

- 第三章中，我们假设了零条件均值假设（MLR. 4）：
$$E(u_i|x_{i1}, x_{i2}, \dots, x_{ik}) = 0$$
- 解释变量与误差项不相关被叫做外生的（exogenous）；外生性服从 MLR. 4
- 解释变量与误差项的相关被叫做内生的（endogenous）；内生性违背了 MLR. 4
- 内生性会导致OLS不具有无偏性与一致性
- 在这一章中，我们讨论内生性下模型的估计问题

JEFFREY M. WOOLDRIDGE

Introductory
Econometrics
A Modern Approach

SIXTH EDITION

Chapter 15

工具变量估计与两阶
段最小二乘法



章节框架

- 在这一章中，我们将探讨内生性下模型的估计问题
- 首先，我们介绍内生性可能的原因以及工具变量的定义
- 之后，我们讨论如何利用工具变量构造估计值
- 最后，我们提出检验内生性的方法

动机：简单回归模型中的遗漏变量

- 内生性问题在社会科学和经济学中很普遍

- (1) 遗漏变量 (omitted variables)
- 例子：工资方程中的教育

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

误差项包含包含一些因素，例如与教育相关的智力，工作经验等

- (2) 测量误差 (measurement error)
- 例子：NBA球员工资与实力 (不能被准确观测)
- (3) 瞬时性 (simultaneity)
- 例子：警察密度与犯罪率

1. Simultaneity bias

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

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?

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。

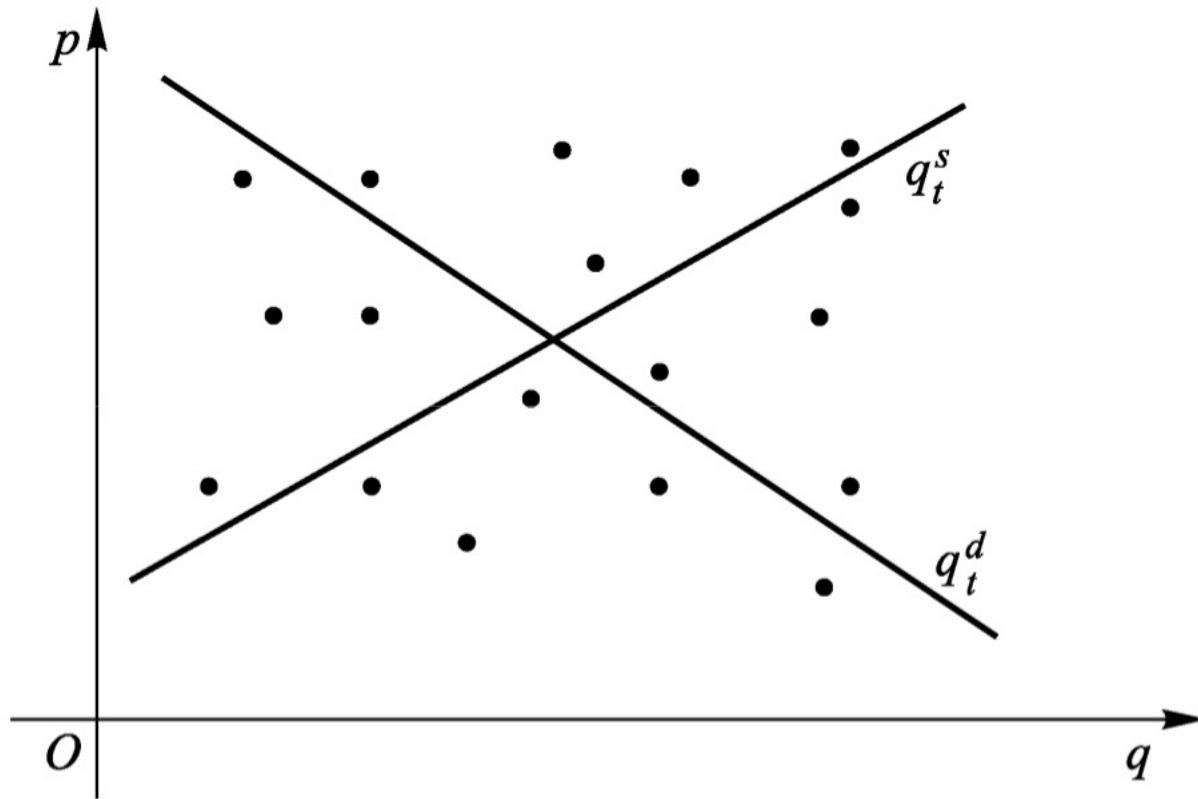


图 10.1 需求与供给决定市场均衡

2. Omitted variable problem

OLS估计量的期望值

真实:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v, E[v|x_1, x_2] = 0$$

- 回归中包含无关变量

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

没问题, 因为 $E(\hat{\beta}_3) = \beta_3 = 0$

$$E[y|x_1, x_2, x_3] = E[y|x_1, x_2]$$

= 0 在总体中

$$E[u|x_1, x_2, x_3] ? = 0$$

但是
不~~用管~~→然而, 包含无关变量可能会增加抽样的方差

(3.4小节)

$$= E[y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 | x_1, x_2]$$

$$\text{因此 } E(\hat{\beta}_3) = \beta_3 = 0.$$

- 遗漏变量: 简单情形

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad \xleftarrow{\text{真实样本(包含}x_1\text{和}x_2\text{)}}$$

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \quad \xleftarrow{\text{估计模型(}x_2\text{是遗漏的)}}$$

$$\textcircled{1} \quad y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_{k-1} x_{ik-1} + \hat{\beta}_k x_{ik} + \hat{u}_i$$

$$\textcircled{2} \quad y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \dots + \tilde{\beta}_{k-1} x_{ik-1} + \tilde{u}_i$$

辅助回归: $x_{ik} = \tilde{\delta}_0 + \tilde{\delta}_1 x_{i1} + \tilde{\delta}_2 x_{i2} + \dots + \tilde{\delta}_{k-1} x_{ik-1} + \tilde{v}_i$

结论: $\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j \quad j=0, \dots, k-1$

对 $j=1$ 证明. $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1$

证明: 辅助回归. $x_{i1} \sim x_{i2}, \dots, x_{ik-1}$, 得残差 \tilde{r}_{ii}

$$\tilde{\beta}_1 = \left(\sum_{i=1}^n \tilde{r}_{ii}^2 \right)^{-1} \left(\sum_{i=1}^n \tilde{r}_{ii} y_i \right) \quad \sum_{i=1}^n \tilde{r}_{ii} = 0, \quad \sum x_{i2} \tilde{r}_{ii} = \dots = \sum x_{ik-1} \tilde{r}_{ii} = 0 \quad (\text{ii})$$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

$$\begin{aligned} \tilde{r}_{ii} &= x_{i1} - \check{y}_0 - \check{y}_1 x_{i2} - \dots - \check{y}_{k-1} x_{ik-1} \\ x_{i1} &= \tilde{r}_{ii} + \check{y}_0 + \check{y}_1 x_{i2} + \dots + \check{y}_{k-1} x_{ik-1} \end{aligned}$$

$$\sum_{i=1}^n \tilde{r}_{ii} y_i = \sum_{i=1}^n \tilde{r}_{ii} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_K x_{iK} + \hat{u}_i)$$

(i) $\sum \tilde{r}_{ii} \hat{\beta}_0 = 0$, (ii) $\sum \tilde{r}_{ii} x_{i2} = \sum \tilde{r}_{ii} x_{i3} = \dots = \sum \tilde{r}_{ii} x_{iK-1} = 0$

$\sum \tilde{r}_{ii} \hat{u}_i = 0$, 因为 \hat{u}_i 和 $1, x_{i1}, \dots, x_{iK}$ 满足 Normal Equation

$$= \sum \tilde{r}_{ii} \hat{\beta}_1 x_{i1} + \sum \tilde{r}_{ii} \hat{\beta}_K x_{iK}$$

$$= \hat{\beta}_1 (\sum \tilde{r}_{ii} x_{i1}) + \hat{\beta}_K (\sum \tilde{r}_{ii} x_{iK})$$

(iii) $\sum \tilde{r}_{ii} x_{i1} = \sum \tilde{r}_{ii}^2$

故 $\hat{\beta}_1 = (\sum \tilde{r}_{ii}^2)^{-1} \cdot (\hat{\beta}_1 \sum \tilde{r}_{ii} + \hat{\beta}_K (\sum \tilde{r}_{ii} x_{iK}))$

$$= \hat{\beta}_1 + \hat{\beta}_K \underbrace{(\sum \tilde{r}_{ii}^2)^{-1} (\sum \tilde{r}_{ii} x_{iK})}_{\equiv \delta_1} = \hat{\beta}_1 + \hat{\beta}_K \delta_1$$

$\hat{\delta}_j$ 理解: x_j 在辅助回归中去掉其它 $x_{(j)}$ 影响的残差和
 x_k 回归. 若 x_j 和 $x_{(j)}$ 无关, 则残差是 x_j . 此时 $\hat{\delta}_j$
成为 x_j 和 x_k 之间的回归系数, 也即是 x_j 和 x_k
的相关.

仅有 $\text{cov}(x_{ik}, x_{ij}) = 0$, 不足以让 $\tilde{\delta}_j = 0$.

OLS估计量的期望值

- 遗漏变量偏误

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

如果 x_1 和 x_2 是相关的，假设一个线性回归关系

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 (\delta_0 + \delta_1 x_1 + v) + u$$

$$= (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (\beta_2 v + u)$$

如果 y 只对 x_1 做回归，
这就是截距项

如果 y 只对 x_1 做回归，
这就是斜率项

误差项

- 结论：所有估计系数都是偏误的 (All estimated coefficients will be biased)

只是简单讨论

- 当遗漏变量 x_2 时 β_1 的估计偏误情况汇总：

	$(\delta_1 > 0)$	$(\delta_1 < 0)$
$\text{Corr}(x_1, x_2) \geq 0$	Positive bias $\hat{\beta}_K \tilde{\delta}_1 > 0$	Negative bias $\hat{\beta}_K \tilde{\delta}_1 < 0$
$\beta_2 > 0$	Positive bias $\hat{\beta}_K \tilde{\delta}_1 > 0$	Positive bias $\hat{\beta}_K \tilde{\delta}_1 > 0$
$\beta_2 < 0$	Negative bias $\hat{\beta}_K \tilde{\delta}_1 < 0$	

- 思考：

此时 $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_K \tilde{\delta}_1$
 $x_K = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 + \tilde{\nu}$.

- 如果 x_1 和 x_2 是相关的，哪条基本假设不成立？没什幺不成立。
- 如果 x_1 和 x_2 是不相关的呢？见上页 $\tilde{\delta}_j$ 理解。

OLS估计量的期望值

- 例子：工资等式中遗漏能力

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

$$abil = \delta_0 + \delta_1 educ + v$$

$$wage = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) educ + (\beta_2 v + u)$$

都是正的

β_1 将被过度估计 (overestimated) 因为 $\beta_2 \delta_1 > 0$ 。这看起来好像受到多年教育的人会挣很高的工资，但这部分是由于平均来说受教育长的人能力也很强。

- 什么时候没有遗漏变量偏误？
 - 如果遗漏变量与解释变量是不相关的

OLS估计量的期望值

- 遗漏变量偏误：更一般的情形

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w$$

$$x_3 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + r$$

$$\text{代入: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (\delta_0 + \delta_1 x_1 + \delta_2 x_2 + r) + u$$

$$\begin{aligned} &= (\beta_0 + \beta_3 \delta_0) + (\beta_1 + \beta_3 \delta_1) x_1 \\ &\quad + (\beta_2 + \beta_3 \delta_2) x_2 \end{aligned}$$

$$+ (\beta_3 r + u)$$

- 很难得到偏误的方向，这是因为 x 间会两两相关
- 如果 x 间不相关，那分析就和之前一样简单
- 例子：工资等式中遗漏能力

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

如果 $exper$ 与($educ$, $abil$)都近似不相关，则遗漏变量的偏误方向就像之前两个变量的简单情形一致： β_1 将被过度估计

但实际上， $exper$ 与 $educ$ 通常多少会负相关

3. Measurement error problem

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：

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

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

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。

Instrumental Variable method (IV)

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：

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

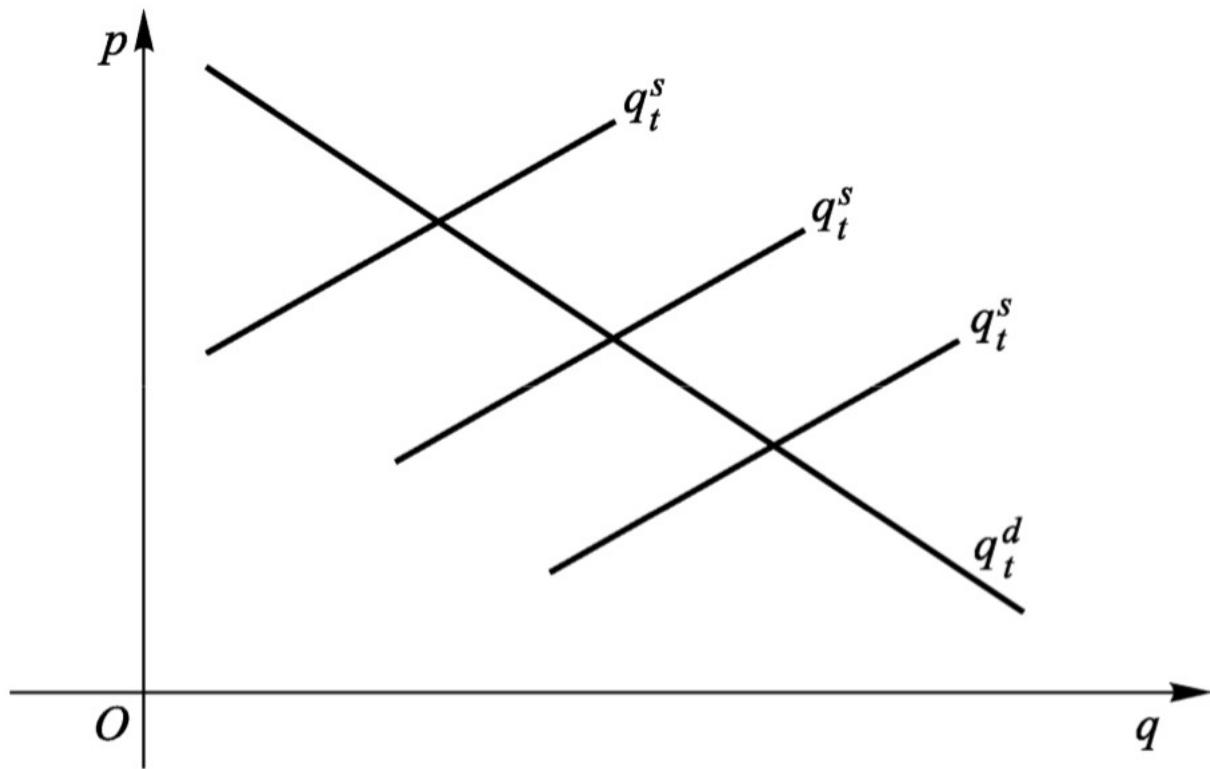


图 10.2 稳定的需求与变动的供给

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人性，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。

动机：简单回归模型中的遗漏变量

- 内生性的影响
 - 考虑一个简单回归模型

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- OLS估计：

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

- 如果 $Cov(x_i, u_i) \neq 0$,

$$E(\hat{\beta}_1) \neq \beta_1,$$
$$plim \hat{\beta}_1 = \beta_1 + \frac{Cov(x_i, u_i)}{Var(x_i)} \neq \beta_1$$

- OLS不具有无偏性与一致性



动机：简单回归模型中的遗漏变量

- 工具变量法 (Instrument Variable, IV)
- IV (z_i) 的定义：
 - 1) 不能出现在回归中
 - 2) 与内生变量高度相关 $Cov(z_i, x_i) \neq 0$
 - 3) 与误差项不相关 $Cov(z_i, u_i) = 0$
- 由性质3) 可得，

$$Cov(z_i, u_i) = 0$$

$$\Leftrightarrow 0 = Cov(z_i, y_i - \beta_0 - \beta_1 x_i) = Cov(z_i, y_i) - \beta_1 Cov(z_i, x_i)$$

- 又因为性质2) ，

$$\Leftrightarrow \beta_1 = Cov(z_i, y_i) / Cov(z_i, x_i)$$

动机：简单回归模型中的遗漏变量

- 从而，我们可以定义如下IV估计量：

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$

- IV估计量满足一致性：

$$\Rightarrow \hat{\beta}_{IV} = \widehat{Cov}(z_i, y_i) / \widehat{Cov}(z_i, x_i) \rightarrow Cov(z_i, y_i) / Cov(z_i, x_i) = \beta_1$$

- 注意：一般情况下IV估计量不具有无偏性

- 如果 $E(u_i^2 | z_1, \dots, z_n) = \sigma^2$ ，可以计算 $\hat{\beta}_{IV}$ 的渐近方差为

$$\frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2}$$

其中 σ_x^2 为 x_i 的总体方差， $\rho_{x,z}^2$ 为 x_i 与 z_i 总体相关系数的平方

动机：简单回归模型中的遗漏变量

- 例子：父亲的教育程度作为IV

OLS: $\widehat{\log(wage)} = - .185 + .109 \text{ educ}$ ← 教育回报可能被高估了
(.185) (.014)

$$n = 428, R^2 = .118$$

$$\widehat{\text{educ}} = 10.24 + .269 \text{ fatheduc}$$

父亲的教育程度是一个好IV吗？

- 没有作为解释变量
- 与教育程度显著相关
- 与误差项不相关（?）

$$n = 428, R^2 = .173$$

IV: $\widehat{\log(wage)} = .441 + .059 \text{ educ}$

教育回报率低了（符合预期）

标准误也大了，不精确了

$$n = 428, R^2 = 1 - RSS_{IV}/TSS = .093$$



动机：简单回归模型中的遗漏变量

- 文献中教育程度的其他IV:
 - 兄弟姐妹的个数
 - 1) 不是工资的决定因素, 2) 与教育程度相关, 因为预算约束, 3) 与先天因素无关
 - 当16岁时与大学的地理远近
 - 1) 不是工资的决定因素, 2) 与教育相关, 因为离大学近将会受到更多教育, 3) 与误差项不相关
 - 出生月份
 - 1) 不是工资的决定因素, 2) 与教育相关, 因为义务教育法, 3) 与误差项不相关

动机：简单回归模型中的遗漏变量

- 低劣工具变量条件下IV的性质

- 如果工具变量不是完全外生的，而且与 x 的相关性较弱，则IV较OLS来说可能非常不一致

$$\text{plim } \hat{\beta}_{1,OLS} = \beta_1 + \text{Corr}(x, u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$\text{plim } \hat{\beta}_{1,IV} = \beta_1 + \frac{\text{Corr}(z, u)}{\text{Corr}(z, x)} \cdot \frac{\sigma_u}{\sigma_x}$$

如果IV是外生的，则没问题。如果不是，则IV与x的相关性越弱，渐近偏差就越大。

IV比OLS更差，如果： $\frac{\text{Corr}(z, u)}{\text{Corr}(z, x)} > \text{Corr}(x, u)$ 例如 $\frac{0.03}{0.2} > 0.1$

- IV与内生变量相关性过于弱导致的IV估计不准确问题被称为弱工具变量问题

多元回归模型的IV估计

- 考虑如下模型：

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

- y_2 为内生变量 (endogenous variable) : $Cov(y_2, u_1) \neq 0$
- z_1 为外生变量 (exogenous variable) : $Cov(z_1, u_1) = 0$

- 工具变量 z_2 的条件：

- 1) 不出现在回归中
- 2) 与误差项不相关: $Cov(z_2, u_1) = 0$
- 3) 与内生解释变量相关: $Cov(y_2, z_2) \neq 0$

- u_1 满足：

$$E(u_1) = 0, \quad E(z_1 u_1) = 0, \quad E(z_2 u_1) = 0$$

多元回归模型的IV估计

- 因为 $u_1 = y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1$,

$$E(y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1) = 0, \quad E(z_1(y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1)) = 0,$$
$$E(z_2(y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1)) = 0.$$

- 样本版本:

$$\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0, \quad \frac{1}{n} \sum_{i=1}^n (z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1})) = 0,$$
$$\frac{1}{n} \sum_{i=1}^n (z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1})) = 0.$$

- 定义 $X_i = (1, y_{i2}, z_{i1})'$, $Z_i = (1, z_{i1}, z_{i2})'$, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$,

$$\hat{\beta} = \left(\sum_{i=1}^n Z_i X_i' \right)^{-1} \sum_{i=1}^n Z_i y_{i1}$$

多元回归模型的IV估计

- 更一般的情况：

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$

内生的 外生变量

- 工具变量 z_k 的条件

- 1) 不出现在回归中
- 2) 与误差项不相关
- 3) 与内生解释变量相关

- u_1 满足：

$$E(u_1) = 0, \quad E(z_j u_1) = 0, \quad j = 1, \dots, k$$

多元回归模型的IV估计

- 样本版本：

$$\frac{1}{n} \sum_{i=1}^n (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \cdots - \hat{\beta}_k z_{ik-1}) = 0,$$

$$\frac{1}{n} \sum_{i=1}^n (z_{ij} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \cdots - \hat{\beta}_k z_{ik-1})) = 0,$$

- 用 $k+1$ 个方程来解 $k+1$ 个估计量 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$
- 定义 $X_i = (1, y_{i2}, z_{i1}, \dots, z_{ik-1})'$, $Z_i = (1, z_{i1}, z_{i2}, \dots, z_{ik})'$, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)'$,

$$\hat{\beta} = \left(\sum_{i=1}^n Z_i X_i' \right)^{-1} \sum_{i=1}^n Z_i y_{i1}$$



多元回归模型的IV估计

- 可以拓展到有多个内生变量与多个工具变量的情况
- 该方法要求工具变量的个数等于内生解释变量的个数
- 如果工具变量的个数小于内生解释变量的个数，则没有办法估计
- 如果工具变量的个数大于(或等于)内生解释变量的个数，则可以采用两阶段最小二乘法 (2SLS)

两阶段最小二乘

- 两阶段最小二乘 (2SLS) 估计

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$

$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

第一阶段 (= 简约型回归) :

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \dots + \hat{\pi}_k z_{k-1} + \hat{\pi}_k z_k$$

用外生信息来解释内生变量 y_2

额外的外生变量 (即工具变量)

第二阶段 (= 用第一阶段 y_2 的预测值来做一个OLS回归) :

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + error$$



两阶段最小二乘

- 为什么2SLS有效?
 - 因为 y_2 由其预测值所替代（是由外生变量预测的），所以在第二阶段的所有变量都是外生的
 - 因为使用由外生信息产生的预测值，所以 y_2 清除了其内生部分（即与误差项相关的部分）
- 2SLS的特征
 - 如果仅有一个内生变量且仅有一个IV，那么2SLS=IV
 - 如果仅有一个内生变量且工具变量有多个，则2SLS估计也可以使用
 - 如果内生变量有多个，第一阶段中对每个内生变量都进行OLS回归
 - 第二阶段的OLS回归的标准误是不准确的
 - 2SLS/IV一般来说精确性不如OLS，因为在第二阶段回归中的多重共线性以及更少的解释变量的变动

两阶段最小二乘

- 例子：在工资等式中使用两个IV的2SLS

第一步回归 (educ对所有外生变量做回归)：

$$\widehat{educ} = 8.37 + .085 exper - .002 exper^2$$

(.27) (.026) (.001)

$$+ .185 \boxed{fatheduc} + .186 \boxed{motheduc}$$

(.024) (.026)

教育程度显著与父母的教育程度相关

2SLS估计结果：

$$\widehat{\log(wage)} = .048 + \boxed{.061} educ + .044 exper - .0009 exper^2$$

(.400) (.031) (.013) (.0004)

$$n = 428, R^2 = 0.136$$

较OLS来说，教育回报低了很多，标准误也大了（不精确了）



两阶段最小二乘

- 检验弱工具变量

$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

- 通过t检验可以验证 z_k 与 y_2 的相关性：

$$H_0: \pi_k = 0$$

- 斯托克和约吉 (Stock and Yogo, 2005) 提出如果t统计量大于3.2 (多个工具变量F统计量大于10) 时, 工具变量可以继续使用。

内生性检验

- 检验解释变量的内生性

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$

怀疑这个变量是内生的

$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

变量 y_2 是外生的，当且仅当 v_2 与 u_1 不相关，即在下面回归中 δ_1 为0:

$$u_1 = \delta_1 v_2 + e_1$$

- 总结：变量 y_2 是外生的，即在下面回归中 δ_1 为0:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + \delta_1 v_2 + e_1$$



内生性检验

- 回归模型:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + \delta_1 \hat{v}_2 + e_1$$

- 检验原假设

$$H_0: \delta_1 = 0$$

- 如果 δ_1 显著异于0， y_2 是外生的原假设被拒绝
- 注意：检验内生性也需要工具变量



本章小节

- 在这一章中，我们介绍了内生性以及解决办法（工具变量方法）
- 我们首先讨论了内生性的成因以及影响
- 工具变量必须具备两个性质:(1)它必须是外生的 (2) 它必须与内生解释变量相关
- 当工具变量个数等于内生变量个数时，我们可以从矩条件出发获得一致估计值
- 当工具变量个数大于或等于内生变量个数时，我们可以使用2SLS方法
- 我们讨论了弱工具变量检验以及内生性检验问题