



# 高级计量经济学 1

## Advanced Econometrics 1

时间: August 11, 2023

# 目录

<b>第 1 章 条件期望与投影</b>	<b>1</b>	4.4.1 回归系数的函数 . . . . .	64
1.1 迭代期望定律 . . . . .	1	4.4.2 $T$ 检验和 Wald 检验 . . .	65
1.2 CEF 误差 . . . . .	4	4.5 异方差的检验 . . . . .	68
1.3 最优预测量与条件方差 . . . . .	6	<b>第 5 章 系统估计</b>	<b>70</b>
1.4 回归导数 . . . . .	8	5.1 回归系统 . . . . .	70
1.5 线性 CEF . . . . .	9	5.2 系统普通最小二乘估计 . . . . .	71
1.6 最优线性预测量 . . . . .	10	5.3 SOLS 的期望与方差 . . . . .	72
1.7 分块回归与遗漏变量偏误 . . .	13	5.4 SOLS 的渐近性质 . . . . .	73
<b>第 2 章 经典线性回归模型</b>	<b>16</b>	5.5 SGLS 的渐近性质 . . . . .	75
2.1 CLRM 假设 . . . . .	16	5.6 似不相关回归 . . . . .	77
2.2 普通最小二乘估计 . . . . .	17	<b>第 6 章 工具变量回归分析</b>	<b>81</b>
2.3 投影矩阵与消灭矩阵 . . . . .	19	6.1 内生性问题 . . . . .	81
2.4 杠杆值与留一回归 . . . . .	20	6.1.1 测量误差 . . . . .	81
2.5 拟合优度 . . . . .	22	6.1.2 联立方程偏误 . . . . .	82
2.6 OLS 的有限样本性质 . . . . .	24	6.2 工具变量 . . . . .	83
2.7 OLS 的协方差矩阵估计 . . . . .	26	6.2.1 工具变量的定义 . . . . .	83
2.8 分块回归 . . . . .	27	6.2.2 结构参数的识别 . . . . .	83
2.9 正态假设下的参数检验 . . . . .	29	6.2.3 工具变量估计量 . . . . .	85
2.9.1 $T$ 检验 . . . . .	30	6.3 二阶段最小二乘估计 . . . . .	86
2.9.2 $F$ 检验 . . . . .	32	6.4 2SLS 的渐近性质 . . . . .	88
2.10 广义最小二乘估计 . . . . .	34	6.5 生成回归元 . . . . .	91
2.11 聚类样本 . . . . .	37	6.6 含期望误差的回归 . . . . .	93
<b>第 3 章 渐近理论基础</b>	<b>40</b>	6.7 控制函数法 . . . . .	95
3.1 收敛概念 . . . . .	40	6.8 模型设定检验 . . . . .	97
3.2 大数定律 . . . . .	46	6.8.1 内生性检验 . . . . .	97
3.3 中心极限定理 . . . . .	49	6.8.2 过度识别检验 . . . . .	99
3.4 一致可积与矩收敛 . . . . .	52	6.8.3 弱工具变量 . . . . .	100
<b>第 4 章 最小二乘的渐近性质</b>	<b>56</b>	<b>第 7 章 广义矩方法</b>	<b>102</b>
4.1 OLS 的一致性 . . . . .	56	7.1 矩方程与 GMM 估计量 . . . . .	102
4.2 OLS 的渐近正态性 . . . . .	57	7.2 GMM 的渐近性质 . . . . .	103
4.3 渐近方差估计量 . . . . .	60	7.2.1 一致性 . . . . .	103
4.4 参数检验 . . . . .	64	7.2.2 渐近正态性 . . . . .	106

7.2.3 二阶段 GMM 估计量 . . .	108	9.3 CMLE 的渐近性质 . . . . .	140
7.3 过度识别检验 . . . . .	109	9.3.1 一致性 . . . . .	140
7.4 系统工具变量 . . . . .	110	9.3.2 得分函数与条件信息矩阵	142
<b>第 8 章 面板数据模型</b>	<b>113</b>	9.3.3 渐近正态性 . . . . .	143
8.1 面板数据 . . . . .	113	9.4 CMLE 的有效性 . . . . .	145
8.2 混合回归 . . . . .	113	9.5 参数检验 . . . . .	147
8.3 随机效应模型 . . . . .	114	9.5.1 Wald 检验 . . . . .	147
8.3.1 RE 估计量 . . . . .	114	9.5.2 似然比检验 . . . . .	148
8.3.2 一般 FGLS 分析 . . . . .	117	9.5.3 Lagrange 乘子检验 . . .	150
8.4 固定效应模型 . . . . .	118	9.6 模型设定检验 . . . . .	150
8.4.1 组内估计 . . . . .	118	9.7 拟极大似然估计 . . . . .	151
8.4.2 虚拟变量回归 . . . . .	120	9.7.1 一般误设 . . . . .	151
8.4.3 一阶差分估计 . . . . .	121	9.7.2 模型选择检验 . . . . .	152
8.4.4 组间估计 . . . . .	122	9.7.3 线性指数族的 QMLE . .	153
8.5 广义离差模型 . . . . .	123	<b>第 10 章 限值因变量模型</b>	<b>155</b>
8.6 FE 的渐近性质 . . . . .	124	10.1 二值响应模型 . . . . .	155
8.7 RE 与 FE 的比较 . . . . .	126	10.1.1 Probit 与 Logit . . . . .	155
8.8 双向误差成分 . . . . .	128	10.1.2 内生性问题 . . . . .	157
8.9 工具变量 . . . . .	129	10.2 断尾回归模型 . . . . .	158
8.10 Hausman-Taylor 模型 . . . . .	130	10.3 归并回归模型 . . . . .	159
8.11 动态面板数据模型 . . . . .	132	10.3.1 Tobit 模型 . . . . .	159
8.11.1 FE 估计的偏误 . . . . .	132	10.3.2 栅栏模型 . . . . .	162
8.11.2 Anderson-Hsiao 估计量	133	10.4 样本选择问题 . . . . .	163
8.11.3 GMM 估计量 . . . . .	134	10.4.1 一致性 OLS 估计 . . . .	163
<b>第 9 章 极大似然估计</b>	<b>137</b>	10.4.2 从属断尾 . . . . .	164
9.1 预备内容 . . . . .	137	<b>参考文献</b>	<b>167</b>
9.2 CMLE 的一般框架 . . . . .	138		



# 第 1 章 条件期望与投影

计量经济学中最主要的工具是回归方法, 用来估计在给定回归元时, 响应变量的条件期望. 本章主要讨论一般条件下的回归分析, 重点放在条件期望模型和线性投影.

## 1.1 迭代期望定律

随机变量  $X : \Omega \rightarrow \mathbb{R}$  为定义在概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$  上的  $\mathcal{F}$ -可测实值函数, 也即对于任意  $a \in \mathbb{R}$  都有

$$X^{-1}((-\infty, a]) \equiv \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$$

$X$  具有累积分布函数

$$F(x) = \mathbb{P}[X \leq x], \quad x \in \mathbb{R}$$

并且  $X$  的期望为

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, dF(x)$$

特别地, 如果  $X$  为离散型随机变量, 则

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} \tau_j \mathbb{P}[X = \tau_j]$$

如果  $X$  为连续型随机变量, 则  $F$  为绝对连续函数,  $X$  具有概率密度函数  $f$ , 则

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, dF(x) = \int_{\mathbb{R}} x f(x) \, dx$$

这里的积分为 Lebesgue 积分. 更一般地, 如果  $h : \mathbb{R} \rightarrow \mathbb{R}$  为 Borel 可测函数,  $Y = h(X)$  且  $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ , 那么

$$\mathbb{E}[Y] = \int_{\mathbb{R}} h(x) \, dF(x)$$

不严格地说, 条件期望函数<sup>1</sup> (Conditional Expectation Function, CEF):

$$m(x) = \mathbb{E}[Y|X = x]$$

是一个关于  $x \in \mathbb{R}^K$  的函数, 它表示  $X = x$  时, 随机变量  $Y$  的期望. 为了估计  $m(x)$  在  $X$  处的值, 有时也将 CEF 视为随机变量  $X$  的函数, 记作  $m(X)$  或  $\mathbb{E}[Y|X]$ , 此时 CEF 是随机的.

假定二元有序对  $(X, Y)$  是定义在乘积空间上的连续型随机变量, 即联合分布函数  $F_{XY}$  绝对连续, 联合概率密度为  $f_{XY}$  存在, 那么  $X$  的边缘概率密度为

$$f_X(x) = \int_{\mathbb{R}} f_{XY}(x, y) \, dy$$

<sup>1</sup>条件期望的正式定义为: 设  $X : \Omega \rightarrow \mathbb{R}$  为定义在概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$  上的随机变量, 并且  $\mathbb{E}|X| < \infty$ , 如果  $\mathcal{F}_0 \subset \mathcal{F}$  是  $\mathcal{F}$  的子  $\sigma$ -代数, 则  $X$  在条件  $\mathcal{F}_0$  下的条件期望为满足以下条件的随机变量  $Y$ : (i)  $Y$  是  $\mathcal{F}_0$ -可测函数, (ii) 对任意的  $A \in \mathcal{F}_0$  都有  $\mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A]$ .

给定  $X = x$  时,  $Y$  的条件概率密度为

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \quad (1.1)$$

$f_{Y|X}(y|x)$  完全描述了  $Y$  对  $x$  的依赖关系. 进一步, 可以将 CEF 表示为

$$m(x) = \mathbb{E}[Y|X = x] = \int_{\mathbb{R}} y f_{Y|X}(y|x) dy$$

然而事实上, 即使条件密度 (1.1) 不是良定义的, 只要  $\mathbb{E}|Y| < \infty$  成立, 任意随机变量  $(X, Y)$  的 CEF 就存在, 并且几乎必然唯一, 它的证明具体可参考 Durrett (2019) 中的 Theorem 4.1.2.

**注** 条件概率分布反映的仅是  $Y$  与  $X$  间的预测关系, 而非因果关系, 换言之它无法刻画  $X$  的变化将如何导致  $Y$  变化, 而因果关系的刻画需要根据经济理论.

更具体地说, 如果  $X$  对  $Y$  具有因果影响, 那么  $X$  对  $Y$  也具有预测能力, 然而  $X$  可以预测  $Y$  却并不意味着  $X$  对  $Y$  之间存在因果关系.

**例 1.1** 假设  $(X, Y)$  的联合概率密度为  $f_{XY} = e^{-y}$ ,  $0 < x < y < \infty$ , 则  $X$  的边际概率密度为

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy = \int_x^{\infty} e^{-y} dy = e^{-x}, \quad x \in (0, \infty)$$

那么对于任意给定的  $x > 0$  都有

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = e^{-(y-x)}, \quad y \in (x, \infty)$$

于是可以计算条件期望

$$\begin{aligned} \mathbb{E}[Y|X = x] &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \\ &= \int_x^{\infty} x e^{-(y-x)} dy \\ &= -e^x \int_x^{\infty} y dy = 1 + x \end{aligned}$$

此时 CEF  $\mathbb{E}[Y|X] = 1 + X$  是  $X$  的线性函数.

现在介绍在概率论和计量经济学中极其重要的迭代期望定律 (Law of Iterated Expectation, LIE).

### 定理 1.1 (简单迭代期望定律)

如果对任意随机向量  $X$  都有  $\mathbb{E}|Y| < \infty$ , 那么

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



**证明** 注意到  $f_{Y|X}(y|x)f_X(x) = f(x, y)$ , 根据 Fubini 定理可得

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Y|X]] &= \int_{\mathbb{R}^K} \mathbb{E}[Y|X] f_X(x) dx = \int_{\mathbb{R}^K} \left( \int_{\mathbb{R}} y f_{Y|X}(y|x) dy \right) f_X(x) dx \\ &= \int_{\mathbb{R}^K} \left( \int_{\mathbb{R}} y f(x, y) dy \right) dx = \int_{\mathbb{R}} y \left( \int_{\mathbb{R}^K} f(x, y) dx \right) dy = \mathbb{E}[Y] \end{aligned}$$

由此证得定理.

简单来说, 随机变量  $Y$  在  $X = x$  下的条件期望的期望等于  $Y$  的无条件期望. 若  $X$  是离散

型的, 则

$$\mathbb{E}[\mathbb{E}[Y|X]] = \sum_{j=1}^{\infty} \mathbb{E}[Y|X = x_j] \mathbb{P}[X = x_j]$$

若  $X$  是连续型的, 则

$$\mathbb{E}[\mathbb{E}[Y|X]] = \int_{\mathbb{R}^K} \mathbb{E}[Y|X = x] f_X(x) dx$$

其中  $f_X$  是  $X$  的概率密度.

**例 1.2** 设  $\mathbb{E}[Y|X] = X^2$ ,  $Y \sim U[0, 1]$ , 现在要求  $\mathbb{E}[X]$ . 尽管无法找到  $X$  的概率密度, 但是  $\mathbb{E}[X]$  可用通过 LIE 得到

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \int_0^1 x^2 dx = \frac{1}{3}$$

现在给出一个条件定理, 它是简单 LIE 的一种扩展.

### 定理 1.2 (条件定理)

对于任意的 Borel 可测函数  $g: \mathbb{R}^K \rightarrow \mathbb{R}$ , 如果  $\mathbb{E}|Y| < \infty$  成立, 那么

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

在此基础上若还有  $\mathbb{E}|g(Y)| < \infty$ , 那么

$$\mathbb{E}[g(X)Y] = \mathbb{E}[g(X)\mathbb{E}[Y|X]] \quad (1.2) \quad \heartsuit$$

**证明** 根据条件期望的定义可知

$$\begin{aligned} \mathbb{E}[g(X)Y|X = x] &= \int_{\mathbb{R}} g(x)y f_{Y|X}(y|x) dy \\ &= g(x) \int_{\mathbb{R}} y f_{Y|X}(y|x) dy = g(x)\mathbb{E}[Y|X = x] \end{aligned}$$

将上式两端取期望并应用简单 LIE 即可得到 (1.2).

最后给出更一般的迭代期望定律, 它允许更多的条件变量存在.

### 定理 1.3 (迭代期望定律)

如果对任意随机向量  $X_1$  和  $X_2$  都有  $\mathbb{E}|Y| < \infty$ , 那么

$$\mathbb{E}[\mathbb{E}[Y|X_1, X_2]|X_1] = \mathbb{E}[Y|X_1] \quad \heartsuit$$

**证明** 首先注意到

$$f(y|x_1, x_2)f(x_2|x_1) = \frac{f(y, x_1, x_2)}{f(x_1, x_2)} \frac{f(x_1, x_2)}{f(x_1)} = f(y, x_2|x_1)$$

以及

$$\mathbb{E}[Y|X_1 = x_1, X_2 = x_2] = \int_{\mathbb{R}} y f(y|x_1, x_2) dy$$

于是

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X_1, X_2]|X = x_1] &= \int_{\mathbb{R}^{K_2}} \mathbb{E}[Y|X_1 = x_1, X_2 = x_2] f(x_2|x_1) dx_2 \\ &= \int_{\mathbb{R}^{K_2}} \left( \int_{\mathbb{R}} y f(y|x_1, x_2) f(x_2|x_1) dy \right) dx_2 \\ &= \int_{\mathbb{R}^{K_2}} \left( \int_{\mathbb{R}} y f(y, x_2|x_1) dy \right) dx_2 = \mathbb{E}[Y|X_1 = x_1]\end{aligned}$$

其中  $x_2 \in \mathbb{R}^{K_2}$ .

条件期望的符号  $\mathbb{E}[Y|X]$  的严格意义是  $\mathbb{E}[Y|\sigma\langle X \rangle]$ , 其中  $\sigma\langle X \rangle$  表示由  $X$  生成的  $\sigma$ -代数, 由于  $\sigma\langle X_1 \rangle \subset \sigma\langle X_1, X_2 \rangle$  (也就是说, 由  $X_1, X_2$  生成的  $\sigma$ -代数包含的信息比  $\sigma\langle X_1 \rangle$  更多), 因此 LIE 表明: 当只有  $\sigma\langle X_1 \rangle$  可用时,  $\mathbb{E}[Y|X_1, X_2]$  的期望不可能从  $\sigma\langle X_1, X_2 \rangle$  包含的信息中得到, 最多只能从  $\sigma\langle X_1 \rangle$  中获得.

最后, 我们给出 LIE 的本质描述: 设  $Y: \Omega \rightarrow \mathbb{R}$  是定义在概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$  上的随机变量, 并且  $\mathbb{E}|Y| < \infty$ , 若子  $\sigma$ -代数  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}$ , 那么

$$\mathbb{E}[\mathbb{E}[Y|\mathcal{F}_2]|\mathcal{F}_1] = \mathbb{E}[Y|\mathcal{F}_1] \quad (1.3)$$

若  $\mathcal{F}_1 = \{\emptyset, \Omega\}$ , 则 (1.3) 变为简单 LIE 的形式:  $\mathbb{E}[\mathbb{E}[Y|\mathcal{F}_1]] = \mathbb{E}[Y]$ .

## 1.2 CEF 误差

利用  $m(X)$  所包含的信息去预测  $Y$  必然存在一定程度上的偏误, 将其定义为 CEF 误差 (又称回归误差)  $e = Y - m(X)$ , 也即  $Y$  与 CEF 在  $X$  处取值之差. 根据 CEF 误差的构造, 可以将其改写为以下形式

$$Y = m(X) + e \quad (1.4)$$

这表明  $e$  的性质来源于  $(X, Y)$  的联合分布, 我们将  $e$  的性质表述为以下定理.

### 定理 1.4

设  $X$  为任意随机向量,  $\mathbb{E}|Y| < \infty$ , 则以下命题成立:

- (1)  $\mathbb{E}[e|X] = 0$ , 称为零条件均值.
- (2)  $\mathbb{E}[e] = 0$ .
- (3) 如果对于某个  $r \geq 1$ ,  $\mathbb{E}|Y|^r < \infty$  成立, 那么  $\mathbb{E}|e|^r < \infty$ .
- (4) 对任意使得  $\mathbb{E}|h(X)e| < \infty$  的可测函数  $h$ ,  $\mathbb{E}[h(X)e] = 0$  成立.



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

(2)  $\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e|X]] = \mathbb{E}[0] = 0$ .

(3) 根据 Minkowski 不等式<sup>2</sup>和条件期望不等式<sup>3</sup>可知

$$(\mathbb{E}|e|^r)^{1/r} \leq (\mathbb{E}|Y|^r)^{1/r} + (\mathbb{E}|m(X)|^r)^{1/r} < 2(\mathbb{E}|Y|^r)^{1/r} < \infty$$

<sup>2</sup>对任意  $m \times n$  矩阵  $X$  和  $Y$ , 都有  $(\mathbb{E}\|X + Y\|^p)^{1/p} \leq (\mathbb{E}\|X\|^p)^{1/p} + (\mathbb{E}\|Y\|^p)^{1/p}$ .

<sup>3</sup>对于任意  $r \geq 1$  以及随机变量  $(Y, X) \in \mathbb{R} \times \mathbb{R}^K$ , 如果  $\mathbb{E}|Y| < \infty$ , 那么  $\mathbb{E}|\mathbb{E}[Y|X]|^r \leq \mathbb{E}|Y|^r < \infty$ .

(4) 根据 (1) 和 LIE 可知

$$\mathbb{E}[h(X)e] = \mathbb{E}[\mathbb{E}[h(X)e|X]] = \mathbb{E}[h(X)\mathbb{E}[e|X]] = \mathbb{E}[0] = 0$$

**注** 等式  $\mathbb{E}[e|X] = \mathbb{E}[e] = 0$  通常称为均值独立, 表明 CEF 误差  $e$  在给定  $X = x$  时的期望为 0 且不依赖于  $X$ , 但这并不代表  $e$  与  $X$  独立, 这是一个非常严格的条件, 要求对于任意有界 Borel 可测函数  $h_1: \mathbb{R}^K \rightarrow \mathbb{R}$ ,  $h_2: \mathbb{R} \rightarrow \mathbb{R}$  都有  $\mathbb{E}[h_1(X)h_2(e)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(e)]$ .

**例 1.3** 设  $e = Xu$ , 其中  $X$  和  $u$  独立, 且都服从标准正态分布  $N(0, 1)$ , 那么在条件  $X$  下  $e$  的分布为  $N(0, X^2)$ , 此时  $\mathbb{E}[e|X] = 0$  且  $e$  不与  $X$  独立.

现在需要进一步测度 CEF 函数的对  $Y$  的偏离程度, 考虑使用 CEF 误差  $e$  的无条件方差

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

称  $\sigma^2$  为回归方差或回归误差的方差, 其大小反映了: 在  $Y$  的变化中, 不能被条件期望  $\mathbb{E}[Y|X]$  所解释的程度. 根据定理 1.4(3), 如果  $\mathbb{E}[Y^2] < \infty$ , 那么显然有  $\sigma^2 < \infty$ .

回归误差的大小依赖于回归元  $X$ , 考虑如下两个回归方程

$$Y = \mathbb{E}[Y|X_1] + e_1$$

$$Y = \mathbb{E}[Y|X_1, X_2] + e_2$$

直观来看,  $X_1, X_2$  包含的信息比  $X_1$  的多, 故而  $\mathbb{E}[Y|X_1, X_2]$  能更多地解释  $Y$  的变化中的线性部分, 因此理应有  $\text{var}(e_2) \leq \text{var}(e_1)$ . 现在将其表述为定理.

### 定理 1.5

设  $X_1$  和  $X_2$  为随机向量, 如果  $\mathbb{E}[Y^2] < \infty$ , 那么

$$\text{var}(Y) \geq \text{var}(Y - \mathbb{E}[Y|X_1]) \geq \text{var}(Y - \mathbb{E}[Y|X_1, X_2])$$



**证明** 根据迭代期望定律  $\mathbb{E}[Y|X_1] = \mathbb{E}[\mathbb{E}[Y|X_1, X_2]|X_1]$  和条件 Jensen 不等式<sup>4</sup>可知

$$(\mathbb{E}[Y|X_1])^2 \leq \mathbb{E}[(\mathbb{E}[Y|X_1, X_2])^2|X_1]$$

两端取无条件期望得到

$$\mathbb{E}[(\mathbb{E}[Y|X_1])^2] \leq \mathbb{E}[(\mathbb{E}[Y|X_1, X_2])^2]$$

同样由  $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X_1]]$  和 Jensen 不等式<sup>5</sup>可知

$$(\mathbb{E}[Y])^2 \leq \mathbb{E}[(\mathbb{E}[Y|X_1])^2] \leq \mathbb{E}[(\mathbb{E}[Y|X_1, X_2])^2] \quad (1.5)$$

注意到  $Y$ ,  $\mathbb{E}[Y|X_1]$  和  $\mathbb{E}[Y|X_1, X_2]$  具有相同的期望  $\mathbb{E}[Y]$ , 于是 (1.5) 意味着

$$0 \leq \text{var}(\mathbb{E}[Y|X_1]) \leq \text{var}(\mathbb{E}[Y|X_1, X_2]) \quad (1.6)$$

定义  $e = Y - \mathbb{E}[Y|X]$  以及  $u = \mathbb{E}[Y|X] - \mu$ , 其中  $u$  和  $\mu$  满足

$$Y - \mu = e + u$$

<sup>4</sup> 设  $g: \mathbb{R}^L \rightarrow \mathbb{R}$  为下凸函数, 那么对任意使得  $\mathbb{E}\|Y\| < \infty$  和  $\mathbb{E}|g(Y)| < \infty$  的随机向量  $(Y, X) \in \mathbb{R}^L \times \mathbb{R}^K$ , 总有  $g(\mathbb{E}[Y|X]) \leq \mathbb{E}[g(Y)|X]$ .

<sup>5</sup> 设  $g: \mathbb{R}^L \rightarrow \mathbb{R}$  为下凸函数, 则对于任意使得  $\mathbb{E}\|X\| < \infty$  和  $\mathbb{E}|g(X)| < \infty$  的随机变量  $X \in \mathbb{R}^L$ , 总有  $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$ .



注意到  $\mathbb{E}[e|X] = 0$  且  $u$  是  $X$  的函数, 根据条件定理可知

$$\mathbb{E}[eu] = \mathbb{E}[u\mathbb{E}[e|X]] = 0$$

又因为  $\mathbb{E}[e] = 0$ , 故而  $e$  和  $u$  不相关, 从而

$$\text{var}(Y) = \text{var}(e) + \text{var}(u) = \text{var}(Y - \mathbb{E}[Y|X]) + \text{var}(\mathbb{E}[Y|X]) \quad (1.7)$$

将 (1.6) 应用于 (1.7) 即可证明定理.

## 1.3 最优预测量与条件方差

对于任意给定的随机向量  $X$ , 现在我们的目标是预测  $Y$ , 我们可以将  $X$  任意的预测量记作  $g(X)$ , 预测误差即为  $Y - g(X)$  的实现值, 可以  $g$  的均方误差来衡量其误差大小:

$$\text{MSE}(g) = \mathbb{E}[(Y - g(X))^2] \quad (1.8)$$

其中  $g: \mathbb{R}^K \rightarrow \mathbb{R}$  为可测函数, 我们定义最优预测量为使得 (1.8) 最小化的  $g(X)$ . 下面将要证明, 对于任意  $(Y, X)$  的联合分布, 最优预测量为 CEF  $m(X)$ .

### 定理 1.6

设  $X$  为任意随机向量, 如果  $\mathbb{E}[Y^2] < \infty$ , 那么对于任意预测量  $g(X)$  都有

$$\mathbb{E}[(Y - g(X))^2] \geq \mathbb{E}[(Y - m(X))^2]$$

其中  $m(X) = \mathbb{E}[Y|X]$ .



**证明** 注意到对任意预测量  $g(X)$  都有

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

其中由定理 1.4(4) 可知  $\mathbb{E}[e(m(X) - g(X))] = 0$ . 当且仅当  $g(X) = m(X)$  时, 上式大于等于可以取到等号.

在上面的论述中, 我们已经用条件期望描述了条件分布的“位置”, 但它没有告诉我们条件分布的分散程度, 为了对其进行刻画, 现在需要引入条件方差的概念.

### 定义 1.1

如果  $\mathbb{E}[W^2] < \infty$ , 那么给定条件  $X = x$  时,  $W$  的条件方差为

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

类似地, 可以将  $\sigma^2(X) = \text{var}[W|X]$  视为一个随机变量.



当 CEF 误差满足矩条件  $\mathbb{E}[e^2] < \infty$  时, 我们可以将条件  $X = x$  下, CEF 误差  $e$  的条件方

差定义为

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

同样可以将  $e$  的条件方差  $\sigma^2(X) = \text{var}(e|X)$  视为随机变量。

关于条件方差, 下面介绍一个重要定理, 称为总方差定律 (Law of Total Variance, LTV), 也称方差分解公式。

### 定理 1.7

如果  $X$  和  $W$  为定义在相同概率空间上的随机变量, 并且  $\mathbb{E}[X^2] < \infty$ , 那么

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

**证明** 根据  $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$  及 LIE 可得

$$\begin{aligned} \text{var}(\mathbb{E}[X|W]) &= \mathbb{E}[(\mathbb{E}[X|W])^2] - (\mathbb{E}[\mathbb{E}[X|W]])^2 \\ &= \mathbb{E}[(\mathbb{E}[X|W])^2] - (\mathbb{E}[X])^2 \end{aligned} \quad (1.9)$$

又根据  $\text{var}(X|W) = \mathbb{E}[X^2|W] - (\mathbb{E}[X|W])^2$  可得

$$\begin{aligned} \mathbb{E}[\text{var}(X|W)] &= \mathbb{E}[\mathbb{E}[X^2|W] - (\mathbb{E}[X|W])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|W])^2] \end{aligned} \quad (1.10)$$

将式 (1.9) 和 (1.10) 相加即可证得定理。

总方差定律将无条件方差分解为了“组内方差”和“组间方差”两个部分。举例而言, 如果  $X$  是受教育程度,  $W$  是年龄, 那么第一项  $\mathbb{E}[\text{var}(X|W)]$  刻画了所有年龄水平上受教育程度差异的均值, 第二项  $\text{var}(\mathbb{E}[X|W])$  则为不同年龄分组间平均受教育程度的差异。

从预测的角度看, 总方差定律表明:  $\mathbb{E}[Y|X]$  变化程度越大, 则它更能够解释  $Y$  的变化, 因而也能更好地预测  $Y$ 。由于  $\text{var}(Y)$  是一个常数, 最优预测量  $m(X)$  的变化越大, 则均方误差的期望越小。当  $\mathbb{E}[Y|X] = Y$  时, 均方误差为 0, CEF 可以完全刻画  $Y$  的变化。事实上, 利用 LTV 可以证明

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

由于回归误差  $e$  具有零条件均值, 所以它的无条件误差方差等于期望条件方差, 也即

$$\sigma^2 = \mathbb{E}[e^2] = \mathbb{E}[\mathbb{E}[e^2|X]] = \mathbb{E}[\sigma^2(X)]$$

也即无条件误差方差相当于是“平均的”条件误差方差。

给定条件方差后, 我们可以定义一个调整误差

$$u = \frac{e}{\sigma(X)} \quad (1.11)$$

由于  $\sigma^2(X)$  是  $X$  的函数, 可以计算出

$$\mathbb{E}[u|X] = \mathbb{E}\left[\frac{e}{\sigma(X)} \middle| X\right] = \frac{1}{\sigma(X)} \mathbb{E}[e|X] = 0$$

以及

$$\text{var}(u|X) = \mathbb{E}[u^2|X] = \mathbb{E}\left[\frac{e^2}{\sigma^2(X)} \middle| X\right] = \frac{1}{\sigma^2(X)} \mathbb{E}[e^2|X] = 1$$

也即  $u$  在给定  $X$  下的条件期望为 0, 条件方差为 1. 注意到 (1.11) 可以改写为

$$e = \sigma(X)u$$

将其代入到 CEF 方程中得到

$$Y = m(X) + \sigma(X)u$$

称为均值-方差形式的 CEF 方程.

最后我们给出同方差和异方差的概念, 这在后面推导参数估计量的渐近性质时具有重要作用.

### 定义 1.2

CEF 误差  $e$  是同方差的, 如果  $\sigma^2(x) = \sigma^2$  不依赖于  $x$ . 若  $\sigma^2(x)$  依赖于  $x$ , 则称  $e$  是异方差的.



**注** 这里定义的同方差和异方差都是条件方差意义下的, 而非无条件方差. 显然, 无条件方差  $\sigma^2$  是一个固定的常数, 它独立于回归元  $X$ , 也就无所谓同方差还是异方差了.

在现实的经济社会中, 同方差是几乎难以达成的苛刻条件, 而异方差才是对回归模型的一个更标准和更稳健的假设. 关于同方差和异方差的图示, 具体可以参考 Wooldridge (2019) 中的 Figure 2.8 与 Figure 2.9.

造成异方差的其中一个原因在于, 如果  $\mathbb{E}[Y|X]$ , 那么  $\text{var}(Y|X)$  和更高阶的条件矩也很有可能依赖于  $X$ . 倘若我们考虑现实的因素, 则还有其他可能的解释, 例如在 OCED 国家中, 政府规模大小和 GDP 波动率具有显著的负相关关系 (Fatás and Mihov, 2001).

## 1.4 回归导数

现在的问题是如何解释在 CEF  $m(x) = \mathbb{E}[Y|X = x]$  中, 回归元  $x$  的边际变化如何导致响应变量  $Y$  的条件期望变化. 通常而言, 如果要考虑单个回归元  $X_1$  边际变化, 此时应该保持其它因素不变. 当  $X_1$  具有连续分布时, 定义  $X_2, \dots, X_k$  不变时  $x_1$  变化的边际效应为

$$\frac{\partial}{\partial x_1} m(x_1, x_2, \dots, x_k)$$

若  $X_1$  是离散型的, 则  $X_1$  的边际效应为离差形式. 例如, 若  $X_1$  为二值变量, 则它在其它条件不变时对 CEF 的边际效应为

$$m(1, x_2, \dots, x_k) - m(0, x_2, \dots, x_k)$$

我们可以用如下记号表述边际效应

$$\nabla_1 m(x) = \begin{cases} \frac{\partial}{\partial x_1} m(x_1, x_2, \dots, x_k), & \text{如果 } X_1 \text{ 为连续型} \\ m(1, x_2, \dots, x_k) - m(0, x_2, \dots, x_k), & \text{如果 } X_1 \text{ 为二值变量} \end{cases}$$

将  $K - 1$  个解释变量的边际效应写成列向量的形式, 则关于  $X$  的回归导数为

$$\nabla m(x) = \begin{bmatrix} \nabla_1 m(x) \\ \nabla_2 m(x) \\ \vdots \\ \nabla_k m(x) \end{bmatrix}$$

当  $x$  中的所有元素都为连续型随机变量时, 回归导数可以简化为  $\nabla m(x) = \partial m(x) / \partial x$ .

当我们计算某个解释变量对 CEF 的边际效应时, 需要保持其它因素不变, 这通常称为 *Ceteris Paribus* 概念. 然而实际上, 我们只能控制那些被包括进了  $\mathbb{E}[Y|X]$  中的其它因素不变, 无法做到保持其它所有一切的因素不变, 因此回归导数依赖于解释变量的选取.

另一方面, 回归导数  $\nabla m(x)$  刻画的是  $Y$  的条件期望的变化, 而非  $Y$  的具体值的变化. 只有当  $e$  不受回归元  $X$  影响时,  $\nabla m(x)$  才是  $Y$  的真实值的变化.

## 1.5 线性 CEF

CEF 中极其特殊的一种情况是线性 CEF, 也即  $m(x) = \mathbb{E}[Y|X = x]$  线性于  $x$ , 此时

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

定义记号  $X = [1, X_1, X_2, \cdots, X_k]'$ , 则线性 CEF 为

$$m(x) = x' \beta \quad (1.12)$$

其中  $\beta = [\beta_0, \beta_1, \cdots, \beta_k]'$  为  $K \times 1$  参数向量. 将线性 CEF 代入  $Y = m(X) + e$  中可得

$$Y = X' \beta + e$$

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

称为  $Y$  对  $X$  的线性回归模型 (Linear Regression Model, LRM).

在线性 CEF 中, 回归导数可以简化为参数向量, 也即  $\nabla m(x) = \beta$ , 这是线性 CEF 一个颇具吸引力的特点. 回归系数  $\beta_j$  可以解释为保持其它条件不变时,  $X_j$  变化对 CEF 带来的边际效应.

**例 1.4** 考虑如下工资方程

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exp} + e$$

其中 **wage** 为工资, **educ** 为受教育程度, **exp** 为工作经验. 那么  $\beta_1$  可以粗略地解释为: 在工作经验相同的情况下, 每增加一年教育水平, 条件期望  $\mathbb{E}[\text{wage}|\text{educ}, \text{exp}]$  会提高  $\beta_1$ <sup>6</sup>.

不仅如此, 在线性 CEF 框架下, 解释变量的非线性影响也能被捕捉到. 假设我们有两个解释变量  $X_1$  与  $X_2$ , CEF 为如下的二次型

$$m(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^3 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 \quad (1.13)$$

<sup>6</sup>更确切地说, 当其它解释变量固定不变时, 对于一个很小的  $\Delta x_j$ , 条件期望的变化为  $\Delta \mathbb{E}[Y|x] \approx \frac{\partial m(x)}{\partial x_j} \Delta x_j$ .

显然, 回归元  $[x_1, x_2]$  线性于  $\beta = [\beta_1, \dots, \beta_6]'$ , 因此我们也称 (1.13) 为线性 CEF. 然而, 如果分别对  $x_1$  和  $x_2$  求偏导

$$\begin{aligned}\frac{\partial}{\partial x_1} m(x_1, x_2) &= \beta_1 + 2\beta_3 x_1 + \beta_5 x_2 \\ \frac{\partial}{\partial x_2} m(x_1, x_2) &= \beta_2 + 2\beta_4 x_2 + \beta_5 x_1\end{aligned}$$

此时回归导数不再是简单的回归系数, 而是关于  $(x_1, x_2)$  的函数, 因此很难单独解释回归系数的意义.

我们称 (1.13) 中的回归系数  $\beta_5$  为交互效应. 如果  $\beta_5 > 0$ , 那么关于  $x_1$  的回归导数随着  $x_2$  的增加而增加, 关于  $x_2$  的回归导数也随着  $x_1$  的增加而增加. 类似地, 如果  $\beta_5 < 0$ , 那么关于  $x_1$  的回归导数随着  $x_2$  的增加而减少, 关于  $x_2$  的回归导数同样随着  $x_1$  的增加而减少.

如果回归元包含虚拟变量, 那么 CEF 也能写成是回归元的线性函数, 现在我们考虑将例 1.4 中的工资方程进行拓展.

**例 1.5** 考虑如下回归模型

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \delta_0 \text{female} + e \quad (1.14)$$

其中 wage 为时薪, educ 为受教育程度, female 为性别, 它是一个虚拟变量:

$$\text{female} = \begin{cases} 1, & \text{性别为女} \\ 0, & \text{性别为男} \end{cases}$$

在零条件均值  $\mathbb{E}[e|\text{female}, \text{educ}] = 0$  的假定下, 回归系数  $\delta_0$  为

$$\delta_0 = \mathbb{E}[\text{wage}|\text{female} = 1, \text{educ}] - \mathbb{E}[\text{wage}|\text{female} = 0, \text{educ}]$$

其含义为: 在相同的受教育水平上, 女性与男性的时薪差异. 此外, 我们还可以在工资方程 (1.14) 中加入交互项得到

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \delta_0 \text{female} + \delta_1 \text{female} \cdot \text{educ} + e$$

于是

$$\mathbb{E}[\text{wage}|\text{female} = 0, \text{educ}] = \beta_0 + \beta_1 \text{educ}$$

$$\mathbb{E}[\text{wage}|\text{female} = 1, \text{educ}] = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) \text{educ}$$

此时  $\delta_1$  衡量的是女性与男性教育回报差异.

## 1.6 最优线性预测量

定理 1.6 表明在所有关于  $X$  的可测函数  $g$  中, CEF  $m(X) = \mathbb{E}[Y|X]$  是对  $Y$  最优的预测量, 然而它的形式通常是未知的. 现在我们考虑线性 CEF 的情形, 尽管它无法捕捉到高维信息和包括所有的交互项, 但它毕竟没有复杂到我们无法对其进行分析.

在进行分析前, 先给出一些正则条件.



**假设 1.1**

设  $Y$  为随机变量,  $X$  为  $K$  维随机向量, 那么:

- (1)  $\mathbb{E}[Y^2] < \infty$ .
- (2)  $\mathbb{E}\|X\|^2 < \infty$ , 其中  $\|\cdot\|$  表示范数<sup>a</sup>.
- (3)  $\mathbf{Q}_{XX} = \mathbb{E}[XX']$  为  $K \times K$  维正定矩阵.

<sup>a</sup>范数  $\|x\| = (x'x)^{1/2}$  表示向量  $x$  的 Euclid 模长.



以上假设的第 (1) 点和第 (2) 点表明  $Y$  和  $X$  具有有限的期望, 方差及协方差, 而第 (3) 点是技术性假设, 它也等价于  $\mathbf{Q}_{XX} = \mathbb{E}[XX']$  是可逆矩阵, 表明线性投影系数  $\beta$  是可识别的 (identified)<sup>7</sup>.

**定义 1.3**

给定条件  $X$  时,  $Y$  的最优线性预测量为

$$\mathcal{P}[Y|X] = X'\beta$$

其中  $\beta$  使得均方误差

$$S(\beta) = \mathbb{E}[(Y - X'\beta)^2] \quad (1.15)$$

最小化, 也即

$$\beta = \arg \min_{b \in \mathbb{R}^K} S(b) \quad (1.16)$$

称  $\beta$  为线性投影系数, 也称最优最小二乘近似系数.



下面将给出关于线性投影模型的显式解和其它性质, 但如果 CEF  $m(X)$  不是线性的, 那么我们通常无法找出  $\beta$  的显式解.

**定理 1.8**

在假设 1.1 下, 以下结论成立:

- (1) 矩  $\mathbb{E}[XX']$  和  $\mathbb{E}[XY]$  存在且它们的元素均有限.
- (2) 线性投影系数 (1.16) 存在且唯一, 等于

$$\beta = \mathbb{E}[XX']^{-1} \mathbb{E}[XY]$$

- (3) 投影误差  $e = Y - X'\beta$  存在, 并且  $\mathbb{E}[e^2] < \infty$ ,  $\mathbb{E}[Xe] = 0$ . 如果  $X$  包含常数项, 那么  $\mathbb{E}[e] = 0$ .

- (4) 若  $r \geq 2$ ,  $\mathbb{E}|Y|^r < \infty$  且  $\mathbb{E}\|X\|^r < \infty$ , 那么  $\mathbb{E}|e|^r < \infty$ .



**证明** (1) 根据假设 1.1(1), (2), 以及期望不等式和范数的性质可知

$$|\mathbb{E}[XX']| \leq \mathbb{E}\|XX'\| = \mathbb{E}\|X\|^2 < \infty$$

<sup>7</sup>一个参数是可识别的, 如果它能被可观测变量的概率分布唯一确定.

再由 Cauchy-Schwarz 不等式得到

$$|\mathbb{E}[XY]| \leq \mathbb{E}[XY] \leq (\mathbb{E}[X]^2)^{1/2}(\mathbb{E}[Y^2])^{1/2} < \infty$$

这表明  $\mathbb{E}[XX']$  和  $\mathbb{E}[XY]$  都是良定义的, 因此假设 1.1(3) 有意义.

(2) 将均方误差 (1.15) 展开, 它可以表示为关于  $\beta$  的二次型:

$$S(\beta) = \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$$

根据矩阵微分求得一阶条件 (First Order Condition, FOC):

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

也即

$$\mathbf{Q}_{XY} = \mathbf{Q}_{XX}\beta \quad (1.17)$$

其中  $\mathbf{Q}_{XY} = \mathbb{E}[XY]$  为  $K \times 1$  向量. 由于  $\mathbf{Q}_{XX}$  是可逆的, 于是

$$\beta = \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} = \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \quad (1.18)$$

最后由二阶条件

$$\frac{\partial}{\partial \beta \partial \beta'} S(\beta) = 2XX'$$

可知 (1.18) 为全局最小化最优解.

(3) 将  $e = Y - X'\beta$  代入到  $\mathbb{E}[e^2]$  中得到

$$\begin{aligned} \mathbb{E}[e^2] &= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\ &= \mathbb{E}[Y^2] - \mathbb{E}[YX'](\mathbb{E}[XX'])^{-1}\mathbb{E}[XY] \\ &\leq \mathbb{E}[Y^2] < \infty \end{aligned}$$

另一方面, 根据期望不等式和 Cauchy-Schwarz 不等式可知

$$|\mathbb{E}[Xe]| \leq \mathbb{E}[Xe] \leq (\mathbb{E}[X]^2)^{1/2}(\mathbb{E}[e^2])^{1/2} < \infty$$

于是  $\mathbb{E}[Xe]$  存在, 它等于

$$\begin{aligned} \mathbb{E}[Xe] &= \mathbb{E}[XY] - \mathbb{E}[XX']\beta \\ &= \mathbb{E}[XY] - \mathbb{E}[XX'](\mathbb{E}[XX'])^{-1}\mathbb{E}[XY] = 0 \end{aligned}$$

将  $X = [X_1, X_2, \dots, X_k]$  中的第一项设置为 1, 根据  $\mathbb{E}[Xe] = 0$  即可知  $\mathbb{E}[e] = 0$ .

(4) 根据 Minkowski 不等式可知

$$\begin{aligned} (\mathbb{E}[e]^r)^{1/r} &= (\mathbb{E}[Y - X'\beta]^r)^{1/r} \\ &\leq (\mathbb{E}[Y]^r)^{1/r} + (\mathbb{E}[X'\beta]^r)^{1/r} \\ &\leq (\mathbb{E}[Y]^r)^{1/r} + (\mathbb{E}[X]^r)^{1/r}\|\beta\| < \infty \end{aligned}$$

**注** 在 1.2 节中, 我们根据 CEF 的定义  $\mathbb{E}[Y|X] = m(X)$  设置了 CEF 误差  $e = Y - \mathbb{E}[Y|X]$ , 并证明了  $\mathbb{E}[e] = 0$ . 而在这里, 我们仅需要假设 1.1 就可以定义线性回归模型

$$Y = X'\beta + e$$

并且投影误差  $e$  满足  $\mathbb{E}[Xe] = 0$ , 而无需用到线性 CEF  $m(X) = X'\beta$  这一更强的假设.

事实上, 根据定理 1.4(4) 可知 CEF 误差满足  $\mathbb{E}[Xe] = 0$ , 从而线性 CEF 是最优线性投影, 然而投影误差并不意味着  $\mathbb{E}[e|X] = 0$ , 因此最优线性投影不一定是线性 CEF. 倘若 CEF 不是线性的, 那我们用线性投影来近似 CEF 将会产生很大误差.

举例而言, 我们假设数据生成过程 (Data Generating Process, DGP) 为  $Y = X + X^2$ , 其中  $X \sim N(0, 1)$ , 此时 CEF  $m(x) = x + x^2$  并不会在预测  $Y$  时产生误差. 现在考虑  $Y$  在  $X$  和常数项上的线性投影  $Y = X'\beta + e$ , 因为  $X \sim N(0, 1)$ , 故而  $X$  与  $X^2 \sim \chi^2(1)$  不相关, 根据定理 1.8(2) 可知线性投影系数为

$$\beta = \begin{bmatrix} \mathbb{E}[X^2] & \mathbb{E}[X^3] \\ \mathbb{E}[X^3] & \mathbb{E}[X^4] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[XY] \\ \mathbb{E}[X^2Y] \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

从而线性投影为  $\mathcal{P}[Y|X] = 1 + X$ , 投影误差为  $e = X^2 - 1$ , 显然它不是 CEF 误差.

在 CEF 模型中, 我们定义了误差方差  $\sigma^2 = \mathbb{E}[e^2]$ , 利用  $e = Y - X'\beta$  可以将其改写为

$$\begin{aligned} \sigma^2 &= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\ &= Q_{YY} - 2Q_{YX}Q_{XX}^{-1}Q_{XY} + Q_{YX}Q_{XX}^{-1}Q_{XX}Q_{XX}^{-1}Q_{XY} \\ &= Q_{YY} - Q_{YX}Q_{XX}^{-1}Q_{XY} \equiv Q_{YY \cdot X} \end{aligned}$$

它揭示了  $Q_{YY \cdot X} = Q_{YY} - Q_{YX}Q_{XX}^{-1}Q_{XY}$  等于  $Y$  在  $X$  上的线性投影误差的方差.

有时候我们会将回归元中的常数项分离出来, 而将线性模型写为以下形式

$$Y = X'\beta + \alpha + e \quad (1.19)$$

其中  $\alpha$  为截距项,  $X$  不包括任何常量. 对上式两端取期望得

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

由于  $\mathbb{E}[e] = 0$ , 故而  $\mu_Y = \mu'_X\beta + \alpha$ , 其中  $\mu_Y = \mathbb{E}[Y]$  而  $\mu_X = \mathbb{E}[X]$ . 将  $\alpha = \mu_Y - \mu'_X\beta$  代入到 (1.19) 可得

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

如果假设 1.1 成立, 那么根据定理 1.8 可知线性投影系数为

$$\beta = [\text{var}(X)]^{-1}\text{cov}(X, Y)$$

其中协方差矩阵  $\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])']$ , 以及  $\text{var}(X) = \text{cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']$ .

更多关于线性投影的内容可以参考 Wooldridge (2010) 的第二章.

## 1.7 分块回归与遗漏变量偏误

现在将回归元  $X$  分割为

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (1.20)$$

再将回归系数分割为  $\beta = [\beta'_1, \beta'_2]'$ , 于是  $Y$  对  $X$  的线性投影为

$$Y = X'_1\beta_1 + X'_2\beta_2 + e \quad (1.21)$$

$$\mathbb{E}[Xe] = 0$$

现在问题是如何推导出子向量  $\beta_1$  和  $\beta_2$ .

首先分割  $\mathbf{Q}_{XX}$  为

$$\mathbf{Q}_{XX} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1X'_1] & \mathbb{E}[X_1X'_2] \\ \mathbb{E}[X_2X'_1] & \mathbb{E}[X_2X'_2] \end{bmatrix}$$

类似可以分割  $\mathbf{Q}_{XY}$  为

$$\mathbf{Q}_{XY} = \begin{bmatrix} \mathbf{Q}_{1Y} \\ \mathbf{Q}_{2Y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1Y] \\ \mathbb{E}[X_2Y] \end{bmatrix}$$

根据分块矩阵求逆公式得

$$\begin{aligned} \mathbf{Q}_{XX}^{-1} &= \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}^{-1} \equiv \begin{bmatrix} \mathbf{Q}^{11} & \mathbf{Q}^{12} \\ \mathbf{Q}^{21} & \mathbf{Q}^{22} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_{11.2}^{-1} & -\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22.1}^{-1}\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22.1}^{-1} \end{bmatrix} \end{aligned} \quad (1.22)$$

其中  $\mathbf{Q}_{11.2} \equiv \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$ ,  $\mathbf{Q}_{22.1} \equiv \mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}$ . 因此

$$\begin{aligned} \beta &= \begin{bmatrix} \mathbf{Q}_{11.2}^{-1} & -\mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22.1}^{-1}\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22.1}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{1Y} \\ \mathbf{Q}_{2Y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_{11.2}^{-1}(\mathbf{Q}_{1Y} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{2Y}) \\ \mathbf{Q}_{22.1}^{-1}(\mathbf{Q}_{2Y} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{1Y}) \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{1Y.2} \\ \mathbf{Q}_{22.1}^{-1}\mathbf{Q}_{2Y.1} \end{bmatrix} \end{aligned} \quad (1.23)$$

因此  $\beta_1 = \mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{1Y.2}$ ,  $\beta_2 = \mathbf{Q}_{22.1}^{-1}\mathbf{Q}_{2Y.1}$ .

特别地, 如果  $\dim(X_1) = 1$ , 也即  $\beta_1 \in \mathbb{R}$ , 此时可以将 (1.21) 改写为

$$Y = X_1\beta_1 + X'_2\beta_2 + e \quad (1.24)$$

再来考虑  $X_1$  对  $X_2$  的线性投影

$$X_1 = X'_2\gamma_2 + u_1$$

$$\mathbb{E}[X_2u_1] = 0$$

于是  $\gamma_2 = \mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$ ,  $\mathbb{E}[u_1^2] = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$ , 以及

$$\begin{aligned} \mathbb{E}[u_1Y] &= \mathbb{E}[(X_1 - \gamma'_2X_2)Y] = \mathbb{E}[X_1Y] - \gamma'_2\mathbb{E}[X_2Y] \\ &= \mathbf{Q}_{1Y} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{2Y} = \mathbf{Q}_{1Y.2} \end{aligned}$$

因此

$$\beta_1 = \mathbf{Q}_{11.2}^{-1}\mathbf{Q}_{1Y.2} = \frac{\mathbb{E}[u_1Y]}{\mathbb{E}[u_1^2]}$$

也即  $\beta_1$  为  $Y$  对  $u_1$  回归得到的回归系数.

注意到  $u_1$  是  $X_1$  对  $X_2$  投影得到的误差, 我们可以认为它是  $X_1$  的变化中无法被  $X_2$  线性

解释的部分, 因此  $Y$  对  $u$  回归就是想用这些部分来线性解释  $Y$  的变化, 于是  $\beta_1$  可以很自然地理解为过滤掉其它回归元对  $X_1$  和  $Y$  的影响后,  $X_1$  对  $Y$  的线性影响.

同样考虑回归元被分割为 (1.20) 那样的形式, 但假设  $X_2$  无法被观测到, 因此我们只考虑  $Y$  在  $X_1$  上的投影:

$$\begin{aligned} Y &= X_1' \gamma_1 + u \\ \mathbb{E}[X_1 u] &= 0 \end{aligned} \tag{1.25}$$

计算线性投影系数  $\gamma_1$  得到

$$\begin{aligned} \gamma_1 &= (\mathbb{E}[X_1 X_1'])^{-1} \mathbb{E}[X_1 Y] \\ &= (\mathbb{E}[X_1 X_1'])^{-1} \mathbb{E}[X_1 (X_1' \beta_1 + X_2' \beta_2 + e)] \\ &= \beta_1 + (\mathbb{E}[X_1 X_1'])^{-1} \mathbb{E}[X_1 X_2'] \beta_2 \equiv \beta_1 + \Gamma_{12} \beta_2 \end{aligned}$$

其中  $\Gamma_{12} = \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}$  是将  $X_2$  投影到  $X_1$  上得到的系数矩阵.

显然  $\gamma_1 = \beta_1 + \Gamma_{12} \beta_2 \neq \beta_1$ , 除非  $\beta_2 = 0$  或  $\Gamma_{12} = 0$ ,  $\gamma_1$  和  $\beta_1$  之间的差  $\Gamma_{12} \beta_2$  称为遗漏变量偏误 (Omitted Variable Bias), 此时模型设定错误, 导致内生性 (endogeneity) 问题. 显然, 如果系数  $\beta_2 > 0$ , 若  $X_2$  对  $X_1$  有正 (负) 向影响, 我们得到的回归系数  $\gamma$  将会高 (低) 估  $\beta$ . 类似可以分析  $\beta_2 < 0$  时的情况.

为了避免遗漏变量偏误, 最好将所有潜在的相关变量纳入到回归模型中, 但显然这极不可能, 我们只能尽可能将合适的变量纳入到回归模型中以减轻遗漏变量偏误的影响.



## 第2章 经典线性回归模型

本章主要介绍有限样本下的经典线性回归模型, 也即 OLS 回归模型, 它是构筑现代计量经济学的基石.

### 2.1 CLRM 假设

设  $\{Y_i, X_i'\}_{i=1}^n$  是一个样本容量为  $n$  的随机样本,  $Y_i$  为一个标量,  $X_i = [1, X_{1i}, \dots, X_{ki}]'$  为  $K \times 1$  维列向量. 我们的目的是通过随机样本  $\{Y_i, X_i'\}_{i=1}^n$  对条件期望  $\mathbb{E}[Y|X]$  建模, 估计未知参数并进行统计推断. 为此我们先给出经典线性回归模型 (Classical Linear Regression Model, CLRM) 假设.

#### 假设 2.1 (线性)

$\{Y_i, X_i'\}_{i=1}^n$  是一个可观测的随机样本, 并且

$$Y_i = X_i' \beta + e_i, \quad i = 1, 2, \dots, n$$

其中  $\beta$  为  $K \times 1$  维未知参数向量,  $e_i$  是不可观测的随机扰动项.

当线性 CEF 是对真实情况的正确设定时, 也即  $\mathbb{E}[e_i|X_i] = 0$  时, 参数  $\beta$  可以解释为  $X_i$  对  $Y_i$  的期望边际效应.

当然, 假设 2.1 并不意味着  $X_i$  和  $Y_i$  存在因果关系, 它描述的仍然是上一章提到的线性预测关系. 无论统计关系多么强和富有启示性, 都无法确立因果关系, 它必须来源于统计学之外的经济理论.

现在我们用定义以下记号

$$Y = [Y_1, Y_2, \dots, Y_n]', \quad n \times 1$$

$$e = [e_1, e_2, \dots, e_n]', \quad n \times 1$$

$$X = [X_1, X_2, \dots, X_n]', \quad n \times K$$

这里  $X$  的第  $i$  行是  $K$  维列向量  $X_i' = [1, X_{1i}, \dots, X_{ki}]$ . 通过以上符号, 我们可以将假设 2.1 简洁地表示为

$$Y = X\beta + e$$

#### 假设 2.2 (严格外生性)

对于一切  $i = 1, 2, \dots, n$  都有:

$$\mathbb{E}[e_i|X] = 0$$

我们可以将该假设简写为  $\mathbb{E}[e|X] = 0$ , 它意味着线性 CEF 是对真实情况的正确设定. 根

据 LIE 可知  $\mathbb{E}[e_i|X_i] = 0$ , 以及  $\mathbb{E}[e_i] = 0$ . 不仅如此, 对于任意  $1 \leq i, j \leq n$ , 都有

$$\mathbb{E}[X_j e_i] = \mathbb{E}[\mathbb{E}[X_j e_i|X]] = \mathbb{E}[X_j \mathbb{E}[e_i|X]] = 0$$

由于  $\mathbb{E}[e_i] = 0$ , 因此假设 2.2 意味着每个解释变量和随机扰动项不相关. 显然, 这一假设排除了动态回归模型, 也即  $X_i$  中包含  $Y_i$  的滞后项的情形.

特别地, 如果  $\{Y_i, X_i\}_{i=1}^n$  是 i.i.d. 随机样本, 那么假设 2.2 等价于  $\mathbb{E}[e_i|X_i] = 0$ .

### 假设 2.3 (非奇异性)

$K \times K$  维方阵  $X'X = \sum_{i=1}^n X_i X_i'$  是正定的.

以上假设中的 (1) 排除了  $K$  个解释变量存在完全多重共线性 (perfect multicollinearity) 的可能, 也即排除了至少存在某个解释变量可以表示为其它  $K - 1$  个解释变量的线性组合的情形.  $X'X$  的非奇异性意味着  $X$  必须是满秩矩阵, 因此解释变量个数  $K$  不能超过样本容量  $n$ .

### 假设 2.4 (球型扰动项)

(1) 同方差性:  $\mathbb{E}[e_i^2|X] = \sigma^2, i = 1, 2, \dots, n$ .

(2) 无自相关:  $\mathbb{E}[e_i e_j|X] = 0, 1 \leq i, j \leq n, i \neq j$ .

我们可以将以上假设简洁地表示为

$$\mathbb{E}[ee'|X] = \sigma^2 I_n$$

其中  $I_n$  是  $n \times n$  维单位矩阵. 鉴于球型扰动项假设难以成立, 我们使用符号  $\Sigma = \mathbb{E}[ee'|X]$  来表示一般情况下的协方差矩阵.

根据 LIE, 假设 2.4 表明, 对于任意  $1 \leq i \leq n$  都有  $\text{var}(e_i^2) = \sigma^2$ , 并且对一切  $i \neq j$  都有  $\mathbb{E}[e_i e_j] = 0$ .

**注** 假设 2.2 和 2.4 无法推出  $e_i$  和  $X$  相互独立, 因为  $e_i$  的条件高阶矩可能依赖于  $X$ .

## 2.2 普通最小二乘估计

下面将介绍普通最小二乘 (Ordinary Least Squares, OLS) 估计方法, 这也是计量经济学最基本的估计方法.

### 定义 2.1

定义线性回归模型  $Y_i = X_i' \beta + e_i$  的残差平方和 (Sum of Squared Residuals, SSR):

$$\text{SSR}(\beta) \equiv (Y - X\beta)'(Y - X\beta)$$

则 OLS 估计量  $\hat{\beta}$  是以下最优化问题的解

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^K} \text{SSR}(\beta)$$

**定理 2.1 (OLS 估计量的存在性)**

在假设 2.1 和 2.3 下, OLS 估计量  $\hat{\beta}$  存在, 且等于

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2.1) \quad \heartsuit$$

**证明** 先将残差平方和展开得到

$$\begin{aligned} SSR(\beta) &= (Y - X'\beta)'(Y - X'\beta) \\ &= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2Y'X\beta + \beta'X'X\beta \end{aligned}$$

其中  $\beta'X'Y = Y'X\beta$  为标量. 进而根据矩阵微分可知 FOC 为

$$\frac{\partial SSR(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta = 0$$

因此 OLS 估计量满足

$$X'Y = X'X\hat{\beta}$$

由于  $X'X$  是可逆矩阵, 于是

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

最后考虑二阶条件

$$\frac{\partial^2 SSR(\beta)}{\partial \beta \partial \beta'} = 2X'X$$

根据  $X'X$  的正定性可知  $\hat{\beta}$  是全局最小化最优解.

注意到

$$\sum_{i=1}^n X_i X_i' = X'X, \quad \sum_{i=1}^n X_i Y_i = X'Y$$

因此 OLS 估计量  $\hat{\beta}$  除了表示为 (2.1), 还可以将其写为

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

这对于分析 OLS 估计量的渐近性质十分有用. 特别地, 对于一元线性回归  $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$ , 其 OLS 估计量为

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

其中  $\bar{Y}$  和  $\bar{X}$  分别是  $Y$  和  $X$  的样本均值.

我们称  $\hat{Y}_i = X_i' \hat{\beta}$  为观测值  $Y_i$  的拟合值或预测值, 而  $\hat{e}_i = Y_i - \hat{Y}_i$  称为  $Y_i$  的 OLS 估计残差, 注意到

$$\hat{e}_i = Y_i - \hat{Y}_i = e_i - X_i'(\hat{\beta} - \beta)$$

下一章将会证明: 随着样本容量  $n$  增大,  $\hat{\beta}$  依概率收敛于  $\beta$ , 从而上式的第二项可以小到忽略不

计. 在推导 OLS 估计量时, 根据 FOC 还可知

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

若  $X_i$  包含截距项, 那么

$$\hat{e}_1 + \hat{e}_2 + \cdots + \hat{e}_n = 0$$

注意, 这一性质是由最小化问题  $\min_{\beta \in \mathbb{R}^K} \text{SSR}(\beta)$  的 FOC 得到的, 不论严格外生性假设成立与否, 该正交条件总是成立的.

此外, 我们还可以使用矩估计的方法获得 OLS 估计量. 考虑线性投影模型

$$Y = X'\beta + e$$

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \quad (2.2)$$

$$\mathbb{E}[Xe] = 0$$

既然线性投影系数使得 MSE 最小化, OLS 估计量使得残差平方和最小化, 那么可以在 (2.2) 中用样本矩替代总体矩得到

$$\hat{\beta} = \left(\frac{1}{n}X'X\right)^{-1} \left(\frac{1}{n}X'Y\right) = (X'X)^{-1}X'Y$$

## 2.3 投影矩阵与消灭矩阵

现在我们介绍投影矩阵与消灭矩阵的概念, 它们对于推导统计量的有限样本性质具有重要作用.

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

$$M = I_n - P$$

它们具有以下良好的代数性质.

### 定理 2.2

对于投影矩阵  $P$  和消灭矩阵  $M$ , 以下命题成立:

- (1)  $PX = X, MX = 0$ .
- (2)  $P\hat{e} = 0, Me = MY = \hat{e}$ .
- (3)  $P^2 = P, M^2 = M$ .
- (4)  $P' = P, M' = M$ .
- (5)  $\text{trace}(P) = K, \text{trace}(M) = n - K$ .
- (6)  $\text{SSR}(\beta) = e'Me$ .



**证明** (1)  $PX = X(X'X)^{-1}X'X = X, MX = X - X = 0$ .

(2) 由  $X'\hat{e} = 0$  即可推知  $P\hat{e} = 0$ , 注意到  $\hat{Y} = X\hat{\beta} = PY$ , 于是

$$\begin{aligned}\hat{e} &= Y - \hat{Y} = (I_n - P)Y = MY \\ &= M(X\beta + e) = MX\beta + Me = Me\end{aligned}$$

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

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

(4) 结论是显然的.

(5) 根据矩阵的迹的性质可知  $\text{trace}(P) = \text{trace}[X(X'X)^{-1}X'] = \text{trace}[(X'X)^{-1}X'X] = \text{trace}(I_K) = K$ , 自然有  $\text{trace}(M) = n - K$ .

(6) 根据  $\hat{e} = Me$  可知

$$\text{SSR}(\beta) = \hat{e}'\hat{e} = (Me)'(Me) = e'M'Me = e'Me$$

由此证得定理.

特别地, 我们可以取  $X = \mathbf{1}_n$ , 它是  $n \times 1$  维元素全为 1 的列向量, 此时  $M = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ , 它可以简化某些求和计算.

## 2.4 杠杆值与留一回归

有时候单个异常值可能会对整个线性回归产生显著影响, 为了克服这种不利影响, 我们通常会使用留一 (Leave-One-Out, LOO) 回归, 为此先介绍杠杆值 (leverage value) 的概念.

首先定义回归矩阵  $X$  的第  $i$  个杠杆值为投影矩阵  $P = X(X'X)^{-1}X'$  的第  $i$  个对角线元素, 也即

$$h_{ii} = X_i'(X'X)^{-1}X_i', \quad 1 \leq i \leq n$$

它是回归向量  $X_i$  的归一化长度, 它衡量了第  $i$  个观测值  $X_i$  相对于其它观测值的差异程度,  $h_{ii}$  越大说明  $X_i$  相对于其它样本值越不同. 我们定义

$$\bar{h} = \max_{1 \leq i \leq n} h_{ii} \quad (2.3)$$

它是所有  $h_{ii}$  中的最大值, 衡量了总体中回归元的变异程度.

现在我们给出有关杠杆值的定理.

### 定理 2.3

对于杠杆值  $h_{ii}$ , 以下命题成立:

- (1)  $0 \leq h_{ii} \leq 1$ .
- (2) 如果  $X_i$  包括截距项, 那么  $h_{ii} \geq 1/n$ .
- (3)  $\sum_{i=1}^n h_{ii} = K$ .



**证明** (1) 显然  $h_{ii} \geq 0$ . 定义  $n \times 1$  维列向量  $s_i$  为

$$s_i = [0, \dots, 1, \dots, 0]'$$



它的第  $i$  个元素为 1, 其它元素为 0. 现在将  $h_{ii}$  改写为  $h_{ii} = s_i' P s_i$ , 利用二次不等式<sup>1</sup>可知

$$h_{ii} = s_i' P s_i \leq s_i' s_i \lambda_{\max}(P) = 1$$

(2) 由于回归元  $X_i$  包含截距项, 故而可以将  $X_i$  分割为  $X_i = [1, Z_i']'$ , 不失一般性, 我们可以用  $Z_i^* = Z_i - \bar{Z}$  替代  $Z_i$ , 于是

$$\begin{aligned} h_{ii} &= \begin{bmatrix} 1 & Z_i^* \end{bmatrix} \begin{bmatrix} n & 0 \\ 0 & Z^{*'} Z^* \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ Z_i^* \end{bmatrix} \\ &= \frac{1}{n} + Z_i^{*'} (Z^{*'} Z^*)^{-1} Z_i^* \geq \frac{1}{n} \end{aligned}$$

(3) 根据杠杆值的定义, 这是显然的.

进一步, 我们用  $\bar{h} = \max_{1 \leq i \leq n} h_{ii}$ , 也即  $h_{ii}$  的最大值来衡量总体的变异程度, 如果一个回归设计是平衡的, 那么每一个  $h_{ii}$  大致应该相等, 在完全平衡的情况下, 根据定理 2.3(3) 可知  $h_{ii} = \bar{h} = k/n$ .

反之, 若回归设计是不平衡的, 则不同的  $h_{ii}$  之间具有很强的差异性, 最极端的情况就是  $\bar{h} = 1$ : 举例而言, 回归元中包括虚拟变量, 并且该虚拟变量仅对唯一的一个观测值取 1, 而对其它观测值都取 0.

现在来看 LOO 回归, 它是利用  $Y$  对缺失第  $i$  个观测值的原样本进行的回归, 此时 OLS 估计量为

$$\begin{aligned} \hat{\beta}_{-i} &= \left( \sum_{j \neq i} X_j X_j' \right)^{-1} \left( \sum_{j \neq i} X_j Y_j \right) \\ &= (X'X - X_i X_i')^{-1} (X'Y - X_i Y_i) \\ &= (X_{-i}' X_{-i})^{-1} X_{-i}' Y_{-i} \end{aligned}$$

定义留一回归的预测值为  $\tilde{Y}_i = X_i' \hat{\beta}_{-i}$ , 预测误差为  $\tilde{e}_i = Y_i - \tilde{Y}_i$ , 有时它可以替代残差  $\hat{e}_i$  来作为  $e_i$  的估计量, 并且具有更好的性质. 由于  $\hat{\beta}_{-i}$  以及  $\tilde{e}_i$  的计算较为复杂, 下面给出一个简化其计算的定理.

#### 定理 2.4

LOO 估计量和预测误差分别等价于

$$\hat{\beta}_{-i} = \hat{\beta} - (X'X)^{-1} X_i \tilde{e}_i \quad (2.4)$$

以及

$$\tilde{e}_i = (1 - h_{ii})^{-1} \hat{e}_i \quad (2.5)$$

其中  $h_{ii}$  为  $X$  的第  $i$  个杠杆值.



**证明** 将 LOO 估计量写为

$$(X_{-i}' X_{-i})^{-1} X_{-i}' Y_{-i}$$

<sup>1</sup>对任意  $n \times 1$  维列向量  $b$  和  $n \times n$  维对称矩阵  $A$ ,  $b' A b \leq \|A\| b' b$  成立, 这里的  $\|A\| = [\lambda_{\max}(A' A)]^{1/2}$ , 表示  $A$  的谱范数.

上式左乘  $(X'X)^{-1}(X'X - X_iX_i')$  得到

$$\hat{\beta}_{-i} - (X'X)^{-1}X_iX_i'\hat{\beta}_{-i} = (X'X)^{-1}(X'Y - X_iY_i) = \hat{\beta} - (X'X)^{-1}X_iY_i$$

也即

$$\hat{\beta}_{-i} = \hat{\beta} - (X'X)^{-1}X_i(Y_i - X_i'\hat{\beta}_{-i}) = \hat{\beta} - (X'X)^{-1}X_i\tilde{e}_i$$

也即 (2.4) 成立, 在 (2.4) 中左乘  $X_i'$  得到

$$X_i'\hat{\beta}_{-i} = X_i'\hat{\beta} - X_i'(X'X)^{-1}X_i\tilde{e}_i = X_i'\hat{\beta} - h_{ii}\tilde{e}_i$$

因此  $\tilde{e}_i = \hat{e}_i + h_{ii}\tilde{e}_i$ , 也即 (2.5) 成立.

## 2.5 拟合优度

现在来分析线性回归模型对数据的拟合程度究竟是好是坏, 也即它对  $\{Y_i\}$  变动的预测能力如何, 为此我们需要设置出一些指标.

### 定义 2.2 (非中心化 $R^2$ )

$$R_{uc}^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{\hat{e}'\hat{e}}{Y'Y}$$

它的含义是  $\{Y_i\}$  的非中心化样本二次型变动可以被预测值  $\{\hat{Y}_i\}$  的非中心化样本二次型变动所预测的比例, 根据定义可知始终有  $0 \leq R_{uc}^2 \leq 1$ .

### 定义 2.3 (中心化 $R^2$ )

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

其中  $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$  是样本均值,  $R^2$  又称可决系数.

倘若我们将投影矩阵  $P = X(X'X)^{-1}X'$  中的  $X$  设置为  $\mathbf{1}_n$ , 于是

$$P = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n' = n^{-1}\mathbf{1}_n'\mathbf{1}_n$$

$$M = I_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$$

从而

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{Y}'M\hat{Y}, \quad \sum_{i=1}^n (Y_i - \bar{Y})^2 = Y'MY$$

下面我们将给出关于可决系数  $R^2$  的一些定理.

### 定理 2.5

如果  $X_i$  中包括截距项  $X_{1i} = 1$ , 那么

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

**证明** 将  $Y_i - \bar{Y}$  进行分解, 于是

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{e}_i^2 + 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y}) \hat{e}_i \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{e}_i^2\end{aligned}$$

其中

$$\begin{aligned}\sum_{i=1}^n (\hat{Y}_i - \bar{Y}) \hat{e}_i &= \hat{\beta}' \sum_{i=1}^n X_i \hat{e}_i - \bar{Y} \sum_{i=1}^n \hat{e}_i \\ &= \hat{\beta}' (X' \hat{e}) - \bar{Y} \sum_{i=1}^n \hat{e}_i = 0\end{aligned}$$

由此证得定理.

由定理2.5可知

$$R^2 = \frac{\hat{Y}' M_0 \hat{Y}}{Y' M_0 Y}$$

由此可知  $0 \leq R^2 \leq 1$  成立, 其中  $M_0 = I_n - \mathbf{1}_n (\mathbf{1}_n' \mathbf{1}_n)^{-1} \mathbf{1}_n'$ . 如果  $X_i$  不包括截距项, 那么

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \neq \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{e}_i^2$$

在这种情况下,  $R^2$  可能为负值.

当  $X_i$  包含截距项时, 中心化  $R^2$  和非中心化  $R_{uc}^2$  有相似的解释, 也即  $R^2$  测度  $\{Y_i\}_{i=1}^n$  的样本方差中可被线性拟合值  $X_i' \hat{\beta}$  所预测的比例.

### 定理 2.6

设  $\hat{\rho}_{Y\hat{Y}}$  为  $Y_i$  和  $\hat{Y}_i$  间的相关系数, 则  $R^2 = \hat{\rho}_{Y\hat{Y}}^2$ .



**证明** 要证  $R^2 = \hat{\rho}_{Y\hat{Y}}^2$ , 只需证明

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

也即证明

$$\hat{Y}' M_0 \hat{Y} = \hat{Y}' M_0 Y$$

根据  $Y = \hat{Y} + \hat{e}$  以及  $\hat{Y}' \hat{e} = 0$  可知

$$\begin{aligned}\text{RHS} &= \hat{Y}' M_0 Y \\ &= \hat{Y}' M_0 (\hat{Y} + \hat{e}) \\ &= \hat{Y}' M_0 \hat{Y} + \hat{Y}' \hat{e} = \text{LHS}\end{aligned}$$

由此证得定理.

**定理 2.7**

随着解释变量个数的增加,  $R^2$  是不减的.



**证明** 考虑以下两个 OLS 回归模型

$$Y = X_1\tilde{\beta} + \tilde{e}$$

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$$

我们只需证明  $\hat{e}'\hat{e} \leq \tilde{e}'\tilde{e}$  即可. 注意到

$$X_1'\tilde{e} = X_1'\hat{e} = X_2'\hat{e} = 0$$

于是

$$\tilde{e}'\hat{e} = (Y - X_1\tilde{\beta})'\hat{e} = (X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e})'\hat{e} = \hat{e}'\hat{e}$$

因此

$$\hat{e}'\hat{e} - \tilde{e}'\tilde{e} = -(\tilde{e} - \hat{e})'(\tilde{e} - \hat{e}) \leq 0$$

由此证得定理.

这表明, 随着解释变量个数增加, 拟合优度不会变差, 因此若我们添加了一些无关变量, 它对因变量没有真正的预测能力, 而  $R^2$  不会对这一行为有任何惩罚. 另一方面, 随着解释变量个数增加, 模型的待估参数越多, 其中可能包含一些数据中不太可能再出现的因素, 因此模型对样本外的  $Y_i$  预测能力越差.

现在我们定义调整  $R^2$  为

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{(n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

它可以对模型中的冗余变量进行惩罚, 也即添加冗余变量会使得调整  $R^2$  下降. 可以证明

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2)$$

因此即使  $X_i$  中包括截距项, 调整  $R^2$  也可能为负值.

最后, 我们应该明确无论是  $R^2$  还是调整  $R^2$ , 它反映的只是统计上的相关性, 而与因果关系无关. 它只衡量模型拟合程度的好坏, 但是无法衡量回归元对响应变量的解释能力的好坏. 事实上, 对于截面数据和面板数据而言,  $R^2$  很少有超过 0.2 的.

## 2.6 OLS 的有限样本性质

**定理 2.8**

在假设 2.1–2.4 成立的条件下, 那么 OLS 估计量满足:

- (1) 无偏性:  $\mathbb{E}[\hat{\beta}|X] = \beta$ , 以及  $\mathbb{E}[\hat{\beta}] = \beta$ .
- (2) 方差:  $\text{var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$ .
- (3) 无相关性:  $\text{cov}(\hat{\beta}, \hat{e}|X) = 0$ .



**证明** (1) 根据  $\hat{\beta} = (X'X)^{-1}X'Y$  可知  $\hat{\beta} - \beta = (X'X)^{-1}X'e$ , 于是

$$\begin{aligned}\mathbb{E}[\hat{\beta} - \beta|X] &= \mathbb{E}[(X'X)^{-1}X'e|X] \\ &= (X'X)^{-1}X'\mathbb{E}[e|X] = 0\end{aligned}$$

也即  $\mathbb{E}[\hat{\beta}|X] = \beta$ , 根据 LIE 可知无条件期望  $\mathbb{E}[\hat{\beta}] = \beta$ .

(2) 由于  $\beta$  是真实参数, 故而  $\text{var}(\hat{\beta}|X) = \text{var}(\hat{\beta} - \beta|X)$ , 于是

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

在假设 2.4 成立的条件时有  $\Sigma = \sigma^2 I_n$ , 此时

$$\begin{aligned}\text{var}(\hat{\beta}|X) &= \sigma^2(X'X)^{-1}X'I_nX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

(3) 首先有  $\mathbb{E}[\hat{e}|X] = M\mathbb{E}[e|X] = 0$ , 再根据协方差的定义可知

$$\begin{aligned}\text{cov}(\hat{\beta}, \hat{e}|X) &= \mathbb{E}\left[\{\hat{\beta} - \mathbb{E}[\hat{\beta}|X]\}\{\hat{e} - \mathbb{E}[\hat{e}|X]\}'|X\right] \\ &= \mathbb{E}[(\hat{\beta} - \beta)\hat{e}'|X] \\ &= \mathbb{E}[(X'X)^{-1}X'ee'M|X] \\ &= (X'X)^{-1}X'\mathbb{E}[ee'|X]M = 0\end{aligned}$$

其中  $\mathbb{E}[ee'|X] = \sigma^2 I_n$  和  $M'X = 0$  保证了上式成立.

现在介绍 Gauss-Markov 定理, 它表明在  $\beta$  的所有线性无偏估计量中, OLS 估计量是最优无偏估计量 (Best Linear Unbiased Estimator, BLUE).

### 定理 2.9 (Gauss-Markov 定理)

在假设 2.1–2.4 下, 如果  $\tilde{\beta}$  是  $\beta$  的线性无偏估计量, 那么  $\text{var}(\tilde{\beta}|X) \geq \text{var}(\hat{\beta}|X)$ .



**证明** 设  $\tilde{\beta} = AY$  为  $\beta$  的任意线性估计量, 由无偏性可知

$$\mathbb{E}[\tilde{\beta}|X] = A'X\beta + A'\mathbb{E}[e|X] = A'X\beta = \beta$$

因此必有  $A'X = I_K$ .

另一方面

$$\tilde{\beta} = A'Y = A'(X\beta + e) = \beta + A'e$$

从而  $\tilde{\beta}$  的条件方差为

$$\text{var}(\tilde{\beta}|X) = \mathbb{E}[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'|X] = \sigma^2 A'A$$

根据定理 2.8(2) 和  $A'X = I_K$  可知

$$\begin{aligned}\text{var}(\tilde{\beta}|X) - \text{var}(\hat{\beta}|X) &= \sigma^2[A'A - (X'X)^{-1}] \\ &= \sigma^2 A'[I_K - X(X'X)^{-1}X']A \\ &= \sigma^2(MA)'(MA)\end{aligned}$$

它是一个半正定矩阵, 定理得证.



**注** Hansen (2022b) 提出了一个现代 Gauss-Markov 定理, 去掉了  $\hat{\beta}$  为线性估计量这一限制, 证明 OLS 估计量为 BUE. 实际上, 该定理与经典 Gauss-Markov 定理并无本质差别. Portnoy (2022) 的结论表明, 估计量的无偏性意味着它是线性的, 因此 BUE 和 BLUE 对于 OLS 而言是一回事.

### 定理 2.10

定义残差方差估计量

$$s^2 = \hat{e}'\hat{e}/(n-K) = \frac{1}{n-K} \sum_{i=1}^n \hat{e}_i^2$$

那么在假设 2.1–2.4 下,  $\mathbb{E}[s^2|X] = \sigma^2$ , 其中  $\sigma^2 = \mathbb{E}[e_i^2]$ .



**证明** 根据  $\hat{e}'\hat{e} = e'Me$ , 以及  $\text{trace}(AB) = \text{trace}(BA)$  可知

$$\begin{aligned} \mathbb{E}[\hat{e}'\hat{e}|X] &= \mathbb{E}[e'Me|X] = \mathbb{E}[\text{trace}(e'Me)|X] \\ &= \mathbb{E}[\text{trace}(ee'M)|X] = \text{trace}(\mathbb{E}[ee'|X]M) \\ &= \sigma^2 \text{trace}(M) = \sigma^2(n-K) \end{aligned}$$

因此

$$\mathbb{E}[s^2|X] = \frac{1}{n-K} \mathbb{E}[\hat{e}'\hat{e}|X] = \sigma^2$$

由此证得定理.

## 2.7 OLS 的协方差矩阵估计

为了进行后续章节的统计推断, 我们需要得到条件方差  $\text{var}(\hat{\beta}|X)$  的估计量. 首先根据  $\hat{e} = Me$  与  $M = I_n - X(X'X)^{-1}X'$  得到

$$\text{var}(\hat{e}|X) = \text{var}(Me|X) = M\Sigma M$$

在同方差条件下

$$\text{var}(\hat{e}|X) = \sigma^2 M$$

并且

$$\text{var}(\hat{e}_i|X) = \mathbb{E}[\hat{e}_i^2|X] = (1 - h_{ii})\sigma^2$$

进一步, 我们定义  $M^*$  为对角矩阵, 其第  $i$  个对角元素为  $(1 - h_{ii})^{-1}$ . 再定义标准化残差

$$\bar{e}_i = (1 - h_{ii})^{-1/2} \hat{e}_i$$

用向量可以表示为

$$\bar{e} = [\bar{e}_1, \bar{e}_2, \dots, \bar{e}_n] = M^{*1/2} Me$$

在同方差条件下

$$\text{var}(\bar{e}|X) = \sigma^2 M^{*1/2} M M^{*1/2}$$

$$\text{var}(\bar{e}_i|X) = \sigma^2$$

因此残差方差估计量  $s^2$  可以写为

$$s^2 = \frac{1}{n} \sum_{i=1}^n \bar{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-1} \hat{e}_i^2$$

根据定理2.10, 我们可以得到 OLS 估计量协方差矩阵  $\text{var}(\hat{\beta}|X)$  的估计量

$$\hat{V}_{\text{OLS}} = s^2 (X'X)^{-1}$$

并且它是无偏的, 因为

$$\mathbb{E}[\hat{V}_{\text{OLS}}|X] = \mathbb{E}[s^2|X](X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

倘若同方差假设不成立, 那么  $\Sigma \neq \sigma^2 I_n$ , 而是等于

$$\Sigma = \mathbb{E}[ee'|X] = \text{diag}\{\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2\}$$

并且

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

由于  $\mathbb{E}[e_i^2|X] = \sigma_i^2$ , 因此  $V_{\text{OLS}}$  的理想无偏估计量为

$$\hat{V}_{\text{OLS}}^{\text{ideal}} = (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' e_i^2 \right) (X'X)^{-1}$$

然而  $e_i$  不可观测, 我们无法得到  $\hat{V}_{\text{OLS}}^{\text{ideal}}$ .

自然而然地, 我们考虑使用残差  $\hat{e}_i$  替代  $e_i$ , 由此可以得到  $V_{\text{OLS}}$  的一个估计量

$$\hat{V}_{\text{OLS}}^{\text{HC0}} = (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \hat{e}_i^2 \right) (X'X)^{-1}$$

然而该统计量未经过自由度调整, 一个更好的估计量是

$$\hat{V}_{\text{OLS}}^{\text{HC1}} = \frac{n}{n-K} (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \hat{e}_i^2 \right) (X'X)^{-1}$$

另外, 我们可以用  $\bar{e}_i$  和  $\tilde{e}_i$  分别替换  $\hat{V}_{\text{OLS}}^{\text{ideal}}$  中的  $e_i$ , 由此得到形式类似的估计量  $\hat{V}_{\text{OLS}}^{\text{HC2}}$  和  $\hat{V}_{\text{OLS}}^{\text{HC3}}$ , 也即

$$\hat{V}_{\text{OLS}}^{\text{HC2}} = (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \bar{e}_i^2 \right) (X'X)^{-1}$$

$$\hat{V}_{\text{OLS}}^{\text{HC3}} = (X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \tilde{e}_i^2 \right) (X'X)^{-1}$$

以上估计量均称为异方差一致性稳健协方差矩阵估计量.

## 2.8 分块回归

首先将  $X$  分割为  $X = [X_1, X_2]$ , 再将  $\beta$  分割为  $\beta = [\beta_1', \beta_2']'$ , 于是回归模型可以写作

$$Y = X_1 \beta_1 + X_2 \beta_2 + e \quad (2.6)$$

$\beta$  的 OLS 估计量可以由  $Y$  对  $X = [X_1, X_2]$  的回归获得, 此时

$$Y = X\hat{\beta} + e = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e \quad (2.7)$$

为了得到  $\hat{\beta} = [\hat{\beta}_1', \hat{\beta}_2']'$  的表达式, 考虑对 (1.23) 进行矩估计, 由此得到

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \hat{Q}_{11 \cdot 2}^{-1} \hat{Q}_{1Y \cdot 2} \\ \hat{Q}_{22 \cdot 1}^{-1} \hat{Q}_{2Y \cdot 1} \end{bmatrix}$$

其中

$$\begin{aligned} \hat{Q}_{11 \cdot 2} &= \hat{Q}_{11} - \hat{Q}_{12} \hat{Q}_{22}^{-1} \hat{Q}_{21} \\ &= \frac{1}{n} X_1' X_1 - \frac{1}{n} X_1' X_2 \left( \frac{1}{n} X_2' X_2 \right)^{-1} \frac{1}{n} X_2' X_1 \\ &= \frac{1}{n} X_1' M_2 X_1 \end{aligned}$$

以及  $M_2 = I_n - X_2(X_2' X_2)^{-1} X_2'$ , 同理可得  $\hat{Q}_{1Y \cdot 2} = \frac{1}{n} X_1' M_2 Y$ . 另一方面

$$\begin{aligned} \hat{Q}_{22 \cdot 1} &= \frac{1}{n} X_2' M_1 X_2 \\ \hat{Q}_{2Y \cdot 1} &= \frac{1}{n} X_2' M_1 Y \\ M_1 &= I_n - X_1(X_1' X_1)^{-1} X_1' \end{aligned}$$

因此可得 OLS 估计量

$$\hat{\beta}_1 = (X_1' M_2 X_1)^{-1} X_1' M_2 Y \quad (2.8)$$

$$\hat{\beta}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 Y \quad (2.9)$$

下面我们介绍著名的 FWL 定理, 它由 Frisch, Waugh 和 Lovell 提出, 该定理给出了计算  $\hat{\beta}_1$  和  $\hat{\beta}_2$  的另一种方法. 考虑 OLS 估计量 (2.9), 注意到  $M_1$  为对称幂等矩阵, 于是

$$\begin{aligned} \hat{\beta}_2 &= (X_2' M_1 X_2)^{-1} X_2' M_1 Y \\ &= (X_2' M_1' M_1 X_2)^{-1} X_2' M_1' M_1 Y \\ &= (\tilde{X}_2' \tilde{X}_2)^{-1} \tilde{X}_2' \tilde{e}_1 \end{aligned}$$

其中  $\tilde{X}_2 = M_1 X_2$ , 以及  $\tilde{e}_1 = M_1 Y$ . 根据定理 2.2(2) 可知,  $\tilde{X}_2$  是  $X_2$  对  $X_1$  回归获得的残差,  $\tilde{e}_1$  是  $Y$  对  $X_1$  回归的残差.

### 定理 2.11 (FWL 定理)

在回归模型 (2.6) 中,  $\beta_2$  的 OLS 估计量可由以下步骤获得:

STEP 1: 使用  $X_2$  对  $X_1$  回归, 获得残差  $\tilde{e}_1$ ;

STEP 2: 使用  $Y$  对  $X_1$  回归, 获得残差  $\tilde{X}_2$ ;

STEP 3: 使用  $\tilde{X}_2$  对  $\tilde{e}_1$  回归, 获得 OLS 估计量  $\hat{\beta}_2$  及残差  $\hat{e}$ .



**注** 类似地,  $X_1$  的回归系数  $\hat{\beta}_1$  也可以通过  $\tilde{e}_2$  对  $\tilde{X}_1$  回归得到, 这里的  $\tilde{e}_2$  是  $X_1$  对  $X_2$  回归的残差, 而  $\tilde{X}_1$  是  $Y$  对  $X_2$  回归的残差.

FWL 定理表明,  $X_2$  的回归系数  $\hat{\beta}_2$  表示的是“过滤掉  $X_1$  影响的  $X_2$ ”对“过滤掉  $X_1$  影响的  $Y$ ”的作用,  $\hat{\beta}_1$  也可以做类似解释.

## 2.9 正态假设下的参数检验

为了检验关于 OLS 估计量的相关假设, 我们需要构建 OLS 估计量在有限样本下的参数检验统计量, 在此之前还需要假定随机扰动项服从条件正态分布.

### 假设 2.5 (正态扰动项)

$$e|X \sim N(0, \sigma^2 I_n).$$

**注** 事实上, 假设 2.5 可以推出假设 2.2 和 2.4 成立, 因此这是个很强的假定.

不仅如此, 我们还需要知道  $\hat{\beta}$  和  $s^2$  的抽样分布, 这就要求先得到  $s^2$  的统计性质, 为此先介绍一个引理.

### 引理 2.1

设  $m \times 1$  维随机向量  $v \sim N(0, I_m)$ ,  $A$  是一个  $m \times m$  维非随机对称幂等矩阵,  $\text{rank}(A) = q \leq m$ , 那么二次型  $v'Av \sim \chi_q^2$ .

**证明** 因为  $A$  是实对称阵, 故而存在正交矩阵  $Q$ , 使得  $Q'AQ = \Lambda$ , 其中  $\Lambda$  为对角矩阵, 其对角元素均为  $A$  的特征值. 由于  $A$  又是幂等矩阵, 故而其特征值只要 0 和 1, 因为  $\text{rank}(A) = q$ , 不妨设  $\Lambda$  对角线上的前  $q$  个元素为 1, 其余元素为 0.

由于  $Q$  是正交矩阵, 其各行各列都是单位向量且两两正交, 故而

$$p = Q'v \sim N(0, I_n)$$

因此二次型

$$v'Av = p'\Lambda p \sim \chi_q^2$$

由此证得引理.

### 定理 2.12

设  $s^2$  为  $\sigma^2$  的残差方差估计量, 那么在假设 2.1, 2.3 和 2.5 下:

$$(1) \frac{(n-K)s^2}{\sigma^2} \bigg| X = \frac{\hat{e}'\hat{e}}{\sigma^2} \bigg| X \sim \chi_{n-K}^2.$$

(2) 在给定  $X$  时,  $s^2$  和  $\hat{\beta}$  独立.

**证明** (1) 根据  $\hat{e}'\hat{e} = eMe$  可知

$$\frac{\hat{e}'\hat{e}}{\sigma^2} = \frac{eMe}{\sigma^2} = \left(\frac{e}{\sigma}\right)' M \left(\frac{e}{\sigma}\right)$$

在假设 2.5 下,  $\frac{e}{\sigma} \bigg| X \sim N(0, I_n)$ , 又因为  $M$  是一个秩为  $n-K$  的幂等矩阵, 根据引理 2.1 可知 (1) 成立.

(2) 注意到  $s^2 = \hat{e}'\hat{e}/(n-K)$ , 只需证明  $\hat{e}$  与  $\hat{\beta}$  相互独立即可. 由于

$$\begin{bmatrix} \hat{e} \\ \hat{\beta} - \beta \end{bmatrix} = \begin{bmatrix} M \\ (X'X)^{-1}X' \end{bmatrix} e \quad (2.10)$$

根据  $e|X \sim N(0, \sigma^2 I_n)$ , 上式在给定  $X$  的条件下也服从正态分布, 由于联合正态分布的不相关意味着独立性, 根据定理 2.8(3) 可知  $\hat{e}$  与  $\hat{\beta}$  相互独立.

现在来构建参数假设检验, 考虑如下线性假设

$$R\beta = r$$

其中  $R$  为  $J \times K$  维选择矩阵,  $r$  是  $J \times 1$  维列向量,  $J$  是参数的限制条件数目.

**例 2.1** 考虑线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

我们想检验原假设  $\mathbb{H}_0: \beta_2 = \beta_3 + 1, \beta_4 = 0$ , 那么

$$R = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

### 引理 2.2

在假设 2.1, 2.3 和 2.5 下, 对于任意非随机  $J \times K$  维矩阵  $R$  都有

$$R(\hat{\beta} - \beta)|X \sim N(0, \sigma^2 R(X'X)^{-1}R')$$



**证明** 在给定  $X$  的条件下,  $\hat{\beta} - \beta$  是  $e$  的线性组合, 因此也服从正态分布. 又因为

$$\mathbb{E}[R(\hat{\beta} - \beta)|X] = R\mathbb{E}[\hat{\beta} - \beta|X] = 0$$

以及

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

因此  $R(\hat{\beta} - \beta)|X \sim N(0, \sigma^2 R(X'X)^{-1}R')$ .

### 推论 2.1

在假设 2.1, 2.3 和 2.5 下, 当原假设  $\mathbb{H}_0: R\beta = r$  成立时有

$$(R\hat{\beta} - r)|X \sim N(0, \sigma^2 R(X'X)^{-1}R')$$



## 2.9.1 $T$ 检验

我们先考虑  $J = 1$  的情况, 也即单个约束条件的情况.

### 定理 2.13 ( $T$ 统计量的分布)

在假设 2.1, 2.3 和 2.5 下, 若  $J = 1$ , 如果原假设  $\mathbb{H}_0: R\beta = r$  成立, 那么  $T$  统计量

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(X'X)^{-1}R'}} \sim t_{n-K} \quad (2.11)$$



**证明** 根据推论2.1, 在原假设  $\mathbb{H}_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$  成立的情况下

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})|\mathbf{X} \sim N(0, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}')$$

从而在给定  $\mathbf{X}$  时有

$$\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}}{\sqrt{\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}} \sim N(0, 1)$$

将 (2.12) 进行等价变换得

$$T = \frac{\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}}{\sqrt{s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}} = \frac{\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}}{\sqrt{\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2}}/(n-K)}$$

根据定理2.12, 由  $t$  分布的定义可知

$$T \sim \frac{N(0, 1)}{\sqrt{\chi_{n-K}^2/(n-K)}} \sim t_{n-K}$$

**注** 事实上, 定理2.13是一个相当一般的形式,  $T$  检验并不局限于检验单个系数, 它的本质实际上是单个约束条件下的检验. 如果考虑特殊的原假设  $\mathbb{H}_0: \hat{\beta}_j = \beta_j$ , 那么对应的  $T$  统计量为

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{jj}^{-1}}}$$

这里的  $\beta_j$  是我们预设的想要检验的真实参数值, 上式的分母称为估计量  $\hat{\beta}_j$  的标准误, 其中  $(\mathbf{X}'\mathbf{X})_{jj}^{-1}$  是矩阵  $(\mathbf{X}'\mathbf{X})^{-1}$  主对角线上的第  $(j, j)$  个元素.

现在来看  $T$  检验的步骤. 给定显著性水平  $\alpha \in (0, 1)$ ,  $|T| > C_{T_{n-K}, \frac{\alpha}{2}}$ , 则拒绝原假设  $\mathbb{H}_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ . 其中,  $C_{T_{n-K}, \frac{\alpha}{2}}$  是  $T_{n-K}$  分布在  $\alpha/2$  水平上的右侧临界值. 用概率的形式可以表述为

$$\mathbb{P}[|T_{n-K}| > C_{T_{n-K}, \frac{\alpha}{2}}] = \alpha$$

反之若  $|T| \leq C_{T_{n-K}, \frac{\alpha}{2}}$ , 则不拒绝原假设  $\mathbb{H}_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ . 注意, 我们通常不说“接受原假设  $\mathbb{H}_0$ ”, 因为即使  $T$  检验统计量不超过对应的临界值, 它也并非是支持  $\mathbb{H}_0$  为真的证据, 只是说没有充分的证据拒绝  $\mathbb{H}_0$  为真.

当我们在检验  $\mathbb{H}_0$  时, 样本的有限性意味着它包含的总体信息也是有限的, 因此可能存在两种错误: I 类错误为原假设为真但被拒绝, 显著性水平  $\alpha$  就是犯第 I 类错误的概率, 也即

$$\mathbb{P}[|T| > C_{t_{n-K}, \frac{\alpha}{2}} | \mathbb{H}_0] = \alpha$$

如果  $\alpha$  显著性水平的功效函数满足

$$\mathbb{P}[|T| > C_{t_{n-K}, \frac{\alpha}{2}} | \mathbb{H}_1] < 1$$

则存在  $\mathbb{H}_0$  为假时被接受的可能性, 称为第 II 类错误. 其中  $\mathbb{H}_1$  为  $\mathbb{H}_0$  的备择假设.

理想的情形是同时最小化 I 类错误和 II 类错误, 然而对于有限样本而言, 这几乎是不可能完成的. 在实际应用中, 我们通常事先设定 I 类错误的水平, 即显著性水平  $\alpha$ , 通常将它选为 10%, 5% 或 1%.

注意, 如果模型中存在近似多重共线性, 则随着样本容量  $n$  增大,  $\text{var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$  并不趋近于 0,



$T$  统计量不显著的可能性增大, 因此若原假设  $\mathbb{H}_0$  为假, 则我们可能会接受它. 换言之, 近似多重共线性会影响  $T$  检验的 II 类错误.

除了使用基于  $T$  检验统计量的方法外, 我们还可以使用基于  $p$  值的判断法则,  $p$  值是可以拒绝原假设  $\mathbb{H}_0$  的最小显著性水平  $\alpha$ , 并且 Stata 之类的软件可以直接给出  $p$  值. 举例而言, 如果  $p = 0.11$ , 那么我们可以在  $\alpha > 0.11$  的水平下拒绝  $\mathbb{H}_0$ , 但是无法在  $\alpha < 0.11$  的水平下拒绝  $\mathbb{H}_0$ . 相较于统计量的临界值,  $p$  值不仅告诉人们是否应该在某一显著性水平下拒绝  $\mathbb{H}_0$ , 还能告诉人们拒绝或不拒绝的程度有多大.

## 2.9.2 $F$ 检验

现在来考虑线性约束个数  $J > 1$  的情况, 为了构造合适的检验统计量, 这里先给出另一个引理.

### 引理 2.3

设  $q \times 1$  随机向量  $Z \sim N(0, V)$ , 其中  $V = \text{var}(Z)$  是一个  $q \times q$  维可逆对称的协方差矩阵, 则  $Z'V^{-1}Z \sim \chi_q^2$ .



**证明** 因为  $V$  的对称正定的, 因此存在可逆对称矩阵  $V^{1/2}$  使得

$$\begin{aligned} V^{1/2}V^{1/2} &= V \\ V^{-1/2}V^{-1/2} &= V^{-1} \end{aligned}$$

现在定义随机变量  $Y = V^{-1/2}Z$ , 则  $\mathbb{E}[Y] = 0$ , 以及

$$\begin{aligned} \text{var}(Y) &= \mathbb{E}[YY'] = V^{-1/2}\mathbb{E}[ZZ']V^{-1/2} \\ &= V^{-1/2}VV^{-1/2} = I_q \end{aligned}$$

从而  $Y \sim N(0, I_q)$ ,  $Y_1, Y_2, \dots, Y_q$  为相互独立的标准正态分布, 于是

$$Z'V^{-1}Z = Y'Y = \sum_{i=1}^q Y_i^2 \sim \chi_q^2$$

由此证得引理.

### 定理 2.14 ( $F$ 统计量的分布)

在假设 2.1, 2.3 和 2.5 下, 若  $J > 1$ , 如果原假设  $\mathbb{H}_0: R\beta = r$  成立, 那么  $F$  统计量

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} \sim F_{J, n-K} \quad (2.12)$$



**证明** 根据推论 2.1, 在原假设  $\mathbb{H}_0: R\beta = r$  成立的情况下

$$(R\hat{\beta} - r)|X \sim N(0, \sigma^2 R(X'X)^{-1}R')$$

由引理 2.3 可知

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi_J^2$$

将 (2.12) 等价变换为

$$F = \frac{\frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{\sigma^2}}{\frac{(n-K)s^2}{\sigma^2}/(n-K)}$$

根据定理 2.12 得到

$$F \sim \frac{\chi_J^2/J}{\chi_{n-K}^2/(n-K)} \sim F_{J,n-K}$$

由此证得定理.

**注** 当  $J = 1$  时,  $F_{1,n-K} = t_{1,n-K}$ , 因此  $T$  检验和  $F$  检验是等价的.

$F$  检验的步骤同  $T$  检验类似但略有不同,  $T$  检验属于双边检验, 在我们预设了显著性水平  $\alpha$  后, 计算的临界值以  $\alpha/2$  为基准. 而  $F$  检验为单边检验, 首先计算出  $F_{J,n-K}$  在  $\alpha$  分位点对应的临界值  $C_{F_{J,n-K},\alpha}$ , 也即

$$\mathbb{P}[F > C_{F_{J,n-K},\alpha}] = \alpha$$

如果  $F$  的值大于临界值  $C_{F_{J,n-K},\alpha}$ , 那么在显著性水平  $\alpha$  下拒绝原假设  $\mathbb{H}_0$ , 否则不拒绝它.

由于按照定义计算  $F$  统计量比较麻烦, 我们在本节的最后部分给出一个更方便的方法来计算它.

#### 定理 2.15

给定假设 2.1 和 2.3, 令  $\hat{e}'\hat{e}$  为以下无约束回归模型的残差平方和

$$Y = X\beta + e$$

再令  $\tilde{e}'\tilde{e}$  为以下有约束回归模型的残差平方和

$$Y = X\beta + e$$

$$\text{s.t. } \mathbf{R}\beta = \mathbf{r}$$

这里  $\tilde{e} = Y - X\tilde{\beta}$ , 而  $\tilde{\beta}$  是有约束回归模型的 OLS 估计量. 那么  $F$  统计量可以写为

$$F = \frac{(\tilde{e}'\tilde{e} - \hat{e}'\hat{e})/J}{\hat{e}'\hat{e}/(n-K)}$$



**证明** 设  $\tilde{\beta}$  是原假设  $\mathbb{H}_0: \mathbf{R}\beta = \mathbf{r}$  成立时有约束模型的 OLS 估计量, 即

$$\tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^K} (Y - X\beta)'(Y - X\beta)$$

现在构建 Lagrange 函数

$$L(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) - 2\lambda'(r - \mathbf{R}\beta)$$

其中  $\lambda$  为  $J \times 1$  维 Lagrange 乘子向量. 根据约束最优化的 FOC 可知

$$\frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \beta} = -2X'(Y - X\tilde{\beta}) + 2\mathbf{R}'\tilde{\lambda} = 0$$

$$\frac{\partial L(\tilde{\beta}, \tilde{\lambda})}{\partial \lambda} = r - \mathbf{R}\tilde{\beta} = 0$$

由于无约束模型的 OLS 估计量为  $\hat{\beta} = (X'X)^{-1}X'Y$ , 于是

$$\begin{aligned}\hat{\beta} - \tilde{\beta} &= (X'X)^{-1}R'\tilde{\lambda} \\ R(X'X)^{-1}R'\tilde{\lambda} &= R(\hat{\beta} - \tilde{\beta})\end{aligned}$$

因此 Lagrange 乘子可以表示为

$$\tilde{\lambda} = [R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

现在将其代入到  $\hat{\beta} - \tilde{\beta}$  的表达式中得到

$$\hat{\beta} - \tilde{\beta} = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

根据定义, 有约束回归模型的残差为

$$\tilde{e} = Y - X\tilde{\beta} = \hat{e} + X(\hat{\beta} - \tilde{\beta})$$

从而

$$\tilde{e}'\tilde{e} - \hat{e}'\hat{e} = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

根据  $F$  统计量的定义得到

$$\begin{aligned}F &= \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/J}{s^2} \\ &= \frac{(\tilde{e}'\tilde{e} - \hat{e}'\hat{e})/J}{\hat{e}'\hat{e}/(n - K)}\end{aligned}$$

由此证得定理.

### 推论 2.2

对于原假设  $\mathbb{H}_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ ,  $F$  统计量为

$$F = \frac{R^2/(K - 1)}{(1 - R^2)/(n - K)}$$

其中  $R^2$  为无约束回归模型的可决系数.



**证明** 在  $\mathbb{H}_0$  成立的条件下,  $\tilde{\beta} = (\bar{Y}, 0, \dots, 0)'$ , 从而

$$\tilde{e}'\tilde{e} = (Y - \mathbf{1}_n\bar{Y})'(Y - \mathbf{1}_n\bar{Y})$$

根据  $R^2$  的定义可知

$$R^2 = 1 - \frac{\hat{e}'\hat{e}}{\tilde{e}'\tilde{e}}$$

于是

$$F = \frac{(\tilde{e}'\tilde{e} - \hat{e}'\hat{e})/(K - 1)}{\hat{e}'\hat{e}/(n - K - 1)} = \frac{R^2/(K - 1)}{(1 - R^2)/(n - K)}$$

## 2.10 广义最小二乘估计

考虑如下矩阵形式的线性回归模型

$$Y = X\beta + e \tag{2.13}$$

随机扰动项之间存在已知形式的相关性和异方差

$$\begin{aligned}\mathbb{E}[e|X] &= 0 \\ \text{var}(e|X) &= \sigma^2 \Sigma\end{aligned}\tag{2.14}$$

这里的  $\Sigma$  是一个  $n \times n$  维对称的正定矩阵, 并且它是关于数据矩阵  $X$  的函数,  $\sigma^2$  仍是一个未知的常数. 现在来研究 OLS 估计量的统计性质.

### 定理 2.16

在假设 2.1, 2.3 和条件 (2.14) 下, OLS 估计量具有以下性质.

- (1) 无偏性:  $\mathbb{E}[\hat{\beta}|X] = \beta$ .
- (2) 方差:  $\text{var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}$ .
- (3) 相关性:  $\text{cov}(\hat{\beta}, \hat{e}|X) \neq 0$ .



**证明** (1) 由  $\hat{\beta} - \beta = (X'X)^{-1}X'e$  可知

$$\mathbb{E}[(\hat{\beta} - \beta)|X] = (X'X)^{-1}X'\mathbb{E}[e|X] = 0$$

(2) 类似之前的证明可得

$$\begin{aligned}\text{var}(\hat{\beta}|X) &= \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] \\ &= \mathbb{E}[(X'X)^{-1}X'ee'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'\mathbb{E}[ee'|X]X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}\end{aligned}$$

(3) 因为  $X'\Sigma M \neq 0$ , 于是

$$\begin{aligned}\text{cov}(\hat{\beta}, \hat{e}|X) &= \mathbb{E}[(\hat{\beta} - \beta)e'|X] \\ &= \mathbb{E}[(X'X)^{-1}X'ee'M|X] \\ &= (X'X)^{-1}X'\mathbb{E}[ee'|X]M \\ &= \sigma^2(X'X)^{-1}X'\Sigma M \neq 0\end{aligned}$$

以上定理表明, 即使随机扰动项存在异方差或自相关, OLS 估计量  $\hat{\beta}$  仍是无偏的, 但由于  $\hat{\beta}$  的方差不再具有简单形式的  $\sigma^2(X'X)^{-1}$ , 因此之前提到的基于简单形式方差的  $T$  检验和  $F$  检验均无效. 此外, 即使使用正确的方差  $\sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}$ ,  $T$  检验和  $F$  检验仍无效, 因为  $\hat{\beta}$  和  $\hat{e}$  存在条件相关性. 另一方面, OLS 估计量此时显然不再是 BLUE.

为了解决上述问题, 现在提出一种新的估计方法, 称为广义最小二乘 (Generalized Least Squares, GLS) 估计.

首先我们知道, 对于任意对称正定矩阵  $\Sigma$ , 总存在矩阵  $\Sigma^{-\frac{1}{2}}$ , 使得  $\Sigma^{-1} = \Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}$ . 现在将  $\Sigma^{-\frac{1}{2}}$  左乘回归方程 (2.13) 得到

$$\tilde{Y} = \tilde{X}\beta + \tilde{e}$$

其中  $\tilde{Y} = \Sigma^{-\frac{1}{2}}Y$ ,  $\tilde{X} = \Sigma^{-\frac{1}{2}}X$ , 以及  $\tilde{e} = \Sigma^{-\frac{1}{2}}e$ . 对上式进行 OLS 回归即可得到所谓的 GLS 估

计量

$$\begin{aligned}\tilde{\beta} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y\end{aligned}$$

根据之前的做法, 容易得到以下定理.

### 定理 2.17

在假设 2.1, 2.3 和条件 (2.14) 成立的情况下有

- (1)  $\mathbb{E}[\tilde{\beta}|X] = \beta$ .
- (2)  $\text{var}(\tilde{\beta}|X) = \sigma^2(X'\Sigma^{-1}X)^{-1}$ .
- (3)  $\text{cov}(\tilde{\beta}, \tilde{e}|X) = 0$ , 其中  $\tilde{e} = \tilde{Y} - \tilde{X}\tilde{\beta}$ .
- (4)  $\tilde{\beta}$  为最佳线性无偏估计量 (BLUE).
- (5)  $\mathbb{E}[\tilde{s}^2|X] = \sigma^2$ , 其中  $\tilde{s}^2 = \hat{e}^{*'}\hat{e}^*/(n-K)$ .



**证明** 变换后的模型  $\tilde{Y} = \tilde{X}\beta + \tilde{e}$  满足 CLRM 假设 2.1–2.4, 并且 GLS 估计量  $\tilde{\beta}$  又是  $\tilde{Y} = \tilde{X}\beta + \tilde{e}$  的 OLS 估计量, 因此根据定理 2.8, 定理 2.9 (Gauss-Markov 定理) 以及定理 2.10 即可证得结论成立.

仅根据定理 2.17 还无法构造新的检验统计量, 如果将条件 (2.14) 强化为

$$e|X \sim N(0, \sigma^2\Sigma) \quad (2.15)$$

那么在假设 2.1, 2.3 和条件 (2.15) 下有

$$\begin{aligned}\tilde{T} &= \frac{R\tilde{\beta} - r}{\sqrt{\tilde{s}^2 R(X'\Sigma^{-1}X)^{-1}R'}} \sim t_{n-K} \\ \tilde{F} &= \frac{(R\tilde{\beta} - r)'[R(X'\Sigma^{-1}X)^{-1}R']^{-1}(R\tilde{\beta} - r)/J}{\tilde{s}^2} \sim F_{J, n-K}\end{aligned}$$

可以看出, 只要是在有限样本条件下, 无论是 OLS 还是 GLS, 在构造检验统计量的时候都必须假设正态随机扰动项. 不仅如此, 使用 GLS 估计必须已知误差项的协方差矩阵的具体形式, 这在实际应用中难以满足, 因此 GLS 的理论意义远高于实践意义.

在一些情形下, 我们可以用估计量  $\hat{\Sigma}$  来替代协方差矩阵  $\Sigma$ , 估计量  $\hat{\Sigma}$  既可以是参数的也可以是非参数的, 由此可以得到以下可行广义最小二乘估计 (Feasible Generalized Least Squares, FGLS)

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

FGLS 的困难在于  $\Sigma$  包含太多参数, 可能多达  $n(n+1)/2$  个, 然而样本容量大小只有  $n$ . 因此如果不对  $\Sigma$  的形式施加一定限制, 那么无法用这  $n$  个数据一致估计出  $\Sigma$ . 例如, 可假设误差项仅存在异方差而不存在自相关, 也即

$$\mathbb{E}[ee'|X] = \text{diag}\{\sigma^2(X_1), \sigma^2(X_2), \dots, \sigma^2(X_n)\}$$

此时得到的 FGLS 估计量  $\tilde{\beta}_F$  的有限样本分布与方差形式  $\Sigma$  已知时得到的 GLS 估计量  $\tilde{\beta}$  的有限样本分布不同, 这是因为  $\hat{\Sigma}$  估计量的抽样误差会对估计量  $\tilde{\beta}_F$  造成一定影响.

然而在一定正则条件下, 可以证明 GLS 估计量和 FGLS 估计量具有相同的渐近性质, 具体

见第五章的相关讨论.

## 2.11 聚类样本

重新回到假设2.1, 它要求个体  $i$  与个体  $j$  的决策互不影响. 然而, 当样本内的个体存在某些联系时, 该假设不成立, 例如: 这些个体间的关系是邻居, 同班同学等.

下面这个例子由 Duflo et al. (2011) 给出. 在 2005 年, Kenya 的 140 座小学获得了一笔资金招募更多的一年级教师以降低班级规模, 其中的 121 座学校原本仅有一个一年级班级, 在招募新教师后, 将这些学校的一年级班级分为两个组别, 其中一个组别由新教师进行授课. 在随机选择的 60 座学校中, 学生进入哪个组别取决于 TA 的初始测验成绩, 我们称其为 Tracking 学校; 而在剩余的 61 座学校中, 学生随机进入两个组别的班级.

现在考虑以下模型

$$\text{TestScore}_{ig} = \alpha + \gamma \text{Tracking}_g + X'_{ig}\beta + e_{ig}$$

其中  $\text{TestScore}_{ig}$  是学校  $g$  中的学生  $i$  在 18 个月后的测验成绩,  $\text{Tracking}_g$  是虚拟变量, 当学校  $g$  为 Tracking 学校时取值为 1,  $X_{ig}$  是一系列控制变量.

我们的目的是估计参数  $\gamma$ , 困难在于学生的测试成绩可能受到同校其他学生的影响, 假设2.1不成立, 因而难以在经典线性回归模型的框架中进行分析. 一个更合理的假设是, 这种相关性不会存在于不同学校中.

现在我们将样本表示为  $\{Y_{ig}, X_{ig}\}$ , 其中  $g = 1, 2, \dots, G$  表示聚类指标,  $i = 1, 2, \dots, n_g$  为第  $g$  个聚类中的个体, 观测值总计为  $n = \sum_{g=1}^G n_g$ . 在 Duflo et al. (2011) 的研究中,  $G = 121$ , 每个学校的样本学生数量从 19 到 62 不等, 观测值总共为  $n = 5795$ .

一般地, 聚类样本中的线性回归模型为

$$Y_{ig} = X'_{ig}\beta + e_{ig}$$

也可以更紧凑地表示为

$$Y_g = X_g\beta + e_g \tag{2.16}$$

其中

$$\begin{aligned} Y_g &= (Y_{1g}, Y_{2g}, \dots, Y_{n_gg})' \\ X_g &= (X_{1g}, X_{2g}, \dots, X_{n_gg})' \\ e_g &= (e_{1g}, e_{2g}, \dots, e_{n_gg})' \end{aligned}$$

而全样本的回归模型仍为  $Y = X\beta + e$ .



根据以上的表示法, 可以写出 OLS 估计量

$$\begin{aligned}\hat{\beta} &= \left( \sum_{g=1}^G \sum_{i=1}^{n_g} X_{ig} X'_{ig} \right)^{-1} \left( \sum_{g=1}^G \sum_{i=1}^{n_g} X_{ig} Y_{ig} \right) \\ &= \left( \sum_{g=1}^G X'_g X_g \right)^{-1} \left( \sum_{g=1}^G X'_g Y_g \right) \\ &= (X'X)^{-1} X'Y\end{aligned}\quad (2.17)$$

为了得到 OLS 估计量的统计性质, 我们给出以下假设.

#### 假设 2.6

聚类样本  $(Y_g, X_g)$  在聚类  $g$  之间是相互独立的.

进一步, 如果以下矩条件成立

$$\mathbb{E}[e_g | X_g] = 0 \quad (2.18)$$

则线性回归模型是正确识别的, 它要求每个聚类中的不可观测因素与回归元  $X_{ig}$  无关.

#### 定理 2.18

在假设 2.6 和条件 (2.18) 下,  $\mathbb{E}[\hat{\beta} | X] = \beta$ .

**证明** 根据 (2.16) 和 (2.17) 可知

$$\hat{\beta} - \beta = \left( \sum_{g=1}^G X'_g X_g \right)^{-1} \left( \sum_{g=1}^G X'_g e_g \right)$$

再由  $\mathbb{E}[e_g | X] = \mathbb{E}[e_g | X_g] = 0$  即可证得定理.

进一步考察  $\hat{\beta}$  的协方差矩阵, 记  $\Sigma_g = \mathbb{E}[e_g e'_g | X_g]$  表示第  $g$  个聚类的  $n_g \times n_g$  维协方差矩阵, 在假设 2.6 下可以得到

$$\begin{aligned}\text{var} \left( \sum_{g=1}^G X'_g e_g \middle| X \right) &= \sum_{g=1}^G \text{var}(X'_g e_g | X_g) \\ &= \sum_{g=1}^G X'_g \mathbb{E}[e_g e'_g | X_g] X_g \\ &= \sum_{g=1}^G X'_g \Sigma_g X_g \equiv \Omega_n\end{aligned}\quad (2.19)$$

此时

$$V_{\text{OLS}} = (X'X)^{-1} \Sigma_n (X'X)^{-1}$$

由于聚类内的观测值存在相关性, 因此我们对协方差矩阵的分析需要特别小心. 考虑一种简单的情形:  $n_g = N$ ,  $\mathbb{E}[e_{ig}^2 | X_g] = \sigma^2$ , 以及  $\mathbb{E}[e_{ig} e_{jg} | X_g] = \sigma^2 \rho$ ,  $i \neq j$ , 此时

$$V_{\text{OLS}} = \sigma^2 [1 + \rho(N-1)] (X'X)^{-1}$$

Arellano (1987) 给出了一般情况下  $\Omega_n$  的聚类稳健估计量

$$\begin{aligned}\hat{\Omega}_n &= \sum_{g=1}^G X'_g \hat{e}_g \hat{e}'_g X_g \\ &= \sum_{g=1}^G \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} X_{ig} X'_{jg} \hat{e}_{ig} \hat{e}_{jg} \\ &= \sum_{g=1}^G \left( \sum_{i=1}^{n_g} X_{ig} \hat{e}_{ig} \right) \left( \sum_{j=1}^{n_g} X_{jg} \hat{e}_{jg} \right)'\end{aligned}$$

此时协方差矩阵估计量为

$$\hat{V}_{OLS} = a_n (X'X)^{-1} \hat{\Omega}_n (X'X)^{-1}$$

其中

$$a_n = \left( \frac{n-1}{n-K} \right) \left( \frac{G}{G-1} \right)$$

为有限样本调整参数.

另一方面, 许多聚类理论都设置了  $n_g = N$  这一条件以简化推导, 但实证分析中很难遇到这种情况. 例如中国西部地区和东部地区的人口数量在省级层面明显不同, 此时聚类稳健推断差不多可以看作是具有极强异质性观测时的异方差稳健推断.

那么什么时候应该使用聚类稳健标准误? Abadie et al. (2023) 在最近的研究中给出了答案, 他们提出了因果聚类方差 (Causal Cluster Variance, CCV) 修正传统的异方差稳健标准误和聚类稳健标准误. 特别是, 如果样本中的聚类数量占相当部分的总体聚类数量时, CCV 将会产生非常显著的修正效果.

## 第3章 渐近理论基础

正如 Wooldridge (2010) 所言, 多数估计量并不像 OLS 估计量那样具有无偏性等性质, 因此我们很难对其进行有限样本分析, 而是将其纳入到大样本框架下得到它的渐近性质.

### 3.1 收敛概念

确定性序列的收敛和有界性在数学分析中已经学过, 这里给出几个新的符号. 如果  $\{a_n\}$  的极限为 0, 则记为  $a_n = o(1)$ ; 如果  $\{a_n\}$  有界, 则记为  $a_n = O(1)$ . 更一般地, 如果  $n^{-\lambda}a_n \rightarrow 0$ , 则  $a_n = o(n^\lambda)$ ; 如果  $\{n^{-\lambda}a_n\}$  有界, 则  $a_n = O(n^\lambda)$ .

设  $\{X_n\}_{n=1}^\infty$  是由定义在同一个概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$  上的随机变量构成的序列<sup>1</sup>, 现在我们给出以下几个收敛的概念.

#### 定义 3.1

(1)  $\{X_n\}$  依概率收敛于  $X$ , 如果对于任意给定的  $\varepsilon > 0$ , 当  $n \rightarrow \infty$  时都有

$$\mathbb{P}[|X_n - X| < \varepsilon] \rightarrow 1$$

记作  $X_n \xrightarrow{p} X$ ,  $X_n - X = o_p(1)$  或  $\text{plim } X_n = X$ .

(2)  $\{X_n\}$  是依概率有界的, 如果对于任意给定的  $\varepsilon > 0$ , 总存在  $0 < \delta < \infty$ , 使得当  $n \rightarrow \infty$  时就有

$$\mathbb{P}[|X_n| > \delta] < \varepsilon$$

记作  $X_n = O_p(1)$ .

(3)  $\{X_n\}$  几乎必然收敛于  $X$ , 如果存在  $A \in \mathcal{F}$ , 使得对于一切  $\omega \in A$  都有  $\mathbb{P}[A] = 1$ , 并且

$$\mathbb{P}\left[\lim_{n \rightarrow \infty} X_n = X\right] = 1$$

记作  $X_n \xrightarrow{a.s.} X$ .


(4)  $\{X_n\}$  依  $L^p$  收敛于  $X$ ,  $p > 0$ , 如果当  $n \rightarrow \infty$  时有

$$\mathbb{E}|X_n - X|^p \rightarrow 0$$

记作  $X_n \xrightarrow{L^p} X$ . 当  $p = 2$  时, 称  $\{X_n\}$  依均方收敛于  $X$ , 记作  $X_n \xrightarrow{q.m.} X$ .

(5)  $\{X_n\}$  依分布收敛于  $X$ , 如果对于一切  $x \in \mathcal{C}(F)$  都有

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

其中  $F_n$  和  $F$  分别为  $X_n$  和  $X$  的累积概率分布,  $\mathcal{C}(F)$  为  $F$  的一切连续点构成的集合. 

**注** 更一般地, 如果  $\{a_n\}$  为非随机序列, 并且  $X_n/a_n = o_p(1)$ , 那么  $X_n = o_p(a_n)$ . 类似可以定义  $O_p(a_n)$  的概念.

<sup>1</sup>事实上, 依分布收敛可以将  $X_n$  定义在不同的概率空间  $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$  上, 但其余的收敛概念只能将  $X_n$  定义在同一个概率空间上.

先来看依概率收敛  $o_p(1)$  和依概率有界  $O_p(1)$  之间的关系.

### 定理 3.1

设  $w_n = o_p(1)$ ,  $x_n = o_p(1)$ ,  $y_n = O_p(1)$ ,  $z_n = O_p(1)$ , 那么

$$(1) w_n + x_n = o_p(1).$$

$$(2) y_n + z_n = O_p(1).$$

$$(3) w_n + y_n = O_p(1).$$

$$(4) w_n x_n = o_p(1).$$

$$(5) y_n z_n = O_p(1).$$

$$(6) w_n y_n = o_p(1).$$



**证明** 这里只证明 (2), (4) 和 (6), 其余命题证法类似.

根据定义可知, 当  $n \rightarrow \infty$  时有

$$\mathbb{P}[|w_n + x_n| > 2\delta] \leq \mathbb{P}[|x_n| > \delta] + \mathbb{P}[|w_n| > \delta] < 2\varepsilon$$

$$\mathbb{P}[w_n x_n > \varepsilon] \leq \mathbb{P}[|w_n| > \varepsilon] + \mathbb{P}[|x_n| > 1] \rightarrow 0$$

以及

$$\begin{aligned} \mathbb{P}[|w_n y_n| > \varepsilon] &\leq \mathbb{P}[|w_n y_n| > \varepsilon, |y_n| < \delta] + \mathbb{P}[|w_n y_n| > \varepsilon, |y_n| \geq \delta] \\ &\leq \mathbb{P}\left[|w_n| > \frac{\varepsilon}{\delta}\right] + \mathbb{P}[|y_n| \geq \delta] < \varepsilon \end{aligned}$$

因为  $\varepsilon$  可以任意小, 因此命题 (6) 成立.

现在我们给出几个关于随机收敛的例子.

**例 3.1** 考虑概率空间  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P})$ , 其中  $\mathcal{B}([0, 1])$  为  $[0, 1]$  上的 Borel 集构成的集族,  $\mathbb{P}$  为  $\mathbb{R}$  上的 Lebesgue 测度,  $s$  在  $[0, 1]$  上均匀分布. 定义随机变量  $X(\omega) = \omega$ , 以及随机变量序列

$$X_n(\omega) = \begin{cases} \omega + \omega^n, & \omega \in [0, 1 - n^{-1}] \\ \omega + 1, & \omega \in (1 - n^{-1}, 1] \end{cases}$$

那么随机变量序列  $\{X_n\}$  依  $L^p$  收敛于  $X$ , 依概率收敛于  $X$ , 并且几乎必然收敛于  $X$ . 这是因为当  $n \rightarrow \infty$  时有

$$\begin{aligned} \mathbb{E}|X_n - X|^p &= \int_0^{1-n^{-1}} \omega^{pn} d\omega + \int_{1-n^{-1}}^1 d\omega \\ &= \frac{1}{pn+1} \left(1 - \frac{1}{n}\right)^{pn+1} + \frac{1}{n} \rightarrow 0 \end{aligned}$$

并且对于任意  $0 < \varepsilon < 1$  都有

$$\mathbb{P}[|X_n - X| < \varepsilon] = \mathbb{P}[X_n = \omega + \omega^n] = 1 - \frac{1}{n} \rightarrow 1$$

由此分别得出  $X_n \xrightarrow{L^p} X$  以及  $X_n \xrightarrow{p} X$ .

最后, 令  $A = \left\{ \omega \in \Omega : \lim_{n \rightarrow \infty} X_n = X \right\}$ , 对于一切  $\omega \in [0, 1)$  都有

$$\mathbb{P}\left[\lim_{n \rightarrow \infty} X_n = X\right] = (1 - n^{-1})\mathbb{P}\left[\lim_{n \rightarrow \infty} \omega^n = 0\right] + n^{-1}\mathbb{P}\left[\lim_{n \rightarrow \infty} 1 = 0\right] = 1$$

故而  $A = \Omega \setminus \{1\}$ , 并且  $\mathbb{P}[A] = 1$ , 因此  $X_n \xrightarrow{a.s.} X$ .

**例 3.2** 对于一切  $n \geq 1$ , 设随机变量  $X_n$  的概率分布为

$$\begin{array}{ccc} X_n & \frac{1}{n} & n \\ \mathbb{P} & 1 - \frac{1}{n} & \frac{1}{n} \end{array}$$

那么  $\{X_n\}$  不依均方收敛于 0, 但是依概率收敛于 0. 这是因为当  $n \rightarrow \infty$  时

$$\mathbb{E}|X_n|^2 = n + \frac{1}{n^2} - \frac{1}{n^3} \rightarrow \infty$$

并且

$$\mathbb{P}[|X_n| > \varepsilon] = \mathbb{P}[X_n = n] = \frac{1}{n} \rightarrow 0$$

**例 3.3** 设  $\{X_n\}$  中的元素两两独立,  $\mathbb{P}[X_n = 1] = 1/n$  并且  $\mathbb{P}[X_n = 0] = 1 - 1/n$ . 对于任意  $0 < \varepsilon < 1/2$ , 可以得到

$$\mathbb{P}[|X_n| > \varepsilon] = \frac{1}{n}$$

因此  $\{X_n\}$  依概率收敛于 0. 由于

$$\sum_{n=1}^{\infty} \mathbb{P}[X_n = 1] = \infty$$

根据 Borel-Cantelli 第二引理<sup>2</sup>可知  $\mathbb{P}[X_n = 1 \text{ i.o.}] = 1$ , 因此  $\{X_n\}$  并不是几乎必然收敛于 0 的.

**例 3.4** 设概率空间为  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P})$ , 各符号含义与例 3.1 中的一致. 考虑随机变量  $X_n$ , 对一切  $\omega \in \Omega$  都有

$$X_{2n-1}(\omega) = 1 - \omega, \quad X_{2n}(\omega) = \omega$$

于是  $X_{2n-1}$  和  $X_{2n}$  具有同样的极限分布, 从而  $\{X_n\}$  依分布收敛于均匀分布  $U[0, 1]$ , 然而它不依概率收敛于任何随机变量.

**例 3.5** 设概率空间为  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P})$ , 各符号含义与例 3.1 中的一致. 定义

$$X_n(\omega) = \begin{cases} 0, & \omega \in [0, 1 - n^{-2}] \\ e^n, & \omega \in (1 - n^{-2}, 1] \end{cases}$$

可以得到

$$\lim_{n \rightarrow \infty} \mathbb{E}|X_n|^p = \lim_{n \rightarrow \infty} \frac{e^{np}}{n^2} = \infty$$

以及

$$\mathbb{P}\left[\lim_{n \rightarrow \infty} X_n(\omega) = 0\right] = 1$$

因此  $\{X_n\}$  不依  $L^p$  收敛于 0, 但几乎必然收敛于 0.

<sup>2</sup>设  $(\Omega, \mathcal{F}, \mathbb{P})$  为概率空间, 集合序列  $\{A_n\} \subset \mathcal{F}$ , 并且它的各个元素两两独立. 如果级数  $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$ , 那么  $\mathbb{P}[A_n \text{ i.o.}] = 1$ , 这里的 i.o. 表示 infinitely often, 相当于发生无数次事件  $A_n \in \mathcal{F}$ . 证明见 Athreya and Lahiri (2006) Theorem 7.2.2.

再考虑  $Y(\omega) = \omega$ , 以及

$$Y_n(\omega) = \begin{cases} 1, & \omega \in \left[\frac{i}{2^k}, \frac{i+1}{2^k}\right], i = n - 2^k, 1 \leq i \leq 2^k \\ \omega, & \text{otherwise.} \end{cases}$$

于是当  $n \rightarrow \infty$  时有

$$\mathbb{E}|Y_n - Y|^p = 1/2^k \rightarrow 0$$

也即  $Y_n \xrightarrow{L^p} Y$ . 然而对于任意  $\omega \in [0, 1]$ , 极限  $\lim_{n \rightarrow \infty} Y_n(\omega)$  不存在, 因此  $\{Y_n\}$  并不几乎必然收敛于  $Y$ .

以上例子主要来源于 Hong (2017), 从中我们发现: 依概率收敛并不意味着依均方收敛, 也不意味着几乎必然收敛, 并且依分布收敛也不意味着依概率收敛, 依  $L^p$  收敛则与几乎必然收敛互不包含. 现在我们将随机收敛的关系用定理表述如下.

### 定理 3.2

- (1)  $X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{p} X$ .
- (2)  $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$ .
- (3)  $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$ .
- (4)  $X_n \xrightarrow{d} X \Rightarrow X_n = O_p(1)$ .



**证明** (1) 对于任意给定的  $\varepsilon > 0$ , 由 Markov 不等式<sup>3</sup>可知当  $n \rightarrow \infty$  时有

$$\mathbb{P}[|X_n - X| > \varepsilon] \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} = 0$$

从而  $X_n \xrightarrow{p} X$ .

(2) 任取  $\varepsilon > 0$ , 可以得到

$$\mathbb{P}[|X_n - X| < \varepsilon] \geq \mathbb{P}[|X_n - X| < \varepsilon, \forall j \geq n] \quad (3.1)$$

也即

$$\{\omega : |X_j(\omega) - X(\omega)| < \varepsilon, \forall j \geq n\} \subset \{\omega : |X_n(\omega) - X(\omega)| < \varepsilon\}$$

因为  $X_n \xrightarrow{a.s.} X$ , 根据几乎必然收敛的定义可知, 存在充分大的正整数  $N$ , 使得对于任意给定的  $\delta > 0$ , 当  $n > N$  时就有  $\mathbb{P}[|X_n - X| < \varepsilon, \forall j \geq n] > 1 - \delta$ . 在 (3.1) 两端取极限得到

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| < \varepsilon] \geq \lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| < \varepsilon, \forall j \geq n] = 1$$

也即  $X_n \xrightarrow{p} X$ .

(3) 设  $F_n$  和  $F$  分别为  $X_n$  和  $X$  的累积分布函数,  $n \geq 0$ , 固定  $x \in \mathcal{C}(F)$ , 于是对于任意

<sup>3</sup> 设  $X$  是  $K \times 1$  维随机向量,  $\phi: \mathbb{R}^K \rightarrow \mathbb{R}_+$  为单调递增函数, 那么对任意  $t > 0$  都有

$$\mathbb{P}[|X| \geq t] \leq \frac{\mathbb{E}[\phi(|X|)]}{\phi(t)}$$



$\varepsilon > 0$  都有

$$\begin{aligned}
 \mathbb{P}[X_n \leq x] &\leq \mathbb{P}[X_n \leq x, X \leq x + \varepsilon] + \mathbb{P}[X_n \leq x, X > x + \varepsilon] \\
 &\leq \mathbb{P}[X_n \leq x] + \mathbb{P}[X_n - X \leq x - X, x - X_n < -\varepsilon] \\
 &\leq \mathbb{P}[X_n \leq x] + \mathbb{P}[X_n - X < -\varepsilon] + \mathbb{P}[X_n - X > -\varepsilon] \\
 &\leq \mathbb{P}[X \leq x + \varepsilon] + \mathbb{P}[|X_n - X| > \varepsilon]
 \end{aligned} \tag{3.2}$$

类似地有

$$\mathbb{P}[X_n \leq x] \geq \mathbb{P}[X \leq x - \varepsilon] - \mathbb{P}[|X_n - X| > \varepsilon] \tag{3.3}$$

根据 (3.2) 和 (3.3) 可知

$$F(x - \varepsilon) - \mathbb{P}[|X_n - X| > \varepsilon] \leq F_n(x) \leq F(x + \varepsilon) + \mathbb{P}[|X_n - X| > \varepsilon]$$

由于  $X_n \xrightarrow{p} X$ , 令  $n \rightarrow \infty$  可知对一切  $\varepsilon \in (0, \infty)$  都有

$$F(x - \varepsilon) \leq \liminf_{n \rightarrow \infty} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(x) \leq F(x + \varepsilon) \tag{3.4}$$

注意到  $x \in \mathcal{C}(F)$ , 故而  $F(x-) = F(x+)$ , 在 (3.4) 中令  $\varepsilon \downarrow 0$  即得  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ .

(4) 由于  $X$  为随机变量, 故而存在  $\delta > 0$ , 使得  $F$  在  $\delta$  和  $-\delta$  处连续, 并且对于任意给定的  $\varepsilon > 0$  都有

$$\mathbb{P}[|X| > \delta] = F(-\delta) + [1 - F(\delta)] < \frac{\varepsilon}{2} \tag{3.5}$$

因为  $X_n \xrightarrow{d} X$ , 因此对于充分大的  $n$  都有

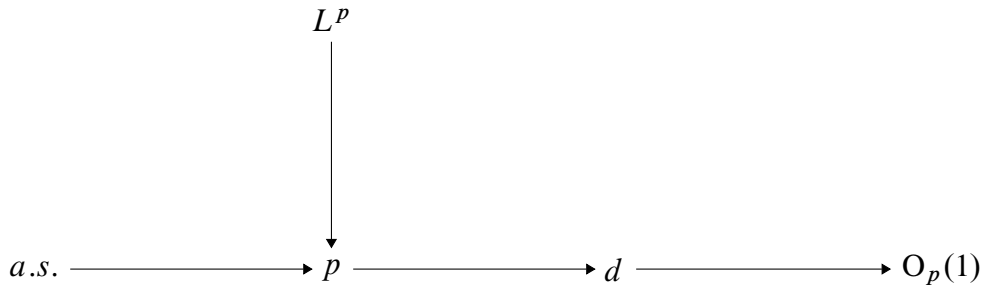
$$|F_n(-\delta) - F(-\delta)| < \frac{\varepsilon}{4} \tag{3.6}$$

$$|F(\delta) - F_n(\delta)| < \frac{\varepsilon}{4} \tag{3.7}$$

将 (3.5)–(3.7) 相加, 由三角不等式即可推知  $\mathbb{P}[|X_n| > \delta] < \varepsilon$ .

**注** 尽管依分布收敛无法推出依概率收敛, 但在特殊情况下是可以的: 如果  $X_n \xrightarrow{d} X$ , 并且存在  $c \in \mathbb{R}$  使得  $\mathbb{P}[X = c] = 1$ , 那么  $X_n \xrightarrow{p} c$ .

我们将上述定理表述成如下示意图:



现在介绍推导渐近多元分布时特别重要的 Slutsky 定理与连续映射定理 (Continuous Mapping Theorem, CMT).

**定理 3.3 (Slutsky 定理)**

设  $\{X_n\}$  和  $\{Y_n\}$  为两个随机变量序列, 并且  $(X_n, Y_n)$  定义在概率空间  $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$  上, 若  $X_n \xrightarrow{d} X$ , 并且存在  $c \in \mathbb{R}$  使得  $Y_n \xrightarrow{p} c$ , 那么

- (1)  $X_n + Y_n \xrightarrow{d} X + c$ .
- (2)  $X_n Y_n \xrightarrow{d} cX$ .
- (3)  $X_n / Y_n \xrightarrow{d} X/c$ , 其中  $c \neq 0$ .



**证明** (1) 设  $F_\infty$  为  $X$  的累积分布函数, 于是  $X + c$  的 cdf 由  $F(x) = F_\infty(x - c)$ ,  $x \in \mathbb{R}$  定义, 固定  $x \in \mathcal{C}(F)$ , 于是  $x - c \in \mathcal{C}(F_\infty)$ .

对于任意给定的  $\varepsilon > 0$ , 可以得到

$$\begin{aligned} \mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n - c| \leq \varepsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n - c| > \varepsilon] \\ &\leq \mathbb{P}[X_n + Y_n \leq x, |Y_n - c| \leq \varepsilon] + \mathbb{P}[|Y_n - c| > \varepsilon] \\ &\leq \mathbb{P}[X_n + c - \varepsilon \leq x] + \mathbb{P}[|Y_n - c| > \varepsilon] \end{aligned} \quad (3.8)$$

类似地有

$$\mathbb{P}[X_n + Y_n \leq x] \geq \mathbb{P}[X_n + c + \varepsilon \leq x] - \mathbb{P}[|Y_n - c| > \varepsilon] \quad (3.9)$$

选取  $x - c, x - c + \varepsilon$  与  $x - c - \varepsilon$ , 使得它们都是  $\mathcal{C}(F_\infty)$  中的元素, 那么由 (3.8) 和 (3.9) 可知

$$\begin{aligned} F_\infty(x - c - \varepsilon) &\leq \liminf_{n \rightarrow \infty} \mathbb{P}[X_n + Y_n \leq x] \\ &\leq \limsup_{n \rightarrow \infty} \mathbb{P}[X_n + Y_n \leq x] \leq F_\infty(x - c + \varepsilon) \end{aligned}$$

现在令  $\varepsilon \rightarrow 0^+$ , 于是

$$\lim_{n \rightarrow \infty} \mathbb{P}[X_n + Y_n \leq x] = F_\infty(x - c) = F(x)$$

(2), (3) 若  $c = 0$ , 根据定理 3.2(3), 只需证明  $X_n Y_n \xrightarrow{p} 0$  即可. 对于任意给定的  $\varepsilon, \delta > 0$ , 可以得到

$$\begin{aligned} \mathbb{P}[|X_n Y_n| > \varepsilon] &= \mathbb{P}[|X_n Y_n| > \varepsilon, |Y_n| \geq \delta] + \mathbb{P}[|X_n Y_n| > \varepsilon, |Y_n| < \delta] \\ &\leq \mathbb{P}[|X_n| \geq \varepsilon/\delta] + \mathbb{P}[|Y_n| \geq \delta] \end{aligned}$$

可以选取  $\pm\varepsilon/\delta$  为  $\mathcal{C}(F_\infty)$  中的元素, 从而

$$0 \leq \liminf_{n \rightarrow \infty} \mathbb{P}[|X_n Y_n| > \varepsilon] \leq \limsup_{n \rightarrow \infty} \mathbb{P}[|X_n Y_n| > \varepsilon] \leq \mathbb{P}[|X_n| > \varepsilon/\delta]$$

令  $\varepsilon \rightarrow 0^+$ , 并且可以将  $\delta$  选得充分小, 使得  $\varepsilon/\delta$  充分大, 因此  $X_n Y_n \xrightarrow{p} 0$ .

若  $c \neq 0$ , 那么  $X_n Y_n = cX_n + X_n(Y_n - c)$ , 上面已经证明了  $X_n(Y_n - c) \xrightarrow{p} 0$ , 因此只需证明  $cX_n \xrightarrow{d} cX$  即可. 当  $c > 0$  时,  $x/c \in \mathcal{C}(F_\infty)$ , 于是

$$\mathbb{P}[cX_n \leq x] = \mathbb{P}[X_n \leq x/c] \rightarrow \mathbb{P}[X \leq x/c] = \mathbb{P}[cX \leq x]$$

也即  $cX_n \xrightarrow{d} cX$ ,  $c < 0$  时的证明类似.

**注** 如果  $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y$ , 那么不一定有  $X_n + Y_n \xrightarrow{d} X + Y$ .

**例 3.6** 设  $X_n$  和  $Y_n$  是相互独立的服从的标准正态分布的随机变量, 那么

$$X_n + Y_n \xrightarrow{d} N(0, 2)$$

另一方面, 如果对于一切  $n \geq 1$  都有  $X_n = Y_n$ , 那么

$$X_n + Y_n = 2X_n \sim N(0, 4)$$

显然  $X_n + Y_n \not\xrightarrow{d} X + Y$ .

#### 定理 3.4 (连续映射定理)

设  $X_1, \dots, X_n$  和  $X$  为  $K \times 1$  维随机向量. 若  $X_n \xrightarrow{d} X$ ,  $f: \mathbb{R}^K \rightarrow \mathbb{R}^L$  为 Borel 可测函数,  $\mathbb{P}[X \in D_f] = 0$ , 其中  $D_f$  为  $f$  的不连续点构成的集合, 那么  $f(X_n) \xrightarrow{d} f(X)$ .



**证明** 见 Klenke (2013) Theorem 13.25.

#### 定理 3.5 (连续映射定理)

设  $X_1, \dots, X_n$  和  $X$  为  $K \times 1$  维随机向量,  $g: \mathbb{R}^K \rightarrow \mathbb{R}^L$  为 Borel 可测函数, 并且在常向量  $\alpha$  处连续, 那么  $X_n \xrightarrow{p} \alpha \Rightarrow g(X_n) \xrightarrow{p} g(\alpha)$ .



**证明** 根据连续函数的定义可知, 对于任意给定的  $\varepsilon > 0$ , 存在  $0 < \delta < \infty$ , 使得当  $\|X_n - \alpha\| < \delta$  时就有

$$\|g(X_n) - g(\alpha)\| < \varepsilon$$

因此

$$\mathbb{P}[\|X_n - \alpha\| < \delta] \leq \mathbb{P}[\|g(X_n) - g(\alpha)\| < \varepsilon]$$

因为  $X_n \xrightarrow{p} \alpha$ , 故而上式左端趋近于 1, 所以  $g(X_n) \xrightarrow{p} g(\alpha)$ .

**注** 如果我们将定理 3.5 中的  $X_n \xrightarrow{p} \alpha$  改为  $X_n \xrightarrow{a.s.} \alpha$ , 其它条件不变, 那么相应的结论变为  $g(X_n) \xrightarrow{a.s.} g(\alpha)$ , 它的证明非常 Trivial.

**例 3.7** 给定假设 2.1, 2.3 和 2.5, 那么  $s^2 \xrightarrow{p} \sigma^2$  且  $s \xrightarrow{p} \sigma$ . 理由如下:

根据相关假设,  $(n - K) \frac{s^2}{\sigma^2} \Big| X \sim \chi_{n-K}^2$ . 进而有

$$\mathbb{E}[s^2 | X] = \sigma^2$$

$$\text{var}(s^2 | X) = \frac{2\sigma^4}{n - K}$$

当  $n \rightarrow \infty$  时,  $\mathbb{E}|s^2 - \sigma^2|^2 = \frac{2\sigma^4}{n - K} \rightarrow 0$  并不依赖于  $X$ . 因此  $s^2 \xrightarrow{q.m.} \sigma^2$ , 由定理 3.2(1) 可知  $s^2 \xrightarrow{p} \sigma^2$ , 根据 CMT 即可推知  $s \xrightarrow{p} \sigma$ .

## 3.2 大数定律

本节主要介绍渐近理论的第二个工具: 大数定律 (Law of Large Numbers, LLN), 它表明 i.i.d. 随机变量的均值将会依概率收敛到一阶矩.

**定理 3.6**

设  $X_1, X_2, \dots, X_n$  是 i.i.d. 随机变量,  $b_n > 0$ ,  $S_n = X_1 + \dots + X_n$ , 并且当  $n \rightarrow \infty$  时以下条件成立: (1)  $b_n \rightarrow \infty$ ; (2)  $\sum_{i=1}^n \mathbb{P}[|X_i| > b_n] \rightarrow 0$ ; (3)  $b_n^{-2} \sum_{i=1}^n \mathbb{E}[X_i^2 \mathbb{1}\{|X_i| \leq b_n\}] \rightarrow 0$ . 那么

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0$$

其中  $a_n = \sum_{j=1}^n \mathbb{E}[X_j \mathbb{1}\{|X_j| \leq b_n\}]$ .



**证明** 设  $Y_j = X_j \mathbb{1}\{|X_j| \leq b_n\}$ ,  $T_n = \sum_{j=1}^n Y_j$ , 于是  $a_n = \mathbb{E}[T_n]$ . 注意到

$$\frac{\sum_{j=1}^n Y_j - a_n}{b_n} = \frac{\sum_{j=1}^n \{Y_j - \mathbb{E}[Y_j]\}}{b_n}$$

的均值为 0, 根据 Chebyshev 不等式<sup>4</sup>可知, 当  $n \rightarrow \infty$  时有

$$\begin{aligned} \mathbb{P}\left[\left|\frac{T_n - a_n}{b_n}\right| > \varepsilon\right] &\leq \varepsilon^{-2} \mathbb{E}\left[\left|\frac{T_n - a_n}{b_n}\right|^2\right] = \varepsilon^{-2} b_n^{-2} \text{var}(T_n) \\ &= (b_n \varepsilon)^{-2} \sum_{j=1}^n \text{var}(Y_j) \leq (b_n \varepsilon)^{-2} \sum_{j=1}^n \mathbb{E}[Y_j^2] \rightarrow 0 \end{aligned}$$

另一方面有

$$\mathbb{P}[S_n \neq T_n] \leq \mathbb{P}\left[\bigcup_{j=1}^n \{X_j \neq Y_j\}\right] \leq \sum_{j=1}^n \mathbb{P}[|X_j| > b_n] \rightarrow 0$$

从而

$$\mathbb{P}\left[\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right] \leq \mathbb{P}[S_n \neq T_n] + \mathbb{P}\left[\left|\frac{T_n - a_n}{b_n}\right| > \varepsilon\right] \rightarrow 0$$

由此证得定理.

上述结论实则是一般化的弱大数定律, 为了不必验证充分条件而更方便地使用它, 我们在此基础上做进一步的推导.

**定理 3.7**

设  $X_1, X_2, \dots, X_n$  是 i.i.d. 随机变量, 当  $x \rightarrow \infty$  时有

$$x \mathbb{P}[|X_i| > x] \rightarrow 0$$

令  $S_n = X_1 + \dots + X_n$ ,  $\mu_n = \mathbb{E}[X_1 \mathbb{1}\{|X_1| \leq n\}]$ , 那么  $S_n/n - \mu_n \xrightarrow{p} 0$ .



**证明** 令  $a_n = n\mu_n$ ,  $b_n = n$ , 显然有

$$\sum_{j=1}^n \mathbb{P}[|X_j| > n] = n \mathbb{P}[|X_i| > n] \rightarrow 0$$

<sup>4</sup>对任意随机变量  $X$  和  $\varepsilon > 0$ , 总有  $\mathbb{P}[|X - \mathbb{E}[X]| > \varepsilon] < \frac{\text{var}(X)}{\varepsilon^2}$ .

当  $n \rightarrow \infty$  时, 根据期望恒等式<sup>5</sup>可知

$$\begin{aligned} b_n^{-2} \sum_{i=1}^n \mathbb{E}[X_i^2 I(|X_i| \leq b_n)] &= \frac{1}{n} \mathbb{E}[X^2 \mathbb{1}\{|X| \leq n\}] \leq \frac{1}{n} \mathbb{E}[\min\{|X|, n\}^2] \\ &= \frac{1}{n} \int_0^\infty 2x \mathbb{P}[\min\{|X|, n\} > x] dx = \frac{1}{n} \int_0^n 2x \mathbb{P}[|X| > x] dx \\ &= \frac{1}{n} \int_M^n 2x \mathbb{P}[|X| > x] dx + o(1) \leq 2 \sup_{x \geq M} x \mathbb{P}[|X| > x] + o(1) \end{aligned}$$

由于  $0 < M < n$  是任意的, 因而  $b_n^{-2} \sum_{i=1}^n \mathbb{E}[X_i^2 \mathbb{1}\{|X_i| \leq b_n\}] \rightarrow 0$ , 由定理3.6可知结论成立.

### 推论 3.1 (Khinchin 弱大数定律)

设  $X_1, X_2, \dots, X_n$  是 i.i.d. 随机变量,  $S_n = X_1 + \dots + X_n$ ,  $\mathbb{E}|X_1| < \infty$  且  $\mu = \mathbb{E}[X_1]$ , 那么  $S_n/n \xrightarrow{P} \mu$ .



**证明** 令  $Y = |X_1| \mathbb{1}\{|X_1| > x\}$ , 显然  $Y > x \mathbb{1}\{|X_1| > x\}$ , 两端取期望得

$$x \mathbb{P}[|X_1| > x] \leq \mathbb{E}[|X_1| \mathbb{1}\{|X_1| > x\}] = \int_x^\infty \mathbb{P}[|X_1| > y] dy$$

于是当  $x \rightarrow \infty$  时,  $x \mathbb{P}[|X_1| > x] \rightarrow 0$ . 另一方面

$$X_1 \mathbb{1}\{|X_1| \leq n\} \rightarrow X_1, \quad n \rightarrow \infty$$

又因为  $\mathbb{E}|X_1| < \infty$ , 根据控制收敛定理<sup>6</sup>(Dominated Convergence Theorem, DCT) 可知, 当  $n \rightarrow \infty$  时有

$$\mu_n = \mathbb{E}[X_1 \mathbb{1}\{|X_1| \leq n\}] \rightarrow \mathbb{E}[X_1] = \mu$$

再根据定理3.7, 对于任意的  $\varepsilon > 0$ , 当  $n \rightarrow \infty$  时有  $\mathbb{P}[|S_n/n - \mu_n| > \varepsilon/2] \rightarrow 0$ , 因此由三角不等式可知

$$\lim_{n \rightarrow \infty} \mathbb{P}[|S_n/n - \mu| > \varepsilon] = 0$$

由此证得推论.

**注** 定理3.7中不需要矩条件  $\mathbb{E}|X_1| < \infty$  成立<sup>7</sup>, 而 Khinchin 弱大数定律则需要这一条件. 另一方面, Khinchin 弱大数定律可以扩展到独立同分布的  $K \times 1$  维随机向量的情形, 只需将矩条件改为  $\mathbb{E}\|X_1\| < \infty$  即可.

**例 3.8 弱大数定律不成立的情况** 设  $X_1, X_2, \dots, X_n$  是独立随机变量, 并且都服从标准 Cauchy 分布, 也即

$$\mathbb{P}[X_1 \leq x] = \int_{-\infty}^x \frac{dt}{\pi(1+t^2)}$$

考虑条件

$$x \mathbb{P}[|X_1| > x] = 2x \int_x^\infty \frac{dt}{\pi(1+t^2)} = x \left(1 - \frac{2 \arctan x}{\pi}\right)$$

<sup>5</sup>对任意非负随机变量  $X$  和  $p > 0$ ,  $\mathbb{E}[X^p] = \int_0^\infty p x^{p-1} \mathbb{P}[X > x] dx$ .

<sup>6</sup>如果  $X_n \xrightarrow{a.s.} X$ , 对于一切  $n = 1, 2, \dots$  都有  $|X_n| \leq Y$ , 并且  $\mathbb{E}[Y] < \infty$ , 那么  $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ .

<sup>7</sup>它所需要的矩条件是更弱的  $\mathbb{E}|X_1|^{1-\varepsilon} < \infty$ , 其中  $0 < \varepsilon < 1$ . 因此不能说定理3.7取消了矩条件这一限制.

根据 L'Hospital 法则可知

$$\lim_{x \rightarrow \infty} x \left( 1 - \frac{2 \arctan x}{\pi} \right) = \lim_{t \rightarrow 0} \frac{2}{\pi(t^2 + 1)} = \frac{2}{\pi}$$

因此定理 3.7 失效. 另一方面

$$\mathbb{E}|X_1| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \lim_{x \rightarrow \infty} \frac{\log(1+x^2)}{\pi} = \infty$$

也意味着 Khinchin 弱大数定律中的矩条件不成立.

实际上, 随机变量  $\bar{X}_n = S_n/n$  的特征函数 (Characteristic Function, CF) 为

$$\phi_{\bar{X}_n}(t) = \mathbb{E}[\exp(it\bar{X}_n)] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(it\frac{X_i}{n}\right)\right] = \left[\exp\left(-\left|\frac{t}{n}\right|\right)\right]^n = \exp(-|t|)$$

因此  $S_n/n$  仍服从标准 Cauchy 分布, 而非依概率收敛于任何常数.

本节讨论的 LLN 都离不开 i.i.d. 这一条件, 后续在进行时间序列数据的计量分析时, 我们会使用遍历定理作为 LLN 的推广, 它允许随机变量间存在相关性.

### 3.3 中心极限定理

现在来看中心极限定理 (Central Limit Theorem, CLT). 我们首先给出 Lindeberg-Feller CLT, 它是一种非常重要且一般化的中心极限定理.

#### 定理 3.8 (Lindeberg-Feller 中心极限定理)

假设对于一切  $n \geq 1$  与  $j = 1, 2, \dots, r_n$ ,  $\{X_{nj}\}$  是独立随机变量构成的序列, 对于一切  $1 \leq j \leq r_n$  都有:

$$\mathbb{E}[X_{nj}] = 0, \quad 0 < \mathbb{E}[X_{nj}^2] \equiv \sigma_{nj}^2 < \infty$$

并且还满足 Lindeberg 条件:

$$\lim_{n \rightarrow \infty} s_n^{-2} \sum_{j=1}^{r_n} \mathbb{E}[X_{nj}^2 \mathbb{1}_{\{|X_{nj}| > \varepsilon s_n\}}] = 0 \quad (3.10)$$

那么当  $n \rightarrow \infty$  时有

$$S_n/s_n \xrightarrow{d} N(0, 1)$$

其中  $S_n = \sum_{j=1}^{r_n} X_{nj}$ , 并且  $s_n^2 = \sum_{j=1}^{r_n} \sigma_{nj}^2$ .



**证明** 见 Athreya and Lahiri (2006) Theorem 11.1.1.

#### 推论 3.2 (Lindeberg-Levy 中心极限定理)

设  $X_1, X_2, \dots, X_n$  为 i.i.d. 随机变量, 如果  $\mathbb{E}[X_1^2] < \infty$ , 那么

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

其中  $\bar{X}_n = n^{-1} \sum_{j=1}^n X_j$ ,  $\mathbb{E}[X_1] = \mu$ , 并且  $\text{var}(X_1) = \sigma^2 < \infty$ .



**证明** 令  $Z_j = (X_j - \mu)/\sigma\sqrt{n}$ ,  $S_n = \sum_{j=1}^n Z_j$ , 以及

$$s_n^2 = \sum_{j=1}^n \text{var}(Z_j) = 1$$

考虑验证 Lindeberg 条件 (3.10). 当  $n \rightarrow \infty$  时有

$$\begin{aligned} s_n^{-2} \sum_{j=1}^n \mathbb{E}[X_j^2 \mathbb{1}\{|X_j| > \varepsilon s_n\}] &= \sum_{j=1}^n \mathbb{E}\left[\left(\frac{X_j - \mu}{\sigma\sqrt{n}}\right)^2 \mathbb{1}\left\{\left|\frac{X_j - \mu}{\sigma\sqrt{n}}\right| > \varepsilon\right\}\right] \\ &= n \left\{ \frac{1}{\sigma^2 n} \mathbb{E}[(X_1 - \mu)^2 \mathbb{1}\{|X_1 - \mu| > \varepsilon\sigma\sqrt{n}\}] \right\} \\ &= \sigma^{-2} \mathbb{E}[(X_1 - \mu)^2 \mathbb{1}\{|X_1 - \mu| > \varepsilon\sigma\sqrt{n}\}] \rightarrow 0 \end{aligned}$$

根据 Lindeberg-Feller CLT 可知  $S_n/s_n \xrightarrow{d} N(0, 1)$ , 再由 CMT 即可推知

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

**注** 以上两个定理的区别在于, Lindeberg-Feller CLT 仅要求独立随机变量, 而 Lindeberg-Levy CLT 在此基础上还要求它们是同分布的.

**例 3.9** 正如例 3.8 提到的那样, 由于 Cauchy 分布的任意阶矩不存在, 因此无法对其使用 CLT, 其样本均值  $\bar{X}_n/n$  也服从标准 Cauchy 分布.

现在我们将要把 CLT 从一元推广至多元情形, 这里仍和前面一样, 需要随机变量序列至少是独立的. 在正式推导之前, 我们直接给出一个引理.

### 引理 3.1 (Cramér-Wold 方法)

设  $\{X_n\}$  是由  $K \times 1$  维随机向量构成的序列, 那么  $X_n \xrightarrow{d} X$ , 当且仅当对于一切  $\lambda \in \mathbb{R}^K$  都有  $\lambda' X_n \xrightarrow{d} \lambda' X$ .



**证明** 见 Athreya and Lahiri (2006) Theorem 10.4.5.

### 定理 3.9 (多元 Lindeberg-Feller 中心极限定理)

设  $n \geq 1$  与  $j = 1, 2, \dots, r_n$ ,  $\{X_{nj}\}$  是由  $K \times 1$  维独立随机向量构成的序列, 对于一切  $1 \leq j \leq r_n$  都有:

$$\mathbb{E}[X_{nj}] = 0, \quad v_n^2 = \lambda_{\min}(\bar{V}_n) > 0$$

并且对于任意  $\varepsilon > 0$  都有

$$\lim_{n \rightarrow \infty} v_n^{-2} \sum_{j=1}^{r_n} \mathbb{E}[||X_{nj}||^2 \mathbb{1}\{||X_{nj}|| \geq \varepsilon v_n\}] = 0 \quad (3.11)$$

那么当  $n \rightarrow \infty$  时

$$\bar{V}_n^{-1/2} \sum_{j=1}^{r_n} X_{nj} \xrightarrow{d} N(0, I_K)$$

其中  $\bar{V}_n = \sum_{j=1}^{r_n} V_{nj}$ , 并且  $V_{nj} = \mathbb{E}[X_{nj} X_{nj}']$ .





**证明** 任取  $\lambda \in \mathbb{R}^K$  使得  $\lambda' \lambda = 1$ . 定义  $U_{nj} = \lambda' \bar{V}_n^{-1/2} X_{nj}$ , 它的方差为

$$\sigma_{nj}^2 = \lambda' \bar{V}_n^{-1/2} V_{nj} \bar{V}_n^{-1/2} \lambda$$

并且

$$s_n^2 = \sum_{j=1}^{r_n} \sigma_{nj}^2 = \lambda' \lambda = 1$$

另一方面, 由 Cauchy-Schwarz 不等式和二次不等式可知

$$\|U_{nj}\|^2 \leq \lambda' \bar{V}_n^{-1} \lambda \|X_{nj}\|^2 \leq \frac{\|X_{nj}\|^2}{\lambda_{\min}(\bar{V}_n)} = \frac{\|X_{nj}\|^2}{v_n^2}$$

也即  $\|U_{nj}\| \leq \|X_{nj}\|/v_n$ . 由条件 (3.11), 当  $n \rightarrow \infty$  时

$$\begin{aligned} s_n^{-2} \sum_{j=1}^{r_n} \mathbb{E}[\|U_{nj}\|^2 \mathbb{1}\{U_{nj} \geq \varepsilon s_n\}] &= \sum_{j=1}^{r_n} \mathbb{E}[\|U_{nj}\|^2 \mathbb{1}\{U_{nj} \geq \varepsilon s_n\}] \\ &\leq v_n^{-2} \sum_{j=1}^{r_n} \mathbb{E}[\|X_{nj}\|^2 \mathbb{1}\{\|X_{nj}\| \geq \varepsilon v_n\}] \rightarrow 0 \end{aligned}$$

根据 Lindeberg-Feller CLT 可知

$$\sum_{j=1}^{r_n} U_{nj} = \lambda' \bar{V}_n^{-1/2} \sum_{j=1}^{r_n} X_{nj} \xrightarrow{d} \lambda' Z$$

其中  $Z \sim N(0, I_K)$ . 根据 Cramér-Wold 方法即可推知

$$\bar{V}_n^{-1/2} \sum_{j=1}^{r_n} X_{nj} \xrightarrow{d} N(0, I_K)$$

由此证得定理.

### 推论 3.3 (多元 Lindeberg-Levy 中心极限定理)

设  $X_1, X_2, \dots, X_n \in \mathbb{R}^K$  为 i.i.d. 随机向量, 如果  $\mathbb{E}\|X_1\|^2 < \infty$ , 那么当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, V)$$

其中  $\mu = \mathbb{E}[X_1]$ , 并且  $V = \mathbb{E}[(X_1 - \mu)(X_1 - \mu)']$ .



**证明** 与一维情形时的类似, 只需令

$$Z_j = \frac{1}{\sqrt{n}} V_j^{-1/2} X_j$$

然后验证 Lindeberg 条件 (3.11), 最后使用 CMT 即可.

最后我们介绍 Delta 方法, 它是概率论意义上的 Taylor 展开, 可以将计量经济学中遇到的光滑可微的非线性统计量线性化, 进而保证 CLT 的使用. 因此, 它可以看作为 CLT 从样本均值到非线性统计量的推广.

### 定理 3.10 (Delta 方法)

设  $g: \Theta \rightarrow \mathbb{R}^L$  为连续可微函数,  $\Theta \subset \mathbb{R}^K$ . 如果

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$$

并且  $\theta$  是  $\Theta$  的内点, 那么当  $n \rightarrow \infty$  时有

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} N(0, \mathbf{G} \mathbf{V} \mathbf{G}') \quad (3.12)$$

其中  $\mathbf{G} = \nabla g(\theta)$  为  $g$  的  $L \times K$  维 Jacobi 矩阵,  $L \leq K$ .



**证明** 根据中值定理可知

$$\begin{aligned} \sqrt{n}[g(\hat{\theta}_n) - g(\theta)] &= g'(\xi) \sqrt{n}(\hat{\theta}_n - \theta) \\ &= \mathbf{G} \sqrt{n}(\hat{\theta}_n - \theta) + [g'(\xi) - \mathbf{G}] \sqrt{n}(\hat{\theta}_n - \theta) \end{aligned}$$

其中  $\xi$  位于  $\hat{\theta}_n$  和  $\theta$  之间. 由于  $\hat{\theta}_n \xrightarrow{p} \theta$ , 故而当  $n \rightarrow \infty$  时  $\xi \xrightarrow{p} \theta$ , 进而由 CMT 可知  $g'(\xi) - \mathbf{G} = o_p(1)$ , 再根据定理 3.2(4) 可知  $\sqrt{n}(\hat{\theta}_n - \theta) = O_p(1)$ , 于是

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{p} \mathbf{G} \sqrt{n}(\hat{\theta}_n - \theta)$$

由 Slutsky 定理可知 (3.12) 成立.

**例 3.10** 假设  $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ ,  $\mu \neq 0$  且  $0 < \sigma < \infty$ , 现在要求出  $\sqrt{n}(\bar{X}_n^{-1} - \mu^{-1})$  的极限分布.

令  $g(\bar{X}_n) = \bar{X}_n^{-1}$ , 由于  $g$  在  $\mu$  处连续可微, 根据一阶 Taylor 展开可知

$$g(\bar{X}_n) = g(\mu) + g'(\bar{\mu}_n)(\bar{X}_n - \mu)$$

其中  $\bar{\mu}_n$  介于  $\bar{X}_n$  和  $\mu$  之间. 由于  $\bar{X}_n \xrightarrow{p} \mu$ , 因此当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\bar{X}_n^{-1} - \mu^{-1}) = -\frac{\sigma}{\bar{\mu}_n^2} \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2/\mu^4)$$

## 3.4 一致可积与矩收敛

本章最后一节给出随机变量序列的一致可积 (Uniformly Integrability, UI) 概念及其性质, 它对于重抽样方法十分重要. 首先, 随机变量  $X$  是可积的, 如果  $\mathbb{E}|X| < \infty$  成立, 或等价地有

$$\lim_{M \rightarrow \infty} \mathbb{E}[|X| \mathbb{1}\{|X| > M\}] = \lim_{M \rightarrow \infty} \int_M^\infty |x| dF(x) = 0$$

下面给出一致可积的概念. 注意, 可积性是关于随机变量的概念, 而一致可积性针对的是随机变量序列.

### 定义 3.2

随机变量序列  $\{X_n\}$  是一致可积的, 如果

$$\lim_{M \rightarrow \infty} \sup_{n \geq 1} \mathbb{E}[|X_n| \mathbb{1}\{|X_n| > M\}] = 0$$

随机向量序列  $\{X_n\}$  是一致可积的, 如果

$$\lim_{M \rightarrow \infty} \sup_{n \geq 1} \mathbb{E}[\|X_n\| \mathbb{1}\{\|X_n\| > M\}] = 0$$



**注** 随机变量序列的一致可积并不意味着序列内的元素都是可积的, 这一点与数学分析中的一

致连续和一致收敛概念具有很大不同.

根据定义可以直接判断, 如果序列  $\{X_n\}$  内的随机向量是独立同分布的, 并且  $\mathbb{E}\|X_1\| < \infty$ , 那么该序列一致可积.

一致可积是更一般化的概念, 对于非同分布的随机向量序列, 对其施加一定正则条件后, i.i.d. 序列的结果即可应用到一致可积的序列上. 例如, 如果各个  $X_n$  是独立的并构成一致可积序列, 那么 WLLN 仍然成立.

另一方面, 我们也可以对随机向量的幂运用一致可积性概念, 假设以下所用符号与多元 Lindeberg-Feller CLT 中的相同. 我们称  $\{X_n^2\}$  是一致平方可积的 (uniformly square integrable), 如果

$$\lim_{M \rightarrow \infty} \sup_{n \geq 1} \mathbb{E}[\|X_n\|^2 \mathbb{1}\{\|X_n\|^2 > M\}] = 0 \quad (3.13)$$

它是一个与 Lindeberg 条件相似, 但稍强一点的条件: 对于任意给定的  $\varepsilon > 0$ , 选取正数  $\delta$  使得  $0 < \delta < v_n^2$ . 定义  $X_n = Z_{nj}$ , 再选取充分大的  $M$  使得 (3.13) 成立, 也即

$$\sup_{n \geq 1} \mathbb{E}[\|X_n\|^2 \mathbb{1}\{\|X_n\|^2 > M\}] < \delta \varepsilon$$

因为当  $n \rightarrow \infty$  时,  $n \varepsilon v_n \rightarrow \infty$ , 因此

$$\frac{1}{n v_n^2} \sum_{j=1}^n \mathbb{E}[\|Z_{nj}\|^2 \mathbb{1}\{\|Z_{nj}\| > n \varepsilon v_n\}] \leq \frac{\delta \varepsilon}{v_n^2} < \varepsilon$$

也即 Lindeberg 条件成立.

直接按照定义来判断随机变量序列是否是一致可积的显然十分麻烦, 现在我们给出一致可积的 Liapunov 充分条件.

### 定理 3.11 (Liapunov 条件)

如果存在  $r > 1$ ,  $\mathbb{E}\|X_n\|^r \leq C < \infty$ , 那么  $\{X_n\}$  一致可积.



**证明** 任取  $\varepsilon > 0$ , 再取  $M \geq (C/\varepsilon)^{1/(r-1)}$ , 于是

$$\begin{aligned} \mathbb{E}[\|X_n\| \mathbb{1}\{\|X_n\| > M\}] &= \mathbb{E}\left[\frac{\|X_n\|^r}{\|X_n\|^{r-1}} \mathbb{1}\{\|X_n\| > M\}\right] \\ &\leq \frac{\mathbb{E}[\|X_n\|^r \mathbb{1}\{\|X_n\| > M\}]}{M^{r-1}} \\ &\leq \frac{\mathbb{E}\|X_n\|^r}{M^{r-1}} \leq \frac{C}{M^{r-1}} < \varepsilon \end{aligned}$$

由于上式对一切  $n$  成立, 因此对上确界也成立.

### 定理 3.12

如果随机向量序列  $\{\|X_i\|^r\}$  是一致可积的, 那么当  $n \rightarrow \infty$  时有

$$n^{-1/r} \max_{1 \leq i \leq n} \|X_i\| \xrightarrow{p} 0 \quad (3.14)$$



**证明** 任取  $\delta > 0$ , 事件  $\left\{\max_{1 \leq i \leq n} \|X_i\| > \delta n^{1/r}\right\}$  表明至少存在某个  $\|X_i\|$  大于  $\delta n^{1/r}$ , 故而它和事件  $\{\cup_{n=1}^{\infty} \{\|X_i\| > \delta n^{1/r}\}\}$  是等价的.

根据测度的有限次可加性得到

$$\begin{aligned}
 \mathbb{P}\left[n^{-1/r} \max_{1 \leq i \leq n} \|X_i\| > \delta\right] &= \mathbb{P}\left[\bigcup_{i=1}^n \{\|X_i\|^r > n\delta^r\}\right] \\
 &\leq \sum_{i=1}^n \mathbb{P}[\|X_i\|^r > n\delta^r] \\
 &\leq \frac{1}{n\delta^r} \sum_{i=1}^n \mathbb{E}[\|X_i\|^r \mathbb{1}_{\{\|X_i\|^r > n\delta^r\}}] \\
 &\leq \delta^{-r} \max_{1 \leq i \leq n} \mathbb{E}[\|X_i\|^r \mathbb{1}_{\{\|X_i\|^r > n\delta^r\}}]
 \end{aligned}$$

根据一致可积性, 当  $n\delta^r \rightarrow \infty$  时, 上式收敛于 0.

(3.14) 表明在一致可积性下, 当  $n \rightarrow \infty$  时,  $\|X_i\|$  最大的观测值收敛的速率低于  $n^{1/r}$ , 并且随着  $r$  增大, 收敛速率越慢.

有时我们对统计量的矩感兴趣, 假设  $X_1, X_2, \dots, X_n$  是均值为  $\mu$ , 方差为  $\sigma^2$  的 i.i.d. 样本, 定义标准化的样本均值统计量

$$Z_n = \sqrt{n}(\bar{X}_n - \mu)$$

则它的均值为 0, 方差为  $\sigma^2$ . 不仅如此, 还可以计算出  $Z_n$  的高阶矩并得到显式表达, 例如

$$\mathbb{E}[Z_n^3] = \frac{\kappa_3}{\sqrt{n}}, \quad \mathbb{E}[Z_n^4] = \frac{\kappa_4}{n} + 3\sigma^2$$

其中  $\kappa_n$  为  $X_1$  的  $n$  阶矩. 进一步, 如果存在  $X_1$  的有限  $r$  阶矩, 那么  $\mathbb{E}[Z_n] \rightarrow \mathbb{E}[Z]$ , 其中  $Z \sim N(0, \sigma^2)$ . 然而对于样本均值的非线性统计量, 则不一定能得到  $Z_n$  高阶矩的表达式.

另一方面, 之前我们给出了依分布收敛  $X_n \xrightarrow{d} X$ , 现在我们想知道如何得到  $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ , 下面先给出一个渐近分布的均值存在的充分条件.

### 定理 3.13

如果  $X_n \xrightarrow{d} X$ , 并且  $\mathbb{E}\|X_n\| \leq C$ , 那么  $\mathbb{E}\|X\| \leq C$ .



**证明** 这里只证明一维形式. 根据期望恒等式, 对于任意非负随机变量  $X$  都有

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > u] du \quad (3.15)$$

设  $F_n$  和  $F$  分别为  $|X_n|$  和  $|X|$  的累积分布, 根据 (3.15) 和 Fatou 引理<sup>8</sup>可知

$$\begin{aligned}
 \mathbb{E}|X| &= \int_0^\infty [1 - F(x)] dx = \int_0^\infty \lim_{n \rightarrow \infty} [1 - F(x)] dx \\
 &\leq \liminf_{n \rightarrow \infty} \int_0^\infty [1 - F_n(x)] dx = \liminf_{n \rightarrow \infty} \mathbb{E}|X_n| \leq C
 \end{aligned}$$

由此证得定理.

值得注意的是, 该定理并不意味着  $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ . 考虑随机变量  $X_n$ , 它的概率分布为  $\mathbb{P}[X_n = n] = 1/n$ , 并且  $\mathbb{P}[X_n = 0] = 1 - 1/n$ , 于是  $X_n \xrightarrow{d} X$  并且  $\mathbb{P}[X = 0] = 1$ , 显然  $\mathbb{E}[Z_n] = 1$  并不收敛于  $\mathbb{E}[Z] = 0$ , 然而此时 (3.15) 也成立. 究其原因, 这是因为累积分布序列

<sup>8</sup>如果  $X_n$  为非负随机变量, 那么  $\mathbb{E}\left[\liminf_{n \rightarrow \infty} X_n\right] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[X_n]$ .

$\{F_n\}$  缺少紧性 (tightness), 解决方法则是一致可积.

### 定理 3.14

如果  $X_n \xrightarrow{d} X$  并且  $\{X_n\}$  是一致可积的, 那么  $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ .



**证明** 这里只证明一维的形式. 记  $X_n = X_n^+ - X_n^-$ , 其中  $X_n^+ = \max\{X_n, 0\}$ ,  $X_n^- = -\min\{X_n, 0\}$ , 因此只需证明  $X_n$  非负的情况. 记  $a \wedge b = \min\{a, b\}$ .

任取  $\varepsilon > 0$ , 根据定理 3.13 可知  $X$  是可积的, 于是

$$\mathbb{E}[X - (X \wedge M)] = \mathbb{E}[(X - M)\mathbb{1}\{X > M\}] \leq \mathbb{E}[X\mathbb{1}\{X > M\}] < \varepsilon$$

根据条件, 随机变量序列  $\{X_n\}$  一致可积, 故而存在正数  $M < \infty$ , 使得对于充分大的  $n$  有

$$\mathbb{E}[X_n - (X_n \wedge M)] = \mathbb{E}[(X_n - M)\mathbb{1}\{X_n > M\}] \leq \mathbb{E}[X_n\mathbb{1}\{X_n > M\}] < \varepsilon$$

由于函数  $X_n \wedge M$  是连续有界的, 又因为  $X_n \xrightarrow{d} X$ , 根据 CMT 和 DCT 可知  $\mathbb{E}[X_n \wedge M] \rightarrow \mathbb{E}[X \wedge M]$ , 也即对充分大的  $n$  有

$$|\mathbb{E}[(X_n \wedge M) - (X \wedge M)]| < \varepsilon$$

根据三角不等式可知

$$\begin{aligned} |\mathbb{E}[X_n - X]| &\leq |\mathbb{E}[X_n - (X_n \wedge M)]| \\ &\quad + |\mathbb{E}[(X_n \wedge M) - (X \wedge M)]| \\ &\quad + |\mathbb{E}[X - (X \wedge M)]| < 3\varepsilon \end{aligned}$$

由此证得定理.

**注** 该定理同样可以推广至随机向量的情形. 另一方面, 根据定理 3.2 可知, 几乎必然收敛, 依  $L^p$  收敛和依概率收敛可以推出依分布收敛, 因此也能得到相同的结论.

可以证明, 随机变量序列  $\{X_n\}$  的一致可积等价于它在  $L^1(\Omega, \mathcal{F}, \mathbb{P})$  上一致有界且一致绝对连续. 另一方面, DCT 中被可测函数所控制的函数序列必定是一致可积的, 因此定理 3.14 是 DCT 的推广, 证明见 Davison (2020) Theorem 12.11.

## 第4章 最小二乘的渐近性质

第二章分析了 OLS 估计量的有限样本性质, OLS 在假设2.1–2.4下是 BLUE, 但是这个结论所需要的假设过于严格. 第三章我们主要介绍了渐近理论基础, 本章主要是渐近理论在 OLS 模型上的应用, 在大样本框架下推导其渐近性质.

### 4.1 OLS 的一致性

这一节将要证明, 在更为宽松的假设条件下, OLS 估计量是对线性投影系数的一致估计, 也即随着样本容量的增大,  $\hat{\beta}$  将会依概率收敛到线性投影系数  $\beta$ .

#### 定义 4.1

如果当  $n \rightarrow \infty$  时有  $\hat{\theta}_n \xrightarrow{P} \theta$ , 则称  $\hat{\theta}_n$  是  $\theta$  的 (弱) 一致估计量.

**例 4.1 有偏但一致的估计量** 设  $\{T_n\}$  为  $\theta$  的一系列估计量构成的集合,  $\delta \neq 0$ , 并且  $T_n$  具有以下概率分布

$$\begin{array}{ccc} T_n & \theta & n\delta + \theta \\ \mathbb{P} & 1 - \frac{1}{n} & \frac{1}{n} \end{array}$$

于是  $T_n \xrightarrow{P} \theta$ , 但是  $\mathbb{E}[T_n] = \theta + \delta$ .

在正式进行推导前, 我们先给出一系列正则条件.

#### 假设 4.1

- (1)  $\{Y_i, X_i\}_{i=1}^n$  是可观测的 i.i.d. 随机样本.
- (2)  $\mathbb{E}[Y_i^2] < \infty$ .
- (3)  $\mathbb{E}[|X_i|^2] < \infty$ .
- (4)  $Q_{XX} = \mathbb{E}[X_i X_i']$  为正定矩阵.

**注** 假设4.1或许和诸多计量经济学教材给出的假设均有所不同, 实际上这里给出的假设是较弱的充分条件.

#### 定理 4.1

在假设4.1下, 当  $n \rightarrow \infty$  时, OLS 估计量  $\hat{\beta} \xrightarrow{P} \beta$ , 也即  $\hat{\beta} = \beta + o_p(1)$ .

**证明** 首先写出  $\hat{\beta} - \beta$  的表达式

$$\begin{aligned} \hat{\beta} - \beta &= \left( n^{-1} \sum_{i=1}^n X_i X_i' \right)^{-1} \left( n^{-1} \sum_{i=1}^n X_i e_i \right) \\ &= \hat{Q}_{XX}^{-1} \hat{Q}_{Xe} \end{aligned}$$

一方面, 根据假设4.1(4) 可知 WLLN 所需要的矩条件成立, 也即

$$\mathbb{E}|X_{ji}X_{li}| \leq (\mathbb{E}[X_{ji}^2])^{1/2}(\mathbb{E}[X_{li}^2])^{1/2} < \infty, \quad \forall 1 \leq j, l \leq K$$

于是根据 WLLN 和 CMT 可知

$$\hat{\mathbf{Q}}_{XX}^{-1} \xrightarrow{p} (\mathbb{E}[X_i X_i])^{-1} = \mathbf{Q}_{XX}^{-1}$$

另一方面, 根据定理1.8(3) 和 WLLN 可知

$$\hat{\mathbf{Q}}_{Xe} \xrightarrow{p} \mathbb{E}[X_i e_i] = 0$$

由于函数  $g(\mathbf{A}, b) = \mathbf{A}^{-1}b$  对一切使得  $\mathbf{A}^{-1}$  存在的  $(\mathbf{A}, b)$  连续, 因此根据 CMT 得到

$$\hat{\beta} - \beta = \hat{\mathbf{Q}}_{XX}^{-1} \hat{\mathbf{Q}}_{Xe} \xrightarrow{p} \mathbf{Q}_{XX}^{-1} 0 = 0$$

由此证得定理.

**注** 假设4.1(2) 中提到了矩条件  $\mathbb{E}[Y_i^2] < \infty$  和  $\mathbb{E}\|X_i\|^2 < \infty$ , 尽管它们并没有在上述定理的证明中使用, 但却是证明定理1.8(3) 的必要条件. 从这个角度可以看出, 证明 OLS 估计量一致性的关键在于  $\mathbb{E}[X_i e_i] = 0$ , 也即每个回归元和误差项不相关.

参考 Hong (2020) 的方法, 我们可以将假设4.1(2), (3) 加强为更直观的条件:  $\mathbb{E}[e_i|X_i] = 0$  且  $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$ . 根据简单 LIE 可知

$$\mathbb{E}[X_i e_i] = \mathbb{E}[\mathbb{E}[X_i e_i|X_i]] = \mathbb{E}[X_i \mathbb{E}[e_i|X_i]] = 0$$

并且 WLLN 所需要的矩条件也成立, 从而可以得到 OLS 一致性的证明.

Amemiya (1985) 则提到了另一种证明 OLS 为一致估计的方法, 只需要当  $n \rightarrow \infty$  时

$$\lambda_{\min}(\mathbf{X}'\mathbf{X}) \rightarrow \infty \quad (4.1)$$

其中  $\lambda_{\min}(\mathbf{X}'\mathbf{X})$  表示矩阵  $\mathbf{X}'\mathbf{X}$  的最小特征根. 根据主成分分析, 特征值可以用来刻画矩阵的信息含量, 因此 (4.1) 表明随着样本容量的增加,  $\mathbf{X}'\mathbf{X}$  中包含的信息对于线性投影系数  $\beta$  的估计精度也会无限增加.

事实上, 根据二次不等式可知对于任意  $\tau \in \mathbb{R}^K$ , 如果  $\tau'\tau = 1$ , 那么当  $n \rightarrow \infty$  时

$$\begin{aligned} \tau' \text{var}(\hat{\beta}|\mathbf{X}) \tau &= \sigma^2 \tau' (\mathbf{X}'\mathbf{X})^{-1} \tau \\ &\leq \sigma^2 \lambda_{\max}[(\mathbf{X}'\mathbf{X})^{-1}] \\ &\leq \sigma^2 \lambda_{\min}^{-1}(\mathbf{X}'\mathbf{X}) \rightarrow 0 \end{aligned}$$

也即 OLS 估计量  $\hat{\beta}$  的条件方差将趋近于 0, 从而依概率收敛到真实参数.

## 4.2 OLS 的渐近正态性

上一节分析了 OLS 估计量  $\hat{\beta}$  的一致性, 但是一致性无法具体描述  $\hat{\beta}$  在大样本下的渐近分布. 为了应用 CLT, 假设4.1不够充分, 为此我们给出以下更强的假设条件.



**假设 4.2**

- (1)  $\{Y_i, X_i\}_{i=1}^n$  是可观测的 i.i.d. 随机样本.
- (2)  $\mathbb{E}[Y_i^4] < \infty$ .
- (3)  $\mathbb{E}\|X_i\|^4 < \infty$ .
- (4)  $\mathbf{Q}_{XX} = \mathbb{E}[X_i X_i']$  为正定矩阵.



在假设4.2中, 我们要求  $Y$  和  $X$  的四阶矩有限, 由于高阶矩有限意味着低阶矩也有限, 因此假设4.2可以推出假设4.1.

**定理 4.2**

在假设4.2下, 协方差矩阵

$$\mathbf{\Omega} \equiv \mathbb{E}[X_i X_i' e_i^2]$$

是有限的, 并且当  $n \rightarrow \infty$  时

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i e_i \xrightarrow{d} N(0, \mathbf{\Omega})$$



**证明** 首先, 根据谱范数的性质, 对任意实矩阵都有

$$\|X_i X_i'\|^{1/2} = \|X_i\|$$

在假设4.2下, 根据 Cauchy-Schwarz 不等式可知

$$\|\mathbf{\Omega}\| \leq \mathbb{E}\|X_i X_i' e_i^2\| = \mathbb{E}[\|X_i\|^2 e_i^2] \leq (\mathbb{E}\|X_i\|^4)^{1/2} (\mathbb{E}[e_i^4])^{1/2} < \infty$$

因为样本  $\{Y_i, X_i\}$  是 i.i.d. 的, 故而  $\{X_i e_i\}$  也是, 又因为  $\mathbb{E}[X_i e_i] = 0$ , 根据多元 Lindeberg-Levy CLT 可知

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i e_i = \sqrt{n} \hat{\mathbf{Q}}_{XX} \xrightarrow{d} N(0, \mathbf{\Omega})$$

其中  $\hat{\mathbf{Q}}_{XX} = n^{-1} \sum_{i=1}^n X_i X_i'$ .

**定理 4.3**

在假设4.2下, 当  $n \rightarrow \infty$  时

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta) \quad (4.2)$$

其中  $\mathbf{Q}_{XX} = \mathbb{E}[X_i X_i']$ ,  $\mathbf{\Omega} = \mathbb{E}[X_i X_i' e_i^2]$ , 并且

$$\mathbf{V}_\beta = \mathbf{Q}_{XX}^{-1} \mathbf{\Omega} \mathbf{Q}_{XX}^{-1}$$



**证明** 注意到

$$\sqrt{n}(\hat{\beta} - \beta) = \left( \frac{1}{n} \sum_{i=1}^n X_i X_i' \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i e_i \right)$$

由于

$$\hat{\mathbf{Q}}_{XX}^{-1} \xrightarrow{p} \mathbf{Q}_{XX}^{-1}$$

其中  $\hat{Q}_{XX} = n^{-1} \sum_{i=1}^n X_i X_i'$ . 再根据定理 4.2 与 Slutsky 定理可知 (4.2) 成立.

**注** 该定理表明  $\hat{\beta} = \beta + O_p(n^{-1/2})$ , 这是比一致性更强的结论.

通常来说, 我们称  $\text{avar}[\sqrt{n}(\hat{\beta} - \beta)] = V_\beta$  为  $\sqrt{n}(\hat{\beta} - \beta)$  的渐近协方差矩阵,  $V_\beta$  的表达式  $Q_{XX}^{-1} \Omega Q_{XX}^{-1}$  呈现出矩阵  $\Omega$  被  $Q_{XX}^{-1}$  所夹住的夹心 (sandwich) 形式.

现在来看  $\hat{\beta}$  在有限样本下的条件方差

$$V_{OLS} = (X'X)^{-1}(X'\Sigma X)(X'X)^{-1} \quad (4.3)$$

由于  $V_{OLS}$  是  $\hat{\beta}$  精确的条件方差, 而  $V_\beta$  是  $\sqrt{n}(\hat{\beta} - \beta)$  的渐近方差, 那么可以得出  $V_\beta \approx nV_{OLS}$ . 现在将 (4.3) 乘以  $n$  倍得到

$$nV_{OLS} = \left(\frac{1}{n}X'X\right)^{-1} \left(\frac{1}{n}X'\Sigma X\right)^{-1} \left(\frac{1}{n}X'X\right)^{-1}$$

当  $n \rightarrow \infty$  时, 确实有  $nV_{OLS} \xrightarrow{p} V_\beta$ .

在某些特殊情况下,  $\Omega$  和  $V_\beta$  的表达式可以简化, 条件为

$$\text{cov}(X_i X_i', e_i^2) = 0 \quad (4.4)$$

此时

$$\Omega = \mathbb{E}[X_i X_i'] \mathbb{E}[e_i^2] = \sigma^2 Q_{XX}$$

$$V_\beta = Q_{XX}^{-1} \Omega Q_{XX}^{-1} = \sigma^2 Q_{XX}^{-1} \equiv V_\beta^0$$

其中  $\sigma^2 = \mathbb{E}[e_i^2]$ . 我们称  $V_\beta^0$  是  $\sqrt{n}(\hat{\beta} - \beta)$  在同方差下的渐近协方差矩阵.

值得注意的是, 尽管我们证明了  $\sqrt{n}(\hat{\beta} - \beta)$  服从渐近正态分布, 但对于任意固定的样本容量  $n$ ,  $\sqrt{n}\hat{\beta}$  的分布也可能与正态分布相距甚远. 随着样本容量  $n$  增大,  $\sqrt{n}\hat{\beta}$  将会越来越接近正态分布, 但多大的  $n$  才能使这种逼近足够好? 仍然没有简单确切的答案. 事实上, 不论样本容量  $n$  有多大, 对于某些满足假设 4.2 的数据, 正态分布的逼近程度可以任意差.

最后我们来看分块回归的情形. 首先将矩阵分割为  $X' = [X_1', X_2']$  及  $\beta = [\beta_1', \beta_2']'$ , 此时

$$\begin{aligned} Y &= X'\beta + e \\ &= X_1'\beta_1 + X_2'\beta_2 + e \end{aligned}$$

再做分割

$$Q_{XX} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}$$

根据 (1.22) 可知

$$Q_{XX}^{-1} = \begin{bmatrix} Q_{11.2}^{-1} & -Q_{11.2}^{-1} Q_{12} Q_{22}^{-1} \\ -Q_{22.1}^{-1} Q_{21} Q_{11}^{-1} & Q_{22.1}^{-1} \end{bmatrix}$$

其中  $Q_{11.2} = Q_{11} - Q_{12} Q_{22}^{-1} Q_{21}$ , 以及  $Q_{22.1} = Q_{22} - Q_{21} Q_{11}^{-1} Q_{12}$ ,  $\Omega$  的各分块由  $Q_{XX}$  的各分块决定. 最终可以将渐近协方差矩阵  $V_\beta$  记作

$$V_\beta = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

可以证明

$$\begin{aligned} V_{11} &= Q_{11 \cdot 2}^{-1}(\Omega_{11} - Q_{12} Q_{22}^{-1} \Omega_{21} - \Omega_{12} Q_{22}^{-1} Q_{21} + Q_{12} Q_{22}^{-1} \Omega_{22} Q_{22}^{-1} Q_{21}) Q_{11 \cdot 2}^{-1} \\ V_{21} &= Q_{22 \cdot 1}^{-1}(\Omega_{21} - Q_{21} Q_{11}^{-1} \Omega_{11} - \Omega_{22} Q_{22}^{-1} Q_{21} + Q_{21} Q_{11}^{-1} \Omega_{12} Q_{22}^{-1} Q_{21}) Q_{11 \cdot 2}^{-1} \\ V_{22} &= Q_{22 \cdot 1}^{-1}(\Omega_{22} - Q_{21} Q_{11}^{-1} \Omega_{12} - \Omega_{21} Q_{11}^{-1} Q_{12} + Q_{21} Q_{11}^{-1} \Omega_{12} Q_{11}^{-1} Q_{12}) Q_{22 \cdot 1}^{-1} \end{aligned}$$

### 4.3 渐近方差估计量

利用大数定律, 我们先将证明在假设4.1下, 估计量  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$  与  $s^2 = e'e/(n-K)$  是对  $\sigma^2$  的一致估计.

#### 定理 4.4

在假设4.1下, 当  $n \rightarrow \infty$  时有  $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$  及  $s^2 \xrightarrow{p} \sigma^2$ .



**证明** 首先将 OLS 估计残差写为

$$\hat{e}_i = Y_i - X_i' \hat{\beta} = e_i - X_i'(\hat{\beta} - \beta)$$

从而

$$\hat{e}_i^2 = e_i^2 - 2e_i X_i'(\hat{\beta} - \beta) + (\hat{\beta} - \beta)' X_i X_i'(\hat{\beta} - \beta) \quad (4.5)$$

以及

$$\begin{aligned} \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n e_i^2 - 2 \left( n^{-1} \sum_{i=1}^n e_i X_i' \right) (\hat{\beta} - \beta) \\ &\quad + (\hat{\beta} - \beta)' \left( n^{-1} \sum_{i=1}^n X_i X_i' \right) (\hat{\beta} - \beta) \end{aligned} \quad (4.6)$$

根据 WLLN 可知

$$\begin{aligned} n^{-1} \sum_{i=1}^n e_i^2 &\xrightarrow{p} \sigma^2 \\ n^{-1} \sum_{i=1}^n e_i X_i' &\xrightarrow{p} \mathbb{E}[e_i X_i'] = 0 \\ n^{-1} \sum_{i=1}^n X_i X_i' &\xrightarrow{p} \mathbb{E}[X_i X_i'] = Q_{XX} \end{aligned}$$

根据 OLS 估计量的一致性可知 (4.6) 依概率收敛于  $\sigma^2$ . 由于当  $n \rightarrow \infty$  时,  $\frac{n}{n-K} \rightarrow 1$ , 因此

$$s^2 = \frac{n}{n-K} \hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

由此证得定理.

定理4.3表明  $\sqrt{n}(\hat{\beta} - \beta)$  服从渐近正态分布并且协方差矩阵为  $V_\beta$ , 出于渐近统计推断的目的, 我们需要得到  $V_\beta$  的一致估计量. 下面将分同方差和异方差的情形进行讨论.

**定理 4.5**

在假设4.1下, 当  $n \rightarrow \infty$  时有

$$\hat{V}_\beta^0 \xrightarrow{p} V_\beta^0$$

其中  $\hat{V}_\beta^0 = s^2 \hat{Q}_{XX}^{-1}$ .



**证明** 根据定理4.4可知  $s^2 \xrightarrow{p} \sigma^2$ , 又因为  $\hat{Q}_{XX} \xrightarrow{p} Q_{XX}$ , 根据 CMT 即可得出结论.

**注** 该定理并不要求回归满足同方差条件, 无论是同方差还是异方差总有  $\hat{V}_\beta^0 \xrightarrow{p} V_\beta^0$ , 只是说在同方差条件下有  $\text{avar}(\sqrt{n}\hat{\beta}) = \sigma^2 Q_{XX}^{-1}$ , 我们才选择  $\hat{V}_\beta^0$  作为它的一致估计量.

现在考虑异方差的情形, 此时协方差矩阵  $V_\beta = Q_{XX}^{-1} \Omega Q_{XX}^{-1}$  无法进一步化简. 正如前面提到的  $\hat{Q}_{XX}^{-1} \xrightarrow{p} Q_{XX}^{-1}$ , 我们只需得到  $\Omega$  的一致估计量即可. 首先有

$$\hat{\Omega}^{\text{ideal}} = \frac{1}{n} \sum_{i=1}^n X_i X_i' e_i^2 \xrightarrow{p} \Omega$$

由于  $e_i$  不可观测, 一个自然的想法是用  $\hat{e}_i$  替代  $e_i$  并得到

$$\begin{aligned} \hat{V}_\beta^{\text{HCO}} &= \hat{Q}_{XX}^{-1} \hat{\Omega} \hat{Q}_{XX}^{-1} \\ &= n(X'X)^{-1} \left( \sum_{i=1}^n X_i X_i' \hat{e}_i^2 \right) (X'X)^{-1} \end{aligned} \quad (4.7)$$

幸运的是, 这样的想法是可行的.

**定理 4.6**

在假设4.2下, 当  $n \rightarrow \infty$  时有  $\hat{\Omega} \xrightarrow{p} \Omega$  以及

$$\hat{V}_\beta^{\text{HCO}} \xrightarrow{p} V_\beta$$



**证明** 首先将  $\hat{\Omega}$  写为

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n X_i X_i' e_i^2 + \frac{1}{n} \sum_{i=1}^n X_i X_i' (\hat{e}_i^2 - e_i^2)$$

在假设4.2下, 相关矩条件成立, 于是根据 WLLN 可知上式第一项

$$\frac{1}{n} \sum_{i=1}^n X_i X_i' e_i^2 \xrightarrow{p} \mathbb{E}[X_i X_i' e_i^2] = \Omega$$

为了证明  $\hat{\Omega}$  是  $\Omega$  的一致估计量, 我们需要证明当  $n \rightarrow \infty$  时有

$$\frac{1}{n} \sum_{i=1}^n X_i X_i' (\hat{e}_i^2 - e_i^2) \xrightarrow{p} 0$$

首先可以得到

$$\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i' (\hat{e}_i^2 - e_i^2) \right\| \leq \frac{1}{n} \sum_{i=1}^n \|X_i X_i' (\hat{e}_i^2 - e_i^2)\| = \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 |\hat{e}_i^2 - e_i^2| \quad (4.8)$$

再对 (4.5) 运用三角不等式和 Schwarz 不等式<sup>1</sup>得到

$$\begin{aligned}
 |\hat{e}_i^2 - e_i^2| &\leq 2|e_i X_i'(\hat{\beta} - \beta)| + |(\hat{\beta} - \beta)' X_i X_i'(\hat{\beta} - \beta)| \\
 &= 2|e_i| |X_i'(\hat{\beta} - \beta)| + |(\hat{\beta} - \beta)' X_i|^2 \\
 &\leq 2|e_i| \|X_i\| \|\hat{\beta} - \beta\| + \|X_i\|^2 \|\hat{\beta} - \beta\|^2
 \end{aligned} \tag{4.9}$$

将 (4.8) 和 (4.9) 结合得到

$$\begin{aligned}
 \left\| \frac{1}{n} \sum_{i=1}^n X_i X_i'(\hat{e}_i^2 - e_i^2) \right\| &\leq 2 \left( \frac{1}{n} \sum_{i=1}^n \|X_i\|^3 |e_i| \right) \|\hat{\beta} - \beta\| + \left( \frac{1}{n} \sum_{i=1}^n \|X_i\|^4 \right) \|\hat{\beta} - \beta\|^2 \\
 &= o_p(1)
 \end{aligned} \tag{4.10}$$

也即  $\hat{\Omega} \xrightarrow{p} \Omega$ .

**注** 式 (4.10) 成立所需要的矩条件

$$\mathbb{E}[\|X_i\|^3 |e_i|] \leq (\mathbb{E}\|X_i\|^4)^{3/4} (\mathbb{E}[e^4])^{1/4} < \infty$$

是根据 Hölder 不等式<sup>2</sup>得到的, 故而  $n^{-1} \sum_{i=1}^n \|X_i\|^3 |e_i| = O_p(1)$ , 因此上式不等号右端的第一项是  $o_p(1)$ .

下面我们将用另一种方法证明一个比  $\hat{\Omega} = \Omega + o_p(1)$  更强的结论. 首先由 Schwarz 不等式可知

$$|\hat{e}_i - e_i| = |X_i(\hat{\beta} - \beta)| \leq \|X_i\| \|\hat{\beta} - \beta\| \tag{4.11}$$

我们先来得到证明上式的有界性.

考虑矩条件  $\mathbb{E}\|X_i\|^r < \infty$ , 根据 Liapunov 充分条件可知随机变量序列  $\{X_n\}$  一致可积, 再根据定理 3.12 得到

$$n^{-1/r} \max_{1 \leq i \leq n} \|X_i\| \xrightarrow{p} 0$$

又因为  $\|\hat{\beta} - \beta\| = O_p(n^{-1/2})$ , 从而

$$\max_{1 \leq i \leq n} |\hat{e}_i - e_i| \leq \max_{1 \leq i \leq n} \|X_i\| \|\hat{\beta} - \beta\| = o_p(n^{-1/2+1/r}) \tag{4.12}$$

由于假设 4.2 要求  $r \geq 4$ , 因此  $q_n = \max_{1 \leq i \leq n} |\hat{e}_i - e_i|$  的收敛率至少为  $o_p(n^{-1/4})$ , 并且随着  $r$  增加, 收敛越快. 进一步根据 (4.12) 得到

$$\begin{aligned}
 \left\| \frac{1}{n} \sum_{i=1}^n X_i X_i'(\hat{e}_i^2 - e_i^2) \right\| &\leq \frac{1}{n} \sum_{i=1}^n \|X_i X_i'\| |\hat{e}_i^2 - e_i^2| \\
 &\leq \frac{2}{n} \sum_{i=1}^n \|X_i\|^2 |e_i| |\hat{e}_i - e_i| + \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 |\hat{e}_i - e_i|^2 \\
 &\leq \frac{2}{n} \sum_{i=1}^n \|X_i\|^2 |e_i| q_n + \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 q_n^2 \leq o_p(n^{-1/4})
 \end{aligned}$$

显然可以推出  $\hat{\Omega} = \Omega + o_p(1)$ .

<sup>1</sup>对于任意  $K \times 1$  维向量  $a$  和  $b$ , 总有  $|a'b| \leq \|a\| \|b\|$ .

<sup>2</sup>设  $p, q > 1$  且  $1/p + 1/q = 1$ , 则对于任意  $K \times L$  维矩阵, 总有  $\mathbb{E}\|X'Y\| \leq (\mathbb{E}\|X\|^p)^{1/p} (\mathbb{E}\|Y\|^q)^{1/q}$ .

本节的最后我们来考虑  $V_\beta$  的其它估计量  $\hat{V}_\beta^{\text{HC1}}$ ,  $\hat{V}_\beta^{\text{HC2}}$  及  $\hat{V}_\beta^{\text{HC3}}$ . 其中

$$\hat{V}_\beta^{\text{HC1}} = \frac{n}{n-K} \hat{V}_\beta^{\text{HC0}}$$

由于  $n \rightarrow \infty$  时有  $n/(n-K) \rightarrow 1$ , 于是  $\hat{V}_\beta^{\text{HC1}} \xrightarrow{p} V_\beta$ .

$\hat{V}_\beta^{\text{HC2}}$  与  $\hat{V}_\beta^{\text{HC3}}$  则与 (4.7) 的形式一样, 只是将  $\hat{\Omega}$  分别替换为

$$\bar{\Omega} = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-1} X_i X_i' \hat{e}_i^2$$

以及

$$\tilde{\Omega} = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-2} X_i X_i' \hat{e}_i^2$$

为了证明  $\hat{V}_\beta^{\text{HC2}}$  与  $\hat{V}_\beta^{\text{HC3}}$  是  $V_\beta$  的一致估计量, 只需证明当  $n \rightarrow \infty$  时,  $\bar{\Omega} - \hat{\Omega}$  与  $\tilde{\Omega} - \hat{\Omega}$  依概率收敛于 0. 在此之前我们先证明一个引理.

#### 引理 4.1

如果  $\{X_n\}$  为 i.i.d. 随机变量序列,  $Q_{XX}$  为正定矩阵, 并且对于某个  $r \geq 2$  有  $\mathbb{E}\|X\|^r < \infty$ , 那么  $\max_{1 \leq i \leq n} h_{ii} = o_p(n^{2/r-1})$ .



**证明** 因为  $\hat{Q}_{XX} \xrightarrow{p} Q_{XX}$ ,  $Q_{XX}$  正定, 根据 CMT 可知

$$\lambda_{\min}(\hat{Q}_{XX}) \xrightarrow{p} \lambda_{\min}(Q_{XX}) > 0$$

根据二次不等式可知

$$\begin{aligned} h_{ii} &= X_i(X'X)^{-1}X_i' \\ &\leq \lambda_{\max}[(X'X)^{-1}](X_iX_i') \\ &= \lambda_{\min}[n^{-1}(X'X)]^{-1}n^{-1}\|X_i\|^2 \\ &\leq [\lambda_{\min}(Q_{XX}) + o_p(1)]^{-1}n^{-1} \max_{1 \leq i \leq n} \|X_i\|^2 \end{aligned} \quad (4.13)$$

根据定理 3.12 可知

$$\max_{1 \leq i \leq n} \|X_i\|^2 = \left( \max_{1 \leq i \leq n} \|X_i\| \right)^2 = o_p(n^{2/r})$$

联系 (4.13) 即可证得定理.

#### 定理 4.7

在假设 4.2 下, 当  $n \rightarrow \infty$  时有  $\bar{\Omega} \xrightarrow{p} \Omega$  与  $\tilde{\Omega} \xrightarrow{p} \Omega$ .



**证明** 根据假设 4.2 和引理 4.1 可知

$$h_n^* = \max_{1 \leq i \leq n} h_{ii} = o_p(1)$$

从而

$$\begin{aligned}\|\bar{\Omega} - \hat{\Omega}\| &\leq \frac{1}{n} \sum_{i=1}^n \|X_i X_i'\| |\hat{e}_i^2| (1 - h_{ii})^{-1} - 1| \\ &\leq \left( \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \hat{e}_i^2 \right) |(1 - h_n^*)^{-1} - 1| \quad (4.14)\end{aligned}$$

按照之前的做法可以得出 (4.14) 不等号右端括号内的部分依概率收敛于  $\mathbb{E}[\|X_i\|^2 e_i^2]$ , 绝对值内的部分为  $o_p(1)$ , 因此 (4.14) 也是  $o_p(1)$ , 从而  $\bar{\Omega} = \hat{\Omega} + o_p(1) \xrightarrow{p} \Omega$ .

类似可以证明

$$\begin{aligned}\|\tilde{\Omega} - \hat{\Omega}\| &\leq \frac{1}{n} \sum_{i=1}^n \|X_i X_i'\| |\hat{e}_i^2| (1 - h_{ii})^{-2} - 1| \\ &\leq \left( \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \hat{e}_i^2 \right) |(1 - h_n^*)^{-2} - 1| = o_p(1)\end{aligned}$$

因此定理成立.

根据以上论述, 异方差稳健的协方差矩阵估计量 HC0, HC1, HC2 及 HC3 都是  $V_\beta$  的一致估计量, 由于它们在大样本下具有相同的渐近分布, 后面我们统一使用  $\hat{V}_\beta$  来表示  $V_\beta$  的一致估计量.

值得注意的是, 尽管  $\hat{V}_\beta$  是渐近方差  $\text{avar}(\sqrt{n}\hat{\beta})$  的一致估计量, 但 “ $\hat{V}_\beta/n$  是  $\text{avar}(\hat{\beta})$  的一致估计” 这一说法却没有什么意义, 因为当  $n \rightarrow \infty$  时  $V/n \rightarrow 0$ . 而不论  $\hat{V}_\beta$  是不是  $V_\beta$  的一致估计, 总有  $\hat{V}_\beta \xrightarrow{p} 0$ .

## 4.4 参数检验

不同于第二章的基于正态随机扰动项的线性假设检验, 本节的内容无需依靠这一过于苛刻的条件, 而是基于大样本理论给出的渐近统计量进行统计推断, 不仅如此, 本节考虑的是一般的非线性假设.

### 4.4.1 回归系数的函数

通常而言, 研究者会对参数向量  $\beta = [\beta_0, \beta_1, \dots, \beta_k]$  的某个具体形式感兴趣, 举例来说就是像单个系数  $\beta_j$  或  $\beta_j/\beta_l$  这样关于  $\beta$  的某个函数  $\theta = R(\beta)$ , 其中  $R: \mathbb{R}^K \rightarrow \mathbb{R}^J$  为某个函数.  $\theta$  的估计量为

$$\hat{\theta} = R(\hat{\beta})$$

如果  $R$  是连续函数, 那么根据 CMT 和  $\hat{\beta} \xrightarrow{p} \beta$  这一事实可以推知  $\hat{\theta}$  是  $\theta$  的一致估计量, 现将其表述为以下定理.

#### 定理 4.8

在假设 4.1 下, 如果当  $n \rightarrow \infty$  时,  $R(\beta)$  在真实参数  $\beta$  处连续, 那么  $\hat{\theta} \xrightarrow{p} \theta$ .





不仅如此, 如果函数  $R : \mathbb{R}^K \rightarrow \mathbb{R}^J$  足够光滑, 由 Delta 法可以得到  $\hat{\theta}$  的渐近正态性, 为此我们给出一个新的假设.

**假设 4.3**

非随机的可测向量值函数  $R : \mathbb{R}^K \rightarrow \mathbb{R}^J$  在真实参数  $\beta$  处连续可微, 并且矩阵  $R = \nabla_{\beta} R(\beta)$  的秩为  $J$ .

**定理 4.9**

在假设 4.2 和 4.3 下, 当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_{\theta}) \quad (4.15)$$

其中  $V_{\theta} = R V_{\beta} R'$ .

出于统计推断的目的, 我们需要渐近协方差矩阵  $V_{\theta} = R V_{\beta} R'$  的一致估计量. 一个直觉是先利用估计量

$$\hat{R} = \nabla_{\beta} R(\hat{\beta}) = \frac{\partial}{\partial \beta} R(\hat{\beta}) \quad (4.16)$$

然后得到  $V_{\theta}$  的估计量  $\hat{V}_{\theta} = \hat{R} \hat{V}_{\beta} \hat{R}'$ , 称其为渐近协方差矩阵估计量. 特别地, 在随机扰动项满足同方差的条件下有

$$\hat{V}_{\theta}^0 = \hat{R} \hat{V}_{\beta}^0 \hat{R}' = s^2 \hat{R} \hat{Q}_{XX}^{-1} \hat{R}'$$

现在我们简单给出这一直觉是正确的证明.

**定理 4.10**

在假设 4.2 和 4.3 下, 当  $n \rightarrow \infty$  时有  $\hat{V}_{\theta} \xrightarrow{p} V_{\theta}$ .

**证明** 根据定理 4.6 可知  $\hat{V}_{\beta} \xrightarrow{p} V_{\beta}$ , 又根据  $\hat{\beta} \xrightarrow{p} \beta$  和  $R = \nabla_{\beta} R(\beta)$  在  $\beta$  处的连续性可知

$$\hat{R} = \nabla_{\beta} R(\hat{\beta}) \xrightarrow{p} \nabla_{\beta} R(\beta) = R$$

于是由 CMT 即可推得定理成立.

**4.4.2  $T$  检验和 Wald 检验**

现在我们想要检验原假设  $\mathbb{H}_0 : R(\beta) = r$ , 这里的  $r$  为一个  $J \times 1$  维非随机向量. 类似第二章的做法, 我们按  $J = 1$  和  $J > 1$  的情况进行分类讨论, 其中  $J = 1$  时的假设对应  $T$  统计量, 而  $J > 1$  时的假设对应 Wald 统计量 (不是之前的  $F$  统计量).

在正式开始之前, 我们还需要一个技术性的假设.

**假设 4.4**

矩阵  $V_{\theta} = R V_{\beta} R'$  正定.

先来看  $J = 1$  的情况, 也即单个约束条件下的参数检验.

**定理 4.11 (渐近  $T$  统计量)**

在假设4.2, 4.3和4.4下, 如果原假设  $\mathbb{H}_0 : R(\beta) = r$  成立, 那么当  $n \rightarrow \infty$  时有

$$T = \frac{\sqrt{n}[R(\hat{\beta}) - r]}{\sqrt{\hat{\mathbf{R}} \hat{\mathbf{V}}_{\beta} \hat{\mathbf{R}}'}} \xrightarrow{d} N(0, 1)$$



**证明** 在原假设  $\mathbb{H}_0$  成立的条件下

$$\begin{aligned} \sqrt{n}[R(\hat{\beta}) - r] &= \sqrt{n}[R(\beta) - r] + \nabla_{\beta} R(\bar{\beta}) \sqrt{n}(\hat{\beta} - \beta) \\ &= \nabla_{\beta} R(\bar{\beta}) \sqrt{n}(\hat{\beta} - \beta) \end{aligned}$$

这里的  $\bar{\beta}$  介于  $\hat{\beta}$  和  $\beta$  之间.

由于  $\mathbf{R} : \mathbb{R}^K \rightarrow \mathbb{R}^J$  连续可微, 并且当  $n \rightarrow \infty$  时有  $\hat{\beta} \xrightarrow{p} \beta$ , 故此时也有

$$\nabla_{\beta} R(\bar{\beta}) \xrightarrow{p} \nabla_{\beta} R(\beta)$$

又根据定理4.3可知

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{V}_{\beta})$$

因此根据 Slutsky 定理有

$$\sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} N(0, \mathbf{R} \mathbf{V}_{\beta} \mathbf{R}')$$

再由定理4.10可得  $\hat{\mathbf{V}}_{\theta} \xrightarrow{p} \mathbf{V}_{\theta}$ , 最后再使用一次 Slutsky 定理即可证得定理.

**注** 在渐近  $T$  检验中, 分母是一个标量, 它被称为  $\hat{\theta}$  的标准误 (standard error).

在  $J > 1$  的多重约束条件下, 上述渐近  $T$  检验将被拓展为 Wald 检验, 但它们在本质上是同样的.

**定理 4.12 (Wald 检验统计量)**

在假设4.2, 4.3和4.4下, 如果原假设  $\mathbb{H}_0 : R(\beta) = r$  成立, 那么当  $n \rightarrow \infty$  时有

$$W = n[R(\hat{\beta}) - r]' [\hat{\mathbf{R}} \hat{\mathbf{V}}_{\beta} \hat{\mathbf{R}}']^{-1} [R(\hat{\beta}) - r] \xrightarrow{d} \chi_J^2$$



**证明** 按照之前的做法可以得到

$$\sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} Z \sim N(0, \mathbf{R} \mathbf{V}_{\beta} \mathbf{R}')$$

根据引理2.3可知  $Z' \mathbf{V}_{\theta}^{-1} Z \sim \chi_J^2$ , 又因为  $\hat{\mathbf{V}}_{\theta}^{-1} \xrightarrow{p} \mathbf{V}_{\theta}^{-1}$ , 使用 Slutsky 定理即可推知

$$\sqrt{n}[R(\hat{\beta}) - r]' \hat{\mathbf{V}}_{\theta}^{-1} \sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} \chi_J^2$$

证毕.

如果将函数  $\mathbf{R} : \mathbb{R}^K \rightarrow \mathbb{R}^J$  设置为线性函数, 也即  $\mathbf{R} = \nabla_{\beta} R(\beta)$  为常矩阵, 那么原假设可以简化为第二章那样的线性假设  $\mathbb{H}_0 : \mathbf{R}\beta = r$ , 此时  $T$  统计量和 Wald 统计量分别为

$$T = \frac{\sqrt{n}(\mathbf{R}\hat{\beta} - r)}{\sqrt{\mathbf{R} \hat{\mathbf{V}}_{\beta} \mathbf{R}'}} \xrightarrow{d} N(0, 1) \quad (4.17)$$

$$W = n(\mathbf{R}\hat{\beta} - r)' (\mathbf{R} \hat{\mathbf{V}}_{\beta} \mathbf{R}')^{-1} (\mathbf{R}\hat{\beta} - r) \xrightarrow{d} \chi_J^2$$

它们是最为常见的检验统计量. 特别地, 如果条件同方差假设成立, 那么可以使用以下更有效的统计量

$$T_{\text{Hom}} = \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2 \mathbf{R}(X'X)^{-1} \mathbf{R}'}} \xrightarrow{d} N(0, 1) \quad (4.18)$$

$$W_{\text{Hom}} = (\mathbf{R}\hat{\beta} - r)'[s^2 \mathbf{R}(X'X)^{-1} \mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - r) \xrightarrow{d} \chi_J^2$$

此时的  $W_{\text{Hom}} = J \cdot F$ , 因此 Wald 统计量在条件同方差下又称  $J \cdot F$  统计量.

特别地, 如果想要检验回归系数是否联合为 0, 可以使用以下  $(n - K)R^2$  检验, 它是 LM 检验 (或称得分检验) 中的一种.

#### 定理 4.13

在假设 4.2 和条件同方差  $\mathbb{E}[e_i^2|X_i] = \sigma^2$  成立的情况下, 如果原假设  $\mathbb{H}_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$  成立, 那么当  $n \rightarrow \infty$  时有

$$(n - K)R^2 \xrightarrow{d} \chi_{K-1}^2$$

**证明** 根据推论 2.2 可知在  $\mathbb{H}_0$  成立的条件下有

$$F = \frac{R^2/(K - 1)}{(1 - R^2)/(n - K)}$$

根据 Wald 检验又可知

$$(K - 1)F = \frac{(n - K)R^2}{1 - R^2} \xrightarrow{d} \chi_{K-1}^2$$

由于  $K$  为常数, 故而  $(K - 1)F = O_p(1)$ , 从而

$$\frac{R^2}{1 - R^2} = O_p(n^{-1}) = o_p(1)$$

也即  $1 - R^2 \rightarrow 1$ . 根据 Slutsky 定理可得

$$(n - K)R^2 = (K - 1)F \cdot (1 - R^2) \xrightarrow{d} \chi_{K-1}^2$$

证毕.

**注** 下面我们给出两点注意事项.

(1). 在线性假设下, 我们使用的是  $\mathbf{R}$  而非  $\hat{\mathbf{R}}$ , 这是因为在非线性假设下,  $\mathbf{R}$  包含了未知的真实参数, 因此需要使用  $\hat{\mathbf{R}}$  才能得到可计算的统计量. 而在线性假设下,  $\mathbf{R}$  为常矩阵, 我们可以直接使用它构造统计量.

(2). 如果在条件同方差的情况下使用 (4.17) 而非 (4.18) 进行参数检验, 尽管这样是可行的, 因为前者更加稳健且适用于一般情况, 但是在较小的样本量下, 由于没有利用同方差这一信息, (4.17) 的真实分布可能与渐近  $\chi^2$  分布相差甚远.

最后, 如果我们想知道使用正态分布来逼近  $T$  统计量的准确程度, 一个直觉是利用 Edgeworth 展开, 具体内容参考 Hansen (2022a) 的 7.19 节内容.

## 4.5 异方差的检验

为了检验条件同方差假设  $\mathbb{E}[e_i^2|X_i] = \sigma^2$ , 可以考虑以下辅助回归

$$\begin{aligned} e_i^2 &= \gamma_0 + \sum_{j=1}^{K-1} \gamma_j X_{ji} + \sum_{1 \leq j \leq l \leq K-1} \gamma_{jl} X_{ji} X_{li} + v_i \\ &= \gamma' \text{vech}(X_i X_i') + v_i \end{aligned}$$

这里的  $\text{vech}$  是一个向量化算子, 它将  $K \times K$  维对称矩阵  $X_i X_i'$  的下三角元素转变为一个  $K(K+1)/2 \times 1$  维向量. 如果原假设  $\mathbb{H}_0: \mathbb{E}[e_i^2|X_i] = \sigma^2$  和条件  $\mathbb{E}[e_i^4|X_i] = \mu$  成立, 那么当  $n \rightarrow \infty$  时有

$$(n - J - 1)R^2 \xrightarrow{d} \chi_J^2$$

其中  $J = K(K+1)/2 - 1$ . 然而  $e_i^2$  是不可观测的, 此时我们可以考虑使用 OLS 残差平方  $\hat{e}_i^2$  将其替代, White (1980) 证明了这种替代不会影响到统计量  $(n - J - 1)R^2$  的渐近分布, 这个检验称为 White 检验, 以下给出它的直观理解.

记  $U_i = \text{vech}(X_i X_i')$ , 则辅助回归方程可以记为

$$e_i^2 = U_i' \gamma + v_i$$

当原假设  $\mathbb{H}_0: \mathbb{E}[e_i^2|X_i] = \sigma^2$  以及  $\mathbb{E}[e_i^4|X_i] = \mu_4$  成立时,  $v_i$  满足条件同方差, 故而

$$\sqrt{n}(\tilde{\gamma} - \gamma) \xrightarrow{d} N(0, \sigma_v^2 \mathbf{Q}_{UU}^{-1})$$

其中  $\tilde{\gamma}$  为 OLS 估计量. 设  $\mathbf{R}$  为  $J \times (J+1)$  维矩阵, 它的第  $(i, i+1)$  个元素为 1, 其中  $i = 1, 2, \dots, J$ , 并且其余元素均为 0. 于是

$$\sqrt{n} \mathbf{R} \tilde{\gamma} = \sqrt{n} \mathbf{R}(\tilde{\gamma} - \gamma) \xrightarrow{d} N(0, \sigma_v^2 \mathbf{R} \mathbf{Q}_{UU}^{-1} \mathbf{R}')$$

从而  $\mathbf{R} \tilde{\gamma} = O_p(n^{-\frac{1}{2}})$ . 注意到原假设可以改写为  $\mathbb{H}_0: \gamma = 0$ , 根据 Wald 检验可知

$$n(\mathbf{R} \tilde{\gamma})'(s_v^2 \mathbf{R} \hat{\mathbf{Q}}_{UU}^{-1} \mathbf{R}')^{-1}(\mathbf{R} \tilde{\gamma}) \xrightarrow{d} \chi_J^2$$

现在用  $\hat{e}_i^2$  替换  $e_i^2$ , 得到新的辅助回归

$$\hat{e}_i^2 = U_i' \gamma + \tilde{v}_i \quad (4.19)$$

记 (4.19) 的 OLS 估计量为  $\hat{\gamma}$ . 注意到

$$\begin{aligned} \hat{e}_i^2 &= [e_i - X_i'(\hat{\beta} - \beta)]^2 \\ &= e_i^2 + (\hat{\beta} - \beta)' X_i X_i' (\hat{\beta} - \beta) - 2(\hat{\beta} - \beta)' X_i e_i \end{aligned}$$

它的第一项为  $e_i^2$ , 第二项和  $U_i = \text{vech}(X_i X_i')$  成正比, 而第三项和  $X_i e_i$  成正比. 因此 OLS 估计量  $\hat{\gamma}$  可以分解为

$$\hat{\gamma} = \tilde{\gamma} + \hat{\delta} + \hat{\eta}$$

其中第一项决定了  $(n - J - 1)R^2$  的渐近  $\chi^2$  分布.

对于第三项, 在模型正确识别的情况下,  $X_i e_i$  与  $U_i$  不相关, 故而  $X_i e_i$  对  $U_i$  回归产生的 OLS 估计量是  $O_p(n^{-\frac{1}{2}})$ , 它乘以因子  $2(\hat{\beta} - \beta)' = O_p(n^{-\frac{1}{2}})$  后以  $n^{-1}$  的速度依概率为 0, 也即

$O_p(n^{-1})$ . 又因为  $R\tilde{\gamma} = O_p(n^{-\frac{1}{2}})$ , 因此  $\hat{\eta}$  对  $\tilde{\gamma}$  的影响可忽略不计.

对于第二项,  $X_i X_i'$  和  $U_i$  完全相关, 故而  $X_i X_i'$  对  $U_i$  回归产生的 OLS 估计量是  $O_p(1)$ , 当它乘以因子  $\|\hat{\beta} - \beta\|^2 = O_p(n^{-1})$  后仍然也以  $n^{-1}$  的速度依概率为 0, 也即  $\hat{\delta} = O_p(n^{-1})$ , 因此  $\hat{\delta}$  对  $\tilde{\gamma}$  的影响也可忽略不计.

**注** 下面给出关于 White 检验的评注.

(1) White 检验本质上检验的是  $\mathbb{E}[e_i^2|X_i]$  是否与  $X_i$  的二次项相关, 故而当  $\mathbb{E}[e_i^2|X_i]$  不依赖于  $X_i$  的二次项, 而依赖于  $X_i$  的更高次项时, White 检验无效.

(2) White 检验如果不能拒绝  $\mathbb{H}_0$ , 只表明没有找到存在条件异方差的证据, 而无法表明条件同方差假设成立.

(3) 当  $\mathbb{E}[e_i^4|X_i] = \mu$  不成立时, 无法使用  $(n - J - 1)R^2$  统计量, 此时应该通过 Wald 原理构造异方差稳健的检验统计量, 这里指的是辅助回归中的误差项  $v_i$  存在异方差.

## 第5章 系统估计

以上章节讨论的都是具有一个方程的线性回归模型,但有时候会出现多个方程的情形,如果多个方程之间存在某种关系,那么联合估计整个系统可能会提高估计的效率.

### 5.1 回归系统

许多单变量线性回归模型中用到的技术都可以用在多方程线性回归模型中,二者最大的区别在于表示矩阵估计量的符号.

假定回归系统包含  $m$  个回归方程,并且每个方程均有  $n$  个观测值

$$\begin{aligned} Y_1 &= X_1' \beta_1 + e_1 \\ &\vdots \\ Y_m &= X_m' \beta_m + e_m \end{aligned} \quad (5.1)$$

也即

$$Y_j = X_j' \beta_j + e_j, \quad j = 1, 2, \dots, m$$

这里的下标  $j$  表示第  $j$  个回归方程,而非第  $j$  个观测值.  $Y_j$  表示第  $j$  个因变量,数据矩阵  $X_j$  和系数向量  $\beta_j$  分别为  $K_j \times n$  和  $K_j \times 1$  维,  $e_j$  为随机误差项. 在整个回归系统中,回归元在每个  $j$  上可以不同也可以相同,共计有  $\bar{K} = \sum_{j=1}^m K_j$  个回归系数.

从一个总体中随机抽样得到的多方程线性回归模型为

$$Y_i = \bar{X}_i \beta + e_i, \quad i = 1, 2, \dots, n \quad (5.2)$$

其中  $Y_i$  和  $e_i$  均为  $m \times 1$  维列向量,并且

$$\bar{X}_i = \begin{bmatrix} X_{i1}' & 0 & \cdots & 0 \\ 0 & X_{i2}' & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & X_{im}' \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

分别为  $m \times \bar{K}$  维矩阵和  $\bar{K} \times 1$  维列向量. 现在我们将这  $n$  个观测堆叠,由此得到

$$Y = \bar{X} \beta + e$$

其中

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \quad \bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_n \end{bmatrix}$$

它们的维数分别为  $mn \times 1$ ,  $mn \times 1$  以及  $mn \times \bar{K}$ .

在许多应用情况下, 每个  $j$  中对应的回归元都是一样的, 此时有  $X_j = X, K_j = K, j = 1, 2, \dots, m$ . 将  $n$  个观测值堆叠后可以得到  $n \times m$  维的表示形式

$$Y = XB + E$$

其中  $B = [\beta_1, \beta_2, \dots, \beta_m]$  为  $K \times m$  维矩阵, 并且

$$Y = \begin{bmatrix} Y'_1 \\ Y'_2 \\ \vdots \\ Y'_n \end{bmatrix}, \quad X = \begin{bmatrix} X'_1 \\ X'_2 \\ \vdots \\ X'_n \end{bmatrix}, \quad E = \begin{bmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_n \end{bmatrix} \quad (5.3)$$

它们的维数分别为  $n \times m, n \times K$  以及  $n \times m$ , 并且每个  $X'_i$  的维数为  $1 \times K, i = 1, 2, \dots, n$ .

除此之外, 回归元相同时的数据矩阵<sup>1</sup>还可以方便地表示为  $m \times mK$  维矩阵

$$\bar{X}_i = \begin{bmatrix} X'_i & 0 & \cdots & 0 \\ 0 & X'_i & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & X'_i \end{bmatrix} = I_m \otimes X'_i$$

其中  $\otimes$  表示 Kronecker 积<sup>2</sup>. 注意, 这里并没有  $\bar{X} = I_m \otimes X$  这一关系.

## 5.2 系统普通最小二乘估计

线性回归系统 (5.1) 可以由 OLS 进行估计, 称为系统普通最小二乘估计 (System Ordinary Least Squares, SOLS), 可以将其表示为

$$\hat{\beta}_j = \left( \sum_{i=1}^n X_{ji} X'_{ji} \right)^{-1} \left( \sum_{i=1}^n X_{ji} Y_{ji} \right)$$

它的堆叠形式为向量

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_m \end{bmatrix}$$

<sup>1</sup>如果不关注观测值, 那么可以将下标  $i$  去掉.

<sup>2</sup>对于任意的矩阵  $A_{m \times n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$  与  $B_{p \times q}$ , Kronecker 积为  $A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}_{mp \times nq}$ .

Kronecker 积的运算律:

- $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ ;
- $(A \otimes B)' = A' \otimes B'$ ;
- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ .



此外还可以用系统符号来表示

$$\hat{\beta} = (\bar{X}'\bar{X})^{-1}\bar{X}'Y = \left(\sum_{i=1}^n \bar{X}_i \bar{X}_i'\right)^{-1} \left(\sum_{i=1}^n \bar{X}_i' Y_i\right) \quad (5.4)$$

其中

$$\bar{X}'\bar{X} = \begin{bmatrix} \sum_{i=1}^n X_{1i} X_{1i}' & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n X_{2i} X_{2i}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^n X_{mi} X_{mi}' \end{bmatrix}, \quad \bar{X}'Y = \begin{bmatrix} \sum_{i=1}^n X_{1i} Y_{1i} \\ \sum_{i=1}^n X_{2i} Y_{2i} \\ \vdots \\ \sum_{i=1}^n X_{mi} Y_{mi} \end{bmatrix}$$

特别地, 如果所有单方程的回归元相同, 那么有

$$\hat{\beta}_j = \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\sum_{i=1}^n X_i Y_{ji}\right)$$

以及

$$B = [\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m] = (X'X)^{-1}X'Y$$

### 5.3 SOLS 的期望与方差

如果以下条件期望假设成立

$$\mathbb{E}[e_i | \bar{X}_i] = 0 \quad (5.5)$$

那么我们可以计算有限样本期望以及  $\hat{\beta}$  的方差. 等式 (5.5) 等价于  $\mathbb{E}[Y_j | X_j] = X_j' \beta_j$ , 这意味着回归模型是正确识别的.

现在将估计量中心化

$$\hat{\beta} - \beta = (\bar{X}'\bar{X})^{-1}\bar{X}'e = \left(\sum_{i=1}^n \bar{X}_i \bar{X}_i'\right)^{-1} \left(\sum_{i=1}^n \bar{X}_i' e_i\right)$$

根据 (5.5), 两端取条件期望即可得到  $\mathbb{E}[\hat{\beta} | \bar{X}] = \beta$ . 因此, 在模型正确识别的情况下, SOLS 是无偏估计.

为了计算 SOLS 估计量的方差, 我们先定义第  $i$  个观测值的误差的条件协方差矩阵为  $\mathbb{E}[e_i e_i' | \bar{X}_i] = \Sigma_i$ . 如果各观测值相互独立, 那么有  $mn \times mn$  维矩阵

$$\mathbb{E}[ee' | \bar{X}] = \text{diag}\{\Sigma_1, \Sigma_2, \dots, \Sigma_n\}$$

于是

$$\text{var}\left(\sum_{i=1}^n \bar{X}_i' e_i \middle| \bar{X}\right) = \sum_{i=1}^n \text{var}(\bar{X}_i' e_i | \bar{X}_i) = \sum_{i=1}^n \bar{X}_i' \Sigma_i \bar{X}_i$$

因此

$$\text{var}(\hat{\beta} | \bar{X}) = (\bar{X}'\bar{X})^{-1} \left(\sum_{i=1}^n \bar{X}_i' \Sigma_i \bar{X}_i\right) (\bar{X}'\bar{X})^{-1}$$

如果每个单方程具有完全相同的回归元, 也即  $\bar{X}_i = I_m \otimes X_i'$ , 那么根据 Kronecker 积的运

算律即可推知

$$\begin{aligned}\bar{X}'\bar{X} &= \sum_{i=1}^n \bar{X}_i' \bar{X}_i = \sum_{i=1}^n (I_m \otimes X_i)(I_m \otimes X_i') \\ &= \sum_{i=1}^n (I_m \otimes X_i X_i') = I_m \otimes \left( \sum_{i=1}^n X_i X_i' \right) \\ &= I_m \otimes (X'X)\end{aligned}$$

从而  $(\bar{X}'\bar{X})^{-1} = I_m \otimes (X'X)^{-1}$ . 另一方面, 根据 Kronecker 积的定义可知  $\Sigma_i = \Sigma_i \otimes 1$ , 再由运算律得到

$$(I_m \otimes X_i)\Sigma_i(I_m \otimes X_i') = \Sigma_i \otimes X_i X_i'$$

于是

$$\text{var}(\hat{\beta}|X) = [I_m \otimes (X'X)^{-1}] \left[ \sum_{i=1}^n (\Sigma_i \otimes X_i X_i') \right] [I_m \otimes (X'X)^{-1}] \quad (5.6)$$

另一方面, 如果误差项满足条件同方差

$$\mathbb{E}[e_i e_i' | \bar{X}_i] = \Sigma, \quad i = 1, 2, \dots, n \quad (5.7)$$

那么协方差矩阵可以简化为

$$\text{var}(\hat{\beta}|X) = (\bar{X}'\bar{X})^{-1} \left( \sum_{i=1}^n \bar{X}_i' \Sigma \bar{X}_i \right) (\bar{X}'\bar{X})^{-1} \quad (5.8)$$

如果以上两种简化条件同时成立, 那么利用 (5.6) 和 (5.8) 可以得到一个相当简单的协方差矩阵表达式

$$\text{var}(\hat{\beta}|X) = \Sigma \otimes (X'X)^{-1} \quad (5.9)$$

## 5.4 SOLS 的渐近性质

前面已经提到, 单方程线性回归模型中的诸多定理都可以普遍应用到回归系统中来, 本节将说明之前证明的 OLS 的渐近性质在 SOLS 中也成立.

### 定理 5.1

如果每个单方程都满足假设 4.2, 那么当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$$

其中  $V_\beta = Q^{-1}\Omega Q^{-1}$ ,  $\Omega = \mathbb{E}[\bar{X}_i e_i e_i' \bar{X}_i'] = \mathbb{E}[\bar{X}_i' \Sigma_i \bar{X}_i]$ , 并且

$$Q = \mathbb{E}[\bar{X}_i' \bar{X}_i] = \begin{bmatrix} \mathbb{E}[X_{1i} X_{1i}'] & & & \\ & \mathbb{E}[X_{2i} X_{2i}'] & & \\ & & \ddots & \\ & & & \mathbb{E}[X_{mi} X_{mi}'] \end{bmatrix}$$



**证明** 根据定理 1.8(3) 可知, 对于一切  $j = 1, 2, \dots, m$  都有

$$\mathbb{E}[X_j e_j] = 0 \quad (5.10)$$

考虑向量

$$\bar{X}'_i e_i = \begin{bmatrix} X_{1i} e_{1i} \\ X_{2i} e_{2i} \\ \vdots \\ X_{mi} e_{mi} \end{bmatrix}$$

在式 (5.10) 下,  $\bar{X}'_i e_i$  具有零均值并且在各观测  $i$  之间是 i.i.d. 的. 根据多元 Lindeberg-Levy CLT 可知

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}'_i e_i \xrightarrow{d} N(0, \Omega)$$

其中

$$\Omega = \mathbb{E}[\bar{X}_i e_i e'_i \bar{X}'_i] = \mathbb{E}[\bar{X}'_i \Sigma_i \bar{X}_i]$$

这里的第二个等号是由 LIE 得到的, 也即

$$\mathbb{E}[\bar{X}_i e_i e'_i \bar{X}'_i] = \mathbb{E}[\bar{X}_i \mathbb{E}[e_i e'_i | X_i] \bar{X}'_i] = \mathbb{E}[\bar{X}'_i \Sigma_i \bar{X}_i]$$

注意到

$$\sqrt{n}(\hat{\beta} - \beta) = \left( n^{-1} \sum_{i=1}^n \bar{X}_i \bar{X}'_i \right)^{-1} \left( n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}'_i e_i \right)$$

参考第四章的做法, 容易证得  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$ .

**例 5.1 协方差矩阵的化简** 在条件同方差假设 (5.7) 下,  $\Omega$  自然而然地化简为

$$\Omega = \mathbb{E}[\bar{X}'_i \Sigma \bar{X}_i]$$

如果所有单方程的回归元相同, 那么

$$\begin{aligned} \Omega &= \mathbb{E}[\bar{X}_i e_i e'_i \bar{X}'_i] \\ &= \mathbb{E}[(I_m \otimes X_i)(e_i e'_i \otimes 1)(I_m \otimes X'_i)] \\ &= \mathbb{E}[e_i e'_i \otimes X_i X'_i] \end{aligned}$$

如果条件同方差和回归元相同这两个假设同时成立, 那么

$$\begin{aligned} \Omega &= \mathbb{E}[(I_m \otimes X_i)(\Sigma \otimes 1)(I_m \otimes X'_i)] \\ &= \mathbb{E}[\Sigma \otimes (X_i X'_i)] \end{aligned}$$

不仅如此, 此时还可以将  $V_\beta$  化简为  $V_\beta = \Sigma \otimes (\mathbb{E}[X_i X'_i])^{-1}$ .

**例 5.2 协方差矩阵的估计** 我们定义第  $i$  个观测值的  $m \times 1$  维残差向量  $\hat{e}_i = Y_i - \bar{X}_i \hat{\beta}$ , 以及误差的  $m \times m$  维协方差矩阵的最小二乘估计量

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \hat{e}_i \hat{e}'_i \quad (5.11)$$

考虑最一般的形式, 直觉告诉我们协方差矩阵  $V_\beta$  的估计量为

$$\hat{V}_\beta = (\bar{X}'\bar{X})^{-1} \left( \sum_{i=1}^n \bar{X}_i' \hat{e}_i \hat{e}_i' \bar{X}_i \right) (\bar{X}'\bar{X})^{-1}$$

在条件同方差假设 (5.7) 下, 协方差矩阵估计量简化为

$$\hat{V}_\beta^0 = (\bar{X}'\bar{X})^{-1} \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma} \bar{X}_i \right) (\bar{X}'\bar{X})^{-1}$$

如果所有单方程的回归元相同, 那么

$$\hat{V}_\beta = [I_m \otimes (X'X)^{-1}] \left[ \sum_{i=1}^n (\hat{e}_i \hat{e}_i' \otimes X_i X_i') \right] [I_m \otimes (X'X)^{-1}]$$

如果条件同方差和回归元相同这两个假设同时成立, 那么

$$V_\beta^0 = \hat{\Sigma} \otimes (X'X)^{-1}$$

容易证明  $n\hat{V}_{OLS}$  和  $n\hat{V}_{OLS}^0$  分别是  $V_\beta$  和  $V_\beta^0$  的一致估计量。

有时候我们对参数  $\theta = R(\beta_1, \dots, \beta_m) = R(\beta)$  感兴趣, 其中  $R$  是关于系数的可测函数, 此时  $\theta$  的最小二乘估计量为  $\hat{\theta} = R(\hat{\beta})$ , 使用 Delta 法可以得到以下定理. 根据这一定理, 我们可以像第四章那样构造适用于参数检验的标准误和检验统计量。

#### 定理 5.2

如果每个单方程都满足假设 4.2 和 4.3, 那么当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta)$$

其中  $V_\theta = R V_\beta R'$ ,  $R = \nabla_\beta R(\beta)$ .



## 5.5 SGLS 的渐近性质

首先推导 SGLS 估计量, 在回归方程

$$Y_i = \bar{X}_i \beta + e_i$$

等号两端左乘  $\Sigma_i^{-\frac{1}{2}}$  得到

$$Y_i^* = \bar{X}_i^* \beta + e_i^*$$

此时误差向量满足  $\mathbb{E}[e_i^* e_i^{*'}] = I_m$ , 使用 SOLS 对以上方程估计即可得到 SGLS 估计量

$$\hat{\beta}_{GLS} = \left( \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} Y_i \right) \quad (5.12)$$

除了用这种方法外, 还可以基于以下矩阵形式的回归系统推导 SGLS 估计量

$$Y = \bar{X} \beta + e$$

这里误差项  $e$  的协方差矩阵为  $\mathbb{E}[ee'] = I_n \otimes \Sigma_i$ , 于是

$$\hat{\beta}_{GLS} = [\bar{X}'(I_n \otimes \Sigma_i^{-1})\bar{X}]^{-1} [\bar{X}'(I_n \otimes \Sigma_i^{-1})Y] \quad (5.13)$$

表达式 (5.12) 和 (5.13) 在代数上是等价的, 证明也非常容易.

不同于之前在定理 5.1 中只需将假设 4.2 扩张到每个单方程成立, 在讨论关于 GLS 估计量的渐近性质之前, 我们需要给出一些更强的正则假设.

### 假设 5.1

$$\mathbb{E}[\bar{X}_i \otimes e_i] = 0.$$

### 引理 5.1

如果假设 5.1 成立, 那么当  $n \rightarrow \infty$  时有  $n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \xrightarrow{p} \mathbb{E}[\bar{X}_i' \Sigma_i^{-1} e_i] = 0$ .

**证明** 该引理的证明需要用到矩阵理论的  $\text{vec}$  算子<sup>3</sup>, 根据它的性质可得

$$\text{vec}(\mathbb{E}[\bar{X}_i' \Sigma_i^{-1} e_i]) = \mathbb{E}[\text{vec}(\bar{X}_i' \Sigma_i^{-1} e_i)] = \mathbb{E}[(e_i' \otimes \bar{X}_i') \text{vec}(\Sigma_i^{-1})] = 0$$

因此由 WLLN 即可得到  $n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \xrightarrow{p} \mathbb{E}[\bar{X}_i' \Sigma_i^{-1} e_i] = 0$ .

**注** 该引理证明的逻辑是, 如果将  $\mathbb{E}[\bar{X}_i \Sigma^{-1} \bar{X}_i']$  中的每一列堆叠起来后为零矩阵, 那么原来的矩阵也必为零矩阵.

下面第二个假设类似于假设 4.2(4), 是为了避免出现奇异矩阵的情况.

### 假设 5.2

协方差矩阵  $\Sigma_i$  正定, 并且  $\mathbb{E}[\bar{X}_i' \Sigma_i^{-1} \bar{X}_i]$  可逆.

除此之外, 我们还假定某些弱的矩条件成立, 正如假设 4.1 和假设 4.2 提到的那样. 首先写出表达式

$$\hat{\beta}_{\text{GLS}} - \beta = \left( n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} \bar{X}_i \right)^{-1} \left( n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \right)$$

根据 WLLN 和 CMT 可知

$$n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} \bar{X}_i \xrightarrow{p} \mathbb{E}[\bar{X}_i' \Sigma_i^{-1} \bar{X}_i]$$

再根据引理 5.1 可得  $n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \xrightarrow{p} 0$ , 从而 GLS 估计量是一致的. 注意到

$$\sqrt{n}(\hat{\beta}_{\text{GLS}} - \beta) = \left( n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} \bar{X}_i \right)^{-1} \left( n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \right)$$

根据 CLT 可知

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma_i^{-1} e_i \xrightarrow{d} N(0, \Omega)$$

<sup>3</sup>对于矩阵  $A_{m \times n} = [a_1 \cdots a_n]$ ,  $\text{vec}(A) = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}_{mn \times 1}$ .

$\text{vec}$  算子具有如下性质: 对于适合的矩阵  $D, E$  和  $F$ ,  $\text{vec}(DEF) = (F' \otimes D) \text{vec}(E)$ .

其中  $\Omega = \mathbb{E}[\bar{X}_i \Sigma_i^{-1} e_i e_i' \Sigma_i^{-1} \bar{X}_i]$ , 于是

$$\sqrt{n}(\hat{\beta}_{\text{GLS}} - \beta) \xrightarrow{d} Q^{-1} \Omega Q^{-1}$$

这里  $Q = \mathbb{E}[\bar{X}_i' \Sigma_i^{-1} \bar{X}_i]$ .

## 5.6 似不相关回归

最后来看如下具有条件期望和条件同方差的系统回归模型

$$Y_i = \bar{X}_i \beta + e_i \quad (5.14)$$

$$\mathbb{E}[e_i | \bar{X}_i] = 0$$

$$\mathbb{E}[e_i e_i' | \bar{X}_i] = \Sigma$$

也即单方程的扰动项不存在自相关且方差相同, 但随机扰动项在不同方程之间存在相关性, 上述模型称为似不相关回归 (Seemingly Unrelated Regression, SUR).

由于  $\Sigma$  的具体形式是未知的, 我们考虑使用 (5.11) 所给出的  $\hat{\Sigma}$  来得到系统 FGLS 估计量

$$\begin{aligned} \hat{\beta}_{\text{SUR}} &= \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} Y_i \right) \\ &= [\bar{X}'(I_n \otimes \hat{\Sigma}^{-1})\bar{X}]^{-1} [\bar{X}'(I_n \otimes \hat{\Sigma}^{-1})Y] \end{aligned} \quad (5.15)$$

称为 SUR 估计量, 由 Zellner (1962) 提出.

上述 SUR 估计量中的  $\hat{\Sigma}$  的构建基于 SOLS 获得的残差, 而一旦我们得到了初始 SUR 估计量, 就能将 (5.11) 中的  $\hat{e}_i$  替换为  $\check{e}_i = Y_i - \bar{X}_i' \hat{\beta}_{\text{SUR}}$ , 并且构造新的协方差矩阵估计量  $\hat{\Sigma} = n^{-1} \sum_{i=1}^n \check{e}_i \check{e}_i'$ , 然后将其迭代到 SUR 估计量中直至收敛.

可以证明, 仅凭每个单方程满足假设 4.2 无法推出  $\mathbb{E}[\bar{X}_i' \Sigma^{-1} e_i] = 0$  成立, 这也是假设 5.1 更强的地方所在.

### 定理 5.3

如果假设 5.1 成立, 并且  $\mathbb{E}[\bar{X}_i' \bar{X}_i]$  是有限且正定的, 那么当  $n \rightarrow \infty$  时有

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \hat{e}_i \hat{e}_i' \xrightarrow{p} \Sigma$$

其中  $\hat{e}_i = Y_i - \bar{X}_i' \hat{\beta}$  为 SOLS 估计残差.

**证明** 注意到  $\hat{e}_i = e_i - \bar{X}_i'(\hat{\beta} - \beta)$ , 于是

$$\hat{e}_i \hat{e}_i' = e_i e_i' - e_i (\hat{\beta} - \beta)' \bar{X}_i' - \bar{X}_i (\hat{\beta} - \beta) e_i' + \bar{X}_i (\hat{\beta} - \beta) (\hat{\beta} - \beta)' \bar{X}_i' \quad (5.16)$$

只需证明后三项的 vec 均值依概率收敛于 0 即可.

首先, 式 (5.16) 第二项的 vec 均值为

$$n^{-1} \sum_{i=1}^n (\bar{X}_i \otimes e_i) \cdot \text{vec}[(\hat{\beta} - \beta)']$$

由于  $\hat{\beta} \xrightarrow{p} \beta$  以及  $n^{-1} \sum_{i=1}^n (\bar{X}_i \otimes e_i) \xrightarrow{p} 0$ , 因此上式为  $o_p(1)$ . 第三项是第二项的转置, 因此也是  $o_p(1)$ . 最后一项的  $\text{vec}$  均值为

$$n^{-1} \sum_{i=1}^n (\bar{X}_i \otimes \bar{X}_i) \cdot \text{vec}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

由于  $n^{-1} \sum_{i=1}^n (\bar{X}_i \otimes \bar{X}_i) = O_p(1)$ , 故而最后一项仍为  $o_p(1)$ . 因此

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n e_i e_i' + o_p(1)$$

因此  $\text{plim } \hat{\Sigma} = \Sigma$ .

以下定理中, 我们都假定适合 CLT 的矩条件成立.

#### 定理 5.4

在假设 5.1 和 5.2 下, 当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\beta}_{\text{SUR}} - \beta) \xrightarrow{d} N(0, V_{\beta}^*)$$

其中  $V_{\beta}^* = Q^{-1} \Omega Q^{-1}$ ,  $Q = \mathbb{E}[\bar{X}_i' \Sigma \bar{X}_i]$ , 以及  $\Omega = \mathbb{E}[\bar{X}_i' \Sigma^{-1} e_i e_i' \Sigma^{-1} \bar{X}_i]$ .



**证明** 注意到

$$\sqrt{n}(\hat{\beta}_{\text{SUR}} - \beta) = \left( n^{-1} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1} \left( n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} e_i \right) \quad (5.17)$$

对上式等号右端的第二项使用  $\text{vec}$  算子, 由于该项本身即为列向量, 因此

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} e_i - n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i = \left[ n^{-\frac{1}{2}} \sum_{i=1}^n (e_i \otimes \bar{X}_i)' \right] \text{vec}(\hat{\Sigma}^{-1} - \Sigma^{-1})$$

根据假设 5.1, CLT 意味着  $n^{-\frac{1}{2}} \sum_{i=1}^n (e_i \otimes \bar{X}_i) = O_p(1)$ , 又因为  $\hat{\Sigma}$  是  $\Sigma$  的一致估计量, 于是

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} e_i = n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i + o_p(1)$$

类似可证  $n^{-1} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i = n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} \bar{X}_i + o_p(1)$ . 因此可将 (5.17) 改写为

$$\sqrt{n}(\hat{\beta}_{\text{SUR}} - \beta) = \left( n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} \bar{X}_i \right)^{-1} \left( n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i \right) + o_p(1)$$

根据 CLT 可知

$$n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i \xrightarrow{d} N(0, \Omega)$$

这里  $\Omega = \mathbb{E}[\bar{X}_i' \Sigma^{-1} e_i e_i' \Sigma^{-1} \bar{X}_i]$ . 根据 WLLN 和 CMT 可知

$$\left( n^{-1} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} \bar{X}_i \right)^{-1} - Q^{-1} = o_p(1)$$

这里  $\mathbf{Q} = \mathbb{E}[\bar{X}_i' \Sigma^{-1} \bar{X}_i]$ . 因为  $n^{-\frac{1}{2}} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i = O_p(1)$ , 于是

$$\sqrt{n}(\hat{\beta}_{\text{SUR}} - \beta) = \mathbf{Q}^{-1} \left( n^{-1/2} \sum_{i=1}^n \bar{X}_i' \Sigma^{-1} e_i \right) + o_p(1)$$

最后使用 Slutsky 定理即可证得结论.

在 FGLS 背景下,  $\mathbf{Q}$  的一致估计量为

$$\hat{\mathbf{Q}} = n^{-1} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i$$

而  $\Omega$  的一致估计量为

$$\hat{\Omega} = n^{-1} \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \hat{e}_i \hat{e}_i' \hat{\Sigma}^{-1} \bar{X}_i$$

其中  $\hat{e}_i = Y_i - \bar{X}_i \hat{\beta}$ , 并且这里的  $\hat{\beta}$  为 SOLS 估计量. 现在可以得到  $\hat{\beta}_{\text{SUR}}$  的渐近协方差矩阵  $\text{avar}(\hat{\beta}_{\text{SUR}})$  的一致估计量

$$\hat{V}_{\text{SUR}} = \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \hat{e}_i \hat{e}_i' \hat{\Sigma}^{-1} \bar{X}_i \right) \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1}$$

特别地, 如果定理 5.4 中的条件还包括 (5.7), 那么

$$\begin{aligned} \mathbb{E}[\bar{X}_i' \Sigma^{-1} e_i e_i' \Sigma^{-1} \bar{X}_i] &= \mathbb{E}[\mathbb{E}[\bar{X}_i' \Sigma^{-1} e_i e_i' \Sigma^{-1} \bar{X}_i | \bar{X}_i]] \\ &= \mathbb{E}[\bar{X}_i' \Sigma^{-1} \mathbb{E}[e_i e_i' | \bar{X}_i] \Sigma^{-1} \bar{X}_i] = \mathbb{E}[\bar{X}_i' \Sigma^{-1} \bar{X}_i] \end{aligned}$$

此时  $V_{\beta}^* = (\mathbb{E}[\bar{X}_i' \Sigma^{-1} \bar{X}_i])^{-1}$ ,  $\text{avar}(\hat{\beta}_{\text{SUR}})$  的一致估计量为

$$\hat{V}_{\text{SUR}}^0 = \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1}$$

### 定理 5.5

在假设 5.1 和 5.2 下, 如果条件 (5.7) 成立, 那么

$$V_{\beta} - V_{\beta}^* \sim \text{p.s.d.}$$

其中  $V_{\beta} = (\mathbb{E}[\bar{X}_i' \bar{X}_i])^{-1} \mathbb{E}[\bar{X}_i' \Sigma \bar{X}_i] (\mathbb{E}[\bar{X}_i' \bar{X}_i])^{-1}$ , 这表明 SUR 估计量比 SOLS 估计量更渐进有效.



**证明** 考虑以下式子

$$\mathbf{A} = (\bar{X}' \bar{X})^{-1} [\bar{X}' (I_n \otimes \Sigma) \bar{X}] (\bar{X}' \bar{X})^{-1} - [\bar{X}' (I_n \otimes \Sigma^{-1}) \bar{X}]^{-1}$$

由于  $\text{plim } n\mathbf{A} = V_{\beta} - V_{\beta}^*$ , 所以只需证明  $\mathbf{A}$  半正定即可. 记  $\mathbf{B} = I_n \otimes \Sigma$ , 于是

$$\mathbf{A} = \mathbf{C} \mathbf{D} \mathbf{C}'$$

其中  $\mathbf{C} = (\bar{X}' \bar{X})^{-1} \bar{X}' \mathbf{B}^{\frac{1}{2}}$ ,  $\mathbf{D} = I_{mn} - \mathbf{B}^{-\frac{1}{2}} \bar{X} (\bar{X}' \mathbf{B}^{-1} \bar{X})^{-1} \bar{X}' \mathbf{B}^{-\frac{1}{2}}$ , 显然  $\mathbf{D}$  是一个对称幂等矩阵, 因此

$$\mathbf{A} = (\mathbf{C} \mathbf{D}) (\mathbf{C} \mathbf{D})' \sim \text{p.s.d.}$$

证毕.



**定理 5.6**

如果以下条件中的任意一个成立, 则 SUR 估计量等价于 SOLS 估计量.

(1) 每个单方程的回归元均相同.

(2) 协方差矩阵  $\Sigma$  为对角矩阵.



**证明** (1) 根据条件可知

$$\begin{aligned}\bar{X}_i' \hat{\Sigma}^{-1} &= (\mathbf{I}_m \otimes X_i) \hat{\Sigma}^{-1} = \hat{\Sigma}^{-1} \otimes X_i \\ &= (\hat{\Sigma}^{-1} \otimes \mathbf{I}_K)(\mathbf{I}_m \otimes X_i) = (\hat{\Sigma}^{-1} \otimes \mathbf{I}_K) \bar{X}_i'\end{aligned}$$

因此

$$\begin{aligned}\hat{\beta}_{\text{SUR}} &= \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' \hat{\Sigma}^{-1} Y_i \right) \\ &= \left[ (\hat{\Sigma}^{-1} \otimes \mathbf{I}_K) \sum_{i=1}^n \bar{X}_i' \bar{X}_i \right]^{-1} \left[ (\hat{\Sigma}^{-1} \otimes \mathbf{I}_K) \sum_{i=1}^n \bar{X}_i' Y_i \right] \\ &= \left( \sum_{i=1}^n \bar{X}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' Y_i \right) = \hat{\beta}\end{aligned}$$

(2) 设  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$ , 于是  $\hat{\Sigma}^{-1} = \text{diag}(\hat{\sigma}_1^{-2}, \dots, \hat{\sigma}_m^{-2})$ , 并且

$$\bar{X}_i' \hat{\Sigma}^{-1} \bar{X}_i = \hat{\Psi}^{-1} \bar{X}_i' \bar{X}_i, \quad \bar{X}_i' \hat{\Sigma}^{-1} Y_i = \hat{\Psi}^{-1} \bar{X}_i' Y_i$$

其中  $\hat{\Psi}$  为分块对角矩阵, 它的第  $j$  个块为  $\hat{\sigma}_j^2 \mathbf{I}_{K_j}$ ,  $j = 1, 2, \dots, m$ . 因此 SUR 估计量

$$\begin{aligned}\hat{\beta}_{\text{SUR}} &= \left( \sum_{i=1}^n \hat{\Psi}^{-1} \bar{X}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \hat{\Psi}^{-1} \bar{X}_i' Y_i \right) \\ &= \left( \sum_{i=1}^n \bar{X}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{X}_i' Y_i \right) = \hat{\beta}\end{aligned}$$

根据定理 5.6, 为了使 SUR 估计量和 SOLS 估计量不同,  $X_j$  必须随着  $j$  的变化而变化, 此时某些解释变量更有可能被遗漏, 而定理 5.5 却表明 SUR 估计量更加渐近有效, 这就需要我们权衡有效性与稳健性.

举例而言, 对于某个双回归方程构成的系统, 倘若我们对第一个方程感兴趣, 而第二个方程出现了模型误设, 那么此时使用 FGLS 得到的 SUR 估计量均是非一致估计量, 而只要  $\mathbb{E}[X_1 e_i] = 0$ , 则对第一个方程使用 OLS 仍能得出一致估计量.

当然, 如果整个回归系统设定正确, 也即假设 (5.1), (5.2) 以及条件 (5.7) 成立, 那么 SUR 估计量不仅是一致的, 且更加渐近有效.

## 第 6 章 工具变量回归分析

在前面各章节的渐近分析中,无一例外都牵扯到了  $\mathbb{E}[Xe] = 0$  这一关键条件,当它不成立时(例如第二章提到的遗漏变量)就会导致微观计量领域中占据重要地位的内生性问题,内生性的解决办法之一就是使用本章讨论的工具变量 (Instrumental Variables, IV).

### 6.1 内生性问题

我们称线性模型

$$Y = X'\beta + e \quad (6.1)$$

存在内生性,如果以下式子成立

$$\mathbb{E}[Xe] \neq 0$$

此时称  $X$  关于  $\beta$  是内生的,  $X$  为内生变量. 为了将 (6.1) 同回归和投影模型进行区分,后面称 (6.1) 为结构方程 (structural equation),  $\beta$  为结构参数 (structural parameter).

如果回归系数由线性投影模型所定义,那么不可能出现内生性,为了看清这一点,我们定义线性投影系数  $\beta^* = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$  以及线性投影模型

$$\begin{aligned} Y &= X'\beta^* + e^* \\ \mathbb{E}[Xe^*] &= 0 \end{aligned}$$

然而在 (6.1) 下,投影系数  $\beta^*$  并不等于结构参数  $\beta$ ,这是因为

$$\begin{aligned} \beta^* &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[X(X'\beta + e)] \\ &= \beta + \mathbb{E}[XX']^{-1}\mathbb{E}[Xe] \neq \beta \end{aligned} \quad (6.2)$$

内生性将导致结构参数  $\beta$  的 OLS 估计量不一致,称为内生性偏误 (endogenous bias). 事实上, i.i.d. 随机样本下的 OLS 估计量相对于投影系数是一致的.

造成内生性的原因主要包括遗漏变量偏误,测量误差以及联立方程偏误,其中遗漏变量偏误在第一章已经介绍过,这里不再赘述.

#### 6.1.1 测量误差

首先考虑被解释变量的测量误差,回归模型具有通常的线性形式

$$Y^* = X'\beta + e \quad (6.3)$$

如果出于某种原因, 我们无法观测到  $y^*$  的真实值, 而只能观测到它的替代值  $y$ , 此时总体误差定义为  $v = y - y^*$ , 将它代入到方程 (6.3) 中得到

$$Y = X'\beta + e + v \quad (6.4)$$

上述模型的自变量和因变量都可观测, 故而可以直接使用 OLS 估计. 倘若最初模型 (6.3) 满足假设 (4.1), 那么在变换后的模型 (6.4) 中,  $e$  仍然和每个回归元无关, 此时只要假设测量误差  $e$  关于每个解释变量无关, 则 OLS 估计量仍然是一致的.

尽管如此, 因变量的测量误差可归为回归扰动项的一部分, 它会增加  $\sqrt{n}(\hat{\beta} - \beta)$  的渐近方差. 换言之, 如果因变量存在测量误差, 那么对  $\beta$  的估计精度会下降.

现在来看含解释变量的测量误差的回归模型

$$Y = X'\beta + e$$

其中解释变量  $X$  无法被观测到, 只能观测到  $X = Z + u$ , 这里  $u$  是测量误差, 并且  $u$  独立于  $Z$  和  $e$ . 此时

$$Y = Z'\beta + e = (X - u)'\beta + e = X'\beta + v \quad (6.5)$$

这里  $v = e - u'\beta$ , 并且

$$\mathbb{E}[Xv] = \mathbb{E}[(Z + u)(e - u'\beta)] = -\mathbb{E}[uu']\beta$$

只要  $\mathbb{E}[ee'] \neq 0$  且  $\beta \neq 0$ , 那么 OLS 估计量是不一致的.

假设只有一个解释变量, 那么根据 (6.2) 和 (6.5) 可知

$$\text{plim } \hat{\beta} = \beta^* = \beta \left( 1 - \frac{\mathbb{E}[u^2]}{\mathbb{E}[X^2]} \right)$$

因为  $0 < \mathbb{E}[u^2]/\mathbb{E}[X^2] < 1$  成立, 所以投影系数总比结构参数更接近于 0, 这称为测量误差偏误 (measurement error bias) 或衰减偏误 (attenuation bias).

### 6.1.2 联立方程偏误

考虑如下简单国民收入决定模型

$$C = \beta_0 + \beta_1 Y + e \quad (6.6)$$

$$Y = C + D$$

其中  $Y$  为国民收入,  $C$  为消费支出,  $D$  为非消费支出, 边际消费倾向  $0 < \beta_1 < 1$ .

假设  $D$  与  $e$  相互独立, 并且  $\mathbb{E}[e] = 0$ . 现在考虑使用 OLS 来估计方程 (6.6) 的消费函数, 注意到

$$\mathbb{E}[Ye] = \mathbb{E}[(C + D)e] = \beta_1 \mathbb{E}[Ye] + \mathbb{E}[e^2]$$

移项后得

$$\mathbb{E}[Ye] = \frac{\mathbb{E}[e^2]}{1 - \beta_1}$$

显然 OLS 估计量也是非一致的. 由于国民收入与消费在这个回归系统内互相决定, 这又称为

联立方程偏误 (simultaneous equations bias).

## 6.2 工具变量

### 6.2.1 工具变量的定义

上面已经提到, 当回归元与误差项相关时会导致内生性, 它的相反概念则是外生性 (exogeneity). 我们称回归元  $X$  关于  $\beta$  是外生的, 如果  $\mathbb{E}[Xe] = 0$  成立, 称  $X$  为外生变量.

在许多情况下, 一个回归模型既有内生变量也有内生变量, 我们将回归元分割为  $X = [X_1', X_2']'$ , 其中  $X_1$  为外生变量而  $X_2$  为内生变量,  $X_1$  和  $X_2$  的维数分别为  $K_1 \times 1$  和  $K_2 \times 1$ . 类似地, 将结构参数分割为  $\beta = [\beta_1', \beta_2']'$ , 于是结构模型变为

$$Y = X_1' \beta_1 + X_2' \beta_2 + e \quad (6.7)$$

由于  $X_2$  是内生变量, 结构参数  $\beta_1$  和  $\beta_2$  的 OLS 估计量均是不一致的, 为了解决这一问题, 我们需要用到工具变量.

#### 定义 6.1

$L \times 1$  维随机向量  $Z$  称为模型 (6.7) 的工具变量, 如果

$$\mathbb{E}[Ze] = 0 \quad (6.8)$$

$$\mathbb{E}[ZZ'] > 0 \quad (6.9)$$

$$\text{rank}(\mathbb{E}[ZX']) = K \quad (6.10)$$

在以上定义中, (6.8) 表明工具变量和误差项不相关, (6.9) 排除了线性相关的冗余工具变量, (6.10) 称为模型可识别的秩条件. 后面我们将看到, 为了使 (6.10) 成立, 一个必要条件是  $L \geq K$ . 如果  $L = K$ , 则称模型恰好识别 (just identified); 如果  $L > K$ , 则称模型过度识别 (over identified).

**例 6.1** Dube and Harish (2020) 的研究表明了 15 至 20 世纪处于女王统治的国家更容易发动战争, 由于王位继承是内生的, 作者选取了上一代君主第一个孩子的性别以及是否有姊妹作为女王统治的工具变量, 这些因素当然和王位继承有关, 但由于性别因素是由遗传学决定的, 因此生男生女不太可能和误差项相关.

### 6.2.2 结构参数的识别

考虑分割工具变量

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ Z_2 \end{bmatrix}$$

由于  $X_1$  是外生的, 它的工具变量  $Z_1 = X_1$ , 而内生变量  $X_2$  的工具变量  $Z_2$  则需来源于模型外, 它的维数为  $L_2 \times 1$  且  $L_2 \geq K_2$ .

现在来考虑关于内生变量  $X_2$  的简约式方程 (reduced form equation)

$$X_2 = \Gamma'Z + u_2 = \Gamma'_1Z_1 + \Gamma'_2Z_2 + u_2 \quad (6.11)$$

它是上一章所提到的多方程回归系统,  $L \times K_2$  维的系数矩阵  $\Gamma$  由线性投影

$$\Gamma = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'_2] \quad (6.12)$$

这意味着  $\mathbb{E}[Zu'_2] = 0$ , 投影系数在条件 (6.9) 下是唯一的. 现在考虑  $Y$  的简约式方程, 将  $X_1 = Z_1$  和 (6.11) 代入到 (6.7) 中得到

$$\begin{aligned} Y &= Z'_1\beta_1 + (\Gamma'_1Z_1 + \Gamma'_2Z_2 + u_2)'\beta_2 + e \\ &= Z'_1\lambda_1 + Z'_2\lambda_2 + u_1 \\ &= Z'\lambda + u_1 \end{aligned}$$

其中

$$\begin{aligned} \lambda_1 &= \beta_1 + \Gamma_1\beta_2 \\ \lambda_2 &= \Gamma_2\beta_2 \\ u_1 &= u'_2\beta_2 + e \end{aligned}$$

此外, 还可以写作

$$\lambda = \bar{\Gamma}\beta \quad (6.13)$$

这里

$$\bar{\Gamma} = \begin{bmatrix} I_{K_1} & \Gamma_1 \\ 0 & \Gamma_2 \end{bmatrix}$$

以上这些式子刻画了结构参数 ( $\beta_1$  和  $\beta_2$ ) 与简约式参数 ( $\Gamma$  和  $\lambda$ ) 之间紧密的联系. 最后, 我们可以写出这个系统的简约式方程

$$Y = \lambda'Z + u_1$$

$$X_2 = \Gamma'Z + u_2$$

一个参数是可识别的, 如果它能被可观测变量的概率分布唯一确定, 一种表明参数可识别的方法是将其实写为总体矩的显式函数 (第一章所使用过的方法).

如果定义 6.1 中的条件被满足, 那么简约形式的参数  $\Gamma$  和  $\lambda$  是可识别的, 因为它们可以写为变量  $(Y, X, Z)$  总体矩的显式函数

$$\Gamma = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'_2] \quad (6.14)$$

$$\lambda = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZY] \quad (6.15)$$

现在我们对结构参数  $\beta$  感兴趣,  $\beta$  与  $(\Gamma, \lambda)$  的关系通过 (6.13) 联系在一起, 如果能被 (6.13) 唯一确定, 那么  $\beta$  是可识别的. 根据线性代数可知, 当且仅当  $\bar{\Gamma}$  满秩时, 也即

$$\text{rank}(\bar{\Gamma}) = K \quad (6.16)$$

$\beta$  可以由 (6.13) 唯一识别.

利用线性代数, 我们可以将矩阵  $\bar{\Gamma}$  表示为  $\bar{\Gamma} = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']$ , 将它和 (6.15) 代入到

(6.13) 中得到

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

这是一个由  $L$  个方程构成的含  $K$  个未知参数的系统, 当且仅当

$$\text{rank}(\mathbb{E}[ZX']) = K$$

时有唯一解, 这也是定义6.1中的秩条件. 换言之, (6.16) 和 (6.10) 是结构参数  $\beta$  可识别的等价条件.

当  $L < K$  时, 整个方程组无解. 当  $L = K$  时, (6.16) 意味着矩阵  $\bar{\Gamma}$  可逆, 此时  $\beta = \bar{\Gamma}^{-1}\lambda$  是 (6.13) 的唯一解. 当  $L > K$  时,  $\bar{\Gamma}$  不再是方阵, 我们可以对  $\lambda = \bar{\Gamma}\beta$  使用系统最小二乘以获得结构参数  $\beta$  的显式表达  $\beta = (\bar{\Gamma}'\bar{\Gamma})^{-1}\bar{\Gamma}'\lambda$ , 其中 (6.16) 同样保证了矩阵  $\bar{\Gamma}'\bar{\Gamma}$  的可逆性.

此外, 秩条件 (6.10) 意味着工具变量  $Z$  与解释变量  $X$  具有相关性. 假设模型只有一个回归元,  $X = (1, X_1)'$ ,  $Z = (1, W_1)'$ , 那么

$$\text{rank}(\mathbb{E}[ZX']) = 2 \Leftrightarrow \mathbb{E}[Z_1W_1] - \mathbb{E}[Z_1]\mathbb{E}[W_1] \neq 0$$

也即  $Z_1$  和  $X_1$  相关.

总结一下定义6.1的三点要求: 一个合适的工具变量需要与误差项不相关, 但要与解释变量相关, 且工具变量之间不存在完全的线性关系. 为了保证结构参数  $\beta$  可识别, 工具变量的个数至少要与解释变量一样多.

### 6.2.3 工具变量估计量

本节考虑模型恰好识别的特殊情况, 也即工具变量个数同解释变量个数一样多. 此时

$$\beta = \bar{\Gamma}^{-1}\lambda = \mathbb{E}[ZX']^{-1}\mathbb{E}[ZY] \quad (6.17)$$

上式的存在性由 (6.9) 和 (6.10) 保证. 除了用这种方法得到  $\beta$  的表达式外, 还可以通过  $e = Y - X'\beta$  和矩条件  $\mathbb{E}[Ze] = 0$  推导.

结构参数  $\beta$  的工具变量估计量  $\hat{\beta}_{IV}$  可以通过矩估计获得, 也即用样本矩替代 (6.17) 中的总体矩, 于是

$$\begin{aligned} \hat{\beta}_{IV} &= \left( n^{-1} \sum_{i=1}^n Z_i X_i' \right)^{-1} \left( n^{-1} \sum_{i=1}^n Z_i Y_i \right) \\ &= \left( \sum_{i=1}^n Z_i X_i' \right)^{-1} \left( \sum_{i=1}^n Z_i Y_i \right) \end{aligned}$$

此外, 我们还可以使用矩阵符号进行推导

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

这里的  $\mathbf{Z}$  和  $\mathbf{X}$  分别是将观测值堆叠起来的矩阵, 其维数分别为  $n \times L$  和  $n \times K$ . 此外, 我们还

能得到

$$\begin{aligned}\frac{\mathbf{Z}'\mathbf{Z}}{n} &= n^{-1} \sum_{i=1}^n Z_i Z_i' = \hat{\mathbf{Q}}_{ZZ} \\ \frac{\mathbf{X}'\mathbf{Z}}{n} &= n^{-1} \sum_{i=1}^n X_i Z_i' = \hat{\mathbf{Q}}_{XZ} \\ \frac{\mathbf{Z}'\mathbf{X}}{n} &= n^{-1} \sum_{i=1}^n Z_i X_i' = \hat{\mathbf{Q}}_{ZX}\end{aligned}$$

它们将被用于后续的渐近推导.

## 6.3 二阶段最小二乘估计

在上文推导 IV 估计量时, 我们考虑的是  $L = K$  这一特殊条件. 由于  $L > K$  时的矩阵  $\bar{\Gamma}$  不再可逆, 因此无法通过  $\lambda = \bar{\Gamma}\beta$  和矩估计获得  $\beta$  的估计量. 一个直觉是将多余的工具变量扔掉来获得 IV 估计量, 尽管这样是可行的, 但同时也会损失许多信息. 在这种情况下, 应该使用二阶段最小二乘 (Two-Stage Least Squares, 2SLS) 来估计  $\beta$ , 下一章将会证明在条件同方差下, 2SLS 估计量在线性工具变量估计量类中是渐近有效的.

**第一阶段:** 使用  $X$  对  $Z$  进行 SOLS 回归得到拟合值  $\hat{X}$ . 考虑辅助线性回归模型

$$X_i = \Gamma' Z_i + u_i, \quad i = 1, 2, \dots, n$$

将其写为堆叠的矩阵形式

$$\mathbf{X} = \mathbf{Z}\Gamma + \mathbf{u}$$

其中  $\mathbf{X}$  为  $n \times K$  维矩阵,  $\mathbf{Z}$  为  $n \times L$  维矩阵,  $\Gamma$  为  $L \times K$  维矩阵,  $\mathbf{u}$  为  $n \times K$  维随机误差项矩阵. 由此得到  $\Gamma$  的 SOLS 估计量

$$\begin{aligned}\hat{\Gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \\ &= \left( \sum_{i=1}^n Z_i Z_i' \right)^{-1} \left( \sum_{i=1}^n Z_i X_i' \right)\end{aligned}$$

进一步得到拟合值

$$\hat{X}_i = \hat{\Gamma}' X_i$$

用矩阵形式表示为

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

**第二阶段:** 使用  $Y$  对拟合值  $\hat{X}$  进行 SOLS 回归. 第二阶段的回归模型可以写为

$$Y_i = \hat{X}_i \beta + v_i, \quad i = 1, 2, \dots, n$$

可以用矩阵形式表示为

$$\mathbf{Y} = \hat{\mathbf{X}}\beta + \mathbf{v}$$

第二阶段  $\beta$  的 OLS 估计量即为 2SLS 估计量

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y\end{aligned}$$

其中  $\hat{X} = Z\hat{\Gamma} = Z(Z'Z)^{-1}Z'X$ . 注意, 第二阶段回归的残差为  $\hat{v} = Y - \hat{X}\hat{\beta}_{2SLS}$ , 而结构方程中的残差为  $\hat{e} = Y - X\hat{\beta}_{2SLS}$ .

上述推导 2SLS 估计量的方法说明了它名称的由来. 除此之外, 我们还可以从简约式方程

$$Y = Z'\lambda + u_1$$

进行推导, 将  $\lambda = \bar{\Gamma}\beta$  代入上式得到

$$Y = Z'\bar{\Gamma}\beta + u_1$$

$$\mathbb{E}[Zu_1] = 0$$

定义  $W = \bar{\Gamma}'Z$ , 我们可以将上式写为

$$Y = W'\beta + u_1$$

$$\mathbb{E}[Wu_1] = 0$$

由此得到  $\beta$  的最小二乘估计量

$$\hat{\beta} = (W'W)^{-1}W'Y = (\bar{\Gamma}'Z'Z\bar{\Gamma})^{-1}\bar{\Gamma}'Z'Y$$

但因为  $\bar{\Gamma}$  是不可知的, 上述估计量不可行, 考虑使用  $\bar{\Gamma}$  的估计量  $\hat{\Gamma} = (Z'Z)^{-1}Z'X$  替代  $\bar{\Gamma}$ , 最终可以得到 2SLS 估计量

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{\Gamma}'Z'Z\hat{\Gamma})^{-1}\hat{\Gamma}'Z'Y \\ &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y\end{aligned}$$

当模型恰好识别时, 2SLS 估计量等价于 IV 估计量. 这是因为

$$[X'Z(Z'Z)^{-1}Z'X]^{-1} = (Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

于是

$$\begin{aligned}\hat{\beta}_{2SLS} &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (Z'X)^{-1}Z'Y = \hat{\beta}_{IV}\end{aligned}$$

因此 2SLS 估计是 IV 估计的一种推广, 而后面将会看到, 2SLS 估计是生成回归元估计的一个特例.

除此之外, 还能通过 FWL 定理来获得 2SLS 估计量. 首先做分割  $X = [X_1, X_2]$ ,  $Z = [Z_1, Z_2]$ , 注意到  $\hat{X}_1 = P_Z Z_1 = Z_1^1$ , 从而  $\hat{X} = [\hat{X}_1, \hat{X}_2] = [Z_1, \hat{X}_2]$ . 于是  $Z_2$  的结构参数  $\beta_2$  的

<sup>1</sup> 因为  $P_Z Z = Z$ , 而  $Z_1$  是  $Z$  的子矩阵.



2SLS 估计量为

$$\begin{aligned}\hat{\beta}_2 &= [\hat{X}_2'(I_n - P_1)\hat{X}_2]^{-1}\hat{X}_2'(I_n - P_1)Y \\ &= [X_2'P_Z(I_n - P_1)P_ZX_2]^{-1}X_2'P_Z(I_n - P_1)Y \\ &= [X_2'(P_Z - P_1)X_2]^{-1}X_2'(P_Z - P_1)Y\end{aligned}$$

其中  $P_Z P_1 = P_1$ .

另一方面, 设  $\tilde{Z}_2 = (I_n - P_1)Z_2$ , 它与矩阵  $Z_1$  正交, 从而可以将  $P_Z$  正交分解为

$$P_Z = P_1 + P_2$$

其中  $P_2 = \tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'$ , 因此又有

$$\begin{aligned}\hat{\beta}_2 &= (X_2'P_2X_2)^{-1}X_2'P_2'Y \\ &= [X_2'\tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'X_2]^{-1}X_2'\tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'Y\end{aligned}\quad (6.18)$$

表达式 (6.18) 在推导面板数据的 FE2SLS 估计量时会很有用.

## 6.4 2SLS 的渐近性质

### 假设 6.1

- (1)  $\{Y_i, X_i, Z_i\}_{i=1}^n$  是可观测的 i.i.d. 随机样本.
- (2)  $\mathbb{E}[Y_i^2] < \infty$ .
- (3)  $\mathbb{E}\|X_i\|^2 < \infty$ .
- (4)  $\mathbb{E}\|Z_i\|^2 < \infty$ .
- (5)  $\mathbb{E}[Z_i Z_i']$  正定.
- (6)  $\text{rank}(\mathbb{E}[Z_i X_i']) = K$ .
- (7)  $\mathbb{E}[Z_i e_i] = 0$ .

以上假设的第 (1)–(4) 点表明所有变量都具有有限方差, 第 (5) 点排除了线性相关的工具变量, 第 (6) 点是结构参数可识别的秩条件, 第 (7) 点保证工具变量和结构误差项不相关. 其中, 第 (5)–(7) 点等价于定义 6.1.

### 定理 6.1

在假设 6.1 下, 当  $n \rightarrow \infty$  时有  $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$ .

**证明** 注意到

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= \left[ \frac{X'Z}{n} \left( \frac{Z'Z}{n} \right)^{-1} \frac{Z'X}{n} \right]^{-1} \frac{X'Z}{n} \left( \frac{Z'Z}{n} \right)^{-1} \frac{Z'e}{n} \\ &= [\hat{Q}_{XZ} \hat{Q}_{ZZ}^{-1} \hat{Q}_{ZX}]^{-1} \hat{Q}_{XZ} \hat{Q}_{ZZ}^{-1} \frac{Z'e}{n}\end{aligned}$$

根据 WLLN 和 CMT 可知

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{p} (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} \mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbb{E}[Z_i e_i] = 0$$

这里

$$\mathbf{Q}_{XZ} = \mathbb{E}[X_i Z_i']$$

$$\mathbf{Q}_{ZZ} = \mathbb{E}[Z_i Z_i']$$

$$\mathbf{Q}_{ZX} = \mathbb{E}[Z_i X_i']$$

其中, 假设 6.1(1) 和 (2)–(4) 保证 WLLN 成立, (5)–(6) 保证 CMT 成立, (7) 保证了最终结论成立.

**注** 尽管 2SLS 估计量在大样本下可以具有一致性, 但在有限样本中它永远不会是无偏估计量. 从这个角度看, 工具变量的数目绝不是越多越好, 因为其数目越多, 出现同质性的工具变量的概率越大, 无法提供更多有效的信息, 反而会增大有限样本偏差. 此外, Kinal (1980) 的结果还表明, 在假设 6.1 和恰好识别情况下的 IV 估计量不具有期望值.

类似地, 如果要将 CLT 应用到 2SLS 估计量上, 我们还需要强化假设 6.1.

### 假设 6.2

在假设 6.1 的基础上, 以下额外条件成立:

- (1)  $\mathbb{E}[Y_i^4] < \infty$ .
- (2)  $\mathbb{E}[\|X_i\|^4] < \infty$ .
- (3)  $\mathbb{E}[\|Z_i\|^4] < \infty$ .
- (4)  $\mathbf{\Omega} = \mathbb{E}[Z_i Z_i' e_i^2]$  正定.

### 定理 6.2

在假设 6.2 下, 当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta)$$

其中

$$\mathbf{V}_\beta = (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{\Omega} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1} (\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1}$$

**证明** 首先写出

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = [\hat{\mathbf{Q}}_{XZ} \hat{\mathbf{Q}}_{ZZ}^{-1} \hat{\mathbf{Q}}_{ZX}]^{-1} \hat{\mathbf{Q}}_{XZ} \hat{\mathbf{Q}}_{ZZ}^{-1} \frac{\mathbf{Z}' e}{\sqrt{n}}$$

根据 Minkowski 不等式可知

$$(\mathbb{E}[e_i^4])^{1/4} \leq (\mathbb{E}[Y_i^4])^{1/4} + \|\beta\|(\mathbb{E}[\|X_i\|^4])^{1/4} < \infty$$

又由 Cauchy-Schwarz 不等式得到

$$\mathbb{E}[\|Z_i e_i\|^2] \leq (\mathbb{E}[\|Z_i\|^4])^{1/2} (\mathbb{E}[e_i^4])^{1/2} < \infty$$

于是由多元 Lindeberg-Levy CLT 的条件成立, 因此

$$\frac{\mathbf{Z}'\mathbf{e}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i e_i \xrightarrow{d} N(0, \mathbf{\Omega})$$

其中  $\mathbf{\Omega} = \mathbb{E}[\mathbf{Z}_i \mathbf{Z}_i' e_i^2]$ . 最后根据 Slutsky 定理得到

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta)$$

结论成立.

特别地, 如果同方差假定成立

$$\mathbb{E}[\mathbf{Z}_i \mathbf{Z}_i' e_i^2] = \sigma^2 \mathbb{E}[\mathbf{Z}_i \mathbf{Z}_i'] \quad (6.19)$$

那么  $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta^0)$ , 这里  $\mathbf{V}_\beta^0 = \sigma^2(\mathbf{Q}_{XZ} \mathbf{Q}_{ZZ}^{-1} \mathbf{Q}_{ZX})^{-1}$ . 同之前的讨论一样, 2SLS 的协方差矩阵的一致估计量为

$$\begin{aligned} \hat{\mathbf{V}}_{2SLS} &= (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \left( \sum_{i=1}^n \hat{X}_i \hat{X}_i' \hat{e}_i^2 \right) (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \\ \hat{\mathbf{V}}_{2SLS}^0 &= s^2 (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \end{aligned}$$

其中  $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ ,  $\hat{e}_i = Y_i - \hat{X}_i' \hat{\beta}_{2SLS}$ , 以及  $s^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$ .

现在来看参数检验, 考虑线性假设  $\mathbb{H}_0: \mathbf{R}\beta = r$ , 在假设 6.2 和 4.3 下, 如果原假设  $\mathbb{H}_0$  为真, 那么当  $n \rightarrow \infty$  时, 稳健 Wald 检验统计量

$$W = n(\mathbf{R}\hat{\beta}_{2SLS} - r)' (\mathbf{R}\hat{\mathbf{Q}}^{-1} \hat{\mathbf{\Omega}} \hat{\mathbf{Q}}^{-1} \mathbf{R}')^{-1} (\mathbf{R}\hat{\beta}_{2SLS} - r) \xrightarrow{d} \chi_J^2$$

其中  $\hat{\mathbf{Q}} = n^{-1} \sum_{i=1}^n \hat{X}_i \hat{X}_i'$ ,  $\hat{\mathbf{\Omega}} = n^{-1} \sum_{i=1}^n \hat{X}_i \hat{X}_i' \hat{e}_i^2$ . 如果条件同方差假设成立, 那么 Wald 检验统计量可以简化为

$$W = \frac{(\mathbf{R}\hat{\beta}_{2SLS} - r)' [\mathbf{R}(\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta}_{2SLS} - r)}{\hat{e}' \hat{e} / (n - K)} \xrightarrow{d} \chi_J^2$$

其中  $\hat{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{2SLS}$ .

尽管 2SLS 框架并没有对内生变量的分布做任何限制, 内生变量既可以是连续的, 也可以是离散的, 这些都不会影响到 2SLS 的一致性. 然而, 使用连续的外生变量作为二元内生变量的工具可能导致弱工具变量问题.

由于 2SLS 一阶段采用的是 OLS 回归, 而内生变量的取值只有 0 和 1, 有的人可能会在 2SLS 的一阶段使用 Probit 模型来获得拟合值, 并将其纳入到第二阶段回归. Angrist and Pischke (2009) 指出这是禁止回归 (forbidden regression), 会导致第二阶段的估计量不一致, 主要原因在于, 期望算子不能穿过非线性函数, 也即  $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ .

然而, 通过 Probit 模型得到拟合值, 并将其作为虚拟内生变量的工具变量, 然后再施行通常的 2SLS 方法是可行的, 特别是当非线性模型对一阶段的 CEF 有更好的近似时, 通过非线性拟合值得到的 2SLS 会比使用线性一阶段的传统 2SLS 更有效 (Angrist and Pischke, 2009; Newey, 1990).

## 6.5 生成回归元

获取 2SLS 估计量的整个流程实际上是“生成回归元 (generated regressors) 估计”的一个例子。我们称回归元是被生成的 (generated), 如果它是理想 (idealized) 回归元的估计量或者它是待估参数的函数。

通常, 一个生成回归元  $\hat{W}$  作为不可观测的理想回归元  $W$  的一个估计量, 那么  $\hat{W}_i$  应该是整个样本的函数, 而非仅是观测  $i$  的函数。由此可见,  $\hat{W}_i$  不再是 i.i.d. 的, 因为  $\hat{W}_i$  的实现值依赖于不同的观测, 这使得经典的 i.i.d. 随机样本假设无效。因此, 回归估计量的样本分布受到影响, 协方差矩阵和标准误的估计将不正确。

生成回归元的计量经济理论由 Pagan (1984) 提出。这里我们关注以下线性模型

$$Y = W'\beta + v \quad (6.20)$$

$$W = A'Z$$

$$\mathbb{E}[Zv] = 0$$

其中可观测值是  $[Y, Z]$ , 并且我们有  $A$  的估计量  $\hat{A}$ 。此时我们可以构造  $W_i$  的生成回归元  $\hat{W}_i = \hat{A}'Z_i$ , 然后在 (6.20) 中用  $\hat{W}_i$  取代  $W_i$ , 由此可以得到  $\beta$  的 OLS 估计量

$$\hat{\beta} = \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{W}_i Y_i \right) \quad (6.21)$$

由于生成回归元本身就是估计量, 因此这里  $\hat{\beta}$  的统计性质和 i.i.d. 随机样本下的 OLS 估计量的不同。

现在回到简约式方程  $Y = Z'\lambda + u_1$ , 定义  $W = \bar{\Gamma}'Z$ ,  $A = \bar{\Gamma}$  以及  $\hat{A} = \hat{\Gamma}$ , 由此可见生成回归元这一框架包含了 2SLS 估计量。

现在的目标是取得  $\hat{\beta}$  的近似分布。首先将 (6.20) 代入到 (6.21) 中得到

$$\hat{\beta} = \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1} \left[ \sum_{i=1}^n \hat{W}_i (W_i' \beta + v_i) \right]$$

注意到  $W_i' \beta = \hat{W}_i' \beta + (W_i - \hat{W}_i)' \beta$ , 于是有

$$\hat{\beta} - \beta = \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1} \left\{ \sum_{i=1}^n \hat{W}_i [(W_i - \hat{W}_i)' \beta + v_i] \right\} \quad (6.22)$$

上式由两个随机项构成, 其中一项可由经典回归理论解决, 而另一项则是生成回归元的部分, 经典回归理论无法解决。

倘若可以使  $(W_i - \hat{W}_i)' \beta$  这部分消失, 那么 (6.22) 将会得到简化。为了使它成立, 就需要生成回归元的回归系数为 0, 由此需要将  $W_i$  进行分割来获得生成回归元那部分的回归系数。

具体而言, 做分割  $W_i = [W_{1i}', W_{2i}']'$ ,  $\hat{W}_i = [\hat{W}_{1i}', \hat{W}_{2i}']'$ , 其中  $W_{1i}$  是可观测的向量,  $\hat{W}_{2i}$  是生成回归元, 再将  $\beta$  分割为  $\beta = [\beta_1', \beta_2']'$ , 此时  $(W_i - \hat{W}_i)' \beta = (W_{2i} - \hat{W}_{2i})' \beta_2$ 。如果  $\beta_2 = 0$  成立,

那么

$$\hat{\beta} - \beta = \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{W}_i v_i \right)$$

进一步, 由于  $\hat{W}_i = \hat{A}' Z_i$ , 故而我们可以将估计量写为样本矩的函数

$$\sqrt{n}(\hat{\beta} - \beta) = \left[ \hat{A}' \left( n^{-1} \sum_{i=1}^n Z_i Z_i' \right) \hat{A} \right]^{-1} \hat{A}' \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i v_i \right)$$

如果  $\hat{A} \xrightarrow{p} A$ , 那么按照标准流程可知  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$ , 其中

$$V_\beta = (A' \mathbb{E}[Z_i Z_i'] A)^{-1} (A' \mathbb{E}[Z_i Z_i' v_i^2] A) (A' \mathbb{E}[Z_i Z_i'] A)^{-1} \quad (6.23)$$

它的渐近协方差矩阵估计量为

$$\hat{V}_\beta = n \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \hat{v}_i^2 \right) \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \right)^{-1}$$


其中  $\hat{v}_i = Y_i - \hat{W}_i' \hat{\beta}$ , 在合适的正则条件下,  $\hat{V}_\beta \xrightarrow{p} V_\beta$ .

最后, 为了检验假设  $\mathbb{H}_0: \beta_2 = 0$ , 我们可以构建 Wald 检验统计量

$$W = n \hat{\beta}_2' [\hat{V}_\beta]_{22}^{-1} \hat{\beta}_2$$

当  $n \rightarrow \infty$  时有  $W \xrightarrow{d} \chi_q^2$ , 其中  $q = \dim(\beta_2)$ .

### 定理 6.3

对于模型 (6.20), 如果  $\mathbb{E}[Y_i^4] < \infty$ ,  $\mathbb{E}||Z_i||^4 < \infty$ ,  $A' \mathbb{E}[Z_i Z_i'] A > 0$ ,  $\hat{A} \xrightarrow{p} A$ , 以及  $\hat{W}_i = (W_{1i}', \hat{W}_{2i}')'$  那么在  $\mathbb{H}_0: \beta_2 = 0$  的情况下, 当  $n \rightarrow \infty$  时有: (1)  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$ , 这里的  $V_\beta$  在 (6.23) 中已经给出. (2) Wald 检验统计量  $W \xrightarrow{d} \chi_q^2$ , 其中  $q = \dim(\beta_2)$ . 

尽管上述理论允许我们检验假设  $\mathbb{H}_0: \beta_2 = 0$ , 但它并没有修正  $\hat{\beta}$  的标准误, 如果我们需要在不施加简化条件  $\mathbb{H}_0: \beta_2 = 0$  的情况下得到  $\hat{\beta}$  的渐近分布, 还需要用到后续章节的广义矩估计方法.

现在考虑另外一种特殊情形, 估计量  $\hat{A}$  可以写成与回归元  $X$  相关的最小二乘形式  $\hat{A} = (Z'Z)^{-1} Z'X$ , 这样的估计量与以下回归系统紧密相关

$$X = A'Z + u \quad (6.24)$$

$$\mathbb{E}[Zu'] = 0$$

显然, 这类估计量  $\hat{A}$  包含了 2SLS 的特殊情形<sup>2</sup>. 此时我们将生成回归元记作  $\hat{W} = Z\hat{A}$ , 首先由

$$\hat{W} = Z(Z'Z)^{-1} Z'(ZA + U)$$

<sup>2</sup> $\hat{\Gamma} = (Z'Z)^{-1} Z'X$  能写成这种形式, 但这种形式的估计量  $\hat{A}$  不一定来源于 2SLS 估计.

可知  $W - \hat{W} = -Z(Z'Z)^{-1}Z'U$ , 于是

$$\begin{aligned}\hat{\beta} - \beta &= (\hat{W}'\hat{W})^{-1}\{\hat{W}'[(W - \hat{W})\beta + v]\} \\ &= (\hat{A}'Z'Z\hat{A})^{-1}\{\hat{A}'Z'[-Z(Z'Z)^{-1}Z'U\beta + v]\} \\ &= (\hat{A}'Z'Z\hat{A})^{-1}[\hat{A}'Z'(-U\beta + v)] \\ &= (\hat{A}'Z'Z\hat{A})^{-1}\hat{A}'Z'e\end{aligned}$$

其中

$$e_i = v_i - u_i'\beta = Y_i - X_i'\beta$$

在合适的正则条件下有  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$ , 其中

$$V_\beta = (A'\mathbb{E}[Z_i Z_i']A)^{-1}(A'\mathbb{E}[Z_i Z_i' e_i^2]A)(A'\mathbb{E}[Z_i Z_i']A)^{-1}$$

特别地, 如果条件同方差假设  $\mathbb{E}[Z_i Z_i' e_i^2] = \sigma^2 \mathbb{E}[Z_i Z_i']$  成立, 那么


$$V_\beta^0 = \sigma^2 (A'\mathbb{E}[Z_i Z_i']A)^{-1}$$

其中  $\mathbb{E}[e_i^2] = \sigma^2$ . 协方差矩阵  $V_\beta$  的估计量为

$$\hat{V}_\beta = n(\hat{W}'\hat{W})^{-1} \left( \sum_{i=1}^n \hat{W}_i \hat{W}_i' \hat{e}_i^2 \right) (\hat{W}'\hat{W})^{-1}$$

其中  $\hat{e}_i = Y_i - X_i'\hat{\beta}$ . 在合适的正则条件下,  $\hat{V}_\beta$  是  $V_\beta$  的一致估计量.

#### 定理 6.4

考虑模型 (6.20) 和 (6.24), 如果  $\mathbb{E}[Y_i^4] < \infty$ ,  $\mathbb{E}||Z_i||^4 < \infty$ ,  $A'\mathbb{E}[Z_i Z_i']A > 0$ ,  $\hat{A} = (Z'Z)^{-1}Z'X$ , 以及  $\hat{A} \xrightarrow{p} A$ , 那么当  $n \rightarrow \infty$  时有  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$ , 并且  $\hat{V}_\beta \xrightarrow{p} V_\beta$ . 

总结一下, 生成回归元和两步估计会影响到估计量的样本分布和协方差矩阵. 一种简化条件是生成回归元的回归系数为 0, 此时经典的样本分布和渐近理论均适用; 另一种重要情况是生成回归元可以写作最小二乘拟合值的形式, 此时估计量的渐近分布仍是经典的.

## 6.6 含期望误差的回归

本节在生成回归元的基础上做进一步扩展, 纳入了一个期望误差项, 具体而言就是

$$Y = W'\beta + u'\alpha + v \quad (6.25)$$

$$W = A'Z$$

$$X = W + u$$

$$\mathbb{E}[Zv] = 0$$

$$\mathbb{E}[uv] = 0$$

$$\mathbb{E}[Zu'] = 0$$

其中  $[Y, X, Z]$  是可观测变量,  $W$  是  $X$  在  $Z$  上的投影,  $u$  为期望误差.

在上述模型中, 矩阵  $A$  的估计量可以由  $X$  对  $Z$  回归得到, 也即  $\hat{A} = (Z'Z)^{-1}Z'X$ , 于是可以计算得到拟合值  $\hat{W}_i = \hat{A}'Z_i$  以及残差  $\hat{u}_i = X_i - \hat{W}_i$ , 最后使用  $Y$  对拟合值  $\hat{W}$  及残差  $\hat{u}$  回归即可估计. 此时

$$Y_i = \hat{W}_i'\beta + \hat{u}_i'\alpha + v_i, \quad i = 1, 2, \dots, n$$

通过第一阶段回归可得  $Z'\hat{U} = 0$ ,  $\hat{W}'\hat{U} = 0$  以及  $W'\hat{U} = 0$ , 于是根据 (2.8) 和 (2.9) 可知

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

注意到

$$Y = \hat{W}\beta + U\alpha + (W - \hat{W})\beta + v$$

通过  $W - \hat{W} = -Z(Z'Z)^{-1}Z'U$  可知

$$\begin{aligned} \hat{\beta} - \beta &= (\hat{W}'\hat{W})^{-1}\hat{W}'[U\alpha + (W - \hat{W})\beta + v] \\ &= (\hat{A}'Z'Z\hat{A})^{-1}\hat{A}'Z'(U\alpha - U\beta + v) \\ &= (\hat{A}'Z'Z\hat{A})^{-1}\hat{A}'Z'e \end{aligned}$$

其中

$$e_i = v_i + u_i'(\alpha - \beta) = Y_i - X_i'\beta$$

另一方面

$$\hat{\alpha} = (\hat{U}'\hat{U})^{-1}\hat{U}'Y$$

又因为  $\hat{U}$  是出自回归  $X = ZA + U$  的残差, 因此

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

于是由  $\hat{U}'W = 0$  和  $\hat{U}'Z = 0$  可知

$$\begin{aligned} \hat{\alpha} - \alpha &= (\hat{U}'\hat{U})^{-1}\hat{U}'[W\beta + (U - \hat{U})\alpha + v] \\ &= (\hat{U}'\hat{U})^{-1}\hat{U}'v \end{aligned}$$

现在我们给出以下定理.

#### 定理 6.5

考虑模型 (6.25), 如果  $\mathbb{E}[Y_i^4] < \infty$ ,  $\mathbb{E}||Z_i||^4 < \infty$ ,  $\mathbb{E}||X_i||^4 < \infty$ ,  $A\mathbb{E}[Z_iZ_i']A' > 0$ , 以及  $\mathbb{E}[u_iu_i'] > 0$ , 那么当  $n \rightarrow \infty$  时

$$\sqrt{n} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{\alpha} - \alpha \end{bmatrix} \xrightarrow{d} N(0, V)$$

其中

$$V = \begin{bmatrix} V_{\beta\beta} & V_{\beta\alpha} \\ V_{\alpha\beta} & V_{\alpha\alpha} \end{bmatrix}$$

以及

$$\begin{aligned} V_{\beta\beta} &= (A' \mathbb{E}[Z_i Z_i'] A)^{-1} (A' \mathbb{E}[Z_i Z_i' e_i^2] A) (A' \mathbb{E}[Z_i Z_i'] A)^{-1} \\ V_{\alpha\beta} &= \mathbb{E}[u_i u_i']^{-1} (\mathbb{E}[u_i Z_i' e_i v_i] A) (A' \mathbb{E}[Z_i Z_i'] A)^{-1} \\ V_{\alpha\alpha} &= \mathbb{E}[u_i u_i']^{-1} \mathbb{E}[u_i u_i' v_i^2] \mathbb{E}[u_i u_i']^{-1} \end{aligned}$$



按照之前的流程, 容易写出估计量  $\hat{\alpha}$  和  $\hat{\beta}$  的协方差矩阵

$$\begin{aligned} \hat{V}_{\alpha\alpha} &= n(\hat{W}' \hat{W})^{-1} \left( \sum_{i=1}^n \hat{w}_i \hat{w}_i' \hat{e}_i^2 \right) (\hat{W}' \hat{W})^{-1} \\ \hat{V}_{\beta\beta} &= n(\hat{U}' \hat{U})^{-1} \left( \sum_{i=1}^n \hat{u}_i \hat{u}_i' \hat{v}_i^2 \right) (\hat{U}' \hat{U})^{-1} \end{aligned}$$

特别地, 如果同方差假设成立

$$\mathbb{E} \left[ \begin{pmatrix} e_i^2 & e_i v_i \\ e_i v_i & v_i^2 \end{pmatrix} \middle| Z_i \right] = C$$

那么渐近协方差矩阵可以简化为

$$\begin{aligned} V_{\beta\beta} &= (A' \mathbb{E}[Z_i Z_i'] A)^{-1} \mathbb{E}[e_i^2] \\ V_{\alpha\alpha} &= \mathbb{E}[u_i u_i']^{-1} \mathbb{E}[v_i^2] \end{aligned}$$

## 6.7 控制函数法

本节介绍另外一种通过最小二乘来得到 2SLS 估计量的方法, 称为控制函数法 (Control Function Approach, CFA), 该方法在非线形回归领域尤其有用. 关于 CFA 的更多内容具体可见 Wooldridge (2015).

首先写出以下结构式和简约式方程

$$\begin{aligned} Y &= X_1' \beta_1 + X_2' \beta_2 + e \\ X_2 &= \Gamma_1' Z_1 + \Gamma_2' Z_2 + u_2 \end{aligned}$$

其中  $X_2$  为内生变量, 通过  $u_2$  和  $e$  的相关性导致内生性,  $Z$  为符合定义 6.1 的工具变量. 现在考虑  $e$  在  $u_2$  上的线性投影

$$\begin{aligned} e &= u_2' \alpha + v \\ \alpha &= \mathbb{E}[u_2 u_2']^{-1} \mathbb{E}[u_2 e] \\ \mathbb{E}[u_2 v] &= 0 \end{aligned}$$



将其代入到结构方程中得到

$$\begin{aligned} Y &= X_1' \beta_1 + X_2' \beta_2 + u_2' \alpha + v \\ \mathbb{E}[X_1 v] &= 0 \\ \mathbb{E}[X_2 v] &= 0 \\ \mathbb{E}[u_2 v] &= 0 \end{aligned} \quad (6.26)$$

注意, 这里  $X_2$  和  $v$  不相关, 这是因为  $X_2$  仅通过  $u_2$  和  $e$  产生相关性, 而  $v$  是  $e$  在  $u_2$  上正交投影后产生的误差.

如果  $u_2$  是可观测的, 那么可以使用最小二乘来估计方程 (6.26). 但由于它实际不可观测, 我们考虑使用简约式残差  $\hat{u}_{2i} = X_{2i} - \hat{\Gamma}_1' Z_{1i} - \hat{\Gamma}_2' Z_{2i}$  替代  $u_{2i}$ , 回归系数  $[\beta_1, \beta_2, \alpha]$  的估计量可由  $Y$  对  $[X_1, X_2, \hat{u}_2]$  的最小二乘获得. 此时

$$Y_i = X_i' \beta + \hat{u}_{2i}' \alpha + \hat{v}_i \quad (6.27)$$

也可以用矩阵形式表示为

$$Y = X\beta + \hat{U}_2\alpha + v$$

下面将会看到这里的  $\hat{\beta}$  在代数形式上和  $\hat{\beta}_{2SLS}$  是一样的.

定义幂等矩阵

$$P_Z = Z(Z'Z)^{-1}Z'$$

于是简约式残差可以记作

$$\hat{U}_2 = (I_n - P_Z)X_2$$

根据 FWL 定理可知

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y \quad (6.28)$$

其中  $\tilde{X} = [\tilde{X}_1, \tilde{X}_2]$ , 并且

$$\tilde{X}_1 = X_1 - \hat{U}_2(\hat{U}_2'\hat{U}_2)^{-1}\hat{U}_2'X_1 = X_1$$

上式最后一个等号成立是因为  $\hat{U}_2'X_1 = 0$ . 另一方面

$$\begin{aligned} \tilde{X}_2 &= X_2 - \hat{U}_2(\hat{U}_2'\hat{U}_2)^{-1}\hat{U}_2'X_2 \\ &= X_2 - \hat{U}_2'[X_2'(I_n - P_Z)X_2]^{-1}X_2'(I_n - P_Z)X_2 \\ &= X_2 - \hat{U}_2 = P_Z X_2 \end{aligned}$$

因此  $\tilde{X} = [X_1, P_Z X_2] = P_Z X$ , 将它代入到 (6.28) 中得到

$$\hat{\beta} = (X'P_Z X)^{-1}X'P_Z Y = \hat{\beta}_{2SLS}$$

然而, 对于非线性模型, 例如

$$\begin{aligned} Y &= Z_1' \delta + \alpha_1 X + \alpha_2 X^2 + e \\ \mathbb{E}[e|Z] &= 0 \end{aligned} \quad (6.29)$$

其中  $Z = [Z_1', Z_2']'$  是外生解释变量,  $X$  是内生解释变量, 并且我们有一个不在  $Z_1$  中的  $Z_2$  作

为  $X$  的工具变量. 于是利用  $[Z_1, Z_2, Z_2^2]$  即可实施标准的 IV 估计, 并且它是一致的. 而如果考虑使用 CF 方法, 则需要做出比 (6.29) 更强的假设, 此时 2SLS 和 CF 得到的估计量不相同, 并且 CF 估计量通常更有效, 但也更缺乏稳健性.

现在我们来考虑 CF 估计量  $[\hat{\beta}, \hat{\alpha}]$  的分布. 令  $W = \bar{\Gamma}'Z$ , 以及  $u = X - W^3$ , 于是回归方程 (6.26) 变为

$$Y = W'\beta + u_2'\gamma + v \quad (6.30)$$

其中  $\gamma = \alpha + \beta_2$ . 作为生成回归元模型的一种, 根据定理 6.5 可知, 如果相关正则条件成立, 那么当  $n \rightarrow \infty$  时有

$$\sqrt{n} \begin{bmatrix} \hat{\beta}_2 - \beta_2 \\ \hat{\gamma} - \gamma \end{bmatrix} \xrightarrow{d} N(0, V)$$

其中

$$V = \begin{bmatrix} V_{22} & V_{2\gamma} \\ V_{\gamma 2} & V_{\gamma\gamma} \end{bmatrix}$$

以及

$$V_{22} = [(\bar{\Gamma}'\mathbb{E}[Z_i Z_i']\bar{\Gamma})^{-1}\bar{\Gamma}'\mathbb{E}[Z_i Z_i' e_i^2]\bar{\Gamma}(\bar{\Gamma}'\mathbb{E}[Z_i Z_i']\bar{\Gamma})^{-1}]_{22}$$

$$V_{\gamma 2} = [\mathbb{E}[u_{2i} u_{2i}']^{-1}\mathbb{E}[u_i Z_i e_i v_i]\bar{\Gamma}(\bar{\Gamma}'\mathbb{E}[Z_i Z_i']\bar{\Gamma})^{-1}]_{\gamma 2}$$

$$V_{\gamma\gamma} = \mathbb{E}[u_{2i} u_{2i}']^{-1}\mathbb{E}[u_{2i} u_{2i}' v_i^2]\mathbb{E}[u_{2i} u_{2i}']^{-1}$$

$$e_i = Y_i - X_i'\beta$$

由此可以推断出  $\hat{\alpha} = \hat{\gamma} - \hat{\beta}_2$  的渐近分布.

#### 定理 6.6

考虑模型 (6.30), 如果  $\mathbb{E}[Y_i^4] < \infty$ ,  $\mathbb{E}\|Z_i\|^4 < \infty$ ,  $\mathbb{E}\|X_i\|^4 < \infty$ ,  $A\mathbb{E}[Z_i Z_i']A' > 0$ , 以及  $\mathbb{E}[u_i u_i'] > 0$ , 那么当  $n \rightarrow \infty$  时

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, V_\alpha)$$

其中

$$V_\alpha = V_{22} + V_{\gamma\gamma} - V_{\gamma 2} - V_{2\gamma}'$$



## 6.8 模型设定检验

### 6.8.1 内生性检验

上文已经提到, 当线性回归模型  $Y = X'\beta + e$  存在内生性时, 我们可以使用 2SLS 估计得到一致估计量, 但模型是否存在内生性往往难以判断, 本节给出 Hausman (1978) 提出的检验内生性的方法.

<sup>3</sup>这里  $u$  可以被分割为  $u = \begin{bmatrix} 0 \\ u_2 \end{bmatrix}$ , 分块矩阵的维数分别为  $K_1 \times 1$  和  $K_2 \times 1$ .

**定理 6.7**

在假设 6.2 和同方差假设  $\mathbb{E}[e_i^2|X_i, Z_i] = \sigma^2$  下, 如果原假设  $\mathbb{H}_0: \mathbb{E}[e_i|X_i] = 0$  成立, 那么当  $n \rightarrow \infty$  时有

$$H = \frac{n(\hat{\beta}_{2SLS} - \hat{\beta})'[(\hat{Q}_{XZ}\hat{Q}_{ZZ}^{-1}\hat{Q}_{ZX})^{-1} - \hat{Q}_{XX}^{-1}](\hat{\beta}_{2SLS} - \hat{\beta})}{s^2} \xrightarrow{d} \chi_K^2$$

其中  $s^2 = \hat{e}'\hat{e}/n$ ,  $\hat{e} = Y - X\hat{\beta}$  为 OLS 估计残差,  $-$  表示广义逆.



**证明** 首先回顾 OLS 估计量

$$\sqrt{n}(\hat{\beta} - \beta) = \hat{Q}_{XX}^{-1}n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i$$

根据 CLT 可知

$$n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i \xrightarrow{d} N(0, \sigma^2 Q_{XX}^{-1})$$

其中  $\mathbb{E}[e_i^2] = \sigma^2$ . 注意到  $n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i = o_p(1)$  以及  $\hat{Q}_{XX}^{-1} \xrightarrow{p} Q_{XX}^{-1}$ , 于是

$$\sqrt{n}(\hat{\beta} - \beta) = Q_{XX}^{-1}n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i + o_p(1)$$

同理可知

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{2SLS} - \beta) &= \hat{A}n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i \\ &= An^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i + o_p(1)\end{aligned}$$

其中

$$\hat{A} = (\hat{Q}_{XZ}\hat{Q}_{ZZ}^{-1}\hat{Q}_{ZX})^{-1}\hat{Q}_{XZ}\hat{Q}_{ZZ}^{-1} \xrightarrow{p} A = (Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}Q_{XZ}Q_{ZZ}^{-1}$$

并且  $n^{-\frac{1}{2}}\sum_{i=1}^n X_i e_i \xrightarrow{d} N(0, \sigma^2 Q_{ZZ})$ .

如果原假设  $\mathbb{H}_0: \mathbb{E}[e_i|X_i] = 0$  成立, 那么当  $n \rightarrow \infty$  时有

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{2SLS} - \hat{\beta}) &= \sqrt{n}(\hat{\beta}_{2SLS} - \beta) - \sqrt{n}(\hat{\beta} - \beta) \\ &= n^{-\frac{1}{2}}\sum_{i=1}^n [(Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1}Q_{XZ}Q_{ZZ}^{-1}Z_i - Q_{XX}^{-1}X_i]e_i + o_p(1) \\ &\xrightarrow{d} N[0, \sigma^2(Q_{XZ}Q_{ZZ}^{-1}Q_{ZX})^{-1} - \sigma^2 Q_{XX}^{-1}]\end{aligned}$$

于是根据 Slutsky 定理可知  $H \xrightarrow{d} \chi_K^2$ .

**注** 尽管 Hausman 检验要求同方差假设成立, 但是构造一个异方差稳健的 Hausman 检验统计量也是可能的, 只是表达式会异常复杂.

另一方面, 我们还可以通过 CF 法来检验内生性. 考虑回归模型

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

我们想检验原假设  $\mathbb{H}_0 : \mathbb{E}[X_2 e] = 0$ , 它等价于  $\mathbb{H}_0 : \mathbb{E}[u_2 e] = 0$ . 根据 CF 法可建立以下模型

$$Y_i = X_i' \beta + \hat{u}_{2i} \alpha + v_i$$

又因为

$$\alpha = \mathbb{E}[u_2 u_2']^{-1} \mathbb{E}[u_2 e]$$

根据定理 6.6, 我们可以构造一个 Wald 检验统计量

$$W = n \hat{\alpha}' [\hat{V}_\alpha]^{-1} \hat{\alpha}$$

以检验假设  $\mathbb{H}_0 : \alpha = 0$ , 并且当  $n \rightarrow \infty$  时有  $W \xrightarrow{d} \chi_{K_2}^2$ .

最后我们强调一点, 这里的内生性检验本质上检验的是 OLS 估计量和 2SLS 估计量 (或 CF 估计量) 在统计意义上是否相距很远. 换言之, Hausman 检验的原理是, 如果模型不存在内生性, 那么这些估计量在数值上应该是差不多的.

## 6.8.2 过度识别检验

除了检验内生性外, 我们还需要检验工具变量的外生性  $\mathbb{E}[Z e] = 0$ . 由于在恰好识别的情况下无法检验, 因此我们将重点放在过度识别检验上. 在同方差假设成立的情况下, 我们可以构造 Sargan (1958) 提出的检验统计量进行检验.

### 定理 6.8

在假设 6.2 和同方差  $\mathbb{E}[e_i^2 | Z_i] = \sigma^2$  下, 如果原假设  $\mathbb{H}_0 : \mathbb{E}[Z_i | e_i] = 0$  成立, 那么当  $n \rightarrow \infty$  时有

$$S = \frac{\hat{e}' Z (Z' Z)^{-1} Z' \hat{e}}{\hat{e}' \hat{e} / n} \xrightarrow{d} \chi_{L-K}^2$$

其中  $\hat{e} = Y - X \hat{\beta}_{2SLS}$ .



**证明** 首先根据 CLT 可知

$$n^{-\frac{1}{2}} \sum_{i=1}^n Z_i e_i \xrightarrow{d} N(0, \Omega)$$

其中  $\Omega = \mathbb{E}[Z_i Z_i' e_i^2] = \sigma^2 \mathbb{E}[Z_i Z_i']$ , 以及  $\sigma^2 = \mathbb{E}[e_i^2]$ . 定义  $\Omega$  的一致估计量

$$\hat{\Omega} = n^{-2} \hat{e}' \hat{e} Z' Z$$

从而  $\hat{\Omega}^{-1} = n^2 (Z' Z)^{-1} / \hat{e}' \hat{e}$ . 根据线性代数可知, 存在矩阵  $\hat{\Omega}^{-\frac{1}{2}}$ , 使得  $\hat{\Omega}^{-1} = \hat{\Omega}^{-\frac{1}{2}} \hat{\Omega}^{-\frac{1}{2}}$ , 故而

$$n^{-\frac{1}{2}} \hat{\Omega}^{-\frac{1}{2}} \sum_{i=1}^n Z_i e_i \xrightarrow{d} N(0, I_l)$$

根据引理 2.1 可知

$$n^{-1} \hat{e}' Z \hat{\Omega}^{-1} Z' \hat{e} \xrightarrow{d} \chi_{L-K}^2$$

也即  $S \xrightarrow{d} \chi_{L-K}^2$ .

当样本容量不大时, Sargan 检验通常会过于拒绝原假设. 此外, Sargan 检验要求条件同方

差假设成立, 如果该假设不成立, 那么我们需要使用 GMM 过度识别检验, Wooldridge (1995) 还介绍了一种在过度识别下的稳健得分检验, 它在数值上同 GMM 过度识别检验统计量相等.

注意, 即便过度识别检验没有拒绝原假设, 也不能说明所有工具变量都是外生的, 只能说明没有找到它们是内生的证据.

### 6.8.3 弱工具变量

最后我们关注弱工具变量 (weak instruments) 问题. 如果工具变量  $Z$  和内生变量  $X$  完全不相关, 则无法识别结构参数. 如果二者仅微弱地相关, 那么  $\mathbb{E}[ZX']^{-1}$  将会很大, 从而影响 2SLS 估计量的渐近方差, 以及导致估计量很不准确并足以使得一切都不显著.

为了简化问题, 我们假设模型中没有外生变量, 此时将  $X_2$ ,  $Z_2$  和  $\beta_2$  分别简写为  $X$ ,  $Z$  和  $\beta$ , 模型为

$$\begin{aligned} Y &= X'\beta + e \\ X &= \Gamma'Z + u_2 \end{aligned}$$

记  $u = [u_1, u_2']'$ , 协方差矩阵为

$$\mathbb{E}[u_i u_i'] = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

结构误差  $e = u_1 - u_2'\beta = u'\gamma$ , 其中  $\gamma = [1, -\beta']'$ . 于是可以定义  $\mathbb{E}[e_i^2|Z_i] = \gamma'\Sigma\gamma$ , 以及  $\Sigma_{2\Sigma} = \mathbb{E}[u_2 e|Z] = \Sigma_{21} - \Sigma_{22}\beta$ .

显然, 当  $\Gamma = 0$  时, 模型完全不可识别, 现在我们想知道  $\Gamma$  很小时会如何. 按照 Staiger and Stock (1997) 介绍的方法, 假定

$$\Gamma = n^{-\frac{1}{2}}C \quad (6.31)$$

这里的  $C$  是一个固定的矩阵.  $\|C\|$  越大表明识别越强, 反之则越弱, (6.31) 称为局部趋零 (local-to-zero) 假设.

对于 2SLS 估计量, 首先由 CLT 得到

$$n^{-\frac{1}{2}} \sum_{i=1}^n Z_i u_i' \xrightarrow{d} \xi = [\xi_1, \xi_2]$$

这里  $\text{vec}(\xi) \sim N(0, \mathbb{E}[u_i u_i' \otimes Z_i Z_i'])$ . 上式同时意味着

$$\frac{1}{\sqrt{n}} Z' e \xrightarrow{d} \xi_e = \xi \gamma$$

根据 Slutsky 定理可知

$$\frac{1}{\sqrt{n}} Z' X = \frac{1}{n} Z' Z C + \frac{1}{\sqrt{n}} Z' U_2 \xrightarrow{d} Q_{ZZ} C + \xi_2$$

以及

$$\begin{aligned} X' P_Z X &= \left( \frac{1}{\sqrt{n}} X' Z \right) \left( \frac{1}{n} Z' Z \right)^{-1} \left( \frac{1}{\sqrt{n}} Z' X \right) \\ &\xrightarrow{d} (Q_{ZZ} C + \xi_2)' Q_{ZZ}^{-1} (Q_{ZZ} C + \xi_2) \end{aligned}$$

还有

$$\mathbf{X}'\mathbf{P}_Z\mathbf{e} = \left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}\right) \xrightarrow{d} (\mathbf{Q}_{ZZ}\mathbf{C} + \xi_2)'\mathbf{Q}_{ZZ}^{-1}\xi_e$$

最后得到

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{e} \\ &\xrightarrow{d} [(\mathbf{Q}_{ZZ}\mathbf{C} + \xi_2)'\mathbf{Q}_{ZZ}^{-1}(\mathbf{Q}_{ZZ}\mathbf{C} + \xi_2)]^{-1}(\mathbf{Q}_{ZZ}\mathbf{C} + \xi_2)'\mathbf{Q}_{ZZ}^{-1}\xi_e\end{aligned}$$

由于固定矩阵  $\mathbf{C}$  的存在, 此时 2SLS 估计量  $\hat{\beta}_{2SLS}$  不是一致的, 并且渐近分布也是非正态的. 因此, 即使拥有相当大的样本量, 使用弱工具变量也难以获得含有较小偏误的估计量.

关于弱工具变量的检验, 参考 Staiger and Stock (1997), Stock and Yogo (2005), 以及 Kleibergen and Paap (2006). 目前关于弱工具变量的许多开放性问题, 例如过度识别模型在异方差下的稳健性推断程序尚未达成一致.

## 第 7 章 广义矩方法

目前在应用计量经济学领域内最流行的估计方法之一就是 Hansen (1982) 建立的广义矩方法 (Generalized Method of Moments, GMM), 它提供了一个统一的计量经济学分析框架, 许多估计量都可以看作是 GMM 框架下的一个特例, 包括之前的 OLS 估计量和 2SLS 估计量, 甚至 ML 估计量也是 GMM 估计量的特例.

### 7.1 矩方程与 GMM 估计量

假设  $\theta \in \Theta$  是  $P \times 1$  维向量, 这里  $\Theta$  为  $P \times 1$  维参数空间. 再定义  $W_i$  为一个  $d \times 1$  维随机向量, 它的支集为  $\mathcal{W}$ , 并且其分布的某一特征可以由  $\theta$  刻画, 并且存在一个  $L \times 1$  维矩函数  $g(W_i, \theta)$  和参数  $\theta_o \in \Theta$ , 使得矩方程

$$\mathbb{E}[g(W_i, \theta_o)] = 0, \quad i = 1, 2, \dots, n \quad (7.1)$$

成立, 下标  $o$  表示该参数是  $\Theta$  内的某个我们感兴趣的真值. 在 IV 模型中,  $g(W_i, \theta) = Z_i(Y_i - X_i'\theta)$ , 这里  $W_i = [Y_i, Z_i', X_i']'$ .

总体而言, 我们称  $\theta_o$  是可识别的, 如果存在一个从数据分布到  $\theta_o$  的唯一映射, 故而在矩方程 (7.1) 的背景下, 这意味着满足它的  $\theta_o$  是唯一的. 由于矩方程 (7.1) 是一个由  $L$  个方程组成的含  $P$  个未知量的系统, 沿用上一章的说法, 我们称  $L = P$  时模型恰好识别,  $L > P$  时模型过度识别,  $L < P$  时模型不可识别.

在恰好识别的情况下, 可以通过 (7.1) 直接解出  $\theta$ . 定义样本矩为

$$\hat{g}(\theta) = n^{-1} \sum_{i=1}^n g(W_i, \theta)$$

再定义矩方法 (Method of Moments, MM) 估计量为使得  $\hat{g}(\theta) = 0$  的那个参数, 例如在 IV 模型中, MM 估计量为

$$\hat{\theta}_{\text{MM}} = (Z'X)^{-1}Z'Y$$

然而在过度识别的情况下, 由于方程数大于未知数的个数, 所以  $\theta$  无解, 此时无法通过矩估计直接解出  $\theta$ . 但是从这个角度出发, 我们期望  $\hat{g}(\theta)$  尽可能小, 由此找到一个合适的估计量. 例如, 通过选择某个  $\hat{\theta}$  来最小化二次型

$$\hat{g}(\theta)' \hat{g}(\theta) = \|\hat{g}(\theta)\|^2$$

然而考虑到  $\hat{g}(\theta)$  内的元素可能存在相关性, 因此引入一个权重矩阵  $\hat{W}$ , 并选择  $\hat{\theta}$  以最小化二次型

$$\begin{aligned} J_n(\theta) &= \hat{g}(\theta)' \hat{W} \hat{g}(\theta) \\ &= \left[ \sum_{i=1}^n g(W_i, \theta) \right]' \hat{W} \left[ \sum_{i=1}^n g(W_i, \theta) \right] \end{aligned}$$

显然, 当  $\hat{W} = I_L$  时, 赋予这  $L$  个样本矩的权重是相同的.

### 定义 7.1

广义矩方法估计量为

$$\hat{\theta}_{\text{GMM}} = \arg \min_{\theta \in \Theta} J_n(\theta)$$

其中  $\hat{W}$  是正定的.



在恰好识别的情况下, GMM 估计量还原为不依赖于  $\hat{W}$  的矩估计量. 而在过度识别的情况下, 权重矩阵  $\hat{W}$  一般会影响到估计的精度, 一个性质良好的  $\hat{W}$  要对方差较大的样本矩赋予小的权重, 并且消除各样本矩之间的相关性. 如果权重矩阵  $\hat{W}$  是固定的, 则通常称  $\hat{\theta}_{\text{GMM}}$  为一步 GMM 估计量 (one-step GMM estimator).

通常而言, 如果矩函数  $g(W_i, \theta)$  是总体参数  $\theta$  的非线性函数, 那么通常无法获得  $\hat{\theta}$  的显式解, 但线性工具变量估计是一个例外. 考虑矩函数

$$g(W_i, \theta) = Z_i(Y_i - X_i'\theta)$$

以及对应的矩条件

$$\mathbb{E}[Z_i(Y_i - X_i'\theta)] = 0$$

其中  $Y_i$  是标量,  $X_i$  和  $\theta$  是  $K \times 1$  维向量,  $Z_i$  是  $L \times 1$  维向量, 并且  $L \geq K$ . 此时

$$\hat{g}(\theta) = n^{-1} \sum_{i=1}^n Z_i(Y_i - X_i'\theta) = \frac{Z'(Y - X\theta)}{n}$$

从而  $J_n(\theta) = n^{-2}(Y - X\theta)'Z\hat{W}Z'(Y - X\theta)$ , 于是可以找到 FOC

$$\frac{\partial}{\partial \theta} J_n(\theta) = -2X'Z\hat{W}Z'(Y - X\theta) = 0$$

如果  $Z'X$  满秩, 则可以找到 GMM 估计量

$$\hat{\theta}_{\text{GMM}} = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'Y$$

特别地, 如果模型恰好识别且  $X'Z$  满秩, 那么

$$\begin{aligned} \hat{\theta}_{\text{GMM}} &= (Z'X)^{-1}\hat{W}^{-1}(X'Z)^{-1}X'Z\hat{W}Z'Y \\ &= (Z'X)^{-1}Z'Y = \hat{\theta}_{\text{IV}} \end{aligned}$$

而在恰好识别的情况下, 如果将权重矩阵选为  $\hat{W} = c(Z'Z)^{-1}$ , 其中  $c \neq 0$ , 那么  $\hat{\theta}_{\text{GMM}} = \hat{\theta}_{\text{2SLS}}$ . 由此我们看到, 2SLS 估计量是线性 IV 估计量的一个特例, 并且后面将会证明, 在条件同方差的假设下, 2SLS 估计量在线性 IV 估计量类中是渐近有效的.

## 7.2 GMM 的渐近性质

### 7.2.1 一致性


为了研究 GMM 估计量的渐近性质, 我们首先要给出新的定义和一些正则条件.




**定义 7.2**

设  $\{g_n(\theta)\}$  为随机向量构成的非负序列, 如果当  $n \rightarrow \infty$  时有

$$\sup_{\theta \in \Theta} \|g_n(\theta) - g(\theta)\| = o_p(1)$$

则称  $g_n(\theta)$  在  $\theta \in \Theta$  中依概率一致收敛于  $g(\theta)$ . 

**假设 7.1**


- (1)  $\{g(W_i, \theta)\}_{i=1}^n$  是可观测的 i.i.d. 随机样本.
- (2)  $P \times 1$  维参数空间  $\Theta$  是紧集.
- (3) 对于任意给定的  $\theta \in \Theta$ , 矩函数  $g(\cdot, \theta)$  是 Borel 可测的, 且对于任意的  $w \in \mathcal{W}$ ,  $g(w, \cdot)$  在  $\Theta$  上为连续函数.
- (4) 样本矩  $\hat{g}(\theta)$  在  $\Theta$  上依概率一致收敛于  $g(\theta) = \mathbb{E}[g(W_i, \theta)]$ , 也即当  $n \rightarrow \infty$  时有
 
$$\sup_{\theta \in \Theta} \|\hat{g}(\theta) - g(\theta)\| \xrightarrow{p} 0$$
- (5)  $\theta$  在  $\Theta$  上可识别, 也即存在唯一的  $\theta_o \in \Theta$  使得  $\mathbb{E}[g(W_i, \theta_o)] = 0$ .
- (6)  $\mathbb{E}[\sup_{\theta \in \Theta} \|g(W_i, \theta)\|] < \infty$ , 又称占优条件.
- (7)  $\hat{W} \xrightarrow{p} W$ , 这里  $W$  是  $L \times L$  维非随机的对称、有限且非奇异的矩阵. 

集合的紧性是一个拓扑概念<sup>1</sup>, 它可以极大简化渐近分析. 根据 Heine-Borel 定理<sup>2</sup>, 我们可以将紧参数空间  $\Theta$  定义为  $\mathbb{R}^P$  空间里的某个非常大的有界闭集, 这样做并不会令人担心. 然而以上假设的第 (5) 点和第 (6) 点通常极难满足, 一般都直接假设其成立.

为了使得一致收敛的假设成立, 我们需要合适的一致弱大数定律 (Uniform Weak Law of Large Numbers, UWLLN).

**引理 7.1 (一致弱大数定律)**

设  $\{W_i\}_{i=1}^n$  是由定义在  $\mathcal{W} \subset \mathbb{R}^d$  上的 i.i.d. 随机向量构成的序列,  $\Theta$  是  $\mathbb{R}^P$  的紧子集,  $q: \mathcal{W} \times \Theta \rightarrow \mathbb{R}$  为一个实值函数. 如果以下条件成立: (1) 对于任意  $\theta \in \Theta$ ,  $q(\cdot, \theta)$  是 Borel 可测的; (2) 对于任意  $w \in \mathcal{W}$ ,  $q(w, \cdot)$  在  $\Theta$  上是连续的; (3) 对于一切  $\theta \in \Theta$ , 存在 Borel 可测函数  $D: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ , 使得对于一切  $w \in \mathcal{W}$  和  $\theta \in \Theta$  都有  $|q(w, \theta)| \leq |D(w)|$  及  $\mathbb{E}[D(W_i)] < \infty$ . 那么

- (1)  $Q(\theta) = \mathbb{E}[q(W_i, \theta)]$  在  $\Theta$  上连续.
- (2) 当  $n \rightarrow \infty$  时有  $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$ , 其中  $\hat{Q}(\theta) = n^{-1} \sum_{i=1}^n q(W_i, \theta)$ . 


**证明** 见 Jennrich (1969).

**注** 由于  $|g(w_i, \theta)| \leq \sup_{\theta \in \Theta} |g(w_i, \theta)|$ , 故而可以直接将  $D(W_i)$  替换为  $\sup_{\theta \in \Theta} |g(W_i, \theta)|$ .

<sup>1</sup> 设  $X$  为一个拓扑空间, 并且  $K \subset X$ , 如果  $K$  的每个开覆盖都有有限子覆盖, 那么  $K$  为紧集.

<sup>2</sup> 有限维  $\mathbb{R}^P$  空间的有界闭子集等价于紧集.

**引理 7.2 (极值估计量的一致性)**

设  $\Theta$  是  $\mathbb{R}^P$  的紧子集,  $\hat{Q} : \Theta \rightarrow \mathbb{R}$  是随机实值函数,  $Q : \Theta \rightarrow \mathbb{R}$  是非随机实值连续函数. 假定对于一切  $\theta \in \Theta$ ,  $\hat{Q}(\cdot)$  是 Borel 可测函数,  $\hat{Q}(\cdot)$  在  $\Theta$  上是连续的, 并且  $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$ . 令  $\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{Q}(\theta)$ , 如果  $Q(\theta)$  只在唯一的  $\theta_o \in \Theta$  处取得在  $\Theta$  上的最大值, 那么  $\hat{\theta} \xrightarrow{P} \theta_o$ . 

**证明** 设  $\mathcal{N} \subset \mathbb{R}^P$  是一个包含  $\theta_o$  的开集, 由于它的补集  $\mathcal{N}^c$  是闭的且  $\Theta$  是紧的, 故而  $\mathcal{N}^c \cap \Theta$  也是紧的. 又因为  $Q : \Theta \rightarrow \mathbb{R}$  是连续函数, 因此由最值定理可以证得  $\max_{\theta \in \mathcal{N}^c \cap \Theta} Q(\theta)$  的存在性. 令

$$\varepsilon = Q(\theta_o) - \max_{\theta \in \mathcal{N}^c \cap \Theta} Q(\theta) \quad (7.2)$$

由于  $\mathcal{N}$  可以选取得任意小, 所以  $\varepsilon > 0$  也可以任意小.

定义事件  $A_n$  为: 对于一切  $\theta \in \Theta$  都有  $|\hat{Q}(\theta) - Q(\theta)| < \varepsilon/2$ , 于是

$$A_n \Rightarrow Q(\hat{\theta}) > \hat{Q}(\hat{\theta}) - \varepsilon/2 \quad (7.3)$$

以及

$$A_n \Rightarrow \hat{Q}(\theta_o) > Q(\theta_o) - \varepsilon/2 \quad (7.4)$$

根据定义有  $\hat{Q}(\hat{\theta}) = \max_{\theta \in \Theta} \hat{Q}(\theta)$ , 故而  $\hat{Q}(\hat{\theta}) \geq \hat{Q}(\theta_o)$ , 从 (7.3) 可以推知

$$A_n \Rightarrow Q(\hat{\theta}) > \hat{Q}(\theta_o) - \varepsilon/2 \quad (7.5)$$

将 (7.4) 和 (7.5) 相加得到

$$A_n \Rightarrow Q(\hat{\theta}) > Q(\theta_o) - \varepsilon \quad (7.6)$$

从 (7.2) 和 (7.6) 可得


$$A_n \Rightarrow \max_{\theta \in \mathcal{N}^c \cap \Theta} Q(\theta) < Q(\hat{\theta})$$

也即  $A_n \Rightarrow \hat{\theta} \in \mathcal{N}$ , 这同时意味着  $\mathbb{P}[A_n] \leq \mathbb{P}[\hat{\theta} \in \mathcal{N}]$ , 根据  $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$  可知

$$\lim_{n \rightarrow \infty} \mathbb{P}[A_n] = 1$$

因此  $\lim_{n \rightarrow \infty} \mathbb{P}[\hat{\theta} \in \mathcal{N}] = 1$ , 也即  $\hat{\theta} = \theta_o + o_p(1)$ .

**定理 7.1**

在假设 7.1 下, 当  $n \rightarrow \infty$  时有  $\hat{\theta} \xrightarrow{P} \theta_o$ . 

**证明** 首先设

$$\hat{Q}(\theta) = -\hat{g}(\theta)' \hat{W} \hat{g}(\theta)$$

$$Q(\theta) = -g(\theta)' W g(\theta)$$

根据三角不等式可知

$$\begin{aligned}
 |\hat{Q}(\theta) - Q(\theta)| &= |\hat{g}(\theta)' \hat{W} \hat{g}(\theta) - g(\theta)' W g(\theta)| \\
 &= |[\hat{g}(\theta) - g(\theta) + g(\theta)]' \hat{W} [\hat{g}(\theta) - g(\theta) + g(\theta)] - g(\theta)' W g(\theta)| \\
 &\leq |[\hat{g}(\theta) - g(\theta)]' \hat{W} [\hat{g}(\theta) - g(\theta)]| \\
 &\quad + 2|g(\theta)' \hat{W} [\hat{g}(\theta) - g(\theta)]| + |g(\theta)' (\hat{W} - W) g(\theta)|
 \end{aligned}$$

假设 7.1 中的占优条件保证了 UWLLN 的条件 (3) 成立, 于是当  $n \rightarrow \infty$  时有

$$\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$$

再由 Cholesky 分解可知存在矩阵  $C$  使得  $W = C' C$ , 如果  $\theta \neq \theta_o$ , 那么由识别条件可知

$$0 \neq W g(\theta) = C' C g(\theta)$$

这同时意味着  $C g(\theta) \neq 0$ , 从而

$$Q(\theta) = -[C g(\theta)]' [C g(\theta)] < Q(\theta_o) = 0$$

于是  $\theta_o$  是  $\Theta$  上唯一使得  $Q(\theta)$  最大化的参数. 又因为  $Q(\theta)$  是连续的, 根据引理 7.2 即可推知  $\hat{\theta} \xrightarrow{p} \theta_o$ .

## 7.2.2 渐近正态性

和之前类似, 我们还需要在假设 7.1 上做出增加额外条件才能得出 GMM 估计量的渐近正态性.

### 假设 7.2

在假设 7.1 的基础上, 以下额外条件成立:

- (1)  $\theta_o \in \text{int}(\Theta)$ .
- (2) 对于一切  $w_i \in \mathcal{W} \subset \mathbb{R}^d$ ,  $g(w_i, \cdot)$  在包含  $\theta_o$  的一个邻域  $\mathcal{N}$  内连续可微.
- (3)  $\mathbb{E}[\sup_{\theta \in \mathcal{N}} \|\nabla_{\theta} g(W_i, \theta)\|] < \infty$ .
- (4)  $\mathbb{E}[\nabla_{\theta} g(W_i, \theta_o)]$  为  $L \times P$  维满秩矩阵.
- (5)  $n^{-\frac{1}{2}} \sum_{i=1}^n g(W_i, \theta_o) \xrightarrow{d} N(0, \Omega_o)$ .

### 定理 7.2

在假设 7.2 下, 当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N[0, (G_o' W G_o)^{-1} G_o' W \Omega_o W G_o (G_o' W G_o)^{-1}]$$

其中  $G_o = \mathbb{E}[\nabla_{\theta} g(W_i, \theta_o)]$ , 并且  $\Omega_o = \mathbb{E}[g(W_i, \theta_o)g(W_i, \theta_o)']$ .

**证明** 因为  $\theta_o \in \text{int}(\Theta)$ , 且当  $n \rightarrow \infty$  时有  $\hat{\theta} \xrightarrow{p} \theta_o$ , 故而此时  $\hat{\theta}$  在  $\Theta$  的内部概率接近于 1. 当  $n$  充分大时, 最大化  $\hat{Q}(\theta) = -\hat{g}(\theta)' \hat{W} \hat{g}(\theta)$  的 FOC 为

$$\left[ \sum_{i=1}^n \nabla_{\theta} g(W_i, \hat{\theta}) \right]' \hat{W} \left[ \sum_{i=1}^n g(W_i, \hat{\theta}) \right] = 0$$

等价地有

$$[\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} \sqrt{n} \hat{g}(\hat{\theta}) = 0$$

定义  $L \times P$  维矩阵

$$G(\theta) = \mathbb{E}[\nabla_{\theta} g(W_i, \theta)]$$

利用三角不等式可知

$$\begin{aligned} \|\nabla_{\theta} \hat{g}(\hat{\theta}) - G_o\| &= \|\nabla_{\theta} \hat{g}(\hat{\theta}) - G(\hat{\theta}) + G(\hat{\theta}) - G_o\| \\ &\leq \|\nabla_{\theta} \hat{g}(\hat{\theta}) - G(\hat{\theta})\| + \|G(\hat{\theta}) - G_o\| \\ &\leq \sup_{\theta \in \Theta} \|\nabla_{\theta} \hat{g}(\hat{\theta}) - G(\hat{\theta})\| + \|G(\hat{\theta}) - G_o\| \xrightarrow{p} 0 \end{aligned}$$

其中 UWLLN 和  $\hat{\theta} \xrightarrow{p} \theta_o$  分别保证了上式最后两行的两项趋于 0. 再根据假设 7.2(5) 可得

$$\sqrt{n} \hat{g}(\theta_o) = n^{-\frac{1}{2}} \sum_{i=1}^n g(W_i, \theta_o) \xrightarrow{d} N(0, \Omega_o) \quad (7.7)$$

其中  $\Omega_o = \mathbb{E}[g(W_i, \theta_o)g(W_i, \theta_o)']$ . 现在将样本矩函数  $\hat{g}(\hat{\theta})$  在  $\theta_o$  处进行一阶 Taylor 展开得

$$\sqrt{n} \hat{g}(\hat{\theta}) = \sqrt{n} \hat{g}(\theta_o) + [\nabla_{\theta} \hat{g}(\bar{\theta})] \sqrt{n}(\hat{\theta} - \theta_o) \quad (7.8)$$

其中  $\bar{\theta}$  在  $\hat{\theta}$  和  $\theta_o$  之间, 根据  $\hat{\theta} \xrightarrow{p} \theta_o$  可知  $\text{plim } \bar{\theta} = \theta_o$ , 于是同理可得

$$\nabla_{\theta} \hat{g}(\bar{\theta}) \xrightarrow{p} G_o$$

现在将条件 (7.8) 代入 FOC 可知

$$[\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} [\sqrt{n} \hat{g}(\theta_o) + [\nabla_{\theta} \hat{g}(\bar{\theta})] \sqrt{n}(\hat{\theta} - \theta_o)] = 0$$

也即

$$\sqrt{n}(\hat{\theta} - \theta_o) = -\{[\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} [\nabla_{\theta} \hat{g}(\bar{\theta})]\}^{-1} [\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} \sqrt{n} \hat{g}(\theta_o)$$

因此当  $n \rightarrow \infty$  时, 根据 (7.7) 和 Slutsky 定理即可证得结论.

一旦获得了 GMM 估计量  $\hat{\theta}$ , 我们就可以对它的渐近方差进行估计, 于是  $\Omega_o$  的一致估计量可以由

$$\hat{\Omega} = n^{-1} \sum_{i=1}^n g(W_i, \hat{\theta}) g(W_i, \hat{\theta})'$$

给出, 而  $G_o$  的一致估计量为

$$\hat{G} = n^{-1} \sum_{i=1}^n \nabla_{\theta} g(W_i, \hat{\theta})$$

于是  $\sqrt{n} \hat{\theta}$  的渐近协方差矩阵估计量为

$$(\hat{G}' \hat{W} \hat{G})^{-1} \hat{G}' \hat{W} \hat{\Omega} \hat{W} \hat{G} (\hat{G}' \hat{W} \hat{G})^{-1}$$

特别地, 如果权重矩阵  $W = \Omega_o^{-1}$ , 此时

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N[0, (G_o' \Omega_o^{-1} G_o)^{-1}]$$

下面我们证明, 这样的  $W$  是最优权重矩阵, 也即对于任意权重矩阵  $W$ , 以下渐近方差之差是半

正定的

$$(G_o'WG_o)^{-1}G_o'W\Omega_oWG_o(G_o'WG_o)^{-1} - (G_o'\Omega_o^{-1}G_o)^{-1} \quad (7.9)$$

### 定理 7.3

在假设 7.2 下, 通过选取  $W = \Omega_o^{-1}$  得到的 GMM 估计量是渐近有效的.



**证明** 令  $V = (G_o'WG_o)^{-1}G_o'W\Omega_oWG_o(G_o'WG_o)^{-1}$ , 以及  $V_o = (G_o'\Omega_o^{-1}G_o)^{-1}$ . 为了证明 (7.9) 成立, 只需证明  $V_o^{-1} - V^{-1}$  半正定即可. 令

$$D = I_L - \Omega_o^{-\frac{1}{2}}WG_o(G_o'W\Omega_oWG_o)^{-1}G_o'W\Omega_o^{\frac{1}{2}}$$

显然  $D$  是一个对称幂等矩阵. 进一步

$$\begin{aligned} V_o^{-1} - V^{-1} &= G_o'\Omega_o^{-1}G_o - G_o'WG_o(G_o'W\Omega_oWG_o)^{-1}G_o'WG_o \\ &= G_o'\Omega_o^{-\frac{1}{2}}D\Omega_o^{-\frac{1}{2}}G_o = (D\Omega_o^{-\frac{1}{2}}G_o)'(D\Omega_o^{-\frac{1}{2}}G_o) \end{aligned}$$

证毕.

**注** 最优权重矩阵  $W$  并不唯一, 对于任意  $c \neq 0$ ,  $c\Omega_o^{-1}$  也是最优的.

现在我们回到线性工具变量估计, 也即矩函数

$$g(W_i, \theta) = Z_i(Y_i - X_i'\theta)$$

如果假设 6.2 成立, 那么可以假设 7.2 也成立. 根据定理 7.2, 线性 GMM 估计量

$$\hat{\theta}_{\text{GMM}} = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'Y$$

在假设 6.2 下具有渐近正态性. 当权重矩阵  $\hat{W} = (Z'Z)^{-1}/n$  时, GMM 估计量变为 2SLS 估计量, 此时  $\hat{W} \xrightarrow{p} W = \mathbb{E}[Z_iZ_i']^{-1}$ , 以及  $\Omega_o = \mathbb{E}[Z_iZ_i'e_i^2]$ .

如果条件同方差假设  $\mathbb{E}[e_i^2|Z_i] = \sigma^2$  成立, 那么  $\Omega_o = \sigma^2\mathbb{E}[Z_iZ_i']$ , 我们可以不失一般性地令  $\sigma^2 = 1$ , 于是根据定理 7.3 可知, 此时 2SLS 估计量在所有线性工具变量估计量类中是渐近有效的. 然而, 如果误差项存在条件异方差, 那么 2SLS 不是渐近有效估计.

### 7.2.3 二阶段 GMM 估计量

定理 7.3 表明, 在假设 7.2 成立的情况下, 按照以下步骤获取的二阶段 GMM 估计量是渐近最优的. 与一步 GMM 估计量不同, 二阶段 GMM 估计量的权重矩阵  $W$  不是固定的.

**第一阶段:** 首先获得某个一致的初始 GMM 估计量  $\tilde{\theta}$ , 这里

$$\tilde{\theta} = \arg \max_{\theta \in \Theta} \hat{g}(\theta)' \tilde{W} \hat{g}(\theta)$$

并且权重矩阵  $\tilde{W}$  依概率收敛于某个有限、对称和正定的矩阵  $W$ , 为简便起见, 可以取  $\tilde{W} = I_L$ . 通常而言,  $\tilde{\theta}$  不是渐近最优的估计量, 但它仍是  $\theta_o$  的一致估计量.

然后, 我们可以构造  $\Omega_o = \text{avar}[\sqrt{n}\hat{g}(\theta_o)]$  的一致估计量

$$\tilde{\Omega} = n^{-1} \sum_{i=1}^n g(W_i, \tilde{\theta})g(W_i, \tilde{\theta})'$$

并选择权重矩阵  $\hat{W} = \tilde{\Omega}^{-1}$ .

第二阶段: 用权重矩阵  $\hat{W} = \tilde{\Omega}^{-1}$  可以得到一个新的 GMM 估计量

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{g}(\theta)' \hat{W} \hat{g}(\theta)$$

这里权重矩阵  $\hat{W}$  不涉及未知参数  $\theta$ , 它是一个随机权重矩阵. 这样得到的二阶段 GMM 估计量是渐近最优的, 这是因为  $\hat{W} \xrightarrow{P} W = \Omega_o^{-1}$ . 此时

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N[0, (G_o' \Omega_o^{-1} G_o)^{-1}]$$

值得注意的是, 尽管二阶段 GMM 估计量是渐近有效的, 但在实际应用中可能需要不断重复估计, 知道 GMM 参数估计值和最小目标函数值收敛为止, 才能消除 GMM 估计量对初始权重矩阵  $\tilde{W}$  的依赖.

最后来看参数检验. 按照之前的做法, 如果原假设  $H_0: R(\theta_o) = r$  成立, 我们可以构造稳健的 Wald 检验统计量

$$W = n[R(\hat{\theta}) - r]' [\hat{R} \hat{V}_{\hat{\theta}} \hat{R}']^{-1} [R(\hat{\theta}) - r]$$

在假设 7.2 成立的条件下, 当  $n \rightarrow \infty$  时有  $W \xrightarrow{d} \chi_J^2$ . 其中  $R: \Theta \rightarrow \mathbb{R}^J$  连续可微,  $J \times P$  维矩阵  $\hat{R} = \nabla_{\theta} R(\hat{\theta})$  的秩为  $J$  并且  $J \leq P$ , 协方差矩阵估计量  $\hat{V}_{\hat{\theta}}$  是基于渐近最优 GMM 估计量来选取的.

## 7.3 过度识别检验

现在我们要检验矩条件  $\mathbb{E}[g(W_i, \theta_o)] = 0$  是否成立, 一个基本思想是构建  $L \times L$  维样本矩

$$\hat{g}(\hat{\theta}) = n^{-1} \sum_{i=1}^n g(W_i, \hat{\theta})$$

并判断它是否显著不等于 0.

首先将  $\sqrt{n}\hat{g}(\hat{\theta})$  在  $\theta_o$  处的一阶 Taylor 展开得到

$$\sqrt{n}\hat{g}(\hat{\theta}) = \sqrt{n}\hat{g}(\theta_o) + \bar{G} \sqrt{n}(\hat{\theta} - \theta_o) \quad (7.10)$$

根据之前的结论又有

$$\sqrt{n}(\hat{\theta} - \theta_o) = -\{[\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} \bar{G}\}^{-1} [\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} \sqrt{n}\hat{g}(\theta_o)$$

将其代入到 (7.10) 中, 并且左乘矩阵  $\hat{W}^{\frac{1}{2}}$  得到

$$\begin{aligned} \hat{W}^{\frac{1}{2}} \sqrt{n}\hat{g}(\hat{\theta}) &= \hat{W}^{\frac{1}{2}} \sqrt{n}\hat{g}(\theta_o) + \hat{W}^{\frac{1}{2}} \bar{G} \sqrt{n}(\hat{\theta} - \theta_o) \\ &= \hat{\Pi} \hat{W}^{\frac{1}{2}} \sqrt{n}\hat{g}(\theta_o) \end{aligned}$$

这里的权重矩阵  $\hat{W} = \tilde{\Omega}^{-1}$  出自第一阶段的渐近最优 GMM 估计, 并且

$$\hat{\Pi} = I_L - \hat{W}^{\frac{1}{2}} \bar{G} \{[\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W} \bar{G}\}^{-1} [\nabla_{\theta} \hat{g}(\hat{\theta})]' \hat{W}^{\frac{1}{2}}$$

根据 (7.7), 由 Slutsky 定理可知

$$\hat{W}^{\frac{1}{2}} \sqrt{n}\hat{g}(\theta_o) \xrightarrow{d} B$$

这里  $\mathbf{B} \sim N(0, \mathbf{I}_L)$ . 当  $n \rightarrow \infty$  时有

$$\hat{\Pi} \xrightarrow{P} \mathbf{I}_L - \mathbf{W}^{\frac{1}{2}} \mathbf{G}_o (\mathbf{G}_o' \mathbf{W} \mathbf{G}_o)^{-1} \mathbf{G}_o' \mathbf{W}^{\frac{1}{2}} = \Pi$$

其中  $\mathbf{W} = \mathbf{\Omega}_o^{-1}$ . 容易验证,  $\Pi$  是一个  $L \times L$  维对称幂等矩阵, 并且  $\text{trace}(\Pi) = L - P$ .

在原假设  $\mathbb{H}_0: \mathbb{E}[g(W_i, \theta_o)] = 0$  成立的情况下, 根据引理 2.1 可知

$$\begin{aligned} n[\hat{g}(\hat{\theta})' \hat{\mathbf{W}} \hat{g}(\hat{\theta})] &= [\hat{\mathbf{W}}^{\frac{1}{2}} \sqrt{n} \hat{g}(\theta_o)]' \hat{\Pi}^2 [\hat{\mathbf{W}}^{\frac{1}{2}} \sqrt{n} \hat{g}(\theta_o)] \\ &= [\hat{\mathbf{W}}^{\frac{1}{2}} \sqrt{n} \hat{g}(\theta_o)]' \hat{\Pi} [\hat{\mathbf{W}}^{\frac{1}{2}} \sqrt{n} \hat{g}(\theta_o)] \\ &\xrightarrow{d} \mathbf{B}' \Pi \mathbf{B} \sim \chi_{L-P}^2 \end{aligned}$$

以上检验方法称为 Hansen 检验.

#### 定理 7.4

在假设 7.2 下, 如果模型是过度识别的且当  $n \rightarrow \infty$  时有  $\hat{\mathbf{W}} \xrightarrow{d} \mathbf{\Omega}_o^{-1}$ . 那么当原假设  $\mathbb{H}_0: \mathbb{E}[g(W_i, \theta_o)] = 0$  成立时, 当  $n \rightarrow \infty$  时有

$$J = n \cdot \hat{g}(\hat{\theta})' \hat{\mathbf{W}} \hat{g}(\hat{\theta}) \xrightarrow{d} \chi_{L-P}^2$$



Hansen 检验基于渐近最优 GMM 估计量, 如果选取的是其它 GMM 估计量, 则无法得到上述结果. 此时需要使用另一个权重矩阵  $\tilde{\mathbf{W}}$  来构造 Wald 检验统计量, 它的渐近分布为  $\chi_L^2$ , 由于  $\chi_{L-P}^2$  的临界值更小, 因此使用渐近最优 GMM 估计量的 Hansen 检验更容易拒绝原假设.

另一方面, Hansen 检验只能在模型过度识别的情况下进行, 如果模型是恰好识别的, 那么无论原假设  $\mathbb{H}_0: \mathbb{E}[g(W_i, \theta_o)] = 0$  是否成立, 样本矩  $\hat{g}(\hat{\theta})$  等于零向量, 从而导致  $J = 0$ .

特别地, 我们可以利用 Hansen 检验来推导定理 6.8. 设矩函数为

$$g(W_i, \theta) = Z_i(Y_i - X_i' \theta)$$

2SLS 残差为  $\hat{e}_i = Y_i - X_i' \hat{\theta}_{2SLS}$ , 选取权重矩阵

$$\hat{\mathbf{W}} = n s^{-2} (\mathbf{Z}' \mathbf{Z})^{-1}$$

这里  $s^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$ , 由此得到的 2SLS 估计量是渐近最优 GMM 估计量, 于是 Hansen 检验统计量为

$$J = n \cdot \hat{g}(\hat{\theta})' \hat{\mathbf{W}} \hat{g}(\hat{\theta}) = \frac{\hat{e}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \hat{e}}{\hat{e}' \hat{e} / n}$$

当条件同方差  $\mathbb{E}[e_i^2 | Z_i] = \sigma^2$  成立时,  $\mathbf{\Omega}_o = \sigma^2 \mathbb{E}[Z_i Z_i']$ , 并且

$$\hat{\mathbf{W}} \xrightarrow{P} \mathbf{\Omega}_o^{-1} = \sigma^{-2} \mathbb{E}[Z_i Z_i']^{-1}$$

因此从定理 7.4 即可推得定理 6.8, 也即 Hansen 检验统计量变为 Sargan 检验统计量.

## 7.4 系统工具变量

本节将处理 SUR 方程组的内生性问题, 这是线性 GMM 的一个应用. 考虑如下线性回归

$$Y_i = \bar{X}_i \beta_o + e_i \quad (7.11)$$



其中  $Y_i$  是  $m \times 1$  维向量,  $\bar{X}_i$  为  $m \times \bar{K}$  维矩阵,  $e_i$  为  $m \times 1$  维误差向量. 可以将其展开为  $j$  个方程构成的 SUR 方程组

$$Y_1 = X_1' \beta_{1o} + e_1$$

$$\vdots$$

$$Y_m = X_m' \beta_{mo} + e_m$$

对于每个  $j = 1, \dots, m$ ,  $X_j$  表示一个既包含内生变量又包含外生变量的  $K_j \times 1$  维向量. 并且对于每个  $j$ , 我们还有一个  $L_j \times 1$  维的工具变量集合  $Z_j$ , 观测值  $i$  上的工具变量矩阵可以记为

$$\bar{Z}_i = \begin{bmatrix} Z_{i1}' & 0 & \cdots & 0 \\ 0 & Z_{i2}' & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_{im}' \end{bmatrix}$$

它的维数为  $m \times \bar{L}$ , 其中  $\bar{L} = \sum_{j=1}^m L_j$ .

现在我们给出以下假设, 并假设相关的矩条件成立.

### 假设 7.3

$$(1) \mathbb{E}[\bar{Z}_i' e_i] = 0.$$

$$(2) \text{rank}(\mathbb{E}[\bar{Z}_i' \bar{X}_i]) = \bar{K}.$$

上述假设意味着对于每个  $j = 1, \dots, m$ , 总有  $\mathbb{E}[Z_j' e_j] = 0$  和  $\text{rank}(\mathbb{E}[Z_{ij} X_{ij}']) = K_j$ . 类似于第六章的讨论, 假设 7.3(2) 是识别的秩条件, 它的必要要求为  $\bar{L} \geq \bar{K}$ .

在假设 7.3 下,  $\beta_o$  是求解以下线性矩条件的唯一  $\bar{K} \times 1$  维向量

$$\mathbb{E}[\bar{Z}_i' (Y_i - \bar{X}_i \beta)] = 0 \quad (7.12)$$

换言之, 对于任意  $\beta \neq \beta_o$ , 总有  $\mathbb{E}[\bar{Z}_i' (Y_i - \bar{X}_i \beta)] \neq 0$ . 当  $\bar{L} = \bar{K}$  时, 模型是恰好识别的, 此时  $\beta$  的系统 IV 估计量为

$$\hat{\beta}_{IV} = \left( \sum_{i=1}^n \bar{Z}_i' \bar{X}_i \right)^{-1} \left( \sum_{i=1}^n \bar{Z}_i' Y_i \right)$$

当  $\bar{L} > \bar{K}$  时, 模型是过度识别的, 可以得到  $\beta$  的 GMM 估计量

$$\hat{\beta}_{GMM} = (\bar{X}' \bar{Z} \hat{W} \bar{Z}' \bar{X})^{-1} \bar{X}' \bar{Z} \hat{W} \bar{Z}' Y$$

其中  $\bar{X}$  和  $\bar{Z}$  分别为  $\bar{X}_i$  和  $\bar{Z}_i$  是在观测值  $i$  上堆叠起来的矩阵, 并且当  $n \rightarrow \infty$  时有  $\hat{W} \xrightarrow{p} W$ , 这里的  $W$  是一个  $\bar{L} \times \bar{L}$  维非随机的对称正定矩阵. 在此基础上还有

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_o) \xrightarrow{d} N(0, V_\beta)$$

其中

$$V_\beta = (Q_{XZ} W Q_{ZX})^{-1} Q_{XZ} W \Omega W Q_{ZX} (Q_{XZ} W Q_{ZX})^{-1}$$



以及

$$\begin{aligned} Q_{XZ} &= \mathbb{E}[\bar{X}_i' \bar{Z}_i] \\ Q_{ZX} &= \mathbb{E}[\bar{Z}_i' \bar{X}_i] \\ \Omega &= \mathbb{E}[\bar{Z}_i' e_i e_i' \bar{Z}_i] \end{aligned}$$

如果选取  $\hat{W} = (\bar{Z}'\bar{Z}/n)^{-1}$ , 那么此时 GMM 估计量为

$$\hat{\beta}_{\text{S2SLS}} = [\bar{X}'\bar{Z}(\bar{Z}'\bar{Z})^{-1}\bar{Z}'\bar{X}]^{-1}\bar{X}'\bar{Z}(\bar{Z}'\bar{Z})^{-1}\bar{Z}'Y$$

称为系统 2SLS 估计量. 由此可以得到残差  $\check{e}_i = Y_i - \bar{X}_i \hat{\beta}_{\text{S2SLS}}$ , 进而可以选取最优权重矩阵

$$\hat{W} = \left( n^{-1} \sum_{i=1}^n \bar{Z}_i' \check{e}_i \check{e}_i' \bar{Z}_i \right)^{-1} \quad (7.13)$$

根据 (7.13) 得到 GMM 估计量  $\hat{\beta}_{\text{GMM}}$  是渐近最优的, 它的协方差矩阵估计量为

$$\left[ \bar{X}'\bar{Z} \left( \sum_{i=1}^n \bar{Z}_i' \hat{e}_i \hat{e}_i' \bar{Z}_i \right)^{-1} \bar{Z}'\bar{X} \right]^{-1} \quad (7.14)$$

其中  $\hat{e}_i = Y_i - \bar{X}_i \hat{\beta}_{\text{GMM}}$ . 从渐近观点上看, 使用第一阶段残差  $\check{e}_i$  来替代  $\hat{e}_i$  不会产生差异, 这个矩阵对角元素的平方根最优 GMM 估计量的渐近标准误.

特别地, 如果  $\bar{Z}_i = \bar{X}_i$  且  $\hat{e}_i$  是 SOLS 残差, 那么表达式 (7.14) 变为 SOLS 的稳健协方差矩阵估计量; 而当  $\bar{Z}_i = \hat{\Sigma}^{-1} \bar{X}_i$ , 且  $\hat{e}_i$  是 SUR 中的 FGLS 残差时, 那么表达式 (7.14) 变为 FGLS 的稳健协方差矩阵估计量.

为了检验原假设  $\mathbb{H}_0: R\beta_o = r$ , 除了使用传统的 Wald 检验外, 还可以使用 GMM 距离检验. 假设  $\hat{W}$  是一致估计出  $W$  的最优加权矩阵,  $\hat{\beta}$  是利用  $W$  获得的无约束 GMM 估计量, 而  $\tilde{\beta}$  表示利用相同的  $\hat{W}$  获得的具有  $J$  个线性约束的 GMM 估计量, 可以证明在  $\mathbb{H}_0$  下, GMM 距离检验统计量

$$n^{-1} \left[ \left( \sum_{i=1}^n \bar{Z}_i' \tilde{e}_i \right)' \hat{W} \left( \sum_{i=1}^n \bar{Z}_i' \tilde{e}_i \right) - \left( \sum_{i=1}^n \bar{Z}_i' \hat{e}_i \right)' \hat{W} \left( \sum_{i=1}^n \bar{Z}_i' \hat{e}_i \right) \right] \xrightarrow{d} \chi_J^2$$

其中残差为  $\tilde{e}_i = Y_i - \bar{X}_i \tilde{\beta}$ , 以及  $\hat{e}_i = Y_i - \bar{X}_i \hat{\beta}$ . 而在另一个假设  $\mathbb{H}_0: \mathbb{E}[\bar{Z}_i' e_i] = 0$  下, 利用 Hansen 检验可以得到

$$n^{-1} \left( \sum_{i=1}^n \bar{Z}_i' \hat{e}_i \right)' \hat{W} \left( \sum_{i=1}^n \bar{Z}_i' \hat{e}_i \right) \xrightarrow{d} \chi_{L-K}^2$$

这里要求模型是过度识别的.

## 第 8 章 面板数据模型

之前我们讨论的内容都是基于截面数据, 本章讨论面板数据的计量模型. 面板数据既包括了截面维度, 也包含了时间维度, 可以在一定程度上缓解由不可观测的时变因素造成的内生性问题, 并且捕捉到更多信息以提高估计的精确度.

### 8.1 面板数据

面板数据 (panel data) 是在一段时期内追踪同一组个体的数据, 我们用下标  $i$  表示第  $i$  个体, 下标  $t$  表示个体位于第  $t$  期, 下标  $it$  可以唯一识别个体  $i$  在第  $t$  期的情况. 此外, 我们  $T$  表示个体  $i$  被观测到的期数, 而用  $S_i$  来表示观测到个体  $i$  的时期组成的集合.

如果在同一时期内, 所有个体数据都能观测到, 则称面板数据是平衡的 (balanced). 假设我们有  $N$  个截面, 以及  $T$  个时期, 那么平衡面板数据共计有  $n = NT$  个观测值. 反之, 如果存在个体在某个时期观测不到的情况, 则称面板数据是非平衡的 (unbalanced), 共计有  $n = \sum_{i=1}^N T$  个观测值. 如果数据是随机缺失的, 则非平衡面板不会对应用产生影响, 只会让推导的符号变复杂, 以下仅讨论平衡面板数据.

除了这种上述方式外, 还可以按短面板和长面板进行分类, 其中短面板数据中的个体维度  $N$  比时间维度  $T$  更大, 而长面板数据的个体维度  $N$  比时间维度  $T$  更小. 一般而言, 微观层面的面板数据都是短面板.

现在来看具体的符号表示. 我们用序对  $(Y_{it}, X_{it})$  表示观测, 其中  $Y_{it}$  是因变量,  $X_{it}$  是  $K \times 1$  维自变量. 此外, 用  $Y_i$  表示由观测值  $Y_{it}$  堆叠起来的  $T \times 1$  维向量,  $X_i$  表示由观测值  $X_{it}'$  堆叠起来的  $T \times K$  维矩阵. 最后, 用记号  $Y = [Y_1', \dots, Y_N']'$  表示  $Y_i$  堆叠起来的  $n \times 1$  维向量,  $X = [X_1', \dots, X_N']'$  的含义相似.

### 8.2 混合回归

最简单的面板数据模型为混合回归 (pooled regression). 考虑以下模型

$$\begin{aligned} Y_{it} &= X_{it}'\beta + e_{it} \\ \mathbb{E}[X_{it}e_{it}] &= 0 \end{aligned} \tag{8.1}$$

其中  $\beta$  是  $K \times 1$  维参数向量,  $e_{it}$  为随机扰动项. 这个模型也可以写作

$$\begin{aligned} Y_i &= X_i\beta + e_i \\ \mathbb{E}[X_i'e_i] &= 0 \end{aligned}$$

这里的  $e_i$  是  $T \times 1$  维的. 最后, 全样本回归可以记作  $Y = X\beta + e$ .

在以上模型, 每一个体的误差项都和解释变量不相关, 这是相当严苛的条件. 如果解释变量包含了因变量  $Y_{it}$  的滞后项, 则 (8.1) 一定不成立. 在混合回归模型中, 估计  $\beta$  的标准方法是

最小二乘, 可以写作

$$\begin{aligned}\hat{\beta}_{\text{POOL}} &= \left( \sum_{i=1}^N \sum_{t=1}^T X_{it} X'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T X_{it} Y_{it} \right) \\ &= \left( \sum_{i=1}^N X'_i X_i \right)^{-1} \left( \sum_{i=1}^N X'_i Y_i \right) \\ &= (X'X)^{-1} X'Y\end{aligned}$$

称  $\hat{\beta}_{\text{POOL}}$  为混合 OLS 估计量, 第  $i$  个个体的残差向量为  $\hat{e}_i = Y_i - X_i \hat{\beta}_{\text{POOL}}$ .

如果比投影条件 (8.1) 更强的均值独立

$$\mathbb{E}[e_{it}|X_i] = 0 \quad (8.2)$$

成立, 那么由

$$\hat{\beta}_{\text{POOL}} = \beta + \left( \sum_{i=1}^N X'_i X_i \right)^{-1} \left( \sum_{i=1}^N X'_i e_i \right)$$

可知  $\mathbb{E}[\hat{\beta}_{\text{POOL}}|X] = \beta$ , 也即  $\hat{\beta}_{\text{POOL}}$  是  $\beta$  的无偏估计量.

如果  $\{e_{it}\}$  满足条件同方差和序列无关, 那么可以  $\hat{\beta}_{\text{POOL}}$  的协方差矩阵估计量具有经典形式. 如果  $\{e_{it}\}$  存在异方差而有序列相关, 那么可以使用异方差稳健的协方差矩阵估计量. 但总体而言, 不同个体的误差项可以假设为不相关, 而同一个体在不同时期的误差项往往相关, 此时应该使用聚类稳健的协方差矩阵估计量

$$\hat{V}_{\text{POOL}} = a_n (X'X)^{-1} \left( \sum_{i=1}^N X'_i \hat{e}_i \hat{e}'_i X_i \right) (X'X)^{-1}$$

其中

$$a_n = \left( \frac{n-1}{n-K} \right) \left( \frac{N}{N-1} \right)$$

为有限样本下的自由度调整.

## 8.3 随机效应模型

### 8.3.1 RE 估计量

显然, 对面板数据进行混合回归要求过于严苛, 并且没有充分利用面板数据的优势. 现在考察面板数据的综合误差项  $e_{it}$  的误差成分结构 (error-components structure), 它的最简形式为

$$e_{it} = u_i + \varepsilon_{it} \quad (8.3)$$

其中  $u_i$  称为个体效应 (individual fixed effects),  $\varepsilon_{it}$  是 i.i.d. 的特质误差项 (idiosyncratic error). 根据 (8.3), 我们可以建立单向误差成分模型 (one-way error components model)

$$Y_{it} = X'_{it} \beta + u_i + \varepsilon_{it} \quad (8.4)$$

上式的堆叠形式为

$$Y_i = X_i \beta + \mathbf{1}_i u_i + \varepsilon_i$$

对它的分析依赖于误差项  $u_i$  和  $\varepsilon_{it}$  的结构.

我们首先给出以下随机效应 (random effects) 假设, 即假定误差结构 (8.3) 具有零条件均值, 并且  $u_i$  和  $\varepsilon_{it}$  都满足条件同方差.

### 假设 8.1

对于误差结构 (8.3), 以下条件成立:

$$\mathbb{E}[\varepsilon_{it} | X_i, u_i] = 0 \quad (8.5)$$

$$\mathbb{E}[\varepsilon_{it}^2 | X_i] = \sigma_\varepsilon^2 \quad (8.6)$$

$$\mathbb{E}[\varepsilon_{it} \varepsilon_{js} | X_i] = 0 \quad (8.7)$$

$$\mathbb{E}[u_i | X_i] = 0 \quad (8.8)$$

$$\mathbb{E}[u_i^2 | X_i] = \sigma_u^2 \quad (8.9)$$

$$\text{rank}(\mathbb{E}[X_i' \boldsymbol{\Omega}^{-1} X_i]) = K \quad (8.10)$$

根据假设以上假设, 通过混合 OLS 方法得到的 POOL 估计量是一致的, 但由于误差项不是球型扰动项, 因而混合回归不是最有效率的方法.

此外, 我们还可以从这些假设中得出误差向量  $e_i$  满足  $\mathbb{E}[e_i | X_i] = 0$ , 以及

$$\begin{aligned} \mathbb{E}[e_i e_i' | X_i] &= \boldsymbol{\Omega} = \sigma_u^2 \mathbf{1}_i \mathbf{1}_i' + \sigma_\varepsilon^2 \mathbf{I}_i \\ &= \begin{bmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\varepsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\varepsilon^2 \end{bmatrix} \end{aligned} \quad (8.11)$$

其中  $\mathbf{1}_i$  是元素全为 1 的  $T \times 1$  维向量,  $\mathbf{I}_i$  为  $T \times T$  维单位矩阵. 对于个体  $i$ , 它的误差项在第  $t$  和第  $s$  期的相关系数为  $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$ .

在假设 8.1 下, 称单向误差成分模型 (8.4) 为随机效应模型. 由于给定了误差项的成分, 如果已知  $\sigma_u^2$  和  $\sigma_\varepsilon^2$ , 那么估计回归系数  $\beta$  的方法为广义最小二乘, 也即

$$\hat{\beta}_{\text{GLS}} = \left( \sum_{i=1}^N X_i' \boldsymbol{\Omega}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \boldsymbol{\Omega}^{-1} Y_i \right)$$

此时

$$\hat{\beta}_{\text{GLS}} - \beta = \left( \sum_{i=1}^N X_i' \boldsymbol{\Omega}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \boldsymbol{\Omega}^{-1} e_i \right)$$

根据假设 8.1 可知  $\mathbb{E}[\hat{\beta}_{\text{GLS}} | \mathbf{X}] = \beta$ , 也即  $\hat{\beta}_{\text{GLS}}$  是  $\beta$  的无偏估计量. 此外还能得到  $\hat{\beta}_{\text{GLS}}$  的条件方

差为

$$V_{\text{GLS}} = \left( \sum_{i=1}^N X_i' \Omega^{-1} X_i \right)^{-1} \quad (8.12)$$

现在来比较估计量  $\hat{\beta}_{\text{GLS}}$  和  $\hat{\beta}_{\text{POOL}}$ , 在假设 8.1 下可知  $\hat{\beta}_{\text{POOL}}$  的条件方差为

$$V_{\text{POOL}} = \left( \sum_{i=1}^N X_i' X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \Omega X_i \right) \left( \sum_{i=1}^N X_i' X_i \right)^{-1}$$

根据 Gauss-Markov 定理可知  $V_{\text{GLS}} \leq V_{\text{POOL}}$ , 因此在假设 8.1 下,  $\hat{\beta}_{\text{GLS}}$  比  $\hat{\beta}_{\text{POOL}}$  更有效. 此外, 当个体效应不存在时有  $\sigma_u^2 = 0$ , 于是  $\Omega = \sigma_\varepsilon^2 I_i$ , 随机效应模型变为混合回归模型, 故而  $V_{\text{GLS}} = V_{\text{POOL}} = \sigma_\varepsilon^2 (X'X)^{-1}$ .

以上的 GLS 分析建立在已知  $\sigma_u^2$  和  $\sigma_\varepsilon^2$  的基础上, 然而在实际情况中, 二者通常都是未知的, 此时仍需使用 FGLS 估计. 假定我们有  $\sigma_u^2$  和  $\sigma_\varepsilon^2$  的一致估计量, 那么可以构建

$$\hat{\Omega} = \hat{\sigma}_u^2 \mathbf{1}_i \mathbf{1}_i' + \hat{\sigma}_\varepsilon^2 I_i$$

于是可行的随机效应估计量为

$$\hat{\beta}_{\text{RE}} = \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} Y_i \right)$$

在假设 8.1 下, 由 FGLS 方法推导而来的随机效应估计量  $\hat{\beta}_{\text{RE}}$  也是一致估计量.

现在的问题是如何获取一致估计量  $\hat{\sigma}_u^2$  和  $\hat{\sigma}_\varepsilon^2$ . 首先定义  $\mathbb{E}[e_{it}^2] = \sigma_\varepsilon^2$ , 根据误差结构 (8.3) 和假设 8.1 可知  $\sigma_e^2 = \sigma_u^2 + \sigma_\varepsilon^2$ , 于是我们只需得到  $\hat{\sigma}_e^2$  和  $\hat{\sigma}_\varepsilon^2$  即可. 记  $\check{e}_{it}$  是出自混合 OLS 的残差, 容易得到

$$\hat{\sigma}_e^2 = \frac{1}{NT - K} \sum_{i=1}^N \sum_{t=1}^T \check{e}_{it}^2 \quad (8.13)$$

另一方面, 我们还需要得出  $\hat{\sigma}_u^2$ . 注意到当  $t \neq s$  时有  $\mathbb{E}[e_{it}e_{is}] = 0$ , 故而

$$\mathbb{E} \left[ \sum_{t=1}^{T-1} \sum_{s=t+1}^T e_{it}e_{is} \right] = \sum_{t=1}^{T-1} \sum_{s=t+1}^T \mathbb{E}[e_{it}e_{is}] = \frac{T(T-1)}{2} \sigma_u^2$$

从而

$$\hat{\sigma}_u^2 = \frac{1}{NT(T-1)/2 - K} \sum_{i=1}^N \sum_{t=1}^{T-1} \sum_{s=t+1}^T \check{e}_{it}\check{e}_{is} \quad (8.14)$$

根据 (8.13) 和 (8.14) 即可得到我们需要的估计量  $\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_e^2 - \hat{\sigma}_u^2$ . 如果面板数据是平衡的, 那么 (8.14) 为

$$\hat{\sigma}_u^2 = \frac{1}{NT(T-1)/2 - K} \sum_{i=1}^N \sum_{t=1}^{T-1} \sum_{s=t+1}^T \check{e}_{it}\check{e}_{is}$$

在假设 (8.1) 下, (8.13) 和 (8.14) 是一致估计量. 除了上述方法外, 还可以通过固定效应来估计特质误差方差  $\sigma_\varepsilon^2$ , 以及通过组间估计来获取  $\sigma_u^2$  的估计量.

最后来看  $\hat{\beta}_{\text{RE}}$  的协方差矩阵估计量. 如果假设 8.1 确实完全成立, 可以直接由 (8.12) 构造出

$$\hat{V}_{\text{RE}}^0 = \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i \right)^{-1}$$

然而假设 8.1 毕竟是很强的条件, 对于  $T$  固定且  $N$  相当大的面板数据而言, 使用以下聚类稳健的估计量通常也不会损失什么

$$\hat{V}_{\text{RE}} = \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} \hat{e}_i \hat{e}_i' \hat{\Omega}^{-1} X_i \right) \left( \sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i \right)^{-1}$$

其中  $\hat{e}_i = Y_i - X_i \hat{\beta}_{\text{RE}}$ .

如果标准的随机效应假设 8.1 成立, 但实际上模型中并没有包括不可观测的个体效应  $u_i$ , 那么混合 OLS 是有效的, 并且所有与混合 OLS 相关的检验统计量也是渐近有效的. 基于此, 如果不可观测效应  $u_i$  不存在, 那么我们可以检验原假设

$$\mathbb{H}_0 : \sigma_u^2 = 0$$

基于此, Wooldridge (2010) 介绍了一种统计量

$$Z = \frac{\sum_{i=1}^N \sum_{t=1}^{T-1} \sum_{s=t+1}^T \hat{e}_{it} \hat{e}_{is}}{\left[ \sum_{i=1}^N \left( \sum_{t=1}^{T-1} \sum_{s=t+1}^T \hat{e}_{it} \hat{e}_{is} \right)^2 \right]^{\frac{1}{2}}}$$

这里的  $\hat{e}_{it}$  为混合 OLS 残差. 在  $\{e_{it}\}$  不存在序列相关的情况下, 这个统计量服从渐近标准正态分布. 然而, 对原假设的拒绝并不意味着随机效应结构是正确的.

### 8.3.2 一般 FGLS 分析

实际上, 随机效应估计是一般 FGLS 分析的一个特例. 如果特质误差  $\{\varepsilon_{it}\}$  存在异方差性且对于不同的  $t$  是序列相关的, 那么在 FGLS 中应该使用更一般的估计量

$$\hat{\Omega} = N^{-1} \sum_{i=1}^N \check{e}_i \check{e}_i'$$

这里的  $\check{e}_i$  是混合 OLS 残差. 只要条件 (8.5) 和 (8.8) 成立, 那么 FGLS 估计量是一致的.

对于  $N$  很大的面板数据而言, 如果假设 8.1 完全成立, 那么一般 FGLS 估计量和 RE 估计量一样渐近有效. 而如果  $\mathbb{E}[e_i e_i' | X_i] = \Omega$ , 并且  $\Omega$  不具有随机效应结构 (8.11), 那么一般 FGLS 估计量更为渐近有效. 那为何还要使用 RE 估计量?

原因在于, 如果  $N$  比  $T$  大不了多少倍, 那么无约束的 FGLS 分析的有限样本性质欠佳, 因为  $\hat{\Omega}$  共计有  $T(T+1)/2$  个待估参数, 而 RE 估计量只需要两个待估参数.

既然一般 FGLS 分析的待估参数太多, 而传统的 RE 分析需要的假设条件过强, 故而可以考虑二者的中间形式. 例如, 特质误差  $\varepsilon_{it}$  服从自相关系数为  $\rho$  且方差为  $\sigma_\varepsilon^2$  的平稳 AR(1) 过程,

那么

$$\Omega = \sigma_u^2 \mathbf{1}_i \mathbf{1}_i' + \mathbb{E}[\varepsilon_i \varepsilon_i']$$

只依赖于参数  $\sigma_\varepsilon^2$ ,  $\sigma_u^2$  以及  $\rho$ , 这些参数通过混合回归就可以得到估计, 于是就可以很容易地使用 FGLS 方法.

## 8.4 固定效应模型

继续考虑具有单向误差结构的回归模型

$$Y_{it} = X_{it}'\beta + u_i + \varepsilon_{it} \quad (8.15)$$

或者

$$Y_i = X_i\beta + \mathbf{1}_i u_i + \varepsilon_i \quad (8.16)$$

在许多情况下, 我们可以将个体效应  $u_i$  解释为不可观测的非时变遗漏变量 (例如在工资方程中,  $u_i$  可能包含不可观测的个体  $i$  的能力), 此时称 (8.15) 和 (8.16) 为固定效应模型 (fixed effects model).

由于  $u_i$  可以被解释为遗漏变量, 自然可以将其视作与解释变量  $X_{it}$  相关, 此时混合 OLS 估计量和 FGLS 估计量均是不一致的.

### 8.4.1 组内估计

#### 假设 8.2

$$\mathbb{E}[\varepsilon_{it}|X_i, u_i] = 0.$$

这个假设称为严格外生性 (strictly exogeneity), 由它可以推出特质误差  $\varepsilon_{it}$  与回归元  $X_{it}$  及个体固定效应  $u_i$  均不相关, 但是反之不行. 由于允许  $\mathbb{E}[u_i|X_i]$  为  $X_i$  的任意函数, 因此 FE 分析比 RE 分析更加稳健.

在假设 8.2 下, 估计  $\beta$  的方法是使用组内变换 (within transformation), 以消去不可观测的个体固定效应  $u_i$ . 首先定义

$$\bar{Y}_i = T^{-1} \sum_{t=1}^T Y_{it} \quad (8.17)$$

它是在给定个体  $i$  的情况下, 对  $Y_{it}$  在时间上取平均. 再定义

$$\dot{Y}_{it} = Y_{it} - \bar{Y}_i$$

称为组内变换, 也即离差形式的  $Y_{it}$ . 继续考虑堆叠形式, 可以将 (8.17) 写为  $\bar{Y}_i = (\mathbf{1}_i \mathbf{1}_i')^{-1} \mathbf{1}_i' Y_i$ ,



从而

$$\begin{aligned}\dot{Y}_i &= Y_i - \mathbf{1}_i \bar{Y}_i \\ &= Y_i - \mathbf{1}_i (\mathbf{1}_i \mathbf{1}_i')^{-1} \mathbf{1}_i' Y_i \\ &= \mathbf{M}_i Y_i\end{aligned}$$

其中  $\mathbf{M}_i = \mathbf{I}_i - \mathbf{1}_i (\mathbf{1}_i \mathbf{1}_i')^{-1} \mathbf{1}_i'$ , 它是一个  $T \times T$  维对称幂等矩阵. 类似可以定义

$$\begin{aligned}\bar{X}_i &= T^{-1} \sum_{t=1}^T X_{it} \\ \dot{X}_{it} &= X_{it} - \bar{X}_i \\ \dot{X}_i &= \mathbf{M}_i X_i\end{aligned}$$

最后可以得到全样本的堆叠表示, 定义  $NT \times NT$  维矩阵  $\mathbf{D} = \text{diag}\{\mathbf{I}_1, \dots, \mathbf{I}_N\}$  以及  $\mathbf{M}_D = \text{diag}\{\mathbf{M}_1, \dots, \mathbf{M}_N\}$ , 于是

$$\mathbf{M}_D \mathbf{Y} = \dot{\mathbf{Y}} = \begin{bmatrix} \dot{Y}_1 \\ \vdots \\ \dot{Y}_N \end{bmatrix}, \quad \mathbf{M}_D \mathbf{X} = \dot{\mathbf{X}} = \begin{bmatrix} \dot{X}_1 \\ \vdots \\ \dot{X}_N \end{bmatrix}$$

现在对方程 (8.15) 在  $t = 1, \dots, T$  上取平均, 由此得到截面方程

$$\bar{Y}_i = \bar{X}_i' \beta + u_i + \bar{\varepsilon}_i \quad (8.18)$$

其中  $\bar{\varepsilon}_i = T^{-1} \sum_{t=1}^T \varepsilon_{it}$ . 使用 (8.15) 减去 (8.18) 可得

$$\dot{Y}_{it} = \dot{X}_{it}' \beta + \dot{\varepsilon}_{it}$$

其中  $\dot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$ . 或者将方程写为

$$\dot{Y}_i = \dot{X}_i \beta + \dot{\varepsilon}_i$$

利用最小二乘即可得到固定效应估计量

$$\begin{aligned}\hat{\beta}_{\text{FE}} &= \left( \sum_{i=1}^N \sum_{t=1}^T \dot{X}_{it} \dot{X}_{it}' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \dot{X}_{it} \dot{Y}_{it} \right) \\ &= \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \right)^{-1} \left( \sum_{i=1}^N \dot{X}_i' \dot{Y}_i \right) \\ &= \left( \sum_{i=1}^N X_i' \mathbf{M}_i X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \mathbf{M}_i Y_i \right)\end{aligned}$$

以及固定效应残差  $\hat{\varepsilon}_{it} = \dot{Y}_{it} - \dot{X}_{it}' \hat{\beta}_{\text{FE}}$ . 由组内变换得到的 FE 估计量又称组内估计量 (within estimator).

**注** 在进行组内变换的时候, 不可观测的个体效应  $u_i$  都被消去, 其本质是消去非时变因素, 因此  $X_{it}$  中不能包含非时变的回归元, 例如截距项、人的性别等变量. 因此在组内估计中,  $X_{it}$  的维数  $K$  是时变因素的数量.

然而我们仍然可以对截距项进行估计, 在模型  $Y_{it} = \alpha + X_{it}' \beta + u_i + \varepsilon_{it}$  中, 估计量  $\hat{\alpha}_{\text{FE}} =$



$\bar{Y} - \bar{X}'\hat{\beta}_{FE}$ , 其中  $\bar{Y}$  和  $\bar{X}$  为对应的全样本均值. 此外, 个体效应估计量可以写为  $\hat{u}_i = \bar{Y}_i - \bar{X}_i'\hat{\beta}_{FE} - \hat{\alpha}_{FE}$ .

利用假设8.2可以得到 FE 估计量的有限样本性质, 首先注意到

$$\hat{\beta}_{FE} - \beta = \left( \sum_{i=1}^N X_i' M_i X_i \right)^{-1} \left( \sum_{i=1}^N X_i' M_i \varepsilon_i \right)$$

立即可得  $\mathbb{E}[\hat{\beta}_{FE} | \mathbf{X}] = \beta$ , 以及<sup>1</sup>

$$\mathbf{V}_{FE} = \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \right)^{-1} \left( \sum_{i=1}^N \dot{X}_i' \Sigma_i \dot{X}_i \right) \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \right)^{-1} \quad (8.19)$$

其中  $\Sigma = \mathbb{E}[\varepsilon_i \varepsilon_i' | X_i]$ . 它的完全稳健估计量为

$$\hat{\mathbf{V}}_{FE} = (\dot{X}' \dot{X})^{-1} \left( \sum_{i=1}^N \dot{X}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i' \dot{X}_i \right) (\dot{X}' \dot{X})^{-1} \quad (8.20)$$

如果特质误差  $\{\varepsilon_{it}\}$  满足条件同方差和序列无关

$$\mathbb{E}[\varepsilon_{it}^2 | X_i] = \sigma_\varepsilon^2 \quad (8.21)$$

$$\mathbb{E}[\varepsilon_{it} \varepsilon_{is} | X_i] = 0 \quad (8.22)$$

这里  $t \neq s$ . 此时 (8.19) 简化为

$$\mathbf{V}_{FE}^0 = \sigma_\varepsilon^2 \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \right)^{-1}$$

容易证明, 如果 (8.21) 和 (8.22) 对混合回归成立, 那么

$$\mathbf{V}_{FE}^0 = \sigma_\varepsilon^2 \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \right) \geq \sigma_\varepsilon^2 \left( \sum_{i=1}^N X_i' X_i \right) = \mathbf{V}_{POOL}$$

表明使用更稳健的 FE 估计的代价是降低了估计量的有效性, 因为在 (8.21) 和 (8.22) 下, 混合 OLS 会利用更多信息. 最后,  $\mathbf{V}_{FE}^0$  的估计量可以表述为

$$\hat{\mathbf{V}}_{FE}^0 = \hat{\sigma}_\varepsilon^2 (\dot{X}' \dot{X})^{-1} \quad (8.23)$$

其中

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1) - K} \sum_{i=1}^N \hat{\varepsilon}_i' \hat{\varepsilon}_i \quad (8.24)$$

这里的  $\hat{\varepsilon}_i = \dot{Y}_i - \dot{X}_i \hat{\beta}_{FE}$  为固定效应残差.

## 8.4.2 虚拟变量回归

另一种得到 FE 估计量的方法是使用  $Y_{it}$  对  $X_{it}$  以及许多虚拟变量进行最小二乘回归. 首先来看没有回归元的方程

$$Y_{it} = u_i + \varepsilon_{it} \quad (8.25)$$

<sup>1</sup> 夹心估计量的中间部分实际为  $\sum_{i=1}^N \dot{X}_i' \mathbb{E}[\hat{\varepsilon}_i \hat{\varepsilon}_i' | X_i] \dot{X}_i$ , 但它在代数上等同于  $\sum_{i=1}^N \dot{X}_i' \mathbb{E}[\varepsilon_i \varepsilon_i' | X_i] \dot{X}_i$ .

考虑对固定效应向量  $u = [u_1, \dots, u_N]'$  的 OLS 估计量  $\hat{u}$ , 由于数据矩阵是  $\mathbf{1}_T$ , 易知  $\hat{u}_i = \bar{Y}_i$ , 以及 OLS 残差  $\hat{\varepsilon}_{it} = Y_{it} - \bar{Y}_i = \dot{Y}_{it}$ .

再令  $d_i$  为  $N$  个虚拟变量构成的向量, 它的第  $i$  个元素为 1 而其它元素为 0. 注意到  $u_i = d_i' u$ , 于是 (8.25) 变为  $Y_{it} = d_i' u + \varepsilon_{it}$ , 可以将其写为堆叠形式  $Y_i = \mathbf{1}_i d_i' u + \varepsilon_i$ , 以及  $Y = \mathbf{D}u + \varepsilon$ , 其中  $\mathbf{D} = \text{diag}\{\mathbf{1}_T, \dots, \mathbf{1}_T\}$ . 由此得到 OLS 估计量

$$\hat{u} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'Y = \text{vec}(\bar{Y}_i)$$

于是可得  $NT \times 1$  维残差向量

$$\hat{\varepsilon} = [\mathbf{I}_{NT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}']Y = \dot{Y}$$

类似地,  $\mathbf{X}$  对  $u$  回归得到的残差为  $\dot{\mathbf{X}}$ .

最后来看带回归元的单向误差成分模型

$$Y_{it} = X_{it}'\beta + d_i' u + \varepsilon_{it}$$

它的矩阵形式为

$$Y = \mathbf{X}\beta + \mathbf{D}u + \varepsilon$$

现在考虑  $(\beta, u)$  的最小二乘估计  $(\hat{\beta}, \hat{u})$ , 称为固定效应模型的最小二乘虚拟变量 (Least Squares Dummy Variable, LSDV) 估计量. 根据 FWL 定理, LSDV 估计量  $\hat{\beta}_{\text{LSDV}}$  可以通过残差  $\dot{Y}$  对  $\dot{\mathbf{X}}$  回归得到, 也即  $\hat{\beta}_{\text{LSDV}} = \hat{\beta}_{\text{FE}}$ .

当  $N$  相当大时, 最好使用组内估计而非 LSDV 方法, 这是因为后者的计算需要消耗大量计算力, 例如在  $T = 10$  和  $N = 10000$  的情况下, 矩阵  $\mathbf{D}$  有 10 亿个元素.

### 8.4.3 一阶差分估计

除了组内变换和虚拟变量回归, 另一种重要的方法为一阶差分 (first differencing). 具体而言, 一阶差分变换为

$$\Delta Y_{it} = Y_{it} - Y_{it-1}$$

也即用第  $t$  期的观测  $Y_{it}$  减去第  $t-1$  期的观测  $Y_{it-1}$ , 这对于  $t = 1$  之外的观测都适用. 用矩阵可以表示为  $\Delta Y_i = \mathbf{D}_i Y_i$ , 其中  $\mathbf{D}_i$  是一个  $(T-1) \times T$  维矩阵差分算子

$$\mathbf{D}_i = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

对 (8.15) 和 (8.16) 实施差分变换得到

$$\Delta Y_{it} = \Delta X_{it}'\beta + \Delta \varepsilon_{it}$$

以及

$$\Delta Y_i = \Delta X_i \beta + \Delta \varepsilon_i \quad (8.26)$$

经过差分变换后, 不可观测的  $u_i$  已经被消去.

现在对回归方程 (8.26) 应用最小二乘可得

$$\begin{aligned}\hat{\beta}_{\text{FD}} &= \left( \sum_{i=1}^N \sum_{t \geq 2} \Delta X_{it} \Delta X'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t \geq 2} \Delta X_{it} \Delta Y_{it} \right) \\ &= \left( \sum_{i=1}^N \Delta X'_i \Delta X_i \right)^{-1} \left( \sum_{i=1}^N \Delta X'_i \Delta Y_i \right) \\ &= \left( \sum_{i=1}^N X'_i \mathbf{D}'_i \mathbf{D}_i X_i \right)^{-1} \left( \sum_{i=1}^N X'_i \mathbf{D}'_i \mathbf{D}_i Y_i \right)\end{aligned}\quad (8.27)$$

我们称 (8.27) 为一阶差分估计量 (first-difference estimator). 此外, 当  $T = 2$  时有  $\hat{\beta}_{\text{FD}} = \hat{\beta}_{\text{FE}}$ , 这是因为

$$\mathbf{D}'_i \mathbf{D}_i = 2\mathbf{M}_i = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

当特质误差  $\{\varepsilon_{it}\}$  是条件同方差的且序列无关, 那么  $\Delta \varepsilon_i = \mathbf{D}_i \varepsilon_i$  具有协方差矩阵  $\sigma_\varepsilon^2 \mathbf{H}$ , 其中  $\mathbf{H} = \mathbf{D}'_i \mathbf{D}_i$ . 另一方面, 此时我们可以用 GLS 来估计方程 (8.26) 并得到

$$\begin{aligned}\tilde{\beta}_{\text{FD}} &= \left( \sum_{i=1}^N \Delta X'_i \mathbf{H}^{-1} \Delta X_i \right)^{-1} \left( \sum_{i=1}^N \Delta X'_i \mathbf{H}^{-1} \Delta Y_i \right) \\ &= \left[ \sum_{i=1}^N X'_i \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i X_i \right]^{-1} \left[ \sum_{i=1}^N X'_i \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i Y_i \right] \\ &= \left( \sum_{i=1}^N X'_i \mathbf{M}_i X_i \right)^{-1} \left( \sum_{i=1}^N X'_i \mathbf{M}_i Y_i \right)\end{aligned}$$

其中  $\mathbf{M}_i = \mathbf{D}'_i (\mathbf{D}_i \mathbf{D}'_i)^{-1} \mathbf{D}_i$ , 这可以由线性代数直接验证.

也就是说, 如果特质误差是 i.i.d. 的, 那么对一阶差分方程使用 GLS 方法得到的估计量是 FE 估计量. 由于 Gauss-Markov 定理表明 GLS 估计量具有渐近有效的方差, 故而在 i.i.d. 特质误差下, FE 估计量比 FD 估计量更有效.

#### 8.4.4 组间估计

所谓的组间估计量 (between estimator) 来源于均值化的回归方程

$$\bar{Y}_i = \bar{X}_i \beta + u_i + \bar{\varepsilon}_i \quad (8.28)$$

对 (8.28) 施加最小二乘可得

$$\hat{\beta}_{\text{BE}} = \left( \sum_{i=1}^N \bar{X}_i \bar{X}'_i \right)^{-1} \left( \sum_{i=1}^N \bar{X}_i \bar{Y}_i \right)$$

在假设8.1下,  $\hat{\beta}_{BE}$  是  $\beta$  的无偏估计量, 并且它的方差为

$$V_{BE} = \sigma^2 \left( \sum_{i=1}^N \bar{X}_i \bar{X}_i' \right)^{-1}$$

其中

$$\sigma^2 = \text{var}(u_i + \bar{\varepsilon}_i) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T}$$

是回归 (8.28) 的误差方差. 然而,  $\hat{\beta}_{BE}$  的有效性不及由 FGLS 估计得到的  $\hat{\beta}_{RE}$ . 此外, 如果  $u_i$  和  $X_i$  任意相关, 那么 BE 估计量必然不是一致的.

根据上述讨论, 我们似乎没有直接使用组间估计的动机, 但组间估计对于构造  $\sigma_u^2$  的估计量很有用. 首先考虑

$$\sigma_b^2 = N^{-1} \sum_{i=1}^N \sigma_i^2 = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T} \quad (8.29)$$

它的估计量可以很自然地考虑为

$$\hat{\sigma}_b^2 = \frac{1}{N-K} \sum_{i=1}^N \hat{e}_{bi}^2 \quad (8.30)$$

其中  $\hat{e}_{bi} = \bar{Y}_i - \bar{X}_i' \hat{\beta}_{BE}$  为组间残差.

我们在之前已经得到了  $\hat{\sigma}_\varepsilon^2$  的表达式 (8.24), 既然它在条件更弱的 FE 估计中适用, 那么它也在 RE 估计中适用. 基于 (8.29) 可得  $\sigma_u^2$  的估计量

$$\hat{\sigma}_u^2 = \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\varepsilon^2}{T}$$

因为它可能是负的, 我们通常使用受限的估计量

$$\hat{\sigma}_u^2 = \max \left\{ 0, \hat{\sigma}_b^2 - \frac{\hat{\sigma}_\varepsilon^2}{T} \right\} \quad (8.31)$$

由此得到的估计量  $\hat{\sigma}_\varepsilon^2$  和  $\hat{\sigma}_u^2$  可用于构建 RE 估计量中的  $\hat{\Omega}$ .

## 8.5 广义离差模型

以上章节介绍了 POOL 估计量、RE 估计量和 FE 估计量, 实际上三者都可以使用 OLS 对广义离差模型 (quasi-demeaned model) 进行估计得到.

考虑具有随机效应结构的综合误差  $e_i$  的协方差矩阵

$$\Omega = \begin{bmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\varepsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

定义矩阵  $P_i = \mathbf{1}_i(\mathbf{1}_i' \mathbf{1}_i)^{-1} \mathbf{1}_i'$ , 于是  $M_i = I_i - P_i$  以及  $P_i Y_i = \mathbf{1}_i \bar{Y}_i$ , 此时

$$\Omega = \sigma_\varepsilon^2 I_i + \sigma_u^2 \mathbf{1}_i \mathbf{1}_i' = \sigma_\varepsilon^2 \left( I_i + \frac{\sigma_u^2 T}{\sigma_\varepsilon^2} P_i \right) = \sigma_\varepsilon^2 (M_i + \rho^{-2} P_i)$$

其中

$$\rho = \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + \sigma_u^2 T}} \quad (8.32)$$

于是  $\mathbf{\Omega}^{-1} = \sigma_\varepsilon^{-2}(\mathbf{M}_i + \rho^2 \mathbf{P}_i)$ , 并且

$$\mathbf{\Omega}^{-\frac{1}{2}} = \sigma_\varepsilon^{-1}(\mathbf{M}_i + \rho \mathbf{P}_i) = \sigma_\varepsilon^{-1}[\mathbf{I}_i - (1 - \rho)\mathbf{P}_i]$$

再令  $\tilde{X}_i = \mathbf{\Omega}^{-\frac{1}{2}} X_i$ ,  $\tilde{Y}_i = \mathbf{\Omega}^{-\frac{1}{2}} Y_i$ , 于是 GLS 估计量为

$$\begin{aligned} \hat{\beta}_{\text{GLS}} &= \left( \sum_{i=1}^N X_i' \mathbf{\Omega}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^N X_i' \mathbf{\Omega}^{-1} Y_i \right) \\ &= \left( \sum_{i=1}^N \tilde{X}_i' \tilde{X}_i \right)^{-1} \left( \sum_{i=1}^N \tilde{X}_i' \tilde{Y}_i \right) \end{aligned} \quad (8.33)$$

其中

$$\tilde{Y}_i = \sigma_\varepsilon^{-1} [Y_i - (1 - \rho) \mathbf{1}_i \bar{Y}_i] \quad (8.34)$$

同理可得

$$\tilde{X}_i = \sigma_\varepsilon^{-1} [X_i - (1 - \rho) \mathbf{1}_i \bar{X}_i] \quad (8.35)$$

$$\tilde{e}_i = \sigma_\varepsilon^{-1} [e_i - (1 - \rho) \mathbf{1}_i \bar{e}_i] \quad (8.36)$$

然而它们是不可行的, 因为  $\rho$  不可观测.

为了得到 FGLS 估计量, 考虑使用估计量  $\hat{\sigma}_\varepsilon^2$  和  $\hat{\sigma}_u^2$  分别替代 (8.32) 中的  $\sigma_\varepsilon^2$  和  $\sigma_u^2$ , 得到估计量

$$\hat{\rho} = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 T}}$$

以及

$$\hat{\theta} = 1 - \frac{\hat{\sigma}_\varepsilon}{\sqrt{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 T}}$$

将其代入到 (8.34), (8.35) 和 (8.36) 中, 于是估计的方程变为

$$Y_{it} - \hat{\theta} \bar{Y}_i = (X_{it} - \hat{\theta} \bar{X}_i)' \beta + e_{it} - \hat{\theta} \bar{e}_i$$

通过对上式进行 OLS 回归来得到 FGLS 估计量  $\hat{\beta}_{\text{FGLS}}$ .

根据之前的讨论, 如果  $\hat{\theta} = 0$ , 那么有  $\tilde{X}_i = X_i$  以及  $\tilde{Y}_i = Y_i$ , 因此  $\hat{\beta}_{\text{FGLS}} = \hat{\beta}_{\text{POOL}}$ . 而当  $\hat{\theta} = 1$  时有,  $\tilde{X}_i = \dot{X}_i$  以及  $\tilde{Y}_i = \dot{Y}_i$ , 故而  $\hat{\beta}_{\text{FGLS}} = \hat{\beta}_{\text{FE}}$ . 最后, 当  $0 < \hat{\theta} < 1$  时,  $\hat{\beta}_{\text{FGLS}}$  就是 RE 估计量  $\hat{\beta}_{\text{RE}}$ .

## 8.6 FE 的渐近性质

之前的部分已经讨论了, 在假设 8.1 成立的条件下, RE 估计量具有一致性, 本节主要讨论 FE 估计量的渐近性质, 为此先给出以下假设.

**假设 8.3**

- (1) 回归模型为  $Y_{it} = X'_{it}\beta + u_i + \varepsilon_{it}$ ,  $i = 1, \dots, N$ ,  $t = 1, \dots, T$  且  $T \geq 2$ .
- (2)  $(\varepsilon_i, X_i)$  是 i.i.d. 的, 其中  $i = 1, \dots, N$ .
- (3) 对所有的  $s = 1, \dots, T$  都有  $\mathbb{E}[X_{is}\varepsilon_{it}] = 0$ .
- (4)  $\mathbf{Q}_T = \mathbb{E}[\dot{X}'_i \dot{X}_i] > 0$ .
- (5)  $\mathbb{E}[\varepsilon_{it}^4] < \infty$ .
- (6)  $\mathbb{E}[\|X_{it}^4\|] < \infty$ .

**定理 8.1**

在假设 8.3 下, 当  $N \rightarrow \infty$  时有

$$\sqrt{N}(\hat{\beta}_{\text{FE}} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta)$$

其中  $\mathbf{V}_\beta = \mathbf{Q}_T^{-1} \mathbf{\Omega}_T \mathbf{Q}_T^{-1}$ , 以及  $\mathbf{\Omega}_T = \mathbb{E}[\dot{X}'_i \varepsilon_i \varepsilon'_i \dot{X}_i]$ .



**证明** 假设 8.2 意味着  $(\dot{X}_i, \varepsilon_i)$  也是 i.i.d. 的, 并且具有有限四阶矩, 根据 WLLN 可知

$$N^{-1} \sum_{i=1}^N \dot{X}'_i \dot{X}_i \xrightarrow{p} \mathbb{E}[\dot{X}'_i \dot{X}_i] = \mathbf{Q}_T$$

另一方面

$$\mathbb{E}[\dot{X}'_i \varepsilon_i] = \sum_{t=1}^T \mathbb{E}[\dot{X}'_{it} \varepsilon_{it}] = \sum_{t=1}^T \mathbb{E}[X_{it} \varepsilon_{it}] - \sum_{t=1}^T \sum_{j=1}^T \mathbb{E}[X_{ij} \varepsilon_{it}] = 0$$

于是由 CLT 可知

$$N^{-\frac{1}{2}} \sum_{i=1}^N \dot{X}'_i \varepsilon_i \xrightarrow{d} N(0, \mathbf{\Omega}_T)$$

其中假设 8.3 的第 (5) 点和第 (6) 点保证了使用 CLT 的前提条件成立. 进一步

$$\sqrt{N}(\hat{\beta}_{\text{FE}} - \beta) = \left( N^{-1} \sum_{i=1}^N \dot{X}'_i \dot{X}_i \right)^{-1} \left( N^{-\frac{1}{2}} \sum_{i=1}^N \dot{X}'_i \varepsilon_i \right) \xrightarrow{d} N(0, \mathbf{V}_\beta)$$

证毕.

之前已经介绍了  $\mathbf{V}_{\text{FE}}$  的完全稳健的估计量 (8.20), 以及在特质误差  $\varepsilon_{it}$  满足同方差和无序列相关时的估计量 (8.23). 现在来看  $\{\varepsilon_{it}\}$  仅存在异方差而无序列相关时的情形, 此时  $\mathbb{E}[\varepsilon_{it}|X_i] = 0$  以及

$$\mathbb{E}[\varepsilon_{it}^2|X_i] = \sigma_{it}^2 \quad (8.37)$$

故而  $\mathbf{\Sigma}_i = \mathbb{E}[\varepsilon_i \varepsilon'_i|X_i] = \text{diag}\{\sigma_{it}^2\}$ , 因此协方差矩阵 (8.19) 变为

$$\mathbf{V}_{\text{FE}} = (\dot{X}' \dot{X})^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \dot{X}_{it} \dot{X}'_{it} \sigma_{it}^2 \right) (\dot{X}' \dot{X})^{-1}$$

使用  $\hat{\varepsilon}_{it}^2$  替代上式中的  $\sigma_{it}^2$ , 再做一个自由度调整可得异方差稳健的协方差矩阵估计量

$$\hat{V}_{FE} = \frac{NT}{N(T-1)-K} (\dot{X}'\dot{X})^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \dot{X}_{it} \dot{X}_{it}' \hat{\varepsilon}_{it}^2 \right) (\dot{X}'\dot{X})^{-1} \quad (8.38)$$

直觉上看, White (1980) 形式的  $\hat{V}_{FE}$  是  $V_{FE}$  合理的估计量, 然而 Stock and Waston (2008) 指出了这种错误.

考虑一种特殊情况,  $\hat{\varepsilon}_{it} = \dot{Y}_{it} - \dot{X}_{it}'\beta$ , 也即残差  $\hat{\varepsilon}_{it}$  通过真实参数  $\beta$  构成, 此时

$$\hat{\varepsilon}_{it} = \dot{\varepsilon}_{it} = \varepsilon_{it} - T^{-1} \sum_{j=1}^T \varepsilon_{ij}$$

在  $\varepsilon_{it}$  仅存在异方差而无自相关的情况下, 根据 (8.22) 和 (8.37) 可得

$$\mathbb{E}[\hat{\varepsilon}_{it}^2 | X_i] = \left( \frac{T-2}{T} \right) \sigma_{it}^2 + \frac{\sigma_i^2}{T}$$

其中  $\sigma_i^2 = T^{-1} \sum_{t=1}^T \sigma_{it}^2$ . 假设  $K=0$ , 将上式代入到 (8.38) 得到

$$\begin{aligned} \mathbb{E}[\hat{V}_{FE} | X] &= \frac{T}{T-1} (\dot{X}'\dot{X})^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \dot{X}_{it} \dot{X}_{it}' \mathbb{E}[\hat{\varepsilon}_{it}^2 | X_i] \right) (\dot{X}'\dot{X})^{-1} \\ &= \left( \frac{T-2}{T-1} \right) V_{FE} + \frac{1}{T-1} (\dot{X}'\dot{X})^{-1} \left( \sum_{i=1}^N \dot{X}_{it} \dot{X}_{it}' \bar{\sigma}_i^2 \right) (\dot{X}'\dot{X})^{-1} \end{aligned}$$

当  $T > 2$  且固定时, 随着  $N \rightarrow \infty$ ,  $\hat{V}_{FE}$  仍是有偏估计. Stock and Waston (2008) 还表明了它不是  $V_{FE}$  的一致估计量 (除非  $T \rightarrow \infty$ ), 并且提出了

$$\begin{aligned} \tilde{V}_{FE} &= \left( \frac{T-1}{T-2} \right) \hat{V}_{FE} - \frac{1}{T-1} \hat{B}_{FE} \\ \hat{B}_{FE} &= (\dot{X}'\dot{X})^{-1} \left( \sum_{i=1}^N \dot{X}_i' \dot{X}_i \hat{\sigma}_i^2 \right) (\dot{X}'\dot{X})^{-1} \\ \hat{\sigma}_i^2 &= \frac{1}{T-1} \sum_{t=1}^T \hat{\varepsilon}_{it}^2 \end{aligned}$$

这里的  $\tilde{V}_{FE}$  是  $V_{FE}$  的无偏估计, 并且在  $N \rightarrow \infty$  但  $T$  固定的情况下也是  $V_{FE}$  的一致估计.

**注** 尽管  $\hat{V}_{FE}$  在上述情况下不是一致估计量, 但如果  $\{\varepsilon_{it}\}$  满足条件同方差和无序列相关, 那么  $\hat{V}_{FE}^0$  仍是一致估计. 此外, 对于异方差和序列相关完全稳健的  $\hat{V}_{FE}$  也是一致的.

## 8.7 RE 与 FE 的比较

在选择 RE 和 FE 的关键考虑是, 不可观测的个体效应  $u_i$  和解释变量  $X_{it}$  是否相关. 如果二者是相关的, 那么 FE 估计量是一致的, 而 RE 估计量不一致. 如果二者不相关, 那么 RE 估计量同样是一致的, 并且渐近有效, 这是因为 RE 估计量是通过 FGLS 获得的.

为了比较这两种估计量, 考虑之前在工具变量章节中使用过的 Hausman 检验, 在此之前先介绍一个引理, 它仍然出自于 Hausman (1978).

**引理 8.1**

设  $\hat{\beta}_0$  和  $\hat{\beta}_1$  都是  $\beta$  的一致估计量, 并且都服从渐近正态分布

$$\sqrt{N}(\hat{\beta}_0 - \beta) \xrightarrow{d} N(0, V_0)$$

$$\sqrt{N}(\hat{\beta}_1 - \beta) \xrightarrow{d} N(0, V_1)$$

定义  $\hat{q} = \hat{\beta}_0 - \hat{\beta}_1$ , 如果  $\hat{\beta}_0$  是渐近有效的, 那么  $\sqrt{N}(\hat{\beta}_0 - \beta)$  和  $\sqrt{N}\hat{q}$  的极限分布的协方差为 0, 也即  $\text{cov}(\hat{\beta}_0, \hat{q}) = 0$ .



**证明** 假设  $C = \text{cov}(\hat{\beta}_0, \hat{q}) \neq 0$ , 定义估计量  $\hat{\beta}_2 = \hat{\beta}_0 + rA\hat{q}$ , 其中  $r$  是一个标量,  $A$  为任意  $K \times K$  维矩阵, 于是

$$\text{var}(\hat{\beta}_2) = \text{var}(\hat{\beta}_0) + rAC + rC'A' + r^2A\text{var}(\hat{q})A'$$

设函数

$$F(r) = \text{var}(\hat{\beta}_2) - \text{var}(\hat{\beta}_0) = rAC + rC'A' + r^2A\text{var}(\hat{q})A'$$

它的一阶导数为

$$F'(r) = AC + C'A' + 2rA\text{var}(\hat{q})A'$$

选取  $A = -C'$ , 注意到协方差矩阵  $C$  是对称的, 因此

$$F'(r) = -2C'C + 2rC'\text{var}(\hat{q})C$$

当  $r = 0$  时,  $F'(0) = -2C'C$  为半负定矩阵, 又因为  $F(0) = 0$ , 因此存在某个充分小的  $r > 0$ , 使得  $F(r) < 0$ . 但由于  $\hat{\beta}_0$  是渐近有效的, 故而  $F(r) \geq 0$ , 产生矛盾, 因此必有  $C = 0$ .

**定理 8.2**

在假设 8.1 和秩条件  $\text{rank}(\mathbb{E}[\dot{X}_i' \dot{X}_i]) = K$  成立的条件下, 当  $N \rightarrow \infty$  时有

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'(\hat{V}_{FE} - \hat{V}_{RE})^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \xrightarrow{d} \chi_K^2$$



**证明** 当  $N \rightarrow \infty$  时有  $\sqrt{N}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \xrightarrow{d} N(0, V)$ , 其中

$$V/n = \text{avar}(\hat{\beta}_{FE}) + \text{avar}(\hat{\beta}_{RE}) - \text{cov}(\hat{\beta}_{FE}, \hat{\beta}_{RE}) - \text{cov}(\hat{\beta}_{RE}, \hat{\beta}_{FE})$$

由于在假设 8.1 下, RE 估计量是渐近有效的, 因此根据引理 8.1 可知

$$\text{cov}(\hat{\beta}_{FE} - \hat{\beta}_{RE}, \hat{\beta}_{RE}) = \text{cov}(\hat{\beta}_{FE}, \hat{\beta}_{RE}) - \text{avar}(\hat{\beta}_{RE}) = 0$$

从而  $V = nV_{FE} - nV_{RE}$ . 根据 Wald 检验原理可知

$$\begin{aligned} H &= n(\hat{\beta}_{FE} - \hat{\beta}_{RE})' \hat{V}^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \\ &= (\hat{\beta}_{FE} - \hat{\beta}_{RE})'(\hat{V}_{FE} - \hat{V}_{RE})^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \end{aligned}$$

因此当  $N \rightarrow \infty$  时有  $H \xrightarrow{d} \chi_K^2$ .

**注** 下面给出关于 Hausman 检验的一些评注.

- Hausman 检验只在假设 8.1 完全成立的情况下才能使用, 也即特质误差  $\{\varepsilon_{it}\}$  不能存在异



方差和序列相关, 否则  $\hat{\beta}_{\text{RE}}$  不再是渐近有效的估计量, 导致无法使用引理 8.1 构造  $H$  统计量.

- 此外, 由于组内变换只能估计时变解释变量的回归系数, 而 RE 方法没有这个限制, 因此  $\hat{\beta}_{\text{RE}}$  的维数可能多于  $\hat{\beta}_{\text{FE}}$  的维数, 此时 Hausman 检验只能对时变的回归元上进行.
- 有时候, Stata 之类的计量软件会给出负的  $H$  统计量, 这可能是由于对  $\hat{\sigma}_\varepsilon^2$  的不同估计导致的, 此时  $H$  统计量中间的那个矩阵不一定正定, 因此最好在  $\hat{V}_{\text{FE}}$  和  $\hat{V}_{\text{RE}}$  中都使用同一个估计量  $\hat{\sigma}_\varepsilon^2$ .

如果特质误差  $\varepsilon_{it}$  存在条件异方差, 那么可以考虑广义离差模型

$$Y_{it} - \hat{\theta}\bar{Y}_i = (X_{it} - \hat{\theta}\bar{X}_i)'\beta + (X_{it} - \bar{X}_i)'\gamma + e_{it} - \hat{\theta}\bar{e}_i$$

对上式进行 OLS 估计后可以获得聚类稳健协方差矩阵估计量, 然后利用 Wald 检验原理即可检验原假设  $\mathbb{H}_0: \gamma = 0$ , 倘若我们拒绝原假设, 则应该拒绝随机效应.

目前的计量分析更倾向于稳健性而非有效性, 通常在线性面板数据模型中直接使用 FE 分析, 而只有在非线性模型中才使用 RE 分析, 因为在非线性模型中通常难以估计出  $\hat{\beta}_{\text{FE}}$ .

## 8.8 双向误差成分

之前讨论的面板数据模型仅将综合误差  $e_{it}$  分解为个体效应  $u_i$  和特质误差  $\varepsilon_{it}$ , 然而实际上可能存在仅随时间变化而不随个体变化的不可观测效应, 由此给出双向误差成分模型 (two-way error component model)

$$Y_{it} = X'_{it}\beta + u_i + v_t + \varepsilon_{it} \quad (8.39)$$

其中  $u_i$  是不可观测的个体效应,  $v_t$  是不可观测的时间效应,  $\varepsilon_{it}$  为特质误差项. 模型 (8.39) 既可以使用 RE 估计也可以使用 FE 估计.

在随机效应框架下, 需要将假设 8.1 拓展到  $v_t$  上, 由于  $e = v \otimes \mathbf{1}_N + \mathbf{1}_T \otimes u + \varepsilon$ , 于是

$$\Omega = (\mathbf{1}_T \otimes \mathbf{1}_N \mathbf{1}_N')\sigma_v^2 + (\mathbf{1}_T \mathbf{1}_T' \otimes \mathbf{1}_N)\sigma_u^2 + \sigma_\varepsilon^2 \mathbf{I}_{NT}$$

它可以在 GLS 中被用于估计  $\beta$ .

在固定效应框架下, 可以使用双向组内变换

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_i - \tilde{Y}_t + \bar{Y}$$

其中

$$\tilde{Y}_t = N^{-1} \sum_{i=1}^N Y_{it}, \quad \bar{Y} = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T Y_{it}$$

类似可以定义  $\ddot{X}_{it} = X_{it} - \bar{X}_i - \tilde{X}_t + \bar{X}$ . 如果  $Y_{it}$  满足 (8.39), 那么

$$\bar{Y}_i = \bar{X}'_i \beta + u_i + \bar{v} + \bar{\varepsilon}_i$$

$$\tilde{Y}_t = \tilde{X}'_t \beta + \bar{u} + v_t + \tilde{\varepsilon}_t$$

$$\bar{Y} = \bar{X}' \beta + \bar{u} + \bar{v} + \bar{\varepsilon}$$

从而

$$\ddot{Y}_{it} = \ddot{X}'_{it}\beta + \ddot{\varepsilon}_{it}$$

对上式使用最小二乘即可得到出双向组内估计量.

除了使用双向组内变换外, 还可以使用 LSDV 方法来估计 (8.39). 设  $\tau_t$  为  $T \times 1$  维向量, 它的第  $t$  个元素为 1, 而其它元素为 0, 再定义时间固定效应向量  $v = [v_1, \dots, v_T]'$ , 于是双向成分模型可以写为

$$Y_{it} = X'_{it}\beta + u_i + \tau'_t v + \varepsilon_{it}$$

注意, 为了使得模型可以识别, 避免完全多重共线性, 需要在  $\tau_t$  中剔除某个基期时间虚拟变量, 然后使用组内估计即可.

如果使用了双向组内变换, 那么  $X_{it}$  中就不能包括任意非时变的回归元  $X_i$ , 以及任意时间序列变量  $X_t$ , 它们在双向组内变换过程中都会被消去. 由于出自 (8.39) 的固定效应估计量关于  $u_i$  和  $v_t$  是不变的, 因此无需新增假设即可分析其渐近性质.

## 8.9 工具变量

考虑固定效应模型

$$Y_{it} = X'_{it}\beta + u_i + \varepsilon_{it} \quad (8.40)$$

我们称  $X_{it}$  是内生的, 如果  $\mathbb{E}[X_{it}X_{is}] \neq 0$ , 这将导致  $\hat{\beta}_{FE}$  是结构参数  $\beta$  的非一致估计量, 我们可以使用工具变量解决内生性问题.

首先令  $Z_{it}$  为  $L \times 1$  维工具向量, 并且  $L \geq K$ . 在截面数据的情况下,  $Z_{it}$  既包括  $X_{it}$  的外生部分, 又包括来自  $X_{it}$  外的排他性外生变量. 然后再定义  $Z_i$  为在个体  $i$  上堆叠起来的  $T \times L$  维工具矩阵,  $Z$  为全样本堆叠起来的工具矩阵. 结构参数  $\beta$  的识别仍依赖于矩条件  $\text{rank}(\mathbb{E}[\dot{Z}'_i \dot{X}_i]) = K$ , 以及排除完全多重共线性所需要的  $\mathbb{E}[\dot{Z}'_i \dot{Z}_i] > 0$ .

再来考虑如下有关固定效应的虚拟变量回归

$$Y_{it} = X'_{it}\beta + d'_i u + \varepsilon_{it}$$

其中  $d_i$  为  $N \times 1$  维虚拟变量向量, 它的全样本形式为

$$Y = X\beta + Du + \varepsilon \quad (8.41)$$

根据之前的讨论, 固定效应估计量  $\hat{\beta}_{FE}$  可以由对上式的 OLS 回归获得, 并且还可以将  $D$  视为外生变量. 于是, 固定效应模型 (8.40) 的 2SLS 估计在代数上, 等于在 (8.41) 中使用  $[Z, D]$  作为工具的  $Y$  对  $[X, D]$  的 2SLS 估计. 根据 (6.18) 可知

$$\hat{\beta}_{FE2SLS} = [X'M_D Z(Z'M_D Z)^{-1}Z'M_D X]^{-1}X'M_D Z(Z'M_D Z)^{-1}Z'M_D Y \quad (8.42)$$

其中  $M_D = I_{NT} - D(D'D)^{-1}D'$ . 注意到  $M_D Y = \dot{Y}$ ,  $M_D X = \dot{X}$ , 以及  $M_D Z = \dot{Z}$ , 于是

$$\hat{\beta}_{FE2SLS} = [\dot{X}'\dot{Z}(\dot{Z}'\dot{Z})^{-1}\dot{Z}'\dot{X}]^{-1}\dot{X}'\dot{Z}(\dot{Z}'\dot{Z})^{-1}\dot{Z}'\dot{Y}$$

如果是双向固定效应模型

$$Y_{it} = X'_{it}\beta + u_i + v_t + \varepsilon_{it}$$

只需要添加  $T - 1$  个时间虚拟变量到回归中, 然后将所有的虚拟变量组合成前面提到的矩阵  $\mathbf{D}$ , 然后利用 (8.42) 即可得到 2SLS 估计量.

为了得到 FE2SLS 估计量的渐近性质, 这里给出一系列正则条件.

#### 假设 8.4

- (1) 回归模型为  $Y_{it} = X'_{it}\beta + u_i + \varepsilon_{it}$ ,  $i = 1, \dots, N$ ,  $t = 1, \dots, T$  且  $T \geq 2$ .
- (2)  $(\varepsilon_i, X_i, Z_i)$  是 i.i.d. 的, 其中  $i = 1, \dots, N$ .
- (3) 对所有的  $s = 1, \dots, T$  都有  $\mathbb{E}[Z_{is}\varepsilon_{it}] = 0$ .
- (4)  $\mathbf{Q}_{ZZ} = \mathbb{E}[\dot{Z}'_i \dot{Z}_i] > 0$ .
- (5)  $\text{rank}(\mathbf{Q}_{ZX}) = K$ , 其中  $\mathbf{Q}_{ZX} = \mathbb{E}[\dot{Z}'_i \dot{X}_i]$ .
- (6)  $\mathbb{E}[\varepsilon_{it}^4] < \infty$ .
- (7)  $\mathbb{E}[\|\dot{X}_{it}^4\|] < \infty$ .
- (8)  $\mathbb{E}[\|\dot{Z}_{it}^4\|] < \infty$ .

#### 定理 8.3

在假设 8.4 下, 当  $N \rightarrow \infty$  时有  $\sqrt{N}(\hat{\beta}_{\text{FE2SLS}} - \beta) \xrightarrow{d} N(0, \mathbf{V}_\beta)$ , 其中

$$\mathbf{V}_\beta = (\mathbf{Q}'_{ZX} \mathbf{Q}^{-1}_{ZZ} \mathbf{Q}_{ZX})^{-1} (\mathbf{Q}'_{ZX} \mathbf{Q}^{-1}_{ZZ} \mathbf{\Omega}_{Z\varepsilon} \mathbf{Q}^{-1}_{ZZ} \mathbf{Q}_{ZX}) (\mathbf{Q}'_{ZX} \mathbf{Q}^{-1}_{ZZ} \mathbf{Q}_{ZX})^{-1}$$

以及  $\mathbf{\Omega}_{Z\varepsilon} = \mathbb{E}[\dot{Z}'_i \varepsilon_i \varepsilon'_i \dot{Z}_i]$ .

它的证明与定理 8.1 的类似, 在此略过. 如果特质误差  $\{\varepsilon_{it}\}$  满足条件同方差和无序列相关, 那么可以将  $\mathbf{V}_\beta$  简化为

$$\mathbf{V}_\beta = \sigma_\varepsilon^2 (\mathbf{Q}'_{ZX} \mathbf{Q}^{-1}_{ZZ} \mathbf{Q}_{ZX})^{-1}$$

遵循标准步骤,  $\hat{\beta}_{\text{FE2SLS}}$  的完全稳健协方差矩阵估计量为

$$\begin{aligned} \hat{\mathbf{V}}_{\text{FE2SLS}} &= [\dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}}]^{-1} \dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \left( \sum_{i=1}^N \dot{Z}'_i \hat{\varepsilon}_i \hat{\varepsilon}'_i \dot{Z}_i \right)^{-1} \\ &\quad \times (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}} [\dot{\mathbf{X}}' \dot{\mathbf{Z}} (\dot{\mathbf{Z}}' \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}' \dot{\mathbf{X}}]^{-1} \end{aligned}$$

其中  $\hat{\varepsilon}_i = \dot{Y}_i - \dot{X}_i \hat{\beta}_{\text{FE2SLS}}$  为 FE2SLS 残差. 根据 Stock and Waston (2008) 的分析, 仍不建议使用异方差稳健的协方差矩阵估计量, 特别是当  $T$  很小的时候.

## 8.10 Hausman-Taylor 模型

FE 分析无法囊括非时变的回归元, 而 RE 分析尽管可以将其纳入, 但所需条件太强, Hausman and Taylor (1981) 提出了介于二者之间的模型, 该模型关于时变变量的假设与 FE 的假设

相同, 而对于非时变变量的假设则更强一些. 考虑模型

$$Y_{it} = X'_{1it}\beta_1 + X'_{2it}\beta_2 + Z'_{1i}\gamma_1 + Z'_{2i}\gamma_2 + u_i + \varepsilon_{it}$$

其中  $X_{1it}$  和  $X_{2it}$  为时变变量,  $Z_{1i}$  和  $Z_{2i}$  为非时变变量. 回归元  $X_{1it}$ ,  $X_{2it}$ ,  $Z_{1i}$ ,  $Z_{2i}$  的维数分别为  $K_1$ ,  $K_2$ ,  $L_1$ ,  $L_2$ . 上述模型的全样本形式为

$$Y = X_1\beta_1 + X_2\beta_2 + Z_1\gamma_1 + Z_2\gamma_2 + u + \varepsilon \quad (8.43)$$

再设  $\bar{X}_1$  和  $\bar{X}_2$  表示在个体上的平均,  $\dot{X}_1 = X_1 - \bar{X}_1$  和  $\dot{X}_2 = X_2 - \bar{X}_2$  为组内变换.

Hausman-Taylor 模型假设所有回归元在每个时间  $t$  上和特质误差  $\varepsilon_{it}$  无关,  $X_{1it}$  和  $Z_{1i}$  关于个体效应  $u_i$  是外生的, 也即

$$\mathbb{E}[X_{1it}u_i] = 0$$

$$\mathbb{E}[Z_{1i}u_i] = 0$$

而回归元  $X_{2it}$  和  $Z_{2i}$  可以与个体效应  $u_i$  产生相关性. 定义矩阵  $X = [X_1, X_2, Z_1, Z_2]$ ,  $\beta = [\beta'_1, \beta'_2, \gamma'_1, \gamma'_2]'$ , 可以将以上假设写为总体矩条件

$$\mathbb{E}[\dot{X}'_1(Y - X\beta)] = 0$$

$$\mathbb{E}[\dot{X}'_2(Y - X\beta)] = 0$$

$$\mathbb{E}[\bar{X}'_1(Y - X\beta)] = 0$$

$$\mathbb{E}[Z'_1(Y - X\beta)] = 0$$

这里的矩条件的个数为  $2K_1 + K_2 + L_1$ , 而回归系数共计  $K_1 + K_2 + L_1 + L_2$  个, 因此为了模型可以识别, 必须保证  $K_1 \geq L_2$ .

由于模型将  $X_1$  和  $Z_1$  视为外生的, 而将  $X_2$  和  $Z_2$  视为内生的, 故而可以使用  $\dot{X}_2$  和  $X_1$  作为它们的工具, 因此使用  $Z = [\dot{X}_1, \dot{X}_2, \bar{X}_1, \dot{Z}_1]$  为工具的 2SLS 方法即可估计出  $\hat{\beta}_{2SLS}$ , 并且它是一致的.

Hausman and Taylor (1981) 在更强的假设条件下获得估计量, 将  $\varepsilon_{it}$  设置为与  $u_i$  均值独立,  $\Omega_i$  具有随机效应结构  $\Omega_i = \sigma_\varepsilon^2 \mathbf{I}_n + \sigma_u^2 \mathbf{1}_i \mathbf{1}'_i$ , 并且  $\Omega = \mathbf{I}_N \otimes \Omega_i$ .

首先做组内变换得到

$$\dot{Y}_{it} = \dot{X}'_{1it}\beta_1 + \dot{X}'_{2it}\beta_2 + \dot{\varepsilon}_{it}$$

取得  $[\beta'_1, \beta'_2]'$  的固定效应估计量  $[\hat{\beta}'_{FE1}, \hat{\beta}'_{FE2}]'$ , 以及残差  $\hat{\varepsilon}_i = \dot{Y}_i - \dot{X}_{1i}\hat{\beta}_{FE1} - \dot{X}_{2i}\hat{\beta}_{FE2}$ , 并且可以构造  $\sigma_\varepsilon^2$  的一致估计量

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)} \sum_{i=1}^N \hat{\varepsilon}'_i \hat{\varepsilon}_i$$

然后定义

$$\tilde{\varepsilon}_i = Y_i - \bar{X}_{1i}\hat{\beta}_{FE1} - \bar{X}_{2i}\hat{\beta}_{FE2}$$

使用  $\tilde{\varepsilon}_i$  对  $Z_{1i}$  和  $Z_{2i}$  进行 2SLS 回归, 工具变量为  $X_1$  和  $Z_1$ , 由此得到估计量  $\hat{\gamma}_{IV1}$  和  $\hat{\gamma}_{IV2}$ . 定义

$$\check{\varepsilon}_{it} = Y_{it} - X'_{1it}\hat{\beta}_{FE1} - X'_{2it}\hat{\beta}_{FE2} - Z'_{1it}\hat{\gamma}_{IV1} - Z'_{2it}\hat{\gamma}_{IV2}$$

以及

$$s^2 = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \tilde{\varepsilon}_{it}^2$$

根据  $\text{plim } s^2 = \sigma_\varepsilon^2 + \sigma_u^2 T$  就能得到估计量  $\hat{\sigma}_u^2 = s^2 - \hat{\sigma}_\varepsilon^2 / T$ , 因此可以定义

$$\hat{\theta}_i = 1 - \frac{\hat{\sigma}_\varepsilon}{\sqrt{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 T}}$$

进一步定义

$$\begin{aligned} Y_{it}^* &= Y_{it} - \hat{\theta}_i \bar{Y}_i \\ X_{it}^* &= X_{it} - \hat{\theta}_i \bar{X}_i \\ Z_{it} &= [(X_{1it} - \bar{X}_{1i})', (X_{2it} - \bar{X}_{2i})', Z'_{1i}, \bar{X}'_{1i}]' \\ X_{it} &= [X'_{1it}, X'_{2it}, Z'_{1i}, Z'_{2i}]' \end{aligned}$$

通过工具变量  $Z_{it}$ , 实施  $Y_{it}^*$  对  $X_{it}^*$  的 2SLS 回归, 即可获得 Hausman-Taylor 估计量  $\hat{\beta}_{\text{HT}}$ . 最后定义全样本数据矩阵  $Y^*, X^*$ , 以及  $Z$ , 那么

$$\hat{\beta}_{\text{HT}} = [X^{*'} Z (Z' Z)^{-1} Z' X^*]^{-1} X^{*'} Z (Z' Z)^{-1} Z' Y^*$$

可以证明, 当模型恰好识别的时候, HT 估计量等同于上述提到的 2SLS 估计量. 而当模型过度识别的时候, 如果更强的假设条件成立, 那么 HT 估计量比 2SLS 估计量更加渐近有效, 而 Amemiya and MaCurdy (1986) 则将 HT 估计量的有效性进一步完善了.

## 8.11 动态面板数据模型

之前考虑的面板数据模型均为静态模型, 也即回归元中没有包括  $Y_{it}$  的滞后项, 然而在许多经济模型中, 当期的决策取决于过去的决策.

面板数据框架下的动态模型由  $\text{AR}(p)$  过程和带有单向误差结构的回归元构成, 也即

$$Y_{it} = \alpha_1 Y_{i,t-1} + \cdots + \alpha_p Y_{i,t-p} + X'_{it} \beta + u_i + \varepsilon_{it} \quad (8.44)$$

其中  $\alpha_j$  为自回归系数,  $X_{it}$  为  $K \times 1$  维回归向量,  $u_i$  为个体效应, 而  $\varepsilon_{it}$  为特质误差. 假定  $u_i$  和  $\varepsilon_{it}$  相互独立,  $\{\varepsilon_{it}\}$  无序列相关且是零均值的, 并且满足  $\mathbb{E}[X_{is} \varepsilon_{it}] = 0$ .

### 8.11.1 FE 估计的偏误

对于一个  $\text{AR}(1)$  过程

$$Y_{it} = \alpha Y_{i,t-1} + u_i + \varepsilon_{it} \quad (8.45)$$

对方程 (8.45) 使用组内变换可得

$$\dot{Y}_{it} = \alpha \dot{Y}_{i,t-1} + \dot{\varepsilon}_{it}, \quad t \geq 2$$

于是个体效应  $u_i$  被消去, 问题的难点在于  $\mathbb{E}[\dot{Y}_{i,t-1} \dot{\varepsilon}_{it}] \neq 0$ .

为了看清 FE 估计的偏误, 考虑一个简单的例子. 假设  $T = 3$ , 于是每个个体  $i$  有两个可观测序对  $[Y_{it}, Y_{i,t-1}]$ , 于是组内估计量为差分估计量. 当  $t = 3$  时, 对 (8.45) 使用差分算子得到

$$\Delta Y_{i3} = \alpha \Delta Y_{i2} + \Delta \varepsilon_{i3} \quad (8.46)$$

对上式使用最小二乘可知

$$\hat{\alpha}_{FE} = \left( \sum_{i=1}^N \Delta Y_{i2}^2 \right)^{-1} \left( \sum_{i=1}^N \Delta Y_{i2} \Delta Y_{i3} \right) = \alpha + \left( \sum_{i=1}^N \Delta Y_{i2}^2 \right)^{-1} \left( \sum_{i=1}^N \Delta Y_{i2} \Delta \varepsilon_{i3} \right)$$

注意到

$$\begin{aligned} \mathbb{E}[\Delta Y_{i2} \Delta \varepsilon_{i3}] &= \mathbb{E}[(Y_{i2} - Y_{i1})(\varepsilon_{i3} - \varepsilon_{i2})] \\ &= \mathbb{E}[Y_{i2} \varepsilon_{i3}] - \mathbb{E}[Y_{i1} \varepsilon_{i3}] - \mathbb{E}[Y_{i2} \varepsilon_{i2}] + \mathbb{E}[Y_{i1} \varepsilon_{i2}] = -\sigma_\varepsilon^2 \end{aligned}$$

当  $|\alpha| < 1$  时, 根据 AR(1) 的方差公式可知

$$\mathbb{E}[(\Delta Y_{i2})^2] = \frac{2\sigma_\varepsilon^2}{1-\alpha^2} - \frac{2\alpha\sigma_\varepsilon^2}{1-\alpha^2} = \frac{2\sigma_\varepsilon^2}{1+\alpha}$$

于是

$$\text{plim}(\hat{\alpha}_{FE} - \alpha) = \frac{\mathbb{E}[\Delta Y_{i2} \Delta \varepsilon_{i3}]}{\mathbb{E}[(\Delta Y_{i2})^2]} = -\frac{1+\alpha}{2}$$

当  $T > 3$  时, 根据 Nickell (1981) 的结论可以得出, 当  $|\alpha| < 1$  时, 固定效应估计量的概率极限为

$$\text{plim}(\hat{\alpha}_{FE} - \alpha) = \frac{1+\alpha}{\frac{2\alpha}{1-\alpha} - \frac{T-1}{1-\alpha^{T-1}}}$$

例如当  $\alpha = 0.5$ ,  $T = 30$  时, 偏误大概为  $-0.056$ .

根据以上分析, 当回归元包括  $Y_{it}$  的滞后项时, FE 估计量不是一致的, 即使时间维度  $T$  非常大.

### 8.11.2 Anderson-Hsiao 估计量

Anderson-Hsiao (1982) 取得了重要突破, 证明了一个简单的工具变量估计量在模型 (8.44) 中是一致的. 首先对 (8.44) 中  $t \geq p+1$  的部分做一阶差分

$$\Delta Y_{it} = \alpha_1 \Delta Y_{i,t-1} + \alpha_2 \Delta Y_{i,t-2} + \cdots + \alpha_p \Delta Y_{i,t-p} + \Delta X'_{it} \beta + \Delta \varepsilon_{it} \quad (8.47)$$

同样有

$$\mathbb{E}[\Delta Y_{i,t-1} \Delta \varepsilon_{it}] = -\sigma_\varepsilon^2$$

但当  $s > 1$  时,  $\mathbb{E}[\Delta Y_{i,t-s} \Delta \varepsilon_{it}] = 0$ , 并且  $\mathbb{E}[\Delta X_{it} \Delta \varepsilon_{it}] = 0$  的严格外生也成立.

由于  $\Delta Y_{i,t-1}$  与  $\Delta \varepsilon_{it}$  之间的相关性导致了内生性问题, 一种解决方法是使用工具变量. Anderson and Hsiao (1982) 指出  $Y_{i,t-2}$  是一个有效工具, 这是因为在  $\{\varepsilon_{it}\}$  序列无关的条件下有

$$\mathbb{E}[Y_{i,t-2} \Delta \varepsilon_{it}] = \mathbb{E}[Y_{i,t-2} \varepsilon_{it}] - \mathbb{E}[Y_{i,t-2} \varepsilon_{i,t-1}] = 0 \quad (8.48)$$

使用  $Y_{i,t-2}$  作为  $\Delta Y_{i,t-1}$  的工具, 由此得到的 IV 估计量即为 Anderson-Hsiao 估计量. 最终, 使用  $[Y_{i,t-2}, \cdots, Y_{i,t-p-1}]$  作为  $[\Delta Y_{i,t-1}, \cdots, \Delta Y_{i,t-p}]$  的工具, 就能一致估计出  $[\alpha_1, \cdots, \alpha_p]$ , 这要

求  $T \geq p + 2$ . 为了看到这一点, 我们同样假定  $T = 3, p = 1$ , 并且没有其它回归元  $X_{it}$ , 此时 Anderson-Hsiao IV 估计量为

$$\hat{\alpha}_{IV} = \left( \sum_{i=1}^N Y_{i1} \Delta Y_{i2} \right)^{-1} \left( \sum_{i=1}^N Y_{i1} \Delta Y_{i3} \right) = \alpha + \left( \sum_{i=1}^N Y_{i1} \Delta Y_{i2} \right)^{-1} \left( \sum_{i=1}^N Y_{i1} \Delta \varepsilon_{i3} \right)$$

如果  $\{\varepsilon_{it}\}$  是序列无关的, 那么 (8.48) 成立, 通常而言  $\mathbb{E}[Y_{i1} \Delta Y_{i2}] \neq 0$ , 于是当  $N \rightarrow \infty$  时有

$$\hat{\alpha}_{IV} \xrightarrow{p} \alpha - \frac{\mathbb{E}[Y_{i1} \Delta \varepsilon_{i3}]}{\mathbb{E}[Y_{i1} \Delta Y_{i2}]} = \alpha$$

从而  $\hat{\alpha}_{IV}$  是  $\alpha$  的一致估计量.

简单总结一下, Anderson-Hsiao 估计量的一致性依赖于两个关键假设, 一是误差项  $\{\varepsilon_{it}\}$  序列无关, 也即  $\Delta \varepsilon_{it}$  不存在二阶自相关, 因此工具变量才能与其无关; 二是工具变量与内生变量存在相关, 在上例中就是  $\mathbb{E}[Y_{i1} \Delta Y_{i2}] \neq 0$ .

### 8.11.3 GMM 估计量

正交条件 (8.48) 是动态面板模型所隐含的众多条件之一, 事实上滞后项  $Y_{i,t-2}, Y_{i,t-3}, \dots$  都是  $\Delta Y_{i,t-1}$  的有效工具变量. 当  $T > p + 2$  时, 尽管 Anderson-Hsiao 估计量是一致的, 但它是缺乏效率的 (受时代所限, GMM 估计在 1982 年才提出), 而这些工具变量可以用来提高估计效率, 这项工作主要由 Arellano and Bond (1991) 发展壮大.

首先将差分回归元  $[\Delta Y_{i,t-1}, \dots, \Delta Y_{i,t-p}, \Delta X'_{it}]$  堆叠为  $T \times (p + K)$  维矩阵  $\Delta X_i$ , 系数向量堆叠为  $\alpha$ , 从而 (8.47) 可以写为  $\Delta Y_i = \Delta X_i \alpha + \Delta \varepsilon_i$ , 全样本回归为  $\Delta Y = \Delta X \alpha + \Delta \varepsilon$ . 可以证明, 当  $t \geq p + 2$  时,  $[Y_{i1}, \dots, Y_{i,t-2}, \Delta X_{it}]$  都是有效的工具变量<sup>2</sup>. 定义矩阵

$$Z_i = \begin{bmatrix} [Y_{i1}, \dots, Y_{ip}, \Delta X'_{i,p+2}] & 0 & \dots & 0 \\ 0 & [Y_{i1}, \dots, Y_{i,p+1}, \Delta X'_{i,p+3}] & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & [Y_{i1}, \dots, Y_{i,T-2}, \Delta X'_{i,T}] \end{bmatrix}$$

它的维数为  $(T - p - 1) \times L$ , 其中  $L = K(T - p - 1) + [(T - 2)(T - 1) - (p - 2)(p - 1)]/2$ . 以上工具矩阵包含了所有的滞后项  $Y_{i,t-2}, Y_{i,t-3}, \dots$ <sup>3</sup>.

根据假设, 我们可以得到  $L$  个矩条件

$$\mathbb{E}[Z'(\Delta Y_i - \Delta X_i \alpha)] = 0$$

如果  $T > p + 2$ , 那么  $L > p$ , 此时模型是过度识别的. 再定义  $L \times L$  维协方差矩阵

$$\Omega = \mathbb{E}[Z'_i \Delta \varepsilon_i \Delta \varepsilon'_i Z_i]$$

令  $Z$  是  $Z_i$  堆叠起来的  $(T - p - 1)N \times L$  维矩阵, 那么一个不可行的有效 GMM 估计量为

$$\hat{\alpha}_{GMM} = (\Delta X' Z \Omega^{-1} Z' \Delta X)^{-1} \Delta X' Z \Omega^{-1} Z' \Delta Y$$

<sup>2</sup>例如对于滞后阶数  $p = 2, t = 5$  的模型, 共有  $K + 3$  个工具变量, 它们是  $[Y_{i1}, Y_{i2}, Y_{i3}, \Delta X'_{it}]$ .

<sup>3</sup>严格外生条件  $\mathbb{E}[X_{is} \varepsilon_{it}] = 0$  可能限制太强, 一个限制性更少的条件是假定回归元都是前定的 (predetermined), 也即对于一切  $s \geq 0$ , 都有  $\mathbb{E}[X_{i,t-s} \varepsilon_{it}] = 0$ , 此时  $X_{it}$  可以与  $\varepsilon_{it}$  的滞后项相关, 由此  $\mathbb{E}[\Delta X_{it} \Delta \varepsilon_{it}] \neq 0$ , 故而应该使用工具  $[X_{i1}, X_{i2}, \dots, X_{i,t-1}]$  替换  $\Delta X_{it}$  并纳入到  $Z_i$  中.



如果  $\{\varepsilon_{it}\}$  是条件同方差的且无序列相关, 那么

$$\Omega = \sigma_\varepsilon^2 \mathbb{E}[Z_i' H Z_i]$$

这里  $H = D_i D_i'$ . 此时

$$\hat{\Omega}_1 = N^{-1} \sum_{i=1}^N Z_i' H Z_i$$

从而可以得到可行的渐近有效 GMM 估计量

$$\hat{\alpha}_1 = (\Delta X' Z \hat{\Omega}_1^{-1} Z' \Delta X)^{-1} \Delta X' Z \hat{\Omega}_1^{-1} Z' \Delta Y$$

称  $\hat{\alpha}_1$  为一步 Arellano-Bond GMM 估计量. 此时  $\hat{\alpha}_1$  的经典协方差矩阵估计量为

$$\hat{V}_1^0 = \hat{\sigma}_\varepsilon^2 (\Delta X' Z \hat{\Omega}_1^{-1} Z' \Delta X)^{-1}$$

这里的  $\hat{\sigma}_\varepsilon^2$  为一步 GMM 残差  $\hat{\varepsilon}_i = \Delta Y_i - \Delta X_i \hat{\alpha}_1$  的样本方差, 而聚类稳健的协方差矩阵估计量为

$$\hat{V}_1 = (\Delta X' Z \hat{\Omega}_1^{-1} Z' \Delta X)^{-1} (\Delta X' Z \hat{\Omega}_1^{-1} Z' \hat{\Omega}_2^{-1} Z \hat{\Omega}_1^{-1} Z' \Delta X) (\Delta X' Z \hat{\Omega}_1^{-1} Z' \Delta X)^{-1}$$

其中

$$\hat{\Omega}_2 = N^{-1} \sum_{i=1}^N Z_i' \hat{\varepsilon}_i \hat{\varepsilon}_i' Z_i$$

是使用一步 GMM 残差构造的  $\Omega$  的聚类稳健估计量. 由此还能得到允许  $\varepsilon_{it}$  存在异方差的两步 Arellano-Bond GMM 估计量

$$\hat{\alpha}_2 = (\Delta X' Z \hat{\Omega}_2^{-1} Z' \Delta X)^{-1} \Delta X' Z \hat{\Omega}_2^{-1} Z' \Delta Y$$

它的聚类稳健估计量为

$$\hat{V}_2 = (\Delta X' Z \hat{\Omega}_2^{-1} Z' \Delta X)^{-1} (\Delta X' Z \hat{\Omega}_2^{-1} Z' \hat{\Omega}_3^{-1} Z \hat{\Omega}_2^{-1} Z' \Delta X) (\Delta X' Z \hat{\Omega}_2^{-1} Z' \Delta X)^{-1}$$

其中

$$\hat{\Omega}_3 = N^{-1} \sum_{i=1}^N Z_i' \tilde{\varepsilon}_i \tilde{\varepsilon}_i' Z_i$$

这里的  $\tilde{\varepsilon}_i = \Delta Y_i - \Delta X_i \hat{\alpha}_2$  为两步 GMM 残差. GMM 估计量可以这样被不断迭代, 直至收敛到一个迭代 GMM 估计量.

Arellano-Bond 估计量相对于 Anderson-Hsiao 估计量的优点是, 当  $T > p + 2$  时, 额外的矩条件可以用来降低估计量的渐近方差. 然而其缺点为当  $T$  很大时, 容易导致弱工具变量问题, 因此有必要限制作为工具变量的滞后项的数量.

对于一步 Arellano-Bond 估计量, 它的权重矩阵  $\hat{\Omega}_1$  并不依赖于 GMM 残差, 也就是说它的随机性要弱于二步 Arellano-Bond 估计量中的权重矩阵  $\hat{\Omega}_2$ , 因此在较小的样本中, 特别是当误差项满足条件同方差时, 一步估计量有更好的性质. 相对地, 在大样本和误差项存在异方差的情形下, 两步估计量具有更好的渐近有效性.

简单总结一下, Arellano-Bond 估计量是通过将 GMM 应用到差分方程 (8.47) 中, 并且选取可用的滞后项  $Y_{i,t-2}, Y_{i,t-3}, \dots$  作为工具变量来获得的一致估计量, 以上方法也叫做差分 GMM



(difference GMM). 尽管有了许多提高, 然而 Blundell and Bond (1998) 指出 Anderson-Hsiao 估计量与 Arellano-Bond 估计量都会面临弱工具变量问题.

考虑 Anerson-Hsiao 估计量的情形, 如果使用  $Y_{i,t-2}$  作为  $\Delta Y_{i,t-1}$  的工具, 我们可以写出简约式方程

$$\Delta Y_{i,t-1} = Y_{i,t-2}\gamma + v_{it}$$

其中简约系数  $\gamma$  可以由线性投影定义. 利用方程  $\Delta Y_{i,t-1} = (\alpha - 1)Y_{i,t-2} + u_i + \varepsilon_{i,t-1}$  以及  $\mathbb{E}[Y_{i,t-2}\varepsilon_{i,t-1}] = 0$ , 从而

$$\gamma = \frac{\mathbb{E}[Y_{i,t-2}\Delta Y_{i,t-1}]}{\mathbb{E}[Y_{i,t-2}^2]} = (\alpha - 1) + \frac{\mathbb{E}[Y_{i,t-2}u_i]}{\mathbb{E}[Y_{i,t-2}^2]}$$

Blundell and Bond (1998) 进一步证明了

$$\gamma = \frac{k(\alpha - 1)}{k + \sigma_u^2/\sigma_\varepsilon^2}$$

其中  $k = (1 - \alpha)/(1 + \alpha)$ . 显然, 当  $\gamma$  接近于 0 时,  $Y_{i,t-2}$  是一个弱工具变量, 此时  $\alpha$  接近于 1, 也即模型是一个随机游走.

对于弱工具变量问题, Arellano and Bover (1995) 以及 Blundell and Bond (1998) 提出了解决办法. 回到之前的水平方程 (8.44) 中, 如果  $\{\varepsilon_{it}\}$  不存在序列相关, 并且  $[\Delta Y_{i,t-1}, \Delta Y_{i,t-2}, \dots]$  与个体效应  $u_i$  不相关, 那么可以使用  $[\Delta Y_{i,t-1}, \Delta Y_{i,t-2}, \dots]$  作为工具变量对模型 (8.44) 进行 GMM 估计, 称为水平 GMM.

进一步, Blundell-Bond 估计量将差分 GMM 和水平 GMM 结合在一起, 在水平 GMM 的假设条件下, 可以得到更有效率的系统 GMM 估计量, 并且一定程度缓解弱工具变量问题.

## 第 9 章 极大似然估计

之前我们考察的全部模型, 没有对以外生变量为条件的内生变量的分布做出任何假设就能通过外生性做出一致估计, 倘若我们给定了内生变量的分布信息, 那么极大似然估计 (Maximum Likelihood Estimation, MLE) 将特别有用, 特别是在许多 (单非全部) 一致且渐近正态的估计量族中, ML 估计量都是渐近有效的, GMM 估计量就是这样的估计量族中的一员.

### 9.1 预备内容

关于  $Y$  的参数模型 (parametric model) 是  $X$  的一个依赖于未知参数  $\theta \in \Theta$  的概率函数, 它表明了  $Y$  的总体分布是某个具体的分布族中的一员. 举例而言, 一个参数模型是  $Y \sim N(\mu, \sigma^2)$ , 它的条件概率密度为

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

其中参数  $\mu \in \mathbb{R}$ , 并且  $\sigma^2 > 0$ . 这个模型刻画了  $Y$  服从由  $\mu$  和  $\sigma^2$  确定的正态分布.

在初等统计学中, 经典的 MLE 正是从对独立同分布的  $\{Y_i\}_{i=1}^n$  的密度族的设定出发, 根据从总体中得到的随机抽样, 就可以构造似然函数和得出一阶条件进行参数估计. 然而在几乎所有的经济应用中, 我们总是对估计条件分布的参数感兴趣.

假定从总体中得到的随机抽样被分割为  $(Y_i, X_i)$ , 其中  $y_i \in \mathcal{Y}$  而  $x_i \in \mathcal{X}$ , 我们对给定  $X_i$  时的  $Y_i$  的条件分布感兴趣, 并且对  $X_i$  的分布不感兴趣, 无需对  $X_i$  的分布做任何具体设定. 从这个角度看, 本章所使用的方法称为条件极大似然估计 (Conditional Maximum Likelihood Estimation, CMLE), 在不引起混淆的情况下简称为 MLE. 通过对  $X_i$  取 0, 就可以得到无条件 MLE 的特殊情形.

为了实施极大似然分析, 我们需要对基本的结构模型, 即给定  $X_i$  时的  $Y_i$  的密度进行设定或推导. 通常而言, 我们假定密度函数具有有限个未知的参数, 因此我们可以得到一个关于  $Y_i$  的参数模型,  $Y_i$  既可以是向量也可以是标量.

为了阐明极大似然的一般理论, 下面给出二值响应模型 (binary response model) 的例子, 它的一个重要分支是概率单位模型 (probit model).

**例 9.1** 假定随机变量  $Y_i^*$  服从

$$Y_i^* = X_i' \theta + e_i$$

其中  $X_i$  和  $\theta$  均为  $K \times 1$  维向量,  $e_i$  与  $X_i$  独立, 并且  $e_i \sim N(0, 1)$ . 我们无法观测到  $Y_i^*$ , 而只能观测到指示  $Y_i^*$  符号的二值变量

$$Y_i = \mathbb{1}[Y_i^* > 0]$$

因此可以写出给定  $X_i$  时  $Y_i$  的分布

$$\begin{aligned}\mathbb{P}[Y_i = 1|X_i] &= \mathbb{P}[Y_i^* > 0|X_i] = \mathbb{P}[X_i'\theta + e_i|X_i] \\ &= \mathbb{P}[e_i > -X_i'\theta|X_i] = 1 - \Phi(-X_i'\theta) = \Phi(X_i'\theta)\end{aligned}\quad (9.1)$$

其中  $\Phi(\cdot)$  为标准正态的累积分布函数. 类似可以得出

$$\mathbb{P}[Y_i = 0|X_i] = 1 - \Phi(X_i'\theta) \quad (9.2)$$

将 (9.1) 和 (9.2) 结合就能得到给定  $X_i$  时  $Y_i$  的条件密度

$$f(y|X_i) = [\Phi(X_i'\theta)]^y [1 - \Phi(X_i'\theta)]^{1-y}$$

显然当  $y \notin \{0, 1\}$  时,  $f(y|X_i) = 0$ .

## 9.2 CMLE 的一般框架

设  $\mathcal{X} \subset \mathbb{R}^K$  和  $\mathcal{Y} \subset \mathbb{R}^G$  分别是随机向量  $X$  和  $Y$  的支集, 再定义给定  $X$  时  $Y$  的条件分布为  $D(Y|X)$ , 对于每个  $X$  而言, 这一分布表示一个概率测度, 并且完整地描述了  $X$  取到某个特殊值时, 随机向量  $Y$  的行为, 这一分布几乎总是由条件密度来描述.

于是我们定义  $p_o(y|X)$  为给定  $X$  时  $Y$  的条件密度, 下标  $o$  表示它为真实的条件密度, 而非某个可能的条件密度, 并且对于所有  $x \in \mathcal{X}$ ,  $p_o(\cdot|X)$  为关于  $\sigma$ -有限测度  $\nu^1$  的密度. 如果  $D(Y|X)$  是离散的, 那么  $\nu$  为计数测度且积分转化为求和; 如果  $D(Y|X)$  是绝对连续的, 那么  $\nu$  为 Lebesgue 测度. 换言之, 如果  $Y_i$  是离散的, 那么  $\nu(dy)$  把积分转换为求和; 如果  $Y_i$  是连续的, 那么就得到了通常的 Lebesgue 积分.

在正式阐述 MLE 的原理前, 还需要用到统计推断领域中极为重要的 Kullback-Leibler 信息准则 (Kullback-Leibler Information Criterion, KLIC). 假定  $f$  和  $g$  是  $\mathbb{R}^M$  上的非负  $\nu$ -可测函数, 定义  $\mathcal{S}_f = \{z \in \mathbb{R}^M : f(z) > 0\}$ , 并且

$$1 = \int_{\mathcal{S}_f} f(z) \nu(dz) \geq \int_{\mathcal{S}_f} g(z) \nu(dz) \quad (9.3)$$

式 (9.3) 中的等式表明  $f$  是  $\mathbb{R}^M$  上的密度, 当  $g$  也是  $\mathbb{R}^M$  上的密度时, 上述不等式成立. 由此产生了一个重要结果

$$\mathcal{J}(f; g) = \int_{\mathcal{S}_f} \log \left[ \frac{f(z)}{g(z)} \right] f(z) \nu(dz) \geq 0 \quad (9.4)$$

其中  $\mathcal{J}(f; g)$  的数量大小即为 KLIC. 式 (9.4) 的另一种表述为

$$\mathbb{E}[\log\{f(Z)\}] \geq \mathbb{E}[\log\{g(Z)\}]$$

CMLE 需要用到条件版本的 (9.4). 设  $\mathcal{Y}(X) = \{y : p(y|X) > 0\}$  为  $Y$  的条件支集,  $\nu$  是不依赖于  $X$  的  $\sigma$ -有限测度, 那么对于任意  $g(\cdot|X) > 0$  都有

$$\mathcal{J}_X(p; g) = \int_{\mathcal{Y}(X)} \log \left[ \frac{p(y|X)}{g(y|X)} \right] p(y|X) \nu(dy) \geq 0$$

<sup>1</sup> 设  $(\Omega, \mathcal{F}, \mu)$  为一个测度空间, 如果存在可数集族  $A_1, A_2, \dots \in \mathcal{F}$ , 使得  $\bigcup_{n=1}^{\infty} A_n = \Omega$ , 并且对一切  $n \geq 1$  都有  $\mu(A_n) < \infty$ , 那么  $\mu$  是  $\sigma$ -有限测度.

上式也可以表述为

$$\mathbb{E}[\log\{p(Y|X)\}|X] \geq \mathbb{E}[\log\{g(Y|X)\}|X]$$

现在任意选取一个非负  $\nu$ -可测函数  $f(\cdot|X)$ , 使得

$$\int_{\mathcal{Y}} f(y|X) \nu(dy) = 1, \quad \forall x \in \mathcal{X} \quad (9.5)$$

根据条件 **KLIC** 可知, 对于任意  $x \in \mathcal{X}$  都有

$$\mathcal{K}(f; X) \equiv \int_{\mathcal{Y}} \log \left[ \frac{p_o(y|X)}{f(y|X)} \right] p_o(y|X) \nu(dy) \geq 0 \quad (9.6)$$

故而当  $f = p_o$  时, 积分恒为 0, 也即 (9.6) 意味着  $\mathcal{K}(f; X)$  在  $f = p_o$  处最小化.

现在将 (9.5) 应用到关于  $p_o(\cdot|x)$  的参数模型上, 即

$$\{f(\cdot|X; \theta) : \theta \in \Theta \subset \mathbb{R}^P\} \quad (9.7)$$

假定对于一切  $x \in \mathcal{X}$  和  $\theta \in \Theta$ ,  $f(\cdot|X; \theta)$  都满足条件 (9.5), 我们称条件密度模型 (9.7) 是正确设定的, 如果存在某个  $\theta_o \in \Theta^2$ , 使得对任意  $x \in \mathcal{X}$  都有

$$f(\cdot|X; \theta_o) = p_o(\cdot|X)$$

对于每个  $x \in \mathcal{X}$ ,  $\mathcal{K}(f; X)$  可以写为

$$\mathbb{E}[\log\{p_o(Y_i|X_i)\}|X_i = X] - \mathbb{E}[\log\{f(Y_i|X_i)\}|X_i = X]$$

因此如果条件密度模型被正确设定, 那么

$$\mathbb{E}[l_i(\theta_o)|X_i] \geq \mathbb{E}[l_i(\theta)|X_i], \quad \theta \in \Theta \quad (9.8)$$

其中

$$l_i(\theta) = l(Y_i, X_i, \theta) = \log f(Y_i|X_i; \theta)$$

表示关于观测值  $i$  的条件对数似然 (conditional log-likelihood), 它是  $\theta$  的一个随机函数. 例如在 Probit 模型中, 观测值  $i$  的对数似然为

$$l_i(\theta) = Y_i \log \Phi(X_i' \theta) + (1 - Y_i) \log [1 - \Phi(X_i' \theta)]$$

在不等式 (9.8) 两端取期望并使用 LIE 可知

$$\theta_o = \arg \max_{\theta \in \Theta} \mathbb{E}[l_i(\theta)] \quad (9.9)$$

使用样本矩替代总体矩, 倘若

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \log f(Y_i|X_i; \theta)$$


存在, 那么称  $\hat{\theta}_{\text{ML}}$  为  $\theta_o$  的条件极大似然估计量. 假如把  $X_i$  处理为常量, 那么就得到经典框架下的 ML 估计量

$$\hat{\theta}_{\text{ML}} = \max_{\theta \in \Theta} \prod_{i=1}^n f(Y_i|X_i; \theta)$$

而 ML 估计量的存在性类似于“紧集上的连续函数存在最值”, 可由以下定理保证.

<sup>2</sup>相对应地, 如果这样的  $\theta_o \in \Theta$  不存在, 则说明模型设定错误.

**定理 9.1**

设  $\Theta \subset \mathbb{R}^P$  为紧集, 假设对于一切  $\theta \in \Theta$ ,  $f(\cdot; \theta)$  都是  $(Y_i, X_i) \in \mathcal{Y} \times \mathcal{X}$  的非负可测函数, 并且对于每个  $i$ , 条件密度  $f(Y_i|X_i; \cdot)$  关于  $\theta \in \Theta$  是连续的, 那么 ML 估计量存在. 

一旦我们得到了 MLE 的存在性, 就可以继续讨论 MLE 的不变性 (invariance). 假设  $\hat{\theta}$  是  $\theta_o \in \Theta$  的 ML 估计量, 我们通过一个定义在  $\Theta$  上的函数  $\tau : \Theta \rightarrow \Lambda$  对模型进行再参数化 (reparametrization), 这里的  $\Lambda$  为值域

$$\Lambda = \{\lambda : \tau(\theta) = \lambda, \theta \in \Theta\}$$

并且规定  $\Lambda$  非空. 如果函数  $\tau : \Theta \rightarrow \Lambda$  为单射, 则它称为再参数化过程. 根据单射的定义, 存在反函数  $\tau^{-1}$ , 使得对于任意  $\theta \in \Theta$  都有  $\tau(\tau^{-1}(\theta)) = \theta$ .

为简单起见, 假设  $Y_i$  是一个随机变量. 定义  $\lambda_o = \tau(\theta_o)$ , 以及  $l(\theta) = \sum_{i=1}^n l_i(\theta)$ , 于是

$$l(\theta) = l(\tau^{-1}(\lambda)) = l^*(\lambda)$$

其中

$$l^*(\lambda) = \sum_{i=1}^n \log f(y_i|X_i; \lambda)$$

由于  $\hat{\theta}$  为  $\theta_o$  的 ML 估计量, 因此对于任意  $\theta \in \Theta$  都有  $l(\hat{\theta}) \geq l(\theta)$ . 定义  $\hat{\lambda} = \tau(\hat{\theta})$ , 对于任意  $\theta \in \Theta$ , 可以得到

$$l(\hat{\theta}) = l(\tau^{-1}(\hat{\lambda})) = l^*(\hat{\lambda}) \geq l(\theta) = l(\tau^{-1}(\lambda))$$

因此对于任意  $\lambda \in \Lambda$  都有

$$l^*(\hat{\lambda}) \geq l^*(\lambda)$$

从而  $\hat{\lambda}$  是  $\lambda_o \in \Lambda$  的 ML 估计量.

**注** 上述证明过程要求  $\tau$  为单射, 这是一个限制性很强的条件. Zehna (1966) 去掉了这一限制, 因此  $\tau : \Theta \rightarrow \Lambda$  可以为任意可测函数, 但是  $\hat{\lambda}$  不再是通常意义下的 ML 估计量.

Zehna (1966) 定义了集合  $\Theta_\lambda = \{\theta : \tau(\theta) = \lambda\}$ , 以及  $M(\lambda) = \sup_{\theta \in \Theta_\lambda} l(\theta)$ , 称  $M(\lambda)$  为由  $\tau$  诱导的似然函数. 此时

$$M(\lambda) = \sup_{\theta \in \Theta_\lambda} l(\theta) \leq \sup_{\theta \in \Theta} l(\theta) = l(\hat{\theta}) = M(\hat{\lambda})$$

对一切  $\lambda \in \Lambda$  成立. 因此  $\hat{\lambda} = \tau(\hat{\theta})$  最大化了由  $\tau$  诱导的似然函数  $M(\lambda)$ . 然而,  $M(\lambda)$  在一般情况下似乎不是任何随机变量的似然函数.

## 9.3 CMLE 的渐近性质

### 9.3.1 一致性

在正式讨论 ML 估计量的渐近性质前, 我们先给出一系列正则假设, 最主要的还是需要用到 UWLLN 和极值估计量的一致性引理.

**假设 9.1**

- (1)  $\{(X_i, Y_i)\}_{i=1}^n$  是一个可观测的随机样本,  $x_i \in \mathcal{X} \subset \mathbb{R}^K$ , 并且  $y_i \in \mathcal{Y} \subset \mathbb{R}^G$ .
- (2)  $P \times 1$  维参数空间是紧集.
- (3) 对于任意  $x \in \mathcal{X}$  和  $\theta \in \Theta$ ,  $f(\cdot|X, \theta)$  是关于  $\sigma$ -有限测度  $\nu(d\mathbf{y})$  的真实密度.
- (4) 存在唯一的  $\theta_o \in \Theta$ , 使得对于一切  $x \in \mathcal{X}$  都有  $p_o(\cdot|X) = f(\cdot|X; \theta_o)$ .
- (5) 对于任意  $\theta \in \Theta$ , 对数似然  $l(\cdot, \theta)$  是  $\mathcal{Y} \times \mathcal{X}$  上的 Borel 可测函数.
- (6) 对于任意  $(y, x) \in \mathcal{Y} \times \mathcal{X}$ , 对数似然  $l(y, x, \cdot)$  是  $\Theta$  上的连续函数.
- (7)  $\mathbb{E}[\sup_{\theta \in \Theta} |l_i(\theta)|] < \infty$ .

在假设 9.1 中, 尽管  $\Theta$  可以不是紧的, 但这就需要做出更多的讨论. 可测性则是技术性假定, 通常无需对其检验. 关键性假设仍是  $\theta_o$  可识别, 占优条件成立, 以及对数似然函数在  $\theta \in \Theta$  上连续.

**定理 9.2**

在假设 9.1 下, 当  $n \rightarrow \infty$  时有  $\hat{\theta}_{ML} \xrightarrow{P} \theta_o$ .

**证明** 定义  $\hat{Q}(\theta) = n^{-1} \sum_{i=1}^n l_i(\theta)$ , 以及  $Q(\theta) = \mathbb{E}[l_i(\theta)]$ , 根据之前证明 GMM 估计量一致性的做法, 由 UWLLN 可以证明  $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$ , 由于真实参数  $\theta_o$  可识别, 最后根据极值估计量一致性引理即可证得结论.

我们将  $\theta_o$  设置为一个有限的向量, 因此可以将  $\Theta$  设置为一个包含  $\theta_o$  的有界闭集, 在 Euclid 空间的意义上, 有界闭集等价于紧集.

对于 Probit 模型, 它关于观测值  $i$  的条件对数似然为

$$l_i(\theta) = Y_i \log \Phi(X_i' \theta) + (1 - Y_i) \log [1 - \Phi(X_i' \theta)]$$

如果  $\mathbb{E}[X_i X_i'] > 0$ , 那么对于任意  $\theta \neq \theta_o$ ,  $X' \theta \neq X' \theta_o$ , 这是因为

$$\mathbb{E}[(X_i' \theta - X_i' \theta_o)^2] = (\theta - \theta_o)' \mathbb{E}[X_i X_i'] (\theta - \theta_o) > 0$$

又由于  $\Phi(\cdot)$  是严格单调函数, 故而当  $\mathbb{E}[X_i X_i']$  正定时, 只有唯一的  $\theta_o \in \Theta$ , 使得  $p_o(\cdot|X_i) = f(\cdot|X_i; \theta_o)$ . 注意到对于任意  $v \in \mathbb{R}$ , 都有不等式

$$|\log \Phi(v)| \leq |\log \Phi(0)| + |v| + |v|^2$$

从而

$$\begin{aligned} |l_i(\theta)| &\leq |Y_i| |\log \Phi(X_i' \theta)| + |1 - Y_i| |\log \Phi(-X_i' \theta)| \\ &\leq |\log \Phi(X_i' \theta)| + |\log \Phi(-X_i' \theta)| \\ &\leq 2[|\log \Phi(0)| + \|X_i\| \cdot \|\theta\| + \|X_i\|^2 \|\theta\|^2] \end{aligned}$$

只要  $\mathbb{E}[X_i X_i']$  的非奇异性意味着  $\mathbb{E}\|X_i\|^2 < \infty$ , 于是  $\mathbb{E}[\sup_{\theta \in \Theta} |l_i(\theta)|] < \infty$  也成立. 因此只要样本  $\{(Y_i, X_i)\}_{i=1}^n$  是 i.i.d. 的, 并且  $\mathbb{E}[X_i X_i']$  正定且有限, 则 Probit 模型的 ML 估计量是一致的.

### 9.3.2 得分函数与条件信息矩阵

首先我们给出一个技术性假设, 也即真实参数  $\theta_o$  位于  $\Theta$  内部, 并且现在对每个观测值  $i$ , 定义一个  $P \times 1$  维向量

$$s_i(\theta) = \nabla_{\theta} l_i(\theta)$$

称为对数似然得分 (score of log-likelihood). 如果条件密度模型  $f(\cdot|X_i; \theta)$  是正确识别的, 那么在  $\theta = \theta_o$  处, 得分函数具有重要的零条件均值性质, 也即

$$\mathbb{E}[s_i(\theta_o)|X_i] = 0 \quad (9.10)$$

换言之, 当我们在  $\theta_o$  处计算  $P \times 1$  维得分时, 并且关于  $f(\cdot|X_i; \theta_o)$  取期望, 则期望值为 0. 根据 LIE 可得  $\mathbb{E}[s_i(\theta_o)] = 0$ , 因此 ML 估计量可以看作是恰好识别情况下的 GMM 估计量.

为了证明条件 (9.10), 对于任意  $\theta \in \Theta$ , 设  $\mathbb{E}_{\theta}[\cdot|X_i]$  表示相对于密度  $f(\cdot|X_i; \theta)$  的条件期望. 于是根据定义可得

$$\mathbb{E}_{\theta}[s_i(\theta)|X_i] = \int_{\mathcal{Y}} s(y, X_i, \theta) f(y|X_i; \theta) v(dy)$$

在一定条件下, 积分和微分在  $\Theta$  内部可交换顺序, 也即

$$\nabla_{\theta} \left[ \int_{\mathcal{Y}} f(y|X_i; \theta) v(dy) \right] = \int_{\mathcal{Y}} \nabla_{\theta} f(y|X_i; \theta) v(dy) \quad (9.11)$$

其中  $x_i \in \mathcal{X}$ , 并且  $\theta$  在参数空间  $\Theta$  内部. 根据条件 (9.5) 可知

$$\int_{\mathcal{Y}} \nabla_{\theta} f(y|X_i; \theta) v(dy) = 0$$

上式等价于

$$\int_{\mathcal{Y}} [\nabla_{\theta} l(y, X_i, \theta)] f(y|X_i; \theta) v(dy) = 0$$

使用  $\theta_o$  替代  $\theta$  即可推出 (9.10) 成立.

**注** 条件密度模型的正确设定是条件 (9.10) 成立的充分不必要条件, 也即 (9.10) 无法推出模型正确设定, 因为模型误设可能存在于高阶矩中.

#### 定理 9.3

在假设 9.1 下, 如果对于任意  $(y, x) \in \mathcal{Y} \times \mathcal{X}$ , 对数似然  $l(y, x, \cdot)$  关于  $\beta \in \text{int}(\Theta)$  连续可微, 那么  $\mathbb{E}[s_i(\theta_o)|X_i] = 0$ .

另一方面, 假定  $\theta_o \in \text{int}(\Theta)$ , 并且  $l_i(\theta)$  在包含  $\theta_o$  的一个邻域  $\mathcal{N}$  内二阶连续可微. 再设观测值  $i$  的 Hessian 矩阵是  $l_i(\theta)$  的二阶偏导数矩阵

$$H_i(\theta) = \nabla_{\theta} s_i(\theta) = \nabla_{\theta}^2 l_i(\theta)$$

它是  $P \times P$  的对称矩阵. 由于 ML 估计量是最大化问题的解, 因此  $H_i(\theta_o)$  的期望是负定的. 定义矩阵

$$H(\theta) = \mathbb{E}[H_i(\theta)]$$

当真实参数  $\theta_o$  可识别时,  $-H_o = \mathbb{E}[H_i(\theta_o)]$  通常是正定的. 可以证明, 矩阵  $-H_o$  等价于  $\Omega_o =$



$\mathbb{E}[s_i(\theta_o)s_i(\theta_o)']$ , 这里的  $\mathbf{\Omega}_o$  称为 Fisher 条件信息矩阵.

在足够的光滑条件下, 积分和微分可交换顺序

$$\nabla_{\theta} \left[ \int_{\mathcal{Y}} s_i(\theta) f(y|X_i; \theta) v(dy) \right] = \int_{\mathcal{Y}} \nabla_{\theta} [s_i(\theta) f(y|X_i; \theta)] v(dy) \quad (9.12)$$

假定  $\theta$  位于  $\Theta$  内部, 根据这一性质, 对恒等式

$$\int_{\mathcal{Y}} s_i(\theta) f(y|X_i; \theta) v(dy) \equiv \mathbb{E}_{\theta}[s_i(\theta)|X_i] = 0$$

求微分可得

$$-\mathbb{E}_{\theta}[H_i(\theta)|X_i] = \text{var}_{\theta}[s_i(\theta)|X_i]$$

于是在  $\theta_o$  处有

$$-\mathbb{E}[H_i(\theta_o)|X_i] = \mathbb{E}[s_i(\theta_o)s_i(\theta_o)'|X_i]$$

根据 LIE 即可推知  $-\mathbf{H}_o = \mathbf{\Omega}_o$ .

#### 定理 9.4

在假设 9.1 下, 如果对于任意  $(y, x) \in \mathcal{Y} \times \mathcal{X}$ , 对数似然  $l(y, x, \cdot)$  关于  $\beta \in \text{int}(\Theta)$  二阶连续可微, 那么  $\mathbb{E}[s_i(\theta_o)s_i(\theta_o)' + H_i(\theta_o)|X_i] = 0$ .



在 Probit 模型中, 得分函数为

$$\begin{aligned} s_i(\theta) &= Y_i \left[ \frac{\phi(X_i'\theta)}{\Phi(X_i'\theta)} \right] - (1 - Y_i) \left[ \frac{\phi(X_i'\theta)}{1 - \Phi(X_i'\theta)} \right] \\ &= \frac{[Y_i - \Phi(X_i'\theta)]\phi(X_i'\theta)}{\Phi(X_i'\theta)[1 - \Phi(X_i'\theta)]} X_i \end{aligned}$$

其中  $\phi(\cdot)$  是标准正态的概率密度函数. 并且 Hessian 矩阵

$$\begin{aligned} H_i(\theta) &= \left\{ - \left[ \frac{Y_i - \Phi(X_i'\theta)}{\Phi(X_i'\theta)(1 - \Phi(X_i'\theta))} \right] [\phi(X_i'\theta)]^2 \right. \\ &\quad \left. + \left[ \frac{Y_i - \Phi(X_i'\theta)}{\Phi(X_i'\theta)(1 - \Phi(X_i'\theta))} \right] \phi'(X_i'\theta) \right\} X_i X_i' \end{aligned}$$

对于 Probit 模型, 还可以证明

$$-\mathbf{H}_o = \mathbf{\Omega}_o = \mathbb{E}[\lambda(X_i'\theta_o)\lambda(-X_i'\theta_o)X_i X_i']$$

其中

$$\lambda(v) = \frac{\phi(v)}{\Phi(v)}$$

称为逆 Mills 比率 (Inverse Mills Ratio, IMR).

### 9.3.3 渐近正态性

跟之前一样, 为了推导 ML 估计量的渐近分布, 我们还需要在假设 9.1 上的基础上增加新的限制条件.



**假设 9.2**

在假设 9.1 的基础上, 以下额外条件成立:

- (1)  $\theta_o \in \text{int}(\Theta)$ .
- (2) 对于任意  $(y, x) \in \mathcal{Y} \times \mathcal{X}$ , 对数似然函数  $l(y, x, \cdot)$  在包含  $\theta_o$  的某个邻域  $\mathcal{N}$  上二阶连续可微.
- (3) 对于一切  $\theta \in \text{int}(\Theta)$ , 式 (9.11) 和 (9.12) 中的微分与积分可交换顺序.
- (4)  $\mathbb{E}[\sup_{\theta \in \mathcal{N}} \|H_i(\theta)\|] < \infty$ .
- (5)  $\mathbb{E}[H_i(\theta_o)]$  为满秩矩阵.
- (6)  $n^{-\frac{1}{2}} \sum_{i=1}^n s_i(\theta_o) \xrightarrow{d} N(0, \mathbf{\Omega}_o)$ .

**定理 9.5**

在假设 9.2 下, 当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, -\mathbf{H}_o)$$

其中  $\mathbf{H}_o = -\mathbb{E}[H_i(\theta_o)]$ .

**证明** 之前已经证明了当  $n \rightarrow \infty$  时有  $\hat{\theta} \xrightarrow{p} \theta_o$ , 因为  $\theta_o \in \text{int}(\Theta)$ , 故而当  $n$  充分大时, ML 估计量  $\hat{\theta}$  也是  $\Theta$  的内点, 且最大化对数似然函数  $n^{-1} \sum_{i=1}^n \log f(Y_i|X_i; \theta)$  的 FOC 为

$$\hat{s}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \nabla_{\theta} l_i(\hat{\theta}) = n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) = 0 \quad (9.13)$$

将  $\hat{s}(\hat{\theta})$  在  $\theta_o$  处一阶 Taylor 展开可得

$$\sqrt{n}\hat{s}(\theta_o) + \hat{H}(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta_o) = 0$$

其中  $\bar{\theta}$  在  $\hat{\theta}$  和  $\theta_o$  之间, 并且由  $\hat{\theta} \xrightarrow{p} \theta_o$  可知  $\bar{\theta} - \theta_o \xrightarrow{p} 0$ . 现在定义 Hessian 矩阵

$$\hat{H}(\theta) = n^{-1} \sum_{i=1}^n \nabla_{\theta}^2 l_i(\theta) = n^{-1} \sum_{i=1}^n H_i(\theta)$$

于是

$$\begin{aligned} \|\hat{H}(\bar{\theta}) - \mathbf{H}_o\| &= \|\hat{H}(\bar{\theta}) - H(\bar{\theta}) + H(\bar{\theta}) - H(\theta_o)\| \\ &\leq \sup_{\theta \in \Theta} \|\hat{H}(\bar{\theta}) - H(\bar{\theta})\| + \|H(\bar{\theta}) - H(\theta_o)\| \xrightarrow{p} 0 \end{aligned}$$

其中 UWLLN 保证第一项趋于 0, 而  $\bar{\theta} \xrightarrow{p} \theta_o$  及  $H(\cdot)$  的连续性保证了第二项趋于 0.

由于  $\mathbf{H}_o = H(\theta_o)$  是非奇异的, 故而当  $n$  充分大时, 矩阵  $\hat{H}(\bar{\theta})$  也是非奇异的. 根据极大似然的 FOC 可知

$$\sqrt{n}(\hat{\theta} - \theta_o) = -\hat{H}^{-1}(\bar{\theta})\sqrt{n}\hat{s}(\theta_o)$$

根据假设 9.2(6) 又可知

$$\sqrt{n}\hat{s}(\theta_o) \xrightarrow{d} N(0, \mathbf{\Omega}_o)$$

其中  $\mathbf{\Omega}_o = \mathbb{E}[s_i(\theta_o)s_i(\theta_o)']$ . 由 Slutsky 定理推得

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, \mathbf{H}_o^{-1} \mathbf{\Omega}_o \mathbf{H}_o^{-1})$$

因为模型是正确设定的, 故而  $-\mathbf{H}_o = \mathbf{\Omega}_o$ , 因此  $\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, -\mathbf{H}_o)$ .

下一节将证明, 在条件密度模型正确设定的情况下, 在很大一类估计量中,  $\hat{\theta}_{\text{ML}}$  都是  $\theta_o$  的渐近有效一致估计量.

然而, 模型误设是一个普遍现象, 会导致 ML 估计量不一致, 因此尽管 ML 估计量更有效, 但是缺乏稳健性. 而之前介绍的 GMM 估计量并不依赖于任何概率分布的假设, 故而比 ML 估计量更稳健.

现在来看  $\hat{\theta}_{\text{ML}}$  的协方差矩阵估计量, 由于在条件密度模型正确设定的情况下有

$$\text{avar}(\sqrt{n}\hat{\theta}) = -\mathbf{H}_o^{-1} = \mathbf{\Omega}_o^{-1}$$

故而有两种方法可以估计  $\text{avar}(\sqrt{n}\hat{\theta})$ . 一种是

$$\hat{\mathbf{V}}_{\text{ML1}} = -\hat{\mathbf{H}}^{-1}(\hat{\theta})$$

其中  $\hat{\mathbf{H}}(\theta) = n^{-1} \sum_{i=1}^n \nabla_{\theta}^2 \log f(Y_i|X_i; \theta)$ . 另一种则是

$$\hat{\mathbf{V}}_{\text{ML2}} = n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})'$$

估计量  $\hat{\mathbf{V}}_{\text{ML1}}$  的一致性可由  $\hat{\theta}_{\text{ML}} \xrightarrow{p} \theta_o$ , 以及假设 9.2(4) 保证. 而估计量  $\hat{\mathbf{V}}_{\text{ML2}}$  的一致性需要做出额外假设, 具体参考 Newey and McFadden (1994).

此外, 估计量  $\hat{\mathbf{V}}_{\text{ML1}}$  具有更好的有限样本性质, 但  $\hat{\mathbf{V}}_{\text{ML2}}$  更加容易计算, 因此在不得不使用数值计算的时候会方便许多.

Newey and McFadden (1994) 证明了在 Probit 模型中有

$$\|\mathbf{H}(\theta)\| \leq 2\|X_i X_i'\|$$

如果  $\mathbb{E}[X_i X_i']$  正定且有限, 那么假设 9.2 的每一个条件都可以满足, 因此 ML 估计量具有渐近正态性.

## 9.4 CMLE 的有效性

本节将证明模型正确设定下的 ML 估计量在相当大的一类渐近正态估计量中都是有效的, 尽管这样的估计量类并没有包括全部的渐近正态估计量. 一个这样的估计量类就包含了 GMM 估计量, 因此 ML 估计量在我们所感兴趣的估计量中都是有效的.

首先需要阐明一个概念, 在全体渐近正态估计量构成的大类中, 不存在渐近有效的估计量. 为了看清这一点, 假设  $\theta_o$  的估计量  $\hat{\theta}$  是渐近正态的, 也即当  $n \rightarrow \infty$  时有

$$\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, \mathbf{V})$$

再定义另一个估计量

$$\tilde{\theta} = \begin{cases} \hat{\theta}, & |\hat{\theta} - \alpha| \geq n^{-\frac{1}{4}} \\ \alpha, & |\hat{\theta} - \alpha| < n^{-\frac{1}{4}} \end{cases}$$

可以证明,  $\tilde{\theta}$  也是  $\theta_o$  的一致估计量, 其中  $\alpha$  是某个给定的常数. 并且当  $\theta_o \neq \alpha$  时有

$$\sqrt{n}(\tilde{\theta} - \theta_o) \xrightarrow{d} N(0, V)$$

而当  $\theta_o = \alpha$  时有

$$n^\beta(\tilde{\theta} - \theta_o) \xrightarrow{d} 0$$

其中  $\beta$  是任意的实数.

换言之, 当  $\theta_o \neq \alpha$  时,  $\tilde{\theta}$  和  $\hat{\theta}$  具有相同的渐近分布, 而当  $\theta_o = \alpha$  时,  $\tilde{\theta}$  收敛到  $\theta_o$  的速率可以任意快, 并且渐近分布具有零方差, 称  $\tilde{\theta}$  关于  $\hat{\theta}$  是超有效的 (superefficient). 总的来看, 超有效性可能只在参数空间  $\Theta$  的 Lebesgue 零测集上才能取得, 因此在实践上意义不大.

回到正题, ML 估计量关于 GMM 估计量的有效性可以通过比较它们的渐近方差来证明. 设  $W_i = [X_i, Y_i]$ , 假定当模型正确设定时有

$$\mathbb{E}[g(W_i, \theta_o)] = 0 \quad (9.14)$$

其中  $g(W, \theta)$  表示一个  $L \times 1$  维向量. 在一定正则条件下, 积分和微分可交换, 按照前面的方法可以证得

$$-\mathbb{E}[\nabla_\theta g_i(\theta_o)] = \mathbb{E}[g_i(\theta_o)s_i(\theta_o)'] \quad (9.15)$$

令  $\hat{\theta}_{\text{GMM}}$  是与正交条件 (9.14) 相联系的 GMM 估计量, 根据之前的结论有

$$\text{avar}(\sqrt{n}\hat{\theta}_{\text{GMM}}) = \mathbb{E}[m_\theta]^{-1} \mathbb{E}[m m'] \mathbb{E}[m_\theta]^{-1}$$

其中

$$m_\theta = \mathbb{E}[\nabla_\theta g(W_i, \theta_o)]' W \nabla_\theta g(W_i, \theta_o)$$

$$m = \mathbb{E}[\nabla_\theta g(W_i, \theta_o)]' W g(W_i, \theta_o)$$

根据 (9.14) 可知  $\mathbb{E}[m_\theta] = -\mathbb{E}[m s']$ , 其中  $s = s_i(\theta_o)$ . 由于 ML 估计量 (在  $\sqrt{n}$  意义下的) 的渐近方差为  $\mathbb{E}[s s']^{-1}$ , 故而

$$\text{avar}(\sqrt{n}\hat{\theta}_{\text{GMM}}) - \text{avar}(\sqrt{n}\hat{\theta}_{\text{ML}}) = \mathbb{E}[m s']^{-1} \mathbb{E}[U U'] \mathbb{E}[m s']^{-1}$$

其中  $U = m - \mathbb{E}[m s'] \mathbb{E}[s s']^{-1} s$ . 显然上式是半正定的, 因此在某些正则条件下, ML 估计量比 GMM 估计量更加渐近有效. 事实上, 渐近方差  $\mathbb{E}[s s']^{-1}$  即为统计学中的 Cramer-Rao 下界.

注意, MLE 的有效性比最优 GMM 的还要强. 最优 GMM 是在给定了矩条件  $\mathbb{E}[g(W_i, \theta_o)] = 0$  的情况下, 通过选取最优权重矩阵获得的. 如果 GMM 要想达到 Cramer-Rao 下界, 则矩条件必须为  $\mathbb{E}[s_i(\theta_o)] = 0$ , 从这个意义上看, 具有最优矩条件的 GMM 渐近等价于 MLE. 此时 GMM 估计量  $\hat{\theta}$  应该满足

$$n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) = 0$$

这正是最大化对数似然函数  $\sum_{i=1}^n \log f(Y_i|X_i; \theta)$  的 FOC.

## 9.5 参数检验

现在考虑当条件密度模型  $f(y_i|X_i; \theta)$  设定正确时, 如何检验原假设

$$\mathbb{H}_0 : R(\theta_o) = r$$

其中  $R : \Theta \rightarrow \mathbb{R}^J$  是非随机的连续可微函数,  $\nabla_{\theta} R(\theta_o)$  是  $J \times P$  维满秩矩阵,  $r$  是  $J \times 1$  维非随机向量, 并且还满足  $J \leq P$ .

下面介绍以 ML 估计量  $\hat{\theta}_{ML}$  为基础的三大统计检验方法, 分别为 Wald 检验, 似然比 (Likelihood Ratio, LR) 检验, 以及 Lagrange 乘子 (Lagrange Multiplier, LM) 检验.

### 9.5.1 Wald 检验

在原假设  $\mathbb{H}_0 : R(\theta_o) = r$  成立的情况下, 根据一阶 Taylor 展开, ML 估计量的渐近正态性, 以及 Slutsky 定理可知

$$\begin{aligned} \sqrt{n}[R(\hat{\theta}) - r] &= \sqrt{n}[R(\theta_o) - r] + \nabla_{\theta} R(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta_o) \\ &= \nabla_{\theta} R(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta_o) \\ &\xrightarrow{d} N\{0, -[\nabla_{\theta} R(\theta_o)]\mathbf{H}_o^{-1}[\nabla_{\theta} R(\theta_o)]'\} \end{aligned}$$

其中  $\bar{\theta}$  位于  $\hat{\theta}$  与  $\theta_o$  之间. 于是二次型

$$n[R(\hat{\theta}) - r]' \{-[\nabla_{\theta} R(\theta_o)]\mathbf{H}_o^{-1}[\nabla_{\theta} R(\theta_o)]'\}^{-1} [R(\hat{\theta}) - r] \xrightarrow{d} \chi_J^2$$

由 Slutsky 定理可以得到 Wald 检验统计量

$$W = n[R(\hat{\theta}) - r]' \{-[\nabla_{\theta} R(\hat{\theta})][\hat{H}^{-1}(\hat{\theta})][\nabla_{\theta} R(\hat{\theta})]'\}^{-1} [R(\hat{\theta}) - r] \xrightarrow{d} \chi_J^2$$

其中

$$\hat{H}(\theta) = n^{-1} \sum_{i=1}^n \nabla_{\theta}^2 l_i(\theta)$$

可以看出, 只需使用无约束的 ML 估计量  $\hat{\theta}$  即可构造 Wald 检验统计量, 这是 Wald 检验的一大优点.

为了使 Wald 检验统计量服从渐近  $\chi^2$  分布, 必须将  $\theta$  限制在  $\Theta$  内部, 而不能在它的边界上. 举例而言, 如果限制  $\theta \in \Theta$  中的全体元素非负, 并且会用到这个约束条件, 那么在  $\mathbb{H}_0 : \theta_o = 0$  下, Wald 检验统计量不会有极限  $\chi^2$  分布.

此外, Wald 统计量的一大缺陷是, 对于以不同方式施加的非线性约束, 它不具有不变性. 考虑经典线性回归模型中的原假设  $\mathbb{H}_0 : \theta_1 = 1$ , 假定  $\theta_1 > 0$ , 渐近  $T$  统计量为  $(\hat{\theta}_1 - 1)/\text{se}(\hat{\theta}_1)$ , 现在定义  $\phi_1 = \log \theta_1$ , 那么原假设可以重新表述为  $\mathbb{H}_0 : \phi_1 = 0$ , 根据 Delta 法可知  $\text{se}(\hat{\phi}_1) = \hat{\theta}_1^{-1} \text{se}(\hat{\theta}_1)$ , 从而基于  $\hat{\phi}_1$  的  $T$  统计量为  $\hat{\phi}_1/\text{se}(\hat{\phi}_1) = \hat{\theta}_1 \log(\hat{\theta}_1)/\text{se}(\hat{\phi}_1) \neq (\hat{\theta}_1 - 1)/\text{se}(\hat{\phi}_1)$ .

由于缺乏不变性, 因此对于非线性假设而言, Wald 检验统计量的有限样本性质可能会很

差, 研究者不得不去研究原假设的各种表述, 以便获得一个令人满意的结果. 为此, 有必要寻找其它的统计量来克服缺乏不变性这个问题.

### 9.5.2 似然比检验

#### 定理 9.6

在假设 9.2 下, 定义对数条件似然样本均值

$$\begin{aligned}\hat{l}(\hat{\theta}) &= n^{-1} \sum_{i=1}^n l_i(\hat{\theta}) \\ \hat{l}(\tilde{\theta}) &= n^{-1} \sum_{i=1}^n l_i(\tilde{\theta})\end{aligned}$$

其中  $\hat{\theta}$  是无约束 ML 估计量, 而  $\tilde{\theta}$  是约束条件  $R(\tilde{\theta}) = r$  下的 ML 估计量. 如果原假设  $\mathbb{H}_0: R(\theta_0) = r$  成立, 那么当  $n \rightarrow \infty$  时有

$$LR = 2n[\hat{l}(\hat{\theta}) - \hat{l}(\tilde{\theta})] \xrightarrow{d} \chi_J^2$$



**证明** 根据之前的讨论, 由于无条件 ML 估计量  $\hat{\theta}$  是问题  $\max_{\theta \in \Theta} \hat{l}(\theta)$  的解, 对应的 FOC 为

$$\hat{s}(\hat{\theta}) = 0$$

其中  $\hat{s}(\theta) = n^{-1} \sum_{i=1}^n \nabla_{\theta} l_i(\theta)$ . 另一方面, 有约束 ML 估计量  $\tilde{\theta}$  是问题

$$\max_{\theta \in \Theta} \hat{l}(\theta) + \lambda'[r - R(\theta)]$$

的解, 其中  $\lambda$  为  $J \times 1$  维 Lagrange 乘子向量. 可以找到 FOC 为

$$\begin{aligned}\hat{s}(\tilde{\theta}) - [\nabla_{\theta} R(\tilde{\theta})]'\tilde{\lambda} &= 0 \\ R(\tilde{\theta}) - r &= 0\end{aligned}\tag{9.16}$$

将  $\hat{l}(\tilde{\theta})$  在  $\hat{\theta}$  处二阶 Taylor 展开可得

$$\begin{aligned}-LR &= 2n[\hat{l}(\tilde{\theta}) - \hat{l}(\hat{\theta})] \\ &= 2n[\hat{l}(\hat{\theta}) - \hat{l}(\hat{\theta})] + 2n[\hat{s}(\hat{\theta})]'(\tilde{\theta} - \hat{\theta}) + \sqrt{n}(\tilde{\theta} - \hat{\theta})'\hat{H}(\bar{\theta}_a)\sqrt{n}(\tilde{\theta} - \hat{\theta}) \\ &= \sqrt{n}(\tilde{\theta} - \hat{\theta})'\hat{H}(\bar{\theta}_a)\sqrt{n}(\tilde{\theta} - \hat{\theta})\end{aligned}$$

其中  $\bar{\theta}_a$  位于  $\hat{\theta}$  和  $\tilde{\theta}$  之间, 从而

$$LR = \sqrt{n}(\tilde{\theta} - \hat{\theta})'[-\hat{H}(\bar{\theta}_a)]\sqrt{n}(\tilde{\theta} - \hat{\theta})$$

现在将  $\hat{s}(\tilde{\theta})$  在无约束 ML 估计量  $\hat{\theta}$  处一阶 Taylor 展开, 再将其代入到 (9.16) 得到

$$\hat{s}(\hat{\theta}) + \hat{H}(\bar{\theta}_b)(\tilde{\theta} - \hat{\theta}) - [\nabla_{\theta} R(\tilde{\theta})]'\tilde{\lambda} = 0$$

其中  $\bar{\theta}_b$  也位于  $\hat{\theta}$  和  $\tilde{\theta}$  之间. 由于  $\hat{s}(\hat{\theta}) = 0$ , 故而

$$\hat{H}(\bar{\theta}_b)\sqrt{n}(\tilde{\theta} - \hat{\theta}) - [\nabla_{\theta} R(\tilde{\theta})]'\sqrt{n}\tilde{\lambda} = 0$$

当  $n$  充分大时, 矩阵  $\hat{H}(\bar{\theta}_b)$  可逆, 于是

$$\sqrt{n}(\tilde{\theta} - \hat{\theta}) = \hat{H}^{-1}(\bar{\theta}_b)[\nabla_{\theta} R(\tilde{\theta})]' \sqrt{n}\tilde{\lambda} \quad (9.17)$$

故而还需推导  $\sqrt{n}\tilde{\lambda}$  的渐近分布.

然后将  $\hat{s}(\tilde{\theta})$  在真实参数  $\theta_o$  处一阶 Taylor 展开, 再将其代入到 (9.16) 得到

$$[\nabla_{\theta} R(\tilde{\theta})]' \sqrt{n}\tilde{\lambda} = \sqrt{n}\hat{s}(\theta_o) + \hat{H}(\bar{\theta}_c)\sqrt{n}(\tilde{\beta} - \beta_o)$$

其中  $\bar{\theta}_c$  位于  $\tilde{\theta}$  和  $\theta_o$  之间. 对于充分大的  $n$  有

$$\hat{H}^{-1}(\bar{\theta}_c)[\nabla_{\theta} R(\tilde{\theta})]' \sqrt{n}\tilde{\lambda} = \hat{H}^{-1}(\bar{\theta}_c)\sqrt{n}\hat{s}(\theta_o) + \sqrt{n}(\tilde{\theta} - \theta_o) \quad (9.18)$$

进一步将  $R(\tilde{\beta}) - r = 0$  在  $\theta_o$  处一阶 Taylor 展开得到

$$\sqrt{n}[R(\theta_o) - r] + [\nabla_{\theta} R(\bar{\theta}_d)]' \sqrt{n}(\tilde{\theta} - \theta_o) = 0 \quad (9.19)$$

其中  $\bar{\theta}_d$  也位于  $\tilde{\theta}$  和  $\theta_o$  之间. 如果原假设  $\mathbb{H}_0 : R(\theta_o) = r$  成立, 那么根据 (9.19) 可得

$$[\nabla_{\theta} R(\bar{\theta}_d)]' \sqrt{n}(\tilde{\theta} - \theta_o) = 0 \quad (9.20)$$

在式 (9.18) 上左乘  $\nabla_{\theta} R(\bar{\theta}_d)$ , 根据 (9.20) 可知

$$[\nabla_{\theta} R(\bar{\theta}_d)][\hat{H}^{-1}(\bar{\theta}_c)][\nabla_{\theta} R(\tilde{\theta})]' \sqrt{n}\tilde{\lambda} = [\nabla_{\theta} R(\bar{\theta}_d)]\hat{H}^{-1}(\bar{\theta}_c)\sqrt{n}\hat{s}(\theta_o)$$

当  $n$  充分大时, 根据假设 9.2(6) 以及 Slutsky 定理可得

$$\begin{aligned} \sqrt{n}\tilde{\lambda} &= \{[\nabla_{\theta} R(\bar{\theta}_d)][\hat{H}^{-1}(\bar{\theta}_c)][\nabla_{\theta} R(\tilde{\theta})]'\}^{-1}[\nabla_{\theta} R(\bar{\theta}_d)]\hat{H}^{-1}(\bar{\theta}_c)\sqrt{n}\hat{s}(\theta_o) \\ &\xrightarrow{d} N(0, \{-[\nabla_{\theta} R(\theta_o)]\mathbf{H}_o^{-1}[\nabla_{\theta} R(\theta_o)]'\}^{-1}) \end{aligned} \quad (9.21)$$

在式 (9.17) 上左乘  $[-\hat{H}(\bar{\theta}_a)]^{\frac{1}{2}}$ , 根据 (9.21) 可知

$$\begin{aligned} [-\hat{H}(\bar{\theta}_a)]^{\frac{1}{2}}\sqrt{n}(\tilde{\theta} - \hat{\theta}) &= [-\hat{H}(\bar{\theta}_a)]^{\frac{1}{2}}\hat{H}^{-1}(\bar{\theta}_b)[\nabla_{\theta} R(\tilde{\theta})]' \sqrt{n}\tilde{\lambda} \\ &\xrightarrow{d} N(0, \mathbf{\Pi}) \sim \mathbf{\Pi}^{\frac{1}{2}}N(0, \mathbf{I}_P) \end{aligned} \quad (9.22)$$

其中

$$\mathbf{\Pi} = \mathbf{H}_o^{-\frac{1}{2}}[\nabla_{\theta} R(\theta_o)]'\{-[\nabla_{\theta} R(\theta_o)]\mathbf{H}_o^{-1}[\nabla_{\theta} R(\theta_o)]'\}^{-1}[\nabla_{\theta} R(\theta_o)]\mathbf{H}_o^{-\frac{1}{2}}$$

为  $P \times P$  维幂等矩阵, 并且秩为  $J$ . 最后根据引理 2.1 可知

$$\begin{aligned} LR &= \sqrt{n}(\tilde{\theta} - \hat{\theta})'[-\hat{H}(\bar{\theta}_a)]^{\frac{1}{2}}[-\hat{H}(\bar{\theta}_a)]^{\frac{1}{2}}\sqrt{n}(\tilde{\theta} - \hat{\theta}) \\ &\xrightarrow{d} \chi_J^2 \end{aligned}$$

证毕.

LR 检验是基于比较原假设  $\mathbb{H}_0 : R(\theta_o) = r$  下的对数似然函数  $\hat{l}(\tilde{\theta})$  和无约束条件下的对数似然函数  $\hat{l}(\hat{\theta})$ . 如果  $\mathbb{H}_0$  成立, 那么无约束模型的  $\hat{l}(\hat{\theta})$  和有约束模型的  $\hat{l}(\tilde{\theta})$  应该相近. 反之, 如果  $\hat{l}(\hat{\theta})$  显著大于  $\hat{l}(\tilde{\theta})$ , 则原假设  $\mathbb{H}_0$  应该为假.

相较于 Wald 检验来说, LR 检验可能更难施行, 毕竟计算起来十分麻烦, 甚至有时无法解析地得到表达式, 但是 LR 检验具有不变性, 这是它的一大优点. 此外, Wald 检验通常比 LR 检验更容易拒绝原假设.

### 9.5.3 Lagrange 乘子检验

除了以上两种方法外, 还可以通过 Lagrange 乘子  $\tilde{\lambda}$  构造 LM 检验, 它又称为 Rao 有效得分检验. 考虑以下有约束的最大化问题

$$\max_{\theta \in \Theta} \hat{l}(\theta) + \lambda'[r - R(\theta)]$$

这里最优 Lagrange 乘子度量了约束条件  $\mathbb{H}_0 : R(\theta_o) = r$  对模型似然函数最大值的影响. 当  $\mathbb{H}_0$  成立时, 施加该约束应该对似然函数最大值的影响不大, 也即  $\tilde{\lambda}$  应该很小. 反之, 如果  $\tilde{\lambda}$  的值很大, 则有理由拒绝原假设.

之前在推导 LR 检验统计量时就已得出

$$\begin{aligned} \sqrt{n}\tilde{\lambda} &= \{[\nabla_{\theta} R(\bar{\theta}_d)][\hat{H}^{-1}(\bar{\theta}_c)][\nabla_{\theta} R(\bar{\theta})]'\}^{-1}[\nabla_{\theta} R(\bar{\theta}_d)]\hat{H}^{-1}(\bar{\theta}_c)\sqrt{n}\hat{s}(\beta_o) \\ &\xrightarrow{d} N(0, \{-[\nabla_{\theta} R(\theta_o)]H_o^{-1}[\nabla_{\theta} R(\theta_o)]'\}^{-1}) \end{aligned}$$

从而二次型

$$n\tilde{\lambda}'\{-[\nabla_{\theta} R(\theta_o)]H_o^{-1}[\nabla_{\theta} R(\theta_o)]'\}\tilde{\lambda} \xrightarrow{d} \chi_J^2$$

最后根据 Slutsky 定理得到 LM 检验统计量

$$LM = -n\tilde{\lambda}'[\nabla_{\theta} R(\tilde{\theta})][\hat{H}^{-1}(\tilde{\theta})][\nabla_{\theta} R(\tilde{\theta})]'\tilde{\lambda} \xrightarrow{d} \chi_J^2$$

简单总结一下, Wald 检验仅使用无约束信息, LM 检验仅使用有约束信息, LR 检验则同时利用了这两种信息, 以上三类检验在大样本下是渐近等价的. Wald 检验使用范围最广, 因为不对条件密度做具体假设, 但 Wald 检验不具有不变性; LR 检验以及某些 LM 检验具有不变性, 但可能很难得到似然函数.

## 9.6 模型设定检验

由于 MLE 依赖于条件密度模型  $f(y|X_i; \theta)$  的正确设定, 如果模型出现误设, 则通常难以得到一致估计量, 即使使用后面介绍的 QMLE 也会造成精度下降. 因此有必要检验原假设

$$\mathbb{H}_0 : \forall x \in \mathcal{X}, \exists \theta_o \in \Theta, \text{ s.t. } f(\cdot|X; \theta_o) = p_o(\cdot|X)$$

White (1982) 提出了通过检验信息矩阵等式

$$\mathbb{E}[s_i(\theta_o)s_i(\theta_o)'] + \mathbb{E}[H_i(\theta_o)] = 0$$

是否成立来检验  $\mathbb{H}_0$  是否成立.

定义  $Q \times 1$  维矩样本均值

$$\hat{g}(\theta) = n^{-1} \sum_{i=1}^n g_i(\theta)$$

其中  $Q = P(P+1)/2$ , 并且

$$g_i(\theta) = \text{vech}[s_i(\theta)s_i(\theta)' + H_i(\theta)]$$



令  $\hat{\theta}$  为 ML 估计量, 在正则条件下可以用 UWLLN 推出

$$\hat{g}(\hat{\theta}) \xrightarrow{p} \mathbb{E}[g_i(\theta_o)]$$

如果信息矩阵等式成立, 那么  $\mathbb{E}[g_i(\theta_o)] = 0$ , 此时  $\hat{g}(\hat{\theta})$  接近于零向量, 因此可以考虑推导  $\sqrt{n}\hat{g}(\hat{\theta})$  的渐近分布. White (1982) 证明了, 一定情况下在  $n \rightarrow \infty$  时有

$$\begin{aligned} \sqrt{n}\hat{g}(\hat{\theta}) &= n^{-\frac{1}{2}} \sum_{i=1}^n [g_i(\theta_o) - \mathbf{G}_o \mathbf{H}_o^{-1} s_i(\theta_o)] + o_p(1) \\ &\xrightarrow{d} N(0, \Sigma) \end{aligned}$$

其中  $\mathbf{G}_o = \mathbb{E}[\nabla_{\theta} g_i(\theta_o)]$ , 渐近协方差矩阵

$$\Sigma = \text{var}[g_i(\theta_o) - \mathbf{G}_o \mathbf{H}_o^{-1} s_i(\theta_o)]$$

从而在模型设定正确时, 信息矩阵统计量

$$IM = n[\hat{g}(\hat{\theta})]' \hat{\Sigma}^{-1} [\hat{g}(\hat{\theta})] \xrightarrow{d} \chi_Q^2$$

其中  $\Sigma$  的一致估计量为

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \hat{M}_i \hat{M}_i'$$

并且  $\hat{M}_i = g_i(\hat{\theta}) - [\nabla_{\theta} \hat{g}(\hat{\theta})] \hat{H}^{-1}(\hat{\theta}) s_i(\hat{\theta})$ .

同样地, IM 检验本质上检验的是信息矩阵等式是否成立. 由于信息矩阵等式成立只是模型设定正确的必要非充分条件, 因此 IM 检验在大样本下不能拒绝原假设  $\mathbb{H}_0$  并不意味着模型设定正确, 只是说没有发现模型误设的证据.

## 9.7 拟极大似然估计

### 9.7.1 一般误设

当条件密度模型  $f(y|X_i; \theta)$  设定错误时, 对于任意的  $\theta \in \Theta$ , 总有  $f(\cdot|X_i; \theta) \neq p_o(\cdot|X_i)$ . 假设伪真值 (pseudo-true value)  $\theta^*$  是问题

$$\max_{\theta \in \Theta} \mathbb{E}[l_i(\theta)]$$

的唯一最优解, 并且  $\hat{\theta}$  是问题

$$\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \log f(Y_i|X_i; \theta)$$

的解, 那么称  $\hat{\theta}$  为拟极大似然估计量 (quasi-maximum likelihood estimator). 在类似 9.2 的正则条件下, 可以通过极值估计量的一致性引理得到

$$\hat{\theta} = \theta^* + o_p(1)$$



但此时不能将  $\theta^*$  解释为真实参数, 事实上它最小化了  $f(\cdot|X_i; \theta)$  和  $p_o(\cdot|X_i)$  之间的距离. 其渐近分布为

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta)$$

这里的渐近协方差矩阵为

$$\begin{aligned} V_\theta &= H_*^{-1} \Omega_* H_*^{-1} \\ &= \mathbb{E}[H_i(\theta^*)]^{-1} \text{avar}[\sqrt{n}\hat{s}(\theta^*)] \mathbb{E}[H_i(\theta^*)]^{-1} \end{aligned}$$

由于条件信息矩阵等式不成立, 因此  $V_\theta$  不能化简, 并且它一般也没有达到 Cramér-Rao 下界, 因此比 MLE 的有效性更低. 此时 QML 估计量的协方差矩阵估计量

$$\hat{V}_{\text{QML}} = \left[ \sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} \left[ \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})' \right] \left[ \sum_{i=1}^n H_i(\hat{\theta}) \right]^{-1} \quad (9.23)$$

在原假设  $\mathbb{H}_0: R(\theta^*) = r$  下, 还可以根据 (9.23) 实施稳健 Wald 检验与稳健 LM 检验, 但是无法构造 LR 检验统计量, 因为条件信息矩阵等式不再成立.

在绝大多数应用中, 我们都无法正确设定条件密度函数, 因此使用的都是 GMM 或者 QMLE 来估计参数. 如果误设程度越低, 也即  $f(\cdot|X_i; \theta)$  与  $p_o(\cdot|X_i)$  越接近, 则 QMLE 的有效性越好.

### 9.7.2 模型选择检验

有时候可能存在多个相互竞争的误设模型, 我们需要在其中选择一个更有吸引力的. 在这方面, Vuong (1989) 基于 LR 检验做出了许多成果.

假设  $f_1(y|X; \theta_1)$  和  $f_2(y|X; \theta_2)$  是条件分布  $D(Y_i|X_i)$  的密度的候选模型,  $\hat{\theta}_1$  和  $\hat{\theta}_2$  分别是收敛于  $\theta_1^*$  和  $\theta_2^*$  的 QMLE. 首先定义条件密度模型

$$\begin{aligned} \mathcal{F}_1 &= \{f_1(y|X; \theta_1) : \theta_1 \in \Theta_1\} \\ \mathcal{F}_2 &= \{f_2(y|X; \theta_2) : \theta_2 \in \Theta_2\} \end{aligned}$$

如果: (i)  $\mathcal{F}_1 \cap \mathcal{F}_2 = \emptyset$ , 那么称  $\mathcal{F}_1$  和  $\mathcal{F}_2$  是严格非嵌套的; (ii)  $\mathcal{F}_2 \subset \mathcal{F}_1$ , 那么称  $\mathcal{F}_2$  嵌套于  $\mathcal{F}_1$ ; (iii)  $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$ ,  $\mathcal{F}_1 \subset \mathcal{F}_2$ , 并且  $\mathcal{F}_2 \subset \mathcal{F}_1$ , 则称  $\mathcal{F}_1$  与  $\mathcal{F}_2$  重叠.

对于严格非嵌套模型, 定义

$$l_m = \sum_{i=1}^n l_{im}(\hat{\theta}_m), \quad m = 1, 2$$

为在对应的估计量处的拟对数似然. 在正则条件下有

$$\begin{aligned} n^{-\frac{1}{2}}(l_1 - l_2) &= n^{-\frac{1}{2}} \sum_{i=1}^n [l_{i1}(\hat{\theta}_1) - l_{i2}(\hat{\theta}_2)] \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n [l_{i1}(\theta_1^*) - l_{i2}(\theta_2^*)] + o_p(1) \end{aligned}$$

在原假设  $\mathbb{H}_0: \mathbb{E}[l_{1i}(\theta_1^*)] = \mathbb{E}[l_{i2}(\theta_2^*)]$  下, 可以证明

$$n^{-\frac{1}{2}} \sum_{i=1}^n [l_{1i}(\theta_1^*) - l_{i2}(\theta_2^*)] \xrightarrow{d} N(0, \eta^2)$$

其中  $\eta^2 = \text{var}[l_{1i}(\theta_1^*) - l_{i2}(\theta_2^*)]$ . 渐近方差  $\eta^2$  的一致估计量为

$$\hat{\eta}^2 = n^{-1} \sum_{i=1}^n [l_{1i}(\hat{\theta}_1) - l_{i2}(\hat{\theta}_2)]^2$$

当原假设  $\mathbb{H}_0$  成立时, 可以得到 VMS 检验统计量

$$n^{-\frac{1}{2}}(l_1 - l_2)/\hat{\eta} \xrightarrow{d} N(0, 1)$$

如果 VMS 统计量显著大于 0, 则说明模型 1 拟合得更好, 显著小于 0 的情况也可以类似地解释为模型 2 拟合得更好, 但这同样无法说明模型 1 和模型 2 是否是正确设定的. 最后, 对于模型嵌套和重叠的情形, 具体见 Vuong (1989).

### 9.7.3 线性指数族的 QMLE

之前讨论的内容允许模型  $f(y|X; \theta)$  没有任何东西是正确设定的, 但在模型设定正确与完全误设之间存在一个中间地带, 允许条件密度的某些特征是正确设定的.

**例 9.2** 假设  $\{(Y_i, X_i)\}_{i=1}^n$  为 i.i.d. 随机样本, 非线性回归模型为

$$Y_i = g(X_i, \beta_o) + e_i$$

并且还满足  $\mathbb{E}[e_i|X_i] = 0$ , 然而我们并不知道  $e_i|X_i$  的条件密度. 为了估计真实参数  $\beta_o$ , 尽管这么做可能不正确, 但仍假设  $e_i|X_i \sim \text{i.i.d. } N(0, \sigma^2)$ , 此时  $Y_i$  的拟条件概率密度为

$$f(y|X_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{[y - g(X_i, \beta)]^2}{2\sigma^2} \right\}$$

其中  $\theta = [\beta', \sigma^2]'$ . 由此可以定义 QML 估计量

$$\hat{\theta} = [\hat{\beta}', \hat{\sigma}^2]' = \arg \max_{\beta, \sigma^2} \sum_{i=1}^n \log f(Y_i|X_i; \theta)$$

则在一般的正则条件下,  $\hat{\beta}$  为  $\beta_o$  的一致估计量.

事实上, 上述例子的结论之所以能成立, 是因为正态分布是线性指数族 (Linear Exponential Family, LEF) 中的一员, 除了正态分布外, LEF 中还包括 Bernoulli 分布, Poisson 分布, Gamma 分布等. 为简单起见, 我们仅考虑响应变量为标量的情形.

Gourieroux et al. (1984) 的结果表明, LEF 中的对数似然可以写为均值的一个函数

$$\log f(y|\mu) = a(\mu) + b(y) + yc(\mu) \quad (9.24)$$

其中  $\mu$  是随机变量  $Y_i$  均值的一个待选值, 并且  $\mu_o = \mathbb{E}[Y_i]$  为真实期望. 例如  $Y$  服从正态分布  $N(\mu, \sigma^2)$ , 则

$$\log f(y|\mu) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - \mu)^2$$

由此可知 (9.24) 中有  $c(\mu) = \mu/\sigma^2$ . 而对于 Bernoulli 分布, 有

$$\log f(y|\mu) = (1-y)\log(1-\mu) + y\log\mu, \quad 0 < \mu < 1$$

因此,  $a(\mu) = \log(1-\mu)$ ,  $b(y) = 0$ ,  $c(\mu) = \log[\mu/(1-\mu)]$ . 事实上, 估计 Probit 模型的正是 Bernoulli QMLE.

现在将  $\mu$  参数化为  $m(X, \theta)$ , 那么条件拟对数似然变为

$$\log f(y|m(X, \theta)) = a(m(X, \theta)) + b(y) + yc(m(X, \theta))$$

假设条件均值是正确设定的, 那么我们可以假定存在  $\theta_o \in \Theta$ , 使得  $\mathbb{E}[Y_i|X_i] = m(X_i, \theta_o)$ . 加上其它技术性条件,  $\hat{\theta}_{\text{QML}}$  是真实参数  $\theta_o$  的一致估计量.

在上面的例子中, 由于  $\mathbb{E}[Y_i|X_i] = g(X_i, \beta_o)$ , 并且  $Y$  服从正态分布, 因此可以通过 QMLE 一致估计出  $\beta_o$ . 不仅如此, 考虑  $Y_i$  为一个非负连续随机变量, 如果条件均值是正确设定的, 那么我们使用 Poisson QMLE 同样也可以得出一致估计量, 唯一的限制是  $\mathbb{E}[Y_i|X_i = x]$  候选值的范围应该与从 LEF 中选取的密度函数所允许的范围相同<sup>3</sup>.

<sup>3</sup>举例而言, 如果  $\mathbb{E}[Y_i|X_i = x]$  的可能值为负, 那么就不能将  $Y_i|X_i$  设置为服从 Poisson 分布, 因为 Poisson 分布的均值不为负.

## 第 10 章 限值因变量模型

上一章提到的 Probit 模型是限值因变量 (Limited Dependent Variable, LDV) 模型中的特例, 它的因变量取值只有 0 和 1. 本章使用极大似然估计法, 简要讨论一些常见的 LDV 模型, 包括二值响应模型, 归并回归模型, 断尾回归模型以及样本选择模型等.

### 10.1 二值响应模型

#### 10.1.1 Probit 与 Logit

在二值响应模型中, 因变量取值范围为  $\{0, 1\}$ , 我们感兴趣的为响应概率

$$P(x) = \mathbb{P}[Y = 1|X = x]$$

以及回归模型

$$Y = P(X) + e \quad (10.1)$$

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

如果有  $P(X) = X'\beta$ , 那么 (10.1) 为线性概率模型 (Linear Probability Model, LPM)

$$Y = X'\beta + e$$

此时 OLS 估计量一定不是一致估计量, 因为  $e$  服从两点分布, 要么为  $1 - X'\beta$ , 要么为  $-X'\beta$ , 故而  $X_i$  与  $e_i$  必定相关. 不仅如此, OLS 的预测值还会超出  $[0, 1]$  这一范围.

为了克服 LPM 的局限性, 选取某个值域严格位于  $(0, 1)$  的实值函数  $G$ , 此时响应概率为

$$\mathbb{P}[Y = 1|X] = G(X'\beta) \quad (10.2)$$

非线性函数  $G$  可以选取为

$$G(z) = e^z / (1 + e^z) = \Lambda(z) \quad (10.3)$$

它是标准 Logistic 随机变量的累积分布函数, (10.1) 为 Logit 模型. 而如果函数  $G$  为

$$G(z) = \int_{-\infty}^z \phi(v) dv = \Phi(z) \quad (10.4)$$

则 (10.1) 为之前提到的 Probit 模型, 其中  $\phi$  为标准正态随机变量的概率密度. 显然, Probit 函数与 Logit 函数都是严格递增的, 并且当  $z \rightarrow -\infty$  时,  $G(z) \rightarrow 0$ , 而当  $z \rightarrow \infty$  时,  $G(z) \rightarrow 1$ .

如同上一章开头的那样, Probit 模型和 Logit 模型可以由潜变量模型得到, 也即

$$Y^* = X'\beta + e, \quad Y = \mathbb{1}[Y^* > 0] \quad (10.5)$$

其中  $Y^*$  是不可观测的, 仅能观测到指示  $Y^*$  符号的二值变量. 此外, 假设  $e$  独立于  $X$ , 并且要么服从标准 Logistic 分布, 要么服从标准正态分布<sup>1</sup>. 无论哪种情况, 对于任意实数  $z$  总有

<sup>1</sup>显然这已经排除了潜变量模型中  $e$  存在异方差的情况, 与通常的线性回归或非线性回归不同, 潜变量模型的异方差性会导致 MLE 不一致. 这部分内容具体见 Wooldridge (2010).

$G(z) = 1 - G(-z)$ . 于是对于模型 (10.5), 响应概率为

$$\begin{aligned}\mathbb{P}[Y = 1|X] &= \mathbb{P}[Y^* > 0|X] = \mathbb{P}[e > -X'\beta|X] \\ &= 1 - G(-X'\beta) = G(X'\beta)\end{aligned}$$

这就得到了式 (10.2) 的情形.

对于二值响应模型, 重要的是解释回归元  $X_j$  对响应概率  $\mathbb{P}[Y = 1|X]$  的影响. 注意到

$$\mathbb{E}[Y|X] = \mathbb{P}[Y = 1|X] = G(X'\beta)$$

因此  $\beta$  不能像线性概率模型那样直接解释. 如果  $X_j$  是连续型的, 则它在  $p(X) = \mathbb{P}[Y = 1|X]$  上的回归导数为

$$\frac{\partial p(X)}{\partial X_j} = g(X'\beta)\beta_j, \quad g(z) = \frac{dG(z)}{dz}$$

这里的  $g$  为  $G$  的导数. 而如果  $X_j$  是二值的, 以  $X_1$  为例, 那么  $X_1$  对  $p(X)$  的影响为

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

利用随机样本, 我们可以对偏效应进行估计. 假设  $\hat{\beta}$  为 ML 估计量, 那么当连续型变量  $X_j$  变动很小时, 它对  $p(X)$  的平均边际效应 (Average Marginal Effect, AME) 为

$$n^{-1} \sum_{i=1}^n g(X_i' \hat{\beta}) \hat{\beta}_j$$

对于二元变量  $X_1$  也可以类似得到

$$n^{-1} \sum_{i=1}^n [G(\hat{\beta}_0 + \hat{\beta}_1 + \cdots + X_{ki} \hat{\beta}_k) - G(\hat{\beta}_0 + X_{2i} \hat{\beta}_2 + \cdots + X_{ki} \hat{\beta}_k)]$$

为了利用 CMLE, 类似于上一章的做法, Probit 模型和 Logit 模型在观测值  $i$  上的对数似然为

$$l_i(\beta) = Y_i \log G(X_i' \beta) + (1 - Y_i) \log[1 - G(X_i' \beta)]$$

根据定义, ML 估计量  $\hat{\beta}$  是问题

$$\max_{\beta \in \Theta} n^{-1} \sum_{i=1}^n \log l_i(\beta)$$

的唯一解.

可以证明, 在模型正确设定的情况下, 如果  $\mathbb{E}[X_i X_i']$  正定且有限, 则这 Probit 模型和 Logit 模型的 ML 估计量都是有效的渐近正态一致估计量, 也即

$$\begin{aligned}\sqrt{n}(\hat{\beta}_P - \beta_P) &\xrightarrow{d} N(0, V_P) \\ \sqrt{n}(\hat{\beta}_L - \beta_L) &\xrightarrow{d} N(0, V_L)\end{aligned}$$

其中

$$\begin{aligned}V_P &= Q_P^{-1} = \mathbb{E}[\lambda(X_i' \beta_P) \lambda(-X_i' \beta_P) X_i X_i']^{-1} \\ V_L &= Q_L^{-1} = \mathbb{E}[\Lambda(X_i' \beta_L) (1 - \Lambda(X_i' \beta_L)) X_i X_i']^{-1}\end{aligned}$$

这里的  $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$  为 IMR. 此时  $\hat{\beta}_P$  和  $\hat{\beta}_L$  的协方差矩阵估计量为

$$\hat{V}_P^0 = \left[ \sum_{i=1}^n \lambda(X_i' \hat{\beta}_P) \lambda(-X_i' \hat{\beta}_P) X_i X_i' \right]^{-1}$$

$$\hat{V}_L^0 = \left\{ \sum_{i=1}^n \Lambda(X_i' \hat{\beta}_L) [1 - \Lambda(X_i' \hat{\beta}_L)] X_i X_i' \right\}^{-1}$$

而如果只有条件均值设定正确, 由于两点分布属于 LEF, 于是在正则条件下的 QML 估计量  $\hat{\beta}_P$  和  $\hat{\beta}_L$  仍是一致的, 但渐近方差  $V_P, V_L$  及其它们的估计量会更为复杂.

### 10.1.2 内生性问题

考虑如下潜变量结构模型

$$Y^* = X_1' \beta_1 + \beta_2 X_2 + e_1 \quad (10.6)$$

$$X_2 = X_1' \gamma_1 + Z' \gamma_2 + e_2 \quad (10.7)$$

$$Y = \mathbb{1}[Y^* > 0] \quad (10.8)$$

其中  $X_2$  是一个内生的连续型解释变量<sup>2</sup>, 而  $X_1$  中的解释变量均是外生的,  $Z$  包含了不在  $X_1$  中的排他性工具, 并且还假定  $[e_1, e_2]$  和  $[X_1, X_2]$  独立.

为了应用 CMLE, 还需要假定扰动项服从联合正态分布, 也即

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \Big| [X_1, Z] \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

此时

$$Y^* = \mu(\theta) + \varepsilon$$

$$\mu(\theta) = X_1' \beta_1 + \beta_2 X_2 + \delta(X_2 - X_1' \gamma_1 - Z' \gamma_2)$$

$$e_1 = \delta e_2 + \varepsilon$$

$$\delta = \sigma_{12}/\sigma_2^2$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$\sigma_\varepsilon^2 = 1 - \sigma_{12}^2/\sigma_2^2$$

并且随机误差项  $\sigma_\varepsilon$  独立于  $e_2$ , 因此也独立于  $X_2$ . 根据以上论述, 可以得到  $[Y, X_2]$  的条件联合概率密度

$$f(y, x_2 | X_1, Z) = \Phi \left( \frac{\mu(\theta)}{\sigma_\varepsilon} \right)^y \left[ 1 - \Phi \left( \frac{\mu(\theta)}{\sigma_\varepsilon} \right) \right]^{1-y} \frac{1}{\sigma_2} \phi \left( \frac{x_2 - X_1' \gamma_1 - Z' \gamma_2}{\sigma_2} \right)$$

<sup>2</sup>由于  $X_2$  在假设下服从正态分布, 因此必须为连续型随机变量, 而不能是离散的.

于是 ML 估计量是以下最大化问题的唯一解

$$\begin{aligned} \max_{\theta \in \Theta} \sum_{i=1}^n & \left\{ Y_i \log \Phi \left( \frac{\mu_i(\theta)}{\sigma_\varepsilon} \right) + (1 - Y_i) \log \left[ 1 - \Phi \left( \frac{\mu_i(\theta)}{\sigma_\varepsilon} \right) \right] \right\} \\ & - \frac{n}{2} \log 2\pi\sigma_2^2 - \frac{1}{2\sigma_2^2} \sum_{i=1}^n (X_{2i} - X'_{1i}\gamma_1 - Z'_i\gamma_2)^2 \end{aligned}$$

以上方法称为 IV Probit.

由于 MLE 在数值计算上可能不易收敛 (尤其是  $e_1$  和  $e_2$  高度相关, 或者含有多个内生变量的时候), 因此 Rivers and Vuong (1988) 提出了两步法估计, 它也是 CF 方法的应用. 设  $\rho$  为  $e_1$  与  $e_2$  的相关系数, 于是  $\sigma_{21} = \rho\sigma_2$ ,  $\delta = \rho/\sigma_2$ , 以及

$$\sigma_\varepsilon^2 = 1 - \delta\sigma_{12} = 1 - \rho^2$$

根据  $e_1 = \delta e_2 + \varepsilon$  可知

$$\begin{aligned} Y^* &= X'_1\beta_1 + \beta_2 X_2 + \delta e_2 + \varepsilon \\ \varepsilon | X_1, X_2, e_2 &\sim N(0, 1 - \rho^2) \end{aligned}$$

可以证明

$$\mathbb{P}[Y = 1 | X_1, X_2, e_2] = \Phi[(X'_1\beta + \beta_2 X_2 + \delta e_2)/(1 - \rho^2)^{\frac{1}{2}}]$$

一旦能够观察到  $e_2$ , 那么  $Y$  对  $X_1, X_2$  与  $e_2$  的 Probit 会产生

$$\left[ \frac{\beta_1}{\sqrt{1 - \rho^2}}, \frac{\beta_2}{\sqrt{1 - \rho^2}}, \frac{\delta}{\sqrt{1 - \rho^2}} \right] \quad (10.9)$$

的一致估计量. 然而  $e_2$  是不可观测的, 此时可以使用 OLS 估计方程 (10.7) 并取得残差  $\hat{e}_2$  (一阶段), 然后用它将  $e_2$  替换掉, 再实施  $Y$  对  $X_1, X_2$  与  $\hat{e}_2$  的 Probit (二阶段), 这样得到的 CF 估计量的渐近性质已由 Rivers and Vuong (1988) 给出. 值得注意的是,  $[e_1, e_2]$  的联合正态对于 CF 估计不是必要的, 但对于 MLE 则必不可少.

相较于直接使用 MLE, 两步法的优点之一是计算简单, 由于一阶段回归的误差被代入到了二阶段, 因此 CF 估计量不如 ML 估计量有效率, 也不易获得合适的标准误. 而一旦潜在的分布假设成立, MLE 就可以取得一致和有效的估计量, 标准误也更容易计算.

此外, 两步法 IV Probit 模型估计的是 (10.9) 而非原本未经缩放的系数, 因此 CF 估计量的绝对值比 ML 估计量的更大, 因此需要通过软件报告出的  $\rho$  值进行调整. 此外, Rivers and Vuong (1988) 还证明了在模型恰好识别的情况下, ML 估计量和 CF 估计量是相同的.

## 10.2 断尾回归模型

考虑经典的线性回归模型

$$\begin{aligned} Y &= X'\beta + e \\ e | X &\sim N(0, \sigma^2) \end{aligned}$$

假设  $c$  为某个已知常数, 我们只能观测到  $Y_i \geq c$  的数据, 而  $Y_i < c$  时的数据是缺失的, 此时称随机变量  $Y$  是断尾的 (truncated). 此时  $Y$  的条件概率密度为

$$f(y|Y > c) = \frac{f(y)}{\mathbb{P}[Y > c]} \quad (10.10)$$

其中  $f(y)$  是  $Y$  不存在断尾时的概率密度. 进一步有

$$\begin{aligned} \mathbb{E}[Y|X, Y > c] &= X'\beta + \sigma \mathbb{E}\left[\frac{e}{\sigma} \mid \frac{e}{\sigma} > \frac{c - X'\beta}{\sigma}\right] \\ &= X'\beta + \sigma \lambda\left(-\frac{c - X'\beta}{\sigma}\right) \end{aligned} \quad (10.11)$$

由于  $Y_i \geq c$  是样本可观测的条件, 因此式 (10.11) 表明直接用 OLS 估计  $Y_i = X_i'\beta + e_i$  是不一致的. 此外, OLS 预测值可能出现  $\hat{Y} \leq c$  的不可能情形. 为了克服这种困难, 我们可以使用 CMLE.

首先注意到  $Y|X$  在断尾前的分布为  $N(X'\beta, \sigma^2)$ , 其概率密度函数为

$$f(y|X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y - X_i'\beta}{\sigma}\right)^2\right] = \frac{1}{\sigma} \phi\left(\frac{y - X_i'\beta}{\sigma}\right)$$

另一方面有

$$\begin{aligned} \mathbb{P}[Y > c|X] &= 1 - \mathbb{P}\left[\frac{Y - X'\beta}{\sigma} \leq \frac{c - X'\beta}{\sigma} \mid X\right] \\ &= 1 - \Phi\left(\frac{c - X'\beta}{\sigma}\right) \end{aligned}$$

根据 (10.10) 可知  $Y$  断尾后的密度为

$$f(y|X_i, Y > c) = \frac{\sigma^{-1} \phi[(y - X_i'\beta)/\sigma]}{1 - \Phi[(c - X_i'\beta)/\sigma]}$$

由此可以得到观测值  $i$  的条件对数似然

$$l_i(\theta) = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2} \left(\frac{Y_i - X_i'\beta}{\sigma}\right)^2 - \log\left[1 - \Phi\left(\frac{c - X_i'\beta}{\sigma}\right)\right]$$

由此可以得到 ML 估计量, 并且在  $\mathbb{E}[X_i X_i']$  非奇异和其它正则条件下, 断尾回归的 ML 估计量是一致和渐近正态的.

## 10.3 归并回归模型

### 10.3.1 Tobit 模型

假设  $Y$  是一个在正值上连续的随机变量, 但它的值取 0 的概率为正, 此时  $Y$  的分布是混合型的. 如果直接使用 OLS, 那么得到的预测值同样可能为负, 并且 OLS 估计量不是一致的.

为了看到这一点, 考虑如下归并回归 (censored regression) 模型

$$Y^* = X'\beta + e \quad (10.12)$$

$$Y = \max\{0, Y^*\}$$

其中回归方程 (10.12) 满足经典的线性模型假设, 并且随机扰动项服从条件正态分布  $e|X \sim$



$N(0, \sigma^2)$ . 这里的数据归并意味着小于 0 的  $Y$  都被压缩到了一个点上, 上述归并回归模型又称为第 I 类 Tobit 模型, 因为它是由 Tobin (1958) 提出的.

首先求  $Y > 0$  的子样本的条件期望

$$\begin{aligned}\mathbb{E}[Y|X, Y > 0] &= \mathbb{E}[Y^*|X, Y > 0] = \mathbb{E}[X'\beta + e|X, Y > 0] \\ &= X'\beta + \mathbb{E}[e|X, e > -X'\beta] = X'\beta + \sigma\lambda(X'\beta/\sigma)\end{aligned}\quad (10.13)$$

其中最后一个等号是通过标准正态的断尾公式<sup>3</sup>得到的. 进一步求全样本的条件期望

$$\begin{aligned}\mathbb{E}[Y|X] &= 0 \cdot \mathbb{P}[Y = 0|X] + \mathbb{E}[Y|X, Y > 0] \cdot \mathbb{P}[Y > 0|X] \\ &= \mathbb{E}[Y|X, Y > 0] \cdot \mathbb{P}[Y > 0|X]\end{aligned}$$

注意到

$$\mathbb{P}[Y > 0|X] = \mathbb{P}[Y^* > 0|X] = \mathbb{P}[e > -X'\beta|X] = \Phi(X'\beta/\sigma)\quad (10.14)$$

因此

$$\begin{aligned}\mathbb{E}[Y|X] &= \Phi(X'\beta/\sigma)[X'\beta + \sigma\lambda(X'\beta/\sigma)] \\ &= X'\beta\Phi(X'\beta/\sigma) + \sigma\phi(X'\beta/\sigma)\end{aligned}\quad (10.15)$$

根据 (10.13) 和 (10.15) 可知, 无论是全样本还是子样本, 对方程 (10.12) 的 OLS 回归都不会产生一致估计量.

然而, 我们可以使用 MLE 对 Tobit 模型进行估计. 类似可以写出

$$\mathbb{P}[Y = 0|X] = \mathbb{P}[Y^* < 0|X] = \mathbb{P}[e < -X'\beta|X] = \Phi(-X'\beta/\sigma)\quad (10.16)$$

此时可以写出  $Y$  的条件概率密度为

$$f(y|X_i) = \left[ \Phi\left(-\frac{X_i'\beta}{\sigma}\right) \right]^{\mathbb{1}_{[y=0]}} \left[ \frac{1}{\sigma} \phi\left(\frac{y - X_i'\beta}{\sigma}\right) \right]^{\mathbb{1}_{[y>0]}}$$

于是 ML 估计量为

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \log f(Y_i|X_i)$$

其中  $\theta = [\beta', \sigma']$ .

Tobit 估计量的渐近性质由 Amemiya (1973) 给出, 就  $\{X_i\}$  取固定常数序列的情况而言, 如果  $X_i$  是有界的并且  $\lim n^{-1} \sum_{i=1}^n X_i X_i'$  非奇异, 则 Tobit 估计量是一致的和渐近正态的<sup>4</sup>. 一旦通过 MLE 获得了估计量和渐近标准误, 三大检验方法都可以根据需要来轻易使用.

现在来看 Tobit 模型的解释, 它和上一节的二值选择模型一样无法直接用回归系数来解释边际效应. 假设  $X_j$  为连续型随机变量, McDonald and Moffitt (1980) 给出了一个特别有用的分

<sup>3</sup> 设  $Y \sim N(0, 1)$ ,  $c$  为任意常数, 那么  $\mathbb{E}[Y|Y > c] = \lambda(-c) = \frac{\phi(c)}{1 - \Phi(c)}$ .

<sup>4</sup> 从这个角度来看, 对于  $X_i$  随机的情况, 一致性和渐近正态性的一个充分条件是  $\mathbb{E}[X_i X_i']$  的非奇异性.

解式

$$\begin{aligned}\frac{\partial \mathbb{E}[Y|X]}{\partial X_j} &= \frac{\partial \mathbb{P}[Y > 0|X]}{\partial X_j} \cdot \mathbb{E}[Y|X, Y > 0] \\ &\quad + \mathbb{P}[Y > 0|X] \cdot \frac{\partial \mathbb{E}[Y|X, Y > 0]}{\partial X_j}\end{aligned}\quad (10.17)$$

对 IMR 求导可得

$$\lambda'(c) = -\lambda(c)[c + \lambda(c)]$$

从而

$$\begin{aligned}\frac{\partial \mathbb{E}[Y|X, Y > 0]}{\partial X_j} &= \beta_j + \beta_j \lambda'(X'\beta/\sigma) \\ &= \beta_j \{1 - \lambda(X'\beta/\sigma)[X'\beta/\sigma + \lambda(X'\beta/\sigma)]\}\end{aligned}\quad (10.18)$$

根据 (10.14) 又可得到

$$\frac{\partial \mathbb{P}[Y > 0|X]}{\partial X_j} = (\beta_j/\sigma)\phi(X'\beta/\sigma)\quad (10.19)$$

将 (10.18) 和 (10.19) 代入到 (10.17) 有

$$\frac{\partial \mathbb{E}[Y|X]}{\partial X_j} = \Phi(X'\beta/\sigma)\beta_j$$

于是 AME 可以通过下式估计而来

$$\left[ n^{-1} \sum_{i=1}^n \Phi(X'_i \hat{\beta}/\hat{\sigma}) \right] \hat{\beta}_j$$

其中  $\hat{\beta}$  和  $\hat{\sigma}$  都是 ML 估计量. 而对于二元变量  $X_j$ , 根据 (10.18) 可知 AME 可估计为

$$n^{-1} \sum_{i=1}^n \{ [\Phi(\hat{w}_{1i}/\hat{\sigma})\hat{w}_{1i} + \hat{\sigma}\phi(\hat{w}_{1i}/\hat{\sigma})] - [\Phi(\hat{w}_{0i}/\hat{\sigma})\hat{w}_{0i} + \hat{\sigma}\phi(\hat{w}_{0i}/\hat{\sigma})] \}$$

其中  $\hat{w}_{1i} = X'_{i(-j)}\hat{\beta}_{-j} + \hat{\beta}_j$ ,  $\hat{w}_{0i} = X'_{i(-j)}\hat{\beta}_{-j}$ , 而下标  $-j$  表示剔除了变量  $X_j$ .

Tobit 潜变量模型中  $e$  的非正态性和异方差性同样会导致 MLE 不一致, Powell (1984) 提出了归并最小绝对偏差 (Censored Least Absolute Deviation, CLAD) 估计量来解决这一问题. 考虑如下归并中值回归

$$Y^* = X'\beta + e$$

$$\text{Med}[e|X] = 0$$

$$Y = \max\{0, Y^*\}$$

这里的  $Y^*$  为潜变量并且满足  $\text{Med}[Y^*|X] = X'\beta$ , 限值因变量  $Y$  归并到 0. 可以证明, 归并中值回归意味着

$$\text{Med}[Y|X] = \max\{0, X'\beta\}$$

Powell (1984) 提出了通过求解

$$\max_{\beta} n^{-1} \sum_{i=1}^n |Y_i - \max\{0, X'_i\beta\}|$$

来估计  $\beta$ . 由于  $|Y - \max\{0, X'\beta\}|$  是关于  $\beta$  的连续函数, 因此 CLAD 估计量在一定正则条件下是一致的. 然而, 由于目标函数并非二阶连续可微, 因此要建立 CLAD 估计量的渐近正态性就十分困难. 当 Tobit 模型设定正确时, CLAD 估计结果应该和 Tobit 估计结果差不多.

### 10.3.2 栅栏模型

根据之前的讨论, 在第 I 类 Tobit 模型中, 一个解释变量对  $\mathbb{P}[Y > 0|X]$  和  $\mathbb{E}[Y|X, Y > 0]$  的边际效应必须有相同的符号, 但我们可能并不想施加这样的约束. 为此, Cragg (1971) 提出了第 I 类 Tobit 模型的扩展, 也即栅栏模型 (hurdle model), 该模型允许分别使用单独的机制来确定参与决策 (即  $Y > 0$  还是  $Y = 0$ ) 与数量决策 ( $Y > 0$  时它的大小).

定义选择指示符  $s$ , 它是一个二值变量, 决定了  $Y = 0$  还是严格为正, 再定义一个具有连续分布的非负潜变量  $W^*$ , 假设  $Y$  由下式生成

$$Y = s \cdot W^*$$

选择指示符  $s$  在观测上等价于  $\mathbb{1}[Y > 0]$ , 而  $W^*$  只有在  $s = 1$  时才能被观测到, 此时  $Y = W^*$ . 进一步假设

$$D[W^*|s, X] = D[W^*|X] \quad (10.20)$$

它表明  $s$  和  $W^*$  以  $X$  为条件相互独立. 换言之, 以可观测的协变量  $X$  为条件, 决定  $s$  和  $W^*$  的机制是相互独立的. 此时

$$\mathbb{E}[Y|X, s] = s\mathbb{E}[W^*|X, s] = s\mathbb{E}[W^*|X]$$

而当  $s = 1$  时有

$$\mathbb{E}[Y|X, Y > 0] = \mathbb{E}[W^*|X]$$

进而

$$\mathbb{E}[Y|X] = \mathbb{P}[s = 1|X]\mathbb{E}[W^*|X]$$

在 Cragg 模型中, 假定条件独立性假设 (10.20) 成立, 并且二值变量  $s$  服从 Probit 模型

$$\mathbb{P}[s = 1|X] = \Phi(X'\gamma)$$

并且还假定潜变量  $W^*$  有一个断尾正态分布. 定义  $W^* = X'\beta + e$ , 其中给定  $X$  时  $e$  服从一个带有断尾点  $-X'\beta$ , 方差为  $\sigma^2$  的正态分布. 因为当  $Y > 0$  时有  $Y = W^*$ , 根据 (10.10) 可知条件概率密度为

$$f(y|X, Y > 0) = \sigma^{-1}[\Phi(X'\beta/\sigma)]^{-1}\phi[(y - X'\beta)/\sigma]$$

进而有

$$f(y|X) = [1 - \Phi(X'\gamma)]^{\mathbb{1}[y=0]}\{\sigma^{-1}\Phi(X'\beta)[\Phi(X'\gamma/\beta)]^{-1}\phi[(y - X'\beta)/\sigma]\}^{\mathbb{1}[y>0]}$$

上式清晰地表明了在没有引入  $s$  和  $W^*$  的情况下是如何设定模型的, 并且当  $\gamma = \beta/\sigma$  时, 断尾正态栅栏 (Truncated Normal Hurdle, TNH) 模型就简约为第 I 类 Tobit 模型.

一旦给定 i.i.d. 随机样本  $\{(X_i, Y_i)\}$ , 则观测值  $i$  的条件对数自然为

$$l_i(\theta) = \mathbb{1}[Y_i = 0] \log[1 - \Phi(X_i' \gamma)] + \mathbb{1}[Y_i > 0] \log \Phi(X_i' \gamma) \\ + \mathbb{1}[Y_i > 0] \{-\log(X_i' \beta / \sigma) + \log \phi[(Y_i - X_i' \beta) / \sigma] - \log \sigma\}$$

由于参数  $\theta = [\gamma', \beta', \sigma']$  允许自由变动, 容易看出 ML 估计量  $\hat{\gamma}$  正好是  $s_i = \mathbb{1}[Y_i > 0]$  对  $X_i$  的 Probit 估计量.

由于条件分布  $D[Y|X, Y > 0]$  在第 I 类 Tobit 模型和 TNH 模型中是相同的, 故而

$$\mathbb{E}[Y|X, Y > 0] = X' \beta + \sigma \lambda(X' \beta / \sigma)$$

所不同的是 TNH 模型允许  $\mathbb{P}[Y > 0|X]$  服从无约束的 Probit 模型, 因此在 TNH 模型中有

$$\mathbb{E}[Y|X] = \Phi(X' \gamma) [X' \beta + \sigma \lambda(X' \beta / \sigma)]$$

于是解释变量  $X_j$  对  $\mathbb{E}[Y|X]$  的边际效应为

$$\frac{\partial \mathbb{E}[Y|X]}{\partial X_j} = \gamma_j \phi(X' \gamma) [X' \beta + \sigma \lambda(X' \beta / \sigma)] + \beta_j \Phi(X' \gamma) \theta(X' \beta / \sigma)$$

其中  $\theta(z) = 1 - \lambda(z)[z - \lambda(z)]$ . 通过 MLE 得到估计量后, 就可以根据上式得出估计 AME 的表达式.

## 10.4 样本选择问题

所谓的选择性样本通常用来描述一个不是从总体中随机抽取得到的样本, 上述提到的数据断尾也是样本选择 (sample selection) 问题的一个特例, 而另一个样本选择问题则是从属断尾 (incidental truncation). 当存在样本选择问题时, 除非在特定类型下, 否则样本无法代表总体, 使用 OLS 就会存在偏差.

### 10.4.1 一致性 OLS 估计

对于总体线性回归模型

$$Y = X' \beta + e \\ \mathbb{E}[e|X] = 0 \quad (10.21)$$

如果数据是完全随机缺失的 (Missing Completely At Random, MCAR), 也即数据缺失的原因在统计上独立于影响  $Y$  的可观测因素与不可观测因素, 那么缺失数据不会在统计上造成任何影响. 从实际上来看, 我们仍可以假定数据是从总体中随机抽样得到的.

另一方面, 如果样本仅由外生解释变量决定, 那么此时就出现外生样本选择问题, 此时也很容易得到使得 OLS 为一致 (甚至无偏) 估计的条件. 在总体中抽取随机样本可得

$$Y_i = X_i' \beta + e_i \quad (10.22)$$

显然, 如果对每个个体  $i$  都能观测到  $Y_i$  和  $X_{ji}$ , 那么就可以使用 OLS. 然而出于某些原因, 某个个体  $i$  的  $Y_i$  或自变量无法观测到, 但至少可以观测到某些个体  $i$  的变量集的全部信息.

对于每一个个体  $i$ , 定义一个选择指标  $s_i$ , 如果观测到了  $[Y_i, X_i]$  的全部信息, 则  $s_i = 1$ , 否则  $s_i = 0$ . 从定义上看,  $s_i = 1$  表示我们将用到这个观测, 而  $s_i = 0$  就表示用不到这一观测. 我们感兴趣的是, OLS 估计量在选择性样本 (即使用  $s_i = 1$  的观测) 上的统计性质. 考虑估计以下方程

$$s_i Y_i = s_i X_i' \beta + s_i e_i \quad (10.23)$$

当  $s_i = 1$  时, 上式就变为 (10.22); 当  $s_i = 0$  时, 上式两端为 0, 这显然没有告诉我们关于  $\beta$  的任何信息. 将  $s_i Y_i$  对  $s_i X_i$  回归, 等同于利用  $s_i = 1$  的观测将  $Y_i$  对  $X_i$  回归. 因此, 可以通过研究一个随机样本来研究 (10.23) 以了解  $\hat{\beta}$  的统计性质. 此时有

$$\hat{\beta} = \beta + \left( n^{-1} \sum_{i=1}^n s_i X_i X_i' \right)^{-1} \left( n^{-1} \sum_{i=1}^n s_i X_i e_i \right)$$

一旦我们假定

$$\mathbb{E}[s_i X_i e_i] = 0 \quad (10.24)$$

以及相关矩条件成立, 则 OLS 估计量是一致的. 关于条件 (10.24) 的一个充分条件是

$$\mathbb{E}[e_i | X_i, s_i] = 0 \quad (10.25)$$

它允许选择指示符  $s$  与可观测变量  $X$  相关, 而与随机误差项  $e$  无关, 这称为外生样本选择 (exogenous sample selection).

一旦条件 (10.21) 成立, 且  $s$  是  $X$  的一个确定性函数, 那么条件 (10.25) 自然也成立. 换言之, 如果存在非随机函数  $h$  使得  $s = h(X)$ , 就产生了外生性抽样并排除了  $s$  受不可观测效应影响的机制. 在外生样本选择的情况下, 条件 (10.25) 意味着

$$\mathbb{E}[Y | X, s] = \mathbb{E}[Y | X] = X' \beta$$

这就表明无论是  $Y$  还是  $X$  发生了数据缺失, 在选择性样本上的 OLS 仍是一致的. 关于 OLS 估计量一致性的结论还可以推广到 2SLS 上, 也即当  $\mathbb{E}[e_i | Z_i, s_i] = 0$  时, 2SLS 通常也是一致的.

## 10.4.2 从属断尾

正如之前提到的那样, 样本选择问题的主要形式为从属断尾. 对于总体线性回归模型

$$Y = X' \beta + e$$

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

假设总能观测到解释变量  $X$ , 但只能观测到总体  $Y$  的一个子集, 并且是否能观测到  $Y$  不取决于  $Y$  的结果. 此时因变量  $Y$  的断尾就是从属的, 它取决于另一个变量<sup>5</sup>. 事实上, 前面提到的 TNH 模型已应用在了缺失数据问题, 但它要求参与决策和数量决策过程条件独立, 并且  $Y$  同样是大于等于 0 的, 而非直接缺失.

利用选择指示符  $s_i$ , 当且仅当  $s_i = 1$  时可以完全观测到  $[Y_i, X_i]$ , 而当  $s_i = 0$  时  $Y_i$  缺失, 于

<sup>5</sup>例如在工资方程中, 如果一个人有工作, 那么确实可以观测到工资, 然而对于失业人群而言, 工资就是缺失值.

是  $Y$  在受选择样本中的条件均值为

$$\mathbb{E}[Y|X, s = 1] = X'\beta + \mathbb{E}[e|X, s = 1] \quad (10.26)$$

由于  $s$  取决于其它变量影响, 故而可以设<sup>6</sup>

$$s = \mathbb{1}[Z'\gamma + u] \quad (10.27)$$

进一步有

$$\mathbb{E}[e|X, s = 1] = \mathbb{E}[e|u > -Z'\gamma]$$

假设  $e$  独立于  $X$ , 并且  $e$  在  $u$  上的线性投影为  $e = \rho u + \varepsilon$ , 并且  $u$  和  $\varepsilon$  独立. 因此方程 (10.26) 变为

$$\mathbb{E}[Y|X, s = 1] = X'\beta + \rho \mathbb{E}[u|u > -Z'\gamma] = X'\beta + \rho g(Z'\gamma)$$

如果  $u \sim N(0, 1)$ , 那么  $g(Z'\gamma) = \lambda(Z'\gamma)$ , 从而

$$\mathbb{E}[Y|X, s = 1] = X'\beta + \rho \lambda(Z'\gamma) \quad (10.28)$$

因此只要  $\rho \neq 0$ , 根据受选择样本获得的 OLS 估计量就不是一致的. 由此可见, 从属断尾算是遗漏变量问题的一个特例.

由于  $\gamma$  是未知的, 我们无法对每个  $i$  计算  $\lambda(Z'_i\gamma)$ , 但根据 (10.27) 和  $u \sim N(0, 1)$  可知

$$\mathbb{P}[s = 1|Z] = \Phi(Z'\gamma)$$

因此可以通过  $s$  对  $Z$  的 Probit 模型来估计  $\gamma$ , 然后再来估计  $\beta$ , 这称为 Heckit 方法 (Heckman, 1976, 1979), 样本选择模型又称为第 II 类 Tobit 模型.

具体而言, Heckit 方法也是两步法估计:

- 利用  $n$  个观测值, 实施  $s_i$  对  $Z_i$  的 Probit 模型并得到  $\hat{\gamma}$ , 然后对每个  $i$  计算  $\text{IMR } \hat{\lambda}_i = \lambda(Z'_i\hat{\gamma})$ .
- 利用受选择样本做  $Y_i$  对  $X_i$  和  $\lambda_i$  的 OLS 回归:  $Y_i = X'_i\beta + \rho\hat{\lambda}_i + e_i$ .

由此得到的 OLS 估计量在标准条件下是一致和渐近正态的, 根据生成回归元的内容, 基于通常的  $T$  检验统计量就能检验  $\mathbb{H}_0: \rho = 0$ . Heckit 方法本质上也是 CF 方法, 因此当  $\rho \neq 0$  时, Heckit 方法得到的 OLS 估计量的标准误存在误差, 此时应该使用 Bootstrap 法对其调整.

为了应用 CMLE, 还需要假定  $[e_1, u]$  服从联合正态分布

$$\begin{bmatrix} e \\ u \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma_{21} \\ \sigma_{12} & 1 \end{bmatrix} \right)$$

可以证明, 此时  $(s, Y)$  的联合概率密度为

$$f(s, y|X_i, Z_i) = [1 - \Phi(Z'_i\gamma)]^{1-s} \left\{ \Phi \left[ \frac{\sigma^2 Z'_i\gamma + \sigma_{21}(y - X'_i\beta)}{\sigma \sqrt{\sigma^2 - \sigma_{21}^2}} \right] \frac{1}{\sigma} \phi \left( \frac{y - X'_i\beta}{\sigma} \right) \right\}^s$$

由此可求得对数似然以及 ML 估计量, 在一定正则条件下, 它是渐近有效的一致估计量.

<sup>6</sup>为了避免高度共线性, 通常要求  $Z$  至少包含一个不在  $X$  中的排他性约束变量, 该变量只影响选择方程而不影响结果方程.

如果在样本选择背景下还存在内生性问题, 假设此时的回归模型为

$$Y = X_1' \beta_1 + \beta_2 X_2 + e \quad (10.29)$$

其中  $Y$  只能在  $s = 1$  时才能观测到,  $X_1$  为外生解释变量,  $X_2$  为内生解释变量, 此时仍可以考虑 CF 方法:

- 实施  $s_i$  对外生解释变量、影响选择方程的排他性约束、以及  $X_2$  的工具变量的 Probit 回归, 计算得到 IMR 后代入到方程 (10.29).
- 对更新后的方程实施标准的 2SLS 回归.

**注** 为了使结果令人信服, 至少应该有两个排他性约束, 一个用于选择方程, 而另一个用于内生解释变量.



## 参考文献

- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, 138(1), 1–35.
- Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41(6), 997–1016.
- Amemiya, T. (1985). *Advanced econometrics*. Cambridge: Harvard University Press.
- Amemiya, T., & MaCurdy, T. E. (1986). Instrumental-variable estimation of an error-components model. *Econometrica*, 54(4), 869–880.
- Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1), 47–82.
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.
- Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4), 431–434.
- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2), 277–297.
- Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68(1), 29–51.
- Athreya, K. B., & Lahiri, S. N. (2006). *Measure theory and probability theory*. New York: Springer.
- Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5), 829–844.
- Davidson, J. (2020). *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford University Press.
- Dube, O., & Harish, S. (2020). Queens. *Journal of Political Economy*, 128(7), 2579–2652.
- Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5), 1739–1774.
- Durrett, R. (2019). *Probability: Theory and examples*. Cambridge: Cambridge University Press.
- Fatás, A., & Mihov, I. (2001). Government size and automatic stabilizers: International and intranational evidence. *Journal of International Economics*, 55(1), 3–28.
- Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3), 681–700.
- Greene, W. H. (2017). *Econometric analysis*. New York: Pearson.



- Hansen, B. E. (2022a). *Econometrics*. Princeton: Princeton University Press.
- Hansen, B. E. (2022b). A modern gauss–markov theorem. *Econometrica*, 90(3), 1283–1294.
- Hansen, B. E. (2022c). *Probability and statistics for economists*. Princeton: Princeton University Press.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029–1054.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.
- Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, 49(6), 1377–1398.
- Hayashi, F. (2000). *Econometrics*. Princeton: Princeton University Press.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. 5(4), 475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Hong, Y. (2017). *Probability and statistics for economists*. Singapore: World Scientific.
- Hong, Y. (2020). *Foundations of modern econometrics: A unified approach*. Singapore: World Scientific.
- Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40(2), 633–643.
- Kinal, T. W. (1980). The existence of moments of k-class estimators. *Econometrica*, 48(1), 241–249.
- Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1), 97–126.
- Klenke, A. (2013). *Probability theory: A comprehensive course*. New York: Springer Science & Business Media.
- McDonald, J. F., & Moffitt, R. A. (1980). The uses of tobit analysis. *Review of Economics and Statistics*, 62(2), 318–321.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99–135.
- Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. MacFadden (Eds.), *Handbook of econometrics* (pp. 2111–2245, Vol. 4). Amsterdam: North Holland.
- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49(6), 1417–1426.
- Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1), 221–247.
- Portnoy, S. (2022). Linearity of unbiased linear model estimators. *American Statistician*, 76(4), 372–375.
- Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3), 303–325.

- Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3), 347–366.
- Rudin, W. (1976). *Principles of mathematical analysis*. New York: McGraw-hill.
- Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3), 393–415.
- Staiger, D. O., & Stock, J. H. (1994). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.
- Stock, J. H., & Watson, M. W. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, 76(1), 155–174.
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models* (pp. 80–108). Cambridge: Cambridge University Press.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.
- Wooldridge, J. M. (1995). Score diagnostics for linear models estimated by two stage least squares. In G. S. Maddala, P. C. B. Phillips, & T. N. Srinivasan (Eds.), *Advances in econometrics and quantitative economics* (pp. 66–87). Oxford: Blackwell.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2), 420–445.
- Wooldridge, J. M. (2019). *Introductory econometrics: A modern approach*. Boston: Cengage Learning.
- Zehna, P. W. (1966). Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics*, 37(3), 744.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298), 348–368.