



# 目录

- 1 概率论与数理统计基础 ..... 1
  - 1.1 参数估计量的性质 ..... 1
  - 1.2 多维随机变量的分布 ..... 2
    - 1.2.1 随机向量及其数字特征 ..... 2
    - 1.2.2 多元正态分布 ..... 4
    - 1.2.3  $\chi^2$  分布 ..... 5
  - 1.3 条件期望 ..... 6
    - 1.3.1 条件期望和条件期望误差 ..... 6
    - 1.3.2 线性条件期望函数 ..... 9
  - 1.4 潜在结果框架 ..... 10
- 2 简单线性回归 ..... 14
  - 2.1 线性投影和线性回归 ..... 14
    - 2.1.1 线性投影模型 ..... 14
    - 2.1.2 线性回归模型 ..... 16
  - 2.2 回归系数的估计 ..... 17
    - 2.2.1 回归系数的期望和方差 ..... 18

2.2.2	估计量的大样本抽样分布	20
2.3	拟合优度	22
2.4	误差方差的估计	23
2.5	置信区间和假设检验	26
3	多元线性回归 .....	27
3.1	回归系数的估计	27
3.1.1	回归系数的期望和方差	28
3.1.2	回归系数的大样本抽样分布	29
3.2	分块回归和偏回归	29
3.2.1	解释变量个数不同时偏回归系数之间的关系	32
3.2.2	偏回归系数的期望和方差	32
3.2.3	偏回归系数的大样本抽样分布	33
3.3	拟合优度和误差方差的估计	33
3.4	正态回归	34
4	工具变量回归 .....	38
4.1	内生性	38
4.2	工具变量估计	39
4.3	两阶段最小二乘	39
4.4	工具变量检验	40
4.4.1	相关性检验	40
4.4.2	外生性检验	40
5	面板数据回归 .....	42
5.1	固定效应模型	42
5.1.1	个体固定效应	42
5.1.2	时间固定效应	43
5.2	随机效应模型	43
5.3	双重差分法	44

# 1. 概率论与数理统计基础

1.1	参数估计量的性质	1
1.2	多维随机变量的分布	2
1.2.1	随机向量及其数字特征	2
1.2.2	多元正态分布	4
1.2.3	$\chi^2$ 分布	5
1.3	条件期望	6
1.3.1	条件期望和条件期望误差	6
1.3.2	线性条件期望函数	9
1.4	潜在结果框架	10

## 1.1 参数估计量的性质

定义 1.1.1 定义

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

为估计量的偏差.

- 当  $\text{bias}(\hat{\theta}) = 0$  时,  $\hat{\theta}$  是  $\theta$  的无偏估计;
- 当  $\lim_{n \rightarrow +\infty} \text{bias}(\hat{\theta}) = 0$  时,  $\hat{\theta}$  是  $\theta$  的渐进无偏估计;
- 当  $\lim_{n \rightarrow +\infty} P\{|\hat{\theta} - \theta| \geq \epsilon\} = 0$  时,  $\hat{\theta}$  是  $\theta$  的一致估计.

定义 1.1.2 定义

$$\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

为估计量的方差.

对于  $\hat{\theta}$  的两个无偏估计量  $\hat{\theta}_1$  和  $\hat{\theta}_2$ , 当  $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ , 则称  $\hat{\theta}_1$  相对于  $\hat{\theta}_2$  是有效的; 若  $\hat{\theta}_0$  是所有  $\hat{\theta}$  的无偏估计量中方差最小的, 则称  $\hat{\theta}_0$  是  $\hat{\theta}$  的最优无偏估计量.

定义 1.1.3 定义

$$\text{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

为估计量的均方误差.

定理 1.1.1 (偏差-方差权衡)

$$\text{mse}(\hat{\theta}) = \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})$$

证

$$\begin{aligned}\text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) &= \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}[\theta])^2 + (\mathbb{E}[\theta])^2 - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 \\ &= \mathbb{E}[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2] \\ &= \text{mse}(\hat{\theta})\end{aligned}$$

□

## 1.2 多维随机变量的分布

### 1.2.1 随机向量及其数字特征

**定义 1.2.1** 如果  $X_1, X_2, \dots, X_n$  都是随机变量, 则称  $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$  为  $n$  维随机向量, 简称随机向量.

**定义 1.2.2** (1) 设  $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$  是离散型随机向量, 则它的分布函数为

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\}$$

式中,  $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ , 并记成  $\mathbf{X} \sim F$ ;

(2) 设  $\mathbf{X} \sim F(\mathbf{x}) = F(x_1, x_2, \dots, x_n)$ , 且  $\mathbf{X}$  为连续型随机向量, 若存在一个非负的函数  $f(\cdot)$ , 使得

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} f(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n$$

对一切  $\mathbf{x} \in \mathbb{R}^n$  成立, 则称  $\mathbf{X}$  有分布密度函数  $f(\cdot)$ .

**定义 1.2.3** 设  $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$  有  $n$  个分量, 若为  $\mathbb{E}[X_i] = \mu_i (i = 1, 2, \dots, n)$  存在, 定义随机向量  $\mathbf{X}$  的均值为

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}$$

$\boldsymbol{\mu}$  是一个  $n$  维向量, 称为均值向量.

**推论 1.2.1** 当  $\mathbf{A}, \mathbf{B}$  为常数矩阵时, 由定义 1.2.3 可推出如下性质:

- (1)  $\mathbb{E}[\mathbf{AX}] = \mathbf{A}\mathbb{E}[\mathbf{X}]$
- (2)  $\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B}$

证

$$\text{设 } \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \mathbf{B} = (b_1, b_2, \dots, b_s),$$

则

$$\begin{aligned} \mathbb{E}[\mathbf{AX}] &= \begin{pmatrix} \sum_{i=1}^n \mathbb{E}[a_{1i}X_i] \\ \sum_{i=1}^n \mathbb{E}[a_{2i}X_i] \\ \vdots \\ \sum_{i=1}^n \mathbb{E}[a_{mi}X_i] \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n a_{1i}\mathbb{E}[X_i] \\ \sum_{i=1}^n a_{2i}\mathbb{E}[X_i] \\ \vdots \\ \sum_{i=1}^n a_{mi}\mathbb{E}[X_i] \end{pmatrix} = \mathbf{A}\mathbb{E}[\mathbf{X}] \\ \mathbb{E}[\mathbf{AXB}] &= \begin{pmatrix} \sum_{i=1}^n \mathbb{E}[a_{1i}X_i b_1] & \sum_{i=1}^n \mathbb{E}[a_{1i}X_i b_2] & \cdots & \sum_{i=1}^n \mathbb{E}[a_{1i}X_i b_s] \\ \sum_{i=1}^n \mathbb{E}[a_{2i}X_i b_1] & \sum_{i=1}^n \mathbb{E}[a_{2i}X_i b_2] & \cdots & \sum_{i=1}^n \mathbb{E}[a_{2i}X_i b_s] \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n \mathbb{E}[a_{mi}X_i b_1] & \sum_{i=1}^n \mathbb{E}[a_{mi}X_i b_2] & \cdots & \sum_{i=1}^n \mathbb{E}[a_{mi}X_i b_s] \end{pmatrix} \\ &= \begin{pmatrix} b_1 \sum_{i=1}^n a_{1i}\mathbb{E}[X_i] & b_2 \sum_{i=1}^n a_{1i}\mathbb{E}[X_i] & \cdots & b_s \sum_{i=1}^n a_{1i}\mathbb{E}[X_i] \\ b_1 \sum_{i=1}^n a_{2i}\mathbb{E}[X_i] & b_2 \sum_{i=1}^n a_{2i}\mathbb{E}[X_i] & \cdots & b_s \sum_{i=1}^n a_{2i}\mathbb{E}[X_i] \\ \vdots & \vdots & & \vdots \\ b_1 \sum_{i=1}^n a_{mi}\mathbb{E}[X_i] & b_2 \sum_{i=1}^n a_{mi}\mathbb{E}[X_i] & \cdots & b_s \sum_{i=1}^n a_{mi}\mathbb{E}[X_i] \end{pmatrix} \\ &= \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B} \end{aligned}$$

□

**定义 1.2.4** (1) 对于随机向量  $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ ,

$$\begin{aligned} \Sigma &= D(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \\ &= \begin{pmatrix} D(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & D(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & D(X_n) \end{pmatrix} \\ &= (\sigma_{ij})_{n \times n} \end{aligned}$$

为  $n$  维随机向量  $\mathbf{X}$  的协方差矩阵,  $|\Sigma|$  为  $\mathbf{X}$  的广义方差.

(2) 对于随机向量  $\mathbf{X} = (X_1, X_2, \dots, X_m)^T$  和  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ , 它们之间的协方差矩阵为

$$\Sigma_{m \times n} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T] = \text{cov}(\mathbf{X}, \mathbf{Y}) = (\text{cov}(X_i, Y_j))$$

$$i = 1, 2, \dots, m; j = 1, 2, \dots, n.$$



若  $\text{cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ , 则称  $\mathbf{X}$  和  $\mathbf{Y}$  不相关.

**推论 1.2.2** 当  $\mathbf{A}, \mathbf{B}$  为常数矩阵时, 由定义 1.2.4 可推出如下性质:

$$(1) D(\mathbf{AX}) = \mathbf{AD}(\mathbf{X})\mathbf{A}^T = \mathbf{A}\Sigma\mathbf{A}^T$$

$$(2) \text{cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{Acov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^T$$

证

$$D(\mathbf{AX}) = \mathbb{E}[(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])^T] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\mathbf{A}^T] = \mathbf{AD}(\mathbf{X})\mathbf{A}^T$$

$$\begin{aligned} \text{cov}(\mathbf{AX}, \mathbf{BY}) &= \mathbb{E}[(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])(\mathbf{BY} - \mathbb{E}[\mathbf{BY}])^T] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T\mathbf{B}^T] \\ &= \mathbf{Acov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^T \end{aligned}$$

□

**定义 1.2.5** 若随机向量  $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$  的协方差矩阵存在, 且每个分量的方差大于零, 则

$$\mathbf{R} = (\text{corr}(X_i, X_j)) = (r_{ij})_{n \times n}$$

为  $\mathbf{X}$  的相关矩阵, 其中

$$r_{ij} = \frac{\text{cov}(X_i, X_j)}{\sqrt{D(X_i)}\sqrt{D(X_j)}}, i, j = 1, 2, \dots, n$$

## 1.2.2 多元正态分布

**定义 1.2.6** 若随机向量  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  的概率密度函数为

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \Sigma > \mathbf{0}$$

则称  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  服从  $p$  元正态分布, 也称  $\mathbf{X}$  为  $p$  元正态变量, 记为:

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$$

当  $p = 2$  时, 可以得到二元正态分布的概率密度函数.

设  $\mathbf{X} = (X_1, X_2)^T$  服从二元正态分布, 则

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2r \\ \sigma_2\sigma_1r & \sigma_2^2 \end{pmatrix}$$

则

$$|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - r^2)$$

$$\Sigma^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - r^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1 \sigma_2 r \\ -\sigma_2 \sigma_1 r & \sigma_1^2 \end{pmatrix}$$

则二元正态变量  $\mathbf{X}$  的概率密度函数为

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2r \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\}$$

当  $r = 1$  时,

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left[ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right] \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left[ -\frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right] \\ &= f(x_1)f(x_2) \end{aligned}$$

此时  $X_1, X_2$  相互独立.

**推论 1.2.3** (多元正态分布的性质) 设  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ ,

(1)  $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}, D(\mathbf{X}) = \Sigma$ ;

(2) 设  $m$  维随机向量  $\mathbf{Z}_{m \times 1} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , 其中  $\mathbf{A}_{m \times p}$  为常数矩阵,  $\mathbf{b}$  为  $m$  维常向量, 则  $\mathbf{Z} \sim N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^T)$ .

### 1.2.3 $\chi^2$ 分布

**定义 1.2.7** 当  $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I}_n)$  时,  $\mathbf{X}^T \mathbf{X} \sim \chi^2(n)$ .

**推论 1.2.4** 若  $\mathbf{A}_{n \times n}$  为对称幂等矩阵,  $r(\mathbf{A}) = r$ , 当  $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I}_n)$  时,  $\mathbf{X}^T \mathbf{A} \mathbf{X} \sim \chi^2(r)$ .

**证** 因为  $\mathbf{A}$  为对称幂等矩阵, 则一定存在可逆矩阵  $\mathbf{P}_{n \times n}$ , 使得

$$\mathbf{A} = \mathbf{P}^T \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \mathbf{P}$$

令  $\mathbf{Y} = \mathbf{P}\mathbf{X}$ , 则

$$\mathbf{X}^T \mathbf{A} \mathbf{X} = \mathbf{X}^T \mathbf{P}^T \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \mathbf{P} \mathbf{X} = \mathbf{Y}^T \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \mathbf{Y} \sim \chi^2(r).$$

□

推论 1.2.5 当  $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$  时,  $(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n)$ .

## 1.3 条件期望

### 1.3.1 条件期望和条件期望误差

**定义 1.3.1** (1) 设  $X$  和  $Y$  的联合分布为离散分布, 对于  $P\{Y = y\} > 0$  的  $y$  值,  $X$  在给定  $Y = y$  之下的条件期望为

$$\mathbb{E}[X|Y = y] = \sum_x x P\{X = x|Y = y\} = \sum_x x \frac{P\{X = x, Y = y\}}{P\{Y = y\}}$$

(2) 设  $X$  和  $Y$  有连续型联合分布, 当  $f_Y(y) > 0$  时,  $X$  在给定  $Y = y$  之下的条件期望为

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx = \int_{-\infty}^{+\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx$$

给定一组随机变量  $Y_i (i = 1, 2, \dots, n)$ , 当  $Y_i$  取遍所有的可能值后, 得到的  $X_i$  的期望值便可以看成  $X_i$  关于  $Y_i$  的函数, 此时称  $\mathbb{E}[X_i|Y_i]$  为  $X_i$  关于  $Y_i$  的**条件期望函数** (conditional expectation function, 简称 CEF).

类似地, 可以定义条件方差:

$$\text{var}(X|Y = y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y = y].$$

#### 推论 1.3.1

$$\mathbb{E}[g(X, Y)|Y = y] = \begin{cases} \sum_x g(x, y) f_{X|Y}(x|y), & X \text{ 是离散型变量} \\ \int_{-\infty}^{+\infty} g(x, y) f_{X|Y}(x|y) dx, & X \text{ 是连续型变量} \end{cases}$$

#### 推论 1.3.2

$$\mathbb{E}[g(X)h(Y)|Y] = h(Y)\mathbb{E}[g(X)|Y]$$

**证** 考虑连续情形:

$$\begin{aligned} \mathbb{E}[g(X)h(Y)|Y = y] &= \int_{-\infty}^{+\infty} g(x)h(y)f_{X|Y}(x|y)dx \\ &= h(y) \int_{-\infty}^{+\infty} g(x)f_{X|Y}(x|y)dx \\ &= h(y)\mathbb{E}[g(X)|Y = y]. \end{aligned}$$



因为  $\mathbb{E}[g(X)h(Y)|Y]$  的每一个实现值 ( $\mathbb{E}[g(X)h(Y)|Y = y]$ ) 与  $h(Y)\mathbb{E}[g(X)|Y]$  的每一个实现值 ( $h(y)\mathbb{E}[g(X)|Y = y]$ ) 总是相等的, 所以有

$$\mathbb{E}[g(X)h(Y)|Y] = h(Y)\mathbb{E}[g(X)|Y]$$

□

**定理 1.3.1 (迭代期望法则)**

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

证 当  $X$  和  $Y$  均为离散型变量时,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \sum_y \sum_x x \frac{P\{X = x, Y = y\}}{P\{Y = y\}} P\{Y = y\} \\ &= \sum_y \sum_x x P\{X = x, Y = y\} \\ &= \sum_x x \sum_y P\{X = x, Y = y\} \\ &= \sum_x x P\{X = x\} \\ &= \mathbb{E}[X]\end{aligned}$$

当  $X$  和  $Y$  均为连续型变量时,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} x \left( \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy \right) dx \\ &= \int_{-\infty}^{+\infty} x f_X(x) dx \\ &= \mathbb{E}[X]\end{aligned}$$

□

**定理 1.3.2 (全方差法则)**

$$\text{var}(X) = \text{var}(\mathbb{E}[X|Y]) + \mathbb{E}[\text{var}(X|Y)]$$

证 注意到

$$\text{var}(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2$$

$$\mathbb{E}[\text{var}(X|Y)] = \mathbb{E}[\mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2] = \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2]$$

因此有

$$\text{var}(\mathbb{E}[X|Y]) + \mathbb{E}[\text{var}(X|Y)] = -(\mathbb{E}[X])^2 + \mathbb{E}[X^2] = \text{var}(X)$$

□

**定义 1.3.2** 定义

$$e = X - \mathbb{E}[X|Y]$$

为条件期望误差 (CEF error).

这里,  $e$  是由  $(X, Y)$  的联合分布决定的随机变量. 同时, 上式可以看作  $X$  的条件期望的分解, 即

$$X = \mathbb{E}[X|Y] + e.$$

**推论 1.3.3** (1)  $\mathbb{E}[e|Y] = \mathbb{E}[e] = 0$

(2)  $\text{cov}(e, g(Y)) = \mathbb{E}[eg(Y)] = \text{cov}(e, \mathbb{E}[X|Y]) = 0$

(3)  $\text{var}(e) = \mathbb{E}[\text{var}(X|Y)]$

**证** (1)  $\mathbb{E}[e|Y] = \mathbb{E}[X - \mathbb{E}[X|Y]|Y] = \mathbb{E}[X|Y] - \mathbb{E}[X|Y] = 0.$

$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e|Y]] = 0.$

(2)  $\text{cov}(e, g(Y)) = \mathbb{E}[eg(Y)] = \mathbb{E}[\mathbb{E}[eg(Y)|Y]] = \mathbb{E}[g(Y)\mathbb{E}[e|Y]] = 0.$

因为  $\mathbb{E}[X|Y]$  是关于  $Y$  的一个函数, 满足  $g(Y)$  的形式, 则有  $\text{cov}(e, \mathbb{E}[X|Y]) = 0.$

(3)  $\text{var}(e) = \mathbb{E}[e^2] = \mathbb{E}[\mathbb{E}[e^2|Y]] = \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]] = \mathbb{E}[\text{var}(X|Y)].$

□

**定理 1.3.3** (条件期望函数的预测性质) 对于给定的  $X$ , 当  $\mathbb{E}[Y^2]$  存在时,  $\mathbb{E}[Y|X]$  是对  $Y$  的最小均方误预测.

证 令  $g(X)$  为任意一个对  $Y$  的预测函数, 则

$$\begin{aligned}
 \text{mse}(g(X)) &= \mathbb{E}[(g(X) - Y)^2] \\
 &= \mathbb{E}[(g(X) - \mathbb{E}[Y|X] - e)^2] \\
 &= \mathbb{E}[(g(X) - \mathbb{E}[Y|X])^2] + 2\mathbb{E}[e(g(X) - \mathbb{E}[Y|X])] + \mathbb{E}[e^2] \\
 &= \mathbb{E}[(g(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[e^2] \\
 &\geq \mathbb{E}[e^2]
 \end{aligned}$$

当且仅当  $g(X) = \mathbb{E}[Y|X]$  时, “=” 成立. 因此有

$$\mathbb{E}[Y|X] = \arg \min \mathbb{E}[(g(X) - Y)^2] = \arg \min \text{mse}(g(X)).$$

□

### 1.3.2 线性条件期望函数

**定义 1.3.3** 当条件期望函数  $m(x) = \mathbb{E}[Y|X = x]$  满足

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = x^T \beta, \text{ 其中 } x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

时, 则称  $m(x)$  为线性条件期望函数 (Linear CEF).

如果关于  $X$  的条件期望函数是线性的, 对  $Y$  进行条件期望分解, 可得

$$Y = X^T \beta + e.$$

现在, 我们介绍 Linear CEF 的一个特例.

若  $X$  为分类数据, 在进行预测时, 我们需要对  $X$  进行赋值, 使  $X$  成为**虚拟变量** (Dummy Variables). 这里,  $X$  是离散型随机变量, 如果对于  $X$  所取的所有可能值,  $\beta$  中均有相应的参数与其对应, 我们称模型达到**饱和**. 此时, 可构造线性饱和虚拟变量模型, 即

$$m(\tilde{X}) = X^T \beta + e$$

其中, 设  $\tilde{X}_i (i = 1, 2, \dots, k)$  有  $n_i$  个取值, 则虚拟随机变量有  $N = \prod_{i=1}^k n_i$  个,  $X$  是一个  $N$  维向量. 这里,  $Y = X^T \beta + e$  称为**饱和虚拟变量回归模型**.

注意, 无论  $Y$  的分布如何, 上述模型都能完美地拟合 Linear CEF, 证明极其复杂, 这里不另加赘述.

## 1.4 潜在结果框架

注

潜在结果框架的三要素是：潜在结果、稳定性假设和随机分配。

在因果推断中，必须有干预，没有干预就没有因果，这里的干预可以是一项政策、一项措施或者一项活动等。以二值的干预变量为例，两个值分别对应于积极的行动和被动的行动，分别称为**处理**和**控制**，受到对应干预的个体分别称为**处理组**（Treatment Group）和**控制组**（Control Group）。

对应于每个干预状态，就有一个**潜在结果**。设  $D_i = 1$  为积极干预状态， $D_i = 0$  为控制干预状态，对应的潜在结果分别为  $Y_{1i}$  和  $Y_{0i}$ ，观测结果为  $Y_i$ ，则有

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = \begin{cases} Y_{1i}, & \text{如果 } D_i = 1 \\ Y_{0i}, & \text{如果 } D_i = 0 \end{cases}$$

**定义 1.4.1** 定义

$$\rho_i = Y_{1i} - Y_{0i}$$

为**个体因果效应**。如果  $\rho_i$  对所有个体而言相同，则称个体因果效应是**同质的**；如果  $\rho_i$  随  $i$  的变化而变化，则称个体因果效应是**异质的**。

然而，个体的潜在结果只能观测到其中一个，个体因果效应无法直接计算，因此我们关注的是平均因果效应。

**定义 1.4.2** 定义

$$\rho = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$$

为**平均因果效应**（Average Treatment Effect, ATE）。

相应地，我们可以定义处理组的平均因果效应为  $\rho_T = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$ ，控制组的平均因果效应为  $\rho_C = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 0]$ 。这样，平均因果效应可以用处理组和控制组平均因果效应的线性加权表示：

$$\rho = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[\mathbb{E}[Y_{1i} - Y_{0i} | D_i]] = P\{D_i = 1\}\rho_T + (1 - P\{D_i = 1\})\rho_C$$

我们可利用样本对三种平均因果效应进行估计，然而在实际观测中，处理组样本的潜在结果  $Y_{0i}$  和控制组样本的潜在结果  $Y_{1i}$  无法观测，因此处理组和控制组的平均因果效应无法观测，从而总体平均因果效应无法观测。此时，我们想到以  $Y_i$  为被解释变量， $D_i$  为解释变量，建立简单的虚拟变量回归模型来识别因果效应参数：

$$Y_i = \alpha + \rho_{ols}D_i + e_i$$

可得  $\hat{\rho}_{ols}$  的值为（第二章中将介绍）

$$\hat{\beta} = \frac{\sum_{i=1}^n (D_i - \bar{D})(Y_i - \bar{Y})}{\sum_{i=1}^n (D_i - \bar{D})^2}$$

处理组和控制组的样本数可进行如下表示：

$$n_T = \sum_{i=1}^n D_i, \quad n_C = \sum_{i=1}^n (1 - D_i)$$

则

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i = \frac{n_T}{n}$$

现在对  $\hat{\rho}_{ols}$  表达式的分子和分母进行重新表示，得

$$\sum_{i=1}^n (D_i - \bar{D})(Y_i - \bar{Y}) = \sum_{i=1}^n (D_i - \frac{n_T}{n})Y_i$$

$$\sum_{i=1}^n (D_i - \bar{D})^2 = \sum_{i=1}^n (D_i^2 - 2\frac{n_T D_i}{n} + \frac{n_T^2}{n^2}) = n_T - \frac{n_T^2}{n} = \frac{n_T n_C}{n}$$

则

$$\begin{aligned} \hat{\rho}_{ols} &= \frac{\sum_{i=1}^n (D_i - \frac{n_T}{n})Y_i}{\frac{n_T n_C}{n}} \\ &= \frac{\sum_{i=1}^n [(n_T + n_C)D_i - n_T]Y_i}{n_T n_C} \\ &= \frac{\sum_{i=1}^n [n_C D_i - n_T (1 - D_i)]Y_i}{n_T n_C} \\ &= \frac{1}{n_T} \sum_{i=1}^n D_i Y_i - \frac{1}{n_C} \sum_{i=1}^n (1 - D_i)Y_i \\ &= \bar{Y}_T - \bar{Y}_C \end{aligned}$$

故而在大样本条件下，

$$\rho_{ols} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$$

现在，我们得到了对平均因果效应的合理估计方法，上式可由处理组的平均因果效应表示为

$$\rho_{ols} = \mathbb{E}[Y_{1i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0] = \rho_T + \mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]$$

此时， $\rho_{ols}$  无法精准反映  $\rho_T$ ，当且仅当施加如下假设时， $\rho_{ols}$  才等于  $\rho_T$ ：

$$\mathbb{E}[Y_{0i} | D_i = 1] = \mathbb{E}[Y_{0i} | D_i = 0]$$

也可由控制组的平均因果效应表示为

$$\rho_{ols} = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] = \rho_C + \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0]$$

此时,  $\rho_{ols}$  无法精准反映  $\rho_C$ , 当且仅当施加如下假设时,  $\rho_{ols}$  才等于  $\rho_C$ :

$$\mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 0]$$

同时可由总体平均因果效应表示为

$$\begin{aligned}\rho_{ols} &= (P\{D_i = 1\} + 1 - P\{D_i = 1\})\mathbb{E}[Y_{1i}|D_i = 1] - (P\{D_i = 1\} + 1 - P\{D_i = 1\})\mathbb{E}[Y_{0i}|D_i = 0] \\ &= P\{D_i = 1\}(\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]) + (1 - P\{D_i = 1\})\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0] + \\ &\quad P\{D_i = 1\}(\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]) + (1 - P\{D_i = 1\})(\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0]) \\ &= \rho + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] + (1 - P\{D_i = 1\})(\rho_T - \rho_C)\end{aligned}$$

此时我们发现,  $\rho_{ols}$  无法精准反映  $\rho$ , 当且仅当施加如下假设时,  $\rho_{ols}$  才等于  $\rho$ :

$$\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0], \rho_T = \rho_C$$

这里, 我们将  $\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$  和  $\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0]$  称为**选择偏差**. 当同时施加选择偏差为 0 和两组因果效应无差异的假设时,  $\rho_{ols}$  才可精准反映总体平均因果效应.

**定理 1.4.1 (稳定个体处理值假设, SUTVA)** 稳定个体处理值假设 (简称稳定性假设) 有两层含义:

**假设 1:** 不同个体的潜在结果之间不会有交互影响;

**假设 2:** 处理水平对所有个体都是相同的.

由上面的推导, 我们知道如果需要进行因果推断, 必须消除选择偏误. 现已知, 实现定理 1.4.1 中的假设即可消除选择偏误, 而若要实现其中假设, 必须了解样本的分配机制. **随机分配**恰恰可以实现这样的假设: 随机的分配个体进入处理组或控制组, 使得无论处理组中的个体, 还是控制组中的个体, 都来自于同一个总体, 从而使得个体的潜在结果独立于  $D$ , 即满足于非混杂性:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i$$

进而均值独立成立:

$$\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0], \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 0]$$

现在验证  $\hat{\rho}_{ols}$  是对  $\rho$  的无偏估计.

#### 注

基于随机分配的因果分析就是**随机控制实验** (Randomized Control Trial, RCT), 简称随机实验.



对  $Y = Y_0 + (Y_1 - Y_0)D$  两边取条件期望, 得

$$\mathbb{E}[Y|D] = \mathbb{E}[Y_0|D] + \mathbb{E}[Y_1 - Y_0|D]D = \mathbb{E}[Y_0] + \rho D$$

设  $\mathbb{E}[Y_0] = \alpha$ , 则有

$$\mathbb{E}[Y|D] = \alpha + \rho D$$

从而

$$Y = \alpha + \rho D + e$$

则  $\hat{\rho}_{ols}$  是对  $\rho$  的无偏估计.

注意到

$$\mathbb{E}[e|D] = \mathbb{E}[Y - \alpha - \rho D|D] = 0$$

因此在随机分配中, 零条件均值假设成立.

如果在随机分配中, 某些变量的特征会使结果具有较大差异, 需要将这些变量作为**控制变量** (协变量) 引入回归模型中, 且需施加**条件独立假设**:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | W_i$$

建立的回归模型如下:

$$Y = \alpha + \rho D + W^T \beta + e$$

此时  $\hat{\rho}_{ols}$  不再是对  $\rho$  的无偏估计, 而是一致估计 (证明略). 注意到

$$\mathbb{E}[Y|D, W] = \alpha + \rho D + W^T \beta + \mathbb{E}[e|D, W], \quad \mathbb{E}[Y|D] = \alpha + \rho D + W^T \beta + \mathbb{E}[e|D]$$

则

$$\mathbb{E}[e|D, W] = \mathbb{E}[e|D]$$

上式称为**条件均值独立假设**.

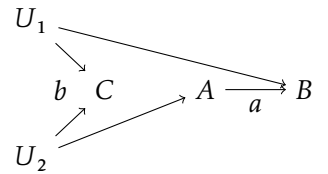
引入控制变量时, 常依据**后门准则**: 所有的后门路径都可被  $Z$  阻断, 当且仅当每条后门路径满足以下任一条件:

- (1) 包含一个中介路径  $A \rightarrow C \rightarrow B$ , 其中中介变量  $C \in Z$ ;
- (2) 包含一个共同原因的路径  $A \leftarrow C \rightarrow B$ , 其中混杂因子  $C \in Z$ ;
- (3) 包含一个共同结果的路径  $A \rightarrow C \leftarrow B$ , 其中对撞因子  $C \in Z$ .

如果同时存在混杂因子和对撞因子, 应选择混杂因子, 不选择对撞因子; 如果  $Z$  阻断了后门路径  $a$ , 但它作为对撞因子打开了后门路径  $b$ , 即产生**样本选择偏差** (如右图所示) 时, 就需要在路径  $b$  上寻找其它可以阻断路径的因子.

**注**

控制变量是进入回归只起辅助作用的变量, 关注的解释变量被称为**关键变量**.



## 2. 简单线性回归

2.1	线性投影和线性回归	14
2.1.1	线性投影模型	14
2.1.2	线性回归模型	16
2.2	回归系数的估计	17
2.2.1	回归系数的期望和方差	18
2.2.2	估计量的大样本抽样分布	20
2.3	拟合优度	22
2.4	误差方差的估计	23
2.5	置信区间和假设检验	26

### 2.1 线性投影和线性回归

#### 2.1.1 线性投影模型

由定理 1.3.6 可知, 条件期望函数  $m(X) = \mathbb{E}[Y|X]$  是对  $Y$  的最优预测函数. 在实际的预测问题中, 我们常常无法获知条件期望函数  $m(\cdot)$  的具体函数形式,  $m(\cdot)$  未必是线性函数.

对于一个复杂的模型, 可以使用**投影**将原始模型的一些特征或变量投影到较小的子空间中, 同时尽量保留原始模型的关键信息. 这种情况下  $Y$  和  $X$  之间的线性预测函数可以构造无数种, 而**线性投影模型**可以从所有线性预测函数中选择最优的预测函数, 构建对条件期望函数的近似.

**定义 2.1.1** 记  $\mathcal{P}[Y|X]$  是对  $\mathbb{E}[Y|X]$  的近似, 对于线性预测函数

$$\mathcal{P}[Y|X] = X^T \beta,$$

当  $\beta$  取  $\beta^* = \arg \min \mathbb{E}[(Y - X^T \beta)^2]$  时, 函数为最优线性预测函数, 此时  $\beta^*$  称为**线性投影系数**.

注意,  $\mathcal{P}[Y|X]$  是关于  $\beta$  的函数, 随  $\beta$  的变化而变化. 因为可构造的线性预测函数有无数种, 对应的  $\beta$  不同, 预测效果则不同.

**定理 2.1.1** 假设  $\mathbb{E}[Y^2]$  和  $\mathbb{E}[\|X\|^2]$  均存在, 且  $\mathbb{E}[XX^T]$  为正定矩阵, 则

$$\beta^* = (\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY].$$

证 设  $S(\beta) = \mathbb{E}[(Y - X^T\beta)^2]$ , 则

$$\begin{aligned} S(\beta) &= \mathbb{E}[(Y - X^T\beta)^T(Y - X^T\beta)] \\ &= \mathbb{E}[(Y^T - \beta^T X)(Y - X^T\beta)] \\ &= \mathbb{E}[Y^2] - 2\beta^T \mathbb{E}[XY] + \beta^T \mathbb{E}[XX^T]\beta \end{aligned}$$

令

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbb{E}[XY] + 2\mathbb{E}[XX^T]\beta = 0$$

得到

$$\beta^* = (\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY].$$

□

根据定理 2.1.1, 最优线性预测函数可以进一步表示成

$$\mathcal{P}[Y|X] = X^T(\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY].$$

与条件期望误差类似, 我们把

$$e = Y - \mathcal{P}[Y|X] = Y - X^T\beta^*$$

称为线性投影误差.

**定理 2.1.2 (线性投影误差的性质)** (1)  $\mathbb{E}[Xe] = 0$ ;

(2) 若  $X$  中存在常数,  $\mathbb{E}[e] = 0$ .

证 (1)  $\mathbb{E}[Xe] = \mathbb{E}[XY] - \mathbb{E}[XX^T]\beta^* = \mathbb{E}[XY] - \mathbb{E}[XX^T](\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY] = 0$ .

(2) 设  $X$  中存在常数  $X_i = m$ , 则有  $\mathbb{E}[me] = 0$ , 即  $\mathbb{E}[e] = 0$ . □

根据以上推导, 我们可以对线性投影模型作出如下定义:

**定义 2.1.2** 对于因变量  $Y$  和自变量  $X = (X_1, X_2, \dots, X_k)^T$ , 如果  $\mathbb{E}[XX^T]$  为正定矩阵, 则称方程组

$$\begin{cases} Y = X^T\beta + e \\ \mathbb{E}[Xe] = 0 \end{cases}$$

为线性投影模型. 其中, 线性投影系数  $\beta$  存在且唯一, 其数学表达式为:

$$\beta = (\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY].$$

 注

$\mathbb{E}[Xe] = 0$  说明  $X$  和  $e$  不相关, 但无法说明任意函数  $h(X)$  与  $e$  不相关 (即  $X$  和  $e$  相互独立). 这里,  $e$  会影响样本  $X$  的分布, 说明  $e$  是人为控制的变量, 不同的  $e$  会生成  $\beta$  的不同取值, 最终影响模型的预测效果.

现在, 考虑自变量中存在常数项的情况, 我们将常数项分离出  $X$ , 模型可调整为

$$\begin{cases} Y = X^T \beta + \alpha + e \\ \mathbb{E}[Xe] = 0 \\ \mathbb{E}[e] = 0 \end{cases}$$

对方程组第一个等式两边取期望, 得

$$\mathbb{E}[Y] = \mathbb{E}[X^T \beta] + \mathbb{E}[\alpha] + \mathbb{E}[e]$$

注

$\mathbb{E}[Xe] = 0$  和  $\beta = (\mathbb{E}[XX^T])^{-1} \mathbb{E}[XY]$  相互等价.

令  $\mu_Y = \mathbb{E}[Y]$ ,  $\mu_X = \mathbb{E}[X]$ , 则有

$$\alpha = \mu_Y - \mu_X^T \beta$$

代入第一个等式, 得

$$Y = X^T \beta + \mu_Y - \mu_X^T \beta + e$$

整理得

$$Y - \mu_Y = (X - \mu_X)^T \beta + e$$

此时

$$\beta = (\mathbb{E}[(X - \mu_X)(X - \mu_X)^T])^{-1} \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = (\text{var}(X))^{-1} \text{cov}(X, Y)$$

因此, 对于含常数项的线性投影模型, 我们可以进行如下数学表示:

$$\begin{cases} Y = X^T \beta + \alpha + e \\ \alpha = \mu_Y - \mu_X^T \beta \\ \beta = (\text{var}(X))^{-1} \text{cov}(X, Y) \end{cases}$$

## 2.1.2 线性回归模型

当条件期望函数  $m(X) = \mathbb{E}[Y|X]$  是线性函数,

$$m(X) = X^T \beta$$

利用条件期望分解, 可以得到

$$Y = X^T \beta + e.$$

由推论 1.3.5(1) 可知  $\mathbb{E}[e|X] = 0$ .

注

线性回归模型的基本假设:  
零条件均值假设:  $\mathbb{E}[e|X] = 0$ ;  
独立同分布假设:  $(X_i, Y_i)$  的观测满足独立同分布;  
同方差假设:  $\text{var}(e|X) = \sigma^2$ ;  
有限峰度假设:  $\mathbb{E}[X_i^4]$  和  $\mathbb{E}[Y_i^4]$  存在.

**定义 2.1.3** 对于因变量  $Y$  和自变量  $X = (X_1, X_2, \dots, X_k)^T$ , 称方程组

$$\begin{cases} Y = X^T \beta + e \\ \mathbb{E}[e|X] = 0 \end{cases}$$

## 为线性回归模型.

这里, 我们将线性回归模型和线性投影模型进行比较.

在线性回归模型中, 由推论 1.3.5(2) 可知,  $\text{cov}(e, h(X)) = 0$ ,  $h(X)$  为任意关于  $X$  的函数, 因此  $e, X$  相互独立. 由  $\text{cov}(e, X) = 0$  可以推得  $\mathbb{E}[Xe] = 0$ , 然而由  $\mathbb{E}[Xe] = 0$  无法推得  $\mathbb{E}[e|X] = 0$ . 这说明线性回归模型的条件较线性投影模型更强, 具有**严格外生性**. 在线性回归模型中,  $e$  是自然设定的变量, 不依赖于  $(X, Y)$  的分布, 该模型适用于  $Y$  和  $X$  因果问题的函数表达; 而在线性投影模型中,  $e$  是人为设定的变量, 该模型适用于  $Y$  和  $X$  相关问题的函数表达.

两个模型基本假设不同, 线性投影模型只要求解释变量和扰动项**正交**, 即  $\mathbb{E}[Xe] = 0$ ; 而线性回归模型在解释变量和扰动项正交的基础上, 还要求具备**严格外生性**, 即  $\mathbb{E}[e|X] = 0$ .

现在, 我们将  $X$  的维数设定为一维, 从最简单的线性回归模型开始讨论.

## 定义 2.1.4 定义

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

为简单线性回归模型, 其中  $Y_i$  为被解释变量,  $X_i$  为解释变量,  $\beta_0 + \beta_1 X_i$  为总体回归线,  $\beta_0$  和  $\beta_1$  分别为总体回归线的截距和斜率.

## 注

线性回归模型中,  $\beta$  中通常包含常数项, 即并非过原点回归. 接下来所有关于线性回归模型的讨论都基于  $\beta$  存在常数项的情况.

## 2.2 回归系数的估计

我们利用**普通最小二乘** (Ordinary Least Squares, 简称 OLS) 对简单线性回归模型中的  $\beta_0$  和  $\beta_1$  进行估计. 令  $\hat{\beta}_0$  和  $\hat{\beta}_1$  分别为  $\beta_0$  和  $\beta_1$  的估计量, 则相应的回归线为  $\hat{\beta}_0 + \hat{\beta}_1 X$ , 故第  $i$  个观测的预测误差为  $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ . 现在需令所有的误差平方和最小, 即

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

最小. 由于预测效果受为  $\hat{\beta}_0$  和  $\hat{\beta}_1$  影响, 上式为关于  $\hat{\beta}_0$  和  $\hat{\beta}_1$  的函数, 故对  $\hat{\beta}_0$  和  $\hat{\beta}_1$  偏导求极值.

$$\begin{cases} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases}$$

我们称上面的方程组为**正规方程组**. 现在对  $\hat{\beta}_0$  和  $\hat{\beta}_1$  进行求解, 将方程组进行整理可得

$$\begin{cases} n\hat{\beta}_0 + \sum_{i=1}^n X_i \hat{\beta}_1 = \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i \hat{\beta}_0 + \sum_{i=1}^n X_i^2 \hat{\beta}_1 = \sum_{i=1}^n X_i Y_i \end{cases}$$

解得

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})X_i} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_{XX}}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

我们称上式使误差平方和最小的截距和斜率为  $\beta_0$  和  $\beta_1$  的 **OLS** 估计量. 此时  $\hat{\beta}_0 + \hat{\beta}_1 X$  称为 **OLS 回归线** 或 **样本回归线**, 基于 OLS 回归线在给定  $X_i$  时  $Y_i$  的预测值为  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ,  $\hat{u}_i = Y_i - \hat{Y}_i$  称为第  $i$  个观测值的残差,  $\sum_{i=1}^n \hat{u}_i^2$  称为残差平方和 (SSR).

此时正规方程组可简化为

$$\begin{cases} \sum_{i=1}^n \hat{u}_i = 0 \\ \sum_{i=1}^n X_i \hat{u}_i = 0 \end{cases}$$

### 定理 2.2.1

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \bar{Y}$$

证 由  $Y_i = \hat{Y}_i + \hat{u}_i$  可得

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n \hat{u}_i$$

由正规方程组可得

$$\sum_{i=1}^n \hat{u}_i = 0$$

因此有

$$\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i = \sum_{i=1}^n \bar{Y}$$

即

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \bar{Y}$$

□

### 2.2.1 回归系数的期望和方差

得到估计量的表达式后, 我们需要对估计量的性质进行评估. 现验证  $\hat{\beta}_1$  的无偏性.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X}) X_i} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X}) (\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^n (X_i - \bar{X}) X_i} \\ &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X}) u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

对  $\hat{\beta}_1 - \beta_1$  取条件期望, 得

$$\mathbb{E}[\hat{\beta}_1 - \beta_1 | X_1, X_2, \dots, X_n] = \frac{\sum_{i=1}^n (X_i - \bar{X}) \mathbb{E}[u_i | X_1, X_2, \dots, X_n]}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0$$



根据迭代期望法则, 有

$$\mathbb{E}[\hat{\beta}_1 - \beta_1] = \mathbb{E}[\mathbb{E}[\hat{\beta}_1 - \beta_1 | X_1, X_2, \dots, X_n]] = 0$$

故

$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

即  $\hat{\beta}_1$  是对  $\beta_1$  的无偏估计. 接下来验证  $\hat{\beta}_0$  的无偏性.

$$\begin{aligned} \mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\bar{Y} - \hat{\beta}_1 \bar{X}] \\ &= \mathbb{E}[\beta_0 + \beta_1 \bar{X} + \bar{u} - \hat{\beta}_1 \bar{X}] \\ &= \beta_0 + \mathbb{E}[(\beta_1 - \hat{\beta}_1) \bar{X}] + \mathbb{E}[\bar{u}] \\ &= \beta_0 \end{aligned}$$

即  $\hat{\beta}_0$  是对  $\beta_0$  的无偏估计.

除了知道  $\hat{\beta}_0$  和  $\hat{\beta}_1$  的估计是无偏以外, 我们关心  $\hat{\beta}_0$  和  $\hat{\beta}_1$  的分散程度, 而方差是衡量分散程度的常用指标. 在同方差假设下,

$$\mathbb{E}[u^2] = \mathbb{E}[\mathbb{E}[u^2 | X]] = \mathbb{E}[u^2 - (\mathbb{E}[u])^2 | X] = \mathbb{E}[\text{var}(u | X)] = \sigma^2$$

$$\text{var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2 = \sigma^2$$

由此可见  $\sigma^2$  是  $u$  的无条件方差.

现在求  $\hat{\beta}_1$  的方差.

$$\mathbb{E}[(\hat{\beta}_1 - \beta_1)^2 | X_1, X_2, \dots, X_n] = \mathbb{E}\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2 \mathbb{E}[u_i^2 | X_1, X_2, \dots, X_n]}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2}\right] = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \mathbb{E}[(\hat{\beta}_1 - \beta_1)^2] - (\mathbb{E}[\hat{\beta}_1 - \beta_1])^2 \\ &= \mathbb{E}[\mathbb{E}[(\hat{\beta}_1 - \beta_1)^2 | X_1, X_2, \dots, X_n]] \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sigma^2}{n\sigma_x^2} \end{aligned}$$

接下来求  $\hat{\beta}_0$  的方差.

$$\text{var}(\hat{\beta}_0) = \text{var}(\beta_0 - (\hat{\beta}_1 - \beta_1)\bar{X} + \bar{u}) = \text{var}(-\hat{\beta}_1 \bar{X} + \bar{u})$$

易证

$$\text{cov}(\hat{\beta}_1, \bar{u}) = 0$$

 注

注意  $\mathbb{E}[u_i | X_1, X_2, \dots, X_n] = 0$  的证明:

由零条件均值假设可知,  
 $\mathbb{E}[u_i | X_i] = 0$ ;

由独立同分布假设可知,  
 $\mathbb{E}[u_i | X_j] = 0, i \neq j$ .

因此

$$\text{var}(\hat{\beta}_0) = \text{var}(\hat{\beta}_1 \bar{X}) + \text{var}(\bar{u}) = \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2$$

由于

$$\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2 + n\bar{X}^2}{n \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

$\hat{\beta}_0$  的方差还可表示成

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

这里，我们对  $\hat{\beta}_1$  和  $\hat{\beta}_0$  的期望和方差进行总结：

$$\mathbb{E}[\hat{\beta}_1] = \beta_1, \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \text{var}(\hat{\beta}_0) = \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2$$

## 2.2.2 估计量的大样本抽样分布

**引理 2.2.1 (斯勒茨基定理, Slutsky's Theorem)** 若  $a_n \xrightarrow{d} a$ ,  $b_n \xrightarrow{p} b$ , 其中  $a$  为随机变量,  $b$  为常数, 则有

- (1)  $a_n + b_n \xrightarrow{d} a + b$ ;
- (2)  $a_n b_n \xrightarrow{d} ab$ ;
- (3)  $\frac{a_n}{b_n} \xrightarrow{d} \frac{a}{b}$ , 满足  $P\{b = 0\} = 0$ .

考察

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X}) u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

令  $v_i = (X_i - \bar{X}) u_i$ ,

$$v_i \xrightarrow{p} \bar{v} = (X_i - \mu_X) u_i$$

$$\mathbb{E}[\bar{v}] = 0$$

又因为  $v_i$  独立同分布,  $\sigma_v^2 = \text{var}[(X_i - \mu_X) u_i]$  非零有限,

$$\sum_{i=1}^n v_i \xrightarrow{d} N(0, n \text{var}[(X_i - \mu_X) u_i])$$

由于

$$\sum_{i=1}^n (X_i - \bar{X})^2 \xrightarrow{p} n \text{var}(X_i)$$

因此

$$\hat{\beta}_1 - \beta_1 \xrightarrow{d} N\left(0, \frac{\text{var}[(X_i - \mu_X)u_i]}{n(\text{var}(X_i))^2}\right)$$

即

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \frac{\text{var}[(X_i - \mu_X)u_i]}{n(\text{var}(X_i))^2}\right)$$

在同方差假设下,  $\text{var}(u_i|X_1, X_2, \dots, X_n) = \sigma^2$ , 则有

$$\begin{aligned}\text{var}[(X_i - \mu_X)u_i|X_1, X_2, \dots, X_n] &= \mathbb{E}[(X_i - \mu_X)^2 u_i^2 - (\mathbb{E}[(X_i - \mu_X)u_i])^2|X_1, X_2, \dots, X_n] \\ &= \mathbb{E}[(X_i - \mu_X)^2 u_i^2|X_1, X_2, \dots, X_n] \\ &= \sigma^2(\text{var}(X_i))\end{aligned}$$

因此有

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \frac{\sigma^2}{n\text{var}(X)}\right)$$

**引理 2.2.2 (连续映射定理, The Continuous Mapping Theorem)** 若  $x_n$  是一个  $k$  维随机变量序列, 向量值函数  $f: \mathbb{R}^k \mapsto \mathbb{R}^m$  在  $x_0$  处连续,

- (1) 当  $x_n \xrightarrow{p} x_0$  时,  $f(x_n) \xrightarrow{p} f(x_0)$ ;
- (2) 当  $x_n \xrightarrow{d} x_0$  时,  $f(x_n) \xrightarrow{d} f(x_0)$ .

现在, 我们来求  $\hat{\beta}_0$  的大样本方差.

对  $\text{var}(\hat{\beta}_0)$  的表达式进行变换, 有

$$\begin{aligned}\text{var}(\hat{\beta}_0) &= \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \\ &= \frac{\frac{1}{n} \sigma^2 \sum_{i=1}^n X_i^2}{n \left[ \frac{1}{n} \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n \frac{1}{n} X_i)^2 \right]}\end{aligned}$$

在大样本条件下,

$$\begin{aligned}\frac{1}{n} X_i &\xrightarrow{p} \mu_X \\ \frac{1}{n} \sum_{i=1}^n X_i^2 &\xrightarrow{p} \mathbb{E}[X_i^2]\end{aligned}$$

因此有

$$\text{var}(\hat{\beta}_0) \xrightarrow{p} \frac{\sigma^2 \mathbb{E}[X_i^2]}{n(\mathbb{E}[X_i^2] - \mu_X^2)}$$

而

$$\frac{\sigma^2 \mathbb{E}[X_i^2]}{n(\mathbb{E}[X_i^2] - \mu_X^2)} = \frac{\sigma^2}{n \left(1 - \frac{\mu_X^2}{\mathbb{E}[X_i^2]}\right)}$$

设

$$H_i = 1 - \frac{\mu_X}{\mathbb{E}[X_i^2]} X_i$$

有

$$H_i^2 = 1 - 2 \frac{\mu_X}{\mathbb{E}[X_i^2]} X_i + \frac{\mu_X^2}{(\mathbb{E}[X_i^2])^2} X_i^2$$

两边同时取期望，得

$$\mathbb{E}[H_i^2] = 1 - \frac{\mu_X^2}{\mathbb{E}[X_i^2]}$$

易知  $H_i$  和  $u_i$  相互独立，计算  $\text{var}(H_i u_i)$ ，得

$$\text{var}(H_i u_i) = \mathbb{E}[H_i^2] \mathbb{E}[u_i^2] = \mathbb{E}[H_i^2] \sigma^2$$

因此

$$\frac{\sigma^2}{n \left(1 - \frac{\mu_X^2}{\mathbb{E}[X_i^2]}\right)} = \frac{\text{var}(H_i u_i)}{n(\mathbb{E}[H_i^2])^2}$$

即

$$\text{var}(\hat{\beta}_0) \xrightarrow{p} \frac{\text{var}(H_i u_i)}{n(\mathbb{E}[H_i^2])^2}$$

因此有

$$\hat{\beta}_0 \xrightarrow{d} N \left( \beta_0, \frac{\text{var}(H_i u_i)}{n(\mathbb{E}[H_i^2])^2} \right)$$

这里，我们对  $\hat{\beta}_1$  和  $\hat{\beta}_0$  的大样本抽样分布进行总结：

$$\hat{\beta}_1 \xrightarrow{d} N \left( \beta_1, \frac{\text{var}[(X_i - \mu_X)u_i]}{n(\text{var}(X_i))^2} \right)$$

$$\hat{\beta}_0 \xrightarrow{d} N \left( \beta_0, \frac{\text{var}(H_i u_i)}{n(\mathbb{E}[H_i^2])^2} \right), \text{ 其中 } H_i = 1 - \frac{\mu_X}{\mathbb{E}[X_i^2]} X_i.$$

## 2.3 拟合优度

**定义 2.3.1** 定义  $Y_i$  与其均值偏差的平方和为总平方和 (TSS),  $\hat{Y}_i$  与  $Y_i$  均值偏差的平方和为解释平方和 (ESS), 回归  $R^2$  为解释平方和与总平方和之比, 即

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$R^2 = \frac{ESS}{TSS}$$

$R^2$  表示可由  $X_i$  解释的  $Y_i$  的方差的比例,  $R^2$  越大, 拟合优度更良好.

#### 定理 2.3.1

$$TSS = ESS + SSR$$

$$R^2 = 1 - \frac{SSR}{TSS}$$

证 由定理 2.1.3 和正规方程组可得

$$\begin{aligned} TSS &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= SSR + 2 \sum_{i=1}^n \hat{u}_i (\hat{Y}_i - \bar{Y}) + ESS \\ &= SSR + 2 \sum_{i=1}^n \hat{u}_i \hat{Y}_i + ESS \\ &= SSR + 2 \sum_{i=1}^n \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1) + ESS \\ &= SSR + 2\hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + 2\hat{\beta}_1 \sum_{i=1}^n X_i \hat{u}_i + ESS \\ &= ESS + SSR \end{aligned}$$

由此可得

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}.$$

□

该结论在多元线性回归模型中同样成立, 我们将在 2.2.2 中进一步证明.

## 2.4 误差方差的估计

为估计误差方差, 需要从残差这一估计量下手.

现考察残差平方和

$$SSR = \sum_{i=1}^n \hat{u}_i^2$$

令  $s_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$ , 则有

$$\begin{aligned} \sum_{i=1}^n \hat{u}_i^2 &= \sum_{i=1}^n [Y_i - (\bar{Y} - \hat{\beta}_1) - \hat{\beta}_1 X_i]^2 \\ &= \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})]^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{s_{XX}} \\ &= \sum_{i=1}^n Y_i^2 - \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \right)^2 - \left( \sum_{i=1}^n \frac{X_i - \bar{X}}{\sqrt{s_{XX}}} Y_i \right)^2 \end{aligned}$$

令  $Y = (Y_1, Y_2, \dots, Y_n)^T$ , 作正交变换  $Z = AY$ , 则有

$$Y^T Y = \sum_{i=1}^n Y_i^2 = \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \right)^2 + \left( \sum_{i=1}^n \frac{X_i - \bar{X}}{\sqrt{s_{XX}}} Y_i \right)^2 + \sum_{i=1}^n u_i^2 = \sum_{i=1}^n Z_i^2 = Z^T Z$$

其中正交矩阵  $A$  满足

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{X_1 - \bar{X}}{\sqrt{s_{XX}}} & \frac{X_2 - \bar{X}}{\sqrt{s_{XX}}} & \cdots & \frac{X_n - \bar{X}}{\sqrt{s_{XX}}} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

且有

$$Z_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i, Z_2 = \sum_{i=1}^n \frac{X_i - \bar{X}}{\sqrt{s_{XX}}} Y_i, Z_k = \sum_{i=1}^n a_{ki} Y_i, k = 3, 4, 5, \dots$$

则

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{k=3}^n Z_k^2$$

由于  $A$  是正交矩阵, 满足  $A^T A = I$ , 所以

$$\sum_{i=1}^n a_{ki} a_{ji} = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}, \sum_{i=1}^n a_{ki}^2 = 1$$



当  $k \geq 3$  时,

取  $j = 1$ , 有

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n a_{ki} = 0, \text{ 即 } \sum_{i=1}^n a_{ki} = 0$$

取  $j = 2$ , 有

$$\frac{X_i - \bar{X}}{\sqrt{s_{XX}}} \sum_{i=1}^n a_{ki} = 0, \text{ 即 } \sum_{i=1}^n X_i a_{ki} = \bar{X} \sum_{i=1}^n a_{ki} = 0$$

则

$$\mathbb{E}[Z_k] = \mathbb{E}\left[\sum_{i=1}^n a_{ki} Y_i\right] = \sum_{i=1}^n a_{ki} \mathbb{E}[\mathbb{E}[Y_i|X_i]] = \sum_{i=1}^n a_{ki} \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 X_i] = \beta_0 \sum_{i=1}^n a_{ki} + \beta_1 \mathbb{E}\left[\sum_{i=1}^n a_{ki} X_i\right] = 0$$

$$\text{var}(Z_k) = \text{var}\left(\sum_{i=1}^n a_{ki} Y_i\right) = \sum_{i=1}^n a_{ki}^2 \text{var}(Y_i) = \sigma^2$$

所以

$$Z_k \sim N(0, \sigma^2)$$

则

$$\frac{\sum_{i=1}^n \hat{u}_i^2}{\sigma^2} \sim \chi^2(n-2)$$

根据  $\chi^2$  分布的性质, 有

$$\mathbb{E}\left[\frac{\sum_{i=1}^n \hat{u}_i^2}{\sigma^2}\right] = n-2$$

整理可得

$$\mathbb{E}\left[\frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}\right] = \sigma^2$$

则

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$$

其中,  $\hat{\sigma}^2$  是  $\sigma^2$  的无偏估计,  $\hat{\sigma}$  称为回归标准误 (SER), 记作

$$SER = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}$$

类似地, 我们在估计  $\hat{\beta}_1$  的方差时, 使用样本方差

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{n \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{(n-2)(\sum_{i=1}^n (X_i - \bar{X})^2)}$$

 注

此处证明使用了正态假设, 目的是利用卡方分布的性质求期望, 在 3.3.2 中我们将介绍无正态假设条件下进行误差方差估计的方法.

来近似,  $\hat{\sigma}_{\hat{\beta}_1}$  称为  $\hat{\beta}_1$  的异方差稳健标准误, 记作

$$SE(\hat{\beta}_1) = \sqrt{\frac{n \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{(n-2)(\sum_{i=1}^n (X_i - \bar{X})^2)^2}}$$

## 2.5 置信区间和假设检验

在 2.1.3 中, 我们证得, 在大样本条件下,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\text{var}[(X_i - \mu_X)u_i]}{n(\text{var}(X_i))^2}\right)$$

现在用  $SE(\hat{\beta}_1)$  来近似  $\hat{\beta}_1$  的方差, 可构造  $t$  统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

**注**

一般来说,  $t$  统计量具有如下形式:

$$t = \frac{\text{估计量} - \text{假设值}}{\text{估计量的标准误}}$$

考虑到大样本情况,  $t$  分布近似为标准正态分布, 即

$$t \sim N(0, 1)$$

我们可以据此对  $\beta_1$  构造置信区间和进行假设检验.

置信区间:

$$(\hat{\beta}_1 - z_{\frac{\alpha}{2}} SE(\hat{\beta}_1), \hat{\beta}_1 + z_{\frac{\alpha}{2}} SE(\hat{\beta}_1))$$

双边检验:

$$H_0: \beta_1 = \beta_{1,0}, H_1: \beta_1 \neq \beta_{1,0}$$

拒绝域为

$$|t| \geq z_{\frac{\alpha}{2}}$$

单边检验 (左侧):

$$H_0: \beta_1 = \beta_{1,0}, H_1: \beta_1 < \beta_{1,0}$$

拒绝域为

$$t \leq z_{\alpha}$$

# 3. 多元线性回归

3.1	回归系数的估计	27
3.1.1	回归系数的期望和方差	28
3.1.2	回归系数的大样本抽样分布	29
3.2	分块回归和偏回归	29
3.2.1	解释变量个数不同时偏回归系数之间的关系	32
3.2.2	偏回归系数的期望和方差	32
3.2.3	偏回归系数的大样本抽样分布	33
3.3	拟合优度和误差方差的估计	33
3.4	正态回归	34

在 2.2.1 中，我们证得

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

因此有

$$\hat{\beta}_1 \xrightarrow{d} \beta_1 + \frac{\text{cov}(X, u)}{\text{var}(X)} = \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_x}$$

如果模型具有内生性，即  $\text{cov}(X, u) \neq 0$ ， $\hat{\beta}_1$  不是  $\beta_1$  的无偏和一致估计，此时简单线性回归出现了遗漏变量偏差。为了解决这个问题，需要采用多元线性回归。

## 3.1 回归系数的估计

定义 3.1.1 定义

$$Y_i = X_i' \beta + u_i$$

为多元线性回归模型。

注意到样本多元回归中，

$$X_i = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}$$

而总体多元回归中，

$$X_{n \times (k+1)} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ 1 & X_{31} & \dots & X_{3k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}$$

其中,  $k$  表示解释变量的个数,  $n$  表示样本数.

类似地, 我们也可得到多元线性回归的正规方程组. 目标函数为

$$\min \sum_{i=1}^n (Y_i - X_i' \hat{\beta})^2$$

对  $\hat{\beta}$  进行求导, 得

$$\nabla = -2 \sum_{i=1}^n X_i' \hat{u}_i = 0$$

即

$$\sum_{i=1}^n X_i' \hat{u}_i = 0$$

将正规方程组写成标量形式, 可得

$$\begin{cases} \sum_{i=1}^n \hat{u}_i = 0 \\ \sum_{i=1}^n X_{i1} \hat{u}_i = 0 \\ \sum_{i=1}^n X_{i2} \hat{u}_i = 0 \\ \dots \\ \sum_{i=1}^n X_{ik} \hat{u}_i = 0 \end{cases}$$

现对正规方程组进行求解:

$$X'(Y - X\hat{\beta}) = 0$$

即

$$\hat{\beta} = (X'X)^{-1}X'Y$$

### 3.1.1 回归系数的期望和方差

现验证  $\hat{\beta}$  的无偏性和一致性.

$$\mathbb{E}[\hat{\beta}|X] = \mathbb{E}[(X'X)^{-1}X'Y|X] = (X'X)^{-1}X'\mathbb{E}[Y|X] = (X'X)^{-1}X'X\beta = \beta$$

由定理 2.1.1 可知,  $\beta = (\mathbb{E}[X'X])^{-1}\mathbb{E}[X'Y]$ , 根据连续映射定理, 可得

$$\hat{\beta} = (X'X)^{-1}X'Y \xrightarrow{p} (\mathbb{E}[X_iX_i'])^{-1}\mathbb{E}[X_iY] = \beta$$

接下来求  $\hat{\beta}$  的条件方差. 设误差方差矩阵为  $D(d_{ii} = \sigma_i^2, d_{ij} = 0, i \neq j)$ , 则

$$\text{var}(\hat{\beta}|X) = \text{var}((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'DX(X'X)^{-1}$$

在同方差条件下, 由于  $D = I_n \sigma^2$ , 则

$$\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

### 3.1.2 回归系数的大样本抽样分布

注意到

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u = \beta + \left( \sum_{i=1}^n X_i X_i' \right)^{-1} \left( \sum_{i=1}^n X_i u_i \right)$$

由于

$$\mathbb{E} \left[ \sum_{i=1}^n X_i u_i | X_i \right] = 0$$

$$\text{var}(X_i u_i) = \mathbb{E}[X_i' X_i u_i^2]$$

则

$$\sum_{i=1}^n X_i u_i \xrightarrow{d} N(0, n \mathbb{E}[X_i' X_i u_i^2])$$

因为

$$\left( \sum_{i=1}^n X_i X_i' \right)^{-1} \xrightarrow{p} \mathbb{E}[n(X_i X_i')]^{-1} = \frac{1}{n} \mathbb{E}[(X_i X_i')]^{-1}$$

则

$$(X'X)^{-1}X'u \xrightarrow{d} N \left( 0, \frac{1}{n} \mathbb{E}[(X_i X_i')]^{-1} \mathbb{E}[X_i' X_i u_i^2] \mathbb{E}[(X_i X_i')]^{-1} \right)$$

即可得到  $\hat{\beta}$  的大样本抽样分布为

$$\hat{\beta} \xrightarrow{d} N \left( \beta, \frac{1}{n} \mathbb{E}[(X_i X_i')]^{-1} \mathbb{E}[X_i' X_i u_i^2] \mathbb{E}[(X_i X_i')]^{-1} \right)$$

## 3.2 分块回归和偏回归

**定义 3.2.1** 定义矩阵

$$P = X(X'X)^{-1}X'$$

为投影矩阵.

注意到  $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$ , 用  $P$  左乘任何向量可得该向量在超平面  $X$  上的投影.

**定理 3.2.1** (投影矩阵的性质) (1)  $P' = P$ ;

(2)  $P^2 = P$ ;

(3)  $PX = X$ ;

(4)  $r(P) = \text{tr}(P) = k + 1$ .

证

$$P' = X((X'X)')^{-1}X' = X'(X'X)^{-1}X = P$$

$$P^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P$$

$$PX = X(X'X)^{-1}X'X = X$$

由于  $P$  是实对称阵, 可进行相似对角化, 即

$$P = Q'I_{k+1}Q$$

则有  $r(P) = k + 1$ . 由幂等矩阵的性质可知,  $\text{tr}(P) = r(P) = k + 1$ . □

**定义 3.2.2** 定义矩阵

$$M = I_n - P$$

为消灭矩阵.

注意到  $\hat{u} = Y - \hat{Y} = Y - PY = MY$ , 用  $M$  左乘任何向量可得该向量投影后的残差向量.

**定理 3.2.2** (消灭矩阵的性质) (1)  $M' = M$ ;

(2)  $M^2 = M$ ;

(3)  $MX = 0$ ;

(4)  $PM = 0$ ;

(5)  $r(M) = \text{tr}(M) = n - k - 1$ .

证

$$M' = I_n' - P' = I_n - P$$

$$M^2 = (I_n - P)(I_n - P) = I_n - 2P + P^2 = I_n - P = M$$

$$MX = X - PX = X - X = 0$$

$$PM = P(I_n - P) = P - P^2 = 0$$

由  $PM = 0$ , 可得  $r(P) + r(M) \leq n$ ; 又由  $P + M = I_n$ , 可得  $n = r(I_n) \leq r(P) + r(M)$ , 所以

$$r(P) + r(M) = n$$



则有  $r(M) = n - k - 1$ . 由幂等矩阵的性质可知,  $\text{tr}(M) = r(M) = n - k - 1$ .

□

利用消灭矩阵的性质, 我们可以将残差写成总体误差项的函数

$$\hat{u} = MY = M(X\beta + u) = Mu.$$

进而将残差平方和也写成总体误差项的函数

$$SSR = \hat{u}'\hat{u} = u'MMu = u'Mu.$$

在多元线性回归中, 只要增加一个变量就会对所有的回归系数产生影响, 然而仅从  $\hat{\beta} = (X'X)^{-1}X'Y$  这一表达式中很难看出不同变量的影响. 为此, 我们将  $X$  进行分块, 即  $X = (X_1 \ X_2)$ , 分别对应于两组解释变量, 此时多元线性回归模型可以改写为

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

$\hat{\beta} = (X'X)^{-1}X'Y$  可改写为

$$\begin{aligned} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \left( \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} (X_1 \ X_2) \right)^{-1} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} Y \\ &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} Y \\ &= \begin{pmatrix} (X_1'M_2X_1)^{-1}X_1'M_2 \\ (X_2'M_1X_2)^{-1}X_2'M_1 \end{pmatrix} Y \end{aligned}$$

令  $r_1 = M_1X_2$ ,  $\tilde{u}_1 = M_1Y$ , 则

$$\hat{\beta}_2 = (\hat{r}_1' \hat{r}_1)^{-1} \hat{r}_1' \tilde{u}_1$$

这里可以看出,  $\hat{r}_1$  是  $X_2$  对  $X_1$  回归得到的残差矩阵,  $\tilde{u}_1$  是  $Y$  对  $X_1$  回归得到的残差矩阵, 而  $\hat{\beta}_2$  恰恰是  $\tilde{u}_1$  对  $\hat{r}_1$  进行回归所得的回归系数. 这就是 **Frisch-Waugh-Lovell 定理** 的内容.

根据这一定理, 我们就可对偏回归系数  $\beta_j$  进行估计. 设  $X_2$  是解释变量  $X_j$  的样本组成的一维行向量. 首先由  $X_j$  对其他所有解释变量进行回归, 得到残差列向量  $r_j$ ; 再由  $Y$  对其他所有解释变量进行回归, 得到残差列向量  $\tilde{u}_j$ , 则

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij} \tilde{u}_{ij}}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

由正规方程组易知

$$\sum_{i=1}^n \hat{r}_{ij} \tilde{u}_{ij} = \sum_{i=1}^n \hat{r}_{ij} (Y_i - X_i^* \tilde{\beta}) \sum_{i=1}^n \hat{r}_{ij} Y_i$$

则有

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij} Y_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

$\beta_j$  是剔除了其他所有解释变量后的残差  $\hat{r}_j$  与  $Y$  的样本协方差和  $\hat{r}_j$  方差之比.

注

$X^*$  为除  $X_j$  以外的其他解释变量组成的行向量,  $\hat{\beta}$  为  $Y$  对除  $X_j$  以外的其他解释变量的回归系数列向量, 显然  $\hat{r}_{ij}$  与  $X^*$  无关.

### 3.2.1 解释变量个数不同时偏回归系数之间的关系

现在, 我们用上述结论探究解释变量个数不同时偏回归系数之间的关系.

先考察简单线性回归  $Y = \gamma_0 + \gamma_1 X_1 + e$  和多元线性回归  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$ .

$$\begin{aligned} \hat{\gamma}_1 &= \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1) Y_i}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} \\ &= \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1) (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik} + \hat{u}_i)}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2} \\ &= \hat{\beta}_1 + \hat{\beta}_2 \gamma_{2,1} + \cdots + \hat{\beta}_k \gamma_{k,1} \end{aligned}$$

$\gamma_{j,1}$  表示  $X_j$  对  $X_1$  的回归系数.

再将条件放宽, 考察短回归  $Y = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + e$  和长回归  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_k X_k + u$ .

$$\begin{aligned} \hat{\gamma}_1 &= \frac{\sum_{i=1}^n \hat{r}_i Y_i}{\sum_{i=1}^n \hat{r}_i^2} \\ &= \frac{\sum_{i=1}^n \hat{r}_i (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik} + \hat{u}_i)}{\sum_{i=1}^n \hat{r}_i^2} \\ &= \hat{\beta}_1 + \hat{\beta}_{p+1} \frac{\sum_{i=1}^n \hat{r}_i X_{ip+1}}{\sum_{i=1}^n \hat{r}_i^2} + \cdots + \hat{\beta}_k \frac{\sum_{i=1}^n \hat{r}_i X_{ik}}{\sum_{i=1}^n \hat{r}_i^2} \\ &= \hat{\beta}_1 + \hat{\beta}_{p+1} \hat{\delta}_{p+1,1} + \cdots + \hat{\beta}_k \hat{\delta}_{k,1} \end{aligned}$$

$\hat{\delta}_{j,1} (p+1 \leq j \leq k)$  表示  $X_j$  对  $X_1$  的回归系数.

### 3.2.2 偏回归系数的期望和方差

现验证  $\hat{\beta}_j$  的无偏性.

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{ij} Y_i}{\sum_{i=1}^n \hat{r}_{ij}^2} = \beta_j + \frac{\sum_{i=1}^n \hat{r}_{ij} u_i}{\sum_{i=1}^n \hat{r}_{ij}^2}$$

对  $\hat{\beta}_j - \beta_j$  取条件期望, 得

$$\mathbb{E}[\hat{\beta}_j - \beta_j | X_{1j}, X_{2j}, \cdots, X_{nj}] = \frac{\sum_{i=1}^n \hat{r}_{ij} \mathbb{E}[u_i | X_{1j}, X_{2j}, \cdots, X_{nj}]}{\sum_{i=1}^n \hat{r}_{ij}^2} = 0$$

则

$$\mathbb{E}[\hat{\beta}_j] = \beta_j$$

即  $\hat{\beta}_j$  是  $\beta_j$  的无偏估计量.

在异方差条件下,

$$\text{var}(\hat{\beta}_j | X_{1j}, X_{2j}, \dots, X_{nj}) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \mathbb{E}[u_i^2 | X_{1j}, X_{2j}, \dots, X_{nj}]}{\sum_{i=1}^n \hat{r}_{ij}^2} = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \sigma_i^2}{(\sum_{i=1}^n \hat{r}_{ij}^2)^2}$$

在同方差条件下,

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{ij}^2} = \frac{\sigma^2}{TSS_j(1 - R_j^2)}$$

其中,  $TSS_j = \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$  是  $X_j$  的总样本变异,  $R_j^2$  则是将  $X_j$  对所有其他自变量 (包括截距项) 进行回归得到的  $R^2$ .

### 3.2.3 偏回归系数的大样本抽样分布

易知

$$\sum_{i=1}^n \hat{r}_{ij} u_i \xrightarrow{d} N(0, n \text{var}(\hat{r}_{ij} u_i))$$

设  $\frac{1}{n} \sum_{i=1}^n \hat{r}_{ij}^2 \xrightarrow{p} a_j^2$ , 则

$$\hat{\beta}_j \xrightarrow{d} N\left(\beta_j, \frac{\text{var}(\hat{r}_{ij} u_i)}{n(a_j^2)^2}\right)$$

在同方差假设下,

$$\hat{\beta}_j \xrightarrow{d} N\left(\beta_j, \frac{\sigma^2}{na_j^2}\right)$$

## 3.3 拟合优度和误差方差的估计

### 拟合优度

在简单线性回归中, 我们在一元条件下证明了定理 2.3.1, 现在我们将该定理推广到多元情况.

$$\begin{aligned} TSS &= Y'Y \\ &= (\hat{Y} + \hat{u})'(\hat{Y} + \hat{u}) \\ &= \hat{Y}'\hat{Y} + 2\hat{Y}'\hat{u} + \hat{u}'\hat{u} \end{aligned}$$

由于

$$\hat{Y}'\hat{u} = Y'PMY = 0$$

则有

$$TSS = \hat{Y}'\hat{Y} + \hat{u}'\hat{u} = SSR + ESS.$$

## 误差方差的估计

考察残差平方和

$$SSR = u'Mu = \text{tr}(u'Mu) = \text{tr}(Mu u')$$

则有

$$\mathbb{E}[SSR|X] = \text{tr}(\mathbb{E}[Mu u'|X]) = \text{tr}(M\mathbb{E}[u u'|X]) = \text{tr}(MD)$$

在同方差假设下,  $D = I_n \sigma^2$ , 则

$$\mathbb{E}[SSR] = \sigma^2(n - k - 1)$$

整理可得

$$\mathbb{E}\left[\frac{\hat{u}'\hat{u}}{n - k - 1}\right] = \sigma^2$$

则

$$\hat{\sigma}^2 = \frac{1}{n - k - 1} \sum_{i=1}^n \hat{u}_i^2$$

$$SER = \sqrt{\frac{1}{n - k - 1} \sum_{i=1}^n \hat{u}_i^2} = \sqrt{\frac{SSR}{n - k - 1}}$$

## 3.4 正态回归

**定义 3.4.1** 当  $u_i|X_i \sim (0, \sigma^2)$  时, 定义

$$Y_i = X_i' \beta + \mu_i$$

为正态回归模型.

该模型在零条件均值、独立同分布、有限峰度、无完全共线假设的基础上, 加入了正态假设, 即  $\mathbb{E}[u|X] = 0$  的假设加强为  $u|X \sim N(0, \sigma^2 I_n)$ . 由此, 我们可以得到  $\hat{\beta}$  和  $\hat{u}$  的条件分布:

因为  $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'u$ , 则有  $\text{var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}X'((X'X)^{-1}X')' = \sigma^2(X'X)^{-1}$ , 那么

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$$

类似地, 由于  $\hat{u} = Mu$ , 则

$$\hat{u}|X \sim N(0, \sigma^2 M)$$

**定理 3.4.1** 在正态回归模型中,  $\hat{\beta}$  和  $\hat{u}$  相互独立.

证 考察  $\hat{\beta}$  和  $\hat{u}$  的联合正态分布.

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{u} \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X' \\ M \end{pmatrix} u$$

$$\text{var} \left( \begin{pmatrix} \hat{\beta} - \beta \\ \hat{u} \end{pmatrix} \middle| X \right) = \sigma^2 \begin{pmatrix} (X'X)^{-1}X' \\ M \end{pmatrix} (X(X'X)^{-1} \quad M) = \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \sigma^2 M \end{pmatrix}$$

由于  $\text{cov}(\hat{\beta} - \beta, \hat{u}) = 0$ , 则  $\text{cov}(\hat{\beta}, \hat{u}) = 0$ , 在正态条件下  $\hat{\beta}$  和  $\hat{u}$  相互独立.  $\square$

#### 定理 3.4.2

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k-1)$$

证

$$\begin{aligned} \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} &= \frac{\hat{u}'\hat{u}}{\sigma^2} \\ &= \left(\frac{u}{\sigma}\right)' M \left(\frac{u}{\sigma}\right) \end{aligned}$$

由定理 3.2.2(5) 可知  $r(M) = n - k - 1$ , 则

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k-1)$$

$\square$

### 置信区间

#### 1. $\beta_j$ 的置信区间

构造  $t$  统计量

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_{jj}}} \sim t(n-k-1)$$

则  $\beta_j$  的置信水平为  $\alpha$  的置信区间为

$$\left( \hat{\beta}_j \pm t_{\frac{\alpha}{2}}(n-k-1) \hat{\sigma} \sqrt{[(X'X)^{-1}]_{jj}} \right)$$

#### 2. $\sigma^2$ 的置信区间

构造  $\chi^2$  统计量

$$\chi^2 = \frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2(n-k-1)$$

则  $\sigma^2$  的置信水平为  $\alpha$  的置信区间为

$$\left( \frac{\hat{u}'\hat{u}}{\chi^2_{\frac{\alpha}{2}}(n-k-1)}, \frac{\hat{u}'\hat{u}}{\chi^2_{1-\frac{\alpha}{2}}(n-k-1)} \right)$$

### 假设检验

#### 1. 偏回归系数 $\beta_j$ 的 $t$ 检验

$$H_0: \beta_j = b, H_1: \beta_j \neq b$$

构造  $t$  统计量

$$t = \frac{\hat{\beta}_j - b}{SE(\hat{\beta}_j)}$$

拒绝域为

$$|t| \geq t_{\frac{\alpha}{2}}(n-k-1)$$

#### 2. 对参数线性组合的 $t$ 检验

$$H_0: \beta_1 = \beta_2, H_1: \beta_1 \neq \beta_2$$

构造  $t$  统计量

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

其中

$$\hat{\beta}_1 - \hat{\beta}_2 = a'\hat{\beta}, a = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}, SE(\hat{\beta}_1 - \hat{\beta}_2) = a'\hat{\sigma}^2(X'X)^{-1}a$$

拒绝域为

$$|t| \geq t_{\frac{\alpha}{2}}(n-k-1)$$

#### 3. 回归系数 $\beta$ 的 $F$ 检验

$$H_0: A\beta = b, H_1: A\beta \neq b$$

设  $r(A) = m$ , 构造  $F$  统计量

$$F = \frac{(A\hat{\beta} - b)'(\hat{\sigma}^2 A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b)}{m} \sim F(m, n-k-1)$$

拒绝域为

$$F \geq F_{\alpha}(m, n-k-1)$$

4. 对冗余变量的  $F$  检验

$$H_0 : \beta_{k-q+1} = \cdots = \beta_k = 0, H_1 : H_0 \text{ 为假}$$

无约束回归  $u$ :  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + \cdots + \beta_k X_k + u$

有约束回归  $r$ :  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + u$

基于残差平方和  $SSR$  的  $F$  检验:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F(q, n-k-1)$$

基于  $R^2$  的  $F$  检验:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n-k-1)} \sim F(q, n-k-1)$$

拒绝域为

$$F \geq F_\alpha(q, n-k-1)$$

# 4. 工具变量回归

4.1	内生性	38
4.2	工具变量估计	39
4.3	两阶段最小二乘	39
4.4	工具变量检验	40
4.4.1	相关性检验	40
4.4.2	外生性检验	40

## 4.1 内生性

在线性回归模型中，我们有很重要的外生性假设，即满足  $\mathbb{E}[Xu] = 0$ ，然而现实中常常会发生内生性问题，即  $\mathbb{E}[Xu] \neq 0$ ，这时线性回归模型便失效了。这便是我们在第三章引言部分介绍的遗漏变量偏误。在这种情况下，为了将  $Y = X'\beta + u$  与线性回归模型的情况进行区分，我们将该式称为**结构方程**。

以下是内生性出现的一些实例：

### 1. 度量误差

假设随机变量  $Y, Z$  满足  $\mathbb{E}[Y|Z] = Z'\beta$ ，然而  $Z$  无法观测，于是我们用  $X = Z + e$  代替  $Z$  进行回归，其中  $e$  是测量误差，且满足  $\mathbb{E}[Ze] = 0$ ，则有

$$Y = Z'\beta + u = (X - e)'\beta + u = X'\beta + v$$

则  $v = u - e'\beta$ 。此时  $\mathbb{E}[Xv] = \mathbb{E}[(Z + e)(u - e'\beta)] = -\mathbb{E}[ee']\beta$ 。用  $Y$  对  $X$  进行偏回归，得到

$$\beta^* = \beta + (\mathbb{E}[XX'])^{-1}\mathbb{E}[Xv] = (1 - \mathbb{E}[XX']^{-1}\mathbb{E}[ee'])\beta = (1 - \mathbb{E}[ZZ' + ee']^{-1}\mathbb{E}[ee'])\beta$$

当测量误差较大，即  $\mathbb{E}[ee']$  较大时， $\beta^* \rightarrow 0$ 。

### 2. 供求曲线

设需求曲线方程为  $Q = -\beta_1 P + u_1$ ，供给曲线方程为  $Q = \beta_2 Q + u_2$ ，写成矩阵形式为

$$\begin{pmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{pmatrix} \begin{pmatrix} Q \\ P \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$



解得

$$\begin{pmatrix} Q \\ P \end{pmatrix} = \frac{1}{\beta_1 + \beta_2} \begin{pmatrix} \beta_2 u_1 + \beta_1 u_2 \\ u_1 - u_2 \end{pmatrix}$$

设  $Q$  对  $P$  的回归方程为  $Q = \beta^* P + u^*$  则

$$\beta^* = \frac{\mathbb{E}[PQ]}{\mathbb{E}[P^2]} = \frac{\beta_2 - \beta_1}{2}$$

可以发现,  $\beta^*$  既不等于  $\beta_1$ , 也不等于  $\beta_2$ .

## 4.2 工具变量估计

**定义 4.2.1** 对于结构方程  $Y = X'\beta + u$ , 其中  $X$  中含  $k$  个解释变量, 若能找到一个变量  $Z_{l \times 1}$ , 满足

(1) 相关性:  $r(\mathbb{E}[ZX']) = k$  (秩条件);

(2) 外生性:  $\mathbb{E}(Zu) = 0$ ,

则称  $Z$  是  $X$  的工具变量 (Instrumental Variable) .

相关性的这一条件我们又称为秩条件, 要满足该条件须有  $l \geq k$ , 即阶条件. 根据是否满足阶条件可分为三种情况: 不可识别 (unidentified):  $l < k$ ; 恰好识别 (just or exactly identified):  $l = k$ ; 过度识别 (overidentified):  $l > k$ .

当  $l = k$  时, 注意到式子

$$\mathbb{E}[ZY] = \mathbb{E}[ZX'\beta] = \mathbb{E}[ZX']\beta$$

则

$$\beta = (\mathbb{E}[ZX'])^{-1} \mathbb{E}[ZY]$$

则  $\beta$  的 IV 估计为

$$\hat{\beta}^{IV} = \left( \sum_{i=1}^n Z_i X_i' \right)^{-1} \left( \sum_{i=1}^n Z_i Y \right) = (Z'X)^{-1} Z'Y$$

显然, 工具变量估计仅适用于恰好识别的情况.

 注

$l = k$  是矩阵  $\mathbb{E}[ZX']$  可逆的必要条件..

## 4.3 两阶段最小二乘

当  $l \geq k$  时, 我们作两阶段最小二乘:

第一阶段回归:  $X$  对  $Z$  进行回归, 建立结构方程  $X = Z'\gamma + u_1$ , 得到

$$\hat{\gamma} = (Z'Z)^{-1} Z'X$$

$$\hat{X} = PX = Z(Z'Z)^{-1} Z'X$$

将  $X$  的结构方程代入  $Y$  的结构方程, 得到

$$Y = X'\beta + u = (\hat{X} + \hat{u}_1)'\beta + u = \hat{X}'\beta + u_2$$

第二阶段回归:  $Y$  对  $\hat{X}$  进行回归, 建立结构方程  $Y = \hat{X}'\beta + u_2$ , 得到

$$\hat{\beta}^{TSLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$

由于  $Y = X'\beta + u = \hat{X}'\beta + (u + (X - \hat{X})'\beta)$ , 则有

$$\mathbb{E}[\hat{X}u_2] = \mathbb{E}[\hat{X}(u + (X - \hat{X})'\beta)] = \mathbb{E}[Z'\hat{\gamma}(u + u_2'\beta)] = \mathbb{E}[Z'\hat{\gamma}u] + \mathbb{E}[(u_2Z)'\beta] = 0$$

此时  $\hat{X}$  和扰动项正交.

当  $l = k$  时, 矩阵  $X'Z$  可逆, 则

$$\hat{\beta}^{TSLS} = (Z'X)^{-1}Z'Y$$

两阶段最小二乘的实质便是: 将内生解释变量  $X$  分成两个部分, 即由工具变量  $Z$  所造成的外生部分  $\hat{X}$  和与扰动项相关的其余部分  $X - \hat{X}$ ; 然后, 被解释变量对外生部分进行回归, 从而满足 OLS 对前定变量的要求而得到一致估计.

## 4.4 工具变量检验

### 4.4.1 相关性检验

易知当  $l = k$  时,

$$\hat{\beta}^{TSLS} = (Z'X)^{-1}Z'Y = \beta + (Z'X)^{-1}Z'u$$

当  $Z$  和  $X$  相关性微弱, 即  $Z$  为弱工具变量时,  $(Z'X)^{-1}Z'u$  无法依概率收敛到 0, 此时  $\hat{\beta}^{TSLS}$  不是  $\beta$  的一致性估计.

此时, 我们用  $F$  检验进行弱工具变量的检验, 可以证明

$$\mathbb{E}[\hat{\beta}^{TSLS}] - \beta \approx \frac{\hat{\beta}^{OLS} - \beta}{\mathbb{E}[F] - 1}$$

其中  $F$  是两阶段最小二乘第一阶段回归中  $\gamma = 0$  这一检验的  $F$  统计量 (回顾 3.4 中回归系数的  $F$  检验).

经验法则: 当仅有一元内生回归变量时, 若  $F < 10$ ,  $TSLS$  估计量有偏, 且  $TSLS$  估计的  $t$  统计量和置信区间均不可靠.

### 4.4.2 外生性检验

当外生性不满足时,  $\mathbb{E}[Z'u] \neq 0$ , 此时  $TSLS$  估计量不一致. 此时, 我们用过度识别约束检验 ( $J$  检验) 对外生性进行检验.

建立  $\hat{u}$  对工具变量的回归方程

$$\hat{u} = \delta_0 + Z'\delta + W'\xi + e$$

令  $F$  为检验  $\delta = 0$  的同方差适用  $F$  统计量, 构造  $J$  统计量

$$J = lF$$

可以证明, 当所有工具变量都为外生的原假设和同方差假设下, 大样本条件下  $J \sim \chi^2(l-k)$ .

# 5. 面板数据回归

5.1	固定效应模型	42
5.1.1	个体固定效应	42
5.1.2	时间固定效应	43
5.2	随机效应模型	43
5.3	双重差分法	44

## 5.1 固定效应模型

### 5.1.1 个体固定效应

定义 5.1.1 定义

$$Y_{it} = X'_{it}\beta + Z'_i\delta + u_i + \epsilon_{it}$$

为个体固定效应模型，其中  $X_{it}$  随时间和个体而变， $Z_i$  不随时间变化但随个体变化， $u_i$  为不可观测的代表个体异质性且与某个解释变量相关的截距项（称为个体固定效应）， $\epsilon_{it}$  为随时间和个体而变的扰动项。

若将个体虚拟变量引入该模型，则上式可写成

$$Y_{it} = X'_{it}\beta + Z'_i\delta + \gamma_2D2_i + \gamma_3D3_i + \cdots + \gamma_NDN_i + \epsilon_{it}$$

注意，引入的虚拟变量个数应为  $N - 1$ ，若引入  $N$  个虚拟变量，则会引起完全多重共线性。

### 固定效应估计

对于固定效应模型，给定个体  $i$ ，将回归方程两边取时间上的平均，得

$$\bar{Y}_i = \bar{X}'_i\beta + Z'_i\delta + u_i + \bar{\epsilon}_i$$

两式相减，得

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + \epsilon_{it} - \bar{\epsilon}_i$$

此时，可以得到  $\beta$  的固定效应估计，记为  $\hat{\beta}_{FE}$ ，即

$$\bar{Y}_{it} = \bar{X}'_{it}\beta + \bar{\epsilon}_{it}$$

在该估计中, 扰动项必须与**各期**解释变量均不相关, 满足严格外生性, 即保证  $X_{it} - \bar{X}_i$  和  $\epsilon_{it} - \bar{\epsilon}_i$  独立.

### 一阶差分估计

对固定效应模型两边进行一阶差分, 得

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + \epsilon_{it} - \epsilon_{i,t-1}$$

此时, 可以得到  $\beta$  的一阶差分估计, 记为  $\hat{\beta}_{FD}$ .

在该估计中, 只要求满足  $X_{it} - X_{i,t-1}$  和  $\epsilon_{it} - \epsilon_{i,t-1}$  这一假设, 此一致性条件比固定效应估计的一致性条件更弱. 可以证明, 当  $T = 2$  时,  $\hat{\beta}_{FE} = \hat{\beta}_{FD}$ ; 当  $T > 2$  时, 若  $\{\epsilon_{it}\}$  独立同分布,  $\hat{\beta}_{FE}$  比  $\hat{\beta}_{FD}$  更有效率.

### 5.1.2 时间固定效应

#### 定义 5.1.2 定义

$$Y_{it} = X'_{it}\beta + \gamma S_t + \epsilon_{it} = X'_{it}\beta + \lambda_t + \epsilon_{it}$$

为**时间固定效应模型**, 其中  $X_{it}$  随时间和个体而变,  $\lambda_t$  为第  $t$  期独有的截距项 (称为**时间固定效应**),  $\epsilon_{it}$  为随时间和个体而变的扰动项.

若将时间虚拟变量引入该模型, 则上式可写成

$$Y_{it} = X'_{it}\beta + \lambda_2 B2_t + \lambda_3 B3_t + \cdots + \lambda_T B T_t + \epsilon_{it}$$

类似地, 引入的虚拟变量个数应为  $T - 1$ . 若将个体固定效应和时间固定效应混合, 可建立模型

$$Y_{it} = X'_{it}\beta + Z'_i\delta + \lambda t + u_i + \epsilon_{it}$$

引入所有虚拟变量, 可得

$$Y_{it} = X'_{it}\beta + Z'_i\delta + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_N D N_i + \lambda_2 B2_t + \lambda_3 B3_t + \cdots + \lambda_T B T_t + \epsilon_{it}$$

#### 注

我们称左式  $\lambda t$  为**时间趋势项**, 若该项写成虚拟变量的形式, 则为**时间固定效应**的一种. 时间趋势项不一定为一次, 可以为高次, 它可以看成是时间固定效应的线性组合, 因此时间固定效应项往往能吸收时间趋势项.

## 5.2 随机效应模型

#### 定义 5.2.1 定义

$$Y_{it} = X'_{it}\beta + Z'_i\delta + u_i + \epsilon_{it}$$

为**随机效应模型**, 其中  $u_i$  和  $X_{it}$ 、 $Z_i$  均不相关.

此时我们可以得到  $\beta$  的 OLS 一致性估计. 然而, 由于扰动项由  $u_i + \epsilon_{it}$  组成, 不是球形扰动项, 因此 OLS 不是最有效率的. 假设不同个体之间的扰动项互不相关, 由于  $u_i$  的存

在，同一个体不同时期的扰动项之间存在自相关，即

### 注

由于面板数据的特点，虽然通常可以假设不同个体之间的扰动项相互独立，但同一个体在不同时期的扰动项之间往往存在自相关，因此对标准误的估计应使用聚类稳健标准误。所谓聚类，就是由每个个体不同时期的所有观测值所组成。

$$\text{COV}(u_i + \epsilon_{it}, u_i + \epsilon_{is}) = \begin{cases} \sigma_u^2, & \text{若 } t \neq s \\ \sigma_u^2 + \sigma_\epsilon^2, & \text{若 } t = s \end{cases}$$

当  $t \neq s$  时，自相关系数满足

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$$

因此，可用 OLS 的残差估计  $\sigma_u^2 + \sigma_\epsilon^2$ ，用 FE 的残差估计  $\sigma_\epsilon^2$ ，用广义最小二乘法来估计，得到随机效应估计量  $\hat{\beta}_{RE}$ ，即

$$Y_{it} - \hat{\theta}\bar{Y}_i = (X_{it} - \hat{\theta}\bar{X}_i)'\beta + (1 - \hat{\theta})Z_i'\delta + [(1 - \hat{\theta})u_i + (\epsilon_{it} - \hat{\theta}\bar{\epsilon}_i)]$$

其中  $\hat{\theta}$  是  $\theta$  的一致估计量， $\theta$  满足

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{T\sigma_u^2 + \sigma_\epsilon^2}}, 0 \leq \theta \leq 1$$

可以证明，新的扰动项是同方差且无序列相关的。

## 5.3 双重差分法

现考虑两期面板模型

$$Y_{it} = \alpha + \gamma D_t + \beta X_{it} + u_i + \epsilon_{it}$$

其中， $D_t$  为实验期虚拟变量， $u_i$  为不可测的个体特征， $X_{it}$  为政策虚拟变量，且有

$$D_t = \begin{cases} 1, & t = 2 \\ 0, & t = 1 \end{cases}, X_{it} = \begin{cases} 1, & \text{若 } i \in \text{处理组, 且 } t = 2 \\ 0, & \text{其他} \end{cases}$$

在 1.4 中，我们介绍了潜在结果框架，了解到在随机实验的条件下，零条件均值假设的成立。然而，如果实验未能完全随机化，则  $X_{it}$  可能与  $u_i$  有关，从而导致 OLS 估计不一致。

因此，我们用一阶差分对上式进行处理，用第二期减去第一期，消去  $u_i$ ，得

$$\Delta Y_{it} = \gamma + \beta \Delta X_{it} + \Delta \epsilon_{it}$$

根据潜在结果框架，

$$\hat{\beta}_{ols} = \Delta \bar{Y}_T - \Delta \bar{Y}_C$$

这种估计方法称为**双重差分估计**（DD），估计量记为  $\hat{\beta}_{DD}$ 。

显然，以  $\Delta Y_{it}$  为被解释变量的双重差分法不适用于多期数据，因此需回到以  $Y_{it}$  为被

解释变量的面板模型. 建立与原模型等价的方程

$$Y_{it} = \beta_0 + \beta_1 G_i \cdot D_t + \beta_2 G_i + \gamma D_t + \epsilon_{it}$$

其中,  $G_i$  为分组虚拟变量, 刻画处理组和控制组本身的差异;  $D_t$  为时间虚拟变量, 刻画实验前后两期本身的差异; 互动项  $G_i \cdot D_t = X_{it}$  刻画处理组的政策效应, 且

$$G_i = \begin{cases} 1, & \text{若 } i \in \text{处理组} \\ 0, & \text{若 } i \in \text{控制组} \end{cases}, D_t = \begin{cases} 1, & t = 2 \\ 0, & t = 1 \end{cases}$$

下面证明上式中的  $\hat{\beta}_1$  就是  $\hat{\beta}_{DD}$ , 即两个方程等价.

令  $t = 1$ , 则有

$$Y_{i1} = \beta_0 + \beta_2 G_i + \epsilon_{i1}$$

令  $t = 2$ , 则有

$$Y_{i2} = \beta_0 + \beta_1 G_i \cdot D_2 + \beta_2 G_i + \gamma + \epsilon_{i2}$$

两式相减, 得

$$\Delta Y_i = \beta_1 G_i \cdot D_2 + \gamma + \Delta \epsilon_i = \beta_1 X_{i2} + \gamma + \Delta \epsilon_i$$

因此  $\hat{\beta}_1 = \hat{\beta}_{DD}$ .

双重差分法的优点在于, 它同时控制了分组效应  $G_i$  和时间效应  $D_t$ . 其隐含假设是, 即使没有政策变化, 控制组和处理组的时间趋势也一样, 在方程中表现为共同的  $\gamma D_t$  这一项.

如果处理组和控制组没有共同的时间趋势, 上述模型需要进行调整, 我们需要在模型中加入时间趋势项或时间固定效应来控制二者的时间趋势. 如果加入时间趋势项, 则引入变量初始值与时间的交乘项; 如果加入时间固定效应, 则引入变量初始值与时间固定效应的交乘项. 另外, 三重差分法也可以控制时间趋势, 但设置的交乘项更为复杂.