

第 10 章 工具变量法

OLS 能够成立的最重要条件是解释变量与扰动项不相关(前定变量或同期外生)。否则，OLS 不一致。

但解释变量与扰动项相关(内生性)的例子比比皆是。

解决内生性的主要方法之一为工具变量法。

内生性的来源包括遗漏变量偏差、联立方程偏差(双向因果关系)，及测量误差偏差(measurement error bias)。

10.1 联立方程偏差

例 考察农产品市场均衡模型：

$$\begin{cases} q_t^d = \alpha + \beta p_t + u_t & (\text{需求}) \\ q_t^s = \gamma + \delta p_t + v_t & (\text{供给}) \\ q_t^d = q_t^s & (\text{均衡}) \end{cases} \quad (10.1)$$

q_t^d 为农产品需求， q_t^s 为农产品供给， p_t 为农产品价格。

市场出清(market clearing)的均衡条件要求 $q_t^d = q_t^s$ 。

令 $q_t \equiv q_t^d = q_t^s$ ，可得

$$\begin{cases} q_t = \alpha + \beta p_t + u_t \\ q_t = \gamma + \delta p_t + v_t \end{cases} \quad (10.2)$$

两个方程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？

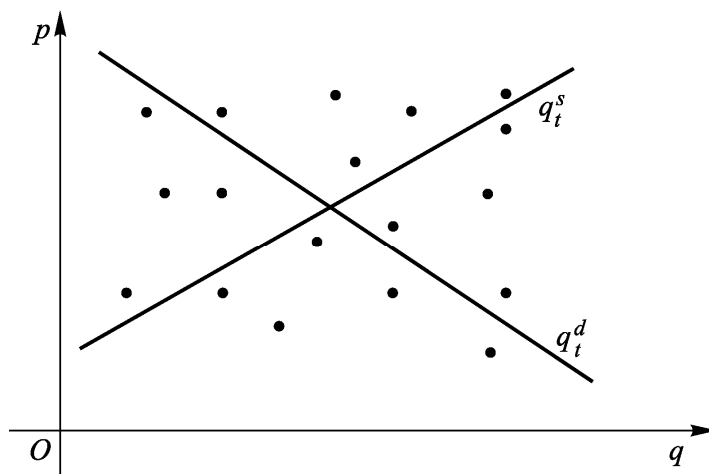


图 10.1 需求与供给决定市场均衡

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。

故解释变量 p_t 与两个方程的扰动项 (u_t, v_t) 都相关, 即 $\text{Cov}(p_t, u_t) \neq 0$, $\text{Cov}(p_t, v_t) \neq 0$ 。

对于需求函数的正冲击($u_t > 0$), 使均衡价格 p_t 上升, 故二者正相关。

对于供给函数的正冲击($v_t > 0$), 使均衡价格 p_t 下降, 故二者负相关。

故 OLS 不一致, 称为“联立方程偏差”(simultaneity bias)或“内生性偏差”(endogeneity bias)。

例 考察宏观经济模型中的消费函数：

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t \\ Y_t = C_t + I_t + G_t + X_t \end{cases} \quad (10.3)$$

Y_t, C_t, I_t, G_t, X_t 分别为国民收入、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。

如单独对消费方程进行 OLS 回归，存在联立方程偏差，得不到一致估计。

10.2 测量误差偏差

内生性的另一来源是解释变量的测量误差(measurement error 或 errors-in-variables)。

例 假设真实模型为

$$y = \alpha + \beta x^* + \varepsilon \quad (10.4)$$

其中, $\beta \neq 0$, $\text{Cov}(x^*, \varepsilon) = 0$ 。

x^* 无法观测, 只能观测到 x , 二者满足如下关系:

$$x = x^* + u \quad (10.5)$$

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

将表达式(10.5)代入方程(10.4)可得

$$y = \alpha + \beta x + (\varepsilon - \beta u) \quad (10.6)$$

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

$$\begin{aligned} \text{Cov}(x, \varepsilon - \beta u) &= \text{Cov}(x^* + u, \varepsilon - \beta u) \\ &= \underbrace{\text{Cov}(x^*, \varepsilon)}_{=0} - \beta \underbrace{\text{Cov}(x^*, u)}_{=0} + \underbrace{\text{Cov}(u, \varepsilon)}_{=0} - \beta \text{Cov}(u, u) \\ &= -\beta \text{Var}(u) \neq 0 \end{aligned} \quad (10.7)$$

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

如果被解释变量存在测量误差，后果却不严重。

比如，只要被解释变量的测量误差与解释变量不相关，则 OLS 依然一致(参见习题)。

10.3 工具变量法

OLS 不一致因内生变量与扰动项相关而引起。

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。

通常借助另外一个“工具变量”实现这种分离。

假设在图 10.1 中，存在某因素使供给曲线经常移动，而需求曲

线基本不动。

可估计需求曲线，参见图 10.2。使得供给曲线移动的变量就是工具变量。

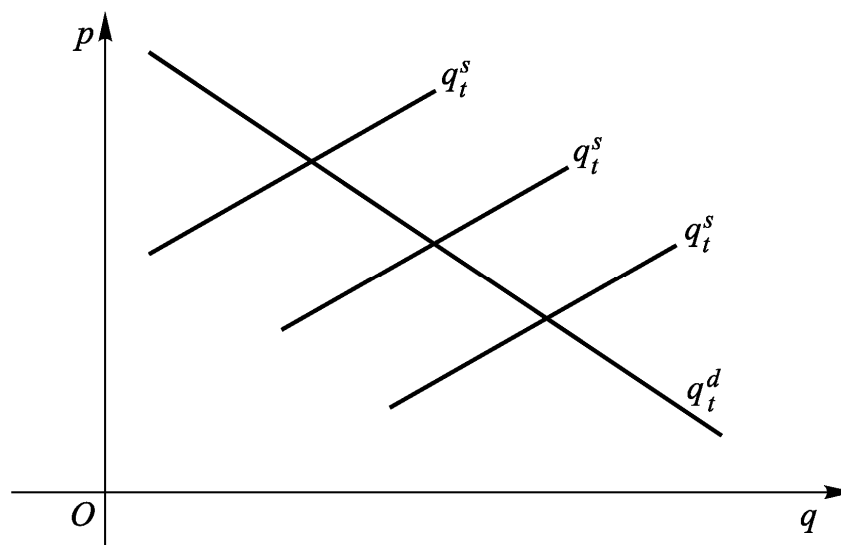


图 10.2 稳定的需求与变动的供给

假设影响供给方程扰动项的因素可分解为两部分，即可观测的气温 z_t 与不可观测的其他因素：

$$q_t^s = \gamma + \delta p_t + \eta z_t + v_t \quad (10.8)$$

假定气温 z_t 是前定变量，与需求方程的扰动项不相关，即 $\text{Cov}(z_t, u_t) = 0$ 。

由于气温 z_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 。

称 z_t 为“工具变量” (Instrumental Variable, 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(relevance): 工具变量与内生解释变量相关, 即 $\text{Cov}(z_t, p_t) \neq 0$ 。

(ii) 外生性(exogeneity): 工具变量与扰动项不相关, 即 $\text{Cov}(z_t, u_t) = 0$ 。

在本例中, 气温 z_t 满足这两个条件。

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

(ii) 外生性: 假设气温 z_t 是前定变量, $\text{Cov}(z_t, u_t) = 0$ 。

利用工具变量的这两个性质,可得到对需求方程回归系数 β 的一致估计。

需求方程为

$$q_t = \alpha + \beta p_t + u_t \quad (10.9)$$

两边同时求与 z_t 的协方差:

$$\begin{aligned} \text{Cov}(q_t, z_t) &= \text{Cov}(\alpha + \beta p_t + u_t, z_t) \\ &= \beta \text{Cov}(p_t, z_t) + \underbrace{\text{Cov}(u_t, z_t)}_{=0} = \beta \text{Cov}(p_t, z_t) \end{aligned} \quad (10.10)$$

由于工具变量的外生性, 故 $\text{Cov}(u_t, z_t) = 0$ 。

根据工具变量的相关性, $\text{Cov}(p_t, z_t) \neq 0$ 。

两边同除 $\text{Cov}(p_t, z_t)$:

$$\beta = \frac{\text{Cov}(q_t, z_t)}{\text{Cov}(p_t, z_t)} \quad (10.11)$$

以样本矩取代总体矩(以样本协方差替代总体协方差), 可得一致的“工具变量估计量”(Instrumental Variable Estimator):

$$\hat{\beta}_{\text{IV}} = \frac{\widehat{\text{Cov}(q_t, z_t)}}{\widehat{\text{Cov}(p_t, z_t)}} = \frac{\sum_{t=1}^n (q_t - \bar{q})(z_t - \bar{z})}{\sum_{t=1}^n (p_t - \bar{p})(z_t - \bar{z})} \xrightarrow{p} \frac{\text{Cov}(q_t, z_t)}{\text{Cov}(p_t, z_t)} = \beta \quad (10.12)$$

$\bar{q}, \bar{p}, \bar{z}$ 分别为 q, p, z 的样本均值。

如工具变量与内生变量无关, $\text{Cov}(z_t, p_t) = 0$, 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, $\text{Cov}(z_t, p_t) \approx 0$, 会导致估计量 $\hat{\beta}_{IV}$ 的方差变得很大, 称为“弱工具变量问题”。

10.4 二阶段最小二乘法

工具变量法一般通过“二阶段最小二乘法”(Two Stage Least Square, 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 $p_t \xrightarrow{\text{OLS}} z_t$, 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一致估计？

首先，把需求方程 $q_t = \alpha + \beta p_t + u_t$ 分解为

$$q_t = \alpha_0 + \beta \hat{p}_t + \underbrace{[u_t + \beta(p_t - \hat{p}_t)]}_{\equiv \varepsilon_t} \quad (10.13)$$

这是第二阶段的回归，扰动项为 $\varepsilon_t \equiv u_t + \beta(p_t - \hat{p}_t)$ 。

命题 在第二阶段回归中， \hat{p}_t 与扰动项 ε_t 不相关。

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

$$\text{Cov}(\hat{p}_t, \varepsilon_t) = \text{Cov}(\hat{p}_t, u_t) + \beta \text{Cov}(\hat{p}_t, p_t - \hat{p}_t) \quad (10.14)$$

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

其次，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 $\text{Cov}(\hat{p}_t, p_t - \hat{p}_t) = 0$ 。

第二阶段回归的解释变量 \hat{p}_t 与扰动项 ε_t 不相关，故 2SLS 一致。

2SLS 的实质：把内生解释变量 p_t 分成两部分，由工具变量 z_t 所造成的外生部分(\hat{p}_t)，及与扰动项相关的其余部分($p_t - \hat{p}_t$)；

把被解释变量 q_t 对 p_t 中的外生部分(\hat{p}_t)进行回归,从而满足 OLS 对前定变量的要求而得到一致估计。

如存在多个工具变量,仍可用 2SLS 法。

假设 z_1 与 z_2 为两个有效工具变量(满足相关性与外生性),则第一阶段回归变为

$$p = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + u \quad (10.15)$$

可得拟合值 $\hat{p} = \hat{\alpha}_0 + \hat{\alpha}_1 z_1 + \hat{\alpha}_2 z_2$,而第二阶段回归不变。

考虑多个内生变量的情形:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (10.16)$$

其中, x_1 与 x_2 均内生,都与 ε 相关。

由于有两个内生变量，至少需要两个工具变量，才能进行 2SLS 估计。

如只有一个工具变量 z ，由第一阶段回归可得， $\hat{x}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 z$ ， $\hat{x}_2 = \hat{\gamma}_0 + \hat{\gamma}_1 z$ 。

将 \hat{x}_1 与 \hat{x}_2 代入原方程：

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + v_t \quad (10.17)$$

\hat{x}_1 与 \hat{x}_2 都是 z 的线性函数，故存在严格多重共线性(参见习题)。

阶条件：进行 2SLS 估计的必要条件是工具变量个数不少于内生解释变量的个数，称为“阶条件”(order condition)。

根据阶条件是否满足可分为三种情况：

(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。

在恰好识别与过度识别的情况下，都可使用 2SLS；

在不可识别的情况下，无法使用 2SLS。

考虑多个内生变量，且包含外生解释变量的情形：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 w + \varepsilon \quad (10.18)$$

x_1 与 x_2 为内生变量， w 为外生变量(与 ε 不相关)。假设有三个有效工具变量 z_1, z_2, z_3 。

在第一阶段回归中，分别将两个内生变量(x_1, x_2)对所有外生变量(包括工具变量 z_1, z_2, z_3 及外生变量 w)回归：

$$x_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 w + u \quad (10.19)$$

$$x_2 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 w + v \quad (10.20)$$

外生变量 w 可视为自己的工具变量，因为满足工具变量的定义。

首先， w 与 w 高度相关，满足相关性。

其次， w 与扰动项 ε 不相关，因为 w 为外生变量。

有时也称 z_1, z_2, z_3 为方程外的工具变量。

将方程(10.19)与(10.20)的拟合值分别记为 \hat{x}_1 与 \hat{x}_2 ，并代入原方程，进行第二阶段回归：

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \beta_3 w + \xi \quad (10.21)$$

ξ 为第二阶段回归的扰动项。记此估计量为 $\hat{\beta}_{IV}$ 。

$\hat{\beta}_{IV}$ 为 β 的一致估计，且渐近正态，可照常进行大样本统计推断。

2SLS 的第二阶段回归就是 OLS，故 $\hat{\beta}_{IV}$ 的协方差矩阵 $\text{Var}(\hat{\beta}_{IV})$ 在形式上与 OLS 估计量相似。

考虑到可能存在异方差，建议使用异方差稳健的标准误。

需要注意，第二阶段回归所得残差为

$$\hat{\xi} \equiv y - (\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_1 + \hat{\beta}_2 \hat{x}_2 + \hat{\beta}_3 w) \quad (10.22)$$

原方程真正的残差却是

$$e_{IV} \equiv y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 w) \quad (10.23)$$

二者并不相等，即 $e_{IV} \neq \hat{\xi}$ 。

进行 2SLS 估计，最好不要自己手工进行两次回归，而直接使用 Stata 命令。

2SLS 的 Stata 命令格式为

```
ivregress 2sls y x1 x2 (x3 = z1 z2),_robust first
```

其中，“y”为被解释变量，“x1 x2”为外生解释变量，“x3”为内生解释变量，而“z1 z2”为方程外的工具变量。

选择项“robust”表示使用异方差稳健的标准误(默认为普通标准误)；

选择项 “first” 表示显示第一阶段的回归结果。

在球形扰动项的情况下，2SLS 是最有效率的工具变量法。

在异方差的情况下，存在更有效率的工具变量法，即“广义矩估计” (Generalized Method of Moments, GMM)。

GMM 是数理统计“矩估计” (Method of Moments, MM) 的推广。

GMM 之于 2SLS，正如 GLS 与 OLS 的关系。

在恰好识别或同方差的情况下，GMM 等价于 2SLS。

10.5 弱工具变量

如工具变量与内生变量仅微弱相关， $\hat{\beta}_{IV}$ 的方差将变得很大。

由于工具变量仅包含极少与内生变量有关的信息，利用这部分信息进行的工具变量法估计就不准确。

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，会导致 $\hat{\beta}_{IV}$ 的小样本性质变得很差，即 $\hat{\beta}_{IV}$ 的小样本真实分布离大样本的渐近正态分布相去甚远，致使基于大样本的统计推断失效。

为检验是否存在弱工具变量，可在第一阶段回归中，检验所有方程外的工具变量的系数是否联合为零。

假设原模型为

$$y = \beta_0 + \beta_1 x + \beta_2 w + \varepsilon \quad (10.24)$$

x 为内生变量， w 为外生变量。

假设有两个有效工具变量 z_1, z_2 ，第一阶段回归为

$$x = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 w + u \quad (10.25)$$

检验 $H_0: \alpha_1 = \alpha_2 = 0$ 。由于工具变量的强弱连续变化，很难确定明确的检验标准。

经验规则：此检验的 F 统计量大于 10 (由于技术性原因，此处使用普通标准误)，则拒绝“存在弱工具变量”的原假设。

在 Stata 作完 2SLS 回归后，可使用以下命令检验弱工具变量：

```
estat firststage
```

此命令将根据第一阶段回归计算一些统计量，包括上文的 F 统计量。

如发现存在弱工具变量，可能的解决方法包括：

(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, LIML)。

在大样本下, LIML 与 2SLS 渐近等价。

在弱工具变量的情况下, LIML 的小样本性质可能优于 2SLS。

LIML 的 Stata 命令为

```
ivregress liml y x1 x2(x3 = z1 z2)
```

此命令在格式上与“ivregress 2s1s”(2SLS)完全相同。

10.6 对工具变量外生性的过度识别检验

工具变量的外生性是保证 2SLS 一致性的重要条件。

如果“工具变量”与扰动项相关,可导致严重的偏差(参见习题)。

在恰好识别的情况下,无法检验工具变量的外生性。

只能进行定性讨论或依赖于专家的意见。

定性讨论: 如果工具变量外生, 则它影响被解释变量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，或对扰动项有影响。

此条件称为“排他性约束”(exclusion restriction)，它排除了工具变量除了通过内生变量而影响被解释变量的其他渠道。

实践中，需找出工具变量影响被解释变量的所有其他可能渠道，然后一一排除，才能说明工具变量的外生性。

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

过度识别检验的大前提(maintained hypothesis)是该模型至少恰好识别，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 ：所有工具变量都外生

如拒绝原假设，则认为至少某个变量与扰动项相关。

假设共有 K 个解释变量 $\{x_1, \dots, x_K\}$ ，其中前 $(K-r)$ 个解释变量 $\{x_1, \dots, x_{K-r}\}$ 为外生变量，而后 r 个解释变量 $\{x_{K-r+1}, \dots, x_K\}$ 为内生变量：

$$y = \underbrace{\beta_1 x_1 + \cdots + \beta_{K-r} x_{K-r}}_{\text{外生}} + \underbrace{\beta_{K-r+1} x_{K-r+1} + \cdots + \beta_K x_K}_{\text{内生}} + \varepsilon \quad (10.26)$$

假设共有 m 个方程外的工具变量 $\{z_1, \dots, z_m\}$, 其中 $m > r$; 则过度识别的原假设为

$$H_0 : \text{Cov}(z_1, \varepsilon) = 0, \dots, \text{Cov}(z_m, \varepsilon) = 0 \quad (10.27)$$

通过 2SLS 的残差 e_{IV} 考察工具变量与扰动项的相关性。

把 e_{IV} 对所有外生变量(所有外生变量与工具变量)进行辅助回归:

$$e_{IV} = \gamma_1 x_1 + \cdots + \gamma_{K-r} x_{K-r} + \delta_1 z_1 + \cdots + \delta_m z_m + error \quad (10.28)$$

原假设(10.27)可写为

$$H_0: \delta_1 = \cdots = \delta_m = 0 \quad (10.29)$$

记辅助回归的可决系数为 R^2 ，Sargan 统计量为

$$nR^2 \xrightarrow{d} \chi^2(m-r) \quad (10.30)$$

Sargan 统计量的渐近分布为 $\chi^2(m-r)$ ，其自由度 $(m-r)$ 是过度识别约束的个数，即方程外工具变量个数 (m) ，减去内生变量个数 (r) ，也就是“多余”的工具变量个数。

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

此检验的直观思想：在过度识别的情况下，可用不同的工具变量组合来进行工具变量法估计；

如果所有工具变量都有效，则这些工具变量估计量 $\hat{\beta}_{IV}$ 都将收敛到相同的真实参数 β 。

可检验不同的工具变量估计量之间的差是否收敛于 0 ；如果不是，则说明这些工具变量不全有效。

在恰好识别的情况下，只有唯一的工具变量估计量，无法进行比较，故过度识别检验失效。

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。

因为过度识别检验成立的大前提是，该模型至少恰好识别。

此大前提无法检验，只能假定成立。

如只有一个内生变量，在进行过度识别检验时，隐含地假定至少有一个工具变量外生，然后检验所有其他工具变量的外生性。

即使不同的工具变量估计量 $\hat{\beta}_{IV}$ 的概率极限相同，并不能保证它们都收敛到真实的参数 β ；也可能都收敛到其他值，比如 $\beta^* \neq \beta$ 。

恰好识别的大前提保证了，在这些工具变量估计量至少有一个收敛到真实参数。

在 Stata 作完 2SLS 估计后，可用以下命令进行过度识别检验：
`estat overid`

10.7 对解释变量内生性的豪斯曼检验：究竟该用 OLS 还是 IV

使用工具变量法的前提是存在内生解释变量。

如何检验解释变量是否内生？

扰动项不可观测，无法直接检验解释变量与扰动项的相关性。

如找到有效工具变量，可借助工具变量来检验。

假设存在方程外的有效工具变量。

如果所有解释变量都外生，则 OLS 与 IV 都一致，但 OLS 比 IV 更有效。

在这种情况下，虽然 IV 一致，但相当于“无病用药”，反而增大估计量的方差。

反之，如果存在内生变量，则 OLS 不一致，而 IV 一致。

“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，以二次型度量此距离：

$$(\hat{\beta}_{IV} - \hat{\beta}_{OLS})' \left[\overline{\text{Var}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})} \right]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \xrightarrow{d} \chi^2(r) \quad (10.31)$$

其中， r 为内生解释变量的个数， $\left[\overline{\text{Var}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})} \right]$ 为 $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$ 的协方差矩阵之样本估计值。

如果此豪斯曼统计量很大，超过了其渐近分布 $\chi^2(r)$ 的临界值，则拒绝“所有解释变量均外生”的原假设，认为存在内生变量，应使用 IV。

豪斯曼检验的 Stata 命令为:

```
reg y x1 x2
```

```
estimates store ols      (存储 OLS 的结果, 记为 ols)
```

```
ivregress 2sls y x1 (x2=z1 z2)      (假设 x2 为内生  
变量, z1,z2 为 IV)
```

```
estimates store iv      (存储 2SLS 的结果, 记为 iv)
```

```
hausman iv ols,constant sigmamore (进行豪斯曼检验)
```

选择项“sigmamore”表示统一使用更有效率的估计量(即 OLS)所对应的残差来计算扰动项方差 $\hat{\sigma}^2$ 。这样有助于保证根据样本数据计算的 $\left[\overline{\text{Var}(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})} \right]$ 为正定矩阵。

选择项“constant”表示 $\hat{\beta}_{\text{IV}}$ 与 $\hat{\beta}_{\text{OLS}}$ 中都包括常数项(默认不含常数项)。

传统豪斯曼检验的缺点是，为简化矩阵 $\left[\overline{\text{Var}(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})}\right]$ 的计算，假设在 H_0 成立的情况下，OLS 最有效率，故不适用异方差的情形(OLS 只在球形扰动项的情况下才最有效率)。

改进的“杜宾-吴-豪斯曼检验”(Durbin-Wu-Hausman Test, DWH)在异方差的情况下也适用。

在 Stata 中作完 2SLS 估计后，可输入以下命令进行异方差稳健的 DWH 检验：

```
estat endogenous
```


10.8 如何获得工具变量

工具变量的两个要求(相关性与外生性)常自相矛盾。

寻找合适的工具变量通常较困难，需要一定的创造性与想象力。

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。

第(ii)步的操作较难，因为扰动项不可观测。

如何判断候选变量(z)是否与扰动项(ε)相关？

由于扰动项是 y 的扰动项，可从 z 与 y 的相关性着手。

z 与 y 相关，因为 z 与内生变量 x 相关。

但 z 对 y 的影响仅通过 x 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。

是否“ z 对 y 的影响仅通过 x 起作用”，可通过定性讨论来确定，即“排他性约束”(exclusion restriction)。

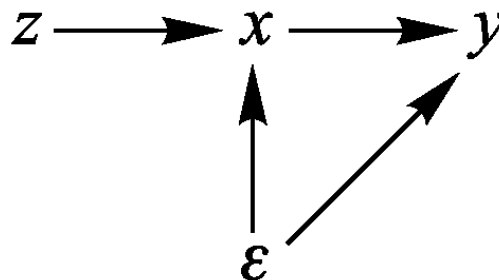


图 10.3 工具变量示意图

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作为工具变量。

相关性：内生解释变量与其滞后变量相关。

外生性：由于滞后变量已经发生(从当期的角度看，其取值已经固定)，故可能“前定”，与当期扰动项不相关。

比如，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。

奖金占员工报酬比重越高，则越能促进生产率的提高。

但生产率越高的企业越有能力给员工发奖金,存在双向因果。

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,当期生产率不可能影响过去的奖金比重,故奖金比重的滞后值(可能)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,犯罪率应越低。

但城市犯罪率高,政府可能增加警察人数。

Levitt(1997)使用“市长选举的政治周期”作为犯罪率的工具变量。

市长竞选连任时，为拉选票，会增加警察人数以保证治安，故满足相关性。

选举周期以机械方式确定，除影响警察人数外，不单独对犯罪率起作用，故满足外生性。

例 制度对经济增长的影响。

好制度促进经济增长，但制度变迁也依赖经济增长。

从历史角度，Acemoglu *et al.* (2001)使用“殖民者死亡率”(settler mortality)作为制度的工具变量。

欧洲殖民者在全球殖民时，由于各地气候及疾病环境(disease environment)不同，殖民者死亡率十分不同。

在死亡率高的地方(比如非洲),殖民者难以定居,在当地建立掠夺性制度(extractive institutions)。

在死亡率低的地方(比如北美),建立有利于经济增长的制度(比如较好的产权保护)。

初始制度上的差异一直延续到今天。故殖民者死亡率与今天的制度相关,满足相关性。

殖民者死亡率除影响制度外,不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？

在美国，电视普及与小儿自闭症发生率的攀升几乎同步。

Waldman *et al.* (2006, 2008)研究看电视是否引发小儿自闭症。

有自闭倾向的儿童可能更经常看电视，存在双向因果关系。

使用降雨量作为电视观看时间的工具变量。

降雨越多的地区，人们呆在室内时间越长，看电视时间也越长，
故相关

降雨量很可能外生（只通过看电视时间而影响被解释变量）。

研究结果支持过多观看电视为小儿自闭症的诱因。

10.9 工具变量法的 Stata 实例

以数据集 `grilic.dta` 为例，继续探讨教育投资回报率。

此数据集的主要变量包括：`lnw`(工资对数)，`s`(教育年限)，`expr`(工龄)，`tenure`(在现单位的工作年数)，`iq`(智商)，`med`(母亲的教育年限)，`kww`(在“knowledge of the World of Work”测试中的成绩)，`rns`(美国南方虚拟变量，住在南方=1)，`smsa`(大城市虚拟变量，住在大城市=1)。

(1) 作为参照系，首先进行 OLS 回归，并使用稳健标准误。

```
. reg lnw s expr tenure rns smsa,r
```

Linear regression					Number of obs = 758	
					F(5, 752) = 84.05	
					Prob > F = 0.0000	
					R-squared = 0.3521	
					Root MSE = .34641	
lnw	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
s	.102643	.0062099	16.53	0.000	.0904523	.1148338
expr	.0381189	.0066144	5.76	0.000	.025134	.0511038
tenure	.0356146	.0079988	4.45	0.000	.0199118	.0513173
rns	-.0840797	.029533	-2.85	0.005	-.1420566	-.0261029
smsa	.1396666	.028056	4.98	0.000	.0845893	.194744
_cons	4.103675	.0876665	46.81	0.000	3.931575	4.275775

教育投资的年回报率高达 10.26%（似乎太高），且在 1%的水平上显著。可能遗漏“能力”，高估了教育的回报率。

(2) 引入智商(iq)作为“能力”的代理变量，再进行 OLS 回归。

```
. reg lnw s iq expr tenure rns smsa,r
```

Linear regression				Number of obs = 758		
				F(6, 751) = 71.89		
				Prob > F = 0.0000		
				R-squared = 0.3600		
				Root MSE = .34454		
lnw	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
s	.0927874	.0069763	13.30	0.000	.0790921	.1064826
iq	.0032792	.0011321	2.90	0.004	.0010567	.0055016
expr	.0393443	.0066603	5.91	0.000	.0262692	.0524193
tenure	.034209	.0078957	4.33	0.000	.0187088	.0497092
rns	-.0745325	.0299772	-2.49	0.013	-.1333815	-.0156834
smsa	.1367369	.0277712	4.92	0.000	.0822186	.1912553
_cons	3.895172	.1159286	33.60	0.000	3.667589	4.122754

教育投资的回报率下降为 9.28%，更为合理些，但仍然过高。

(3) 由于用 iq 度量能力存在“测量误差”，故 iq 是内生变量。

使用变量(med, kww)作为 iq 的工具变量。

母亲的教育年限(med)与 KWW 测试成绩(kww)都与 iq 正相关；
并假设 med 与 kww 为外生。

进行 2SLS 回归，使用稳健标准误，显示第一阶段的回归结果。

```
. ivregress 2sls lnw s expr tenure rns smsa  
(iq=med kww),r first
```

First-stage regressions

Number of obs = 758
 F(7, 750) = 47.74
 Prob > F = 0.0000
 R-squared = 0.3066
 Adj R-squared = 0.3001
 Root MSE = 11.3931

iq	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
s	2.467021	.2327755	10.60	0.000	2.010052	2.92399
expr	-.4501353	.2391647	-1.88	0.060	-.9196471	.0193766
tenure	.2059531	.269562	0.76	0.445	-.3232327	.7351388
rns	-2.689831	.8921335	-3.02	0.003	-4.441207	-.938455
smsa	.2627416	.9465309	0.28	0.781	-1.595424	2.120907
med	.3470133	.1681356	2.06	0.039	.0169409	.6770857
kww	.3081811	.0646794	4.76	0.000	.1812068	.4351553
_cons	56.67122	3.076955	18.42	0.000	50.63075	62.71169

Instrumental variables (2SLS) regression					Number of obs = 758	
					Wald chi2(6) = 370.04	
					Prob > chi2 = 0.0000	
					R-squared = 0.2775	
					Root MSE = .36436	
lnw	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
iq	.0139284	.0060393	2.31	0.021	.0020916	.0257653
s	.0607803	.0189505	3.21	0.001	.023638	.0979227
expr	.0433237	.0074118	5.85	0.000	.0287968	.0578505
tenure	.0296442	.008317	3.56	0.000	.0133432	.0459452
rns	-.0435271	.0344779	-1.26	0.207	-.1111026	.0240483
smsa	.1272224	.0297414	4.28	0.000	.0689303	.1855146
_cons	3.218043	.3983683	8.08	0.000	2.437256	3.998831
Instrumented: iq						
Instruments: s expr tenure rns smsa med kww						

教育投资回报率降为 6.08%，且在 1% 水平上显著；比较合理。

(4) 过度识别检验:

```
. estat overid
```

```
Test of overidentifying restrictions:
```

```
Score chi2(1)          =   .151451   (p = 0.6972)
```

p 值为 0.697，故接受原假设，认为(med, kww)外生。

(5) 工具变量与内生变量的相关性。

从第一阶段的回归结果可知，工具变量(med, kww)对内生变量 iq 有较好解释力， p 值都小于 0.05。

正式检验须计算第一阶段回归的普通(非稳健) F 统计量。

使用普通标准误重新进行 2SLS 估计。

```
. quietly ivregress 2sls lnw s expr tenure rns  
smsa (iq=med kww)
```

```
. estat firststage
```

First-stage regression summary statistics					
Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(2,750)	Prob > F
iq	0.3066	0.3001	0.0382	14.9058	0.0000

Minimum eigenvalue statistic = 14.9058

Critical Values	# of endogenous regressors:	1
Ho: Instruments are weak	# of excluded instruments:	2

	5%	10%	20%	30%
2SLS relative bias	(not available)			
	10%	15%	20%	25%
2SLS Size of nominal 5% Wald test	19.93	11.59	8.75	7.25
LIML Size of nominal 5% Wald test	8.68	5.33	4.42	3.92

由于 F 统计量为 14.91，超过 10，故认为不存在弱工具变量。

(6) 使用对弱工具变量更不敏感的有限信息最大似然法(LIML):

```
. ivregress liml lnw s expr tenure rns smsa
(iq=med kww),r
```

Instrumental variables (LIML) regression					Number of obs = 758	
					Wald chi2(6) = 369.62	
					Prob > chi2 = 0.0000	
					R-squared = 0.2768	
					Root MSE = .36454	
lnw	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
iq	.0139764	.0060681	2.30	0.021	.0020831	.0258697
s	.0606362	.019034	3.19	0.001	.0233303	.0979421
expr	.0433416	.0074185	5.84	0.000	.0288016	.0578816
tenure	.0296237	.008323	3.56	0.000	.0133109	.0459364
rns	-.0433875	.034529	-1.26	0.209	-.1110631	.0242881
smsa	.1271796	.0297599	4.27	0.000	.0688512	.185508
_cons	3.214994	.4001492	8.03	0.000	2.430716	3.999272
Instrumented: iq						
Instruments: s expr tenure rns smsa med kww						

LIML 估计值与 2SLS 非常接近,侧面印证“不存在弱工具变量”。

(7) 使用工具变量法的前提是存在内生解释变量。为此进行豪斯曼检验，原假设为“所有解释变量均为外生”。

```
. quietly reg lnw iq s expr tenure rns smsa  
  
. estimates store ols  
  
. quietly ivregress 2sls lnw s expr tenure rns  
smsa (iq=med kww)  
  
. estimates store iv  
  
. hausman iv ols,constant sigmamore
```

传统的豪斯曼检验假定同方差，故在回归中未使用稳健标准误。

Note: the rank of the differenced variance matrix (1) does not equal the number of coefficients being tested (7); be sure this is what you expect, or there may be problems computing the test. Examine the output of your estimators for anything unexpected and possibly consider scaling your variables so that the coefficients are on a similar scale.

	—— Coefficients ——			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	iv	ols	Difference	S.E.
iq	.0139284	.0032792	.0106493	.0054318
s	.0607803	.0927874	-.032007	.0163254
expr	.0433237	.0393443	.0039794	.0020297
tenure	.0296442	.034209	-.0045648	.0023283
rns	-.0435271	-.0745325	.0310054	.0158145
smsa	.1272224	.1367369	-.0095145	.0048529
_cons	3.218043	3.895172	-.6771285	.3453751

b = consistent under Ho and Ha; obtained from ivregress
 B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
 = 3.84
 Prob>chi2 = 0.0499
 (V_b-V_B is not positive definite)

p 值(Prob>chi2)为 0.0499, 可在 5%水平上拒绝“所有解释变量均为外生”的原假设, 认为 iq 内生。

传统的豪斯曼检验在异方差下不成立，下面进行异方差稳健的 DWH 检验：

```
. estat endogenous
```

Tests of endogeneity			
Ho: variables are exogenous			
Durbin (score) chi2(1)	=	3.87962	(p = 0.0489)
Wu-Hausman F(1,750)	=	3.85842	(p = 0.0499)

上表提供了 F 统计量与 χ^2 统计量，二者在大样本下渐近等价。

二者的 p 值都小于 0.05，故认为 iq 内生。

(8) 汇报结果：将以上各种估计法的系数及标准误列在同一表格中，可使用以下命令：

```
. qui reg lnw s expr tenure rns smsa,r
. est sto ols_no_iq
. qui reg lnw iq s expr tenure rns smsa,r
. est sto ols_with_iq
. qui ivregress 2sls lnw s expr tenure rns smsa
(iq=med kww),r
. est sto tsls
. qui ivregress liml lnw s expr tenure rns smsa
(iq=med kww),r
. est sto liml
. estimates table ols_no_iq ols_with_iq tsls
liml,b se
```

其中，选择项“b”表示显示回归系数，“se”表示显示标准误。

Variable	ols_no_iq	ols_with~q	tsls	liml
s	.10264304	.09278735	.06078035	.06063623
	.00620988	.00697626	.01895051	.01903397
expr	.0381189	.03934425	.04332367	.04334159
	.00661439	.00666033	.00741179	.0074185
tenure	.03561456	.03420896	.02964421	.02962365
	.00799884	.00789567	.00831697	.00832297
rns	-.08407974	-.07453249	-.04352713	-.04338751
	.02953295	.02997719	.03447789	.03452902
smsa	.13966664	.13673691	.12722244	.1271796
	.02805598	.02777116	.02974144	.02975994
iq		.00327916	.01392844	.01397639
		.00113212	.00603931	.00606812
_cons	4.103675	3.8951718	3.2180433	3.2149943
	.08766646	.11592863	.39836829	.40014925

legend: b/se

用一颗星表示 10%的显著性，两颗星表示 5%的显著性，三颗星表示 1%的显著性，可使用如下命令：

```
. estimates table ols_no_iq ols_with_iq tsls
liml,star(0.1 0.05 0.01)
```

Variable	ols_no_iq	ols_with_iq	tsls	liml
s	.10264304***	.09278735***	.06078035***	.06063623***
expr	.0381189***	.03934425***	.04332367***	.04334159***
tenure	.03561456***	.03420896***	.02964421***	.02962365***
rns	-.08407974***	-.07453249**	-.04352713	-.04338751
smsa	.13966664***	.13673691***	.12722244***	.1271796***
iq		.00327916***	.01392844**	.01397639**
_cons	4.103675***	3.8951718***	3.2180433***	3.2149943***

legend: * p<.1; ** p<.05; *** p<.01

Stata 官方命令 “estimates table” 无法同时显示回归系数、标准误与表示显著性的星号。

下载非官方命令 “estout”。

```
. ssc install estout
```

```
. esttab ols_no_iq ols_with_iq tsls liml,se r2  
mtitle star(* 0.1 ** 0.05 *** 0.01)
```

选择项“se”表示在括弧中显示标准误(默认显示 t 统计量, 如果使用选择项“p”则显示 p 值)。

选择项“r2”表示显示 R^2 。

选择项“mtitle”表示使用模型名称(model title)作为表中每列的标题(默认使用被解释变量作为标题)

选择项“star(* 0.1 ** 0.05 *** 0.01)”表示以星号表示显著性水平。

	(1) ols_no_iq	(2) ols_with_iq	(3) tsls	(4) liml
s	0.103*** (0.00621)	0.0928*** (0.00698)	0.0608*** (0.0190)	0.0606*** (0.0190)
expr	0.0381*** (0.00661)	0.0393*** (0.00666)	0.0433*** (0.00741)	0.0433*** (0.00742)
tenure	0.0356*** (0.00800)	0.0342*** (0.00790)	0.0296*** (0.00832)	0.0296*** (0.00832)
rns	-0.0841*** (0.0295)	-0.0745** (0.0300)	-0.0435 (0.0345)	-0.0434 (0.0345)
smsa	0.140*** (0.0281)	0.137*** (0.0278)	0.127*** (0.0297)	0.127*** (0.0298)
iq		0.00328*** (0.00113)	0.0139** (0.00604)	0.0140** (0.00607)
_cons	4.104*** (0.0877)	3.895*** (0.116)	3.218*** (0.398)	3.215*** (0.400)
N	758	758	758	758
R-sq	0.352	0.360	0.278	0.277
Standard errors in parentheses				
* p<0.1, ** p<0.05, *** p<0.01				

如要将上表输出到 Microsoft Word 文档，并以文件名 iv 来命名此文档，可运行如下命令：

```
. esttab ols_no_iq ols_with_iq tsls liml using  
iv.rtf,se r2 mtitle star(* 0.1 ** 0.05 *** 0.01)
```

```
(output written to iv.rtf)
```

其中，“iv.rtf”的扩展名“rtf”表示“rich text format”。

点击输出结果中的“iv.rtf”链接，即可打开此文件，然后可在 Word 中继续编辑此文件。