

第 10 章 工具变量, 2SLS 与 GMM

10.1 解释变量与扰动项相关的例子

例 农产品市场均衡模型

$$\begin{cases} q_t^d = \alpha_0 + \alpha_1 p_t + u_t & (\text{需求}) \\ q_t^s = \beta_0 + \beta_1 p_t + v_t & (\text{供给}) \\ q_t^d = q_t^s & (\text{均衡}) \end{cases}$$

令 $q_t \equiv q_t^d = q_t^s$, 可得

$$\begin{cases} q_t = \alpha_0 + \alpha_1 p_t + u_t \\ q_t = \beta_0 + \beta_1 p_t + v_t \end{cases}$$

两个方程中的被解释变量与解释变量完全一样。

如直接作回归 $q_t \xrightarrow{\text{OLS}} p_t$, 估计的是需求函数还是供给函数?

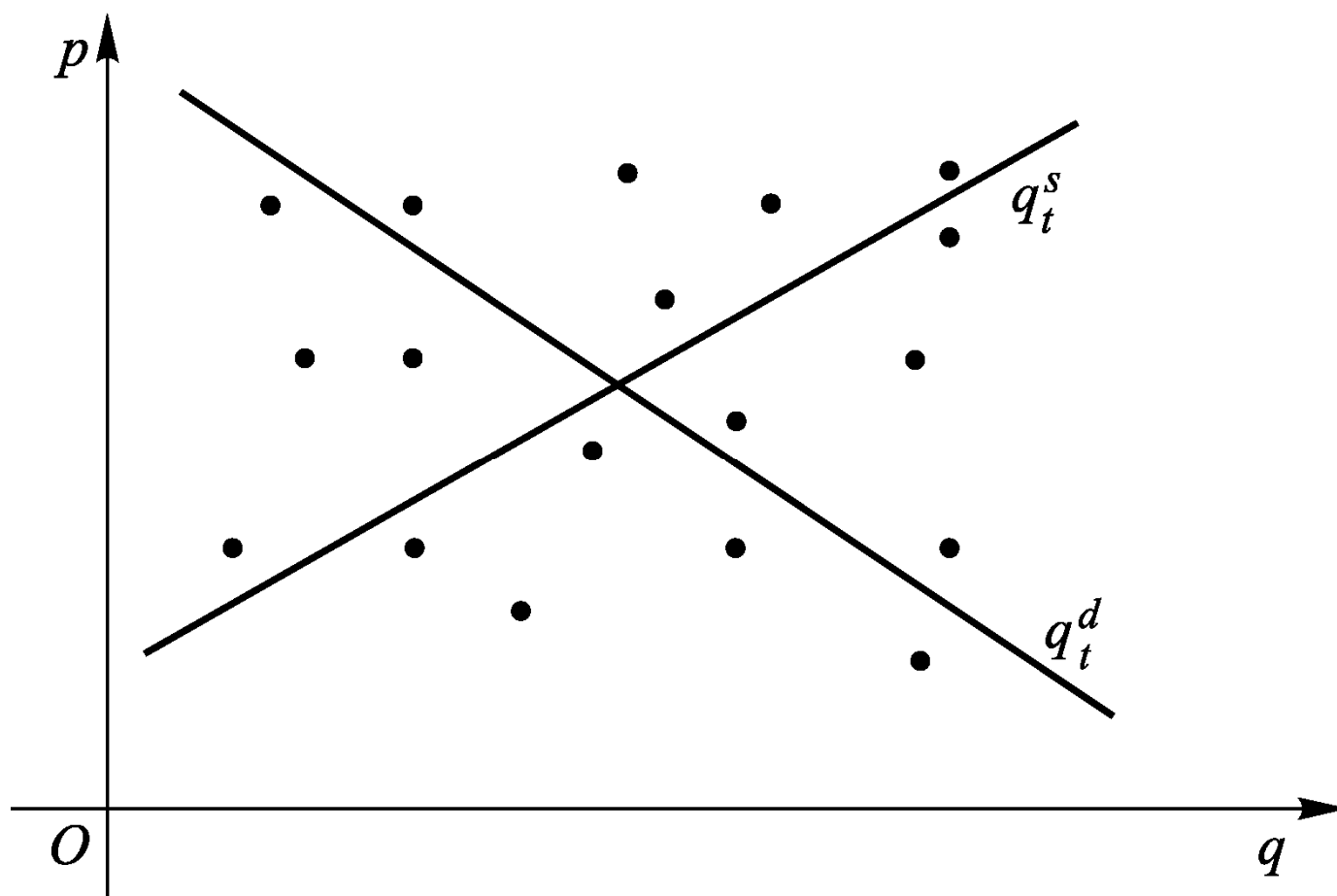


图 10.1 需求与供给决定市场均衡

把线性方程组中的 (p_t, q_t) 看成是未知数(内生变量), 把 (u_t, v_t) 看作已知, 可求解 (p_t, q_t) 为 (u_t, v_t) 的函数:

$$\begin{cases} p_t = p_t(u_t, v_t) = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_t - u_t}{\alpha_1 - \beta_1} \\ q_t = q_t(u_t, v_t) = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_t - \beta_1 u_t}{\alpha_1 - \beta_1} \end{cases}$$

由于 p_t 为 (u_t, v_t) 的函数, 故 $\text{Cov}(p_t, u_t) \neq 0$, $\text{Cov}(p_t, v_t) \neq 0$ 。

OLS 估计值 $\hat{\alpha}_1, \hat{\beta}_1$ 不是 α_1, β_1 的一致估计量。

称这种偏差为“联立方程偏差”(simultaneity bias)或“内生变量偏差”(endogeneity bias)。

如能将内生变量分成两部分，一部分与扰动项相关，另一部分与扰动项不相关，可用与扰动项不相关的那部分得到一致估计。

这种分离常借助另一“工具变量”来实现。

假设在图 10.1 中，存在某个因素(变量)使得供给曲线经常移动，而需求曲线基本不动，则可估计需求曲线，参见图 10.2。

这个使得供给曲线移动的变量就是工具变量。

假设供给方程的扰动项可分解为两部分，即可观测的气温 x_t 与不可观测的其他因素：

$$q_t^s = \beta_0 + \beta_1 p_t + \beta_2 x_t + v_t$$

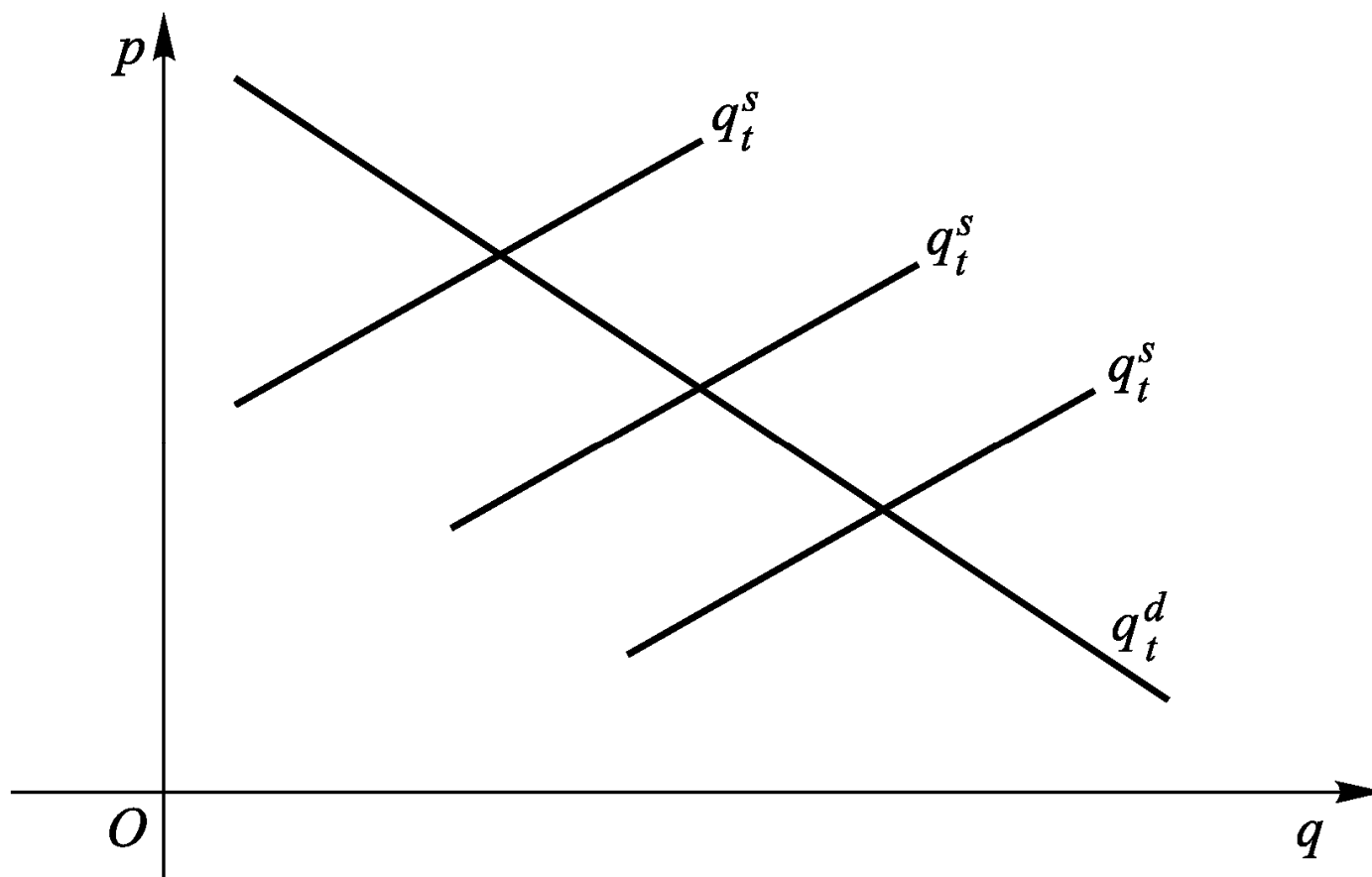


图 10.2 稳定的需求与变动的供给

假定气温 x_t 是前定变量，与两个扰动项都不相关，即 $\text{Cov}(x_t, u_t) = 0$ ， $\text{Cov}(x_t, v_t) = 0$ 。

由于气温 x_t 的变化使得供给函数 q_t^s 沿着需求函数 q_t^d 移动，故可估计需求函数 q_t^d 。

此时，称 x_t 为“工具变量” (Instrumental Variable，简记 IV)。

在回归方程中(此处为需求方程)，一个有效(valid)的工具变量应满足以下两个条件。

- (i) 相关性：工具变量与内生解释变量相关，即 $\text{Cov}(x_t, p_t) \neq 0$ 。
- (ii) 外生性：工具变量与扰动项不相关，即 $\text{Cov}(x_t, u_t) = 0$ 。

工具变量的外生性也称“排他性约束”(exclusion restriction), 因为外生性意味着, 工具变量影响被解释变量的唯一渠道是通过与其相关的内生解释变量, 它排除了所有其他的可能影响渠道。

在本例中, 气温 x_t 满足这两个条件。

(i) 相关性: 从联立方程组可解出 $p_t = p_t(x_t, u_t, v_t)$, 故 $\text{Cov}(x_t, p_t) \neq 0$ 。

(ii) 外生性: 因为气温 x_t 是前定变量, 故 $\text{Cov}(x_t, u_t) = 0$ 。

利用工具变量的这两个性质, 可得到对 α_1 的一致估计。

同时对需求方程 $q_t = \alpha_0 + \alpha_1 p_t + u_t$ 两边求与 x_t 的协方差：

$$\begin{aligned}\text{Cov}(q_t, x_t) &= \text{Cov}(\alpha_0 + \alpha_1 p_t + u_t, x_t) \\ &= \alpha_1 \text{Cov}(p_t, x_t) + \underbrace{\text{Cov}(u_t, x_t)}_{=0} = \alpha_1 \text{Cov}(p_t, x_t)\end{aligned}$$

根据工具变量的相关性， $\text{Cov}(p_t, x_t) \neq 0$ ，可把上式两边同除以 $\text{Cov}(p_t, x_t)$ ：

$$\alpha_1 = \frac{\text{Cov}(q_t, x_t)}{\text{Cov}(p_t, x_t)}$$

使用对应的样本值，可得一致的“工具变量估计量”
(Instrumental Variable Estimator)：

$$\hat{\alpha}_{1,IV} = \frac{\widehat{\text{Cov}(q_t, x_t)}}{\widehat{\text{Cov}(p_t, x_t)}} \xrightarrow{p} \frac{\text{Cov}(q_t, x_t)}{\text{Cov}(p_t, x_t)} = \alpha_1$$

如果工具变量与内生变量无关, $\text{Cov}(x_t, p_t) = 0$, 则无法定义工具变量法。

如果工具变量与内生变量的相关性很弱, $\text{Cov}(x_t, p_t) \approx 0$, 会导致估计量 $\hat{\alpha}_{1,IV}$ 的方差变得很大, 称为“弱工具变量问题”。

传统的工具变量法通过“二阶段最小二乘法”(Two Stage Least Square, 简记 2SLS 或 TSLS)来实现。

第一阶段回归：用内生解释变量对工具变量回归，即 $p_t \xrightarrow{\text{OLS}} x_t$ ，得到拟合值 \hat{p}_t 。

第二阶段回归：用被解释变量对第一阶段回归的拟合值进行回归，即 $q_t \xrightarrow{\text{OLS}} \hat{p}_t$ 。

为什么这样做能得到好结果？把需求方程 $q_t = \alpha_0 + \alpha_1 p_t + u_t$ 分解

$$q_t = \alpha_0 + \alpha_1 \hat{p}_t + \underbrace{[u_t + \alpha_1(p_t - \hat{p}_t)]}_{\equiv \varepsilon_t}$$

命题 在第二阶段回归中， \hat{p}_t 与新扰动项 $\varepsilon_t \equiv u_t + \alpha_1(p_t - \hat{p}_t)$ 不相关。

证明：由于 $\varepsilon_t \equiv u_t + \alpha_1(p_t - \hat{p}_t)$ ，故

$$\text{Cov}(\hat{p}_t, \varepsilon_t) = \text{Cov}(\hat{p}_t, u_t) + \alpha_1 \text{Cov}(\hat{p}_t, p_t - \hat{p}_t)$$

首先，由于 \hat{p}_t 是 x_t 的线性函数(\hat{p}_t 为第一阶段回归的拟合值)，而 $\text{Cov}(x_t, u_t) = 0$ (工具变量的外生性)，故上式右边的第一项 $\text{Cov}(\hat{p}_t, u_t) = 0$ 。

其次，由于在第一阶段回归中，拟合值 \hat{p}_t 与残差 $p_t - \hat{p}_t$ 正交(OLS 的正交性)，故上式右边的第二项 $\text{Cov}(\hat{p}_t, p_t - \hat{p}_t) = 0$ 。

由于 \hat{p}_t 与 ε_t 不相关，故 2SLS 一致。

例 宏观经济模型中的消费函数

$$\begin{cases} C_t = \alpha_0 + \alpha_1 Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases}$$

其中， Y_t, C_t, I_t, G_t, X_t 分别代表国民收入、总消费、总投资、政府净支出与净出口。

第一个方程为消费方程，第二个方程为国民收入恒等式。

如果单独对消费方程进行 OLS 估计，将不一致。

例 解释变量测量误差(measurement error or errors-in-variables)。

假设真实模型为

$$y = \alpha + \beta x^* + \varepsilon, \quad \text{Cov}(x^*, \varepsilon) = 0, \quad \beta \neq 0$$

但 x^* 无法精确观测，而只能观测到 x ，二者满足如下关系：

$$x = x^* + u, \quad \text{Cov}(x^*, u) = 0, \quad \text{Cov}(u, \varepsilon) = 0$$

其中，测量误差 u 与被测量变量 x^* 不相关，也与扰动项 ε 不相关。

代入可得：

$$y = \alpha + \beta x + (\varepsilon - \beta u)$$

新扰动项 $(\varepsilon - \beta u)$ 与解释变量 x 存在相关性:

$$\begin{aligned}\text{Cov}(x, \varepsilon - \beta u) &= \text{Cov}(x^* + u, \varepsilon - \beta u) \\ &= \underbrace{\text{Cov}(x^*, \varepsilon)}_{=0} - \beta \underbrace{\text{Cov}(x^*, u)}_{=0} + \underbrace{\text{Cov}(u, \varepsilon)}_{=0} - \beta \text{Cov}(u, u) \\ &= -\beta \text{Var}(u) \neq 0\end{aligned}$$

故 OLS 不一致, 称为“测量误差偏差”(measurement error bias)。

可确定此偏差的方向:

$$\begin{aligned}
\hat{\beta} &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \\
&\xrightarrow{p} \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} = \frac{\text{Cov}(x_i^* + u, \alpha + \beta x_i^* + \varepsilon)}{\text{Var}(x_i^* + u)} \\
&= \frac{\beta \text{Var}(x_i^*)}{\text{Var}(x_i^*) + \text{Var}(u)} = \beta \cdot \frac{1}{1 + \left(\sigma_u^2 / \sigma_{x^*}^2 \right)}
\end{aligned}$$

由于 σ_u^2 与 $\sigma_{x^*}^2$ 一定为正，故 $0 < \frac{1}{1 + \left(\sigma_u^2 / \sigma_{x^*}^2 \right)} < 1$ 。

无论真实参数 β 大于或小于 0, 此偏差总是使得 $\hat{\beta}$ 的绝对值变小而趋向于 0。

故也称为“衰减偏差”(attenuation bias)或“向 0 衰减”(attenuation toward zero)。

相对于 x_i^* 的方差 $\sigma_{x^*}^2$, 如果测量误差 u_i 的方差 σ_u^2 越大, 则 $(\sigma_u^2 / \sigma_{x^*}^2)$ 越大, $\frac{1}{1 + (\sigma_u^2 / \sigma_{x^*}^2)} \rightarrow 0$, 则向 0 衰减的偏差越严重。

如果被解释变量存在测量误差，后果却不严重。真实模型：

$$y^* = \beta x + \varepsilon, \quad \text{Cov}(x, \varepsilon) = 0, \quad \beta \neq 0$$

y^* 无法精确观测，只能观测到 y ，二者满足如下关系：

$$y = y^* + v$$

其中， v 为测量误差。代入可得：

$$y = \beta x + (\varepsilon + v)$$

只要 $\text{Cov}(x, v) = 0$ ，则 OLS 一致，但可能增大扰动项的方差。

10.2 工具变量法作为一种矩估计

1. 矩估计 (Method of Moments, MM)

假设随机变量 $x \sim N(\mu, \sigma^2)$ ，其中 μ, σ^2 为待估参数。

因为有两个待估参数，故使用两个总体矩条件 (population moment conditions):

一阶矩: $E(x) = \mu$

二阶矩: $E(x^2) = \text{Var}(x) + [E(x)]^2 = \sigma^2 + \mu^2$

用对应的样本矩 (sample moments) 来替代总体矩条件可得:

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n x_i = \hat{\mu} \\ \frac{1}{n} \sum_{i=1}^n x_i^2 = \hat{\mu}^2 + \hat{\sigma}^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

其中， $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ 为样本均值。推导中用到，
 $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ 。

任何随机向量 \mathbf{x} 的函数 $f(\mathbf{x})$ 的期望 $E[f(\mathbf{x})]$ 都称为“总体矩”。

OLS 也是矩估计。利用解释变量与扰动项的正交性：

$$E[\mathbf{x}_i \varepsilon_i] = \mathbf{0} \Rightarrow E[\mathbf{x}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})] = \mathbf{0} \quad (\text{代入 } \varepsilon_i = y_i - \mathbf{x}_i' \boldsymbol{\beta})$$

$$\Rightarrow E(\mathbf{x}_i y_i) = E(\mathbf{x}_i \mathbf{x}_i') \boldsymbol{\beta} \quad (\text{展开、移项})$$

$$\Rightarrow \boldsymbol{\beta} = [E(\mathbf{x}_i \mathbf{x}_i')]^{-1} E(\mathbf{x}_i y_i) \quad (\text{假设 } [E(\mathbf{x}_i \mathbf{x}_i')]^{-1} \text{ 存在, 求解 } \boldsymbol{\beta})$$

以样本矩替代总体矩，可得矩估计：

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}}$$

2. 工具变量法作为一种矩估计

假设回归模型为

$$y_i = \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{i, K-1} + \beta_K x_{iK} + \varepsilon_i$$

只有最后一个解释变量 x_{iK} 为内生变量，即 $\text{Cov}(x_{iK}, \varepsilon_i) \neq 0$ ，故 OLS 不一致。

假设有一个有效工具变量 w 满足 $\text{Cov}(x_{iK}, w_i) \neq 0$ (相关性)，以及 $\text{Cov}(w_i, \varepsilon_i) = 0$ (外生性)。

由于 x_1, \dots, x_{K-1} 不是内生变量，可把自身作为自己的工具变量(满足工具变量法的两个条件)。

记解释向量 $\mathbf{x}_i \equiv (x_{i1} \cdots x_{i,K-1} \ x_{iK})'$ ，则原模型为 $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ 。

记工具向量 $\mathbf{z}_i \equiv (z_{i1} \cdots z_{i,K-1} \ z_{iK})' \equiv (x_{i1} \cdots x_{i,K-1} \ w_i)'$ 。

定义 $\mathbf{g}_i \equiv \mathbf{z}_i \varepsilon_i$ 。由于工具向量与扰动项正交，故 $E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 为“总体矩条件”或“正交条件” (orthogonality condition)。

$$E(\mathbf{z}_i \varepsilon_i) = \mathbf{0} \quad \Rightarrow \quad E[\mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})] = \mathbf{0} \quad (\text{代入 } \varepsilon_i = y_i - \mathbf{x}_i' \boldsymbol{\beta})$$

$$\Rightarrow E(\mathbf{z}_i y_i) = [E(\mathbf{z}_i \mathbf{x}_i')] \boldsymbol{\beta} \quad (\text{展开、移项})$$

$$\Rightarrow \boldsymbol{\beta} = [E(\mathbf{z}_i \mathbf{x}_i')]^{-1} E(\mathbf{z}_i y_i) \quad (\text{假设 } [E(\mathbf{z}_i \mathbf{x}_i')]^{-1} \text{ 存在})$$

以样本矩代替上式中的总体矩，可得 IV 估计量，

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i y_i \right) = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}$$

其中， $\mathbf{Z} \equiv (\mathbf{z}_1 \cdots \mathbf{z}_{n-1} \mathbf{z}_n)'$ 。

OLS 也是一种工具变量法。

如果 \mathbf{x}_i 全部是前定变量，可将自己作为工具变量，即 $\mathbf{z}_i = \mathbf{x}_i$ ， $\mathbf{Z} = \mathbf{X}$ 。因此，

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}}$$

命题 如果秩条件“ $\text{rank}[\text{E}(\mathbf{z}_i \mathbf{x}_i')] = K$ ”成立(即方阵 $\text{E}(\mathbf{z}_i \mathbf{x}_i')$ 满秩), 则在一定的正则条件下, $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 是 $\boldsymbol{\beta}$ 的一致估计, 且 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 服从渐近正态分布。

证明: 因为 $\text{rank}[\text{E}(\mathbf{z}_i \mathbf{x}_i')] = K$, 故 $[\text{E}(\mathbf{z}_i \mathbf{x}_i')]^{-1}$ 存在。

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta} &= (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\boldsymbol{\varepsilon} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i \right) = \mathbf{S}_{\text{ZX}}^{-1} \bar{\mathbf{g}} \xrightarrow{p} [\text{E}(\mathbf{z}_i \mathbf{x}_i')]^{-1} \underbrace{\text{E}(\mathbf{g}_i)}_{=0} = \mathbf{0}\end{aligned}$$

其中, $\mathbf{S}_{\text{ZX}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}_i'$, $\bar{\mathbf{g}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i$ 。

根据与第 5 章类似的推导可知：

$$\sqrt{n} \bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$$

其中， $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}_i') = E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i')$ 。工具变量估计量 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 渐近地服从正态分布：

$$\sqrt{n} (\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\boldsymbol{\beta}}_{\text{IV}}))$$

其中，渐近方差矩阵 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = [E(\mathbf{z}_i \mathbf{x}_i')]^{-1} \mathbf{S} [E(\mathbf{x}_i \mathbf{z}_i')]^{-1}$ 。

秩条件 $\text{rank}[E(\mathbf{z}_i \mathbf{x}_i')] = K$ 意味着，工具变量 w_i 与解释变量 x_i 相关。

如果不相关，则秩条件无法满足。

以一元回归为例，此时， $K = 2$ ， $\mathbf{x}_i = (1 \ x_i)'$ ， $\mathbf{z}_i = (1 \ w_i)'$ ，则

$$\mathbf{E}(\mathbf{z}_i \mathbf{x}_i') = \mathbf{E} \left[\begin{pmatrix} 1 \\ w_i \end{pmatrix} (1 \ x_i) \right] = \mathbf{E} \begin{bmatrix} 1 & x_i \\ w_i & w_i x_i \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{E}(x_i) \\ \mathbf{E}(w_i) & \mathbf{E}(w_i x_i) \end{bmatrix}$$

$$\text{rank}[\mathbf{E}(\mathbf{z}_i \mathbf{x}_i')] = K = 2 \quad \Leftrightarrow \quad \text{行列式} \begin{vmatrix} 1 & \mathbf{E}(x_i) \\ \mathbf{E}(w_i) & \mathbf{E}(w_i x_i) \end{vmatrix} \neq 0$$

$$\Leftrightarrow \mathbf{E}(w_i x_i) - \mathbf{E}(w_i) \mathbf{E}(x_i) \neq 0$$

$$\Leftrightarrow \text{Cov}(w_i, x_i) \neq 0, \text{ 即 } w_i \text{ 与 } x_i \text{ 相关。}$$

阶条件：满足秩条件的必要条件是 z_i 中至少包含 K 个变量，即不在方程中出现的工具变量个数不少于方程中内生解释变量的个数。称此条件为“阶条件” (order condition)。

根据是否满足阶条件可分为三种情况：

- (1) 不可识别(unidentified)：工具变量个数小于内生解释变量个数；
- (2) 恰好识别(just or exactly identified)：工具变量个数等于内生解释变量个数；
- (3) 过度识别(overidentified)：工具变量个数大于内生解释变量个数。

以上介绍的工具变量法仅适用于“恰好识别”的情形。

在“过度识别”的情况下， $\mathbf{Z}'\mathbf{X}$ 不是方阵， $(\mathbf{Z}'\mathbf{X})^{-1}$ 不存在，也就无法定义工具变量估计量 $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 。

解决方法之一是扔掉“多余”的工具变量，但被扔掉的工具变量包含有用的信息。

有效率的做法是 2SLS。

10.3 二阶段最小二乘法

多个工具变量的线性组合仍然是工具变量，因为仍满足工具变量的两个条件(相关性与外生性)。

如果生成工具变量的 K 个线性组合，则又回到恰好识别的情形。

在球型扰动项的假定下，由 2SLS 所提供的工具变量线性组合是所有线性组合中最渐近有效的。

第一阶段(分离出内生变量的外生部分)

将解释变量 $\mathbf{x}_1, \dots, \mathbf{x}_K$ 分别对所有 L 个工具变量 $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_L\}$ 作 OLS 回归, 其中第 k 个解释变量 $\mathbf{x}_k \equiv (x_{1k} \cdots x_{nk})'_{n \times 1}$ 。

得到拟合值:

$$\hat{\mathbf{x}}_1 = \mathbf{P}\mathbf{x}_1, \quad \hat{\mathbf{x}}_2 = \mathbf{P}\mathbf{x}_2, \quad \dots, \quad \hat{\mathbf{x}}_K = \mathbf{P}\mathbf{x}_K$$

其中, $\mathbf{P} \equiv \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ 为对应于 \mathbf{Z} 的投影矩阵。定义

$$\hat{\mathbf{X}} \equiv (\hat{\mathbf{x}}_1 \quad \hat{\mathbf{x}}_2 \quad \cdots \quad \hat{\mathbf{x}}_K) = \mathbf{P}(\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_K) = \mathbf{P}\mathbf{X} = \mathbf{Z}\left[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]$$

第二阶段(使用此外生部分进行回归)

由于 \hat{X} 是 $\{z_1, z_2, \dots, z_L\}$ 的线性组合(参见第一阶段回归), 故 \hat{X} 恰好包含 K 个工具变量。

使用 \hat{X} 为工具变量对原模型 $y = X\beta + \varepsilon$ 进行工具变量法估计:

$$\hat{\beta}_{IV} = (\hat{X}'X)^{-1} \hat{X}'y = (\hat{X}'\hat{X})^{-1} \hat{X}'y$$

因为 $\hat{X}'\hat{X} = (PX)'(PX) = X'P'PX = X'P'X = \hat{X}'X$, 投影矩阵 P 为对称幂等矩阵, $P' = P$, $P^2 = P$ 。

可将 $\hat{\beta}_{IV}$ 视为把 y 对 \hat{X} 进行 OLS 回归而得到, 故名“二阶段最小二乘法”。

第二阶段回归所得到的残差为 $\mathbf{e}_2 \equiv \mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_{2\text{SLS}}$ ，而原方程残差却是 $\mathbf{e} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{2\text{SLS}}$ 。

执行 2SLS 最好直接使用 Stata 的命令。

在同方差的假设下， $\hat{\boldsymbol{\beta}}_{\text{IV}}$ 的协方差矩阵估计量为 $\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}})} = s^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$ ，其中 $s^2 \equiv \frac{\mathbf{e}'\mathbf{e}}{n-K}$ 。

在异方差的情况下，应使用稳健的协方差矩阵估计量，即 $\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}})} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \left(\sum_{i=1}^n e_i^2 \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i' \right) (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$ 。

将 $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ 代入方程，可得 2SLS 的最终表达式：

$$\hat{\beta}_{2\text{SLS}} = (X'PX)^{-1} X'Py = \left[X'Z(Z'Z)^{-1} Z'X \right]^{-1} X'Z(Z'Z)^{-1} Z'y$$

10.4 有关工具变量的检验

1. 不可识别检验

使用工具变量法的前提之一是秩条件成立，即 $\text{rank}[E(\mathbf{z}_i \mathbf{x}_i')] = K$ (满列秩)，其中 $\mathbf{z}_i \equiv (z_{i1} \cdots z_{iL})'$ (L 个工具变量)， $\mathbf{x}_i \equiv (x_{i1} \cdots x_{iK})'$ (K 个解释变量)， \mathbf{z}_i 与 \mathbf{x}_i 可有重叠元素，且 $L \geq K$ (阶条件)。

如果矩阵 $E(\mathbf{z}_i \mathbf{x}_i')$ 的列秩小于 K ，则不可识别。

对于秩条件是否成立，可进行“不可识别检验” (underidentification test)。

原假设为 “ $H_0: \text{rank}[E(\mathbf{z}_i \mathbf{x}_i')] = K - 1$ ” (不可识别)，而替代假设为 “ $H_1: \text{rank}[E(\mathbf{z}_i \mathbf{x}_i')] = K$ ”。

在扰动项为 iid 的假设下(不存在异方差)，可使用“Anderson LM 统计量” (Anderson, 1951)，其渐近分布为 $\chi^2(L - K + 1)$ 。

如果不作 iid 扰动项的假设(允许存在异方差)，应使用

“Kleibergen-Paap rk LM 统计量(Kleibergen-Paap, 2006), 其渐近分布也是 $\chi^2(L-K+1)$ 。

2. 弱工具变量检验

如果工具变量 \mathbf{z} 与内生解释变量 \mathbf{x} 完全不相关, 则无法使用工具变量法, 因为 $[E(\mathbf{z}_i \mathbf{x}_i')]^{-1}$ 不存在。

如果 \mathbf{z} 与 \mathbf{x} 仅仅微弱地相关, 可认为 $[E(\mathbf{z}_i \mathbf{x}_i')]^{-1}$ 很大, 导致 IV 估计量的渐近方差 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = [E(\mathbf{z}_i \mathbf{x}_i')]^{-1} \mathbf{S} [E(\mathbf{z}_i \mathbf{x}_i')]^{-1}$ 变得很大。

由于 \mathbf{z} 中仅包含很少与 \mathbf{x} 有关的信息, 利用这部分信息进行的 IV 估计不准确, 即使样本容量很大也很难收敛到真实参数值。这种

工具变量称为“弱工具变量”(weak instruments)。

弱工具变量的后果类似于样本容量过小，导致 $\hat{\beta}_{IV}$ 的小样本性质变得很差，而 $\hat{\beta}_{IV}$ 的大样本分布也可能离正态分布相去甚远，致使基于大样本理论的统计推断失效。

判断弱工具变量的方法主要有以下四种。方法之一为使用“偏 R^2 ”。假设回归模型为

$$y = \mathbf{x}_1' \boldsymbol{\beta}_1 + x_2 \beta_2 + \varepsilon$$

其中，只有 x_2 为内生解释变量。记工具变量为 $(\mathbf{x}_1 \mathbf{z}_2)$ ，其中 \mathbf{z}_2 为方程外的工具变量。

在 2SLS 的第一阶段回归中， $x_2 \xrightarrow{\text{OLS}} \mathbf{x}_1, \mathbf{z}_2$ ，其 R^2 包含了内生

变量 x_2 与工具变量 z_2 相关性的信息，但也可能由于 x_2 与 x_1 的相关性所造成。应使用滤去 x_1 影响的“偏 R^2 ”(partial R^2)，记为 R_p^2 。

具体步骤：首先把 x_2 对 x_1 回归， $x_2 \xrightarrow{\text{OLS}} x_1$ ，记其残差为 e_{x_2} ，代表 x_2 中不能由 x_1 解释的部分；

其次，把 z_2 对 x_1 回归， $z_2 \xrightarrow{\text{OLS}} x_1$ ，记其残差为 e_{z_2} ，代表 z_2 中不能由 x_1 解释的部分；

最后，对两个残差进行回归，即 $e_{x_2} \xrightarrow{\text{OLS}} e_{z_2}$ ，所得可决系数就是 R_p^2 。具体 R_p^2 多低才构成弱工具变量，目前尚无共识。

Shea (1997)将 R_p^2 推广到多个内生解释变量的情形，Stata 称为

“Shea’s partial R^2 ”。

判断弱工具变量的方法之二为，在第一阶段回归中， $x_2 = \mathbf{x}'_1\boldsymbol{\gamma}_1 + \mathbf{z}'_2\boldsymbol{\gamma}_2 + error$ ，检验原假设“ $H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}$ ”。

经验规则(rule of thumb)是，如果此检验的 F 统计量大于 10，则可拒绝“存在弱工具变量”的原假设，不必担心弱工具变量问题。

在多个内生解释变量的情况下，有多个第一阶段回归，故有多个 F 统计量。可使用 Stock and Yogo (2005)提出的“最小特征值统计量”(minimum eigenvalue statistic)。

如果只有一个内生解释变量，该统计量还原为 F 统计量。

判断弱工具变量的方法之三为，如果假设扰动项为 iid，可使用“Cragg-Donald Wald F 统计量” (Cragg and Donald, 1993)，其临界值由 Stock and Yogo (2005)提供。

判断弱工具变量的方法之四为，如果不作 iid 扰动项的假设，则应使用“Kleibergen-Paap Wald rk F 统计量”，其临界值也来自 Stock and Yogo (2005)。

解决弱工具变量问题的方法：

- (i) 寻找更强的工具变量；
- (ii) 使用对弱工具变量更不敏感的“有限信息最大似然估计法”

(Limited Information Maximum Likelihood Estimation, 简记 LIML); 在大样本下, LIML 与 2SLS 是渐近等价的。在存在弱工具变量的情况下, LIML 的小样本性质可能优于 2SLS。

(iii) 如果有较多工具变量, 可舍弃弱工具变量。在选择舍弃哪个工具变量时, 可进行“冗余检验”(redundancy test)。

冗余工具变量(redundant instruments): 使用这些工具变量不会提高估计量的渐近效率(asymptotic efficiency)。

具体表现: 在第一阶段回归中, 这些工具变量的系数不显著; 或者这些工具变量与内生解释变量的“偏相关系数”(partial correlations)为 0 或接近于 0(偏相关系数的平方即为偏 R^2)。

基于偏相关的思想，该检验考察内生变量与可能的冗余工具变量之间在过滤掉(partial out)其他工具变量影响之后的矩阵交叉乘积(matrix cross product)的秩是否为 0。

该冗余检验的原假设是，指定的工具变量为多余的。

该检验统计量的渐近分布为 χ^2 分布，自由度为“内生变量个数”乘以“冗余工具变量个数”。

3. 过度识别检验

在恰好识别的情况下，无法检验 IV 的外生性，只能定性讨论。

如果 IV 是外生的，则其对被解释变量发生影响的唯一渠道就是通过内生变量，除此以外别无其他渠道。

由于此唯一渠道(内生变量)已包括在回归方程中，故工具变量不会再出现在被解释变量的扰动项中，或对此扰动项有影响。此条件被称为“排他性约束”(exclusion restriction)。

需要找出 IV 影响被解释变量的所有其他可能渠道，一一排除，才能信服地说明工具变量的外生性。

在过度识别的情况下,可进行“过度识别检验”(overidentification test)。

此检验的大前提(maintained hypothesis)是该模型至少是恰好识别的,即有效工具变量至少与内生解释变量一样多。

在此大前提下,过度识别检验的原假设为“ H_0 :所有工具变量都是外生的”。

如果拒绝该原假设,则认为至少某个变量与扰动项相关。

假设前 $(K-r)$ 个解释变量 $\{x_1, \dots, x_{K-r}\}$ 为外生解释变量,而后 r 个解释变量 $\{x_{K-r+1}, \dots, x_K\}$ 为内生解释变量。

假设共有 m 个方程外的工具变量 $\{z_1, \dots, z_m\}$ ，其中 $m > r$ 。

把工具变量法的残差对所有外生变量(即所有外生解释变量与 IV)进行以下辅助回归：

$$e_{i, IV} = \gamma_1 x_{i1} + \dots + \gamma_{K-r} x_{i, K-r} + \delta_1 z_{i1} + \dots + \delta_m z_{im} + error_i$$

原假设可写为 “ $H_0: \delta_1 = \dots = \delta_m = 0$ ”。记此辅助回归的可决系数为 R^2 ，则 Sargan 统计量为

$$nR^2 \xrightarrow{d} \chi^2(m-r)$$

其中， χ^2 分布的自由度 $(m-r)$ 为过度识别约束的个数(即“多余”的工具变量个数)。

如果恰好识别，则 $m - r = 0$ (自由度为 0)， $\chi^2(0)$ 无定义，故无法使用这个“过度识别检验”。

即使接受了过度识别的原假设，也并不能证明这些工具变量的外生性。

过度识别检验的成立有一个大前提，即至少该模型是恰好识别的。此大前提无法检验，只能假定其成立。

4. 究竟该用 OLS 还是工具变量法：对解释变量内生性的检验

“豪斯曼检验” (Hausman specification test)(Hausman, 1978)的原假设为“ H_0 ：所有解释变量均为外生变量”。

如果 H_0 成立，则 OLS 与 IV 都一致，在大样本下 $\hat{\beta}_{IV}$ 与 $\hat{\beta}_{OLS}$ 都收敛于真实的参数值 β ，故 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 依概率收敛于 $\mathbf{0}$ 。

反之，如果 H_0 不成立，则 IV 一致而 OLS 不一致，故 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 不会收敛于 $\mathbf{0}$ 。

如果 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 的距离很大，则倾向于拒绝原假设。

根据沃尔德检验原理，以二次型来度量此距离：

$$(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})' \mathbf{D}^- (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) \xrightarrow{d} \chi^2(r)$$

其中, $\mathbf{D} \equiv \widehat{\text{Var}(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})}$, 而 \mathbf{D}^- 为 \mathbf{D} 的广义逆矩阵(因为 \mathbf{D} 不一定可逆; 但如果 \mathbf{D} 正定, 则可逆)。当 \mathbf{D} 可逆时, 广义逆矩阵就是一般的逆矩阵, 即 $\mathbf{D}^- = \mathbf{D}^{-1}$ 。

r 为内生解释变量的个数(不包括外生解释变量)。

$$\text{Var}(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) = \text{Var}(\hat{\beta}_{\text{IV}}) + \text{Var}(\hat{\beta}_{\text{OLS}}) - 2\text{Cov}(\hat{\beta}_{\text{IV}}, \hat{\beta}_{\text{OLS}})$$

如果在 H_0 成立的情况下, OLS 最有效率, 则可证明 $\text{Cov}(\hat{\beta}_{\text{IV}}, \hat{\beta}_{\text{OLS}}) = \text{Var}(\hat{\beta}_{\text{OLS}})$, 故

$$\mathbf{D} = \widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}})} - \widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}})}$$

上述检验的缺点是，它假设在 H_0 成立的情况下，OLS 最有效率。

如果存在异方差，OLS 并不最有效率(不是 BLUE)。故传统的豪斯曼检验不适用于异方差的情形。

解决方法之一为，通过“自助法”(bootstrap)，即计算机模拟“再抽样”(resampling)的方法来计算 $\mathbf{D} \equiv \widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})}$ 。

解决方法之二为，使用“杜宾-吴-豪斯曼检验”(Durbin-Wu-Hausman Test, 简记 DWH)，该检验在异方差的情况下也适用，更为稳健。

假设模型为, $y = \mathbf{x}'_1\boldsymbol{\beta}_1 + x_2\beta_2 + \varepsilon$, 其中 x_2 为唯一的内生解释变量。

记工具变量为 $\mathbf{z} \equiv (\mathbf{x}_1 \ z_2)$, 其中 z_2 为方程外的工具变量。

考虑 2SLS 的第一阶段回归, 即 $x_2 = \mathbf{x}'_1\boldsymbol{\gamma} + \mathbf{z}'\boldsymbol{\delta} + v$ 。

由于工具变量 \mathbf{z} 与 ε 不相关, 故

$$E(x_2\varepsilon) = E[(\mathbf{x}'_1\boldsymbol{\gamma} + \mathbf{z}'\boldsymbol{\delta} + v)\varepsilon] = \underbrace{E(\mathbf{x}'_1\boldsymbol{\gamma}\varepsilon)}_{=0} + \underbrace{E(\mathbf{z}'\boldsymbol{\delta}\varepsilon)}_{=0} + E(v\varepsilon) = E(v\varepsilon)$$

这意味着

$$“x_2 \text{ 为内生变量}” \Leftrightarrow “E(x_2\varepsilon) \neq 0” \Leftrightarrow “E(v\varepsilon) \neq 0”$$

故只需检验第一阶段回归的扰动项 v 是否与原模型的扰动项 ε 相关即可：

$$\varepsilon = \rho v + \xi$$

如果 ε 与 v 不相关，则 $\rho = 0$ 。代入原模型可得

$$y = \mathbf{x}'_1 \boldsymbol{\beta}_1 + x_2 \beta_2 + \rho v + \xi$$

使用第一阶段回归的残差 \hat{v} 来代替 v ，进行辅助回归：

$$y = \mathbf{x}'_1 \boldsymbol{\beta}_1 + x_2 \beta_2 + \hat{v} \rho + error$$

对原假设“ $H_0 : \rho = 0$ ”进行 t 检验。如果拒绝“ $H_0 : \rho = 0$ ”，则

认为存在内生解释变量；否则，认为所有解释变量均为外生。

考虑到可能存在异方差，则在作 t 检验时使用稳健标准误。

如果存在多个内生解释变量，即 $y = \mathbf{x}'_1\boldsymbol{\beta}_1 + \mathbf{x}'_2\boldsymbol{\beta}_2 + \varepsilon$ 。

则在第一阶段回归中可得到与内生解释变量 \mathbf{x}_2 相对应的多个残差 $\hat{\mathbf{v}}$ ，进行以下辅助回归，

$$y = \mathbf{x}'_1\boldsymbol{\beta}_1 + \mathbf{x}'_2\boldsymbol{\beta}_2 + \hat{\mathbf{v}}'\boldsymbol{\rho} + error$$

对原假设 “ $H_0: \boldsymbol{\rho} = \mathbf{0}$ ” 进行 F 检验即可。如果担心存在异方差，则可在作 F 检验时使用稳健标准误。

10.5 GMM 的假定

在球型扰动项的假定下，2SLS 最有效率。如果存在异方差或自相关，则存在更有效的方法，即“广义矩估计”(Generalized Method of Moments, 简记 GMM)。

GMM 之于 2SLS，正如 GLS 之于 OLS。

假定 10.1 线性假定(linearity)

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n)$$

其中， $\mathbf{x}_i \equiv (x_{i1} \ x_{i2} \ \cdots \ x_{iK})'$ 为第 i 个观测数据。

假定 10.2 渐近独立的平稳过程

记 L 维工具变量为 \mathbf{z}_i (可能与 \mathbf{x}_i 重叠), \mathbf{w}_i 由 $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}$ 中不重复的变量构成且不含常数项。随机过程 $\{\mathbf{w}_i\}$ 为渐近独立的平稳过程。

假定 10.3 工具变量的正交性

所有工具变量 \mathbf{z}_i 均为“前定”，即与同期扰动项正交。定义 L 维列向量 $\mathbf{g}_i \equiv \mathbf{z}_i \varepsilon_i$ ，则 $E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 。

假定 10.4 秩条件

$L \times K$ 维矩阵 $E(\mathbf{z}_i \mathbf{x}_i')$ 满列秩，即 $\text{rank}[E(\mathbf{z}_i \mathbf{x}_i')] = K$ 。记 $\boldsymbol{\Sigma}_{ZX} \equiv E(\mathbf{z}_i \mathbf{x}_i')$ 。

假定 10.5 $\{\mathbf{g}_i\}$ 为鞅差分序列，其协方差矩阵 $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}_i') = E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i')$ 为非退化矩阵。

假设 10.6 四阶矩 $E[(x_{ik} z_{ij})^2]$ 存在且有限， $\forall i, j, k$ (finite fourth moments)。

10.6 GMM 的推导

与总体矩条件 $E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 相对应的样本矩条件为

$$\mathbf{g}_n(\hat{\boldsymbol{\beta}}) \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) = \mathbf{0}$$

此联立方程组，未知数 $\hat{\boldsymbol{\beta}}$ 有 K 个，方程个数为 L 个 (\mathbf{z}_i 的维度)。

如果 $L < K$ ，为不可识别，则 $\hat{\beta}$ 有无穷多解。

如果 $L = K$ ，为恰好识别，则 $\hat{\beta}$ 有唯一解，即 $\hat{\beta}_{IV}$ 。

如果 $L > K$ ，为过度识别，则 $\hat{\beta}$ 无解。传统的矩估计法行不通。

虽然无法找到 $\hat{\beta}$ 使得 $\mathbf{g}_n(\hat{\beta}) = \mathbf{0}$ ，总可以找到 $\hat{\beta}$ ，使得向量 $\mathbf{g}_n(\hat{\beta})$ 尽可能地接近 $\mathbf{0}$ ，比如，使二次型 $(\mathbf{g}_n(\hat{\beta}))'(\mathbf{g}_n(\hat{\beta}))$ 最小。

更一般地，用“权重矩阵”(weighting matrix) W 来构成二次型。

假设 \hat{W} 为 $L \times L$ 维对称正定矩阵(可依赖于样本), 且 $\text{plim}_{n \rightarrow \infty} \hat{W} = W$, 其中 W 为非随机的对称正定矩阵。定义最小化的目标函数为

$$\min_{\hat{\beta}} J(\hat{\beta}, \hat{W}) \equiv n \left(g_n(\hat{\beta}) \right)' \hat{W} \left(g_n(\hat{\beta}) \right)$$

其中, 因子 n 不影响最小化。定义“GMM 估计量”为此无约束二次型最小化问题的解(Hansen, 1982):

$$\hat{\beta}_{\text{GMM}}(\hat{W}) \equiv \argmin_{\hat{\beta}} J(\hat{\beta}, \hat{W})$$

GMM 估计量取决于权重矩阵 \hat{W} 。对 \hat{W} 的自由选择是 GMM 的最大优点之一, 可通过最优地选择 \hat{W} 使 $\hat{\beta}_{\text{GMM}}$ 最有效。

不同矩条件的强弱程度一般不同，强的矩条件意味着其对应的方差较小(矩阵 $\mathbf{S} = \mathbf{E}(\mathbf{g}_i \mathbf{g}_i')$ 的主对角线元素)，是比较紧的约束，故会通过 $\hat{\mathbf{W}}$ 得到较大的权重。

$\mathbf{g}_n(\hat{\boldsymbol{\beta}})$ 是 $\hat{\boldsymbol{\beta}}$ 的一次函数，故 $J(\hat{\boldsymbol{\beta}}, \hat{\mathbf{W}})$ 是 $\hat{\boldsymbol{\beta}}$ 的二次(型)函数，通过向量微分可得到其最小化问题的解：

$$\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) = (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{Zy}}$$

$$\text{其中, } \mathbf{S}_{\text{ZX}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{x}_i', \quad \mathbf{S}_{\text{Zy}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i y_i \circ$$

秩条件 $\text{rank}[\mathbf{E}(\mathbf{z}_i \mathbf{x}_i')] = K$ 及 $\hat{\mathbf{W}}$ 正定保证在大样本下， $(\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1}$ 存在。

在恰好识别情况下， \mathbf{S}_{ZX} 为方阵，GMM 还原为普通的 IV 法：

$$\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) = \mathbf{S}_{ZX}^{-1} \underbrace{\hat{\mathbf{W}}^{-1} \mathbf{S}_{ZX}'^{-1} \mathbf{S}_{ZX}' \hat{\mathbf{W}}}_{=I} \mathbf{S}_{Zy} = \mathbf{S}_{ZX}^{-1} \mathbf{S}_{Zy} = \hat{\boldsymbol{\beta}}_{\text{IV}}$$

故 GMM 确实是矩估计的推广。

只有在过度识别的情况下，才有必要使用 GMM。

10.7 GMM 的大样本性质

定理(GMM 估计量的大样本性质)

(1) ($\hat{\beta}_{\text{GMM}}$ 为一致估计) 在假定 10.1-10.4 之下,
 $\text{plim}_{n \rightarrow \infty} \hat{\beta}_{\text{GMM}}(\hat{W}) = \beta$ 。

(2) ($\hat{\beta}_{\text{GMM}}$ 为渐近正态) 如果假定 10.3 (即 $E(\mathbf{g}_i) = \mathbf{0}$) 强化为假定 10.5 (即 $\{\mathbf{g}_i\}$ 为鞅差分序列), 则

$$\sqrt{n}(\hat{\beta}_{\text{GMM}} - \beta) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\beta}_{\text{GMM}}))$$

其中, $\text{Avar}(\hat{\beta}_{\text{GMM}}) = (\Sigma'_{\text{ZX}} W \Sigma'_{\text{ZX}})^{-1} \Sigma_{\text{ZX}} W S W \Sigma_{\text{ZX}} (\Sigma'_{\text{ZX}} W \Sigma_{\text{ZX}})^{-1}$,
 $S = E(\mathbf{g}_i \mathbf{g}_i') = E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i')$, $\Sigma_{\text{ZX}} \equiv E(\mathbf{z}_i \mathbf{x}_i')$ 。

(3) ($\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$ 的一致估计量)如果 $\hat{\mathbf{S}}$ 是 \mathbf{S} 的一致估计量,则在假定 10.2 下, $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$ 的一致估计量为

$$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})} = (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \hat{\mathbf{S}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}} (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1}$$

证明: (1) 抽样误差可写为

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) - \boldsymbol{\beta} &= (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i y_i \right) - \boldsymbol{\beta} \\ &= (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (\mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i) \right) - \boldsymbol{\beta} \\ &= (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \left(\mathbf{S}_{\text{ZX}} \boldsymbol{\beta} + \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i \right) - \boldsymbol{\beta} \\ &= (\mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \mathbf{S}_{\text{ZX}})^{-1} \mathbf{S}'_{\text{ZX}} \hat{\mathbf{W}} \bar{\mathbf{g}} \end{aligned}$$

其中, $\bar{\mathbf{g}} \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i$, $\mathbf{g}_i \equiv \mathbf{z}_i \varepsilon_i$ 。由于 $(\mathbf{S}'_{ZX} \hat{\mathbf{W}} \mathbf{S}_{ZX})^{-1} \xrightarrow{p} (\boldsymbol{\Sigma}'_{ZX} \mathbf{W} \boldsymbol{\Sigma}_{ZX})^{-1}$, $\mathbf{S}'_{ZX} \hat{\mathbf{W}} \xrightarrow{p} \boldsymbol{\Sigma}'_{ZX} \mathbf{W}$, 而 $\bar{\mathbf{g}} \xrightarrow{p} E(\mathbf{g}_i) = E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$ 。因此, $\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$ 。

保证 GMM 一致性的最重要条件仍是 $E(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$, 即工具变量与扰动项正交。

(2) 由于抽样误差 $\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) - \boldsymbol{\beta} = (\mathbf{S}'_{ZX} \hat{\mathbf{W}} \mathbf{S}_{ZX})^{-1} \mathbf{S}'_{ZX} \hat{\mathbf{W}} \bar{\mathbf{g}}$, 故 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{W}}) - \boldsymbol{\beta}) = (\mathbf{S}'_{ZX} \hat{\mathbf{W}} \mathbf{S}_{ZX})^{-1} \mathbf{S}'_{ZX} \hat{\mathbf{W}} (\sqrt{n} \bar{\mathbf{g}})$ 。

根据假定 10.5 及鞅差分序列的中心极限定理, $\sqrt{n} \bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$, 其中 $\mathbf{S} \equiv E(\mathbf{g}_i \mathbf{g}_i') = E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i')$ 。

由于 $\sqrt{n}(\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta)$ 是 $\sqrt{n} \bar{g}$ 的线性组合，故 $\sqrt{n}(\hat{\beta}_{\text{GMM}}(\hat{W}) - \beta) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\beta}_{\text{GMM}}))$ 。由于 $(S'_{ZX} \hat{W} S_{ZX})^{-1} \xrightarrow{p} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$ ， $S'_{ZX} \hat{W} \xrightarrow{p} \Sigma'_{ZX} W$ ，故

$$\text{Avar}(\hat{\beta}_{\text{GMM}}) = (\Sigma'_{ZX} W \Sigma_{ZX})^{-1} \Sigma'_{ZX} W S W \Sigma_{ZX} (\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$$

其中， $(\Sigma'_{ZX} W \Sigma_{ZX})^{-1}$ 为对称矩阵。

(3) 由于 $\hat{S} \xrightarrow{p} S$ ，而且 $S_{ZX} \xrightarrow{p} \Sigma_{ZX}$ ， $\hat{W} \xrightarrow{p} W$ ，故估计量 $\widehat{\text{Avar}(\hat{\beta}_{\text{GMM}})} = \underbrace{(S'_{ZX} \hat{W} S_{ZX})^{-1} S'_{ZX} \hat{W}}_{\text{面包}} \underbrace{\hat{S} \hat{W} S_{ZX}}_{\text{菜}} \underbrace{(S'_{ZX} \hat{W} S_{ZX})^{-1}}_{\text{面包}}$ 是 $\text{Avar}(\hat{\beta}_{\text{GMM}})$ 的一致估计量。形式上也是一个夹心估计量。

命题 在假定 10.1、假定 10.2 与假定 10.6 下(四阶矩存在), 对于 $\boldsymbol{\beta}$ 的任何一致估计量 $\hat{\boldsymbol{\beta}}$, 定义残差 $e_i \equiv y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$, 则 $s^2 \equiv \frac{1}{n} \sum_{i=1}^n e_i^2$ 是 $\sigma^2 \equiv E(\varepsilon_i^2)$ 的一致估计, 而且 $\hat{\mathbf{S}} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}_i'$ 是 $\mathbf{S} \equiv E(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i')$ 的一致估计。

命题 使 $\text{Avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}})$ 最小化的“最优权重矩阵”(optimal weighting matrix)为 $\hat{\mathbf{W}} = \hat{\mathbf{S}}^{-1}$ 。

定义 使用 $\hat{\mathbf{S}}^{-1}$ 为权重矩阵的 GMM 估计量被称为“效率 GMM”(efficient GMM)或“最优 GMM”(optimal GMM)。

由于 2SLS 一致，用 2SLS 的残差来计算 $\hat{\mathbf{S}} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}_i'$ 也一致。

两步最优 GMM 估计：

第一步：使用 2SLS，得到残差，计算 $\hat{\mathbf{S}} \equiv \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{z}_i \mathbf{z}_i'$ 。

第二步：最小化 $J(\hat{\boldsymbol{\beta}}, \hat{\mathbf{S}}^{-1})$ ，得到 $\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{S}}^{-1})$ 。

实际操作中，常使用“迭代法”(iterative GMM)直至估计值收敛，即用第二步所获残差再来计算 $\hat{\mathbf{S}}$ ，然后再求 $\hat{\boldsymbol{\beta}}_{\text{GMM}}(\hat{\mathbf{S}}^{-1})$ ，以此类推。

命题 在条件同方差的情况下，最优 GMM 就是 2SLS。

证明：假设 $E(\varepsilon_i^2 | \mathbf{z}_i) = \sigma^2 > 0$ (条件同方差)，则根据迭代期望定律，

$$\mathbf{S} \equiv E(\mathbf{z}_i \mathbf{z}_i' \varepsilon_i^2) = E_{\mathbf{z}_i} E(\mathbf{z}_i \mathbf{z}_i' \varepsilon_i^2 | \mathbf{z}_i) = E_{\mathbf{z}_i} \left[\mathbf{z}_i \mathbf{z}_i' E(\varepsilon_i^2 | \mathbf{z}_i) \right] = \sigma^2 E(\mathbf{z}_i \mathbf{z}_i')$$

因此， $\tilde{\mathbf{S}} \equiv s^2 \mathbf{S}_{ZZ}$ 是 \mathbf{S} 的一致估计量，其中 $\mathbf{S}_{ZZ} \equiv \frac{1}{n} \mathbf{Z}'\mathbf{Z}$ 。

使用 $\tilde{\mathbf{S}}^{-1} = (s^2 \mathbf{S}_{ZZ})^{-1}$ 为最优权重矩阵，则最优 GMM 估计量为

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GMM}}(\tilde{\mathbf{S}}^{-1}) &= \left(\mathbf{S}'_{ZX} (s^2 \mathbf{S}_{ZZ})^{-1} \mathbf{S}_{ZX} \right)^{-1} \mathbf{S}'_{ZX} (s^2 \mathbf{S}_{ZZ})^{-1} \mathbf{S}_{Zy} \\ &= \left(\mathbf{S}'_{ZX} \mathbf{S}_{ZZ}^{-1} \mathbf{S}_{ZX} \right)^{-1} \mathbf{S}'_{ZX} \mathbf{S}_{ZZ}^{-1} \mathbf{S}_{Zy} \end{aligned}$$

由于 $\mathbf{S}_{ZX} \equiv \frac{1}{n} \mathbf{Z}'\mathbf{X}$, $\mathbf{S}_{ZZ} \equiv \frac{1}{n} \mathbf{Z}'\mathbf{Z}$, $\mathbf{S}_{Zy} \equiv \frac{1}{n} \mathbf{Z}'\mathbf{y}$, 故

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GMM}}(\tilde{\mathbf{S}}^{-1}) &= \left(\frac{1}{n} \mathbf{X}'\mathbf{Z} \cdot n(\mathbf{Z}'\mathbf{Z})^{-1} \cdot \frac{1}{n} \mathbf{Z}'\mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}'\mathbf{Z} \cdot n(\mathbf{Z}'\mathbf{Z})^{-1} \frac{1}{n} \mathbf{Z}'\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y} \equiv \hat{\boldsymbol{\beta}}_{2\text{SLS}}\end{aligned}$$

在条件同方差的情况下，两步最优 GMM 可省略为一步。因为两步最优 GMM 中第一步的目的只是得到 $\hat{\mathbf{S}}^{-1}$ ，而在条件同方差假定下，可直接令 $\hat{\mathbf{S}}^{-1} = \mathbf{S}_{ZZ}^{-1}$ 。

故 2SLS 也称为“一步 GMM”。

GMM 的过度识别检验(Overidentification Test or Hansen's J Test)

在恰好识别的情况下，GMM 最小化的目标函数 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1}) = 0$ 。

在过度识别的情况下，如果所有的过度识别约束都成立，则目标函数 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1})$ 应该离 0 不远。

如果 $J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1})$ 大于 0 很多，则可倾向于认为某些过度识别约束不成立。

在原假设“ H_0 : 所有矩条件均成立”的情况下，目标函数本身就是检验统计量

$$J(\hat{\beta}_{\text{GMM}}, \hat{S}^{-1}) \xrightarrow{d} \chi^2(L-K)$$

其中， $(L-K)$ 为过度识别约束的个数，因为在估计 $\hat{\beta}_{\text{GMM}}$ 的过程中失去了 K 个自由度。

在条件同方差的情况下， J 统计量与 Sargan 统计量相等。

检验部分工具变量的正交性(Testing Subsets of Orthogonality Condition)

如果过度识别检验拒绝“所有工具变量均为外生”的原假设，则怀疑部分工具变量不满足正交性(外生性)。

假设在 L 个工具变量 \mathbf{z}_i 中，已知前 L_1 个工具变量 \mathbf{z}_{i1} 满足正交性，而怀疑后 $(L - L_1)$ 个工具变量 \mathbf{z}_{i2} 不满足正交性。

检验原假设 “ $H_0: E(\mathbf{z}_{i2}\varepsilon_i) = 0$ ”。

进行此检验的前提条件是 $L_1 \geq K$ ，保证即使仅用前 L_1 个工具变量 \mathbf{z}_{i1} 进行估计，该模型也为恰好识别。

如果 $L_1 \geq K$ ，则可分别用所有 L 个工具变量 \mathbf{z}_i 或前 L_1 个工具变量 \mathbf{z}_{i1} 进行 GMM 估计，记相应的 J 统计量为 J 与 J_1 。

如果将后 $(L - L_1)$ 个工具变量 \mathbf{z}_{i2} 也用于 GMM 估计，使得 J 统计量大大增加，则倾向于拒绝原假设 “ $H_0: E(\mathbf{z}_{i2}\varepsilon_i) = 0$ ”。

可以证明：

$$C \equiv J - J_1 \xrightarrow{d} \chi^2(L - L_1)$$

其中， C 统计量称为“GMM 距离” (GMM distance) 统计量或“Sargan 差” (difference-in-Sargan) 统计量，因为它是两个 GMM 估计的 Sargan-Hansen 统计量之差。

自由度为 $(L - L_1)$ ，即怀疑正交性不成立的工具变量个数。

在存在自相关的情况下使用 GMM

在时间序列数据中，即使存在自相关，也仍可使用 GMM，只要采用异方差自相关稳健的标准误来进行统计推断就行。

10.8 如何获得工具变量

寻找工具变量的步骤大致可以分为两步：

- (i) 列出与内生解释变量(x)相关的尽可能多的变量的清单；
- (ii) 从这一清单中剔除与扰动项相关的变量。
- (ii) 的操作有一定难度，因为扰动项不可观测。

由于扰动项是被解释变量(y)的扰动项，故可从该候选变量与被解释变量的相关性着手。

显然 z 与 y 相关，因为 z 与内生解释变量 x 相关。

重要的是， z 对 y 的影响仅仅通过 x 来起作用，因为如果 z 与 ε 相关，则 z 对 y 的影响必然还有除 x 以外的渠道，参见图 10.3。

至于是否“ z 对 y 的影响仅仅通过 x 来起作用”，有时可以通过定性的讨论来确定。这就是“排他性约束” (exclusion restriction)。

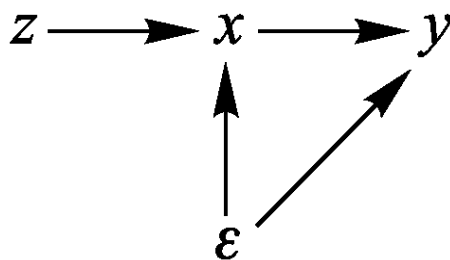


图 10.3 工具变量示意图

例 滞后变量。对于时间序列或面板数据，常使用内生解释变量的滞后变量作为工具变量。内生解释变量与其滞后变量相关。由于滞后变量已经发生，故为“前定”，可能与当期扰动项不相关。

Groves et al (1994)考察国企改革(员工奖金激励制度)对企业生产率的作用。一般来说，奖金占员工中报酬比重越高，则越能促进生产率的提高。另一方面，生产率越高的企业越有能力给员工发奖金，存在双向因果关系。

Groves et al (1994)使用奖金比重的滞后值作为当期奖金比重的工具变量。二者的相关性显然。另一方面，当期的生产率不可能影响过去的奖金比重，故奖金比重的滞后值具有外生性。

例 警察人数与犯罪率。警察人数越多，执法力度越大，则犯罪率应该越低。但警察人数是内生变量，如果某城市的犯罪率很高，则市政府通常会增加警察人数。

Levitt (1997)使用“市长选举的政治周期”作为犯罪率(包括 7 种类型的犯罪)的工具变量。在任市长在竞选连任时，为拉选票，会增加警察人数，故满足相关性。

另一方面，选举周期一般以机械的方式确定，除了对警察人数有影响外，不会单独地对犯罪率起作用，故满足外生性。

例 国际贸易与经济增长。实证研究国际贸易对经济增长的促进作用，面临内生解释变量的问题，因为经济增长可反作用于国际贸易。

Frankel and Romer (1999)使用地理因素作为工具变量。首先，国际贸易受地理因素的影响(比如，距离较近的国家之间的贸易量较大)，故满足相关性。

其次，地理因素对经济增长的影响可能仅仅通过国际贸易这个渠道来实现。

例 制度对经济增长的影响。好的制度能促进经济增长，但制度变迁常常也依赖于经济增长。

Acemoglu et al (2001)使用“殖民者死亡率”(settler mortality)作为制度的工具变量。当近代欧洲的殖民者在全世界进行殖民时，由于各地的气候及疾病环境(disease environment)不同，欧洲殖民者的死亡率十分不同。在死亡率高的地方(比如，非洲)，殖民者难

以定居，故在当地建立掠夺性制度(extractive institutions)。而在死亡率低的地方(比如，北美)，则建立有利于经济增长的制度(比如，较好的产权保护)。这种初始制度上的差异一直延续到今天。因此，殖民者死亡率与今天的制度相关，满足相关性。

另一方面，殖民者死亡率除了对制度有影响外，不再对当前的经济增长有任何直接影响，故满足外生性。

例 看电视过多引发小儿自闭症？Waldman et al (2006, 2008)研究过多观看电视是否引发小儿自闭症。但有自闭倾向的儿童可能更经常看电视，存在双向因果关系。

Waldman et al (2006, 2008)使用降雨量作为电视观看时间的工具变量。二者存在相关性，即降雨越多的地区，人们呆在室内的

时间越长，故看电视时间也越长；而降雨量很可能是外生的(只通过看电视时间而影响被解释变量)。

例 学区竞争与教育质量。是否一个城市的学区越多，学区间竞争越激烈，越有利于提高教育质量？内生性问题：在学区形成的过程中，效率高的学区会变得更大，或许兼并相邻的学区。

Hoxby (2000)使用一个城市河流的数目作为该城市学区个数的工具变量。历史上，如果一个城市的河流越多，则妨碍交通的自然障碍越多，导致城市设立更多的学区；故河流数目满足相关性。

另一方面，河流数目很可能不会直接影响教育质量，故满足外生性。