsymbol{\beta}}\mid X)) = \boldsymbol{\beta}.$$

引理. (矩阵向量求导)

(1) 若$\mathbf{a}, \mathbf{x} \in R^n$, 则$\frac{\partial(\mathbf{a}^\top \mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$;

(2) 若A为$n \times n$对称矩阵, $\mathbf{x} \in R^n$, 则$\frac{\partial(\mathbf{x}^\top A\mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x}$。

证法2: 求导

目标函数$Q(\boldsymbol{\beta}) = \| \mathbf{y} - X\boldsymbol{\beta} \|^2 = \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}$,

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2(\mathbf{y}^\top X)^\top + 2X^\top X\boldsymbol{\beta} = -2X^\top(\mathbf{y} - X\boldsymbol{\beta}) = -2X^\top \boldsymbol{\varepsilon}$$

令之为0, 得正则方程

$$X^\top \boldsymbol{\varepsilon} = X^\top(\mathbf{y} - X\boldsymbol{\beta}) = 0 \iff X^\top X\boldsymbol{\beta} = X^\top \mathbf{y}$$

$\Rightarrow$ LS估计$\hat{\boldsymbol{\beta}} = (X^\top X)^- X^\top \mathbf{y}$.

注: 也可以展开平方和求导 (避免矩阵向量求导)

记X的各行为$X = (\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_n)^\top$,

$$Q(\boldsymbol{\beta}) = \| \mathbf{y} - X\boldsymbol{\beta} \|^2 = \sum_{i=1}^n \left(y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} \right)^2, \quad \text{令} \frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^n \tilde{\mathbf{x}}_i \left(y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} \right) = 0$$

即 $\sum_{i=1}^n \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} = \sum_{i=1}^n \tilde{\mathbf{x}}_i y_i \Rightarrow \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \right)^- \left(\sum_{i=1}^n \tilde{\mathbf{x}}_i y_i \right) = \left(X^\top X \right)^- X^\top \mathbf{y}$

正则方程

LS估计是正则方程

$$X^{\top}\boldsymbol{\varepsilon} = X^{\top}(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$$

的解，该方程与模型假设$X \perp\!\!\!\perp \boldsymbol{\varepsilon}$是相容的， 这是因为

$$X \perp\!\!\!\perp \boldsymbol{\varepsilon} \Rightarrow E(X^{\top}\boldsymbol{\varepsilon}) = 0 \Rightarrow \text{方程} X^{\top}\boldsymbol{\varepsilon} = \mathbf{0}\text{合理}$$

一般地，假设基于n个独立样本的估计方程为

$$g_n(\theta) = 0$$

该方程的解$\hat{\theta}$为参数θ的估计（比如似然方程 $g_n(\theta) = \frac{\partial logL(\theta)}{\partial \theta} = 0$）。

$\hat{\theta}$是无偏估计或近似无偏估计的一个必要条件是

$$Eg_n(\theta) = 0$$

直观上，$g_n(\hat{\theta}) = 0$且$E(\hat{\theta}) = \theta \Rightarrow g_n(\theta) \approx 0$，平均来看$Eg_n(\theta) = 0$.

细节：假设 $g_n(\theta) = 0$的解为 $\hat{\theta}$，Taylor展开

$$\begin{aligned} 0 = g_n(\hat{\theta}) &\approx g_n(\theta) + g_n'(\theta)(\hat{\theta} - \theta) \\ &\approx g_n(\theta) + G_n(\theta)(\hat{\theta} - \theta) \end{aligned}$$

其中$g_n'(\theta)$一般是n项独立和，由大数定律$g_n'(\theta) \approx E[g_n'(\theta)] \triangleq G_n(\theta)$

$\Rightarrow \hat{\theta} - \theta \approx G_n(\theta)^{-1}g_n(\theta)$，$E(\hat{\theta} - \theta) \approx G_n(\theta)^{-1}E(g_n(\theta)) = 0$.

LS估计- 矩方法观点（moment method)

模型假设：$y_i = \beta_0 + \mathbf{x}_i^{\mathsf{T}}\mathbf{b} + \varepsilon_i = \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i$，$\varepsilon_i$ 与$\tilde{\mathbf{x}}_i$独立，$\varepsilon_i \sim (0, \sigma^2)$ 。

$\Leftrightarrow$以矩阵-向量表达：　$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$, $\boldsymbol{\varepsilon}$与X独立。

矩方法

因为 $E(\tilde{\mathbf{x}}_i\varepsilon_i) = 0$, 矩方法令样本矩$\sum_{i=1}^{n} \tilde{\mathbf{x}}_i\varepsilon_i / n$等于0:

$$令 \sum_{i=1}^{n} \tilde{\mathbf{x}}_i\varepsilon_i = X^{\mathsf{T}}\boldsymbol{\varepsilon} = X^{\mathsf{T}}(\mathbf{y} - X\boldsymbol{\beta}) = 0,$$

即LS的正则方程 $X^{\mathsf{T}}\mathbf{y} = X^{\mathsf{T}}X\boldsymbol{\beta} = 0 \Rightarrow \hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$

LS方法或矩方法也可简单地描述为:

模型 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 两边同时左乘X^{T}

$$X^{\mathsf{T}}\mathbf{y} = X^{\mathsf{T}}X\boldsymbol{\beta} + X^{\mathsf{T}}\boldsymbol{\varepsilon}$$

并舍弃$X^{\mathsf{T}}\boldsymbol{\varepsilon}$一项$(令X^{\mathsf{T}}\boldsymbol{\varepsilon} = \mathbf{0})$，得正则方程$X^{\mathsf{T}}\mathbf{y} = X^{\mathsf{T}}X\boldsymbol{\beta}$。

其它估计方程

除了同时左乘X^{T}，是否还有其它选择？

显然方程两边同时左乘$p\times n$矩阵$Z^{\mathsf{T}}=f(X^{\mathsf{T}})$:

$$Z^{\mathsf{T}}\mathbf{y}=Z^{\mathsf{T}}X\boldsymbol{\beta}+Z^{\mathsf{T}}\boldsymbol{\varepsilon}$$

并令$Z^{\mathsf{T}}\boldsymbol{\varepsilon}=0$,即$Z^{\mathsf{T}}\mathbf{y}=Z^{\mathsf{T}}X\boldsymbol{\beta}$，也是一个合理的方程，这是因为

$\boldsymbol{\varepsilon}$与$Z^{\mathsf{T}}=f(X^{\mathsf{T}})$独立$\Rightarrow E(Z^{\mathsf{T}}\boldsymbol{\varepsilon})=0$。

事实上，任何外生的Z (即Z与$\boldsymbol{\varepsilon}$独立)都满足$E(Z^{\mathsf{T}}\boldsymbol{\varepsilon})=0$,导致合理的方程:

$$Z^{\mathsf{T}}\boldsymbol{\varepsilon}=0\Leftrightarrow Z^{\mathsf{T}}\mathbf{y}=Z^{\mathsf{T}}X\boldsymbol{\beta}$$

进一步，若$C(Z^{\mathsf{T}}X)=C(Z^{\mathsf{T}})$,则方程有解；

若$Z^{\mathsf{T}}X$可逆，则有唯一解

$$\tilde{\boldsymbol{\beta}}=(Z^{\mathsf{T}}X)^{-1}Z^{\mathsf{T}}\mathbf{y}$$

它是无偏的。

$$E(\tilde{\boldsymbol{\beta}}\,|\,Z,X)=(Z^{\mathsf{T}}X)^{-1}Z^{\mathsf{T}}E(X\boldsymbol{\beta}+\boldsymbol{\varepsilon}\,|\,Z,X)$$
$$=\boldsymbol{\beta}+(Z^{\mathsf{T}}X)^{-1}Z^{\mathsf{T}}E(\boldsymbol{\varepsilon}\,|\,Z,X)=\boldsymbol{\beta}.$$

有趣的是，即使线性模型的Gauss-Markov假设不成立（一般如此!），上述矩方法仍能产生合理的估计方程，并得到具有优良性质的估计。

工具变量法

假设 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 中X与$\boldsymbol{\varepsilon}$不独立，此时正则方程$X^{\mathsf{T}}\boldsymbol{\varepsilon} = \mathbf{0}$不合理。

假设存在外生的 Z（称为工具变量）满足

$$Z \perp\!\!\!\perp \boldsymbol{\varepsilon}，且\ Z与X有关$$

则$E(Z^{\mathsf{T}}\boldsymbol{\varepsilon}) = \mathbf{0}$成立，那么下述估计方程是合理的

$$Z^{\mathsf{T}}\boldsymbol{\varepsilon} = Z^{\mathsf{T}}(\mathbf{y} - X\boldsymbol{\beta}) = 0$$

由此解得$\tilde{\boldsymbol{\beta}} = (Z^{\mathsf{T}}X)^{-1}Z^{\mathsf{T}}\mathbf{y}$，称为工具变量最小二乘估计(参见附录)。


Ill. Niklas Elmehed © Nobel Prize Outreach.
David Card
Prize share: 1/2


Ill. Niklas Elmehed © Nobel Prize Outreach.
Joshua D. Angrist
Prize share: 1/4


Ill. Niklas Elmehed © Nobel Prize Outreach.
Guido W. Imbens
Prize share: 1/4

2021诺贝尔经济奖：

J. Agrist, G.Imbens, D.Card.

对于特定的经济社会问题，发现了一些工具变量。

LS估计 - 解方程的观点

我们也可以将线性模型参数估计问题看作是解近似方程的问题，这与矩方法类似，但可以借鉴解方程的方法处理传统矩方法或LS方法难以处理的问题，比如$p > n$的情形。

超定方程
$n > p$

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$可看作是超定方程(over-determined system):

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1}, \quad n\text{个方程} > p\text{个未知量}$$

只有当$\mathbf{y} \in C(X)$时才有解(一般不可能)。

- 我们可求最优近似解使误差$\|\mathbf{y} - X\boldsymbol{\beta}\|^2$ 最小，这是LS方法。

- 也可以设法减少方程个数，即通过整合方程(比如合并某些方程、删除冗余方程,称为Pre-conditioning）的方式减少方程个数，　这等价于方程两边同时做某个Z^{T}-线性变换（Z^{T}：$k\times n$），得

$$Z^{\mathsf{T}}\mathbf{y} = Z^{\mathsf{T}}X\boldsymbol{\beta},$$

由前述讨论知Z的选取不能太随意：Z应与X有关，且Z外生（随机）。

欠定定程
$n < p$

$p > n$时，模型$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$中求解参数$\boldsymbol{\beta}$可看作是解欠定方程(underdetermined system)

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1}, \quad n < p,$$

通常有无穷多解，为了求出有意义的解，通常对解施加某些限制，比如

- 不定方程：数论中限制线性方程的解为正整数或有理数；
- 主成分回归：极小化欧氏模长$\|\boldsymbol{\beta}\|_2$，$\tilde{\boldsymbol{\beta}} = X^{+}\mathbf{y}$;
- 压缩感知或lasso：极小化$\|\boldsymbol{\beta}\|_0$ (非0的个数)或其放松$\|\boldsymbol{\beta}\|_1$。

误差方差的LS估计

拟合值 残差

- 投影$\hat{\mathbf{y}} = P_X\ \mathbf{y} = X\hat{\boldsymbol{\beta}} = X(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$,称为拟合值向量。
- $\mathbf{e} = \mathbf{y}^{\perp} = \mathbf{y} - \hat{\mathbf{y}} = (I_n - P_X)\mathbf{y}$称为残差向量。

$\hat{\boldsymbol{\beta}}$满足正则方程：$X^{\mathsf{T}}(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = X^{\mathsf{T}}\mathbf{e} = 0$, 即$\mathbf{e} \perp C(X)$,
特别地$\mathbf{e} \perp \mathbf{1}$, 即$\bar{e} = \mathbf{1}^{\mathsf{T}}\mathbf{e}/n = 0$ (样本均值为0)

误差方差估计

残差平方和：$RSS = \|\mathbf{e}\|^2 = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 = \mathbf{y}^{\mathsf{T}}(I_n - P_X)\mathbf{y}$

σ^2的LS估计定义为：$\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n-p} = \dfrac{1}{n-p}\|\mathbf{e}\|^2$。

LS估计的统计性质

模型假设$\varepsilon_1, \ldots, \varepsilon_n$ iid $\sim(0, \sigma^2)$ 反映了我们对误差的认知：

$\varepsilon_1, \ldots, \varepsilon_n$大致相同，$\varepsilon_i \approx 0$且$|\varepsilon_i| \approx \sigma$

即使没有这些概率模型假设，我们也能操作LS。有了这些假设，我们可评价LS方法，比如，LS估计$\hat{\boldsymbol{\beta}}$在平均意义下是否等于或接近真正的$\boldsymbol{\beta}$？其精度/方差如何？逼近误差RSS大概多大？

第6讲引理1(2)
$RSS = s_{yy} - s_{xy}^2/s_{xx}$
$= s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2/s_{xx}$
从投影的观点是显然的

例如，$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$中，

$$\mathrm{RSS} = \|\mathbf{e}\|^2 = \mathbf{y}^{\mathsf{T}}(I_n - P_X)\mathbf{y} = (X\boldsymbol{\beta} + \boldsymbol{\varepsilon})^{\mathsf{T}}(I_n - P_X)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^{\mathsf{T}}(I_n - P_X)\boldsymbol{\varepsilon},$$

平均来看（当X给定时）：

$$E(\mathrm{RSS}) = E\boldsymbol{\varepsilon}^{\mathsf{T}}(I_n - P_X)\boldsymbol{\varepsilon} = tr\big((I_n - P_X)\,\mathrm{var}(\boldsymbol{\varepsilon})\big) = (n - r)\sigma^2$$

其中$r = tr(P_X) = rank(X).$

引理（下页）

引理1：若$\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$，则$E(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = \boldsymbol{\mu}^{\mathsf{T}} A\boldsymbol{\mu} + tr(A\Sigma).$

证明：$E(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = Etr(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = Etr(A\mathbf{x}\mathbf{x}^{\mathsf{T}}) = trE(A\mathbf{x}\mathbf{x}^{\mathsf{T}}) = trAE(\mathbf{x}\mathbf{x}^{\mathsf{T}}),$

因为$\Sigma = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} = E(\mathbf{x}\mathbf{x}^{\mathsf{T}}) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}$

所以$E(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = trA(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}) = tr(A\Sigma) + trA\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}} = tr(A\Sigma) + \boldsymbol{\mu}^{\mathsf{T}} A\boldsymbol{\mu}$

无偏性

定理1. 假设线性模型$\mathbf{y} = X_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，$E\boldsymbol{\varepsilon} = 0, \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$，其中$\boldsymbol{\varepsilon}$与$X$独立。假设$X$是列满秩的（$n \geq p$, 且$X^{\mathsf{T}}X$可逆），则

(1) LS估计的无偏性：$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$

（2）LS估计的方差：$\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \sigma^2 \left(X^{\mathsf{T}}X\right)^{-1}.$

（3）误差方差估计的无偏性：$E(\hat{\sigma}^2) = \sigma^2$。

证明：因为$\boldsymbol{\varepsilon}$与X独立, 所以

- $E(\mathbf{y} \mid X) = E(X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \mid X) = X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon} \mid X) = X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = X\boldsymbol{\beta}$
- $\mathrm{var}(\mathbf{y} \mid X) = \mathrm{var}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \mid X) = \mathrm{var}(\boldsymbol{\varepsilon} \mid X) = \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$

(1) $E(\hat{\boldsymbol{\beta}} \mid X) = E\left(\left(X^{\top}X\right)^{-1}X^{\top}\mathbf{y} \mid X\right) = \left(X^{\top}X\right)^{-1}X^{\top}\left[E(\mathbf{y} \mid X)\right] = \boldsymbol{\beta} \Rightarrow E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$

（2） $\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \mathrm{var}(\left(X^{\top}X\right)^{-1}X^{\top}\mathbf{y} \mid X)$

$= \left(X^{\top}X\right)^{-1}X^{\top}\left[\mathrm{var}(\mathbf{y} \mid X)\right]X\left(X^{\top}X\right)^{-1} = \sigma^2\left(X^{\top}X\right)^{-1}$

（3）前一页我们已经证明了$E(RSS \mid X) = (n-p)\sigma^2$,

$\Rightarrow E(RSS) = (n-p)\sigma^2$

$\Rightarrow E(\hat{\sigma}^2) = E(RSS/(n-p)) = \sigma^2.$

$\hat{\beta}$的方差估计

在方差公式$\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \sigma^2\left(X^{\top}X\right)^{-1}$中代入（Plug-in）$\sigma^2$的估计,

即得到$\hat{\boldsymbol{\beta}}$方差的估计：

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}} \mid X) = \hat{\sigma}^2\left(X^{\top}X\right)^{-1}$$

例1(续). 假设 (y_i, x_i)，$i = 1,2...,n$ 独立，满足简单线性回归模型：

$$y_i = a + bx_i + \varepsilon_i ,$$

第6讲命题1、2中我们已经求得LS估计及其方差，这里我们以矩阵向量形式再次计算如下：

记 $\boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \mathbf{y} = (y_1,..., y_n)^{\mathsf{T}}$，设计阵 $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = (\mathbf{1}, \mathbf{x})$，$\mathbf{x} \neq \mathbf{1}c$,

模型为 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}a + \mathbf{x}b + \boldsymbol{\varepsilon}$，LS估计

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \Sigma x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ \Sigma x_i y_i \end{pmatrix}$$

$$= \frac{1}{n(\Sigma x_i^2 - n\bar{x}^2)} \begin{pmatrix} \Sigma x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \Sigma x_i y_i \end{pmatrix} = \begin{pmatrix} \bar{y} - \bar{x}s_{xy}/s_{xx} \\ s_{xy}/s_{xx} \end{pmatrix}$$

$$\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \sigma^2 (X^{\mathsf{T}}X)^{-1} = \frac{\sigma^2}{ns_{xx}} \begin{pmatrix} \Sigma x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

$$= \sigma^2 \begin{pmatrix} 1/n + \bar{x}^2/s_{xx} & -\bar{x}/s_{xx} \\ -\bar{x}/s_{xx} & 1/s_{xx} \end{pmatrix}$$

定义：对任何两个对称 $n\times n$矩阵A,B，若$A-B\geq 0$ (非负定)，则称在Loewner偏序意义下A不小于B，记作$A\geq B$.

性质：若$A\geq B$，则对任何$k\times n$矩阵C，$CAC^{\mathsf{T}}\geq CBC^{\mathsf{T}}$

LS估计的最优性

定理2(Gauss - Markov定理). 线性模型 $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$中,假设$X_{n\times p}$列满秩，则$\boldsymbol{\beta}$的最小二乘估计$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$是最优无偏线性估计(BLUE, best linear unbiased estimate,),即对任何$\boldsymbol{\beta}$的线性无偏估计$\tilde{\boldsymbol{\beta}}=C\mathbf{y}$

$$\mathrm{var}(\tilde{\boldsymbol{\beta}})\geq \mathrm{var}(\hat{\boldsymbol{\beta}}),$$

其中C是仅与X有关的$p\times n$常数矩阵。

证明：由$\tilde{\boldsymbol{\beta}}=C\ \mathbf{y}$的无偏性(注意$C$仅与$X$有关)

$$\boldsymbol{\beta}=\mathrm{E}(\tilde{\boldsymbol{\beta}}\mid X)=E(C\mathbf{y}\mid X)=CE(\mathbf{y}\mid X)=CX\boldsymbol{\beta},$$

上式对任何$\boldsymbol{\beta}$成立，故$CX=I_p$。

因为$\mathrm{var}(\tilde{\boldsymbol{\beta}}) = E\mathrm{var}(\tilde{\boldsymbol{\beta}} \mid X) + \mathrm{var}(E(\tilde{\boldsymbol{\beta}} \mid X)) = E\mathrm{var}(\tilde{\boldsymbol{\beta}} \mid X)$，

$\mathrm{var}(\hat{\boldsymbol{\beta}}) = E\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X)$，我们只需要证明：

$\mathrm{var}(\tilde{\boldsymbol{\beta}} \mid X) = C\,\mathrm{var}(\mathbf{y} \mid X)C^{\mathsf{T}} = \sigma^2 CC^{\mathsf{T}} \geq \mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \sigma^2 (X^{\mathsf{T}}X)^{-1}$。

即$CC^{\mathsf{T}} \geq (X^{\mathsf{T}}X)^{-1}$.

因为$CX = I_p$，$P_X \leq I_n$,所以

$$(X^{\mathsf{T}}X)^{-1} = CX(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}C^{\mathsf{T}} = CP_X C^{\mathsf{T}} \leq CC^{\mathsf{T}}.$$

所以$\mathrm{var}(\tilde{\boldsymbol{\beta}} \mid X) \geq \mathrm{var}(\hat{\boldsymbol{\beta}} \mid X)$，$\mathrm{var}(\tilde{\boldsymbol{\beta}}) \geq \mathrm{var}(\hat{\boldsymbol{\beta}})$.

推论：假设$A_{k\times p}$是仅与X有关的常数矩阵，则$A\hat{\boldsymbol{\beta}}$是$A\boldsymbol{\beta}$的BLUE。

特别地，$\hat{\beta}_k$是β_k的BLUE，$\hat{\beta}_k - \hat{\beta}_j$是$\beta_k - \beta_j$的BLUE，等等。

不确定性原理

对于简单线性模型 $y_i = a + bx_i + \varepsilon_i$ 的任一无偏估计 $\tilde{b}$，由GM定理

$$\mathrm{var}(\tilde{b} \mid x) \geq \mathrm{var}(\hat{b} \mid x) = \sigma^2 / s_{xx}$$

记样本方差 $s_x^2 = s_{xx}/(n-1)$，上式等价于

$$\mathrm{var}(\tilde{b} \mid x) \times s_x^2 \geq \sigma^2/(n-1),$$

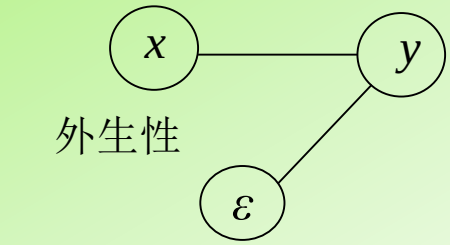
这说明左端两个方差不可能同时很小，即 x 和 $\tilde{b}$ 不可能同时精确测量 (Uncertainty principle, 不确定性原理, 测不准原理).

附录: 工具变量法

工具变量法（Instrumental Variable method）试图在线性模型中误差与自变量不独立的情况下，求解回归系数的无偏估计，发现因果。

外生变量

如果线性模型 $y = a + bx + \varepsilon$ 中 x 不是研究对象本身固有而是外界随机赋予的，称x是外生的（exogenous），此时x与ε独立。外生性是推断x和y之间因果关系的关键条件。



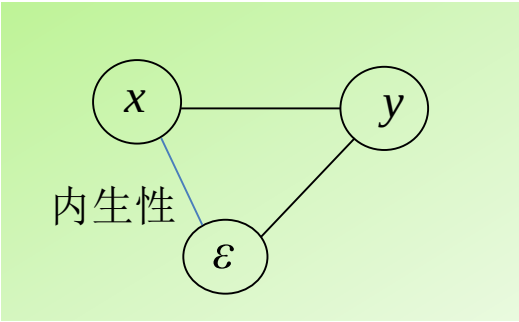
例2. 第1讲霍乱案例中，比较两个自来水公司客户的问题可表述成：

$$y = a + bx + \varepsilon$$

其中y为霍乱状态，x为饮用水来源（是否污染）， Snow 医生论证了x 是外生变量（随机取值,自然试验），因而b的LS估计是无偏的。

内生变量

观察研究中，模型 $y = a + bx + \varepsilon$ 中x是研究对象本身固有的，一般与ε不独立，称x是内生的 (endogenous)。此时如何推断x和y之间因果关系？



当x是内生变量时，可以尝试在回归模型中控制所有与x, y都有关系的干扰因素（非常困难）。计量经济学家发明了工具变量法（Wright, 1928），其关键在于发现恰当的自然试验 (natural experiment)，当然，这也是非常困难的。

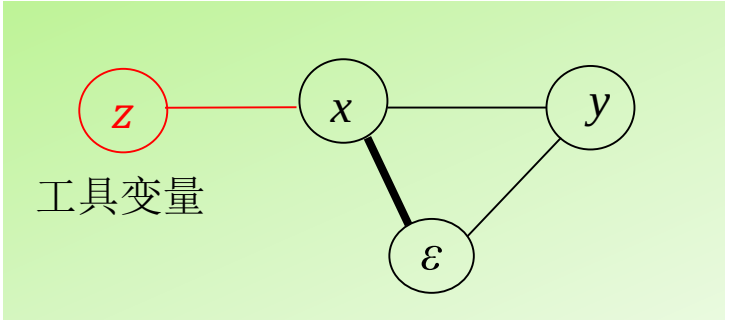
工具变量

线性模型$y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2)$, x与ε不独立（内生）。

假设存在一个自然试验(natural experiment)产生的外生变量z, 满足

z与 ε 独立（外生）， z与x相关，

z称为工具变量（instrumental variable，IV）。



工具变量z 满足条件：

(1) z与ε独立；

(2) z与x相关；

x与ε之间有联系，但z与ε独立。

工具变量法在研究y与x的关系时，以$\hat{x} = P_z x$替代x, 换言之，利用z 将x中与ε 有关的成分消除/清洗掉。

工具变量最小二乘

假设模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$, ε_i和x_i相关, $i = 1,\ldots,n$。

假设z_i是工具变量，满足条件：(1) z_i与ε_i独立；(2) z_i与x_i相关；

记$\mathbf{y} = (y_i)^{\mathsf{T}}, \mathbf{x} = (x_i)^{\mathsf{T}}, \mathbf{z} = (z_i)^{\mathsf{T}}, \boldsymbol{\varepsilon} = (\varepsilon_i)^{\mathsf{T}}, Z = (\mathbf{1}, \mathbf{z}), X = (\mathbf{1}, \mathbf{x})$,

模型：$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}a + \mathbf{x}b + \boldsymbol{\varepsilon}$。

$\boldsymbol{\varepsilon}$与Z独立$\Rightarrow E(Z^{\mathsf{T}}\boldsymbol{\varepsilon}) = 0$。

令$Z^{\mathsf{T}}\boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{1}^{\mathsf{T}}(\mathbf{y} - \mathbf{1}a - \mathbf{x}b) \\ \mathbf{z}^{\mathsf{T}}(\mathbf{y} - \mathbf{1}a - \mathbf{x}b) \end{pmatrix} = 0$，得工具变量最小二乘估计$(IVLS)$:

$$\tilde{b}_{IVLS} = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})} = \frac{s_{yz}}{s_{xz}}, \quad \tilde{a}_{IVLS} = \bar{y} - \tilde{b}\bar{x}.$$

等价地，也可从总体模型求出参数的表达，再代入样本：

总体模型$y = a + bx + \varepsilon$，因为z与ε独立，所以

$$0 = \mathrm{cov}(\varepsilon, z) = \mathrm{cov}(y - a - bx, z) = \mathrm{cov}(y, z) - b\,\mathrm{cov}(x, z)$$

所以，$b = \mathrm{cov}(y, z) / \mathrm{cov}(x, z)$,代入样本方差：$\tilde{b} = s_{yz} / s_{xz}$.

性质：$\hat{b}_{IV} = \frac{s_{yz}}{s_{xz}}$ 渐近无偏，即$E(\hat{b}_{IV}) \to b, n \to \infty$

两步估计1

注意到IVLS有下述表示，

$$\hat{b}_{IV} = \frac{s_{yz}}{s_{xz}} = \frac{s_{yz}/s_{zz}}{s_{xz}/s_{zz}},$$

所以，IVLS估计可由下述两步LS方法得到：

(1) $lm(x \sim z)$: $x = c + dz + \varepsilon^{(1)} \Rightarrow LS估计 \hat{d} = s_{xz}/s_{zz}$

(2) $lm(y \sim z)$: $y = e + fz + \varepsilon^{(2)} \Rightarrow LS估计 \hat{f} = s_{yz}/s_{zz}$

$\Rightarrow \hat{b}_{IV} = \hat{f}/\hat{d}.$

两步估计2

注意到IVLS估计也可表示如下

$$\hat{b}_{IV}=\frac{s_{yz}}{s_{xz}}=\frac{s_{yz}}{s_{zz}\left(s_{xz}/s_{zz}\right)}=\frac{s_{yz}}{s_{zz}\hat{d}}=\frac{\Sigma(\hat{d}\,z_i-\hat{d}\,\bar{z})(y_i-\bar{y})}{\Sigma(\hat{d}\,z_i-\hat{d}\,\bar{z})^2}=\frac{s_{y\hat{x}}}{s_{\hat{x}\hat{x}}}$$

其中$\hat{d}=s_{xz}/s_{zz}$，$\hat{x}_i=\hat{d}\,z_i$

所以，IVLS估计看作是下述两步LS估计：

---

(1) $lm(x\sim z)$：$x=c+dz+\varepsilon^{(1)}\Rightarrow LS$估计$\hat{d}=s_{xz}/s_{zz}$，

令$\hat{x}_i=\hat{d}z_i$

(2) $lm(y\sim\hat{x})$：$y=e+f\,\hat{x}+\varepsilon^{(2)}\Rightarrow LS$估计$\hat{f}=s_{y\hat{x}}/s_{\hat{x}\hat{x}}$这就是$\hat{b}_{IV}$

该观点反映了工具变量法的本质：

(1) 回归$y\sim x$不正确，因为x是内生的（x与ε相关）；

(2) 以$\hat{x}$替代x（$\hat{x}$与ε独立）；

(3) $y\sim\hat{x}$

例3. 2021诺贝尔经济奖获得者Agrist 研究了教育程度（x，受教育时长，月）是否与收入（y）存在因果关系(Angrist and Krueger, 1991) 。考虑线性模型

$$y = a + bx + \varepsilon,$$

与收入有关的其它因素（比如能力，家庭因素，努力程度等）无法全部测量或精确测量到，我们将它们放到 ε 中。显然，这些因素与自变量x有关，因此基于上述简单模型无法正确地（无偏）估计x的效应。

Angrist and Krueger 注意到小学入学时所有儿童年龄最大有1年的差距（当年12月达到6岁即可入学)

出生季节	1 (10-12月)	2 (1-3月)	3 (4-6月)	4 (7-9月)
入学年龄	5 ¾ -6	6 -6 ¼	6 ¼ - 6 ½	6 ½ - 6 ¾

而义务教育法规定青少年在16岁生日之前必须在校学习，所以对于16岁生日当天离开学校的那些学生而言，其受教育时长是由其生日 z 决定的，所以生日与x有关，但与ε无关（天然试验）, 所以 z 可作为工具变量

$$y = earnings$$
$$x = years\ of\ education$$
$$z = 出生季节$$

例4. 研究服兵役与否x (0或1) 对生活质量y的影响，建立模型

$$y = a + bx + \varepsilon$$

其中ε代表与y有关的个人和环境因素。志愿征兵制下入伍是个人的选择，因此x是内生的，基于LS方法得到的b的估计是有偏的/不正确的。

Angrist (1990) 发现了一个天然试验：1970's 越战期间，美国推行了基于“抽签”的强制征兵制度（draft lottery）。每个19-26岁符合兵役条件的男性被随机分配了一个号码RSN（1-365）。征兵前国家公布一个数字 T ，小于 T 的人将被强制列入入伍候选，再进行体检等其它程序。 1970，71，72年分别T=195，125，95。变量 $z = 1_{(RSN<T)}$是工具变量，它与兵役状态x相关，但与研究对象独立。

参考文献：

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*. June, 80:3, pp. 313–36.

Angrist, Joshua D. and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*. November, 106:4, pp. 979–1014.

Angrist, J. D., and Krueger, A. B. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *The Journal of Economic Perspectives* (15:4), pp. 69-85.

其它工具变量的例子(Angrsit and Krueger 2001)

Natural and Randomized Experiments

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*	*Reference*
		1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling-Sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)

常见的内生性原因

（0） 观察研究基本都有内生性。

(1) 丢失变量(Omitted variable,回归方程中没有控制相关变量)

假设正确模型为: $y = a + bx + cz + \delta, \delta \perp x$（$\perp$ 代表独立，下同）

其中 z 与 x 相关。如果我们没有测量,或者没有在上述

模型中控制z，那么工作模型 $y = a + bx + \varepsilon$中 $\varepsilon = cz + \delta$ 与 x 不独立。

（2） 因果颠倒(reverse causation,响应变量是自变量的原因)

假设正确模型为 $y = a + bx + \varepsilon$, $\varepsilon \perp x$

但实际操作中工作模型取为： $x = c + dy + \delta,$

从正确模型我们知道 $x = -a/b - y/b - \varepsilon/b,$ 所以工作模型中$\delta = -\varepsilon/b$与y有关。

(3) 自变量带误差模型 (Error in Variable，EV模型)

正确模型： $y = a + bx_0 + \varepsilon_0$， $x_0 \perp \varepsilon_0$. 假设 对于x_0的测量有误差，即我们

只能测量到: $x = x_0 + \delta,$其中$x_0 \perp \delta$.则

$$y = a + b(x - \delta) + \varepsilon_0 = a + bx + (\varepsilon_0 - b\delta) \hat{=} a + bx + \varepsilon$$

显然$x = x_0 + \delta$与$\varepsilon = \varepsilon_0 - b\delta$都含$\delta$，它们相关.